**2020-2021 Data Analytics**

**Semester 1 Exam Information**

On Semester 1 Exam day (December 9th for A1 and December 10th for B1), you are expected to demonstrate your understanding of three standards (Analysis, Evaluation, and Coding Language). You will be completing an R programming task, and write an analysis and evaluation for one specific representation.

In order to prepare for the exam, you are required to:

1. Read Chapter 3 (page 50-67) of the Weapons of Math Destruction by Cathy O’Neil.
2. Familiarize yourself with the IPEDS data and US News Ranking data
   1. IPEDS is the Integrated Postsecondary Education Data System. The dataset contains information from the year 2013. It has information from 1534 universities in the United States. The dataset has 145 columns.
   2. In 1983, U.S. News & World Report began to evaluate 1800 colleges and universities throughout the United States and rank them for excellence. It continues to do so until now.

During our three-day exam preparation sessions, we will be using the two datasets presented above.

Before exam day, please make sure you complete the following:

1. Go to bit.ly/2020sasda\_rstudio and click on the “Semester 1 Exam - Student Documents” folder
2. Download the following files and place them in the folder for which you will set your RStudio working directory on the day of the exam.
   1. Sem1Exam.R
   2. IPEDS\_data.csv
   3. ranking\_data.csv
3. Open Sem1Exam.R in your RStudio
4. Run all 25 lines of codes
   1. Out of the 1534 universities listed, how many schools have been in US News & World Report Top 50 ranking?
   2. Out of the 1534 universities listed, how many schools have not been in the Top 50 ranking?
5. Familiarize yourself with the IPEDS dataset and US News & World Report dataset.
   1. What do the columns represent?
   2. What information is provided for you?
   3. What do these columns look like in R?
   4. Which column will tell you whether or not the university is classified as a “historically black colleges and universities”?
   5. Which column will give you the percentage of women enrolled in the school?
   6. Which column will give you the percentage of international students in the school?
   7. What information does the column X2013 give you?
6. Should you have questions about the dataset, please ask prior to the day of the exam. Any questions about the structure of the dataset will not be answered on the day of the exam.

**Session 35:** Final Exam Preparation Pre-Practice

1. Run the following line. What information does this line give you? What does it mean?

|  |
| --- |
| table(uni$ranking, uni$Historically.Black.College.or.University) |

1. Run the following line. What information does this line give you? What does it mean?

|  |
| --- |
| prop.table(table(uni$ranking, uni$Historically.Black.College.or.University),1)\*100 |

1. Is it true that the tuition and fees for ranked universities are higher than the tuition and fees for non-ranked universities? Use the boxplot to explain your answer.

|  |
| --- |
| boxplot(uni$Tuition.and.fees..2013.14~uni$ranking) |

1. Explain the difference between the following boxplot functions.

[Do not try to run the codes in R. It won’t work. The x, y, x1, x2 are only examples in this question.]

|  |  |
| --- | --- |
| boxplot(x~y) | boxplot(x1, x2) |

1. What information does the following lines of code give you? What conclusions can you draw from the representation? What are the limitations?

|  |
| --- |
| boxplot(uni$Tuition.and.fees..2010.11, uni$Tuition.and.fees..2011.12,  uni$Tuition.and.fees..2012.13, uni$Tuition.and.fees..2013.14) |

1. Label the following boxplot with a title, x-axis label, and y-axis label so that the representation provides sufficient information for understanding

|  |
| --- |
| boxplot(uni$Tuition.and.fees..2010.11, uni$Tuition.and.fees..2011.12,  uni$Tuition.and.fees..2012.13, uni$Tuition.and.fees..2013.14) |

1. What conclusions can you make from this data representation?

|  |
| --- |
| plot(uni$X2013, uni$Tuition.and.fees..2013.14) |

1. What conclusions can you make from this data representation? What are the limitations?

|  |
| --- |
| highest\_degree <- table(uni$ranking, uni$Highest.degree.offered)  barplot(highest\_degree,  main="Highest Degrees Offered in Top 50 vs. Not Ranked Universities",  xlab="Highest Degrees Offered", col=c("darkblue","red"),  ylab="Number of Universities",  ylim=c(0,600),  names.arg=c("Bachelor's Degree", "D-Professional",  "D-Other", "D-Research",  "D-Professional&Research", "Master's"), cex.names=0.8)  legend("topleft", col=c("darkblue","red"),  c("Not ranked", "Top 50"), pch = 15, inset = .04) |

1. Run the linear regression model below.

|  |
| --- |
| uni$ranking <- as.factor(uni$ranking)  color\_type = c("red", "blue")[uni$ranking]  plot(uni$Tuition.and.fees..2013.14,  uni$Graduation.rate...Bachelor.degree.within.4.years..total,  col=color\_type)  model1 <- lm(uni$Graduation.rate...Bachelor.degree.within.4.years..total~  uni$Tuition.and.fees..2013.14)  summary(model1)  abline(model1) |

1. Summarize the general trend that you see in the representation.
2. Interpret the slope of the line of best fit.
3. What questions might you have about the representation and dataset?
4. What return value does the following code give? Explain what the code does.

|  |
| --- |
| subset(subset(uni, ranking == "Top 50"),  Graduation.rate...Bachelor.degree.within.4.years..total <20) [,2] |

1. Run the following linear regression.

|  |
| --- |
| plot(uni$Tuition.and.fees..2013.14,  uni$Graduation.rate...Bachelor.degree.within.5.years..total,  col=color\_type)  model2 <- lm(uni$Graduation.rate...Bachelor.degree.within.5.years..total~  uni$Tuition.and.fees..2013.14)  summary(model2)  abline(model2) |

1. Summarize the general trend that you see in the representation.
2. How is this different from the scatterplot and linear regression in #9?
3. Run the linear regression.

|  |
| --- |
| plot(uni$X2013,  uni$SAT.Math.75th.percentile.score)  model3 <- lm(uni$SAT.Math.75th.percentile.score~uni$X2013)  abline(model3)  summary(model3) |

1. Interpret the slope.
2. What conclusions can you make from this data representation? What are the limitations?