Syllabus

Course Overview

Course Overview

This course provides you with both practical and theoretical aspects of data analytics using descriptive, inferential, diagnostic, and predictive statistical methods.

Today's business climate demands a proactive approach, going beyond Business Intelligence (BI) acquired through basic data mining. That new approach is analytics. This course introduces analytics, starting with the analytics lifecycle, basic data management techniques, and the use of descriptive statistics techniques. After an introduction to the concept of Key Performance Indicators (KPIs), you will explore the challenges of data sources, the use of ETL processes to improve data, and an elementary example of data mining. Then you will return to KPIs at a more sophisticated level, learning how they interact with data mining and more advanced analytical techniques. At the end of the course, you will produce both group and individual solutions to an analytics problem. Throughout the course, you will progress in your use of SAS for data analysis.

Becoming a data analyst allows you to be a part of the analytics revolution that is bringing data-driven solutions to broad sectors of society including, but not limited to, business, medicine, and education. This course is a survey course, intended to give you the training, time, and opportunities to begin learning SAS, while also gaining a high-level understanding of several different concepts in data analysis. Each of the topics in this course will be covered in more detail in future courses, so be sure to save all of your work and SAS code for use in future projects.

About Vila Health

In this program, a simulation of a fictitious hospital organization called Vila Health allows you to solve real problems with real analytical solutions. The media interactions can help you understand a business problem and the steps you may take to solve it. It also lets you practice the role of analyst and suggests how you can articulate a solution in a manner that others can understand.

Technology Resources

If you require the use of assistive technology or alternative communication methods to participate in course activities, please contact Disability Services to request accommodations.

Course Competencies

(Read Only)

To successfully complete this course, you will be expected to:

1. Apply data management fundamentals.
2. Describe the concepts of ETLs and data warehouses.
3. Revise the data mining concepts.
4. Evaluate applied analytics in professional domains.
5. Justify the use of resources available within a collaborative environment.
6. Communicate the decision-making process to stakeholders.

Course Prerequisites

Prerequisite(s): Completion of or concurrent registration in PM5331 or ANLT5002.

Syllabus >> Course Materials

Required

The materials listed below are required to complete the learning activities in this course.

Library

The following required readings are provided in the Capella University Library or linked directly in this course. To find specific readings by journal or book title, use Journal and Book Locator. Refer to the Journal and Book Locator library guide to learn how to use this tool.

- Anderson, A. (2014). Business Statistics For Dummies John Wiley & Sons: Skillsoft Collection
- Apgar, D. (2015). The false promise of big data: Can data mining replace hypothesis-driven learning in the identification of predictive performance metrics? *Systems Research and Behavioral Science, 32*(1), 28–49.
- Badawy, M., El-Aziz, A. A., Idress, A. M., Hefny, H., & Hossam, S. (2016.). A survey on exploring key performance indicators. *Future Computing and Informatics Journal.*
- Bell, P. C. (2015). Sustaining an analytics advantage [PDF]. *MIT Sloan Management Review, 56*(3), 21–24.
- Buglear, J. (2001). Stats Mean Business: A Guide to Business Statistics Taylor and Francis: Skillsoft Collection
- Gholami, R., Higón, D. A., & Emrouznejad, A. (2015). Hospital performance: Efficiency or quality? Can we have both with IT? *Expert Systems With Applications, 42*(12), 5390–5400.
- Goben, A., & Raszewski, R. (2015). The data life cycle applied to our own data. *Journal of the Medical Library Association, 103*(1), 40–44.
- Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Waltham, MA: Elsevier.
- Hui, E. (2019). Learn R for Applied Statistics: With Data Visualizations, Regressions, and Statistics Apress: Skillsoft Collection
- Larose, D. T., & Larose, C. D. (2014). *Discovering knowledge in data: An introduction to data mining* (2nd. ed.). Hoboken, NJ: Wiley.
- Lee, G. (2015). Business Statistics Made Easy in SAS SAS Institute: Skillsoft Collection
- Nelson, S.(2016). Excel Data Analysis For Dummies, 3rd Edition John Wiley & Sons: Skillsoft Collection
- Parmenter, D. (2015). *Key performance indicators (KPI): Developing, implementing, and using winning KPIs* (3rd. ed.). Hoboken, NJ: Wiley.
- Rivera, J., & Delaney, S. (2015). Using business analytics to improve outcomes. *Healthcare Financial Management, 69*(2), 64–67.
- Schmuller, J.(2017). Statistical Analysis with R For Dummies John Wiley & Sons: Skillsoft Collection
- Siemens, G. (2013). Learning analytics: The emergence of a discipline [PDF]. *American Behavioral Scientist, 57*(10), 1380–1400.

External Resource

Please note that URLs change frequently. While the URLs were current when this course was designed, some may no longer be valid. If you cannot access a specific link, contact your instructor for an alternative URL. Permissions for the following links have been either granted or deemed appropriate for educational use at the time of course publication.

- Anitha, J., & Babu, M. S. P. (2014). ETL work flow for extract transform loading [PDF]. *International Journal of Computer Science and Mobile Computing, 3*(6), 610–617. Retrieved from http://ijcsmc.com/docs/papers/June2014/V3I6201481.pdf
- Brown, B. (2015). SAS University Edition debuts new features. Retrieved from https://communities.sas.com/docs/DOC-19735
- Horner, P. (2015, December). Q&A: The state of INFORMS and the profession. *ORMS Today, 40*(6). Retrieved from https://www.informs.org/ORMS-Today/Public-Articles/December-Volume-40-Number-6/Q-A-The-state-of-INFORMS-and-the-profession
- SAS. (n.d.). SAS support communities. Retrieved from https://communities.sas.com/https://communities.sas.com/

Suggested

The following materials are recommended to provide you with a better understanding of the topics in this course. These materials are not required to complete the course, but they are aligned to course activities and assessments and are highly recommended for your use.

Optional

The following optional materials are offered to provide you with a better understanding of the topics in this course. These materials are not required to complete the course.

External Resource

Please note that URLs change frequently. While the URLs were current when this course was designed, some may no longer be valid. If you cannot access a specific link, contact your instructor for an alternative URL. Permissions for the following links have been either granted or deemed appropriate for educational use at the time of course publication.

- Python. (n.d.). Retrieved from https://www.python.org/
- R: The R Project for Statistical Computing. (n.d.). Retrieved from https://www.r-project.org/

Unit 1 ❯❯ Descriptive Analytics

Introduction

Case Study

Tiffany has moved from business process to the data analytics team at Vila Health. As she has been making the transition, so have a few other members of Vila Health. One of the challenges they have run into is that not all of the members of the newly formed team have a strong understanding of the underlying data analytics. As a starting point to get everyone on the same page, Tiffany would like the new team to review some of the basics of data analytics including the measure of central tendency and dispersion.

Analysts help make sense of what happened, what is happening, how often it happens on average, and what is the likelihood of something happening in the future. Furthermore, with some modeling techniques, an analyst can recommend what should happen in the future. This week you will focus on descriptive statistics, confidence intervals, and hypothesis tests, with an eye towards predictive analytics.

To-Do List:

- **Discussion:** Participate in a discussion to explore measures of central tendency, dispersion and association are all important when it comes to data analytics.
- **Assignment:** Calculate basic statistics for a given scenario and write an essay to answer the questions presented in the assignment.
- **Interactive Learning Module**: Complete an interactive module to learn how descriptive statistics are used to summarize data.
- **Interactive Learning Module:** Complete an interactive module to learn the details of the business problem for your virtual internship as a data analyst at Vila Health.
- **What You need to Know:** Study introductory concepts for the application of statistics in business.

- **Prepare:** Download a data analytics software such as SAS or R for use in the course. You will select one of these data analytics software and use the same software for the duration of this course.

Learning Activities

u01s1 - Activity Overviews

Discussion Overview

In the discussion for this week, you will explore that measures of central tendency, dispersion and association are all important when it comes to data analytics.

Assignment Overview

In this assignment for this week, you will calculate statistics for a given scenario and write an essay to answer the questions presented.

u01s2 - What You Need to Know

Business Statistics

In the Capella Library, read the following:

- Anderson, A. (2014). Business Statistics For Dummies John Wiley & Sons: Skillsoft Collection.
  - Chapter 2 focuses on various graphical representations of the data including histograms and analyzing the distribution of data. Chapter 3 focuses on the measure of central tendency.
  - Chapter 4 focuses on the measure of dispersion in a data set.
- Nelson, S.(2016). Excel Data Analysis For Dummies, 3rd Edition John Wiley & Sons: Skillsoft Collection.
  - Chapter 9: "Using the Statistics Functions". This chapter will cover how to use Excel to do some of the basic statistics functions including counting items, finding the mean, mode and median. In addition, other items including the standard deviation, variance, distributions and correlation is also covered. This will be quite useful with your assignment this week.
  - Chapter 10: "Descriptive Statistics", in particular, focus on the methods to create a histogram and how to rank items by percentiles.
- Buglear, J. (2001). Stats Mean Business: A Guide to Business Statistics Taylor and Francis: Skillsoft Collection.
  - Chapter 2 explores the different types of qualitative and quantitative data.
  - Chapter 3 explores the various measures. Note that in this text, they reference the measure of location for central tendency and measure of spread for dispersion.

u01s3 - Prepare: Software Preparation and Technology Access

In this course, you will be using software and technology that is needed to complete designated activities and assignments. There is no additional cost for this software and technology. Some software packages will be made available to you at no additional cost through Capella's subscription with Microsoft, while other software packages are available for free download through open-source licensing.

Capella University requires learners to meet certain minimum computer requirements. Please note that some software required for a course may exceed these minimum requirements. Check the requirements for the software you may need to download and install to make sure it will work on your device. Most software will require a Windows PC. If you use a Mac, refer to Installing a Virtual Environment and Windows on a Mac.

The software and technologies below are strongly recommended to support you in completing the course objectives. If you have access to other tools that you believe may still meet the requirements of this course, please discuss your selected alternatives with your instructor.

If you use assistive technology or any alternative communication methods to access course content, please contact Disability Services with any access-related questions or to request accommodations.

For this course, follow the instructions provided through the links below to download and install software or register for an account, as required.

SAS Statistical Software

SAS OnDemand for Academics (SODA).

- Download the SAS data files for use in the assignments: ANLT5010 Data Files [ZIP].
- Open the file and take some time to explore this dataset to see how it was constructed.

Open Source Statistical Software

R and Python are the two open source software that can be applied as an alternative to SAS to complete the assignments in this course:

- R: Go to the Download page of the Getting Started section of The R Project for Statistical Computing home page to download the latest version of R.
- Python: Go to the Download section of the Python Beginners Guide to download the latest version of Python.

Selecting a Statistical Software

There are three commonly used statistical software for data analytics. R and Python are open source statistical software that are free. SAS offers commercial statistical software to purchase, and offers free or low cost statistical software to academic, noncommercial users. R and Python provide support through their open source web-based platforms. SAS provides support provide through several channels including telephone and web-based resources. You may already use one of these statistical software where you work or for personal use. For this course you need to select a statistical software for data analytics and plan to use the same software for all of the assignments throughout this course. Either SAS, R, or Python software can be used to complete the assignments in this course.

Microsoft Software

1. Visit Capella's Microsoft Software page for instructions on obtaining free Microsoft software.
2. Identify the version of MS Visio / Project / Access / Visual Studio Enterprise / SQL Server / Etc. that is compatible with your operating system.
3. Download and install.

Note: As a Capella learner, you have access to IT online resources through Capella's Skillsoft subscription, where you can find helpful materials

u01v1 - Interactive Learning Module: Descriptive Statistics

A population is the complete set of data that we want to understand and learn about. However, this population can be quite large. When this is the case, we tend to analyze a subset of that data called a sample. For example, if we wanted data on the starting salaries of Capella graduates in their degree field for all of the learners that graduated from Capella last year it would be quite time-consuming to take a census of the entire population. We could however take a sample of those graduates across all programs and obtain and record the starting salaries from the sample to help describe the entire population.

Data collected based on categories can generally be summarized based on the frequency or the percentage of items in each category. Numerical data can be represented by counts or measures. Actual counts with underlying meaning can be summarized using measures beyond just the count or percentage. Some of these measures include summarized data based on:

- Measures of central tendency:
  - Mean
  - Mode
  - Median
- Measures of spread or dispersion:
  - Range
  - Variance
  - Standard deviation
- Measure of association:
  - Covariance
  - Correlation

This interactive learning module will help you review measures used for descriptive statistics

u01v2 - Interactive Learning Module: Analytics Internship - Data Types and Sources

An integral part of this course and Capella University's Data Analytics program is your virtual internship with Vila Health, a fictional health care system operating hospitals and other health care facilities throughout the upper Midwestern United States. In this first challenge, you will learn the details of the business problem that will be the focus for your report in the coming weeks.

Course Resources

Analytics Internship: Data Types and Sources | Transcript

u01d1 - Write Your Discussion Post

Measures

Understanding the measures of central tendency, dispersion and association are all important when it comes to data analytics. Choose one of the three measures and explain how the measure that you have chosen is important to describe the data. Provide an example of how to calculate and determine this measure.

Response Guidelines

Respond to at least two other learners and share with them the information in their initial posts that best helped you in understanding the concepts.

Course Resources

Graduate Discussion Participation Scoring Guide

u01s4 - Prepare: Programming Instructions

Using R

Access the Using R page on Campus for resources on:

- How to get started with R.
- How to read in a file in R.
- Getting started with Descriptive Statistics in R.

Using Python

Access the Using Python page on Campus for resources on:

- How to get started with Python.
- How to read in a file in Python.
- Getting started with Descriptive Statistics in Python.

Using SAS

Access the Using SAS page on Campus for resources on:

- How to get started with SAS.
- How to read in a file in SAS.

- Getting started with Descriptive Statistics in SAS.

u01a1 - Basic Statistics

Overview

In this assignment, you will calculate statistics for this scenario and write an essay to answer the questions presented. Include both your calculations and your essay in your assignment submission.

Scenario

Vila Health has been working on improving patient wait times at Cengaly Hospital. Typical wait times during busy periods under the current system have been around 10 minutes. Vila Health is hoping that with the new system, they will be able to lower the typical waiting times to less than 6 minutes and would like to evaluate the new system as it has been put in place for 6 months and has been operating consistently. You would like to take a sample of 100 patients during peak hours. For this assignment, you will be working in the Discovery phase of the data analytics lifecycle.

Instructions

Calculations

For this assignment use cf_ANLT_W1_WaitingTimes.csv file with a the dataset of randomly selected times across multiple days to complete the following calculations.

1. Calculate the mean, median and mode for the patient waiting times:
    - How do they compare?
    - Is it possible patient waiting times during peak hours is less than six minutes?
2. Calculate the standard deviation of the sample of 100 patients.
3. Using the standard deviation, calculate the estimates of tolerance intervals of all the patient waiting times containing 68.26%, 95.44% and 99.73% of all patient waiting times.

Essay

Examine the histogram of the data in figure 1 for this scenario to address the following questions in your essay.

Figure 1. Histogram of patient waiting times for Cengaly Hospital

1. Explain if the Empirical rule can be used to describe the patient waiting times including why or why not.

2. Does the estimate of a tolerance level of 68.26% of all patient waiting times provide evidence that at least two-thirds of all patients will have to wait less than 8 minutes? Explain your answer.

Your assignment will be graded on the following criteria:

- Calculate and compare the mean, median and mode.
- Discuss the patient waiting times based on the measure of central tendency.
- Calculate the standard deviation and estimated tolerance intervals of all of the patient waiting times.
- Explain the use of the Empirical rule and tolerance levels.
- Use communication style and vocabulary appropriate for the target audience.

Refer to the Data Management Fundamentals Scoring Guide for details.

Submission Requirements

- **Written communication**: Written communication is free of errors that detract from the overall message.
- **Length of paper**: 3–5 typed double-spaced pages.
- **Resources:** At least three scholarly resources. Include a reference page at the end of the paper.
- **APA guidelines:** Double-spaced paragraph formatting in the body of the paper. When appropriate, use APA-formatted headings. Resources and citations are formatted according to current APA style and format.
- **Font and font size:** Times New Roman, 11 point font.

Competencies Measured

By successfully completing this assignment, you will demonstrate your proficiency in the following course competencies and assignment criteria:

- **Competency 1: Apply data management fundamentals.**
  - Calculate the standard deviation and estimated tolerance intervals of all of the patient waiting times.
  - Calculate and compare the mean, median and mode.
- **Competency 4: Evaluate applied analytics in professional domains.**
  - Explain the use of the Empirical rule and tolerance levels.
  - Discuss the patient waiting times based on the measure of central tendency.
- **Competency 6: Communicate the decision-making process to stakeholders.**
  - Use communication style and vocabulary appropriate for the target audience.

Unit 2 » Data Analytics

Introduction

Case Study

Tiffany has been pleased with the progression of the data analytics team with how they have been able to understand the basic business statistic rules. The next phase that Tiffany would like the team to learn is, more about the measures of association. As usual, as it may be to understand the measure of central tendency and dispersion, Tiffany understands that it is important to define how various data points are related. For example, Vila Health may want to know how if patient length of stays are directly related to the geographical area that the patient is in.

For this week, you will continue to work with Cengaly Hospital data to further analyze aspects to measure association, skewness and kurtosis. These measures of association are important as they help define how closely two variables follow each other.

To-Do List

- **Discussion:** Examine correlation and create a hypothesis between two variables with and explanation of how you would measure the correlation between the variables.
- **Assignment:** Calculate the frequency distribution and correlation statistics for a given scenario and write an essay to answer the questions presented in the assignment.
- **What You need to Know:** Study the concepts and programming steps used in the application of measures of association in descriptive statistics. **Additional Statistics**: Study the concepts of skewness and kurtosis of data set distributions in the application of descriptive statistics.

Learning Activities

u02s1 - Activity Overviews

Discussion Overview

In the discussion for this week, you will examine correlation and create a hypothesis between two variables with an explanation of how you would measure the correlation between the variables.

Assignment Overview

In the assignment for this week, you will calculate statistics for a given scenario and write an essay to answer the questions presented.

u02s2 - What you Need to Know

Measures of Association

Read the following from the Capella Library:

- Anderson, A. (2014). Business Statistics For Dummies John Wiley & Sons: Skillsoft Collection.
  - Chapter 5: Measuring How Data Sets are related to Each Other. Focus on methods to calculate the covariance and correlation as well as how to interpret the correlation coefficient.
- Nelson, S.(2016). Excel Data Analysis For Dummies, 3rd Edition John Wiley & Sons: Skillsoft Collection.
  - Chapter 9, Using the Statistics Functions. Focus on the normal distributions, t-distributions and f-distributions as well as the correlation.
- Buglear, J. (2001). Stats Mean Business: A Guide to Business Statistics Taylor and Francis: Skillsoft Collection.
  - Chapter 4. Chapter 4 goes into correlation and regression as well as how to summarize data collected over time.

Additional Descriptive Statistics

Skewness

Skewness describes the distribution of a data set. With a data set, we will typically have symmetrical or skewed data. When the data set is symmetric is when we have the mean to equal or close to the median and mode. If the data set is skewed, then, the mean is not equal or close to the median or mode. Note that we can have a skewed distribution that is positively or negatively skewed.

- A positively skewed distribution will have a larger value of mean and a smaller value of mode where the median is between the mean and mode.
- A negatively skewed distribution will have a smaller value of mean, a larger value of mode and the median having a value between the two.

Figure 1. Skewness

Looking at the graph, we can see that the positively (or right skewed distribution) has the right side tail of the distribution curve longer and most the data value are distributed on the left side of the distribution curve. When it comes to the skewness coefficient on a dataset, we can interpret as the following:

- If the skewness coefficient is equal to 0, the distribution is symmetric.
- If the skewness coefficient is greater than 1, the distribution is strongly skewed positively to the right.
- If the skewness coefficient is less than -1, the distribution is strongly skewed negatively to the left.
- If the skewness coefficient is between -0.5 and -1, the distribution is moderately skewed negatively to the left.
- If the skewness coefficient is between 0.5 and 1, the distribution is moderately skewed positively to the right.
- If the skewness coefficient is between -0.5 and 0.5, the distribution is approximately symmetrical.
- The skewness measure the degree of asymmetry of a probability distribution.
  - In the example of a symmetrical distribution, the values are evenly distributed above and below the mean.

Kurtosis

Kurtosis describes the flatness or amount of peakedness of the distribution for a data set. Kurtosis measures the likelihood of extreme outcomes in a distribution relative to the normal distribution. For a normal distribution, the value of the kurtosis measure is 0.

Figure 2. Kurtosis

When interpreting kurtosis, we can look at the following:

- If the distribution is neither flat nor peaked at the center, it is called a normal, symmetric or mesokurtic curve.
- If the distribution is flatter or has a lower and broader peak with thinner and shorter tails, it is called a platykurtic curve.
  - This is the case when there is less area.
- If the distribution is more peaked and has thicker and longer tails than a normal curve, it is called a leptokurtic curve.
  - This is because more area in the tails of the distribution and less is near the mean.

u02s3 - Prepare: Programming Instructions

Using R

Access the Using R page on Campus for resources on:

- Creating a Dataset/Data Cleaning in R.
- Getting started with Descriptive Statistics in R.

Using Python

Access the Using Python page on Campus for resources on:

- Creating a Dataset/Data Cleaning in Python.
- Getting started with Descriptive Statistics in Python.

Using SAS

Access the Using SAS page on Campus for resources on:

- Creating a Dataset/Data Cleaning in SAS.
- Getting started with Descriptive Statistics in SAS.

u02d1 - Write Your Discussion Post

Correlation

For this week's initial post, address the following:

Correlation helps to define the association between two variables. There can be assumptions in place to define how correlated two variables are but calculating their correlation can help ensure that the assumptions are correct. Consider your field and create a hypothesis between two variables and explain how you would measure the correlation between the variables. For example, we could have a hypothesis that there's a correlation between speed limits and highway accidents. How would we be able to validate that hypothesis?

Response Guidelines

Respond to at least two other learners and share with them the information in their initial posts that best helped you in understanding the concepts.

Course Resources

Graduate Discussion Participation Scoring Guide

u02a1 - Additional Statistics

Overview

In this assignment, you will calculate statistics for this scenario and write an essay to answer the questions presented. Include both, your calculations and your essay in your assignment.

Scenario

Vila Health has been working on a new billing system to be deployed to all hospitals. Before they release the billing system to all hospitals, they have deployed it to Cengaly Hospital first. With the older billing system, the hospital system mostly had payments from their patients taking 40 days or more. Vila Health hopes that with the new billing system, it will substantially reduce the amount of time it takes patients to make their payments.

There is a random sample of 65 invoices from 15025 invoirces that were processed in the first two months of the new system's operation. If the sample can be used to establish that the new billing system substantially reduces payment time, Vila Health does plan to quickly deploy it to the other hospitals within their system.

By looking at the payment times, we can see that the shortest payment time is 10 days and that the longest is 29 days. This does not tell us much information so it is important to perform a frequency distribution of the data and then graph the distribution by constructing a histogram.

Instructions

Calculations

For this assignment, the dataset is available as cf_ANLT_W2_InvoiceDayAndAmount.csv of the invoice payment days and the amount:

1. Determine the frequency distribution and construct the histogram:
    - Find the number of classes.
    - Find the class length.
    - Form the non-overlapping classes of equal width by defining the boundaries of the classes.
    - Tally and count the number of measurements in each class.
    - Graph the histogram.
2. Determine the skewness and kurtosis of the payment days.
3. Calculate the correlation between the payment days and the amount owed.

Essay

1. Explain how the correlation calculation shows the correlation between payment days and the amount owed.
2. Explain if the new billing system is more efficient than the old billing system based on the sample of the new system.

Your assignment will be graded on the following criteria:

- Create the frequency distribution and histogram.
- Explain the skewness and kurtosis on the sample data set.
- Calculate the correlation of two variables.
- Explain how the correlation calculation shows the correlation between two variables.
- Explain if one billing system approach is more efficient than another based upon statistical analysis of a data set.
- Provide a logical argument in support of conclusions or recommendations.

Refer to the Additional Statistics Scoring Guide for details.

Submission Requirements

- **Written communication**: Written communication is free of errors that detract from the overall message.
- **Length of paper**: 3–5 typed double-spaced pages.
- **Resources:** At least three scholarly resources. Include a reference page at the end of the paper.
- **APA guidelines:** Double-spaced paragraph formatting in the body of the paper. When appropriate, use APA-formatted headings. Resources and citations are formatted according to current APA style and format.
- **Font and font size:** Times New Roman, 11 point font.

Competencies Measured

By successfully completing this assignment, you will demonstrate your proficiency in the following course competencies and assignment criteria:

- **Competency 1: Apply data management fundamentals.**
    - Explain how the correlation calculation shows the correlation between two variables.
    - Calculate the correlation of two variables.
    - Explain the skewness and kurtosis on the sample data set.
    - Create the frequency distribution and histogram.
- **Competency 4: Evaluate applied analytics in professional domains.**
    - Explain if one billing system approach is more efficient than another based upon statistical analysis of a data set.
- **Competency 6: Communicate the decision-making process to stakeholders.**
    - Provide a logical argument in support of conclusions or recommendations.

Unit 3 >> Data Identification

Introduction

Case Study

Up to this point, Tiffany has mostly focus on having the team analyze fairly simple data sets. However, Tiffany knows that her team needs to become more versed in using industry standard languages and tools to analyze much larger data sets. Although there are many choices available, SAS, R and Python are some of the most widely used options to analyze big data. She would like her team to start using some of those choices and implementing them.

This week introduces getting started using SAS or R and provides instructions for data analysis project set-up and programming steps for fundamental data management using either of these data analytics software. The readings for this week provide additional information on data quality, analytic and data lifecycles, and the importance of understanding the data—specifically, understanding the quality issue and source of the data itself. Data quality may be compromised during data collection. Missing data, erroneous data, and erroneously formatted data are some factors of poor-quality data. Data sources can be from external or internal systems. Learning how data can be purified and integrated from multiple sources and used to meet the objectives of the analytics project is one of the biggest challenges for an analyst.

To-Do List:

- **Discussion:** Explore SAS and R data analytics software and provide arguments for using with one or the other, or using a completely different software for data analytics.
- **Assignment:** Write data analysis code for fundamental data management and write an executive summary to answer the questions presented in the assignment.
- **What You need to Know:** Conduct initial programming steps to become familiar with your selected data analytics software, and study the data analytics life cycle.

Learning Activities

u03s1 - Activity Overviews

Discussion Overview

In the discussion for this week, you will explore SAS and R data analytics software and provide arguments for using with one or the other, or using a completely different software for data analytics.

Assignment Overview

In the assignment for this week, you will write data analysis code and write an executive summary to answer the questions presented.

u03s2 - What You Need to Know

Introduction to SAS and R

SAS

Use the Capella Library to read the following if you are choosing to use SAS for your coding:

- Lee, G. (2015). Business Statistics Made Easy in SAS SAS Institute 2015 : Skillsoft Collection
  - Chapter 5: Introduction to SAS
  - Chapter 6: Basics of SAS Programs, Data Manipulation, Analysis and Reporting. Although the earlier chapters can be useful to review, the core around SAS should start here. Pay specific attention to the three big tasks in Business Statistics.

R

Use the Capella Library to read the following if you are choosing to use R for your coding:

- Hui, E. (2019). Learn R for Applied Statistics: With Data Visualizations, Regressions, and Statistics Apress: Skillsoft Collection
  - Chapter 2 covers the IDE for R and how to install and setup RStudio.
  - Chapter 3 goes through some of the basic syntax of R.
  - Chapter 4 is focused on descriptive statistics and should be your main focus to become familiar with R.
- Schmuller, J.(2017). Statistical Analysis with R For Dummies John Wiley & Sons: Skillsoft Collection
  - Chapters 3-5 focus on the coding aspects and functions to calculate the measures for the assignment.

Data Analytics Lifecycle and Data Identification

This week you will study the concept of the analytic lifecycle. In general, lifecycle is divided into eight areas:

- Problem Identification and Definition.
- Design and Build.
- Data Acquisition.
- Exploration and Reporting (Visualization).
- Modeling (Churn Model, Risk Scoring Model).
- Actionable Analytics.
- Feedback.

The analytic lifecycle greatly depends on data. Data itself has a lifecycle, for which the analytic lifecycle adapts. Both lifecycles start with the understanding of the objective, goal, and problem. Data quality is the determinant factor in value and applicability of the analytic method and the usability and applicability of the resulting recommendation and course of action.

The analytic lifecycle typically starts with problem analysis and then moves to data analysis, data collection, and data preparation. For this reason, the analyst must understand the data and its quality, where the data came from, and how the data works together from different data sources before creating the solution. It is interesting how different disciplines and industries have a similar analytic lifecycle approach.

Use the Capella University Library to read the following:

- Siemens, G. (2013). Learning analytics: The emergence of a discipline [PDF]. *American Behavioral Scientist, 57*(10), 1380–1400.
  - This article provides a view of analytical project organization and scope from an academic researcher's perspective. An analytical model highlights what and how data is sourced, used, and measured to include a feedback loop. A data source model provides examples of internal and external source systems.
- Goben, A., & Raszewski, R. (2015). The data life cycle applied to our own data. *Journal of the Medical Library Association, 103*(1), 40–44.
  - This article provides an overview of a data analytical lifecycle from two librarian scientists' perspectives.

You may view the following walkthrough to help you understand the life cycle of analytics and data identification:

- Life Cycle of Analytics and Data Identification Walkthrough.

u03s3 - Prepare: Programming Instructions

Using R

Access the Using R page on Campus for resources on:

- Using Confidence Intervals in R.
- Hypothesis Testing in R.

Using Python

Access the Using Python page on Campus for resources on:

- Using Confidence Intervals in Python.
- Hypothesis Testing in Python.

Using SAS

Access the Using SAS page on Campus for resources on:

- Using Confidence Intervals in SAS.
- Hypothesis Testing in SAS.

u03d1 - Write Your Discussion Post

SAS or R

For this week's initial post, consider the following decision. For the rest of the course, you have the option to use SAS or R to perform data analytics.

- From an organizational standpoint, what arguments would you make for using with one or the other, or using a completely different software for data analytics?

Response Guidelines

Respond to at least two other learners and share with them the information in their initial posts that best helped you in understanding the concepts.

Course Resources

Graduate Discussion Participation Scoring Guide

u03a1 - Data Management Fundamentals

Overview

In this assignment, you will write data analysis code and write an executive summary to answer the questions presented. Include both your data analysis code and your executive summary in your assignment submission.

For this assignment, you will be working in the Discovery phase of the data analytics lifecycle. You will create statistical analysis software code and a process or plan to identify the sources of data for a data analytics project, and begin to evaluate the quality of that data. In this assignment process, you will create a reusable data cleansing strategy for use in statistical analysis software. The univariate analysis function generally includes minimum, maximum, mean, median and other variable details.

Instructions

Data Analysis Programming

For this assignment, you will be using the cf_ANLT5010_W3_Grades.csv data set that is available in the course files. Explore the existing data and conduct a quality control check on missing data, erroneous data, and data with an incompatible format.

1. Using the **cf_ANLT5010_W3_Grades.csv** data set, complete the data analysis programming steps as listed in the table.
2. Identify a strategy for data cleansing by explaining the steps you took, in the form of a reusable data cleansing strategy or plan, including any code and instructions.

Data Analysis Programming Steps

If you are using SAS:	**If you are using R:**
• **Import the data into SAS EM.**	• Load the data into R (use the read.csv() or read.csv2() functions).
• **Print the first five observations using PROC.**	• Print the first five observations using the head function.
• **Print, and note if anything looks out of place.**	• Note if anything looks out of place.

If you are using SAS:	If you are using R:
• **Run a PROC Frequency (PROC Freq) on at least one of the qualitative variables.**	• Run the frequency (freq() function) on at least one of the qualitative variables.
• **Run a PROC Univariate on the quantitative variable and summarize your findings.**	• Run the univariate (univariate() function) on the quantitative variable and sumamrize your findigns.

Executive Summary:

1. Write an executive summary of your work that includes the following items:
   - Identify the sources of data and attributes that are available to you using the data set provided.
   - Define the data set and variables, including any attributes about the variables that you have discovered.
   - Describe the process used to explore the existing data set and identify what variables, and associated attributes, exist.
   - Explain your strategy for data cleansing in case of quality issues.
   - Introduce your preferred method of working with raw data and which type of data sets, formats, you prefer to work with.
2. Describe your strengths and opportunities to learn using tools to import and work with multiple data sources.
3. Include as an appendix to your executive summary the output of your data analysis programming code.

Your assignment will be graded on the following criteria:

- Identify data sources and attributes.
- Select methods for working with raw data of different types.
- Address the issue of data quality.
- Discuss personal strengths and weaknesses in working with data sources.
- Create an executive summary with appropriate focus and level of detail.

Refer to the Data Management Fundamentals Scoring Guide for details.

Submission Requirements

- **Written communication**: Written communication is free of errors that detract from the overall message.
- **Length of paper**: 2–3 typed double-spaced pages.
- **Resources:** At least three scholarly resources. Include a reference page at the end of the paper.
- **APA guidelines:** Double-spaced paragraph formatting in the body of the paper. When appropriate, use APA-formatted headings. Resources and citations are formatted according to current APA style and format.
- **Font and font size:** Times New Roman, 11 point font.

Competencies Measured

By successfully completing this assignment, you will demonstrate your proficiency in the following course competencies and assignment criteria:

- **Competency 1: Apply data management fundamentals.**
  - Select methods for working with raw data of different types.
  - Identify data sources and attributes.
  - Address the issue of data quality.
- **Competency 4: Evaluate applied analytics in professional domains.**
  - Discuss personal strengths and weaknesses in working with data sources.
- **Competency 6: Communicate the decision-making process to stakeholders.**
  - Create an executive summary with appropriate focus and level of detail.

Unit 4 >> Data Manipulation

Introduction

Case Study

For much of the data that Tiffany has shared with the team, the data has been cleansed already. Having the data cleansed means that the data has gone through various checks and balances to ensure that the data is free from errors. These checks are helpful when evaluating the compatibility of data formats and fields from different data sources and when identifying issues within data sources that require cleansing.

These data issues may be the result of data entry errors, other transcription errors, or even simply the relatively frequent changes in the demographic information that naturally occurs for customers. She would like her team to focus on identifying those types of errors and recommending solutions for them.

For this week, you will continue to work with some basic data quality checks using data analytical software.

To-Do List:

- **Discussion:** Examine types of regression and consider what preference is given to quantitative versus qualitative attributes when choosing proper variables for a regression-type data analysis approach.
- **Assignment:** Write data analysis code to apply the data cleansing methods methods you learned last week to check for data quality and write an essay to answer the questions presented in the assignment.
- **What You need to Know:** Study concepts and coding for conducting data quality checks and data cleansing.

Learning Activities

u04s1 - Activity Overviews

Discussion Overview

In this week's discussion, you will examine types of regression and consider what preference is given to quantitative versus qualitative attributes when choosing proper variables for a regression-type data analysis approach.

Assignment Overview

In the assignment for this week, you will write data analysis code to apply the methods you learned last week to check for data quality and write an essay to answer the questions presented.

u04s2 - What You Need to Know

Data Quality Checks

Data quality checks are helpful when evaluating the compatibility of data formats and fields from different data sources and when identifying issues within data sources that require cleansing. You will continue to focus on data quality and data cleansing using data analytical software by researching options for addressing the data issues that you find as part of your data quality checks.

You will also compare methods for handling customer demographics attributes, which are commonly full of data issues. These data issues may be the result of data entry errors, other transcription errors, or even simply the relatively frequent changes in the demographic information that naturally occur for customers. You will identify the types of errors that may arise in this type of data and recommend solutions for handling these errors.

SAS

In the Capella Library, read the following if you are choosing to use SAS for your coding:

- Lee, G. (2015). Business Statistics Made Easy in SAS SAS Institute 2015 : Skillsoft Collection.
  - Chapters 7–9. Chapter 9 in particular will cover the basic statistics to check and fix data if there are issues including missing data.

R

In the Capella Library, read the following if you are choosing to use R for your coding:

- Hui, E. (2019). Learn R for Applied Statistics: With Data Visualizations, Regressions, and Statistics Apress: Skillsoft Collection.
  - Chapters 4 and 5. Chapter 4 in particular will help with some of the common errors and issues with the data.
- Schmuller, J.(2017). Statistical Analysis with R For Dummies John Wiley & Sons: Skillsoft Collection
  - Chapters 6–8. These chapters will looking some of the standard scores and being able to summarize all of the data. These chapters also explain the details of what normal is when it comes to coding.

u04v1 - Interactive Learning Module: Analytics Internship - ETL and Data Warehousing

In this Vila Health activity, you will have the opportunity to speak with key stakeholders from Clarion Court in order to learn more about the data you will need to complete your assignments. The Vila Health director of IT will also be available to provide insight into the structure of the Vila Health data warehouse.

Course Resources

Analytics Internship: ETL and Data Warehousing | Transcript

u04d1 - Write Your Discussion Post

Regression

Regression is the broad term used to describe a way to measure relationships between variables. The simplest case of regression is simple linear regression, which is used to compare the relationship of one quantitative variable to another quantitative variable. More complex applications of regression include applications of logistic regression, multiple regression, and multivariate regression.

- **Quantitative versus qualitative analysis:** Considering all the types of regression mentioned above, in choosing proper variables that an analyst considers for a regression-type approach:
  - What preference is given to quantitative versus qualitative attributes?
  - Can both types of variables be considered in one model? If so, how?

Response Guidelines

Respond to at least two other learners and share with them the information in their initial posts that best helped you in understanding the concepts.

Course Resources

Graduate Discussion Participation Scoring Guide

u04a1 - Data Cleansing

Overview

In this assignment, you will write data analysis code to apply the methods you learned last week to check for data quality and write an essay to answer the questions presented. You will use statistical software to test the data cleansing plan you created last week on real data in this week's assignment. Include both your data analysis code and your essay in your assignment submission.

Instructions

Data Analysis Programming

For this assignment, you will use the data set cf_ANLT5010_W4_Airbnb.csv found in the course files.

1. Using the **cf_ANLT5010_W4_Airbnb.csv** data set, complete the data analysis programming steps as listed in the table.
2. For each issue that you found in the data, use the data step as appropriate to fix the data issues. Save your code to include in the data cleansing strategy in your assignment and also for your own future reference.

Data Analysis Programming Steps

If you are using SAS	If you are using R
• **Import the data into SAS EM.**	• Load the data into R (use the read.csv() or read.csv2() functions).
• **Print the first five observations using PROC Print, and note if anything looks out of place.**	• Print the first five observations using the head function, and note if anything looks out of place.
• **Run a PROC Frequency (PROC Freq) on at least one of the qualitative variables.**	• Run the frequency (freq() function) on at least one of the qualitative variables.
• **Run a PROC Univariate or a PROC Means on the quantitative variable and summarize your findings**	• Run the univariate (univariate() function) on the quantitative variable and summarize your findings.

Essay

1. Write a 3–5 page summary paper that discusses how you applied your data cleansing strategy or plan, what types of data issues your cleansing strategy identified, and the gaps in your current strategy. Include a section on what you feel are your strengths and where you feel you have an opportunity to learn using data analytical tools to import and work with multiple data sources.
2. Include as an appendix of your essay the results of your data cleansing reports in whatever format you detailed in your Week 3 plan, for example an output or a summary table, your code, and your written evaluation.

Your assignment will be graded on the following criteria:

- Describe methods for working with raw data of different types.
- Apply integration of different types of data cleansing.

- Discuss personal strengths and weaknesses in working with data sources.
- Organize content so ideas flow logically with smooth transitions.

Refer to the Data Cleansing Scoring Guide for more details.

Additional Requirements

- **Written communication:** Written communication is free of errors that detract from the overall message.
- **Length of paper**: 3–5 typed double-spaced pages.
- **Resources:** At least three scholarly resources. Include a reference page at the end of the paper.
- **APA guidelines:** Double-spaced paragraph formatting in the body of the paper. When appropriate, use APA-formatted headings. Resources and citations are formatted according to current APA style and format.
- **Font and font size:** Times New Roman, 11 point font.

Competencies Measured

By successfully completing this assignment, you will demonstrate your proficiency in the following course competencies and assignment criteria:

- **Competency 1:Apply data management fundamentals.**
  - Describe methods for working with raw data of different types.
  - Discuss personal strengths and weaknesses in working with data sources.
- **Competency 2: Describe the concepts of ETLs and data warehouses.**
  - Apply integration of different types of data cleansing.
- **Competency 6: Communicate the decision-making process to stakeholders.**
  - Organize content so ideas flow logically with smooth transitions.

Course Resources

Academic and Professional Document Guidelines [PDF]

Unit 5 >> Confidence Levels

Introduction

Case Study

Tiffany has been having quite a bit of success getting her team prepared with data analytics. She frequently hears about her team members not entirely understanding confidence intervals and how confidence intervals are used to determine the true mean of the population of a data set and the reliability of the data. The confidence interval helps the data analyst determine the variable characteristics of a data set. Confidence intervals are also used in hypothesis testing.

It is important to note that the confidence interval is not the range and that it contains 95% of the values when the confidence level is set at 95%. It is correct to note that the calculated confidence interval contains the true population mean. The reason for this is that the population mean only has one value. The confidence interval computed depends upon the data collected and the sample size. If the data sampling was repeated, the confidence interval would almost certainly be different.

Tiffany also brings up the fact that in most cases, 95% confidence level is used just because of tradition but there is nothing special about that number. In some cases 90% or 99% confidence level may be applied. Lower confidence levels such as 90%, allows more variation in a data set and more leniency in hypothesis testing.

The computerization of our society has substantially enhanced our ability to generate and collect data from different sources. As a result, we encounter a deluge of data—transactional data, medical data, demographic data, finance data, and marketing data. In order to generate value from this data, we must classify, summarize, and analyze it to discover trends or anomalies. Confidence intervals provide one means to analyze trends and anomalies in a data set. An inferential statistical technique, hypothesis testing, provides a way to test for statistically significant differences between a parameters, such as average sales, and a constant, such as the break-even point, or the difference between a parameter for one population, such as average sales for Store A, versus another population, such as average sales for Store B.

To-Do List:

- **Discussion:** Report on the progress of your Vila Health report and analytics solution proposal that is due in Week 10 of this course.
- **Assignment:** Write data analysis code to determine confidence intervals and conduct hypothesis testing, and write an essay analysis to answer the questions presented.
- **What You need to Know:** Study the concepts and application of confidence levels in descriptive statistics.

Learning Activities

u05s1 - Activity Overviews

Discussion Overview

In the discussion for this week, you will report on the progress of your Vila Health report and analytics solution proposal that is due in Week 10 of this course.

Assignment Overview

In the assignment for this week, you will write data analysis code and write an essay analysis to answer the questions presented.

u05s2 - What You Need to Know

Confidence Levels

View the following walkthrough to help you understand the application of confidence levels in descriptive statistics:

- Descriptive Statistics and Confidence Levels Walkthrough.

From Capella Library, read the following:

- Anderson, A. (2014). Business Statistics For Dummies John Wiley & Sons: Skillsoft Collection.
  - Chapter 11: Confidence Intervals and the Student's t-Distribution. This will help describe the need for the estimated range of the values.

SAS

From Capella Library, read the following if you are choosing to use SAS for your coding:

- Lee, G. (2015). Business Statistics Made Easy in SAS SAS Institute: Skillsoft Collection.
  - Chapter 7 with the focus on assessing the shape of the variables's distribution.
  - Chapter 10 will focus on various methods of graphing in SAS.

R

From Capella Library, read the following if you are choosing to use R for your coding:

- Schmuller, J.(2017). Statistical Analysis with R For Dummies John Wiley & Sons: Skillsoft Collection.
  - Chapter 9. In particular, focus on the central limit theorem and the confidence limits.

u05d1 - Write Your Discussion Post

Vila Health Progress Report

You should be working through your Vila Health final report that is due in Week 10 of this course. Although you do have until week 10 to complete the work, answer the following questions:

- Share with the class what you've been able to complete so far.
- What challenges have you run into with the data?
- What questions do you have that you need answers to?
- What suggestions or tips would you have for others?

Response Guidelines

Respond to at least two other learners and share with them the information in their initial posts that best helped you in understanding the concepts.

Course Resources

Graduate Discussion Participation Scoring Guide

u05a1 - Statistical Analysis of Database Solutions

Overview

In this assignment you will write data analysis code and write an essay analysis to answer the questions presented. Include both your data analysis code and your essay analysis in your assignment submission. Once you understand what values the variables in your data contain and identify potential data issues and solutions, you can begin basic statistical analyses.

Descriptive statistical analysis can reveal more about the distribution of valid values a variable can take on as opposed to simply identifying erroneous or missing values. Moving on from descriptive analysis, you can also estimate average values and proportions, such as average sales in this assignment, for example. Using an additional inferential technique, hypothesis testing, you can test for statistically significant differences between a parameter, such as average sales, and a constant, such as the break-even point, or the difference between a parameter for one population, such as average sales for Store A, versus another population, such as average sales for Store B.

Instructions

Data Analysis Programming

Use the data set cf_ANLT5010_W5 Sales_Data.csv for this assignment from the ANLT5010 Data Files (ZIP) package that you downloaded in Week 1:

1. Using the **cf_ANLT5010_W5 Sales_Data.csv** data set, complete the data analysis programming steps as listed in the table.

Data Analysis Programming Steps

If you are using SAS	If you are using R
• **Use measures of central tendency (use the PROC univariate) and variability to create a descriptive analysis of the daily sales for each store.**	• Use measures of central tendency (use the univariate function) and variability to create a descriptive analysis of the daily sales for each store.
• **Create a 95% confidence interval (by default, this is set at 95%) for the average daily sales for each store (Store A and tore B) for the year.**	• Create a 95% confidence interval for the average daily sales for each store (Store A and Store B) for the year.
• **Choose a confidence level for your hypothesis tests. (Common confidence levels include 90%, 95%, and 99%, but you can choose a different level if you desire.)**	• Choose a confidence level for your hypothesis tests. (Common confidence levels include 90%, 95%, and 99%, but you can choose a different level if you desire.)

If you are using SAS	If you are using R
- **Conduct a test of hypothesis to determine whether Store A's average daily sales for the year significantly exceeded the break-even threshold of $4900.**	- Conduct a test hypothesis to determine whether Store A's average daily sales for the year significantly exceeded the break-even threshold of $4900./
- **Conduct a test of hypothesis to determine whether Store B's average daily sales for the year significantly exceeded the break-even threshold of $4900.**	- Conduct a test of hypothesis to determine whether Store B's average daily sales for the eyar significantly exceeded the break-even threshold of $4900.
- **Conduct an additional test of hypothesis to determine whether Store A's average daily sales was significantly lower than Store B's average daily sales.**	- Conduct an additional test of whether Store A's average daily sales was significantly lower than Store B's average daily sales.

Essay

Write a 3–6-page summary paper that includes your analysis of the sales data. Your analysis should include:

- The output and interpretation of the statistical analyses listed above.
- Factors you determined to be relevant in your analysis, such as the source fields or attributes you used.
- Identified potential inherent risks of the methods you used in your analysis.
- Checks to validate all assumptions necessary for the hypothesis testing and confidence interval methods you selected.
- Create a persuasive argument advocating for a viewpoint or recommendation.
- Use communication style and vocabulary appropriate for the target audience.

Your assignment will be graded on the following criteria:

- Apply statistical analysis.
- Identify factors relevant to assumptions, risks, and source fields.
- Present a well-supported position.
- Communicate to stakeholders clearly and succinctly.

Refer to the Statistical Analysis of Database Solutions Scoring Guide for more details.

Additional Requirements

- **Written communication**: Written communication is free of errors that detract from the overall message.
- **Length of paper**: 3–6 typed double-spaced pages.
- **Resources:** At least three scholarly resources. Include a reference page at the end of the paper.
- **APA guidelines:** Double-spaced paragraph formatting in the body of the paper. When appropriate, use APA-formatted headings. Resources and citations are formatted according to current APA style and format.
- **Font and font size:** Times New Roman, 11 point font.

Competencies Measured

By successfully completing this assignment, you will demonstrate your proficiency in the following course competencies and assignment criteria:

- **Competency 1: Apply data management fundamentals.**
  - Identify factors relevant to assumptions, risks, and source fields.
- **Competency 3: Revise data mining concepts.**
  - Apply statistical analysis.
- **Competency 6: Communicate the decision-making process to stakeholders.**
  - Create a persuasive argument advocating for a viewpoint or recommendation.
  - Use communication style and vocabulary appropriate for the target audience.

Course Resources

Academic and Professional Document Guidelines [PDF]

Unit 6 >> Key Performance Indicators

Introduction

Case Study

Tiffany realizes her team needs to understand the significance of Key Performance Indicators (KPI) to the organization and what that means for a data analyst. To an analyst, indicators are more than targets. An aspect of the role of an analyst is to assist with the improvement efforts of the organization.

The analyst is usually in the position to assess the variability of some process, product, or service with constant monitoring, evolution, and assessments. However, to the management of the organization, indicators become the elements of a report card about how it appears in static reports or dashboards.

Tiffany reinforces how KPIs are key measures a data analyst can work with to help in an organization's performance reports and dashboards.

Regardless of the intent of an analyst in creating or modifying an indicator, the makeup of the indicator, its understanding by users, and its acceptance by stakeholders are important. This week, you will learn about key performance indicators: what they are, how they differ from one industry to another, and how they are used in quantitative and qualitative research.

To-Do List:

- **Discussion:** Select an industry that is familiar or of interest to you and analyze its key performance indicators.
- **Assignment:** Identify data sources and create a data set to calculate a Key Performance Indicator (KPI), create a step-by-step guide for identifying, sourcing, and calculating a new KPI, and make recommendations for best practices.
- **What You need to Know:** Study how KPI measures are commonly used to help a business define and evaluate how successful it is, typically in terms of meeting the milestones and objectives towards its long-term organizational goals.

Learning Activities

u06s1 - Activity Overviews

Discussion Overview

In the discussion for this week, you will select an industry that is familiar or of interest to you and analyze its key performance indicators.

Assignment Overview

In the assignment for this week, you will:

- Identify data sources and create a data set to calculate a KPI.
- Create a step-by-step guide for identifying, sourcing, and calculating a new KPI.
- Make recommendations for best practices.

u06s2 - What You Need to Know

Introduction to Key Performance Indicators (KPI)

The complicated path of modern business projects reflects a business environment that is continuously growing in complexity. Factors impacting a project's progress, such as new advancements in computer technology, an unpredictable economy, and the increase in stakeholder involvement makes metrics and Key Performance Indicators (KPIs) for project management important focal points. KPI measures are commonly used to help a business define and evaluate how successful it is, typically in terms of meeting the milestones and objectives towards its long-term organizational goals.

Not all indicators are KPIs. Some indicators are building blocks or variables for a composite indicator. There are indicators that are relevant to the present and future state of the business. Other indicators are selected based on personal preferences of stakeholders, and may or may not help the organization directly.

Access the following reading:

- Parmenter, D. (2015). *Key performance indicators (KPI): Developing, implementing, and using winning KPIs* (3rd. ed.). Hoboken, NJ: Wiley.
    - Read Chapter 1, "The Great KPI Misunderstanding," pages 3–23.
        - Refer to Appendix E: Performance Measures Database for examples of KPIs.

The following walkthrough will help you understand key performance indicators:

- Key Performance Indicators Walkthrough.

u06d1 - Write Your Discussion Post

KPI Measures and Improvement Strategies

For this discussion, you will select an industry that is familiar or of interest to you and analyze its KPI.

- Use the Capella University Library and the Internet to find the KPIs commonly used in your selected industry, and provide an overview of a minimum of 3–5 of them.
- Select one KPI to focus on for the remainder of this KPI analysis.
- Define the business problem that is being measured by your selected KPI, and select a goal, or optimal value, for the KPI, along with references supporting that goal.
- Identify the data types and data sources that would be required to analyze and report on your selected KPI.
- Compare and contrast the quantitative and qualitative research methods that might be used to assist in improving your selected KPI and identify the advantages of each method with supporting references.

Response Guidelines

Respond to at least two other learners and share with them the information in their initial posts that best helped you in understanding the concepts.

Course Resources

Graduate Discussion Participation Scoring Guide

u06a1 - Data Sources and KPIs

Overview

In this assignment, you will:

- Identify data sources and create a data set to calculate a KPI.
- Create a step-by-step guide for identifying, sourcing, and calculating a new key performance indicator.
- Make recommendations for best practices.

Include both your data analysis code and your guide in your assignment submission.

KPIs usually do not emerge from source systems directly as they are. They are usually a calculation, or series of calculations, made on a combination of variables from a variety of sources. When first identifying a KPI, you need to figure out how to measure it, where the data will be sourced, and how to manipulate the data to translate it into a meaningful indicator. By doing this assignment, you will gain an understanding of what work is needed "behind the scenes" to build KPIs and related reporting. Since KPIs are often the main point of contact an executive has with the organization's data, it is important for analysts to understand what the KPIs are composed of, and what factors might contribute to the change in value for a particular KPI. Analysts can then answer the "why" questions asked by executives regarding the degradation or improvement in performance on a particular KPI.

Instructions

Use the industry and the industry-specific KPI from your initial discussion post for this week's discussion as you complete the following tasks and analysis:

- Identify the potential source systems for the data you need to calculate this KPI, including definitions of the data source(s) and the variables from those data sources you will need to calculate the KPI.
- Create a sample data set that includes the variables identified above as inputs to the KPI calculation/manipulation, and 10 sample records.
- Calculate the identified KPI using the sample data set and included variables. Use the necessary procedures needed to prepare the sample data set.
- Based on your work in this assignment, create a step-by-step guide for identifying, sourcing, and calculating a new key performance indicator, and make recommendations for best practices related to these steps.
- Include your code, output, and a screenshot of the data set in your submission.

Your assignment will be graded on the following criteria:

- Identify KPIs relevant to a professional domain.
- Identify data sources to use in calculating a KPI.
- Create data set from scratch.
- Identify calculations and manipulations needed for translating source data into a KPI.
- Provide a logical argument in support of conclusions or recommendations.

Refer to the Data Sources and KPIs Scoring Guide for more details.

Additional Requirements

- **Written communication**: Written communication is free of errors that detract from the overall message.
- **Length of paper**: 4–6 typed double-spaced pages.
- **Resources:** At least three scholarly resources. Include a reference page at the end of the paper.
- **APA guidelines:** Double-spaced paragraph formatting in the body of the paper. When appropriate, use APA-formatted headings. Resources and citations are formatted according to current APA style and format.
- **Font and font size:** Times New Roman, 11 point font.

Competencies Measured

By successfully completing this assignment, you will demonstrate your proficiency in the following course competencies and assignment criteria:

- **Competency 1: Apply data management fundamentals.**
  - Identify data sources to use in calculating a KPI.
  - Create a data set from scratch.
- **Competency 2: Describe the concepts of ETLs and data warehouses.**
  - Identify calculations and manipulations needed for translating source data into a KPI.
- **Competency 4: Evaluate applied analytics in professional domains.**
  - Identify key performance indicators related to a professional domain.
- **Competency 6: Communicate the decision-making process to stakeholders.**
  - Provide a logical argument in support of conclusions or recommendations.

Course Resources

Academic and Professional Document Guidelines [PDF]

Unit 7 >> Data Warehouse Design and Mining

Introduction

Case Study

At Vila Health, data has been stored in transactional databases for quite some time. Although Tiffany has not worked with the data warehouses specifically, she does know that this is an important transition that her team needs to be involved in. Transitioning data from many different sources into a cohesive data warehouse is on the books for Tiffany so it is important that she becomes in tune and aware of the changes and needs.

The readings in this unit, provides information on analytical processes, data warehousing, ETL design, and work flows. They describe how the data analyst may approach the task of defining data and how data may work together as information contained in a data mart.

In this week, you will locate an article or case on how an organization applied a data mining model or algorithm to solve a problem and you will highlight the challenges they may have faced along the way. Specifically, you may choose an operations scenario in your industry using public domain data.

To-Do List:

- **Discussion:** Search for and summarize an article published within in the past three years that applies data mining methods to solve an organizational need in an industry of your choice.
- **Assignment:** Summarize best practices for data analytics programming.
- **What You need to Know:** Study data warehousing techniques and the extract, transform, and load (ETL), is a process most commonly used with data warehousing, and how data mining can be used for decision making.

Learning Activities

u07s1 - Activity Overviews

Discussion Overview

In the discussion for week, you will search for and summarize an article published within in the past three years that applies data mining methods to solve an organizational need in an industry of your choice.

Assignment Overview

In the assignment for this week, you will summarize best practices for data analytics programming.

u07s2 - What You Need to Know

Introduction to Data Warehousing

Data science literature seems to leave out a number of challenges analysts face on a daily basis. The three most basic and frustrating of these challenges are knowing which data sources to use, obtaining the permissions to access those source systems, and determining how the data is supposed to work together.

Often, the time of the analyst is dedicated to data analysis. That task is facilitated when the organization has a data governance program, enabling analysts and other users of the data to understand the definition, source, and use of the data elements themselves. Even so, the next activity for most analysts is data prepping (data cleansing and data conforming). All the while, the analyst continues to perform analysis, learning about the data needed to solve a problem.

Extract, transform, and load (ETL), is a process most commonly used with data warehousing. A major goal of a data warehouse is to enable an organization to create the information it needs to reach its goals. Although there are many ways to access and use the data in the data warehouse, the analyst may find that the information is incomplete, not entirely valid, or containing wrong information.

An analyst may discover data and information errors within a data warehouse. When this happens, the analyst may develop a process to obtain the data needed for a project by reaching out to various parts of an organization. By reaching out and working with other professionals the analyst may enhance the current data warehouse solution.

This week's resources provide an overview of data warehousing for data mining:

- Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Waltham, MA: Elsevier.

    - In Chapter 4, "Data Warehousing and Online Analytical Processing:"
        - Read Section 4.1, "Data Warehouse: Basic Concepts," pages 125–135.
        - Read Section 4.2, "Data Warehouse Modeling: Data Cube and OLAP," pages 135–150.
        - Read Section 4.3, "Data Warehouse Design and Usage," pages 150–156.
- Anitha, J., & Babu, M. S. P. (2014). ETL work flow for extract transform loading [PDF]. *International Journal of Computer Science and Mobile Computing, 3*(6), 610–617. Retrieved from http://ijcsmc.com/docs/papers/June2014/V3I6201481.pdf

Introduction to Data Mining for Decision Making

A major goal of data analytics is to measure and reduce uncertainty to help improve decision making by an organization (Albright & Winston, 2013). To accomplish this goal, organizations use traditional statistics, descriptive and predictive analytic algorithms with data mining, and other analytical methods. Note that here we refer to mining as unsupervised analytics, equivalent to a descriptive process. Data mining will lead us to low-hanging fruit. The next step would be exploitation of statistical methods for diagnostic, predictive, and prescriptive analytics.

This week's resources provide an overview of how data mining can be used to make decisions:

- Larose, D. T., & Larose, C. D. (2014). *Discovering knowledge in data: An introduction to data mining* (2nd. ed.). Hoboken, NJ: Wiley.
    - Read Chapter 1, "An Introduction to Data Mining."

The following walkthrough to help you understand data mining and decision making.

- Data Mining and Decision Making Walkthrough.

References

Albright, S. C., & Winston, W. L. (2013). Business analytics: Data analysis and decision making. Stamford, CT: Cengage Learning.

Brohman, M. K., & Watson, H. J. (2006). Maximizing the return on OLAP and data mining analysts. Business Intelligence Journal, 11(3), 30–36.

Moody, D. L., & Kortink, M. A. R. (2003). From ER models to dimensional models: Bridging the gap between OLTP and OLAP design, Part I. Business Intelligence Journal, 8(3), 7–24.

Saunders, T. (2009). Cooking up a data warehouse. Business Intelligence Journal, 14(2), 16–22.

Sherman, R. (2007). Trial-and-error method of ETL. DM Review, 17(4), 27.

u07d1 - Write Your Discussion Post

Data Mining Applications

Use the Capella University Library: Journal and Book Locator, to search for each of the journal publications listed in this discussion, then search for an article within one of these publications to summarize in this discussion. Select an article that was published within in the past three years that applies data mining methods to solve an organizational need in an industry of your choice from one of these journal publications.

- *ACM Transactions on Knowledge Discovery From Data.*
- *Business Intelligence Journal.*
- *Harvard Business Review.*
- *Information Systems Management*.
- *International Journal of Data Warehousing and Mining.*
- *International Journal of Business Analytics and Intelligence*.
- *Journal of Knowledge Management*.
- *Journal of Forecasting*.
- *Journal of Marketing Analytics*.
- *MIT Sloan Management Review*.

Summarize the application of data mining methods in the article you selected. Use the following questions as applicable to guide your summary of the article:

- What journal did you select you article from, and why did you select an article from this journal?
- Why did you select this article to summarize?
- What organizational need was addressed with data mining?
- What data mining methods were applied to address the organizational need?

- What were the challenges of the applied data mining methods?
- What were the positive outcomes of the of the applied data mining methods?

Response Guidelines

Respond to at least two other learners and share with them the information in their initial posts that best helped you in understanding the concepts.

Course Resources

Graduate Discussion Participation Scoring Guide

u07s3 - Prepare: Programming Instructions

Using R

Access the Using R page on Campus for resources on using Simple Linear Regression in R.

Using Python

Access the Using Python page on Campus for resources on using Simple Linear Regression in Python.

Using SAS

Access the Using SAS page on Campus for resources on using Simple Linear Regression in SAS.

u07a1 - Programming Best Practices

Overview

In completing your previous assignments for this course, you used data analytics programs to extract and transform data in preparation for analysis. In this assignment, you will summarize best practices for data analytics programming.

Instructions

Create a 1–2-page executive summary of best practices for data analytics programming, including the following portions:

- Best practices for testing your written program.
- Best practices for submitting and troubleshooting your written program.
- Best practices for documentation and communication regarding the intention and execution of your program, as well as what is included in your program.

Your assignment will be scored on the following criteria:

- Recommend best practices for testing written programs.
- Recommend best practices for submitting and troubleshooting written programs.
- Recommend best practices for documentation and communication for written programming.
- Create a persuasive argument advocating for a viewpoint or recommendation.
- Create an executive summary with appropriate focus and level of detail.

Refer to the Programming Best Practices Scoring Guide for more details.

Additional Requirements

- **Written communication**: Written communication is free of errors that detract from the overall message.
- **Length of document**: 1–2 typed double-spaced pages.

- **Resources:** At least three scholarly resources. Include a reference page at the end of the paper.
- **APA guidelines:** Double-spaced paragraph formatting in the body of the paper. When appropriate, use APA-formatted headings. Resources and citations are formatted according to current APA style and format.
- **Font and font size:** Times New Roman, 11 point font.

Competencies Measured

By successfully completing this assignment, you will demonstrate your proficiency in the following course competencies and assignment criteria:

- **Competency 1: Apply data management fundamentals.**
    - Recommend best practices for testing programs.
- **Competency 3: Revise data mining concepts.**
    - Recommend best practices for submitting and troubleshooting programs.
- **Competency 5: Justify the use of resources available within a collaborative environment.**
    - Recommend best practices for documentation and communication for programming.
- **Competency 6: Communicate the decision-making process to stakeholders.**
    - Create a persuasive argument advocating for a viewpoint or recommendation.
    - Create an executive summary with appropriate focus and level of detail.

Course Resources

Academic and Professional Document Guidelines [PDF]

Unit 8 >> Troubleshooting and Optimizing the ETL Process

Introduction

Case Study

Tiffany has started to explore the data warehouse now and there have been quite a few challenges for the organization. In part, having to manage all of the data from various sources has introduced a lot of issues. She now has to train her team to troubleshoot the ETL process and go through the proper data cleansing processing.

The week introduces regression analysis concepts and programming steps. You will also examine the importance of troubleshooting and optimizing the ETL (extract, load, transfer) process to improve your data analysis.

To-Do List:

- **Discussion:** Share best practices you have learned for programming a data cleansing or ETL (extract, load, transfer) solution.
- **Assignment:** Write data analysis code for an elementary data model and write an essay that analyzes the data mining method you conducted for this assignment.
- **What You need to Know:** Study regression analysis concepts and programming steps, and examine the importance of troubleshooting and optimizing the ETL (extract, load, transfer) process to improve your data analysis.

Learning Activities

u08s1 - Activity Overviews

Discussion Overview

In the discussion for this week, you will share best practices you have learned for programming a data cleansing or ETL (extract, load, transfer) solution.

Assignment Overview

In the assignment for this week, you will write data analysis code and write an essay that analyzes the data mining method you conducted for this assignment.

u08s2 - What You Need to Know

Regression Analysis

Use the Capella Library to read the following:

- Anderson, A. (2014). Business Statistics For Dummies John Wiley & Sons: Skillsoft Collection
  - Chapter 15: Simple Regression Analysis. This chapter will explain the need for regression and why it is performed.

Troubleshooting and Optimizing ETL

Refining and enhancing an analytic solution is common. Analysts may spend most of their time understanding a problem, analyzing the relevant data, and preparing it for user consumption and action. Data bottlenecks, different data definitions for the same data attribute, and untimely—or worse, unnoticed— changes that were not accounted for in the data of a source system can cause issues at the end, when it is time for the analysis to make sense. When these data problems emerge, analysts engage in troubleshooting their ETL and data cleansing and manipulation processes. In addition, stakeholders often request ongoing updates on these analyses. Whether those include updates to the actual analytical model or simply ongoing model validation, a more reliable and optimized ETL process is necessary to ensure the quality of those updates and to minimize the time spent waiting by the analyst for the regular process to complete.

Use the Capella Library to read or review the following:

- Apgar, D. (2015). The false promise of big data: Can data mining replace hypothesis-driven learning in the identification of predictive performance metrics? *Systems Research and Behavioral Science, 32*(1), 28–49.
- Badawy, M., El-Aziz, A. A., Idress, A. M., Hefny, H., & Hossam, S. (2016, April). A survey on exploring key performance indicators. *Future Computing and Informatics Journal.*
- Gholami, R., Higón, D. A., & Emrouznejad, A. (2015). Hospital performance: Efficiency or quality? Can we have both with IT? *Expert Systems With Applications, 42*(12), 5390–5400.
- Rivera, J., & Delaney, S. (2015). Using business analytics to improve outcomes. *Healthcare Financial Management, 69*(2), 64–67.

SAS

Use the Capella Library to read the following if you are choosing to use SAS for your coding:

- Lee, G. (2015). Business Statistics Made Easy in SAS SAS Institute: Skillsoft Collection
  - Chapter 13. This chapter focus on linear regression and how to implement it in SAS.

R

Use the Capella Library to read the following if you are choosing to use R for your coding:

- Hui, E. (2019). Learn R for Applied Statistics: With Data Visualizations, Regressions, and Statistics Apress: Skillsoft Collection
  - Chapters 6. The end of chapter 6 looks at linear regressions in R.
- Schmuller, J.(2017). Statistical Analysis with R For Dummies John Wiley & Sons: Skillsoft Collection
  - Chapter 14 explores regression using R.

Course Resources

Anderson, A. (2014). Business Statistics For Dummies John Wiley & Sons: Skillsoft Collection

u08d1 - Write Your Discussion Post

ETL Lessons Learned

For this discussion, share the best practices you have learned for programming a data cleansing or ETL (extract, load, transfer) solution. Your best practices should answer the following questions:

- What steps should you take to test your code prior to submitting it against the full data set?
- How do you tell whether there was an issue or error in running your code?
- What are some ways you can ensure that the next analyst to run your program can understand and effectively troubleshoot your code?
- What else do we need to consider as the volume of your data grows larger and larger.

Response Guidelines

Respond to at least two other learners and share with them the information in their initial posts that best helped you in understanding the concepts.

Course Resources

Graduate Discussion Participation Scoring Guide

u08v1 - Interactive Learning Module: Analytics Internship - Performance Management and KPIs

In this final Vila Health activity for this course, you will receive a request that will significantly change the work and recommendations you have been preparing. Your will need to discuss what the changes mean and how they will impact the analytic approach. As you gather information from your mentor and the stakeholders, consider what other data might now be required and what questions you will need to address in order to make your final recommendations.

Course Resources

Analytics Internship: Performance Management and KPIs | Transcript

u08a1 - Elementary Data Mining Modeling

Overview

For this assignment, you will write data analysis code and write an essay that analyzes the data mining method you conducted for this assignment. You will conduct a regression analysis that includes performing diagnostic checks, creating a linear regression equation, and conducting a sensitivity analysis on the resulting linear regression equation. This assignment allows you to understand the framework for conducting a regression analysis, which is an elementary form of data mining.

Instructions

Before you look at the data set, cf_ANLT5010_W8_ Height_Shoe_Size.txt from the ANLT5010 Data Files (ZIP) package. Read the course file, cf_ANLT5010_W8_ChallengeScenario.docx provided in the resources for this assignment

Data Analysis Programming

- Conduct a simple linear regression analysis using SAS or R (lm command).
- Check that the assumptions of the linear regression analysis methods are met.
- Create and specify the equation of the linear regression model.
- Use the linear regression equation to predict the value of the dependent variable given the independent variable.
- Conduct a sensitivity analysis on your simple linear regression model.

Essay

- Analyze the applicability and limitations of regression analysis, in general, and how to measure the quality of the regression model.
- Explain potential dangers of extrapolation using this regression model, in particular, and also using any regression model, in general.
- Address the following:
  - What types of challenges did you have in developing this data mining solution?
  - What other types of challenges might organizations have while developing, implementing, and/or using data mining solutions like this?
  - How might a data mining solution like this be used by an organization to address its organizational challenges?
- Write a 4–7-page summary paper of your analysis including appropriate graphs, program output, or screenshots to support your analysis.

Your assignment will be scored on the following criteria:

- Analyze the application of an analytics project to organizational challenges.
- Describe the challenges pertaining to developing or implementing data mining solutions.
- Explain the challenges pertaining to data domain expert and predictive models using regression.
- Conduct sensitivity analysis on predictive outcome based on changes in independent variables.
- Present analysis of qualitative information to effectively explain or justify a viewpoint or recommendation.

Refer to the Elementary Data Mining Modeling Scoring Guide for more details.

Additional Requirements

- **Written communication**: Written communication is free of errors that detract from the overall message.
- **Length of paper**: 4–7 typed double-spaced pages.
- **Resources:** At least three scholarly resources. Include a reference page at the end of the paper.
- **APA guidelines:** Double-spaced paragraph formatting in the body of the paper. When appropriate, use APA-formatted headings. Resources and citations are formatted according to current APA style and format.
- **Font and font size:** Times New Roman, 11 point font.

Competencies Measured

By successfully completing this assignment, you will demonstrate your proficiency in the following course competencies and assignment criteria:

- **Competency 3: Revise data mining concepts.**
  - Explain challenges pertaining to data domain expert and predictive models using regression.
  - Conduct sensitivity analysis on predictive outcome based on changes in independent variables.
  - Conduct sensitivity analysis on predictive outcome based on changes in independent variables.
  - Describe challenges pertaining to developing or implementing data mining solutions.
- **Competency 6: Communicate the decision-making process to stakeholders.**
  - Present analysis of qualitative information to effectively explain or justify a viewpoint or recommendation.

Course Resources

Academic and Professional Document Guidelines [PDF]

Unit 9 >> Analytics Solution Strategies

Introduction

Case Study

Now that Tiffany has her team working through the ETL processes, she has been introduced to various departments within Vila Health. They have been looking for specific answers based on data. This in part has made the organization define specific indicators to be measured for performance. Understanding that also helps Tiffany ensure that the correct variables and data sets are being logged.

Once initial analysis is completed on a KPI and ongoing reporting is set up, it is often necessary to investigate and troubleshoot drastic changes in performance on a particular KPI or set of KPIs. This may be as simple as adding variables to the reporting solution for that KPI, so that performance on that KPI can be evaluated from different perspectives. However, this is not always enough. Sometimes, so little is known about what contributes to, or impacts, the performance on that KPI that either elementary or more advanced data mining needs to be conducted to identify the factors that contribute to the KPI's performance. In this week, you will learn about the interaction between KPIs and data mining, discuss how data mining can be used to identify KPIs and to address poor performance on KPIs, and create a proposal for doing so in an industry of your choice.

According to Strome (2013, p. 29):

> *An analytics strategy is more than simply a data utilization strategy, a data analysis strategy, a technology strategy, or a quality improvement strategy. In fact, elements of all these are required for an effective analytics strategy. An analytics strategy is necessary to ensure that an organization's analytics capabilities are aligned with its quality and performance improvement needs.*

Resources

Strome, T. L. (2013). *Healthcare analytics for quality and performance improvement.* Hoboken, NJ: Wiley & Sons.

To-Do List:

- **Discussion:** Research and describe essential elements to include in an analytic solution proposal.
- **Assignment:** Conduct research for a key performance indicator (KPI) to determine which potential statistical, analytical, or data mining software and procedures may be used for data mining on this KPI, and you will and write an essay about your research.
- **What You need to Know:** There are no assigned readings for this unit. Please look ahead to Week 10 and prepare for the upcoming assignment.

Learning Activities

u09s1 - Activity Overviews

Discussion Overview

In the discussion for this week, you will research and describe the essential elements to be included in an analytic solution proposal.

Assignment Overview

In the assignment for this week, you will conduct research for a key performance indicator (KPI) to determine which potential statistical, analytical, or data mining software and procedures may be used for data mining on this KPI, and you will and write an essay about your research.

u09s2 - What You Need to Know

There are no assigned readings for this unit. Please look ahead to Week 10 and prepare for the upcoming assignment.

u09d1 - Write Your Discussion Post

Analytic Solution Proposal Key Elements

In preparation for this unit’s assignment, use the Capella University Library or the Internet to find examples of analytic solution proposals. Based on the examples you found in your searches and your experience, discuss what you believe are the essential elements to include in every analytic solution proposal.

Response Guidelines

Respond to at least two other learners. Share with them any elements you believe are missing from their lists and any that you consider unnecessary. Be sure to support your assertions with examples or references.

Course Resources

Graduate Discussion Participation Scoring Guide

u09a1 - Data Mining and KPIs

Overview

In this assignment, you will conduct research for a Key Performance Indicator (KPI) to determine which potential statistical, analytical, or data mining software and procedures may be used for data mining on this KPI, and you will write an essay that describes your research and includes sample code for data mining.

As you learned in previous weeks of this course, there are recommended or best practice KPIs in various industries. Once an organization has identified a KPI, and has initially analyzed and created reporting for that KPI, the organization continues to monitor performance on that KPI. When unexpected performance changes occur on that KPI, further analysis needs to be completed to understand the factors and variables that are contributing to that unexpected performance. As advanced analytical methods are adopted by more and more industries and organizations, they have begun to turn to data mining methods for evaluating performance and researching performance issues on those KPIs. As a data analyst or data scientist, you will need to understand how to identify KPIs and code the manipulations and calculations for them. Going beyond that, you also will identify which data mining methods might be appropriate for researching unexpected KPI performance changes, and determine how to conduct those methods using your organization's selected statistical, analytical, or data mining software.

Instructions

To complete this assignment, conduct research and write a paper based on the following steps:

- Select an industry-standard KPI for an industry of your choice. This KPI could be the same as the one you selected for your previous assignment or a different one.
- Explain how data mining could be used to research poor performance or big changes in performance of your selected KPI.
- Choose one specific data mining method and describe how that method, specifically, might be used to investigate performance changes or poor performance on your selected KPI.
- Research on the Internet or in the Capella University Library to find what statistical, analytical, or data mining software and procedures might be used for your specified data mining method.
- Include sample code, a description of how the code might be used, and supporting references, where appropriate.

Your assignment will be scored on the following criteria:

- Identify key performance indicators.
- Explain how data mining methods can be used to research unexpected performance on a key performance indicator.
- Recommend appropriate software and procedures for using data mining to evaluate and improve performance on a key performance indicator.
- Recommend performance measurements and a management plan for improvement.
- Cite reliable external sources that explain or justify a viewpoint or recommendation.

Refer to the Data Mining and KPIs Scoring Guide for more details.

Additional Requirements

- **Written communication:** Written communication is free of errors that detract from the overall message.
- **Length of paper:** 4–6 typed double-spaced pages.
- **Resources:** At least three scholarly resources. Include a reference page at the end of the paper.
- **APA guidelines:** Double-spaced paragraph formatting in the body of the paper. When appropriate, use APA-formatted headings. Resources and citations are formatted according to current APA style and format.
- **Font and font size:** Times New Roman, 11 point font.

Competencies Measured

By successfully completing this assignment, you will demonstrate your proficiency in the following course competencies and assignment criteria:

- **Competency 3: Revise data mining concepts.**
    - Recommend appropriate software and procedures for using data mining to evaluate and improve performance on a key performance indicator.
    - Explain how data mining methods can be used to research unexpected performance on a key performance indicator.
- **Competency 4: Evaluate applied analytics in professional domains.**
    - Recommend performance measurements and a management plan for improvement.
    - Identify the key performance indicators.
- **Competency 6: Communicate the decision-making process to stakeholders.**
    - Cite reliable external sources that explain or justify a viewpoint or recommendation.

Course Resources

Academic and Professional Document Guidelines [PDF]

Unit 10 >> Human Intervention and Successful Data Solutions

Introduction

Case Study

Tiffany has truly understood the need for the role of a domain expert as she has progressed with her team. There is a need for the team to focus on balance, emphasizing that discussions may play a role in obtaining optimal results in an analytics project. This is the theory she put into practice in coming up with a solution for the Vila Health report.

As a reminder, your final report should contain the summary of what you worked on throughout the course, from data identification and manipulation, data warehousing and data mining, to specific analytics employed. Your diagnostic and inferential analyses and your proposed solution should be an integral part of this report.

To-Do List:

- **Discussion:** Reflect on what you learned during this course and how you may apply what have learned.
- **Assignment:** Write an analytics solution proposal that highlights the various data analysis methods you learned during this course to recommend a diagnostic solution for identified Key Performance Indicators (KPIs).
- **What You need to Know:** Review examples of the application of sustained and successful data solutions.

Learning Activities

u10s1 - Activity Overviews

Discussion Overview

In the discussion for this week, you will reflect on what you learned during this course and how you may apply what have learned.

Assignment Overview

In the assignment for this week, you will write an analytics solution proposal that highlights the various data analysis methods you learned during this course to recommend a diagnostic solution for identified KPIs.

u10s2 - What You Need to Know

Sustaining Successful Data Solutions

Use the Capella University Library to read the following:

- Bell, P. C. (2015). Sustaining an analytics advantage [PDF]. *MIT Sloan Management Review, 56*(3), 21–24.

Use the Internet to read or review the following:

- Brown, B. (2015). SAS University Edition debuts new features. Retrieved from https://communities.sas.com/docs/DOC-19735
- Horner, P. (2015, December). Q&A: The state of INFORMS and the profession. *ORMS Today, 40*(6). Retrieved from https://www.informs.org/ORMS-Today/Public-Articles/December-Volume-40-Number-6/Q-A-The-state-of-INFORMS-and-the-profession
- SAS. (n.d.). SAS support communities. Retrieved from https://communities.sas.com/https://communities.sas.com/

u10d1 - Write Your Discussion Post

Reflection

For this discussion:

- Reflect on the past 10 weeks of this course.
- Elaborate on what you have found most informative and how you can apply what you have learned in future endeavors.
- Discuss your challenges and opportunities you encountered.
  - What was the most challenging aspect of the group report?

Response Guidelines

Responses to this discussion are optional but welcomed.

Course Resources

Graduate Discussion Participation Scoring Guide

u10a1 - Vila Health Final Report

Overview

For this final assignment you will compile an analytics solution proposal report of no more than 20 pages. In this report you will highlight the various data analysis methods you learned during this course to recommend a diagnostic solution for identified Key Performance Indicators (KPIs) including charts, graphs, and code as appropriate.

Instructions

For this assignment use the cf_ANLT5010_W10_Penalties_ClarionCourt.csv data set. You will also use the cf_ANLR5010_W10_2004ResidentFile_DataDictionary_061609.pdf data dictionary. The key information to include in the report are listed here.

- List all the data fields available to be used for evaluating Clarion Court's performance on falls and pressure ulcers.
- Propose, and support with references, appropriate KPIs for evaluating Clarion Court's performance with respect to falls and pressure ulcers.
- Identify calculations and transformation rules to create these KPIs using the existing **cf_ANLT5020_W10_Penalties_ClarionCourtcsv** data, and discuss whether sufficient data is currently available to do so.
  - If sufficient data does not currently exist to create the KPIs, identify what data is needed.

- Include a description of the records and the variables that are needed for the creation of these KPIs.

- Describe how your data cleansing strategy may be incorporated into an ETL (extract, transform, load) process for loading data into a data mart to allow for future reporting and analysis of falls and pressure ulcers.
- Summarize the steps used to identify the sources of data, investigate data quality issues, and identify data attributes and variables.
- Include in your analysis an assessment of the present and future state of the Vila Health data and your proposed data analytics solution.
- Provide details of your predictive model and assumptions, and describe your diagnostic recommendation.
    - Identify whether or not the recommendation addresses the current business issue at Vila Health.
    - If not, identify what would be needed in order to do so.

Your assignment will be scored on the following criteria:

- Summarize the steps to identify data sources, attributes, and variables.
- Describe a predictive model and diagnostic recommendations.
- Identify key performance indicators (KPIs) used by various industries and/or professional domains.
- Integrate visual elements that clarify or highlight key points.
- Provide a logical argument in support of conclusions or recommendations.
- Use communication style and vocabulary appropriate for the target audience.

Refer to the Vila Health Final Report Scoring Guide for more details.

Additional Requirements

- **Written communication:** Written communication is free of errors that detract from the overall message.
- **Length of paper:** No more than 20 typed double-spaced pages.
- **Resources:** At least three scholarly resources. Include a reference page at the end of the paper.
- **APA guidelines:** Double-spaced paragraph formatting in the body of the paper. When appropriate, use APA-formatted headings. Resources and citations are formatted according to current APA style and format.
- **Font and font size:** Times New Roman, 11 point font.

Competencies Measured

By successfully completing this assignment, you will demonstrate your proficiency in the following course competencies and assignment criteria:

- **Competency 1: Apply data management fundamentals.**
    - Summarize the steps to identify data sources, attributes, and variables.
- **Competency 3: Revise data mining concepts.**
    - Describe a predictive model and diagnostic recommendations.
- **Competency 4: Evaluate applied analytics in professional domains.**
    - Identify Key Performance Indicators (KPIs) used by industries and/or professional domains.
- **Competency 6: Communicate the decision-making process to stakeholders.**
    - Integrate visual elements that clarify or highlight key points.
    - Provide a logical argument in support of conclusions or recommendations.
    - Use communication style and vocabulary appropriate for the target audience.