## Problem statement

*There has been feedback about **delays** in Grab transport services in both Jakarta and Singapore (late driver arrival → late arrival at destination). This has led to a decrease in customer satisfaction and trust, which might increase customer turnover. Upon further research, it was found that this could be due to inaccurate ETA prediction, bad route planning, or other factors. Using the above dataset of GPS pings, what insights can your team obtain to help Grab alleviate these issues?*

Before we start tackling on the problem, we would first like to breakdown the problem statement into smaller segments. As users of Grab ourselves, we understand the frustration users face when there are delays, as it severely impacts user experience.

To further narrow it down, we considered splitting each trip into two. Pre-trip, consists of the waiting time from when a user is matched with a driver to when a user has been successfully picked up by the driver. Post-trip, consists of the travelling time from the pick-up point to the destination. However, we realised that most of the factors which may have caused delays overlap between the two. After all, it makes sense as delays are mainly caused by factors while the driver is on the road.

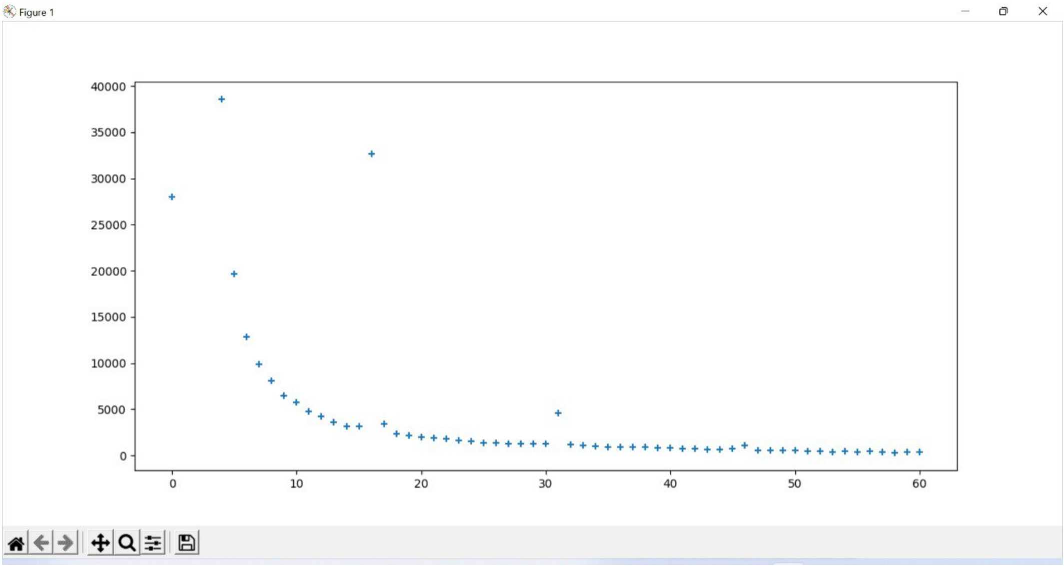
Prior to performing analysis on the data, we must first clean it as it may contain undesirable noise which could skew our results.

At first, our first layer of filter would be to remove any trajectories that has non-continuous pings, i.e. does not have complete data every second. While we might be able to predict the data if only a small portion is missing, ultimately, we would not know for sure what happened within that time frame. Hence, to ensure accuracy of our analysis, we will remove them.

However, after trying to do so, we realised that we only have 30 trajectories out of the 28000 trajectories in the Singapore dataset which fulfills this requirement.

This would remove too many datapoints and leave us with too few data to work on. Thus, we decided to use a different approach.

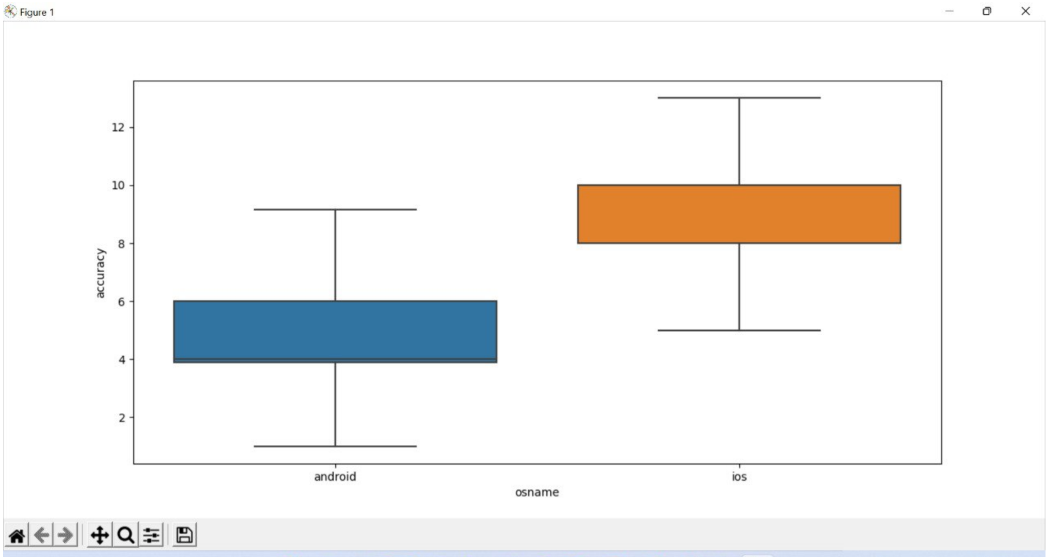
We realised that by having our filter criteria to be having continuous ping, it is too stringent. Thus, we needed to have a criteria that is more relaxed. However, we do not know how many seconds of time difference should we allow, i.e. how long should we allow the trajectory to have a gap in their data set. Hence, we sorted the dataset according to their trajectory id and time. From there, we added in a column that shows the time difference between each sample, in seconds, and we plotted a graph.



We have the number of seconds in time difference on the x-axis, as well as the number of samples that had that particular time difference relative to the previous time sample. We have omitted out the datapoint for 1 seconds and 2 seconds as it is too high which will cause the rest of the graph to be too small. As we can see, above the 30 seconds mark, the graph starts to plateau.

Hence, in the end, we decided to set our cut-off at 30 seconds. This is also because we feel that at this mark, we are retaining sufficient dataset for our analysis. Also, any higher than that, the vehicle would have travelled long enough a distance such that the estimation of its whereabouts as well as the cruising speed would be inaccurate.

The second layer of filter will be removing those data points with an accuracy level of higher than 10m.



We chose to set our threshold at 10m because it is the 75th percentile of the accuracy levels of ios devices. In addition, we believe that accuracy levels that are higher than that will have too much deviation from the actual path taken by the vehicle.

While we are indeed cutting out a lot of datapoints with this filter, we believe that it is better to work with quality data of a smaller quantity, rather than a larger but noisy dataset.

After the two layers of filtering, we will now breakdown the possible causes which led to delays while the driver is on the road, that could be explained by the dataset of GPS pings.

Inaccurate ETA Prediction

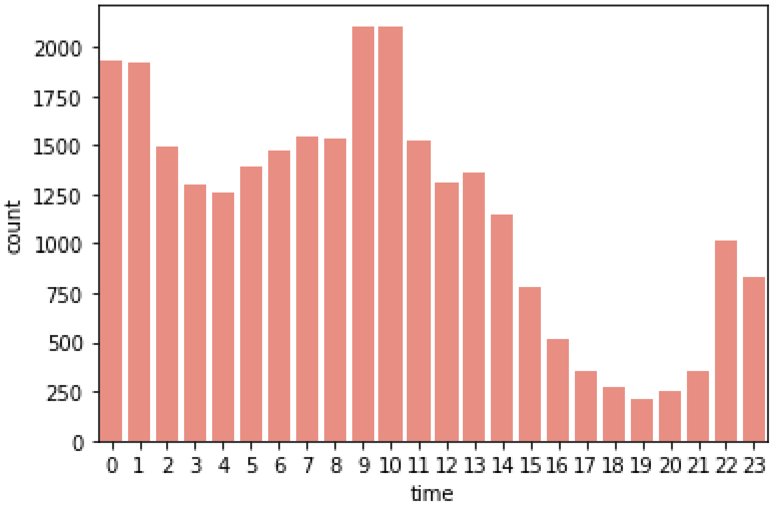
Firstly, the delays could be due to inaccurate ETA prediction. Delays are a result of the actual time taken for the trip being longer than the ETA given to the user. We assume that most of the trips taken are of the optimal route and free of traffic accidents. After all, traffic accidents are unpredictable and delays caused by them are unavoidable. With that in mind, we have broken down the factors which could have resulted in the inaccurate ETA prediction below:

1) ETA given to user is based on average speed of drivers as a collective, as opposed to the average speed of the individual driver giving them a ride, thus it does not provide a personalised ETA based on the driving patterns of each driver.

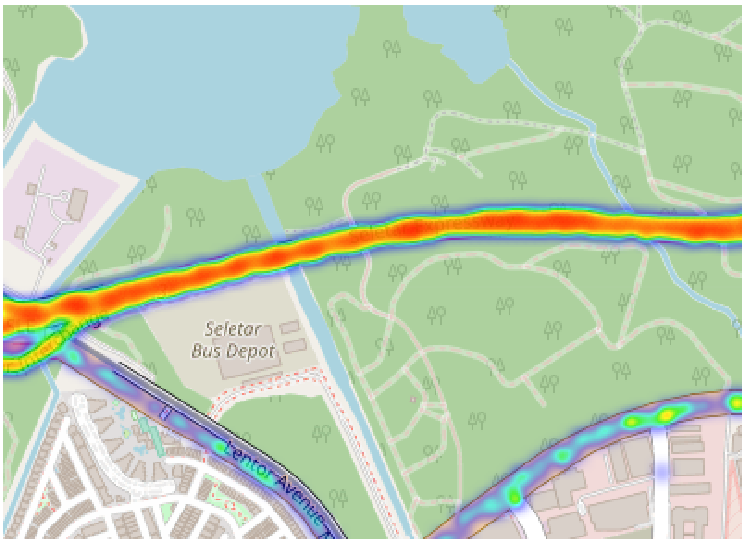
2) ETA should be adjusted based on the time of the day, as traffic conditions vary significantly throughout different times of the day.

For the first point, this is based on the assumption that when Grab calculates the ETA, it does not take into account the driver's driving patterns. After all, it is hard to take into account each driver's driving patterns at an individual level. For instance, some tend to go way above the speed limit on expressways, pulling the average speed up and the average time needed down. This means that for law-abiding drivers, the average ETA would not be a good prediction of their arrival time, as the arrival time would be artificially lowered by irresponsible drivers. This is just one example of variation in driving patterns. Some others could include, one's tendency to overtake other vehicles, one's tendency to attempt to beat the red light, as well as one's familiarity of the location.

To test this hypothesis out, we shall use the average speed of the trajectories along a busy highway to find out if there is a significant spread between drivers. If there is, it makes sense for us to adjust the ETA based on each individual's average speed as opposed to the collective's average speed. This is under the assumption that drivers do indeed have a certain driving pattern and will stick to it on most days and in most situations. This will account for the anonymity of the dataset whereby we are unable to tell if the same driver drove at different speeds on different days.



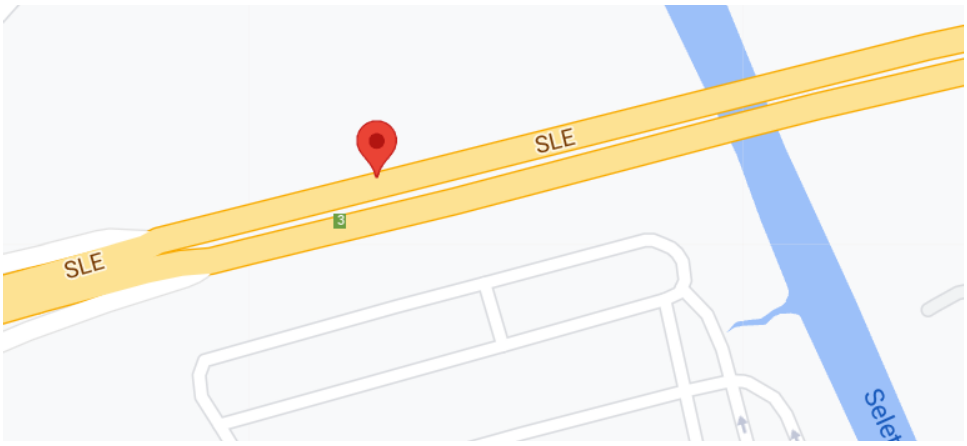
The graph above shows the number of trajectories per hour. We choose the 9am to 10am timeslot as it is the timeslot where there is the highest number of trajectories.



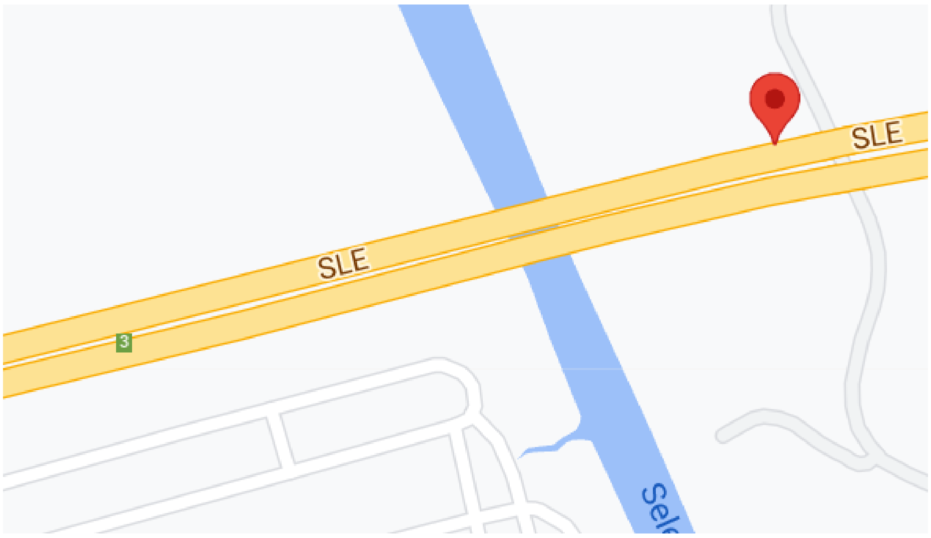
The heatmap above shows that the paths of Grab drivers from 9 am to 10am along SLE.

We choose this path as it is being used by a lot of drivers and it is also relatively straight, hence would faciliate our analysis. We tried creating a heatmap using all 30 million data points in the Singapore dataset. However, due to a lack of quality computing resources, the file would crash due to a lack of ram, even with the use of Google Colab. Thus, we decided to use just 3 million data points from one of the parquet files within the Singapore dataset.

We would need to choose an entry point and an exit point on the road so that we can use them as our reference.

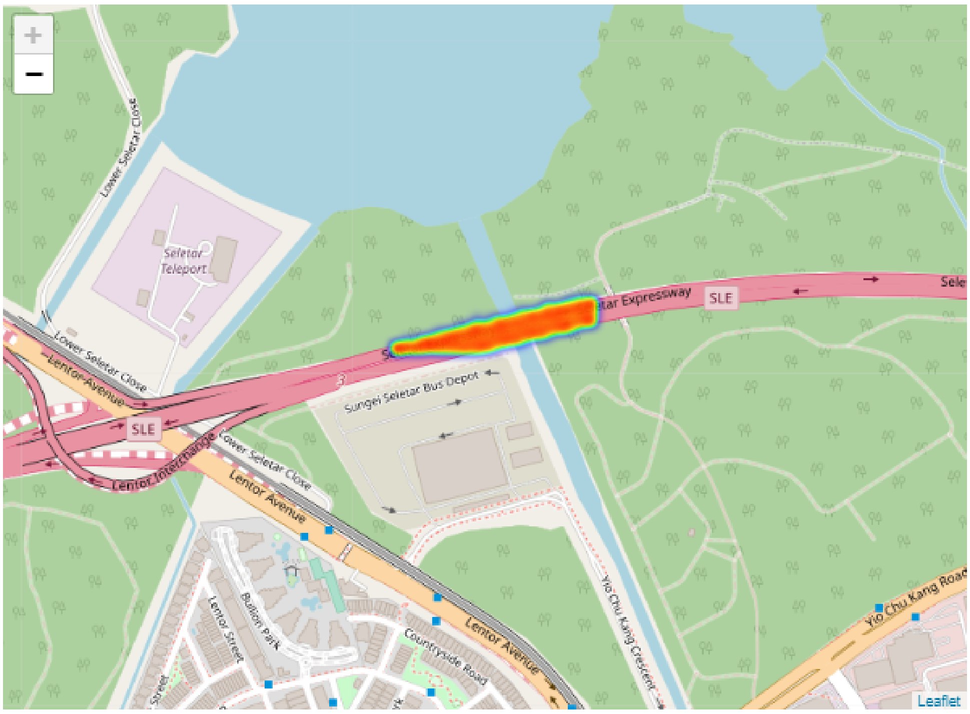


To facilitate our analysis, we have chosen our entry point to be 1.395492, 103.837983 as shown above.



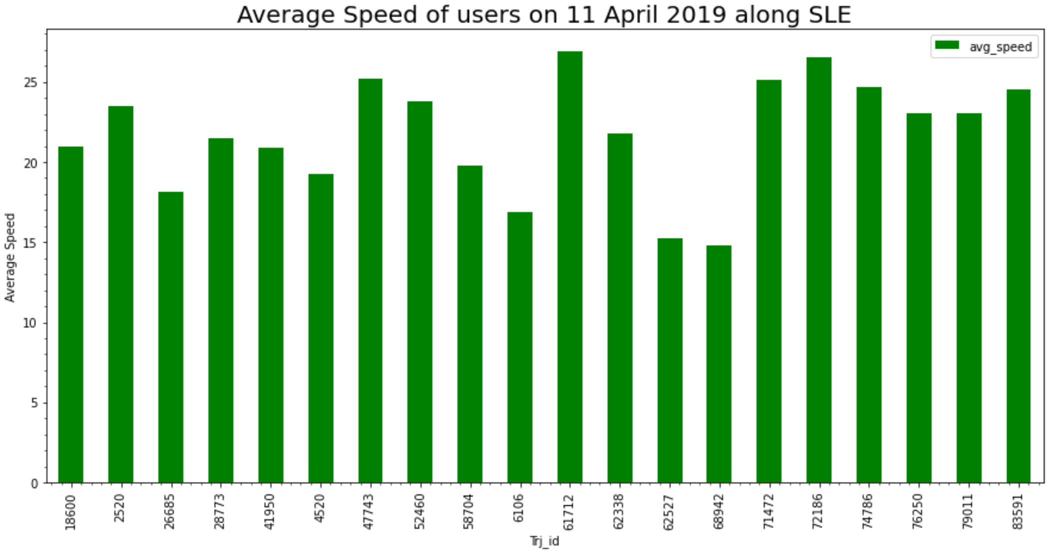
Our exit point would be 1.396340, 103.841326 as shown above.

We first filtered the dataset by keeping those trajectories that contained these two points in their samples between 9am and 10am.



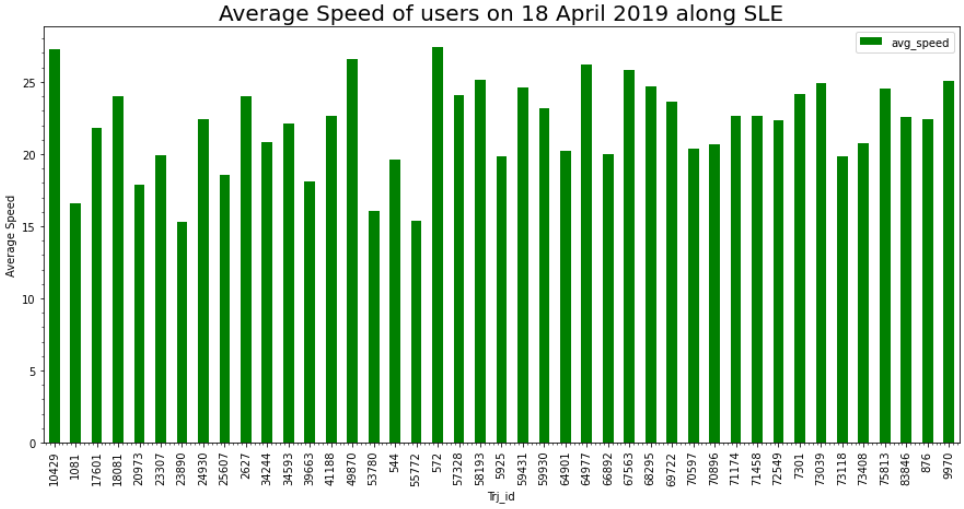
After filtering, we would end up with the heatmap above.

Since our aim is to find out if there is a spread of different average speeds between drivers who drive along this particular road during the 9am to 10am time period, we decided to choose a particular day, which in this case is Thursday, 11 April 2019, and analyse it. We will plot a graph showing the average speed of each trajectory that is in our filtered dataset on 11 April from 9am to 10am on this path.



From the graph above, we can tell that there is some difference in the average speed of the trajectories. In fact, there is a standard deviation of about 3.42km/h. These data are collected from Grab drivers on the same day, same time of day, as well as along the same stretch of road. Furthermore, since they are all from Grab drivers, rationally speaking, they should probably be driving as fast as possible to complete more orders. Thus, while there are certainly other factors involved, we assume that a significant contributing factor is due to their individual's driving pattern.

To better conclude that this is not an one-off event, we did the same process on the same dataset, but for 18 April 2019, which is also a Thursday, so that we keep the other variables constant.



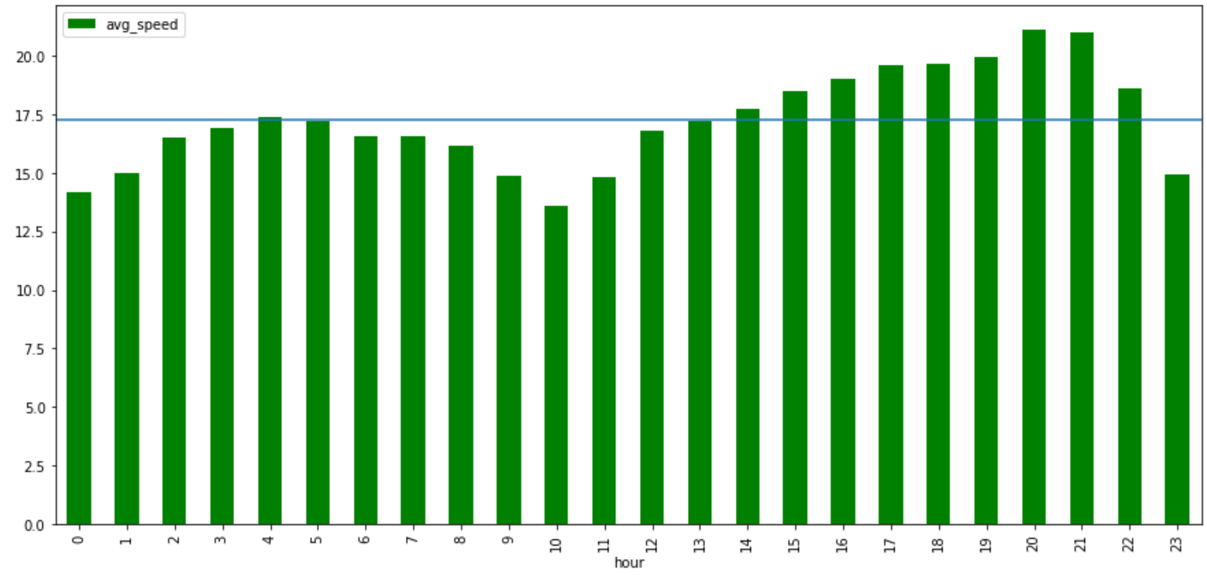
We attained the graph above which has a standard deviation of 3.12km/h.

Thus, we can conclude that different individuals have different driving patterns, and thus there should not be an one size fit all algorithm to predict the ETA of the ride. Rather, we should provide a personalised ETA for each driver based on their own data and usual driving speed along different road conditions.

This would help reduce the discrepancy between the ETA and the actual driving time, thus reducing possible delays.

For the second point, we believe that there will be different traffic conditions during different parts of the day. As a result, the ETA of each trajectory should be adjusted accordingly. Here, we shall assume that the number of trajectories per hour corresponds to the live road traffic in real time. We will break down the day by hours to find out the relative traffic conditions.

To test this hypothesis out, we shall separate the trajectories by hour. For instances whereby the trajectory spans over 2 separate hour intervals, we will split the trajectory into two, each corresponding to the hour it belongs to. After which, we will compute the average speed over the 24 hour intervals.



From all the 30 million data points from the Singapore dataset, we managed to plot this graph based on the method mentioned above. The y-axis is in terms of km/h, which shows the average speed of all of the trajectories in that hour. The x-axis is split into 24 hours, with 0 hours referring to 12am.

From this graph, we can tell that the mean is at 17.24 km/h, and there is a spread of different average moving speeds throughout the day. Thus, this supports our hypothesis that there will be different traffic conditions throughout different times of the day. As a result, to result for the spread in average moving speeds, the algorithm which Grab uses to calculates its ETA should also be adjusted in proportion to these moving speeds so that the ETA would be more accurate.

Bad Route Planning

Secondly, the delays could be due to bad route planning. With regards to bad route planning, we will rule out congestion caused by live accidents which occurs too near to our Grab driver. This is because there is no other path the driver can take to detour to take a less congested path.

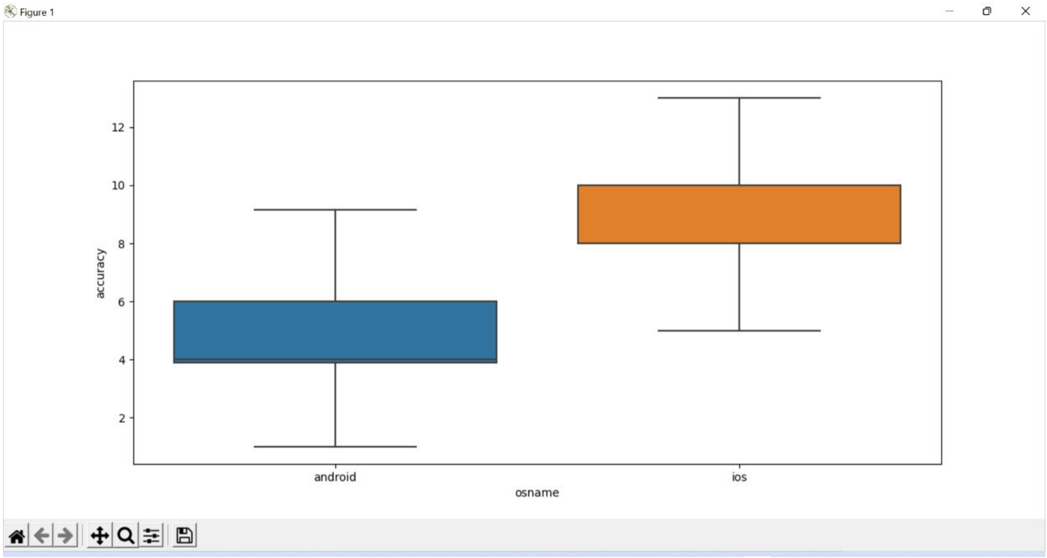
We assume and strongly believe that the current route planning by Grab takes into account road repairs, road closures, which are all pre-determined, as well as congestion caused by accidents and due to the sheer number of cars, which are live information, plus congestion in the future due to peak hours, based on historical data.

Thus, we would like to look at it from a different perspective. When two routes take roughly the same amount of time, we should allocate the driver a route that is known to take less time historically. For instance, one of the routes contain a road that is more prone to accidents based on historical records, possibly due to road structure or blind spots. Considering that both routes take a similar time, we should try to avoid the route which contains such a road, as there is a higher chance of a congestion happening due to a higher chance of an accident occurring. However, as such data is not provided in the GPS Pings dataset, we are unable to validate this hypothesis.

Another avenue in which we can improve on route planning would be to reduce allocating routes which has poor signals. This is because we are unable to track the driver's location accurately and give live feedback on the route taken and ETA. Without an accurate GPS signal received from the driver, we are unable to track the ride accurately. Thus, we can unable to reflect changes in ETA in real time when the driver changes his or her speed. When the car reaches a location with better signal and the ETA is updated, there could be a huge difference with the original ETA, significantly reducing user satisfaction. However, this is based on the assumption that given two routes with same estimated travelling time, we should prioritise choosing roads with better signals. Another assumption would be that on any given day at any given time, the road that historically has poor signal reception will still have poor signal reception.

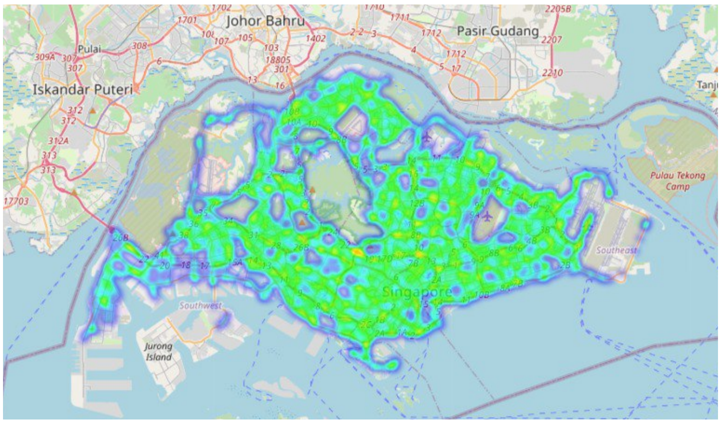
With these assumptions in place, we can then move on to the analysis. We would first like to find out which are the roads that have poor signals. Since we do not have concrete data on this, we assume that the quality of the signals is highly correlated to the accuracy level, hence we will use that as a gauge.

We will create a heat map of each GPS ping sample that have high accuracy levels from the Singapore dataset. We realised that ios and android device have different accuracy levels, thus we cannot set the same threshold for the definition of a having a high accuracy level across all devices.



The graph above shows the spread of accuracy levels for android and ios devices. The threshold that we will set is the 75th percentile of each device type, which is 6m for android devices and 10m for ios devices.

We will then plot a heatmap for those datapoints which are above the threshold . From there, we could then see if there are any overlaps which would be shown through the darker portions on the map.

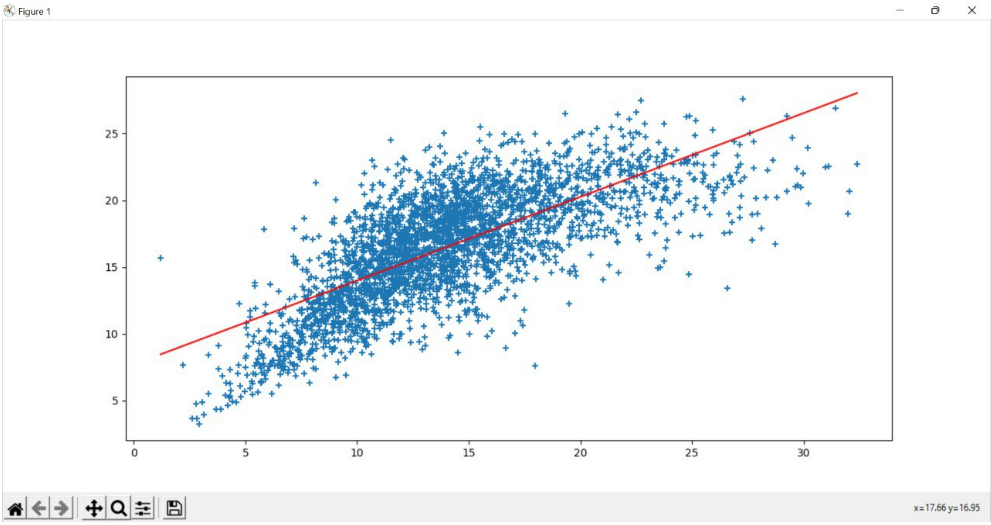


The graph above shows that there is actually no significant difference in the areas in which high accuracy levels are detected. There is generally an even spread. Hence, our hypothesis is incorrect and thus this is not a significant factor in causing the delay of ETA.

Analysis Of Outliers

With the analysis of the factors above done, we would like to see if we can spot any trends from trips which take a longer time relative to ones of a similar distance. For this, we assume that for trips with longer distances, in order to reach the destination faster, usage of the expressways is typically more frequent as they have higher speed limits as well as fewer traffic lights. This is also true for the converse, whereby for shorter distances, normal and minor roads are mainly used to navigate around neighbourhoods, which have lower speed limits and typically more traffic lights. We believe that this is a valid assumption given that Singapore is relatively small and has good traffic infrastructure. The assumption should also hold for Jakarta, as it is relatively more sparse as compared to Singapore. Thus, there might be a correlation between the distance of the trip and the average speed of the car. If we were to prove this correlation, we could find out the average speed of cars in each distance category. With this information, we can then filter out those trips which have a slower average speed within each distance category and see if we can spot any trends from it.

We will first compute the average speed of each trajectory and plot a scatter plot to see if there is a correlation between the average speed of trajectories and the distance of trajectories.



The graph above shows a positive relationship between the distance of each trajectory and the average speed. This finding confirms our hypothesis. Each driver may have different driving patterns throughout the ride. For instance, for a long ride which typically involves first moving to the nearest highway from a neighbourhood, cruising through the highway, and lastly exiting the highway to reach another neighbourhood, we hypothesize that the same driver could have different driving patterns for each of these segments. Thus, if we could find out more about the driving patterns of a typical driver in each of these situations, we could better understand how different situations on the road will impact the ride duration, and thus aid us in providing a more accurate ETA.

## Conclusion

In conclusion, we attempted to tackle the issue of delays from various perspectives. We highlighted various pointers which we believed could have been a cause for the delay and provided a solution in which we hope can help Grab improve on the ETA prediction algorithm.