The first question came to my mind after opening the file was why this data was acquired. Also, what is this data and how it was acquired. From my experience, it seems that the data was acquired for these purposes:

1. To analysis racecar function
2. To analysis driver performance
3. To use for racecar development and improvement

Although I am a fan of racecars and Formula 1, I have no prior experience with racecars and its data, so before moving forward and make a problem statement, I did a bit of search through internet about racecars and its data acquisition and analysis (even I found a good movie named “Rush”). In racing, one of the most important questions is that how fast are you and speed data is one the best channels to answer this question. However, it is not rescannable to use only one channel. There are other helpful channels such as braking, acceleration, driver behavior and performance and so on. In fact, there are three main categories for data analysis of racecars, powertrain analysis (RPM, oil pressure, water temperature), chassis analysis (lateral and longitudinal G-forces, wheel speed) and driver (steering angle, throttle, brake pressure).

For an example, to analysis speed, the comparison between different laps can be useful which helps to see how a setup or component change affects the performance on a particular sector of the lap. In addition, driver performance on different laps, or performance between different drivers can be analyzed. Figure 1 shows the variation of speed in different laps for data set 1. There is a large variation and low mean in lap 22 which is because of outgoing time. However, the speeds in laps 23, 24 and 25 are similar to each other. There is big decrease of speed in the last lap. Figure 2 shows the map and laps of the first data set. It seems from this figure that the GPS data is noisy or maybe the GPS was not calibrated and if we can use google map information, we can clean the data. A few more plot are shown in the appendix. All in all, speed is what really matters when it comes to racing, and throttle and RPM are directly connected to the development of speed and using of steering angle together with throttle angle can give useful information.

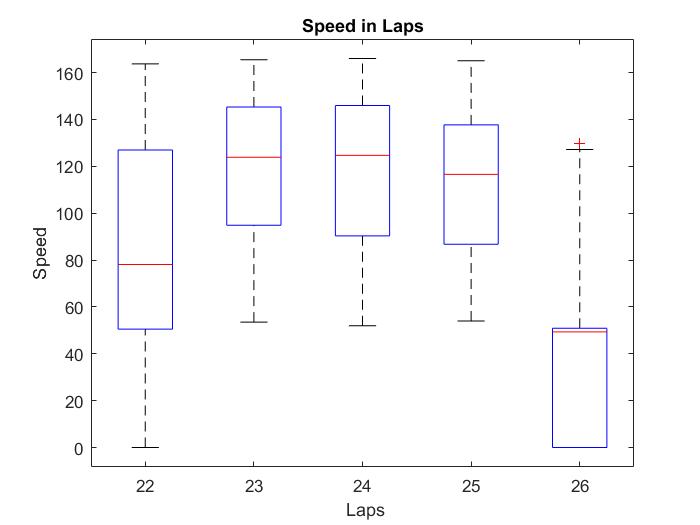


Figure : Speed variation in different laps



Figure : Map of the lap

What I have done so far:

1. Studying of racecar variables (channels) and searching for information about them and how they are related to each other. I found two books but I have no access to them, “A practical guide to race car data analysis” and “analysis techniques for racecar data acquisition”. It will be helpful if you have them and share them with me.
2. Import data set 1 into MATLAB, clean and normalize it. For the laps, I assumed that the lap’s number does not change until the next lap number is coming. So, missing values (NaN) are replaced by the previous value of the lap number. For speed and steering, linear interpolation is used to estimate the missing values. In RTC, real time clock, there are 600 missing values between each two recorded values. Each reported value in RTC, is in HHMMSS format and it is increasing by one minute which means those 600 missing values are related to 60 seconds (1 value for each 0.1 second).
3. Basic analysis of data in Matlab such as normailization, mean, plot and so on. However, processing in MATLAB is really slow and it is not the perfect software for big data. So, I decided to use big data tools such as Cloudera.
4. Using Cloudera. FTP the data to CHD (Cloudera distributed Hadoop) local file system using filezilla. I did it using VM VirtualBox in windows. Importing data from local file system to HDFS. Using spark-sell in CDH. Creating RDD (resilient distributed dataset) based on input data.

What else I can do:

1. To further analyze the speed, I will use Multi covariance plot and perform a t-test significant analysis and multivariate regression to show how the speed varies with different factors
2. If you have any suggestions or problem statements.

There are a few questions so far before moving forward:

1. There are two excel files and there are 17 channels in data set 1, but 20 channels in data description file and data set 2. What are the differences and similarities between these two data sets? For example, these are for two different drivers or two different days? Usually, in machine learning, we have two data sets, one (big data set) for training and one (small data set) for validation (to prevent overfitting and evaluate developed algorithms). I am not sure if it is possible to use data set 1 as validation (which is small) and data set 2 as training. Can you give me a hint here please.
2. There are a lot of missing data. There are various methods to handle missing data but those method are dependent on the problem and data. Why data is missing and what is the distribution of missing data are important to handle them. For example, for Lap channel, it seems that it is possible find missing values based on the previous values. If the current Lap is 22, this value will be valid for next missing values until it changes. Is there a measured variable that indirectly can predict the probability or value of missing data? There are imputation methods such as mean substitution or regression, or model based methods like maximum likelihood but I am not sure if it make sense to use them on these channels.
3. I could not find much information about the relation between different channels. For example, I can say there is relation between speed and Hitoe ecg, but it is not clear for most of them. More information in this regard, can be helpful for more processing.

Appendix

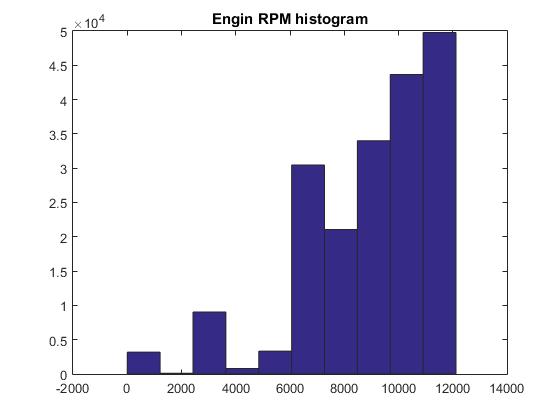


Figure 3: Histogram of Engin RPM

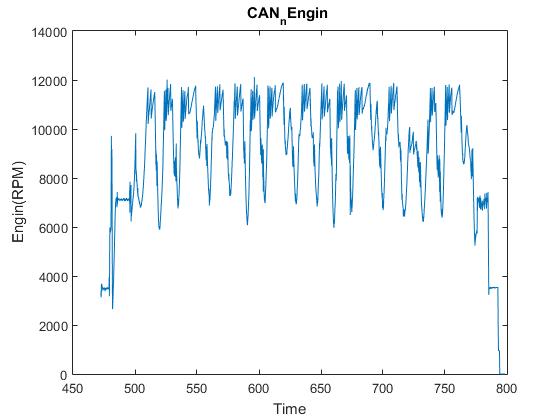


Figure 4: Engin RPM

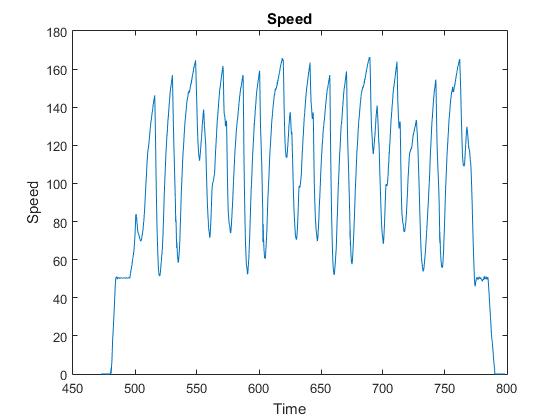


Figure 5: Speed channel

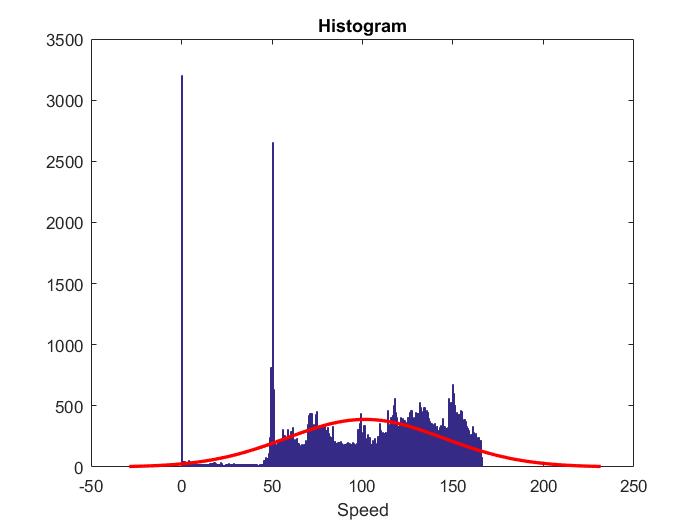


Figure 6: Histogram of Speed. A Gaussian (in red) is mapped on the speed to show the distribution of speed

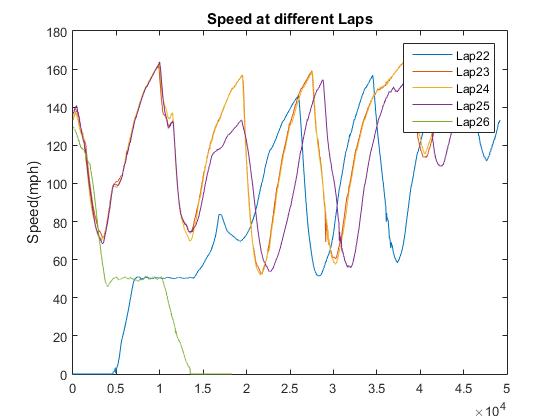


Figure 7: Speed in different Laps. There are 5 laps in data set 1. In Lap 22, speed starts from 0 which is because of outing time. The last lap, lap 26, ends in 0 which means the car stops in this lap.

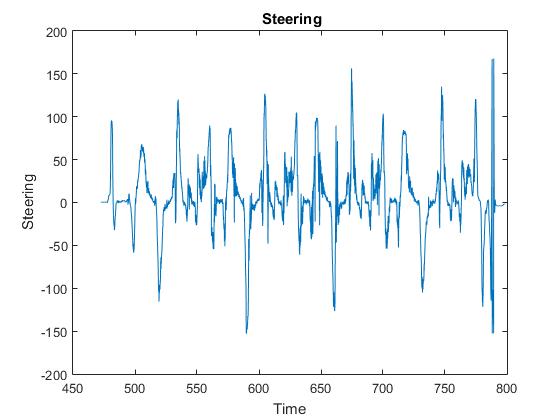


Figure 8: Steering angle. It seems to be noisy and there are outliers which should be removed.