Assignment module one

Mayel Espino

2 University of San Diego

# Author Note

# Abstract

This is the assignment exercises and answers for module one of ADS500B.

*Keywords:*

# Assignment exercise one

The Exercise 1 Dataset (located in your assignment prompt in Blackboard) contains an imaginary dataset of auto insurance providers and their ratings as provided by the latest three customers. Now if you had to choose an auto insurance provider based on these ratings, which one would you opt for and why?

# Assignment exercise two

The Exercise 2 Dataset (located in your assignment prompt in Blackboard) for this problem contains statistics in arrests per 100,000 residents for assault and murder, in each of the 50 US states, in 1973. Also given is the percentage of the population living in urban areas. Use the pre-processing techniques at your disposal to prepare the dataset for analysis.

2.1: Resolve the missing values ​without deleting rows​. ​

2.2: Look for outliers and smooth noisy data ​without deleting rows​. ​

2.3: Prepare the dataset to establish a relation between an urban population category and a crime type. Submit your final dataset with imputed and smoothed values. [Hint: Convert the urban population percentage into categories, for example, small (<50%), medium (<60%), large (<70%), and extra-large (70% and above) urban population.] ​

# Assignment exercise three

The Exercise 3 Dataset (located in your assignment prompt in Blackboard) for this problem is a dataset of bridges in Pittsburgh. The original dataset was prepared by Yoram Reich and Steven J. Fenves, Department of Civil Engineering and Engineering Design Research Center, Carnegie Mellon University. Use this dataset to complete the following tasks:

3.1: Look for outliers and smooth noisy data.

3.2: Resolve all the missing values ​without deleting rows​.

3.3: Prepare the dataset to establish a relation among:

Length of the bridge and its purpose.

Number of lanes and its materials.

Span of the bridge and number of lanes.

# Assignment exercise four

A community library has decided to limit all future procurement of books either to hardback or to softback copies. The library also plans to convert all the existing books to one cover type later. Fortunately, to help you decide, the library has gathered a small sample of data (Exercise 4 Dataset located in your assignment prompt in Blackboard) that gives measurements on the volume, area (only the cover of the book), and weight of 15 existing books, some of which are softback (“pb”) and the rest are hardback (“hb”) copies.

The Exercise 4 Dataset represents that the dataset has 15 instances of the following four attributes:

Volume: Book volumes in cubic centimeters Area: Total area of the book in square centimeters

Weight: Book weights in grams

Cover: A factor with levels; Hb for hardback, and Pb for paperback

Now use this dataset to decide which type of book you want to procure in the future. Here is how you are going to do it. Determine:

4.1: The mode of the book covers. ​

4.2: The mean of the book weights by book covers. ​

4.3: The variance in book volumes. ​

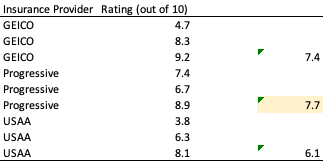
4.4: Use the above values to decide which book cover types the library should opt for in the future.

# Results

# Answer for assignment exercise one

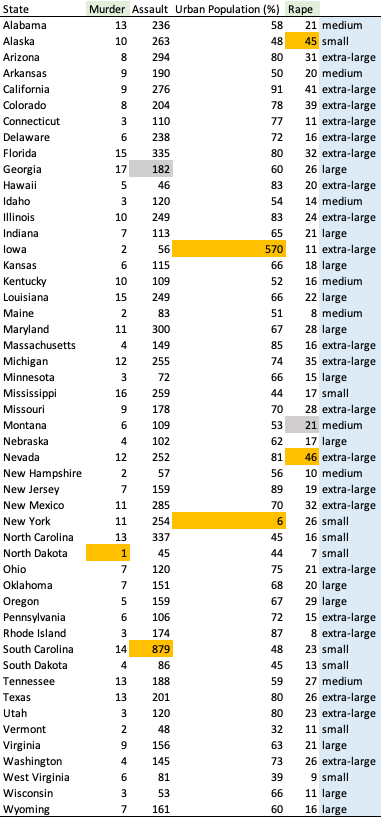
**Table 1**

Average customer rating for Insurance providers



Based on the average customer rating I would select Progressive.

# Answer for assignment exercise two



**Table 2**

statistics in arrests per 100,000 residents for assault and murder, in each of the 50 US states, in 1973

# Table 2 is the result of processing done to the data set provided. The following are the steps taken:

For columns murder and rape, I removed all the decimal points.

For columns assault and rape, to find the missing values without removing them, I calculated the average for each column. The value I replaced is highlighted in grey.

To identify the outliers, highlighted in orange, I compared the extreme values to the average and the frequency of similar values. So for example Murder has an outlier value of 1, while most of the smaller values are 2 and 3.

The following are the outliers per column:

Murder outlier: 1

Assault outlier: 879

Urban population outliers: 570 and 6

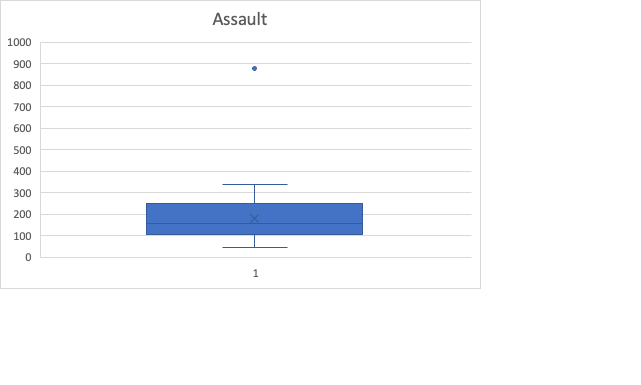
Rape outliers: 45 and 46

To support my findings, below are the boxplots for each of columns:



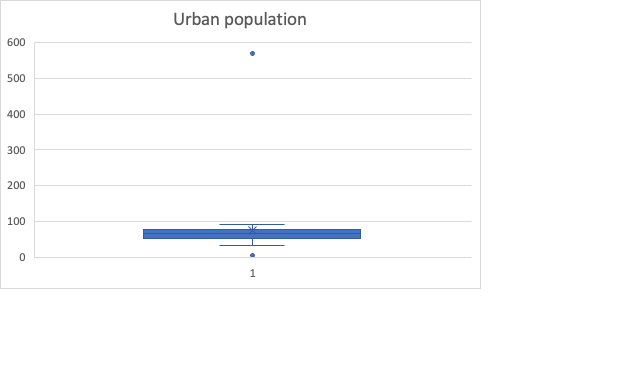
**Figure 1**

Murder boxplot



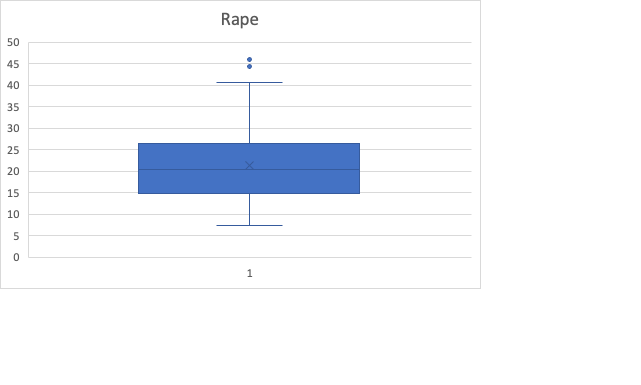
**Figure 2**

Assault boxplot.



**Figure 3**

Urban population boxplot



**Figure 4**

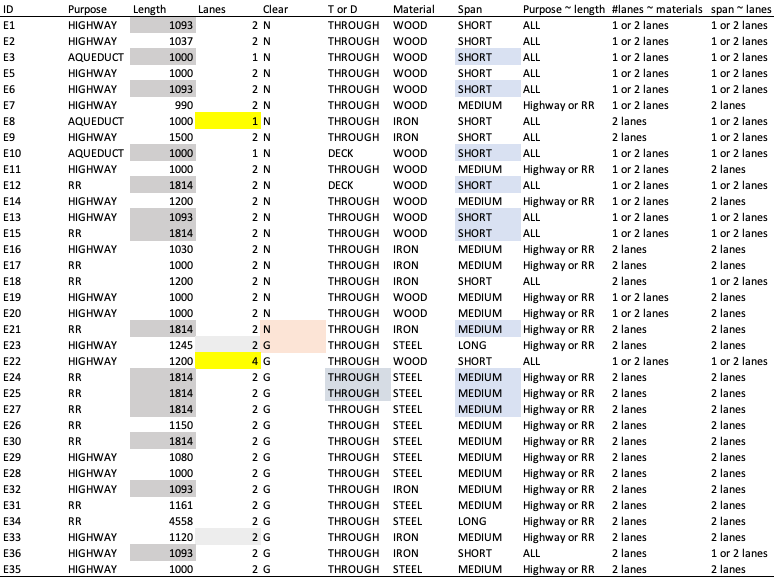
Rape boxplot

Finally the column highlighter in blue is the categorization to establish a relation between an urban population category and a crime type. This was done as per the instructions in the assignment, categorizing the populations in to small, large and extra-large.

# Answer for assignment exercise three

**Table 3**

Bridges in Pittsburgh



Steps taken to complete the assignment:

To replace the empty values in Length, I used the average of all the lengths, highlighted in grey.

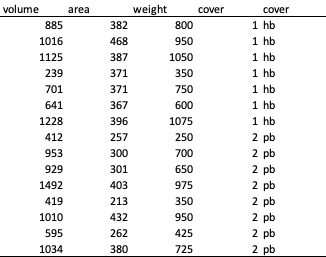
To replace the empty values in Lanes, I used the most common value of 2, highlighted in light grey.

To replace the empty values in Clear, since G and N are evenly distributed, I replaced the blank values following the same distribution, since there were two blanks, I used one G and one N. Highlighted in pink.

To replace the empty values in column “T or D” I used the most common value of TRHOUGH, highlighted in blue-grey.

To replace the empty values in span, I used the most common value based on each category: I replaced all blanks for wooden spans with short and steel and iron with medium which occurs more often than does long.

# Answer for assignment exercise four



**Table 4**

Sample data for library

Answer 1: I assigned 1 to SB and 2 to PB. The mode of book covers is PB, or 2. Based on the exploratory statistics function in Excel.

Answer 2: The mean for HB is 796.429 and for PB is 628.125. Based on the exploratory statistics function in Excel.

Answer 3: The sample variance is 115518.3524, I used the VAR function in Excel.

Answer 4: The answer to this assignment is my recommendation: buy more hard cover books. My answer is based on the fact that the mean for total area of hard cover books is 796.429 and the mean for total area of 628.125. So if there has to be a choice between the two we choose the type of cover which is most common in the population. The fact that hard cover is more common can obey to a couple of factors: popularity or durability. In either case it is a good option to select hard cover.