**Transporter – GPS Data**

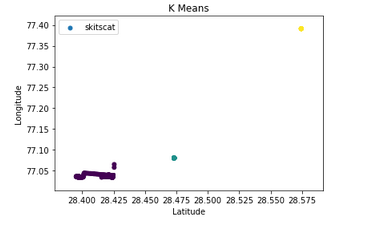
To clean the data (remove noise) from it, I first used various visualizations primarily scatter plots using matplotlib and tableau.

Just by looking at the visualizations it was clear that data had noise.

I modeled velocity of the driver.

To do this,

0. I calculated the distance covered by user(truck) in every timestamp and created a column for it. I have also calculated the time difference between each record.

1. I first clustered the latitude and longitudes to remove noise that is coming from UP and another place. (Observed it when visualized data on tableau).

Plot is made using matplotlib and clearly shows the noise in two clusters

2. Using Haversine formula, latitudes and longitudes of locations I computed the distance between them. I did that for all records.

3. If there is a huge time difference between record generation(time stamp) then it is assumed that GPS is not available.

4. If the distance is huge then it is assumed that it is because of tower jump.

5. Stationery points were also removed, these are the points where the vehicle didn’t move and distance covered between two time-stamps is 0.

After removing such data points, I have clustered the driving velocity in 3 different areas as 3 different clusters as shown in the code. The source code with visualizations are available in the file.

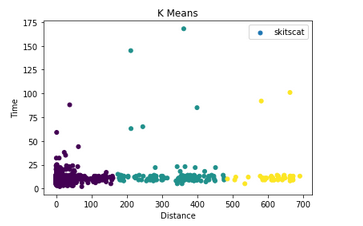
Based on the above clustered points using a cluster a plot has been made with tableau which shows a path.

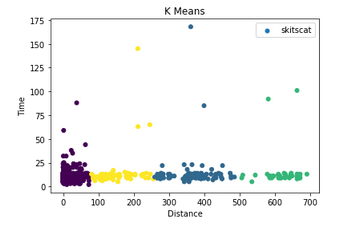
I have used K-Means for clustering, other machine learning models can be applied and compared for better understanding of the data and there by removing the noise in it.

The clean dataset that I have attached contains data which is filtered based on machine-learning technique (clustering). I have used time-difference and distance covered between timestamps to build the model.

As mentioned above, I have used K-Means clustering with k = 3 and the clusters were formed based on the relative speed of the vehicle at different points.

Below is the plot of it:



When K is 4 then the cluster formed is shown below:

The Plots clearly show that there is noise in the data. There are outliers as well in the data which imply the huge time-stamp difference between records.

Based on the clusters formed from k=3 means cluster a map plot has been made using tableau is show below which shows a path formed by removing noise .

I have attached the cleansed data(which can still be cleaned) along with other attributes that I computed for better understanding of it and to remove noise.

A lot of other things can be used as well to remove noise. Filters such as:

i) Speed of a record greater than 25m/s can be removed to reduce noise (as mentioned in the problem statement)

ii) Average speed at which vehicle moved between two timestamps can be calculated from the computed fields of distance covered by the vehicle between the time-stamps there by removing the records which contains noise.

iii) Methods like distance-based filtering can also be used to remove noise, where in we take 3 points p1, p2 and p3 and compute distance between dist(p1-p2), dist(p2-p3), dist(p1-p3) and consider point p2 is noise if summation of dist(p1-p2) + dist(p2-p3) > K\*dist(p1-p3) where k>=2.

iv) Bearing in the GPS data can also be used remove noise. Whenever it abruptly changes it direction we can consider the data point as noise.

Depending upon business requirements many other filters can be used to remove noise from the data.