OCADO TECHNOLOGY INTERNSHIP TASK

Stack used : Python, Jupyter Notebook, SQL

INFO : All the required tasks are done at the end of this PDF file.

Python code : LINK

Name: Wiktor Łach

Location : Kraków

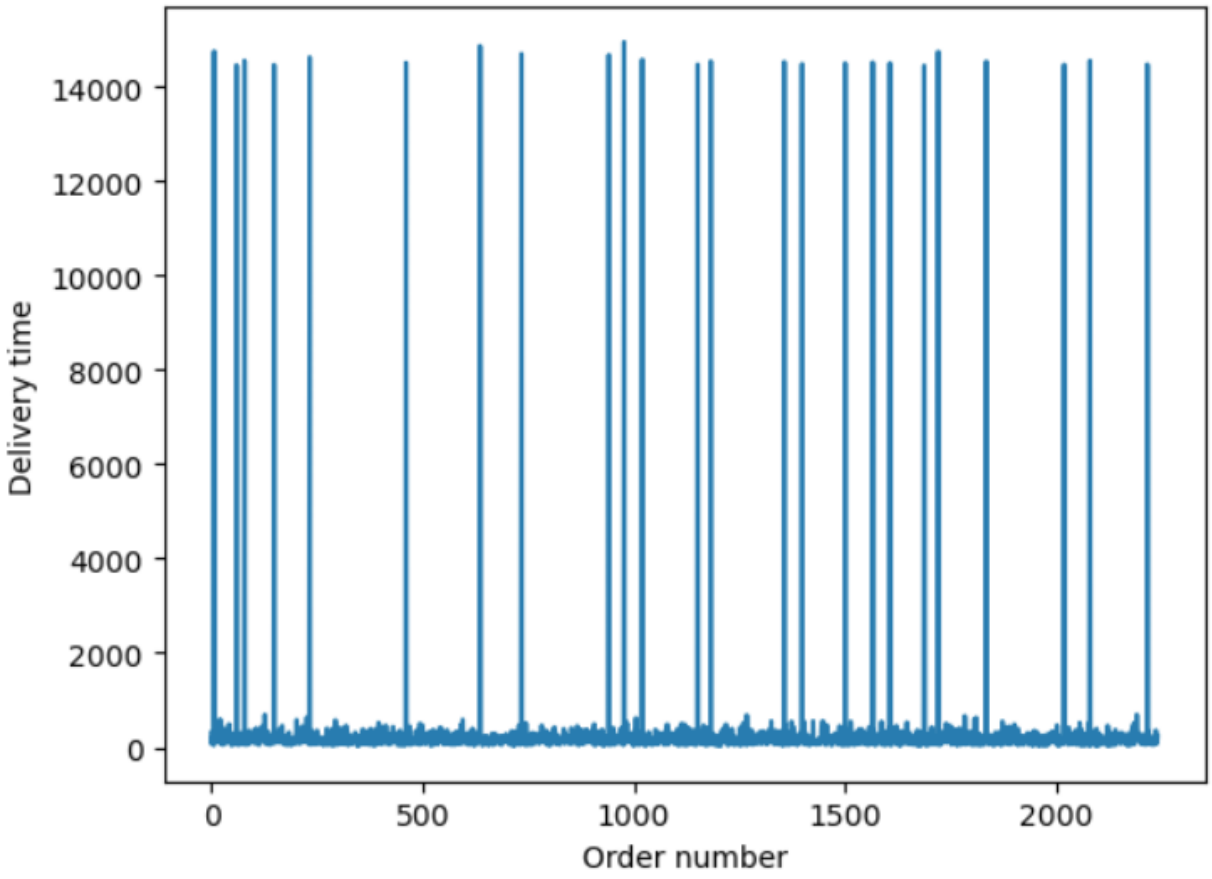
wiktorr.lach@gmail.com

After exporting data from SQL as CSV I create additional column in data frame which is delivery time (the difference between start and end of delivery). Then I begin cleaning data.

First of all I notice that delivery_duration column has some negative values in it. There are 35 rows where the start time is later than end time so I delete them. It's important to understand why this error occurred, because potentially the rest of the data can also be influenced. For example GPS might not work properly or partitioning algorithm has some bugs.

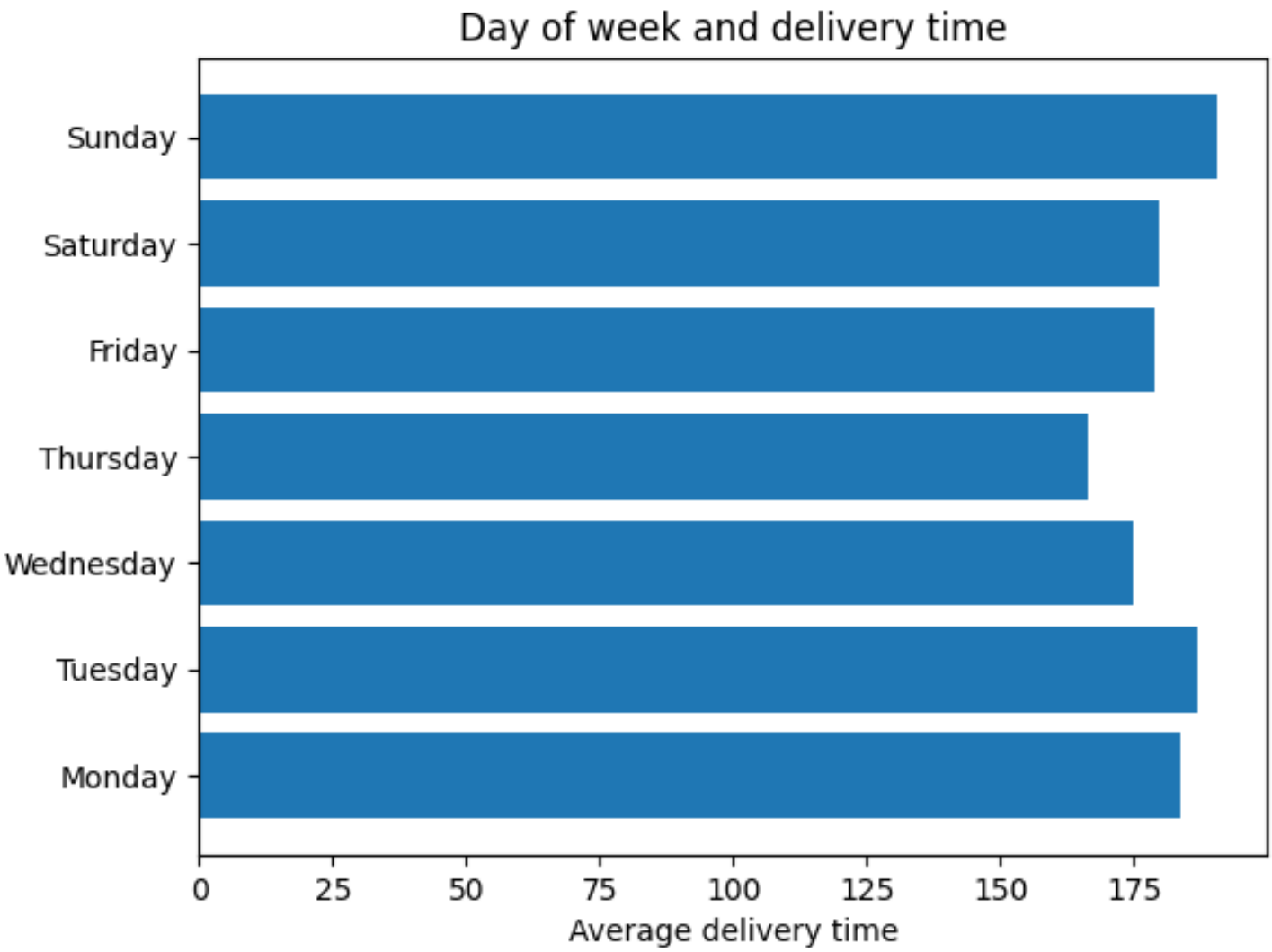
segment_type	segment_start_time	segment_end_time	delivery_duration
STOP	2024-02-24 14:44:13	2024-02-24 14:39:53	-260

After displaying plot with delivery times I see that there are some outstanding records. I assume that it was some error or some unexpected accident happened (4 hours is too long delivery time). I delete rows with incorrect data once again by deleting all rows where delivery time is over one hour.

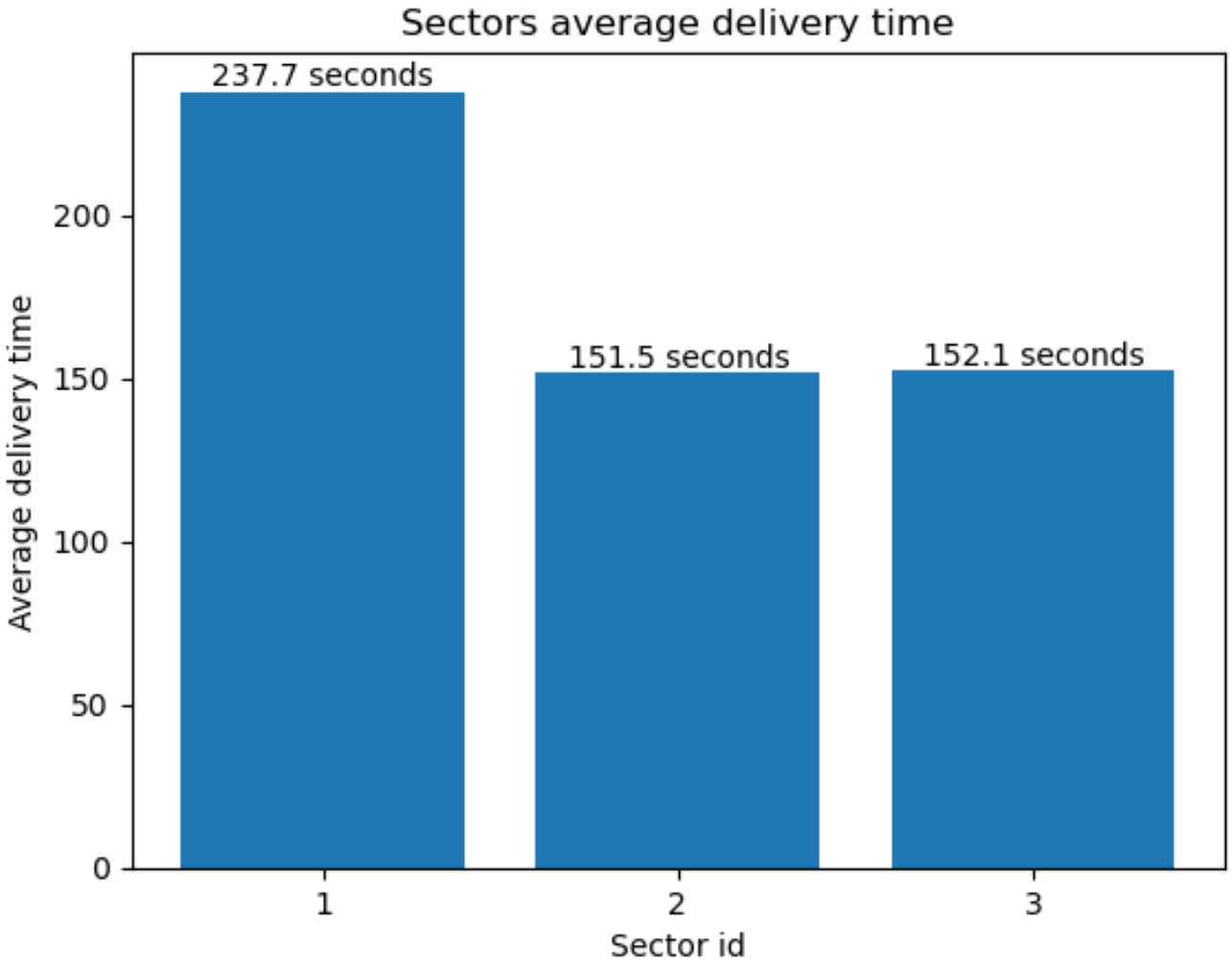


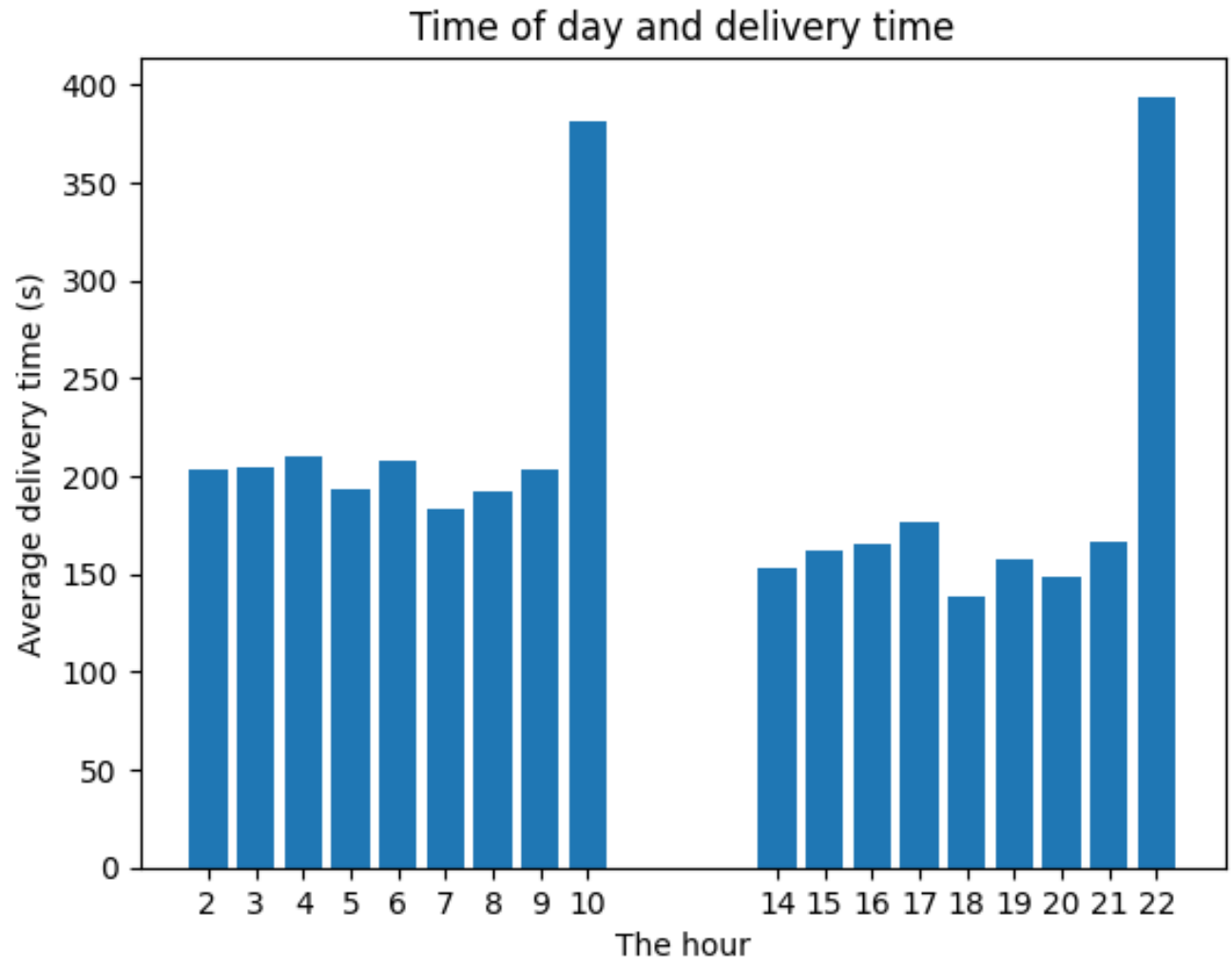
After cleaning the data I don’t see any inconsistencies in data so I begin to transform the data frame. There are multiple rows of the same order if there is more than one product, so I leave only one and I add a column with total count of all products ordered.

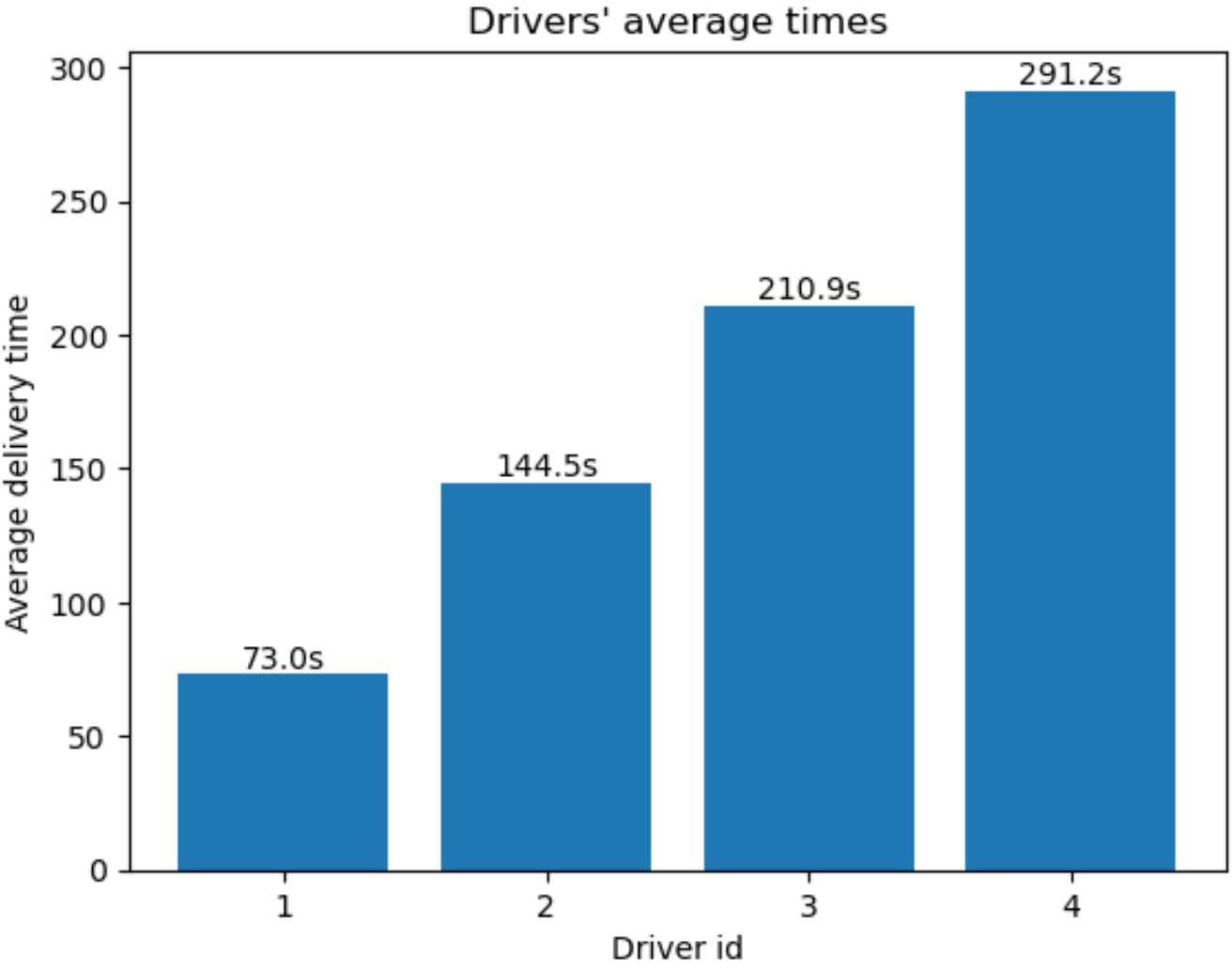
After that I want to see some correlations between delivery duration and other data so I run multiple plots to see if I can notice some patterns.




Number of ordered products and delivery time
Average delivery time
Number of ordered products
0
50
100
150
200
250
300
0
2
4
6
8
10
12
14


Sectors average delivery time
237.7 seconds
151.5 seconds
152.1 seconds
Average delivery time
Sector id
0
50
100
150
200
1
2
3


Time of day and delivery time
Average delivery time (s)
0
50
100
150
200
250
300
350
400
2 3 4 5 6 7 8 9 10
14 15 16 17 18 19 20 21 22
The hour


Drivers' average times
Average delivery time
0
50
100
150
200
250
300
73.0s
144.5s
210.9s
291.2s
1
2
3
4
Driver id

Conslusions (task 4):

As we can see the type of sector and the number of products in order have the highest impact on delivery time. Sector number one might have the problems with traffic jams, or it might be placed the furthest from the grocery. I can suspect that the number of ordered products impacts the delivery time, because for example the delivery man might not be able to pick up all the products at once (high volume).

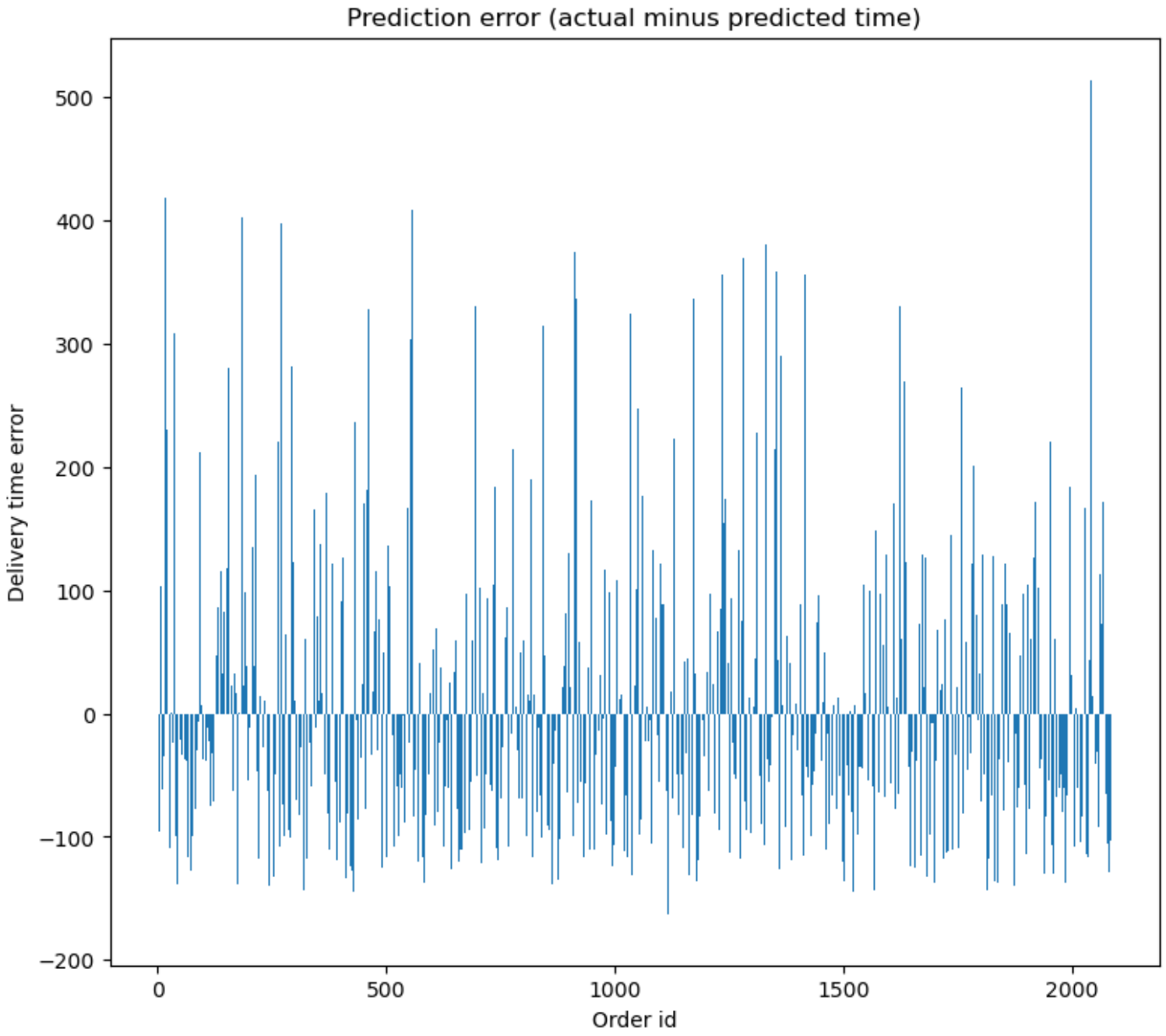
Big differences in delivery time are also noticeable when we take a look at the time of day. Between 11:00-14:00, and between 23:00-2:00 there are no deliveries made. For example it might be break caused by the change of people in the shift. We can notice that at the end of those potential shifts the delivery times are the highest. It would be important to get inside info and understand why it happens.

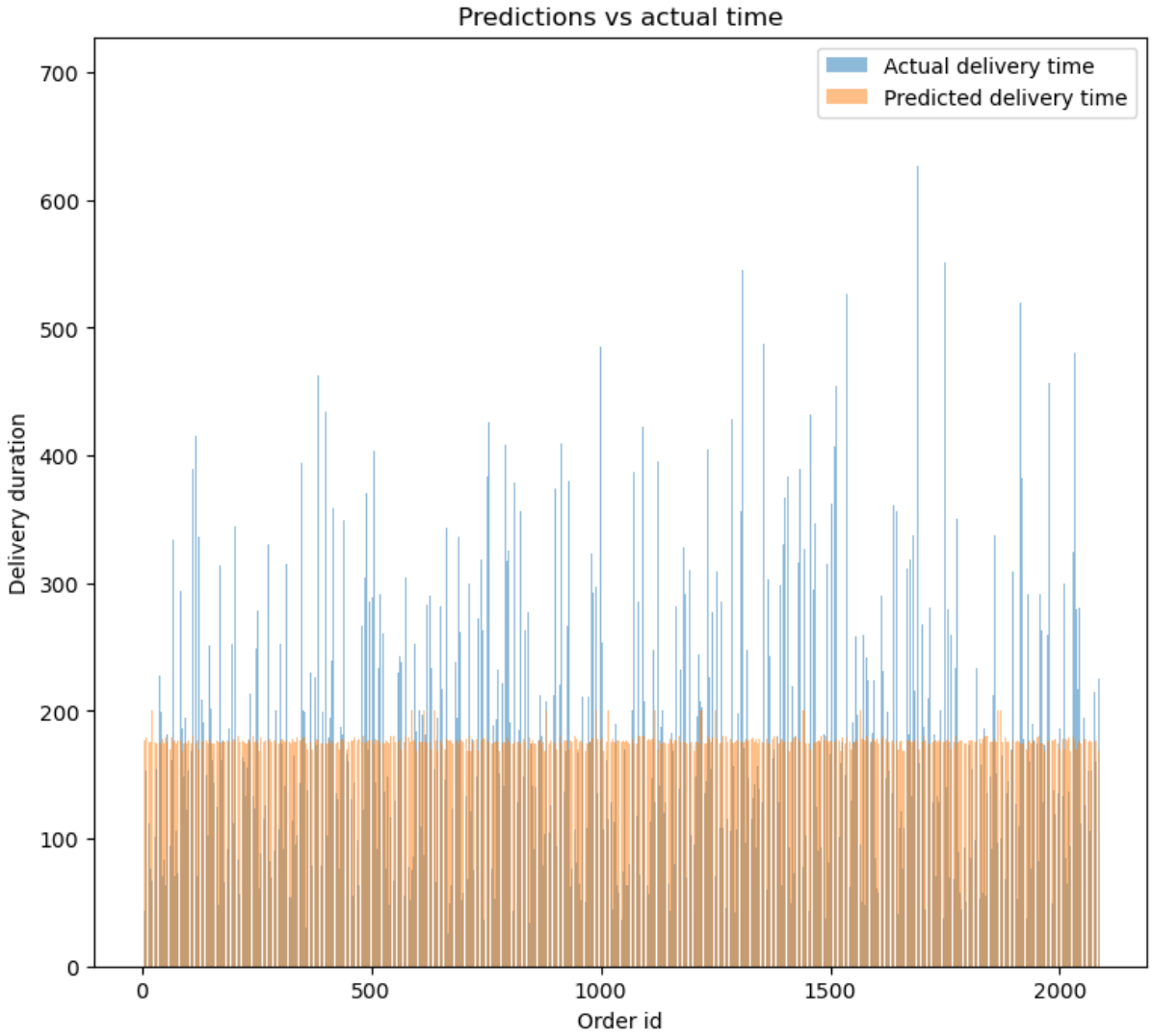
Drivers' average times are pretty weird, because driver number 1 delivers products four times faster than the 4th one, but I have no idea about the work organization, so maybe it is caused by not choosing random to specific order by random. It isn't caused by statistical error, because the numbers of deliveries for every driver are almost the same as shown on the picture below:

driver_id	
1	1648
3	1630
4	1579
2	1472

There are some differences between day of week. On Sunday the average delivery times are the highest. They aren't that significant but they could help with making predictions anyway.

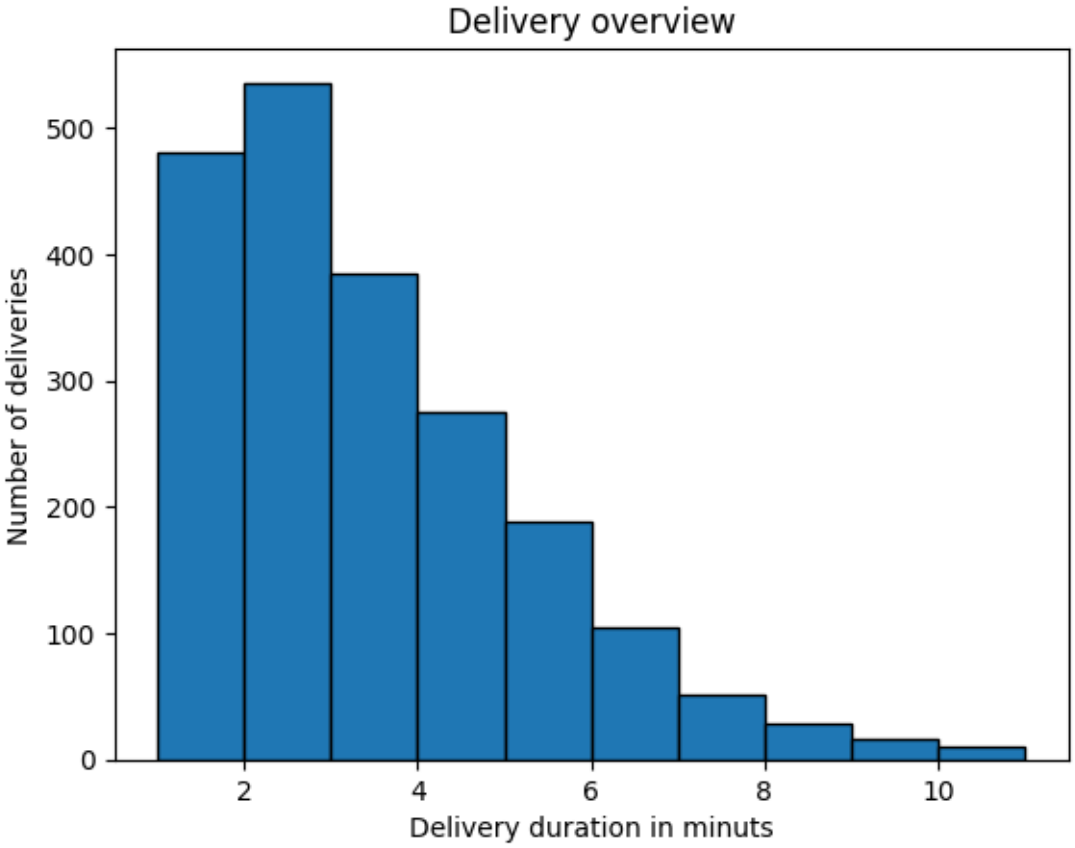
Below there are some graphs made for analyzing the accuracy of model for predicting delivery times that has been used in the past.



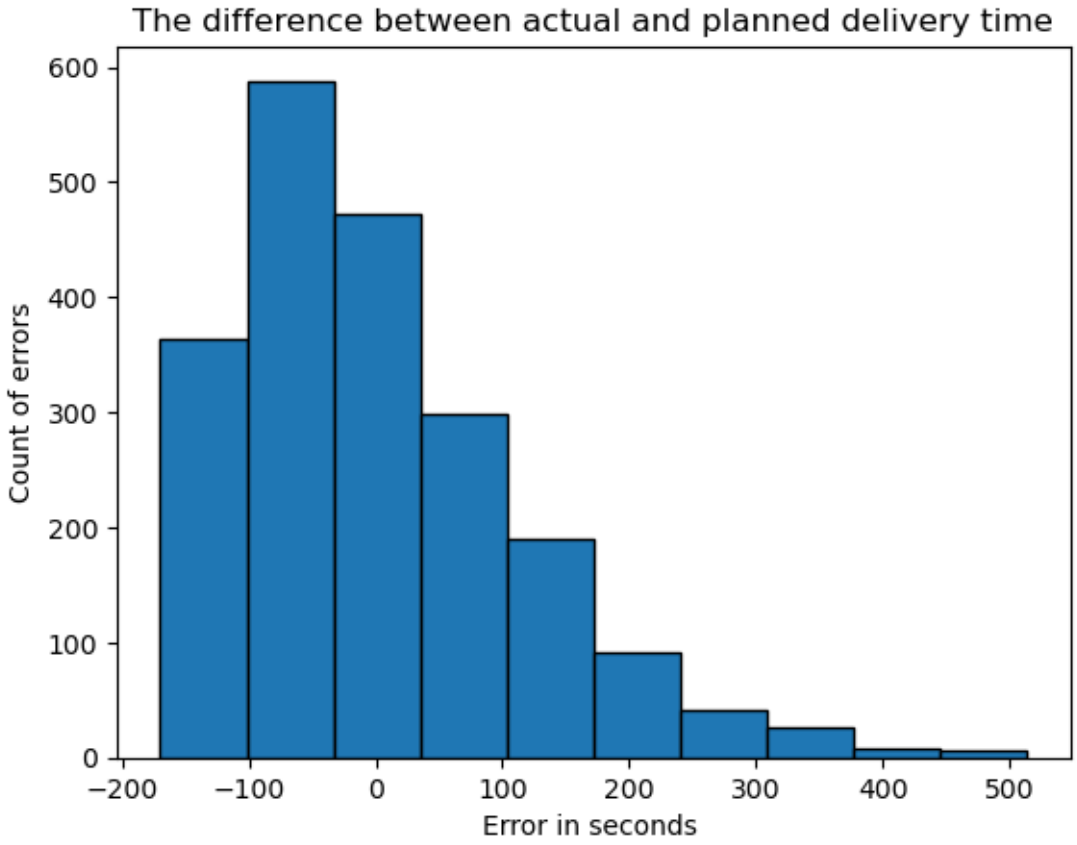


The predictions are almost the same for every order which is made. The errors are pretty big. Often orders take way more time that it was predicted.

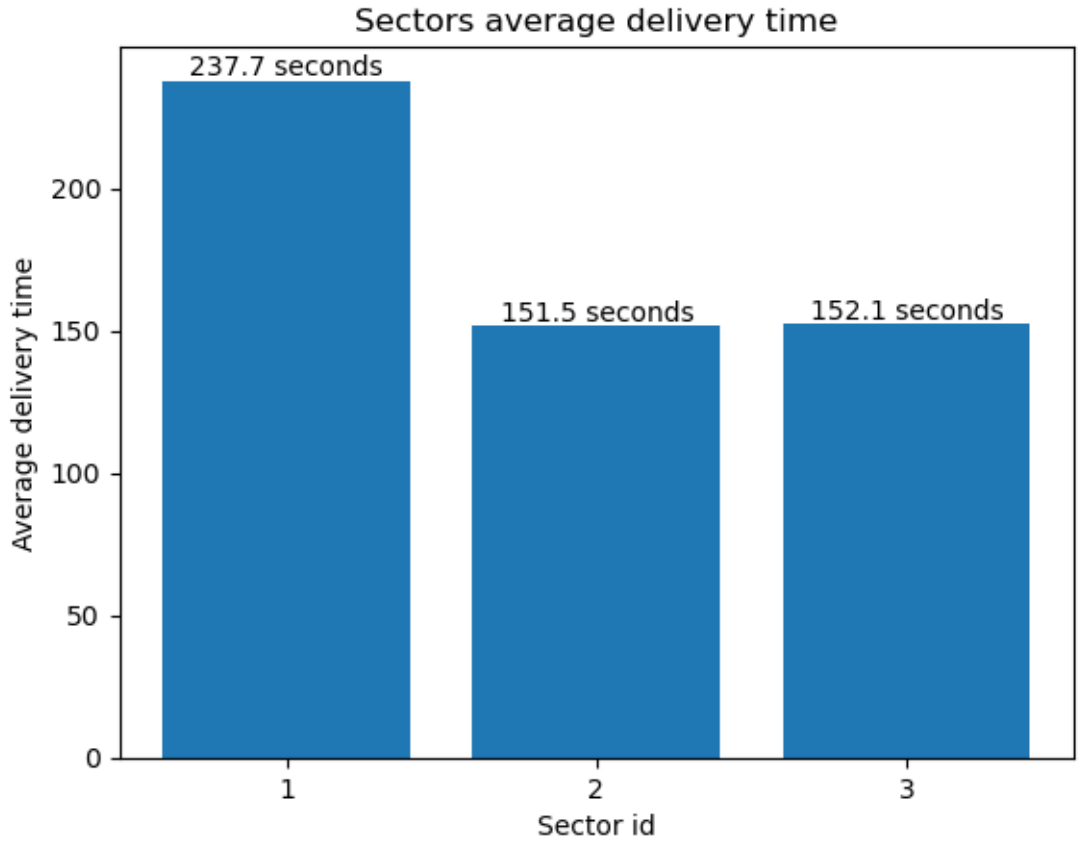
1. Generate a histogram showing the actual delivery length with 1 minute granularity (rounded up).



2. Generate a histogram showing prediction error (difference between planned and actual delivery times).



3. We received insight from our drivers that delivering in one of the sectors is significantly longer than in other sectors. Generate a chart to visualise this hypothesis.



4. Play with the data by grouping, aggregating and remodelling it. Are you able to find any correlations or trends that could be valuable for prediction quality improvement? Describe briefly your findings and visualise them on charts.

   Already done in previous part of this PDF.