**Homework #9 (Due Oct 29 11:59 PM)**

IST 3420 - Fall 2017, Chen

**Name**: \_\_\_\_\_\_\_\_Adam Forestier\_\_\_\_\_\_\_\_\_\_\_\_

**Correlation, Missing Data, Outliers (10 points)**

***Tasks:***

1. Download the data file “data.xlsx”.
2. Install package “xlsx” in order to read the Excel file into R.
3. Use the read.xlsx() function in the “xlsx” package to read the data file. (1 point)
4. Use cor() function with the whole dataset as the parameter to calculate the pair-wise correlation coefficients. Paste your correlation matrix below. (1 point)

|  |
| --- |
| x y z  x 1.0000000 0.5835267 0.4552331  y 0.5835267 1.0000000 0.5238227  z 0.4552331 0.5238227 1.0000000 |

1. Use the boxplot rule to detect outliers in x variable. List outliers in x variable below. (2 points)

|  |
| --- |
| The outliers are: 11.23 and 9.56 |

6. Use dplyr::filter() to remove all outliers of x from the dataset. (1 point)

7. Use cor() function to re-calculate the pair-wise correlation coefficients on the updated dataser. Paste your correlation matrix below. (1 point)

|  |
| --- |
| x y z  x 1.0000000 0.7622366 0.7060406  y 0.7622366 1.0000000 0.5304971  z 0.7060406 0.5304971 1.0000000 |

8. Are the two correlation coefficient matrices different? Explain the reasons below. (2 points)

|  |
| --- |
| The correlation coefficient matrices are different. This is because two outlier values in x were filtered out so they no longer effected the correlation matrix |

9. Paste your R Markdown script code in the following box (1 point).

|  |
| --- |
| Your R Markdown Code:  ---  title: "Homework 9"  author: "Adam Forestier"  date: "October 23, 2017"  output: html\_document  ---  ```{r setup, include=FALSE}  knitr::opts\_chunk$set(echo = TRUE, message = F)  ```  Clean the current environement  ```{r}  rm(list = ls())  ```  Load Packages  ```{r}  # Used openxlsx because rJava would not load, I believe this is due to an update because I do have Java on my pc  library(openxlsx)  library(dplyr)  ```  # Read the data file  ```{r}  # Read in data  xlsx\_file <- read.xlsx("data.xlsx")  # Show data  str(xlsx\_file)  ```  # Calculate the pair-wise correlation coefficients  ```{r}  cor(xlsx\_file)  ```  # boxplot rule to detect outliers in x variable  ```{r}  box <- boxplot(xlsx\_file$x)  box  ```  # remove all outliers of x from the dataset  ```{r}  xlsx\_file <- dplyr::filter(xlsx\_file, !(x %in% box$out))  ```  # re-calculate the pair-wise correlation coefficients on the updated dataset  ```{r}  cor(xlsx\_file)  ``` |

Upload this document with your answers to “Homework 9” on Canvas.

Upload your R Markdown script file to “Homework 9” on Canvas. (1 point)