**Volvo Truck Analytics**

*This report is for the progress period ending on* ***11/05/2019*** (originally 10/31/2019)*.*

*Report is based on the guidelines under the "Report" section provided in the rubric.*

**Ioannis Batsios**

*Vehicle Temp Analysis (Around 12 hours)*

The goal of this task was to determine if the external temperature affected the performance of the vehicle. The results showed that the external temperature was in line with the temperature of the internal temperature readings of the vehicle. After seeing this result, I decided to do some research and discovered that the external temperature does affect performance and the feature (or part) that is affected the most is the turbocharger, because it is compressing the hot intake air. If the ambient air is increased, the turbocharger has to do more work in order to prove the engine with the same amount of oxygen. If the temperature continues to increase, the turbocharger will eventually hit a maximum speed and will no longer provide enough oxygen to the engine resulting in a loss of power. Unsure what more is needed to be done with this information.

**Bill Downs**

Identifying similarities in distribution and scatter plots in APU metrics. Also, linear regression, covariance and correlation is analyzed. Obvious physical attributes of alternator and APU are noted. In general, when the APU charges the batteries, the amps will drop on the alternator sensor while increasing the voltage in the APU battery bank. This was my main idea because this is where Volvo plans on removing connection from the alternator to the batter and change with a photovoltaic system. Result should give an idea of how much batteries should take.

(about 48+ hours together)

**Wahab Ehsan**

*Display Data by day (about 3 hours)*

Function implementation of display data by day was created to sort all data by average for each day for a certain attribute. Goal was to create a cleaner way of looking at data to find something interesting. We were able to find out that data for Truck one data was the week of data or seven days which was expected, but for truck two there was only 3 days’ worth of data which was less than we were told. This means we'll have to decide which days to compare when comparing the two trucks and which days will be not useful.

*Outlier Detection data (about 12 hours)*

This function was made to detect all the outliers in the data for a certain attribute and have a boxplot made for visuals. Even though the data for truck one was 7 days’ worth, the averages were very far apart making me think if there are a lot of outliers and after running it through the function I was able to find out that the last 3 days of data for truck one was outliers. Meaning there were three days’ worth of no useful data for Speed (km/hr).

*Determining Estimator (about 3 hours)*

Was able to determine to use the Kernel Density Estimation (KDE) because of it being non-parametric. For the bandwidth I decided to use a larger bandwidth since our data was scattered around. Use of this function will be beneficial in the future. We used this because our data had large number of outliers and was not a normal distribution. Maybe after removing the outliers we will switch to MLE. And we didn’t choose MoM because that is also for normal distributions and not as effective in our case because of our several outliers in the data.

**James Polk**

*Aggregation of CPU Usage of Truck 1 (14+ hours)*

The goal of this task is to provide aggregation and time-series representation of the CPU load of Truck 1. I contained my observational data to the first three days of logging in order to conform to the data of Truck 2. The time-series representation of the three days shows periods of sensor dropout for the CPU. There is no particular pattern that can be gleaned from the surrounding data. The analysis shows that the CPU load of each of the three days follows a similar bimodal distribution with peaks around 45% and 56%. Combining the three days’ worth of data provided a mean of 47.948982. Performing a simple sampling of 1000 random rows resulted in a similar bimodal distribution with a mean of 47.579, a difference of 0.189982. Performing a sampling distribution by taking a sample of 1000 random rows and making 200-point estimates of the mean resulted in a normal distribution with a mean of 47.931810000000006; a difference of 0.017172. Calculating a 95% confidence interval for the aggregate data resulted in a z-critical value of 1.959963984540054, a sample mean of 48.356 and a confidence interval of (47.9081645837598, 48.803835416240204). Plotting of 25 confidence intervals showed 24 intersecting with the true population mean.

*Calculation of Distance Traveled by Truck 1 (6.5 hours)*

Since the data does not include latitude and longitudinal information, I sliced the DataFrame by day, for the first three days, and resampled the data to 1 second intervals by mean. I then used the formula of sum(Vehicle Speed (Wheel-Based; km/hr))/3600 to determine the total distance traveled. The results were 872.5 km, 371.3 km, and 781.8 km for days 1, 2, and 3 respectively. This allowed us to determine that Truck 1 was a long-haul truck.

**Chris T.**

*GPS Speed vs. Wheel-Based Speed Analysis & Determination of Long/Short Haul Trucks (16+ hours)*

The goal of this task was to compare GPS Speed and Wheel-Based Vehicle Speed for Truck 1 and Truck 2 internally through visual representation and percent change, and then compare the results for each truck to each other. The results revealed a major issue with the speed data: they are bimodal. The GPS Speed readings for Truck 2 are significantly lower than its Wheel-Based counterpart, but this isn't true for Truck 1's data: they are fairly consistent with each other in terms of hard-value readings. However, the trends are the same between the two columns in Truck 2, which indicates that the GPS Speed component was misconfigured. Finally, this initial and basic analysis of these two columns has indicated that the long-haul truck is Truck 1 and the short-haul truck is Truck 2.

*GitHub Repository Management (estimated 7+ hours)*

For most of Project Progress 2, I have been given the responsibility of managing the GitHub repository in a multitude of ways. This includes updating the home page README, creating, updating, and closing Issues, creating and assigning Labels to those Issues, managing and assigning group members to those Issues, and trying to keep the general repository structure simple and clean.