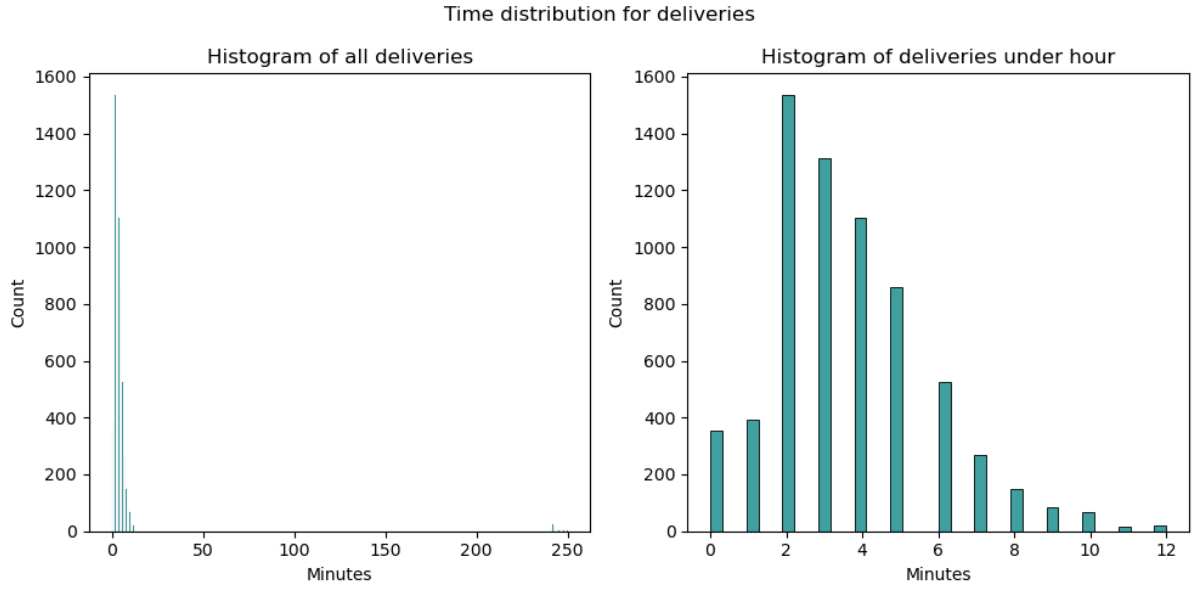
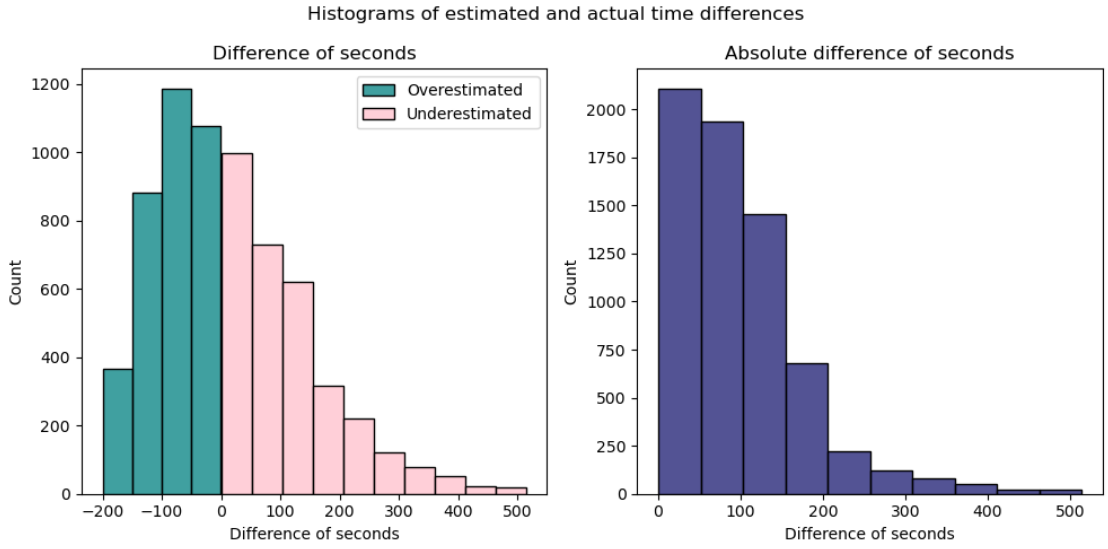
Delivery Time Analysis

I started with exploring the data, first I calculated the delivery time, on which the whole analysis is based. I then found out that some of the times were negative, so I deleted them as time can not be negative.
Then I calculated that out of 6752 records, more than half came on time. The highest planned delivery duration is 200 seconds, however, in reality, 40% of deliveries exceed this time. This indicates a significant discrepancy between planned and actual delivery durations. Because planned delivery is a mean of all of the records, it is not personalized to the individual record.



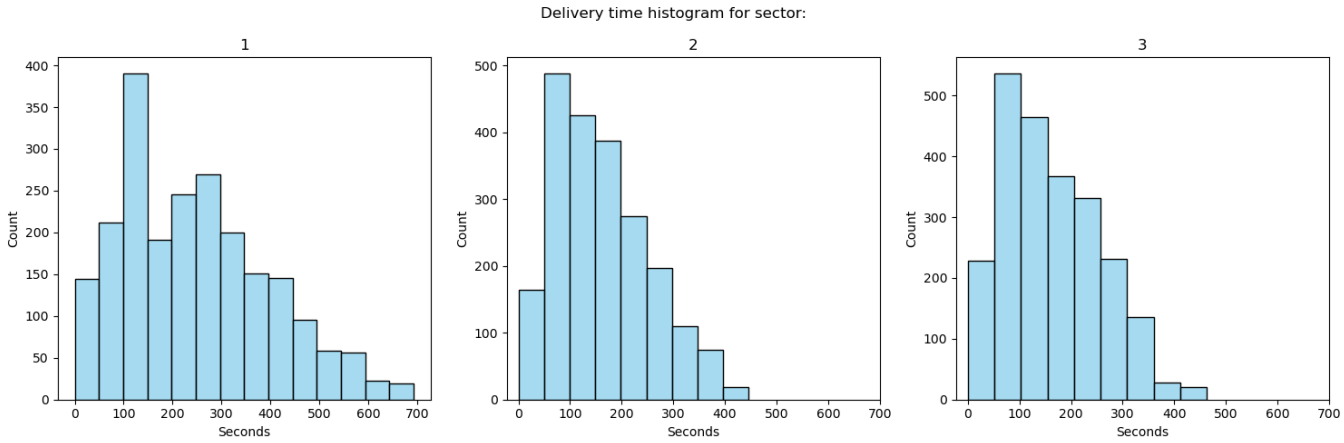
As it can be seen on the plot, there is a significant gap between 15 and 240 minutes, this may for example indicate a scenario where delivery drivers recorded all of their deliveries as being completed at the end of their shift, possibly during a break, without accurately logging the exact delivery times. After considering only the first part we can see a Histogram with a distribution resembling Gaussian but with majority of the deliveries clustered near the lower range. Given the presence of these high times, I suspect that these records may not represent actual delivery behavior but rather could be errors or anomalies in the data. I will exclude these outlier values from further analysis to ensure more reliable insights. I was already suspicious of the records in the earlier part (describe table) but this visualization has solidified my understanding.

My next step was exploring the differences between predicted and estimated times.



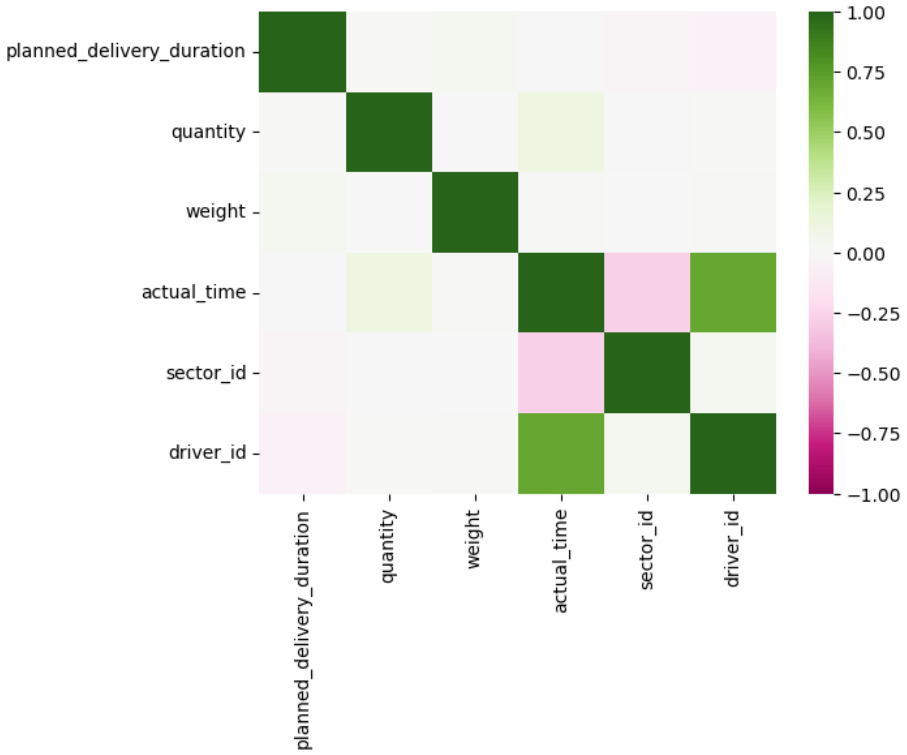
In the first plot we can see that the data is clearly split into two categories based on whether the actual delivery time was overestimated (actual time is shorter than planned) or underestimated (actual time is longer than planned), overestimated category has less bins, with maximum difference of 200, while underestimated has a wider range to over 500. The second plot, which shows the absolute differences, helps to clarify the overall trend, the majority clusters around 0, with higher counts up to 200 second time difference. As the differences increase, the number of deliveries decreases.

I proceeded with the analysis and discovered that the average mistake of prediction the delivery time variable in the first sector is over a minute, while in the rest two it's close to 15 seconds. Also the average time of delivery is half longer. Each sector has similar amount of records.



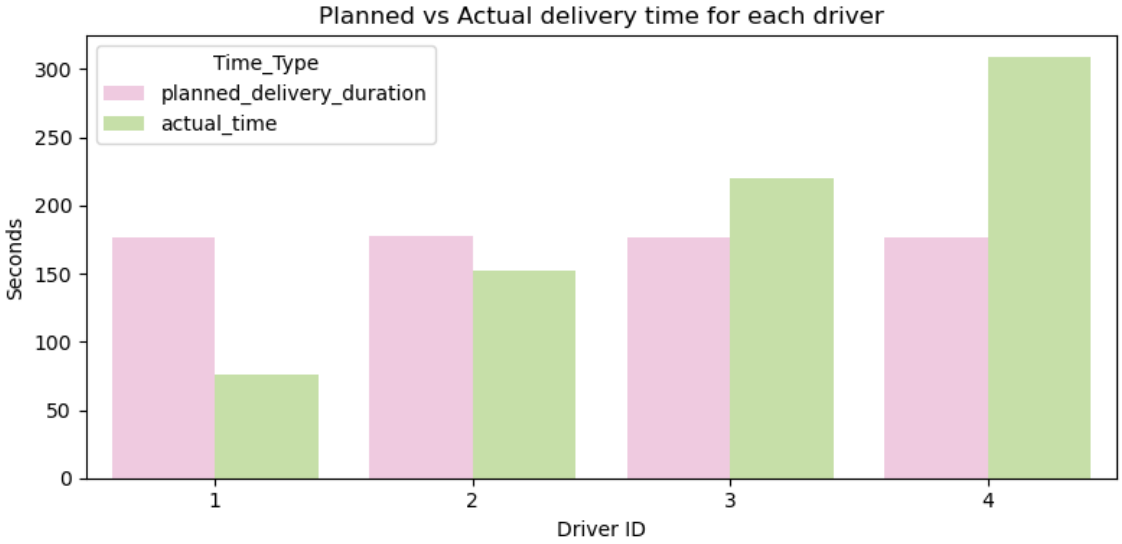
As we can see from the plots, the delivery time range for the first sector is noticeably wider—by almost 300 seconds—compared to the others.
Another interesting observation is that the first sector is the only one where the mean prediction error is positive. Moreover, this error exceeds one minute, which indicates that the planned delivery time is significantly underestimated. This misestimation may be skewing the overall performance results and suggests that the prediction model should be adjusted for this specific sector.

I decided to explore the correlations between different variables.



The only somehow relevant correlations are driver_id and sector_id with the actual delivery time, I already focused on the sectors, so let's explore drivers. I made a calculation, by grouping the data by driver_id, and calculated the mean time of planned delivery duration as well as actual time.

driver_id	planned_delivery_duration	actual_time
1	177.046180	75.935576
2	177.388996	151.961612
3	176.397668	220.027988
4	176.644888	308.990926



As can be seen both from the numbers and the plot, Driver with ID 1 is the quickest. There is a visible difference of delivery time depending on driver, even though the driver with id 1 was in the hardest sector with ID 1 as much as others, he was still the fastest one.

Conclusion

Planned delivery times are frequently underestimated, around 40% of deliveries exceed the maximum planned duration. This issue is particularly visible in Sector 1, where the average prediction error exceeds one minute and the spread of delivery times is significantly wider compared to other sectors. It takes me to conclusion that sector 1 needs to be prioritized in model adjustment, as it is skewing model performance the most. There is also a noticeable variation in performance between drivers, driver 1 was the fastest, despite working in the most difficult sector just as often as others. This should also be considered in prediction model. Other variables such as quantity or weight show practically no correlation and likely don't impact delivery time.

I would focus on sector 1 and refining the prediction there to better reflect its longer delivery durations, as the current estimates are consistently too low. Generally since delivery performance clearly varies between both sectors and drivers, future predictions could incorporate these 2 factors. This would allow the system to better adapt and result in more accurate estimates.