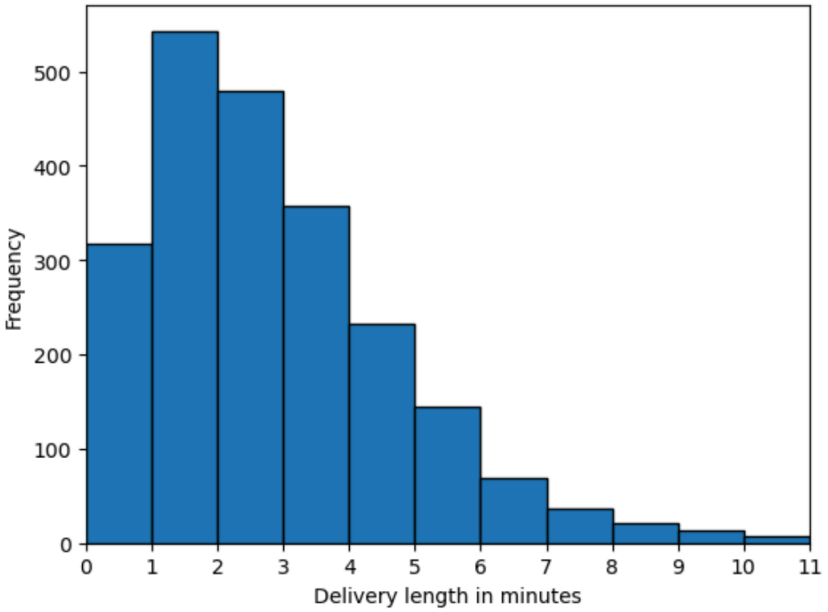
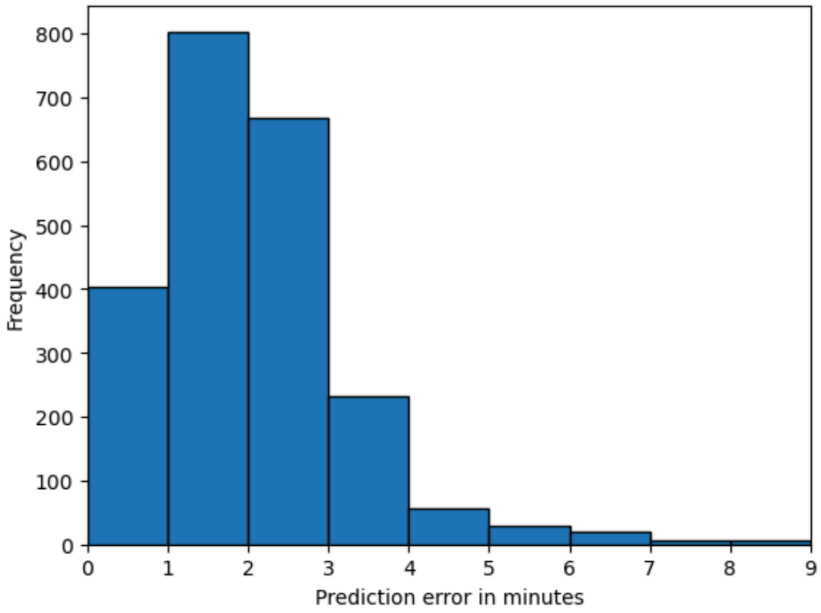
Data analysis and visualization

Below I present insight into the droptime database. Charts below were created from almost all samples. Samples excluded from charts and further analysis include samples where delivery time was below zero and samples where delivery time was longer than 30 minutes. The reason for excluding these samples was because I decided it was erroneous data. Delivery time cannot be smaller than 0. There were almost 30 samples where delivery time was around 4 hours and the next highest value of delivery time after samples where delivery times was close to 4 hours was 11 minutes so I decided that those samples are erroneous. I also excluded samples from table route_segments where order_id was empty, as I decided to analyze those samples separately. In parts 1, 2 and 3 I used only samples where segment type was equal to 'STOP'.

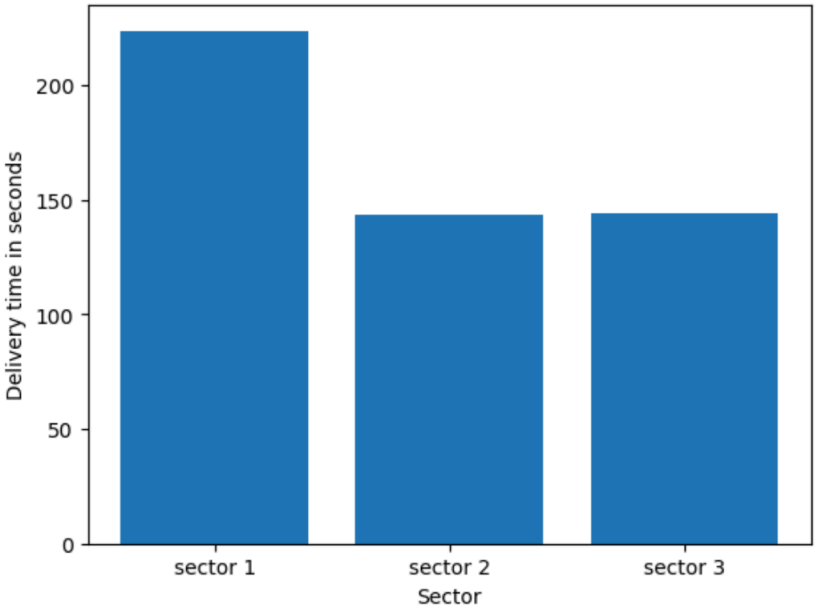
1. Histogram showing actual delivery length.



2. Histogram showing prediction error.



3. Bar chart showing delivery time depending on sector.



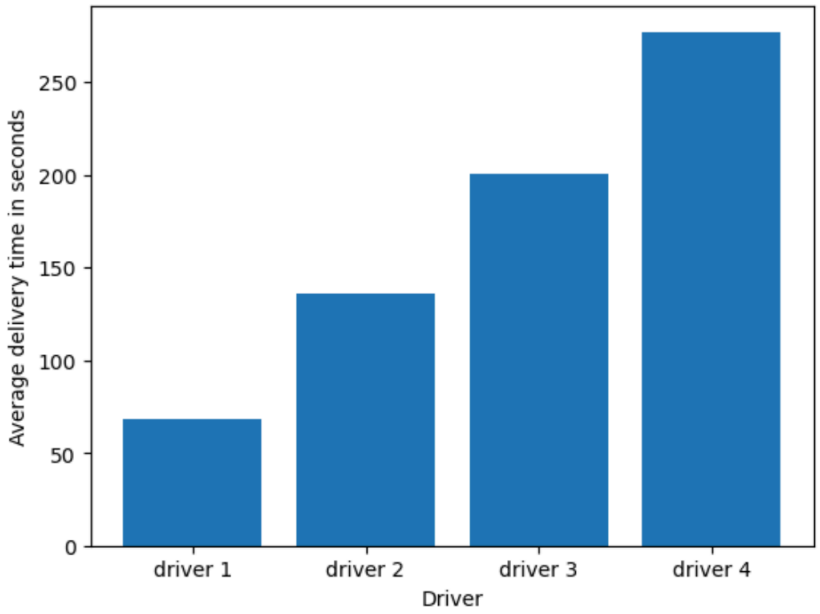
This chart proves the hypothesis that indeed in one of the sectors delivery times are longer. It is visible that delivery times in sectors 2 and 3 are around 160 seconds, while in sector 3 it is around 220 seconds.

4. Other trends and correlations:

- ***Correlation between actual delivery time and driver.***

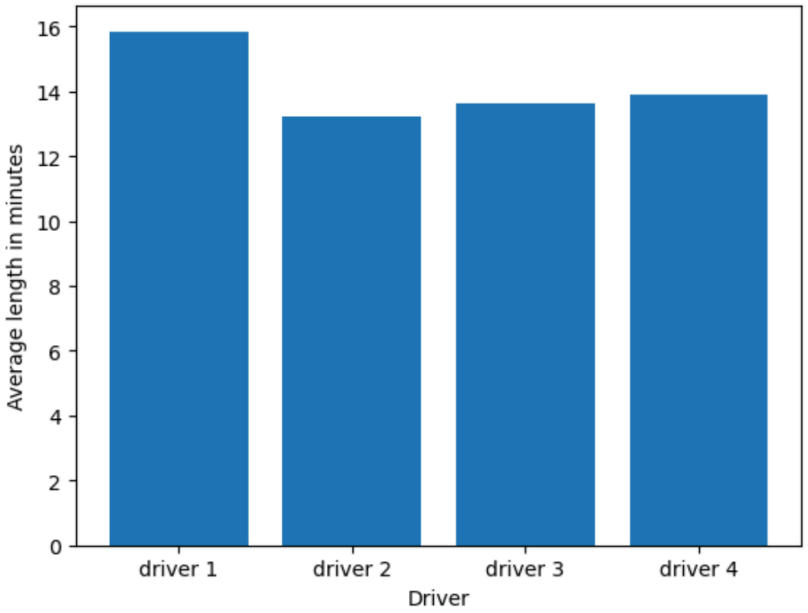
I decided to investigate the hypothesis that it might take longer for some drivers to complete deliveries.

Average time of STOP segments where order id is not missing:



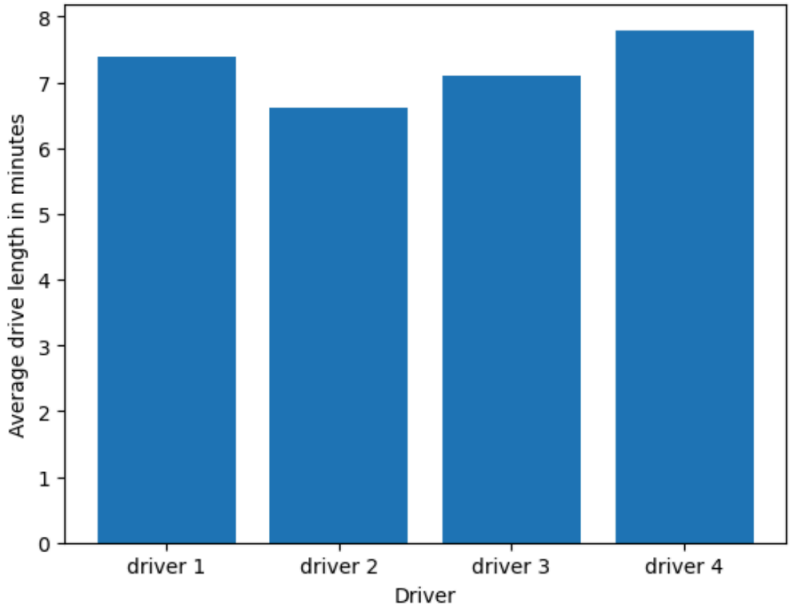
There is a big difference between delivery times between drivers during stop segments.

Average time of stop segment where order id is missing:

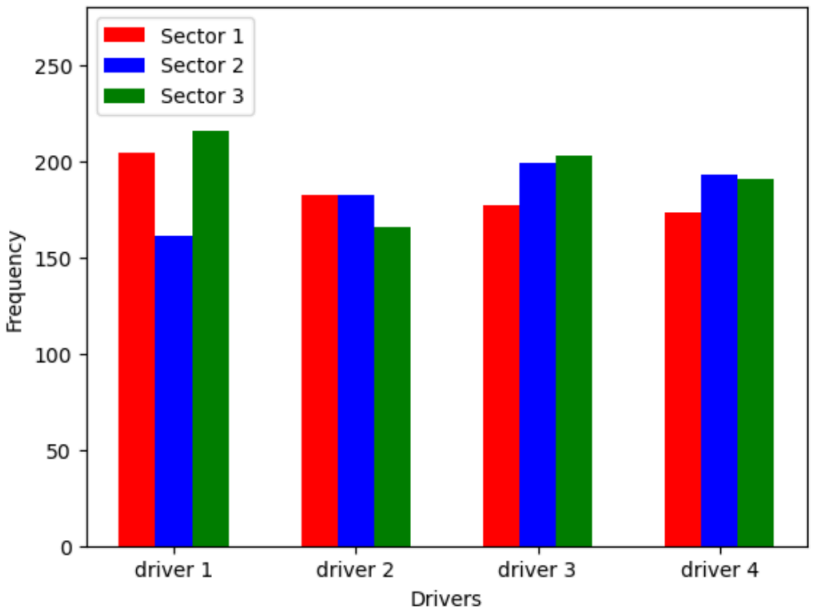


Average stop time during stop segments where order id is missing is slightly bigger for driver 1.

Average drive time:

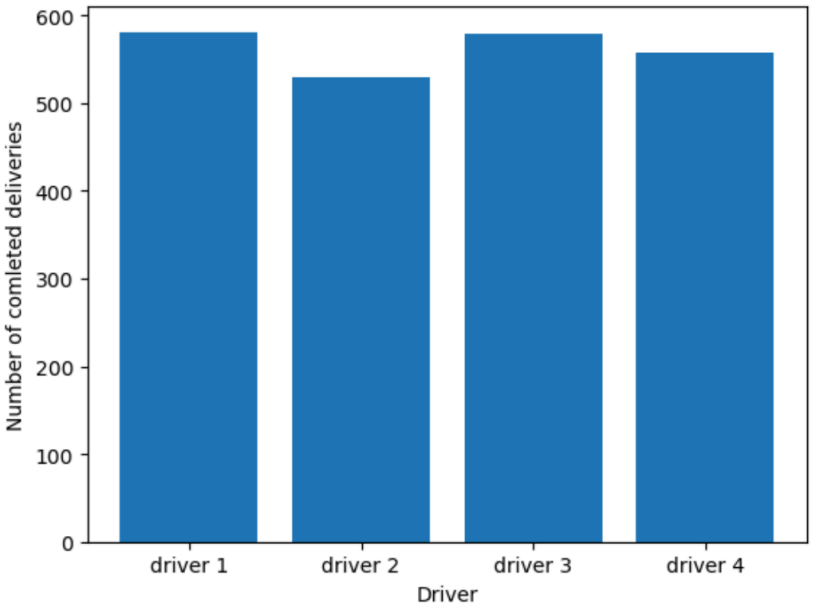


Average drive time is slightly longer for driver 4.

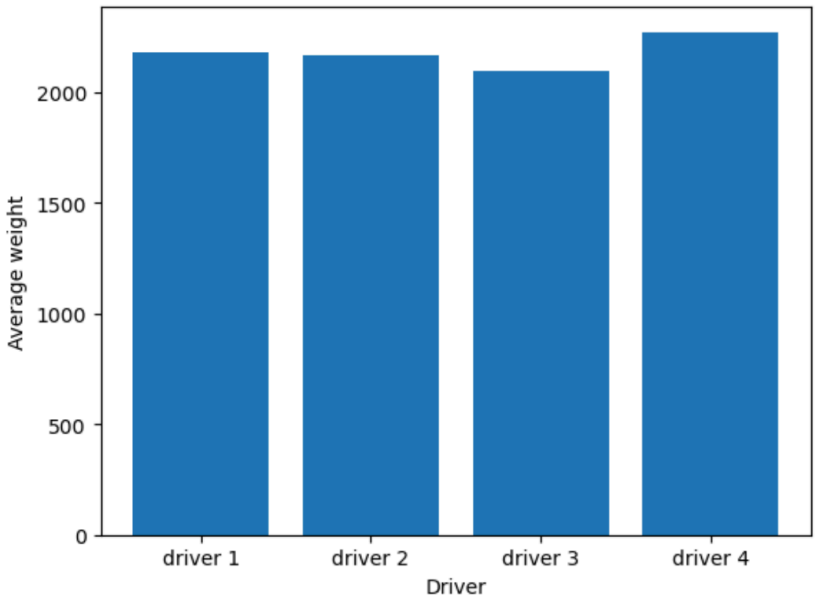


Driver 1 delivers mostly in sectors 1 and 3 while others deliver more or less the same amount in each sector.

Amount of deliveries completed by each driver:

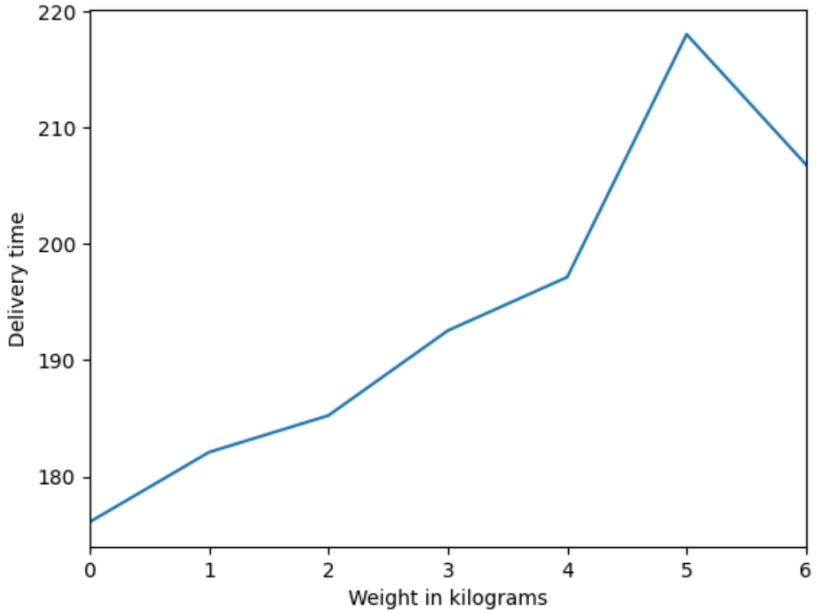


Average weight of deliveries per driver:



Conclusion: It is visible that driver 1 is quicker than other drivers. Most probably it is not because of the sector where he or she delivers because driver 1 delivers more in sector 1 than other drivers and deliveries in sector 1 take longer as shown earlier. It is most likely not because of the average weight of deliveries as average weight is similar for all drives. Average stop time in samples where order id is missing is longer for driver 1; those might be personal brakes, if that is the case it might be beneficial to encourage drivers to rest between deliveries.

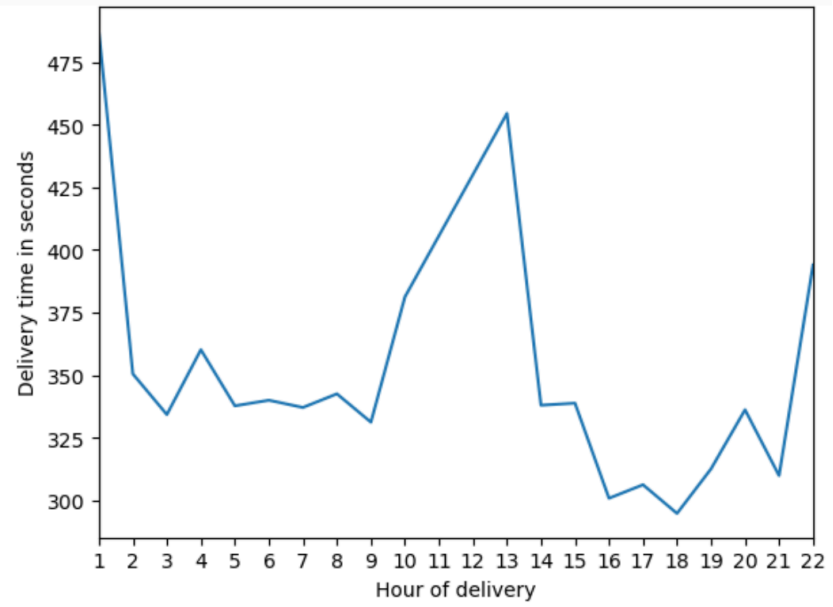
- ***Correlation between actual delivery time and weight of delivery.***



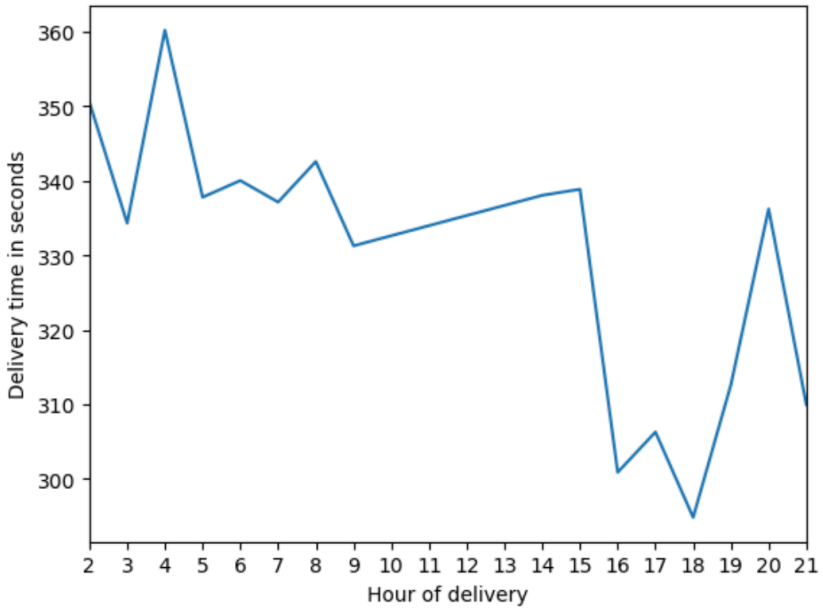
This chart shows a growing trend of delivery time depending on weight.
Delivery time is inclining with weight with slight decline after 5 kilograms. This may be because of the fact that there is less data about deliveries heavier than 5 kilograms.

- ***Correlation between hour of delivery and delivery time.***

First chart shows delivery time depending on hour of delivery, including all data (both segment types)



It is visible that deliveries take more time at around 1am, 1pm and 11pm. Deliveries take the shortest amount of time at around 6pm. Although the amount of data about deliveries during several hours is significantly smaller than in others. Below is a chart showing delivery times depending on hour, but containing only data where number of samples is bigger than 100:



In this chart we can see that deliveries take more at 4am. As in the previous chart it is visible that deliveries take the shortest amount of time at around 6pm

It is hard to conclude when deliveries take the longest as there is no sufficient data, but it is visible that deliveries at around 6pm take the shortest. It is visible on both charts.