COLORADO STATE UNIVERSITY
— GLOBAL —

MIS470: DATA SCIENCE FOUNDATIONS

Credit Hours: 3

Contact Hours: This is a 3-credit course, offered in accelerated format. This means that 16 weeks of material is covered in 8 weeks. The exact number of hours per week that you can expect to spend on each course will vary based upon the weekly coursework, as well as your study style and preferences. You should plan to spend 14-20 hours per week in each course reading material, interacting on the discussion boards, writing papers, completing projects, and doing research.

Faculty Information: Faculty contact information and office hours can be found on the faculty profile page.

COURSE DESCRIPTION AND OUTCOMES

Course Description:

This course provides an overview of the tools and techniques for analyzing data using statistics, R Programming, and SQL. Topics include data storage, linear regression, classification, linear models, tree-based learning, R programming, and SQL basic commands.

Course Overview:

Data Science is fundamentally important in many different fields. However, the focus in the course is from the business perspective. The aim of the course is to expose students in data science by providing necessarily tools through statistics, R programing, and SQL language. Descriptive as well as inferential statistics are examined. These are crucial in decision making. Predictive modeling and basic programming will be examined through R. The students will gain these skills through examples and step-by-step procedures. Lastly, the students will be able to manage and manipulate data using SQL.

Course Learning Outcomes:

1. Explain effective data storage methods.
2. Demonstrate descriptive data analysis including visualization.
3. Demonstrate linear regression analysis.
4. Apply classification to a dataset.
5. Demonstrate predictive data analysis including modeling.
6. Demonstrate the use of R programming for data analysis.
7. Explore databases using SQL commands.

PARTICIPATION & ATTENDANCE

Prompt and consistent attendance in your online courses is essential for your success at CSU-Global Campus. Failure to verify your attendance within the first 7 days of this course may result in your withdrawal. If for some reason you would like to drop a course, please contact your advisor.

Online classes have deadlines, assignments, and participation requirements just like on-campus classes. Budget your time carefully and keep an open line of communication with your instructor. If you are having technical problems, problems with your assignments, or other problems that are impeding your progress, let your instructor know as soon as possible.

Course Materials

Required:

This course utilizes a custom zyBook ebook that will be integrated within your course.

***NOTE:** All non-textbook required readings and materials necessary to complete assignments, discussions, and/or supplemental or required exercises are provided within the course itself. Please read through each course module carefully.*

Course Schedule

Due Dates

The Academic Week at CSU-Global begins on Monday and ends the following Sunday.

- **Discussion Boards:** The original post must be completed by Thursday at 11:59 p.m. MT and peer responses posted by Sunday at 11:59 p.m. MT. Late posts may not be awarded points.
- **Opening Exercises:** Take the Opening Exercise before reading each week's content to see which areas you will need to focus on. You may take these exercises as many times as you need. The Opening Exercises will not affect your final grade.
- **Mastery Exercises:** Students may access and retake Mastery Exercises through the last day of class until they achieve the scores they desire.
- **Critical Thinking:** Assignments are due Sunday at 11:59 p.m. MT.

Weekly Reading and Assignment Details

Module 1

Readings

- zyBooks – Module 1
- *An Introduction to R*: Chapters 1, 2, and 3

Opening Exercise (0 points)

Discussion (25 points)

Critical Thinking (55 points)

Choose one of the following two assignments to complete this week. Do not do both assignments. Identify your assignment choice in the title of your submission.

Option #1: R Installation and swirl

The installation instructions for R are listed in the module with this assignment: MIS470_R_Installation_Instructions.

The purpose of this assignment is to verify that you have successfully downloaded and installed R and RStudio in your computer. Also, you will gain the basics of R programming through an R package called **swirl**. First, download and install R and RStudio in your computer. The following website shows the installation steps (Steps 1 and 2): https://swirlstats.com/students.html

Complete the following using RStudio:

a. Install and start the function **swirl()**. These are Steps 3 and 4 from the website above.
b. In the R console, type swirl(). This will start the swirl package.
c. When prompted, type your name. Then, choose "2: No. Let me start something new." You should be able to see 15 lessons.
d. Complete the Lessons 1-10, 11, 13, and 15. Complete each lesson and take the screenshot of each lesson showing a 100% completion rate. After the last lesson, take a screenshot of the current date in R. You can use the function **Sys.Date()**.

Copy and paste all 13 + 1 screenshots (along with current date) into a Word document. Submit the Word file in Canvas for grading.

Option #2: RStudio Installation and swirl

The installation instructions for R are listed in the module with this assignment: MIS470_R_Installation_Instructions.

The purpose of this assignment is to verify that you have successfully downloaded and installed R and RStudio in your computer. Also, you will gain the basics of R programming through an R package called **swirl**. First, download and install R and RStudio in your computer. The following website shows the installation steps (Steps 1 and 2): https://swirlstats.com/students.html

Complete the following steps using RStudio:

a. Install and start the function **swirl()**. These are Steps 3 and 4 from the website above.
b. In the R console, type swirl(). This will start the swirl package.
c. When prompted, type your name. Then, choose "2: No. Let me start something new." You should be able to see 15 lessons.
d. Complete the Lessons 1-10, 12, 14, and 15. Complete each lesson and take a screenshot of each lesson showing a 100% completion rate. After the last lesson, take a screenshot of the current date in R. You can use the function **Sys.Date()**.

   Copy and paste all 13 + 1 screenshots (along with current date) into a Word document. Submit the Word file in Canvas for grading.

Mastery Exercise (10 points)

Module 2

Readings

- zyBooks – Module 2

Opening Exercise (0 points)

Discussion (25 points)

Critical Thinking (55 points)

Choose one of the following two assignments to complete this week. Do not do both assignments. Identify your assignment choice in the title of your submission.

Option #1: Confidence Intervals

One of the most popular ways to estimate an unknown parameter is the confidence interval. Instead of just using a point estimate, it gives an interval estimate. This gives more information about the unknown parameter. In this assignment, we will examine the confidence interval.

a. Download the dataset *JNJ.csv* from the course website.
b. Using R, compute the sample mean of the closing price of Johnson and Johnson's (JNJ) daily stock prices.
c. Using R, compute the standard deviation. Also, compute the 90%, 95%, and 99% confidence intervals for the true mean closing JNJ stock price.
d. Based on c, discuss the pattern you see from these three confidence intervals and explain why such a pattern exists.
e. What is the advantage(s) and disadvantage(s) of increasing the confidence levels? Explain.
f. Interpret the confidence intervals you've calculated from part c.

Using a Word document, show the relevant calculations as well as the answers/solution. Copy the screenshots of the R output and paste them into the Word document. Your screenshot must include the current date using the function Sys.Date(). Submit the Word file in Canvas for grading.

Option #2: Hypothesis Testing

Hypothesis testing allows us to make decision(s) based on data. In this assignment, we wish to determine whether the mean closing price of Johnson and Johnson's daily stock price is greater than $125.

a. Download the dataset JNJ.csv from the course website.
b. Write the null and alternative hypotheses.
c. Using R, compute the test statistic.
d. Using R, calculate the p-value.
e. Using the significance level of 0.05, what is the decision? Explain how you've reached the decision.
f. Interpret your findings.

Using a Word document, show the relevant calculations as well as the answers/solution. Copy the screenshots of the R output and paste them into the Word document. Your screenshot must include the current date using the function Sys.Date(). Submit the Word file in Canvas for grading.

Mastery Exercise (10 points)

Module 3

Readings

- zyBooks – Module 3

Opening Exercise (0 points)

Discussion (25 points)

Critical Thinking (55 points)

Choose one of the following two assignments to complete this week. Do not do both assignments. Identify your assignment choice in the title of your submission.

Option #1: Regression Model—A Paper

How does Gross Domestic Product (GDP) affect currency exchange rates? One way to answer such a question is by using regression analysis. Complete the following:

a. Download the dataset GDP.xls from the course website.
b. Using R, create a scatter plot for GDP vs. US/EUR. Comment on the relationship.
c. Fit a linear regression model.
d. How good is the fit of the model from part c?
e. Describe how you can predict the US/EUR exchange rate?

Copy and paste the R outputs into a Word document and label each section clearly. If you take a screen shot, make sure that it shows the current date. Additionally, ensure you have answered all of the questions. Your well-written paper should be between 2-4 pages in length. Follow APA format, according to CSU-Global Guide to Writing and APA. Include a title page and reference page. Cite at least one outside academic source other than the textbook, course materials, or other information provided as part of the course materials.

Option #2: Regression Model—A Presentation

How does Gross Domestic Product (GDP) affect currency exchange rates? One way to answer such a question is by using regression analysis. Complete the following:

Download the dataset GDP.xls from the course website. See above.

a. Using R, create a scatter plot for GDP vs. US/EUR.
b. Comment on the relationship.
c. Fit a linear regression model.
d. How good is the fit of the model from part c?
e. Describe how you can predict the US/EUR exchange rate?

Copy and paste the R graph and output into an MS PowerPoint presentation. If you take a screen shot, make sure that it shows the current date. Your presentations should have six to ten slides. Every slide must contain a “Notes Section” with full explanation of the slide’s content.

Follow APA format, according to CSU-Global Guide to Writing and APA. Include a title slide and a slide citing references. These two slides are in addition to the 6-10 used in the presentation. Cite at least one outside academic source other than the textbook, course materials, or other information provided as part of the course materials.

For help with constructing an effective presentation, see the visual presentations page in the CSU-Global Writing Center. Submit either the PowerPoint file or a Word document containing the link to your online presentation.

Mastery Exercise (10 points)

Module 4

Readings

- zyBook – Module 4
- Loh, W. (2014). Fifty years of classification and regression trees. *International Statistical Review*, *82*(3), 329–348.

Opening Exercise (0 points)

Discussion (25 points)

Critical Thinking (55 points)

Choose one of the following two assignments to complete this week. Do not do both assignments. Identify your assignment choice in the title of your submission.

Option #1: Logistic Regression—A Paper

In this assignment, we examine classification using logistic regression. In R console, type *mtcars*. The dataset *mtcars* is a generic dataset in R. This dataset comprises of fuel consumption and 10 aspects of automobile design and performance for 32 automobiles. Using only the variables *am* (0 = automatic, 1 = manual) and *mpg*, your task is to fit a logistic regression model. Complete the following steps using R.

a. Create a scatter plot of am vs. mpg. Describe the relationship and explain why a simple linear regression model may not be suitable.
b. Using the variables *am* and *mpg*, fit a logistic regression model. Use the function glm().
c. Write out the estimated model.
d. Suppose a car has 16 mpg. How would you classify the transmission: automatic or manual? Explain and show how you classified the transmission.

Copy and paste the appropriate R graph and output into a Word document. If you take a screen shot make sure that it shows the current date. Additionally, ensure you have answered all of the questions. Your well-written paper should be between 2-4 pages in length. Follow APA format, according to CSU-Global Guide to Writing and APA. Include a title page and reference page. Cite at least one outside academic source other than the textbook, course materials, or other information provided as part of the course materials.

Option #2: Logistic Regression—A Presentation

In this assignment, we examine classification using the logistic regression. In R console, type *mtcars*. The dataset *mtcars* is a generic dataset in R. This dataset comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles. Using only the variables *am* (0 = automatic, 1 = manual) and *mpg*, your task is to fit a logistic regression model. Complete the following steps using R.

a. Create a scatter plot of am vs. mpg. Describe the relationship and explain why a simple linear regression model may not be suitable.
b. Using the variables *am* and *mpg*, fit a logistic regression model. Use the function glm().
c. Write out the estimated model.

Suppose a car has 16 mpg. How would you classify the transmission: automatic or manual? Explain and show how you classified the transmission.

Copy and paste the appropriate R graph and output into an MS PowerPoint presentation. If you take a screenshot make sure that it shows the current date. Your presentations should have six to ten slides. Every slide must contain a “Notes Section” with full explanation of the slide’s content.

Follow APA format, according to CSU-Global Guide to Writing and APA. Include a title slide and a slide citing references. These two slides are in addition to the 6-10 used in the presentation. Cite at least one outside academic source other than the textbook, course materials, or other information provided as part of the course materials.

For help with constructing an effective presentation, see the visual presentations page in the CSU-Global Writing Center. Submit either the PowerPoint file or a Word document containing the link to your online presentation.

Mastery Exercise (10 points)

Portfolio Milestone (50 points)

Choose one of the following two assignments to complete this week. Do not do both assignments. Identify your assignment choice in the title of your submission.

Option #1: House Price Prediction

Please read the final Portfolio Project in Module 8. You will need to choose one of the two options. If you choose to do Portfolio Project Milestone Option #1, then you need to complete the same option in Module 8. Download the dataset house.training.csv. This dataset contains 25 quantitative explanatory variables describing many aspects of residential homes in Ames, IA. The response variable is the sale price. More description is available from:

https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data

Using R, calculate the summary statistics (minimum, maximum, mean, median, and standard deviation) and create a histogram of sale price for this dataset. Describe the summary and shape of the distribution of sale price.

Copy and paste all the R output for summary statistics and histogram into a Word document. If you take a screenshot, make sure that it shows the current date. Submit the file in Canvas for grading.

Option #2: A Research Paper in Prediction/Forecasting

Please read the final Portfolio Project in Module 8. You will need to choose one of the two options. If you choose to do Portfolio Project Milestone Option #2, then you need to complete the same option in Module 8. This option of the Portfolio Project gives flexibility in choosing your own work. However, the focus of the project must be on predictions and/or forecasting. For instance, you may investigate how prediction or forecasting is used in the music industry.

Using a Word document, please provide the following:

1. The objective and/or motivation of the study.
2. A background of the study.
3. A summary on what you will investigate. This includes a description of dataset (including the location of dataset) and methodologies you will use.
4. At least two citations.

Submit a 1-2-page summary in a Word document in Canvas. Watch for feedback from your instructor to help further guide your process as you conduct your analyses.

Module 5

Readings

- zyBook – Module 5
- Koch, R. (2015). From business intelligence to predictive analytics. *Strategic Finance*, *96*(7), 56-58.

Opening Exercise (0 points)

Discussion (25 points)

Critical Thinking (55 points)

Choose one of the following two assignments to complete this week. Do not do both assignments. Identify your assignment choice in the title of your submission.

Option #1: Predicting Behavior with Logistic Regression

Customer churn occurs when customers stop doing business with a company. Retaining existing customers is less expensive than it is to acquire new customers and hence, building a good predictive model for customer churn is of importance to many companies. Download the dataset *Telco.customer.csv* from the course website. Through this dataset, we attempt to predict behavior to retain customers using logistic regression.

Follow the steps below and create a PowerPoint presentation.

1. Using R, partition the dataset into training and testing sets by using the code: (YOURDATA is the name of your dataset in R.)

   ```
   intrain<- createDataPartition(YOURDATA$Churn,p=0.7,list=FALSE)
   set.seed(2017)
   training<- YOURDATA[intrain,]
   testing<- YOURDATA[-intrain,]
   ```

2. 2. Fit a logistic regression model by using the code:

   ```
   glm(Churn ~ .,family=binomial(link="logit"),data=training)
   ```

3. Examine the resulting fitted model. What are the significant factors that affect customer churn? Explain how and why they are significant.

4. Now, let’s examine how the model fits using the following code.

   ```
   testing$Churn <- as.character(testing$Churn)
   testing$Churn[testing$Churn=="No"] <- "0"
   testing$Churn[testing$Churn=="Yes"] <- "1"
   fitted.results <- predict(LogModel,newdata=testing,type='response')
   fitted.results <- ifelse(fitted.results > 0.5,1,0)
   misClasificError <- mean(fitted.results != testing$Churn)
   print(paste('Logistic Regression Accuracy',1-misClasificError))
   ```

   This provides the accuracy of the model.

5. How can you make a customer churn prediction from the model you fitted? Explain. Although it is not required, the actual calculation may help your understanding.

Provide the output from each set of code and your answer and/solution to the questions in an MS PowerPoint presentation. Your presentations should have eight to twelve slides. Every slide must contain a "Notes Section" with full explanation of the slide's content.

Follow APA format, according to CSU-Global Guide to Writing and APA. Include a title slide and a slide citing references. These two slides are in addition to the eight to twelve used in the presentation. Cite at least one outside academic source other than the textbook, course materials, or other information provided as part of the course materials.

For help with constructing an effective presentation, see the visual presentations page in the CSU-Global Writing Center. Submit either the PowerPoint file or a Word document containing the link to your online presentation.

Option #2: Predicting Behavior with Decision Trees

Customer churn occurs when customers stop doing business with a company. Retaining existing customers is less expensive than it is to acquire new customers and hence, building a good predictive model for customer churn is of importance to many companies. Download the dataset *Telco.customer.csv* from the course website. Through this dataset, we attempt to predict behavior to retain customers using classification tree.

Follow the steps below and create a PowerPoint presentation.

1. Using R, partition the dataset into training and testing sets by using the code: (YOURDATA is the name of your dataset in R.)

   ```
   intrain<- createDataPartition(YOURDATA$Churn,p=0.7,list=FALSE)
   set.seed(2017)
   training<- YOURDATA[intrain,]
   testing<- YOURDATA[-intrain,]
   ```

2. Fit a decision tree model by using the code:

   ```
   tree <- ctree(Churn~gender+MultipleLines+StreamingTV+StreamingMovies+Contract, training)
   plot(tree)
   ```

3. Examine the resulting fitted model. What are the significant (most important) factors that affect customer churn? Explain how and why they are significant (important).

4. Now, let's examine how the model fits using the following code.

   ```
   p1 <- predict(tree, training)
   tab1 <- table(Predicted = p1, Actual = training$Churn)
   pred_tree <- predict(tree, testing)
   tab2 <- table(Predicted = pred_tree, Actual = testing$Churn)
   print(paste('Decision Tree Accuracy',sum(diag(tab2))/sum(tab2)))
   ```

This provides the accuracy of the model.

5. How can you make a customer churn prediction from the model you fitted? Explain. Although it is not required, the actual calculation may help your understanding.

Provide the output from each set of code and your answer and/solution to the questions in an MS PowerPoint presentation. Your presentations should have 8-12 slides. Every slide must contain a "Notes Section" with full explanation of the slide's content.

Follow APA format, according to CSU-Global Guide to Writing and APA. Include a title slide and a slide citing references. These two slides are in addition to the 8-12 used in the presentation. Cite at least one outside academic source other than the textbook, course materials, or other information provided as part of the course materials.

For help with constructing an effective presentation, see the visual presentations page in the CSU-Global Writing Center. Submit either the PowerPoint file or a Word document containing the link to your online presentation.

Mastery Exercise (10 points)

Module 6

Readings

- Chapters 6, 7, 9, 10 in An Introduction to R
- Kopf, D. (2018, November 17). What's the best way to learn the programming language R? (Preferably, for free). Retrieved from https://qz.com/1464525/whats-the-best-way-to-learn-the-programming-language-r-preferably-for-free/
- Machlis, S. (2017, August 18). Beginner's guide to R: Introduction. Retrieved from https://www.computerworld.com/article/2497143/business-intelligence/business-intelligence-beginner-s-guide-to-r-introduction.html

Opening Exercise (0 points)

Discussion (25 points)

Critical Thinking (55 points)

Choose one of the following two assignments to complete this week. Do not do both assignments. Identify your assignment choice in the title of your submission.

Option #1: Custom Functions in R - Histograms

In this assignment, we will build custom functions in R. As an example, the following function called *addPercent* converts a value into a percentage with one decimal place.

```
addPercent <- function(x){
  percent <- round(x*100, digits = 1)
  result <- paste(percent, "%", sep = "")
  return(result)
}
```

Below are a few output results from this function.

```
> addPercent(.1)
[1] "10%"
> addPercent(10)
```

```
[1] "1000%"
> addPercent(10.1)
[1] "1010%"
> addPercent(0.1443)
[1] "14.4%"
```

1. Write a custom R function that inputs a temperature in Fahrenheit F^{o} and converts to Celsius C^{o}. The relationship is C^{o} = 5(F^{o} – 32)/9.
2. Write a custom R function that computes the sum of squares of two numbers.
3. Write a custom R function that takes any univariate dataset and calculates the mean, minimum, maximum, and standard deviation.
4. In statistics, a dataset needs to be transformed in order to meet certain assumptions. Write a custom R function that takes any univariate dataset and creates a histogram of the raw dataset and a histogram of the log-transformed dataset.
5. Write a custom R function of your own. Describe what your function does and produce the output.

Using a Word document, include your all your functions along with least four different output results from each of the functions 1 through 5. You may either copy and paste the codes and output results on a Word document or take the screen shots. If you take the screen shots, make sure to show the current date.

Option #2: Custom Functions in R - Boxplots

In this assignment, we will build custom functions in R. As an example, the following function called *addPercent* converts a value into a percentage with one decimal place.

```
addPercent <- function(x){
  percent <- round(x*100, digits = 1)
  result <- paste(percent, "%", sep = "")
  return(result)
}
```

Below are a few output results from this function.

```
> addPercent(.1)
[1] "10%"
> addPercent(10)
[1] "1000%"
> addPercent(10.1)
[1] "1010%"
> addPercent(0.1443)
[1] "14.4%"
```

1. Write a custom R function that inputs a temperature in Celsius C^{o} and converts into Fahrenheit F^{o}. The relationship is F^{o} = (9/5)C^{o} + 32.
2. Write a custom R function that computes the difference of squares of two numbers.
3. Write a custom R that takes any univariate dataset and calculate the mean, minimum, maximum, and standard deviation.
4. In statistics, a dataset often needs to be transformed in order to meet certain assumptions. Write a custom R function that takes any univariate dataset and creates a boxplot of the raw dataset and a histogram of the square root transformed dataset.

5. Write a custom R function of your own. Describe what your function does and produce the output.

Using a Word document, include your all your functions along with least four different output results from each of the functions 1 through 5. You may either copy and paste the codes and output results on a Word document or take the screen shots. If you take the screen shots, make sure to show the current date.

Mastery Exercise (10 points)

Module 7

Readings

- zyBook – Modules 7 & 8
- A Relational Database Overview. Retrieved from https://docs.oracle.com/javase/tutorial/jdbc/overview/database.html
- PostGreSQL Installation

Opening Exercise (0 points)

Discussion (25 points)

Critical Thinking: Networking (40 points)

Connections to other professionals and organizations are important in the world of business and performance management. Though credentials and experience are important, many opportunities arise because of who you know—those individuals and organizations that have become part of your professional network.

How do you build a strong professional network and brand? Review the following resources and the material in the CSU-Global Career Center for information on LinkedIn and professional networking:

- How to Build a Powerful Professional Network (Links to an external site.)
- LinkedIn 201: How to Cultivate a Powerful Network (Links to an external site. Don't miss the second page of this article!)
- Networking on LinkedIn (Video from Lynda.com; sign in with your CSU-Global credentials.)
- Why You Should Be Using LinkedIn Groups (links to an external site).

For this assignment, you need to complete three networking tasks:

- Build a 90% complete profile on LinkedIn to include your background, education, experience, skills, and accomplishments. Ensure your professional brand is consistent across social and professional platforms. (Consult the CSU-Global Career Center for more help with LinkedIn and networking).
- Explore professional organizations within the degree field and consider the value of becoming a member.
- Find and establish a mentor/coach in the field. Make contact with this professional.

To complete this assignment, submit the following in a one-page document: the URL to your LinkedIn public profile, a summary of and value assessment of one or more organizations in your field, and the name of your mentor/coach. Also, provide the reason why you chose this mentor/coach.
Review the rubric for specific grading criteria.

Mastery Exercise (10 points)

Module 8

Readings

- zyBook – Module 7 & 8
- Parker, Z., Poe, S., & Vrbsky, S. V. (2013, April). Comparing NoSQL MongoDB to an SQL DB. In *Proceedings of the 51st ACM Southeast Conference*. ACM.

Opening Exercise (0 points)

Discussion (25 points)

Mastery Exercise (10 points)

Portfolio Project (300 points)

Choose one of the following two assignments to complete this week. Do not do both assignments. Identify your assignment choice in the title of your submission.

Option #1: House Price Prediction

In real estate, housing market prediction (forecasting) is crucial. There are many factors that may influence the house prices. The datasets *housing.training.csv* and *housing.testing.csv* contain 25 quantitative explanatory variables describing many aspects of residential homes in Ames, IA.

The goal of this project is to predict house prices. To this end, we will be using regression analysis.

1. In Week 4 Portfolio Milestone, you've examined housing.training.csv dataset. Now, examine housing.testing.csv dataset and perform the same tasks as given in Week 4 Portfolio Milestone. Using R, calculate the summary statistics (minimum, maximum, mean, median, and standard deviation) and create a histogram of sale price for each dataset. Comparing with *housing.training,csv* dataset, describe the similarities and/or differences.
2. Combine the two datasets *housing.training.csv* and *housing.testing.csv*. This can be done in R by using the function *combine()*. Create a histogram of sale prices for the combined dataset and compare it with the histograms from training and testing datasets. Describe the similarities and differences.
3. Using only the dataset *housing.training.csv*, fit a linear regression model using all the explanatory variables and SalePrice as the response variable.
4. What are the significant factors? How do these variables relate to the sale price? Interpret your estimated model.
5. Remove all the rows with missing values (NA) from the dataset *housing.testing.csv*. The function *complete.cases()* can be used. Using only the first 20 rows from *housing.testing.csv*, predict the sale price. The R function *predict()* can perform this task. You should have 20 predicted sale prices.
6. Compare the predicted sale prices to the actual sale prices from the *housing.testing.csv* dataset (the first 20 rows). How good is your prediction?

For each R output result, you may either type directly into a Word document or take a screenshot. If you take the screenshot, make sure that the current date is shown.

Ensure everything is clearly labeled. The report must be 10-12 pages long, *including* a title page and reference page (the report itself should be 8-10 pages). Cite three scholarly sources other than the textbook, course materials, or other information provided as part of the course materials. Follow APA format, according to CSU-Global Guide to Writing and APA.

Option #2: A Research Paper in Prediction/Forecasting

Using the topic(s) you’ve chosen in Module 4 Portfolio Project Milestone, continue and finish the project.

As an example, “What are the factors that may affect the gasoline price at the pump in the US? How can we make gasoline price predictions?” To answer such questions, examine what has been found in the literature using scholarly citations. Below is an example of scholarly citation.

Liddle, B. (2009). Long-run relationship among transport demand, income, and gasoline price for the US. *Transportation Research Part D: Transport and Environment*, *14*(2), 73-82.

Use APA-style references wherever necessary to support your findings and discussions. See the MIS Library Guide for access to relevant databases and research.

At a minimum, the report must contain the following:

1. A relevant title.
2. An abstract giving the overview of the whole study.
3. A body describing the dataset, method, and results from the method by using your dataset.
4. Conclusion/Discussion summarizing the study.
5. At least ten academic scholarly citations excluding blogs and personal or commercial websites.

The written report must be 10-12 pages long, *including* a title page and reference page (the report itself should be 8-10 pages). It should have an introduction stating the problem definition or questions, a body that contains the supporting analysis (including descriptive and graphical summaries using R), and a conclusion paragraph that addresses your findings. Cite three scholarly sources other than the textbook, course materials, or other information provided as part of the course materials. Follow APA format, according to CSU-Global Guide to Writing and APA.

COURSE POLICIES

Grading Scale	
A	95.0 – 100
A-	90.0 – 94.9
B+	86.7 – 89.9
B	83.3 – 86.6
B-	80.0 – 83.2
C+	75.0 – 79.9
C	70.0 – 74.9
D	60.0 – 69.9
F	59.9 or below

Course Grading

20%	Discussion Participation
0%	Opening Exercises
8%	Mastery Exercises
37%	Critical Thinking Assignments
35%	Final Portfolio Project

In-Classroom Policies

For information on late work and incomplete grade policies, please refer to our **In-Classroom Student Policies and Guidelines** or the Academic Catalog for comprehensive documentation of CSU-Global institutional policies.

Academic Integrity
Students must assume responsibility for maintaining honesty in all work submitted for credit and in any other work designated by the instructor of the course. Academic dishonesty includes cheating, fabrication, facilitating academic dishonesty, plagiarism, reusing /repurposing your own work (see CSU-Global Guide to Writing & APA for percentage of repurposed work that can be used in an assignment), unauthorized possession of academic materials, and unauthorized collaboration. The CSU-Global Library provides information on how students can avoid plagiarism by understanding what it is and how to use the Library and internet resources.

Citing Sources with APA Style
All students are expected to follow the CSU-Global Guide to Writing & APA when citing in APA (based on the most recent APA style manual) for all assignments. A link to this guide should also be provided within most assignment descriptions in your course.

Disability Services Statement
CSU-Global is committed to providing reasonable accommodations for all persons with disabilities. Any student with a documented disability requesting academic accommodations should contact the Disability Resource Coordinator at 720-279-0650 and/or email ada@CSUGlobal.edu for additional information to coordinate reasonable accommodations for students with documented disabilities.

Netiquette
Respect the diversity of opinions among the instructor and classmates and engage with them in a courteous, respectful, and professional manner. All posts and classroom communication must be conducted in accordance with the student code of conduct. Think before you push the Send button. Did you say just what you meant? How will the person on the other end read the words?

Maintain an environment free of harassment, stalking, threats, abuse, insults, or humiliation toward the instructor and classmates. This includes, but is not limited to, demeaning written or oral comments of an ethnic, religious, age, disability, sexist (or sexual orientation), or racist nature; and the unwanted sexual advances or intimidations by email, or on discussion boards and other postings within or connected to the online classroom. If you have concerns about something that has been said, please let your instructor know.