**Student name: Trevor Cardoza SMC username: TMC12**

Instructions: Fill out your full name and SMC username above. Answer each question in its respective following space on this answer sheet. **Give a sufficient yet succinct answer for each question, and show figures / tables / graphs on this same sheet when appropriate or explicitly requested.** After completion, rename and save this file as **“Assignment 3\_BUSAD 040\_*your full name*” in either MS Word (.docx or .doc) or PDF (.pdf) format**, then upload and submit it on Moodle by the due date. No other file formats will be accepted. Late submission will not be accepted.

Note: This **answer sheet** together with the **one** completed **Excel file** “**2011Movies.xlsx**” and **one R script file** created and named “**hw3.R**” by yourself are the three files you need to submit on Moodle after completion.

**Total possible points: 100 points**

1. **Chapter 3 questions - Motion Picture Industry (50 points in total)**

The motion picture industry is a competitive business. More than 50 studios produce several hundred new motion pictures each year, and the financial success of the motion pictures varies considerably. The opening (weekend) gross sales, the total gross sales, the number of theaters the movie was shown in, and the number of weeks the motion picture was in release are the four common variables used to measure the success of a motion picture. Data on the top 100 grossing motion pictures released in 2011 (Box Office Mojo website, March 17, 2012) are contained in a provided accompanying Excel file named **2011Movies.xlsx**. Note that some movies, such as War Horse, were released late in 2011 and continued to run in 2012.

Use the given Excel file with the appropriate **Excel functions** to complete the following questions. Save your completed Excel file and upload it together with your answer sheet on Moodle.

1.1 Calculate and show the descriptive statistics of **mean**, **standard deviation**, as well as the **five-number summary** (i.e., minimum, first quartile, median, third quartile, and maximum) for **EACH** of the **four variables** in the Excel dataset in the following space. Examining the calculated descriptive statistics for the four variables, do you think whether there was a wide variation in the performance of motion pictures? Why or why not? (15 points)

Yes the standard deviation is wide and describes the major difference between the 1st and 3rd quartile.

1.2 List and compare the **mean** and **median** (from the descriptive statistics you calculated in part 1.1) for **EACH** of the **four variables** in the following space. Do you see that all the means are greater than their corresponding medians? If so, what do you think caused the inflation of the means? Which measure, the means or the medians, do you think can give us a better picture of the middle or more typical performance characteristics in the motion picture industry? Why? (10 points)

The mean, because it takes every single point into consideration for the middle ground.

1.3 What motion pictures, if any, should be considered **high-performance outliers** in terms of the variable of total gross sales with the outlier detection method of Q3 + 1.5(IQR)? (Hint: you can generate a box plot for the variable in Excel and change the box plot’s design to show the outliers’ numbers by selecting the box plot first, then clicking Design (under the Chart Tools contextual menu), then choosing Quick Layout (under the Chart Layouts group), and finally choosing Layout 2, and now you may find and list the high-performance movie outlier(s)’ names according to their shown total gross sales numbers from the Excel dataset.) (15 points)

Harry Potter is way higher than the norm and should be considered an outlier.

1.4 Calculate and show the **sample correlation coefficients** between **total gross sales** and **EACH** of **the other three variables**. Are the correlation coefficients positive or negative? What do they mean (sign and magnitude) for the relationship between the corresponding two variables? Among the other three variables, which would be the best predictor of total gross sales? Why? (10 points)

They are positive

There are positive correlations between all of the variables and total gross sales, meaning they have relation in some capacity.

Opening Gross sales would be the greatest indicator

1. **Numerical measures in R (50 points in total)**

For the following questions, you should develop and run the R commands in a **R Script file** created and named “**hw3.R**” **by yourself** in RStudio. Save your completed R script file “**hw3.R**” and upload it together with your answer sheet on Moodle. You should also **copy and paste the relevant original R commands as well as the run results** for each question in its respective following space on this answer sheet.

Note: The given CSV file “**Colleges.csv**” is only used for data importing into your RStudio session, so you don’t need to submit it again for this assignment.

2.1 You are going to **use R** to compute some **numerical measures** of descriptive statistics for some variable(s) from the external CSV data file called “**Colleges.csv**”. So the first thing in your “**hw3.R**” script file you need to do is to import the external CSV data file into your current RStudio session and store it into a new variable called “**colleges**”. Write down and run the R code you would use for this importing and storage purpose. Now the data is in the variable “**colleges**” which is a data frame. Then use and run the function **head( )** on the variable “**colleges**” to show the **first 10 rows only** in the data frame. Copy, paste and show all the relevant R commands and results here in the following space. (5 points)

*Hint: if you are unsure about how to use the head( ) function, you can use the help( ) function to learn more about the head( ) function.*

colleges <- read.csv("Colleges.csv", header=TRUE, sep=",")

head(colleges)

YearFounded Tuition GraduateRate

1 1893 36697 79

2 1845 29754 70

3 1951 23680 68

4 1904 13572 37

5 1863 40542 91

6 1839 39864 84

2.2 Now you are ready to use R to compute some numerical measures for a variable/column in the data frame “**colleges**”. First, compute the (sample) **mean** for the variable/column “**Tuition**” in the data frame “**colleges**” and assign the result to a new variable called “**average**”. Then, compute the **median** for this same variable and assign the result to a new variable called “**tu.median**” (which means tuition’s median). Copy, paste and show all the relevant R commands and results here in the following space. So, what are the mean and median for Tuition? Are the two values (mean and median) equal? If not, which value is greater? What can you interpret from the (possible) difference between the mean and median, that is, is the shape of Tuition skewed? If so, is it skewed to the right or left? And why? (10 points)

average <- mean(colleges$Tuition)

tu.median <- median(colleges$Tuition)

average 32759.17

tu.median 33005L

they are not equal and the median is bigger skewed to the left because the numbers lean more on the lower end of the Median then the higher end.

2.3 This time, compute the **skewness** numerical measure on **Tuition** by using the **Skew( )** function available in an external package called “**DescTools**”. Install the package first and attach it to your current RStudio session. Then run the **Skew( )** function on **Tuition** with the optional argument **method = 2** and assign it to a new variable called “**skewness**”. Run the variable **skewness** and show the result. Copy, paste and show all the relevant R commands and results here in the following space. What is the skewness value for Tuition? Is it positive, zero, or negative? How do you interpret it in terms of the skewness of the shape of Tuition? Does it agree with the interpretation you had in part 2.2? (10 points)

library(DescTools)

skewness <- Skew(colleges$Tuition,method = 2)

skewness -0.880601

it is negative which leans towards the lower end of the values to the left. This matches my interpretation.

2.4 Compute the three quartiles **Q1**, **Q2** and **Q3** for **Tuition** **simultaneously** using the **quantile( )** function and store the result into a new variable called “**percentiles**”. Copy, paste and show all the relevant R commands and results here in the following space. Run the variable **percentiles** and show the result. Copy, paste and show all the relevant R commands and results here in the following space. What is the value of **Q3** and how would you interpret the value of **Q3**? (5 points)

percentiles <- quantile(colleges$Tuition)

0% 25% 50% 75% 100%

4560.0 28357.0 33005.0 39097.5 43866.0

Q3 equals 33005.0

2.5 Use R to compute the **variance** as well as the **standard deviation** of **Tuition**, and assign them into two new variables called “**variance**” and “**std**” respectively. Run the two new variables to show the values of variance and standard deviations in the RStudio console. Copy, paste and show all the relevant R commands and results here in the following space. (5 points)

variance <- Var(colleges$Tuition)

std <- SD(colleges$Tuition)

Variance = 54908100

Std = 7410

2.6 Compute the **five-number summary** on **Tuition** using the **summary( )** function. Copy, paste and show the relevant R commands and results here in the following space. Do they agree to the relevant numbers you computed before in the previous questions? (5 points)

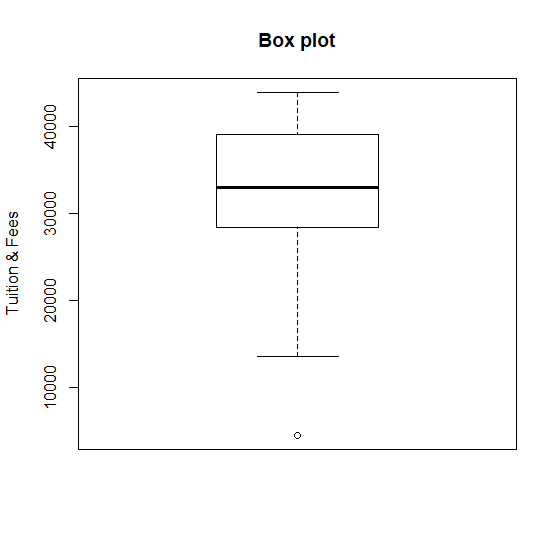
Min. 1st Qu. Median Mean 3rd Qu. Max.

4560 28357 33005 32759 39098 43866

Yes they are exactly the same

2.7 Develop a **box plot** on **Tuition** using the **boxplot( )** function. **Assign the main title as “Box plot” and the y axis label as “Tuition & Fees”.** Copy, paste and show the relevant R command and the box plot as an image here in the following space. According to the box plot, is there any outlier detected? If so, which tuition value is taken as the outlier? (5 points)

boxplot(colleges$Tuition,main = "Box plot",ylab = "Tuition & Fees")



Outlier detected at value 4560

2.8 Use R to compute the **covariance** and the **correlation coefficient** between **Tuition** and **GraduateRate**, and assign them into two new variables called “**covariance**” and “**correlation**” respectively. Run and show the results of the two new variables. Copy, paste and show all the relevant R commands and results here in the following space. What does the **correlation coefficient** value tell you? Is the linear relationship between tuition and graduate rate positive, none or negative? Is it a weak or strong linear relationship? Do your conclusions agree with the **covariance** value? (5 points)

covariance <- Cov(colleges$Tuition,colleges$GraduateRate)

correlation <- Cor(colleges$Tuition,colleges$GraduateRate)

covariance = 77889.42

correlation = 0.7235392

It tells me that as tuition goes up, so does the graduation rate at the college.

The value is positive.

Yes they are both positive