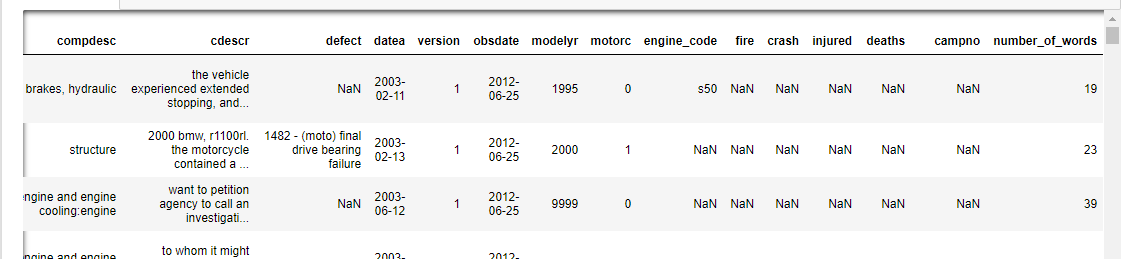
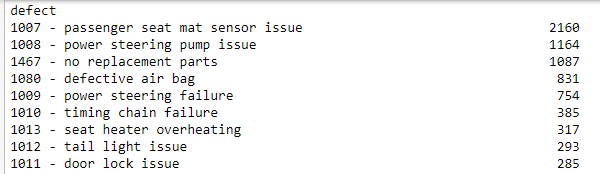
Capstone Project 1 – Inferential Statistics

A number of exploratory statistics were done in order to understand if trends in the data could be predicted. However, because of the nature of this problem (clustering text data), the process of determining correlation between variables was not as necessary. In this project, the main field of interest is the narrative and other parameters of the data are not used, except to validate the data. For example, when clustering is carried out and cluster IDs or topic models are assigned to the data, it is possible to inspect the work of the model to determine if it created a grouping similar to the labels that have been manually assigned to the data. Because this process has been done already (labeling) and is a luxury that is usually not present in clustering problems, it is prudent to use it as a check.

One question that was investigated was if length of comments seems to have any bearing on the type of complaint. To investigate this question, the document corpus was tokenized and a new column was written into the dataframe specifying the length of that voq (column entitled: number\_of\_words).



Once this column was included, the highest frequency complaint categories were investigated. Perhaps the larger categories (more complaints logged) have a typical length of complaint that is different from the average length of a complaint in the entire dataset (93.83 words). The following output provided an idea of which categories to investigate first.



The average number of words in a passenger seat mat sensor complaint was 88.31 while the average number of words in a power steering pump complaint was 88.18. Finally, the average number of words in a no replacement part complaint was 76.61.

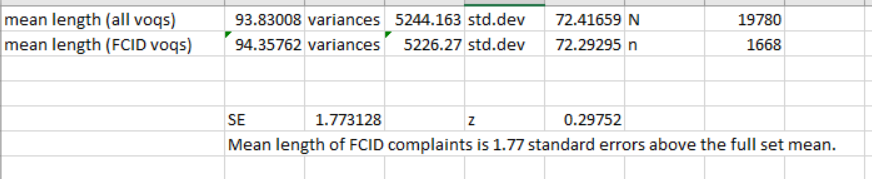
Deliberating further, perhaps complaints that are about a more dangerous issue are explained in more detail and thus have more words. The fire, crash, injury, and death columns in the data were used to as surrogates for severity. These four columns were aggregated into one column to determine which defect category had the highest counts of either fires, crashes, injuries or deaths.

The complaint categories with the largest safety-relevant issues (FCID) are listed along the x-axis above. It is possible that customers who undergo a stressful situation use their complaints as a means to vent. This may increase the length of their complaints and if true, length of complaints may be an indicator of severity.

To verify if this claim, hypothesis testing was employed:

H0: The difference between the mean length of comments between the full set and the complaints flagged in the fire, injury, crash and death column is NOT significant.

H1: The difference between the mean length of comments between the full set and the complaints flagged in the fire, injury, crash and death column is significant.



To test this hypothesis, a Z-test was done because our n was larger than 30. There were 1668 complaints that mentioned a fire, crash, injury or death in them out of the total 19,780 complaints. The standard error and p-value was calculated assuming a 95% confidence interval to be 0.7661066 using the stats package in scipy. Because of the large value of p, we cannot reject the null hypothesis and therefore can’t say that there is a statistically significant difference between the word count of complaint when there is a safety issue mentioned versus when there is not.

The last exploratory statistic done was the most beneficial for the purposes of clustering. Below are the top 10 most frequent defect categories:

1007 - passenger seat mat sensor issue 2160

1008 - power steering pump issue 1164

1467 - no replacement parts 1087

1080 - defective air bag 831

1009 - power steering failure 754

1010 - timing chain failure 385

1013 - seat heater overheating 317

1012 - tail light issue 293

1011 - door lock issue 285

1014 - engine stalled while driving 275

Individual dataframes were created for each of these complaint categories and complaints were tokenized in order to understand if the most common words in the datasets actually relate to the defect names. For example, for clustering to be effective, it would be nice if when looking at all of the complaints about “passenger seat mat sensor issue”, words like “passenger”, “seat” and “mat” show up in highest frequency. If not, further investigation would need to be done to understand what words contribute to a specific complaint’s grouping in a cluster.

Below is the output of the top words in the passenger seat mat sensor complaints. We see that “passenger”, and “seat” are both included. Another step could be to remove stop words and then try this again, but the initial results are promising. If similar results are found for the other top 10 categories, then it will be decided that clustering/ topic modeling is useful.

Passenger seat mat = [('the', 16750),

('was', 4750),

('and', 4278),

('to', 4238),

('a', 4087),

('that', 3074),

('not', 3016),

('passenger', 3008),

('is', 2844),

('in', 2772),

('i', 2676),

('on', 2507),

('bmw', 2489),

('of', 2374),

('seat', 2266),

('vehicle', 2187),

('this', 2173),

('air', 2120),

('recall', 1786),

('my', 1766)]

All categories except “seat heater overheating” and “engine stalled while driving” had terms that appear in the top 20 words that described the issue. For “seat heater overheating” and “engine stalled while driving”, descriptive words were found in the top 30 words, shown below. This is encouraging for our analysis because if the top 30 words for each category are studied, after stop word removal, lemmatization and other pre-processing techniques, we should be able to categorize VOQs that fall into the ten largest complaint categories at the very least. This would still save us a tremendous amount of time over manual coding, so this initial statistical exploration of the data has led to more belief in the utility of the proposed analysis.

Engine stalled while driving = [('the', 2900),

('to', 940),

('was', 874),

('and', 823),

('a', 603),

('vehicle', 541),

('i', 477),

('in', 334),

('of', 332),

('on', 325),

('that', 325),

('not', 325),

('car', 317),

('it', 316),

('engine', 310),

('contact', 298),

('failure', 291),

('bmw', 277),

('this', 214),

('for', 205),

('stalled', 203),

('while', 199),

('my', 199),

('driving', 196),

('is', 186),

('at', 178),

('dealer', 177),

('with', 152),

('be', 140),

('have', 137)]