Capstone 1: Inferential analysis

One of the main variable studied in this dataset is the duration of delay that the accident causes. The dataset was narrowed to the ten cities with the most accidents. Out of this group, various variables were tested against each other, with duration being of prime interest. Pearson Correlation tests were conducted on three combinations of variables: duration vs. street type, duration vs severity, and severity vs street type. For all three of these variations, the tests returned that the variables were probably related. However, it must be noted that the relation was not very strong for duration vs. street type and duration vs severity. On the other hand, severity vs. street type returned a much stronger correlation which implies that of the three, this one is the most likely to be actually correlated.

In addition, bootstrapping was used to determine if the true value of the mean of duration was acquired from the dataset. In order to do this, the mean was calculated for each of the top ten cities directly from the data. Next, the data was cleaned further by removing obvious outliers. For instance, the ECDF of the data showed a datapoint at 6000 minutes. This is a very clear outlier from the rest of the data points and was easily removed. To remove outliers, a z test was used. If the z-score was above 3, the data point was removed from the set. Many points were removed which hundreds or thousands of percentiles removed.

Bootstrapping was then employed in order to replicate 100,000 samples for each city of the same length of the original sample from that city. For instance, Dallas had a sample length of approximately 64,000, so each bootstrap sample for Dallas was also 64,000 rows long. The mean of each replicate was calculated and stored. The gathered data was then graphed along with a 95% confidence interval. The means of each city were stored and compared to the raw means collected from the dataset.

The figures below show these two charts:

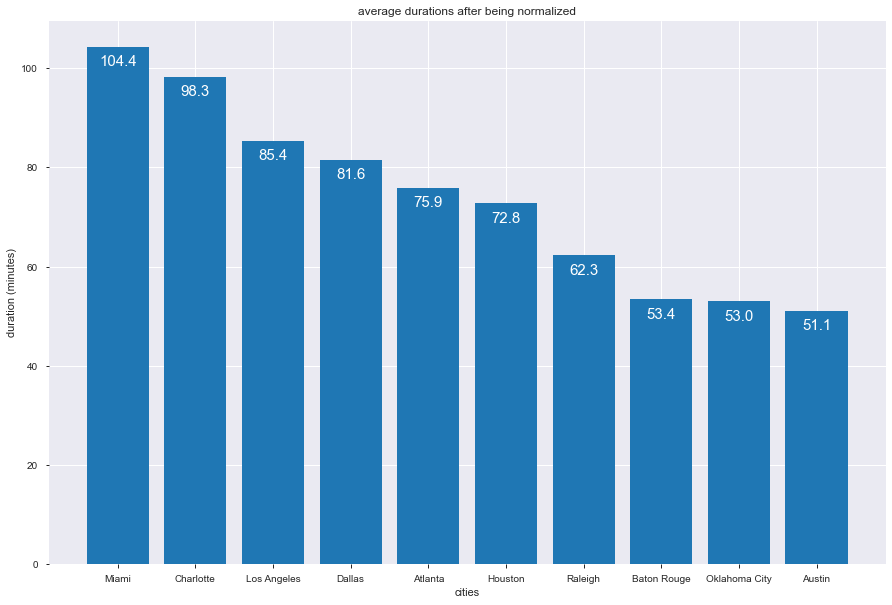


Fig 1. Average duration after being normalized (bootsrapping)

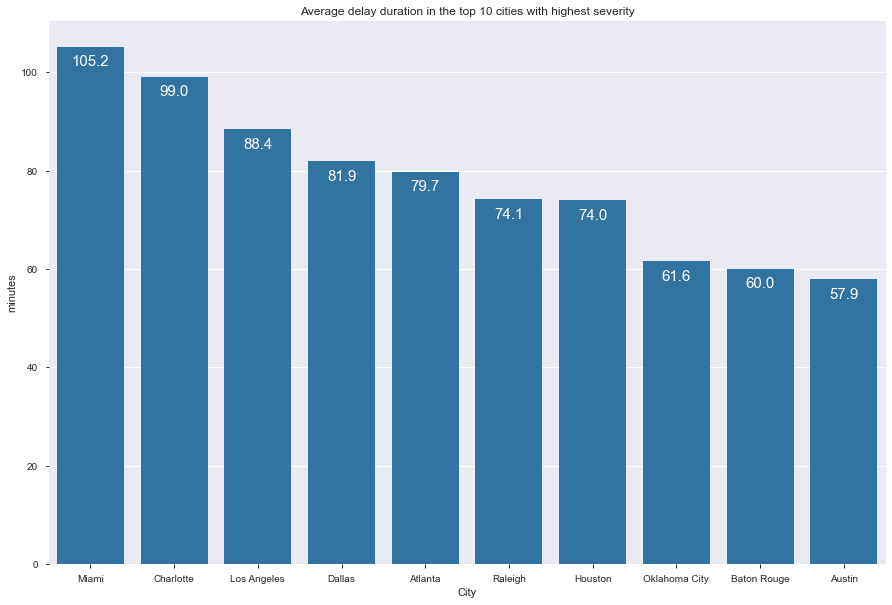


Fig 2. Raw mean durations from dataset (no normalization)

As can be seen from these charts, the normalization did correct the mean by approximately one and a half minutes. However, some cities showed considerable change. For instance, Baton Rouge and Oklahoma City both showed corrections in the 7 minute range, a substantial change.

Conclusion

It is quite clearly seen that performing a z-test was beneficial to the dataset. On average, only 30 to 50 records were removed for each city implying that the majority of data was in an acceptable range. Most of the records removed were extreme outliers in the range of 6000 minutes all the way to 20000 minutes. In addition, the Pearson Correlation Coefficient implied that there was, indeed, a relationship between duration, the street type, and the severity of the accident.