Jordan Mayer
01/10/2018
STAT 350, 1:30, with Fan Wu
Lab 071

Lab 1 Report

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).

   This data set contains the following **20** variables:

Variable	**Type (C = Categorical, N = Numeric)**
State	C
Region	C
CountyIndex	C
UrbanIndicator	C
Population	N
LandArea	N
PopulationDensity	N
PercentMaleDivorce	N
PercentFemaleDivorce	N
MedianIncome	N
IncomeCategory	C
PercentCollegeGraduates	N
MedianHouseAge	N
RobberiesPerPopulation	N
AssaultsPerPopulation	N
BurglariesPerPopulation	N
LarceniesPerPopulation	N
EducationSpending	N
EducationSpendingP2	N
TestScore	N

2. (16 points) 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.

   1) **What sort of correlation exists between median household income and average test score?**
   2) **What sort of correlation exists between median household income and percentage of divorced males?**

Jordan Mayer
01/10/2018
STAT 350, 1:30, with Fan Wu
Lab 071

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.

```
# set working directory and import US Data
setwd("//myhome.itap.purdue.edu/puhome/pu.data/Desktop/mayer15/STAT 350/Lab 01")
USData <- read.table("USData_Spring.txt", header=TRUE, sep="\t")
```

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

```
# clean US Data
USData_clean <- USData[complete.cases(USData),]
```

   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.)

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

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

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

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

1103 – 1098 = **5 rows removed**

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

//myhome.itap.purdue.edu/puhome/pu.data/Desktop/mayer15/STAT 350/Lab 01

Jordan Mayer
01/10/2018
STAT 350, 1:30, with Fan Wu
Lab 071

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 "U0ban" 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.
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."

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.”
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.
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.