Project 1

Basic Univariate Statistics, Graphical Methods, and Communication of Data

Ngan Le Thi Thai (Crystal)

Northeastern University – College of Professional Study

ALY6010 21420 Probability Theory and Intro Statistics – SEC Winter 2019

Instructor: Dee Chiluiza (PhD.)

27th February 2019

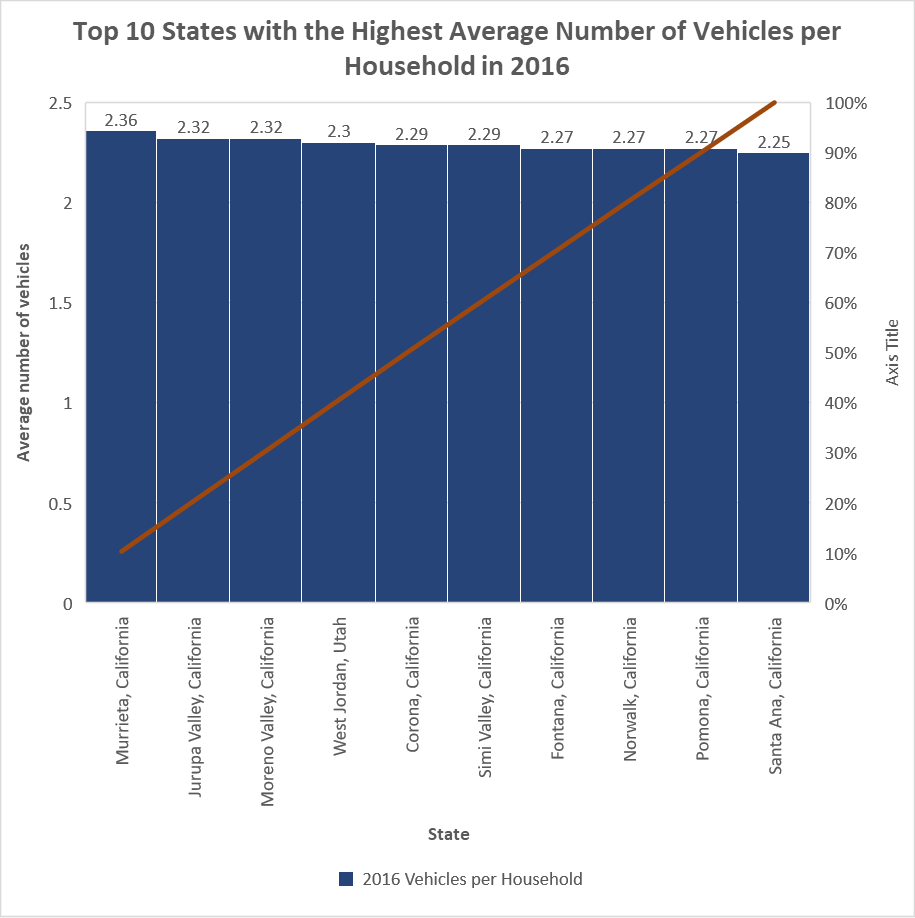
Introduction

People are all familiar with data in daily life by storing images or videos in their phones, downloading information on the internet. Speaking of data statistics, they tend to think about application in large scale with complicated calculation. However, data statistic has been used in different fields such as finances, human resources, biotechnology, and so forth by the most common software such as Microsoft Excel. This essay would analyze the data about vehicle status of 317 jurisdictions in the US between 2015 and 2016 based on the knowledge of Basic Univariate Statistics, Graphical Methods, and Communication of Data (week one module).

This essay will focus on the methods of descriptive and numerical analysis. Therefore, it is limited on the object of study (the percentage of non-vehicle households and number of vehicles per household only). I believe that It is enough to show many insights about means of transport in the US. We can see the trend by observe data within two years and evaluate personal vehicles situation in the country.

About Excel, it is a software which is developed by Microsoft. This is the most friendly and common application with people coming from different majors. Unlike normal usage, adds-on tools developed by Microsoft Excel help user to execute the advance analysis and statistic that they do not needed to be a professional in the major.

Analysis

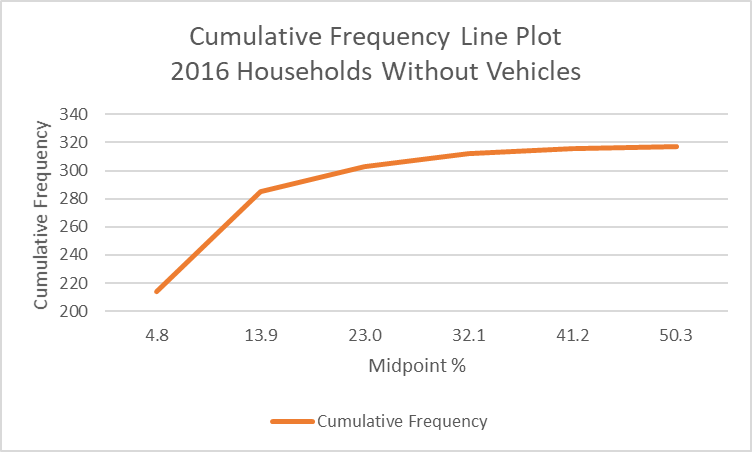
The first Pareto graph provides the top ten states having the highest number of personal vehicles per household in 2016. The shape of columns remains nearly unchanged.

According to the purposes of Pareto chart, we can see the decreasing value from the left to the right. It can be easily seen that value each section does not fluctuate significantly, and ten cities are of California (except for West Jordan, Utah). Therefore, we can conclude that California is the state where each family has the highest number of personal vehicles (about 2 or 3 cars per family) compared to other states.

Regarding the number of households without vehicles, most non-vehicle families in states took from 4.76% to 13.88%. There are a few states which has more than 22.9% number of families without cars.

Based on the shape of above chart, we can see the frequency of data lies on two left columns which have value of 4.76 and 13.88 in average. This distribution might be explained by the outliners which is mention later in this essay.

In term of the Line Plot below, the cumulative frequency of families without vehicles increased sharply interval of 4.8% to 13.9%. This means that people the majorities non-cars families constituted 4.8% to 13.9% in the states.



Below is the numerical descriptive statistic of quantity of vehicles in the national wide.



Each digit itself speak for themselves. For instance, each household possesses 1.72 vehicles in average the US while the maximum is 2.36 cars are hold by a family.

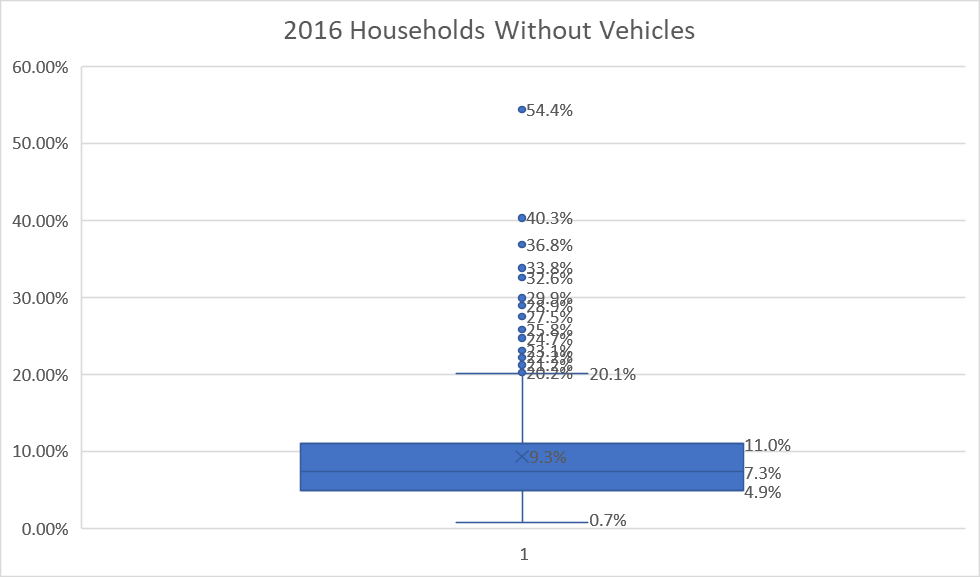
Another factor influences the accuracy of data is outliners which is calculated by following measurements (with data of 2016 Percentage of households Without Vehicles)



Then, outliners are defined by intervals of upper bound and lower bound. Here are outliners are calculated (27 outliners)

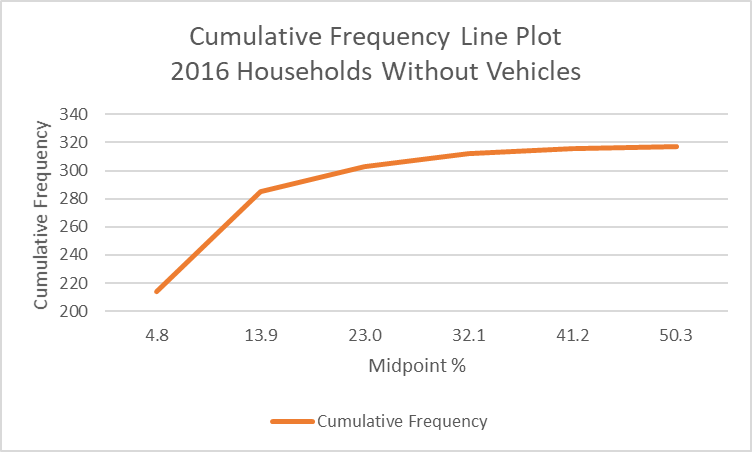
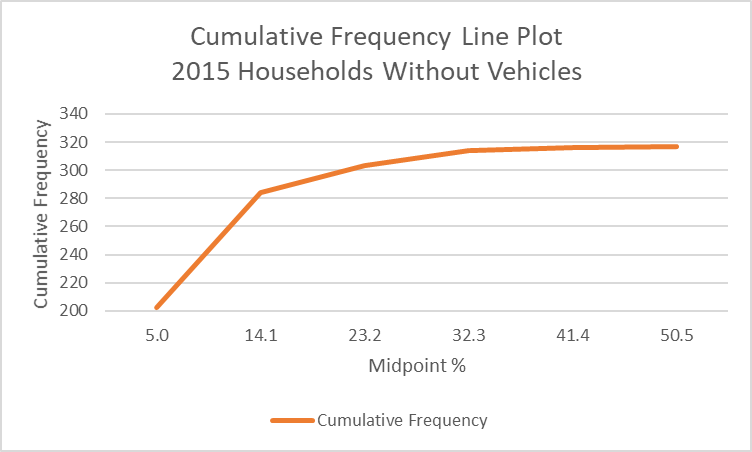


Another way to figure out outliners is Box and Whisker graph which is draw below



Comparison between 2015’s and 2016’s data in the fraction of non-vehicle households.

In 2015, 5% people without vehicles in states made up the most frequency in the country. Then, this number decrease to 4.76% to 2016. This means that American people tend to possess more vehicles in 2016.



Overall, the shape of two charts are identical because they have appropriately similar numerical description (1.72 in 2016 and 1.69 2015). Hence, it is considered as a slight move in the value from 2015 to 2016

Consider number of people having personal vehicles, there is also small change in numbers of given time. Mean, mode, median in two tables are almost equal and have the almost same distant with max and min.



The next table describe the average vehicles per household in states between 2015 and 2016.In general, Montana took the lead in 2015 but was surpassed by Utah in 2016 with two vehicles per family.



Another graph to show the connection between data in two compared time is scatter plot

As mentioned in the numerical of number of people having vehicles, the data in 2015 and 2016 are identical. Therefore, the distribution of data has the same patterns.

Conclusion

There are two way of illustrating the data which are graphical and numerical descriptive statistic. They are all valuable. However, choosing the method of description depends on purposes of analyzing the data. Histograms could show the frequency of data while Box and Whisker graphs are used to detect the outliners.

It is essential to consider the appearance of outliners which is the extreme value. They affect the shape of graphs. Besides, outliners can make numerical measurement error.

Pivot table is also a useful method to summarize and do statistics while scatter plot is best for indicate the correlation between compared time.

References

Absentdata. (n.d.). *How to Find Outliners in Excel.* Retrieved February 25th, 2019 from <https://www.absentdata.com/how-to-find-outliers-in-excel/>