**DAD 215 Final Project: Data Analysis Report**

**By Alexander Varljen**

**August 14, 2022**

1. **Introduction:**

This report uses data sets from New York Cities transportation department to answer the following research question: “During the months of June and July, males under the age of 25 travel for longer time periods than females in the same age group. These riders also have higher utilization of the subscription passes than the customer passes. Present results as an entire population and by available borough. Use the birth year and the current year to calculate age.”. To answer this question, I created my own data set from merging four Citi Bike Data sets from the year 2015 and the months January, December, July, and June, added variables to this data set, created bar graphs from this data set, and used proc report to make calculations. The sections beyond this introduction will describe these steps and the conclusions reached after the analysis.

1. **Creating the Variables Month and Year:**

Code:

Graphical user interface, text, application, email

Description automatically generated

Log:

Graphical user interface, text, application

Description automatically generated

Results:

Graphical user interface, table

Description automatically generated

**Explanation:**

This is the code, code log, and printed results of taking my Citi Bike Data set for July 2015 and adding the variables month and year to the data set. I repeated the process across all four of my data sets. As far as what challenges I faced, I realized that I needed an id for the merging of the data sets, so I had to create four new data sets from the originals other than the ones that I already made, as they excluded the “bikeid”. I did not run into any significant issues outside of that.

1. **Making One Data Set Out of the Four City Bike Data Sets:**

Code:

Graphical user interface, text, application

Description automatically generated

Text

Description automatically generated with low confidence

Log:

Text

Description automatically generated

Results:

Table

Description automatically generated

**Explanation:**

Here is the code, code log, and printed results of the merging of the four data sets by the id “bikeid”. To do this, I first merged the December and January data sets into a data set called “winter”, and then merged the July and June data sets into a data set called “summer”. After this, I merged the winter and summer data sets into a data set called “cityBikeData”. I ran into some simple errors, which I could quickly identify and fix using the code log.

1. **Breaking Up Trip Duration into Three Categories:**

Code:

Graphical user interface, text, application

Description automatically generated

Log:

Text

Description automatically generated

Results:

Table

Description automatically generated

**Explanation:**

Here is the code, code log, and printed results for some code meant to create three categories of trip duration: one for 300 or less called “short”, one for 300 to 700 called “med”, and one for 700 and over called “long”. I decided to create a new data set while doing this, as I ran into some issues constructing the program. However, the program worked after formatting the code to make a new data set.

1. **Summing Trip Durations:**

Code:

Text

Description automatically generated

Log:

Graphical user interface, text, application

Description automatically generated

Results for June:

Diagram

Description automatically generated

Results for July:

Diagram

Description automatically generated

Results for December:

Diagram

Description automatically generated with medium confidence

Results for January:

A picture containing diagram

Description automatically generated

**Explanation:**

Here is the code, code log, and printed results of a program designed to sum the trip durations by gender for those under 25 who’s trip duration is 700 or over. While the code and code log are for participants trip durations during the month of June, I also wrote code for the other three months and have shown the results. I ran into no major issues but decided to confine my trip duration sums to trip durations 700 and over.

1. **Finding the Mean of Trip Durations:**

Code:

Text

Description automatically generated

Log:

Text

Description automatically generated

Results for June:

Table

Description automatically generated

Results for July:

Table

Description automatically generated

Results for December:

Table

Description automatically generated

Results for January:

Table

Description automatically generated

**Explanation:**

Here is the code, code log, and printed results of the code constructed to produce a report calculating the mean trip duration for each gender during the months of January, December, July, and June. I ran into no major issues other than some small errors, which I could easily fix after looking at the code log and fixing some semantic errors.

1. **Counting Subscription Passes:**

Code:

Text

Description automatically generated

Log:

Graphical user interface, text, application, email

Description automatically generated

Results for Subscription Passes in June:

A picture containing diagram

Description automatically generated

Results for Subscription passes in July:

Diagram

Description automatically generated

**Explanation:**

Here is the code, code log, and printed results of a program designed to sum the subscriber passes of the Citi Bike users by gender. I first found the results for the month of June, and then changes the program slightly to find the results for the month of July. I ran into no major issues and was able to print the results successfully.

1. **Final Notes and Conclusions of Analysis:**

To import the data sets that were used during this analysis report, I first went to “file” in SAS, then press “import data”. Then I selected “Microsoft Excel Workbook” and hit next, which brought up a window that allowed me to browse for the file that I wanted to import. Then, I selected the tables that I want to import, and then the library (I chose \_EXP0\_ for all four of my files), and then a new data set name. Then, I selected next, and the page made me a PROC IMPORT program to permanently keep the data set.

While programming for this report, I decided to use “bikeid” as the I.D. that I would merge the four data sets by. I formatted my mean calculations to allow for five figures before the decimal place and up to 3 after, as I thought that this would allow for any mean derived from this data set to fit. I also formatted the graph pattern to L3, just to make the bar graphs more visible. The functions used to transform the data included if-statements to confine the temporary data sets being analyzed to specific constraints, such as ages bellow 25 and known genders. To analyze the data, I decided to sum trip durations greater than or equal to 700 by gender and display the results via horizontal bar graphs, use proc report to find the mean of the trip durations for each gender and sum the number of subscriber pass users via a horizontal bar chart and a new variable that denotes each subscriber pass user with a 1 that can be summed to count the total.

The results of the analysis are as follows. While there seems to be twice to a third more males using city bikes during June and July, females seem to travel slightly longer on average during these months than their male counterparts. When we look to the month’s farthest away on the calendar for comparison, we see a similar pattern. As far as subscription passes are concerned, more than twice as many males seem to have subscription passes as females during the months of July and June.