**Emre Yener**

**6/6/2019**

**Summary of Work**

Our dataset from Kaggle consisted of Los Angeles parking citations with the information regarding the citation itself and what kind of a car had the citation. To start with, the full dataset is about 9.4 million rows and it takes a huge amount of time to compute a data frame or a plot every time. Therefore, we have decided to take samples of our own if and when we see fit. The random sample Emre have used to answer all questions is 1/200th of the entire data set picked at random and issued a “random state” to keep the consistency going.

In an ideal situation, one hopes to find the correct information entered into the system however, this is a very susceptible issue to human error. The parking officers might not have been paying attention, not all information from the car might be clear are two examples of human error. In efforts to create my own “ideal” world, Emre analyzed on where those errors might be. When Emre was looking at the entire data, a few things stood out the most. There are a lot of car models and unfortunately there are more than 50 states listed.

To go more into detail, the “ideal” data world that Emre intended to create had to start by cleaning up the data first. Firstly, the states. There are 50 states in the United States and 9 commonwealths /territories. This amounts to a maximum of 59 unique acronyms in the “RP State Plate” which represents the license plate states. The cleanup started with dropping the not available values which no states have been entered. Next up, Emre had to filter the actual state and commonwealth acronyms against the wrong ones. Unfortunately, this was a manual procedure to find the indices and drop them. Considering there were 66 entries, it was not very rigorous to filter and get it down to 51 including District of Columbia. After this, the percentages of occurrences were calculated and a data frame has been formed. Bar chart visualization is the most efficient way to see and compare this type of a data frame with 51 things to represent.

Secondly, the car models. This column in the actual data set consisted of 2314 unique car models after dropping the not available values. Fortunately, Emre’s sample data came up with only 153 unique values. The procedure between models and states in terms of analyzing and creating a new data frame out of percentages of occurrences are the same. The procedure is as follows: group by the model, find a total count of how many times a model has been entered, find the individual counts for all models, divide by total count and multiply by 100 to enter a percentage value representing the model. The main difference of analysis here was the abundance of unique values and the way human error occurs. For states, it was a completely wrong state entered with no indication of what that acronym might mean since it does not resemble any already existing state. However, the models have more resemblance in the entries therefore there can be a matching.

The matching questioned on the above paragraph can be done in many ways. One easy way to do it, should you succeed against the complicated syntax, is called Regular Expressions. It is a library on Python which helps you match a user identified partial string to be looked into a list of strings by the way user intends. Emre’s way of looking was to primarily look for the starting first, first two or first three letters to find resemblance. The similar strings then got compared in terms of the percentages that exist for them in the data frame. If it did not amount to too much, it was ignored. The top nine contributors to this caused received the following additional procedure to further clean up the data frame: indices for similar strings have been located, the percentages for all similar strings have been summed, the redundant ones are dropped and the new percentage value has been implemented to the only representative left of that bunch.

For visualization purposes, a horizontal bar chart for the models and a regular bar chart for the states seemed the best fit. The only reason for a horizontal bar chart is that at the end of all cleaning and merging, there were still 134 unique values to represent in a single graph. When the bars were put at the x axis, the visual aid becomes a hindrance to vision. The reasoning to choose a bar chart above a pie chart is again the abundance of unique values to represent. The reason to choose a bar chart over the scatter plot is that accept for the top 15 percentages of unique states or models, the percentages drop down to a very insignificant level where there is a culmination of scatter point all along a line and it does not help to show the difference between each value.