Lab 1 (100 points): Introduction to Statistical Packages
Objectives: Loading files, cleaning and manipulating the data.

A. (10 pts.) Online Prelab

B. (90 points) US Demographic, Crime, and Test-Score Data. This semester, we are going to be exploring some Demographic, Crime, and Test-Score data for counties across the United States. The data we will analyze are in the data set "USData.txt". The variable names and definitions are listed in the file "US_Data set_Definition.pdf". In this lab, we are going to explore what is included in the data set, load it into the software package, and do some basic manipulations.

1. (10 points) How many variables does this data set contain? Which are categorical or qualitative variables and which are quantitative or numeric variables? Besides looking at the documentation file provided, you might want to look at the data file itself in a spreadsheet, notepad or the software package (R only).

Solution:

This data set contains 20 variables.

- Categorical. There are 5: State, Region, CountyIndex, UrbanIndicator, and IncomeCategory.
- Quantitative. There are 15: Population, LandArea, PopulationDensity, PercentMaleDivorce, PercentFemaleDivorce, MedianIncome, PercentCollegeGraduates, MedianHouseAge, RobberiesPerPopulation, AssaultsPerPopulation, BurglariesPerPopulation, LarceniesPerPopulation, EducationSpending, EducationSpendingP2, and TestScore

2. (16 pts.) Write two analysis questions that can be answered from the data provided. In the project due at the end of the semester, your group will have to pose general questions that can be answered by three different statistical methods. You will be allowed to change the questions when you start the project, but this will get you thinking of possibilities.

Solution:

There are many possible answers, here are three examples:

- Does the number of robberies per 100,000 people differ significantly among the different regions of the US? Similar questions can be asked for the other crime variables.
- Is more Education Spending associated with higher Average Test Scores?
- Is a larger Percentage of College Graduates positively associated with Median Income?

3. (20 points) Load the data into your software package, and provide the programming code used to do so. If you used menu options to load the data, rather than code, please describe the procedure you followed. No output is required.

Solution:

Method 1

```
setwd("W:\\Labs")
USData <- read.table("USData.txt", header = TRUE, sep = "\t")
```

Method 2

Import Data set → From CSV → Browse to "USData.txt" → Change Delimiter to Tab → Click "Import."

```
str(USData)
USData <- as.data.frame(USData)
```

Remember that the output is not supposed to be included because it was not asked for.

4. (19 points) Are there missing values (NA) in the data set? If so, please create a new data set by removing any rows that contain one or more NAs from the original data set. Please save this new data set to your computer and/or ITaP folder; this will be the data set that you will be using for the rest of the semester.

   a. (5 pts.) Code

Solution:

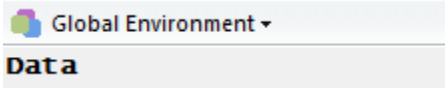
```
# Remove missing data
USData_cleaned <- USData[complete.cases(USData), ]
# Save data set
write.table(USData_cleaned, file = "USData_cleaned.txt", row.names = FALSE,
            sep = "\t")
# Determine rows
nrow(USData)
nrow(USData_cleaned)
```

   b. (9 pts.) We want to know how many rows were removed, so please answer:

      i. How many observations are there in the original data set? (The output is all that is required.)

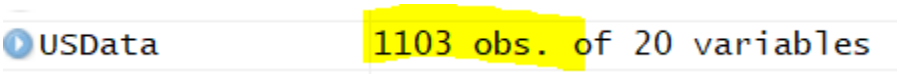
Solution:

Note: The second method is found under Global Environment on the upper right of RStudio.




```
> nrow(USData)
[1] 1103
```

OR

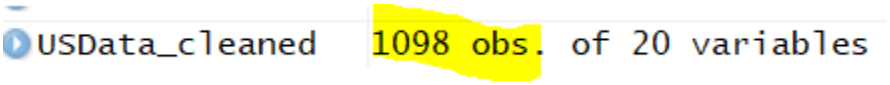


There were 1,103 observations in the original data set.

ii. How many observations are there after removing the incomplete data? (The output is all that is required.)

Solution:

```
> nrow(USData_cleaned)
[1] 1098
```



There are 1,098 observations after removing the incomplete data.

iii. How many rows were removed (show the work, even though it is a quick calculation)?

Solution:

1103 -1098 = 5

c. (5 pt.) In which directory did you save your cleaned data set?

Solution:

W:\
It doesn't matter what this answer here. I just asked the question so that you can refer to it for the rest of the semester.

5. (10 points) For readability, we want to transform the values of "UrbanIndicator" from a number to what the number represents. That is, please create a new variable called "UrbanNew" such that:
   If UrbanIndicator is "1", UrbanNew is "Urban" and
   If UrbanIndicator is "1", UrbanNew is ""Rural"

a. (5 pts.) Code. Remember that all code needed to answer part b) needs to be included in this part.

Solution:

```
USData_cleaned$UrbanNew <- as.character(USData_cleaned$UrbanIndicator)
USData_cleaned$UrbanNew[USData_cleaned$UrbanIndicator=="1"] <- "Urban"
USData_cleaned$UrbanNew[USData_cleaned$UrbanIndicator=="0"] <- "Rural"
USData_cleaned[c(5,55,355,555),c("State","CountyIndex","UrbanIndicator",
     "UrbanNew")]
```

b. (5 pts.) Print or display the data set (on the computer, not to physical paper), and take screen clippings which demonstrate the following rows: 5, 55, 355, and 555. Please highlight or somehow indicate the changes. These rows will prove that your code worked correctly. To save space, you are permitted to restrict the data set to show only the relevant columns and the columns for "State" and "CountyIndex."

Solution:

```
> USData_cleaned[c(5,55,355,555),c("State","CountyIndex","UrbanIndicator",
    "UrbanNew")]
# A tibble: 4 × 4
       State CountyIndex UrbanIndicator UrbanNew
       <chr>       <int>          <int>    <chr>
1    Alabama           5              0    Rural
2 California          11              1    Urban
3       Iowa           9              0    Rural
4   Missouri           7              1    Urban
```

Note that it is possible that the original numbers will print out.

6. (15 points) We are going to show that "PopulationDensity" can be calculated from other variables in the data set.
   a. (5 pts.) Write down the equation relating "PopulationDensity" to “Population” and “LandArea.”

Solution:

$$PopulationDensity = \frac{Population}{LandArea}$$

b. (5 pts.) Write code (and provide it here) to create a new variable called "PopulationDensityNew" which implements the calculation described in part a). Remember that all code needed to answer part c) needs "to be included in this part.

Solution:

```
USData_cleaned$PopulationDensityNew <-
     USData_cleaned$Population/USData_cleaned$LandArea
USData_cleaned[1:6, c("PopulationDensity", "PopulationDensityNew")]
```

c. (5 pts.) Show that your code is correct by displaying the original variable "PopulationDensity" and “PopulationDensityNew”. Please only print out the first 6 rows. To save space, you are permitted to restrict the data set to show only the relevant columns.

Solution:

```
> USData_cleaned[1:6, c("PopulationDensity", "PopulationDensityNew")]
# A tibble: 6 × 2
  PopulationDensity PopulationDensityNew
              <dbl>                <dbl>
1         1650.8333            1650.8333
2         1273.8278            1273.8278
3         1199.6209            1199.6209
4          626.8847             626.8847
5          532.6939             532.6939
6         3485.7143            3485.7143
```

Our calculation matches the true value.