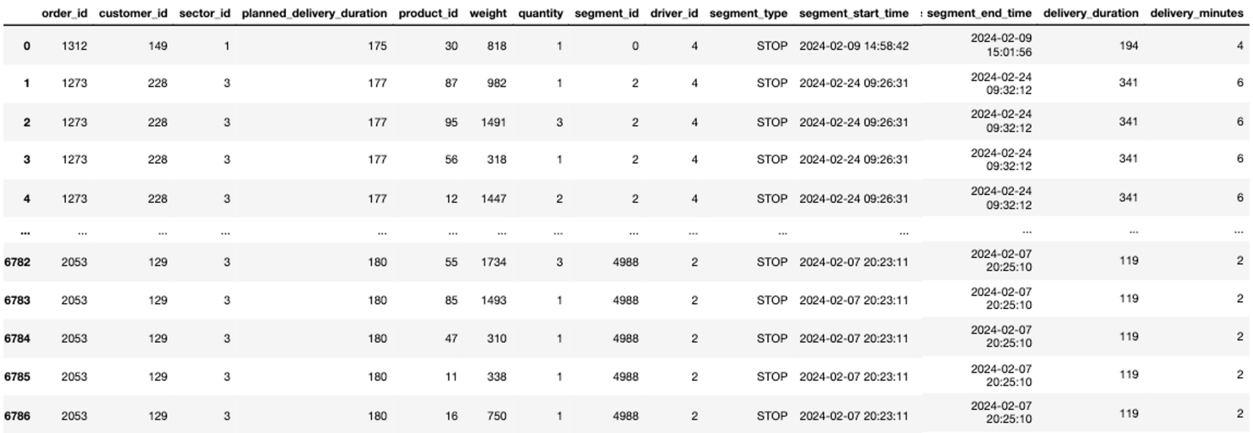
Delivery Duration Analysis

Zuzanna Jarlaczyńska

April 7, 2024

1 Data overview



	order_id	customer_id	sector_id	planned_delivery_duration	product_id	weight	quantity	segment_id	driver_id	segment_type	segment_start_time	segment_end_time	delivery_duration	delivery_minutes
0	1312	149	1	175	30	818	1	0	4	STOP	2024-02-09 14:58:42	2024-02-09 15:01:56	194	4
1	1273	228	3	177	87	982	1	2	4	STOP	2024-02-24 09:26:31	2024-02-24 09:32:12	341	6
2	1273	228	3	177	95	1491	3	2	4	STOP	2024-02-24 09:26:31	2024-02-24 09:32:12	341	6
3	1273	228	3	177	56	318	1	2	4	STOP	2024-02-24 09:26:31	2024-02-24 09:32:12	341	6
4	1273	228	3	177	12	1447	2	2	4	STOP	2024-02-24 09:26:31	2024-02-24 09:32:12	341	6
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
6782	2053	129	3	180	55	1734	3	4988	2	STOP	2024-02-07 20:23:11	2024-02-07 20:25:10	119	2
6783	2053	129	3	180	85	1493	1	4988	2	STOP	2024-02-07 20:23:11	2024-02-07 20:25:10	119	2
6784	2053	129	3	180	47	310	1	4988	2	STOP	2024-02-07 20:23:11	2024-02-07 20:25:10	119	2
6785	2053	129	3	180	11	338	1	4988	2	STOP	2024-02-07 20:23:11	2024-02-07 20:25:10	119	2
6786	2053	129	3	180	16	750	1	4988	2	STOP	2024-02-07 20:23:11	2024-02-07 20:25:10	119	2

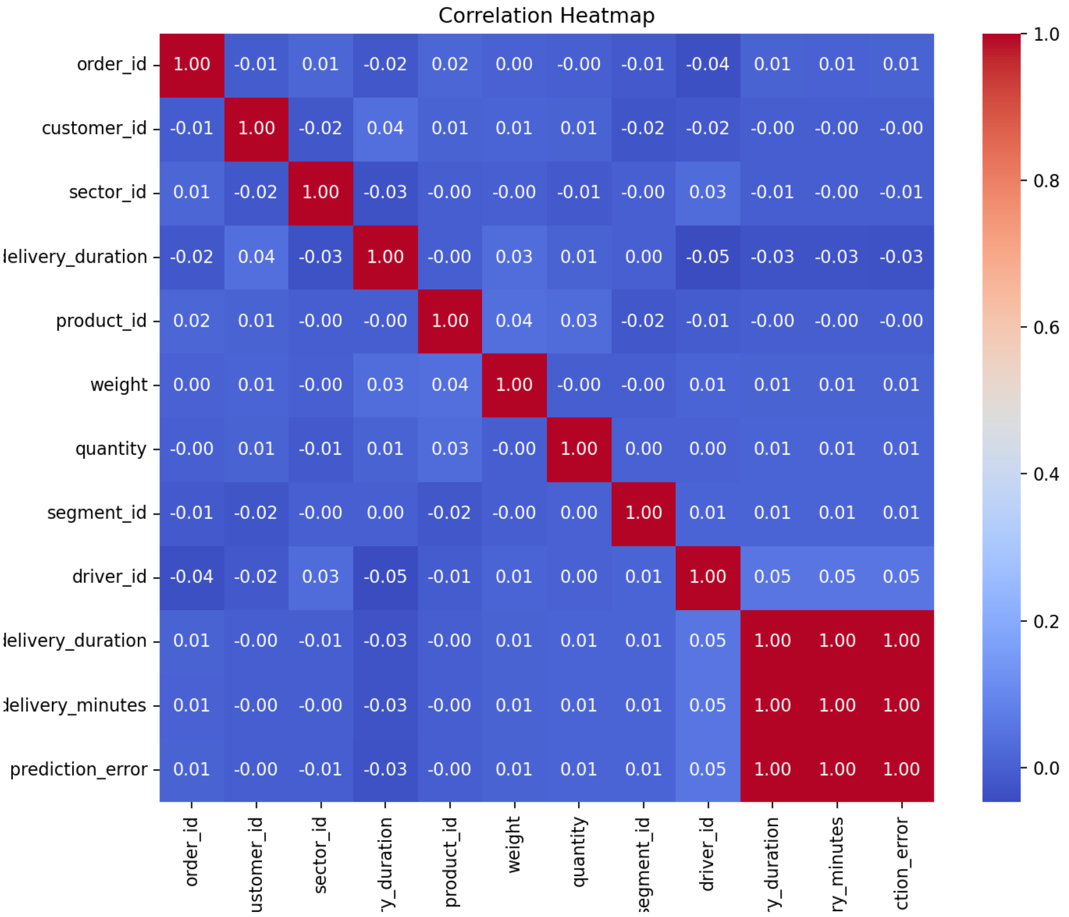
Firstly, we load the database from .csv file. As we can see, the obtained dataset presents every unique pair of product and order.

1.1 Data preprocessing

To improve the quality of our analysis and extract important features, we should preprocess the data. That will include transforming columns 'segment_start_time' and 'segment_end_time'. Even though the whole date may not be very helpful with predicting delivery duration, the hour of delivery seems to be important. As the segment_id is equal to "STOP" for every row and doesn't carry any important information, we can drop this column as well.

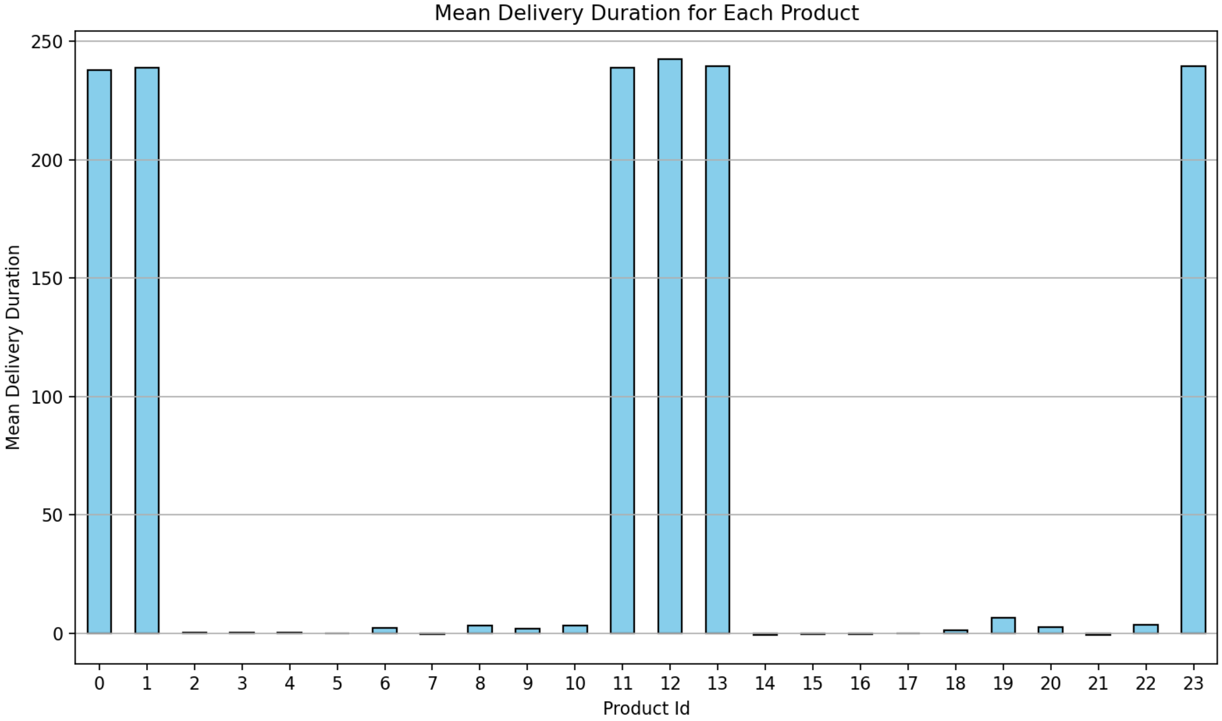
1.2 Features dependencies

To see if there are any vivid correlations in our data, we can create a heatmap for nummeric columns.



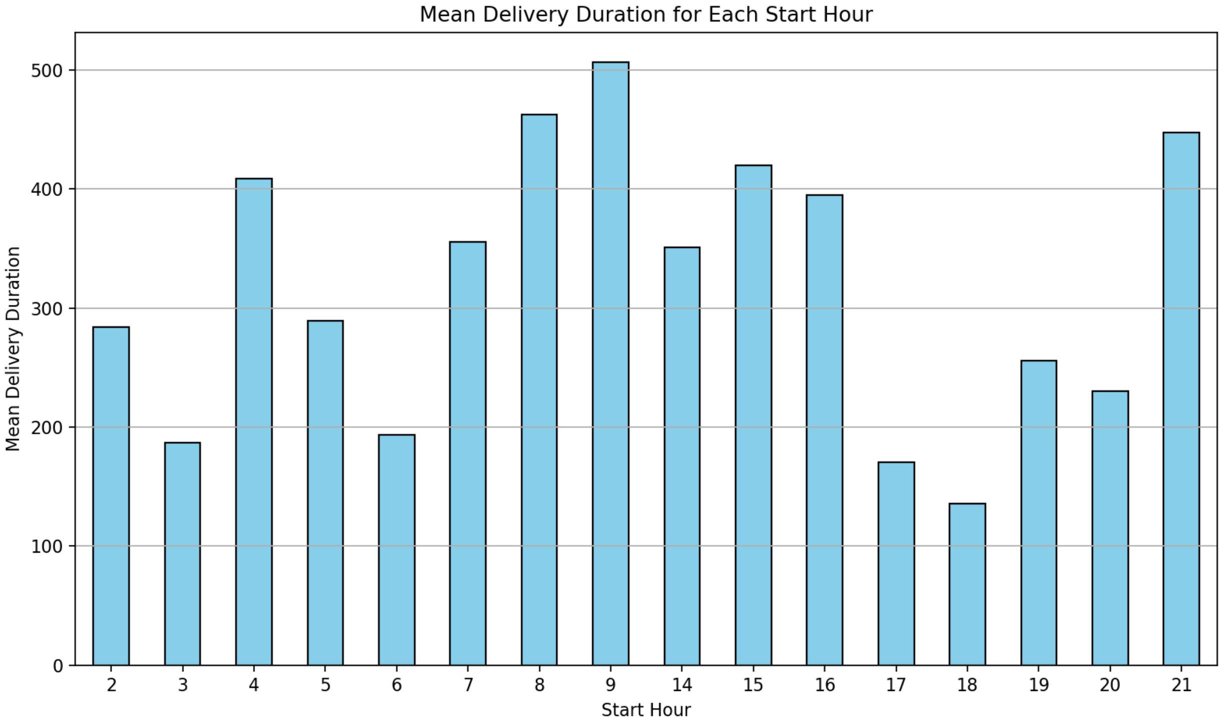
On the created correlation plot we can't see any important correlations. However, we need to bear in mind that the dependency may not be linear and simple correlation won't show it. Instead of correlation plot, we now will use grouping to see how each feature impacts delivery time.

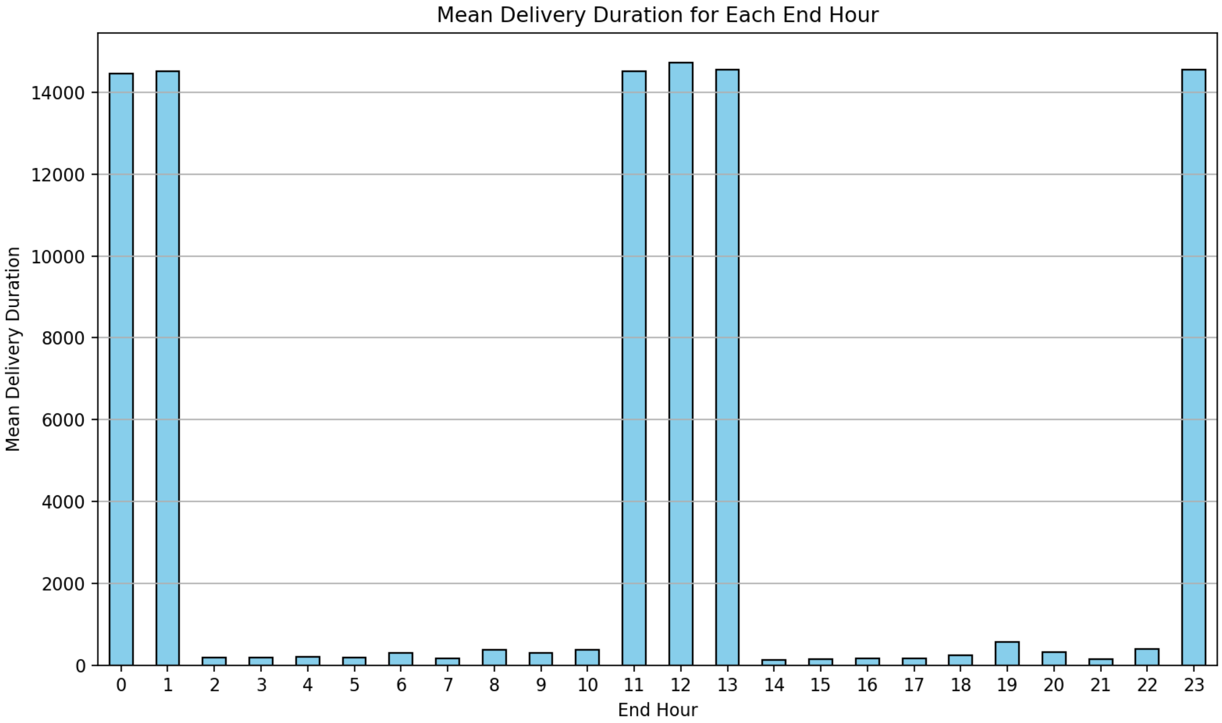
1.2.1 Product Id

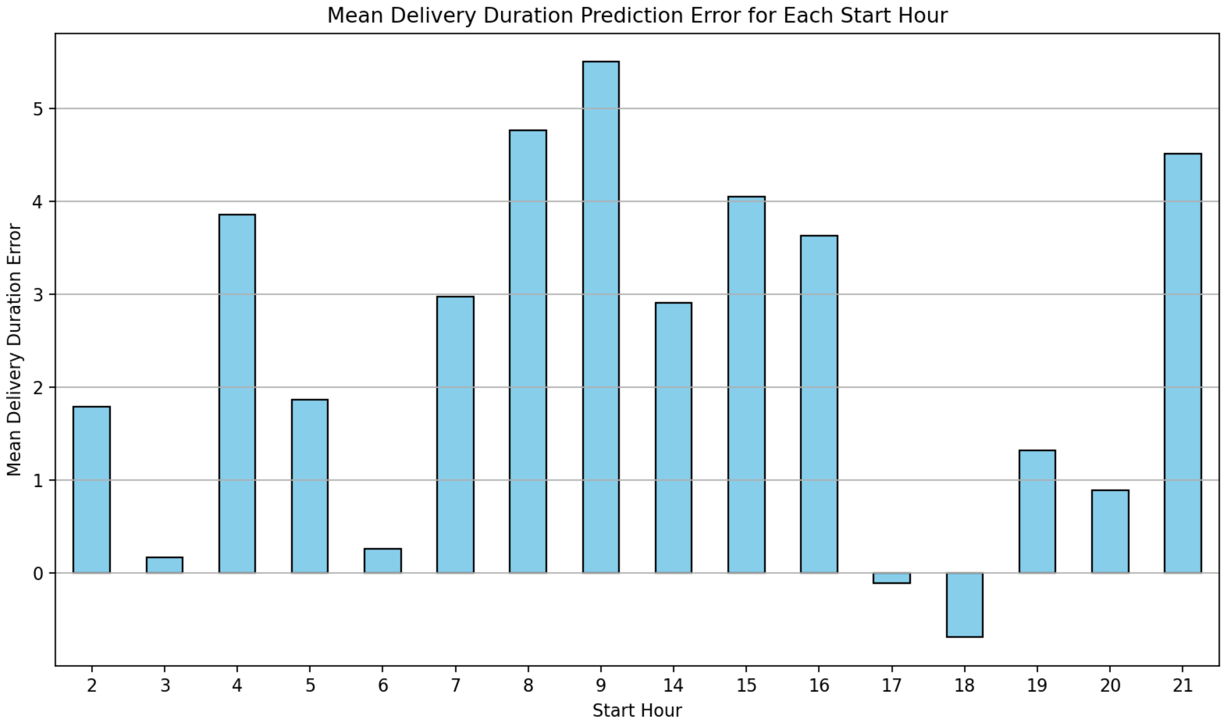


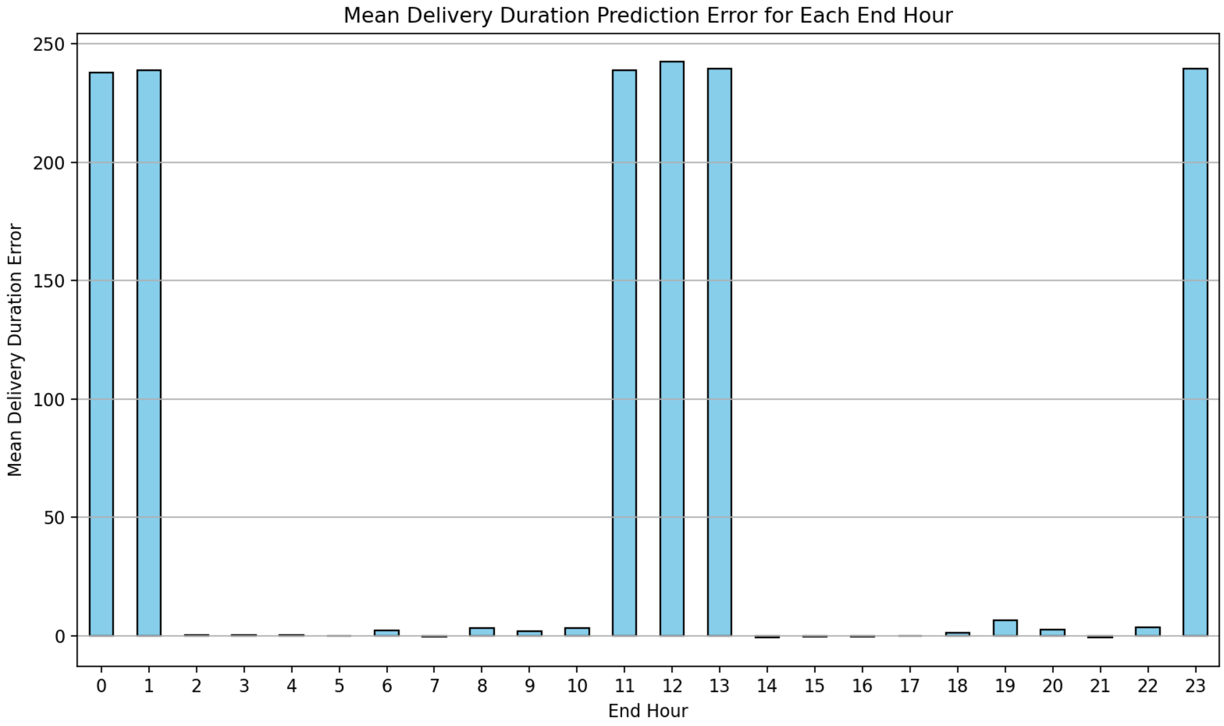
On the plot above we can see that the delivery of some products take much more time. The reason for that may be for example the size of the product. Type of product delivered should be taken into account while making predictions. Now, instead of considering every product included in order, we take into account only total weight of an order.

1.2.2 Start/End hour






Mean Delivery Duration Prediction Error for Each Start Hour
Mean Delivery Duration Error
0
1
2
3
4
5
2
3
4
5
6
7
8
9
14
15
16
17
18
19
20
21
Start Hour


Mean Delivery Duration Prediction Error for Each End Hour
Mean Delivery Duration Error
0
50
100
150
200
250
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
End Hour

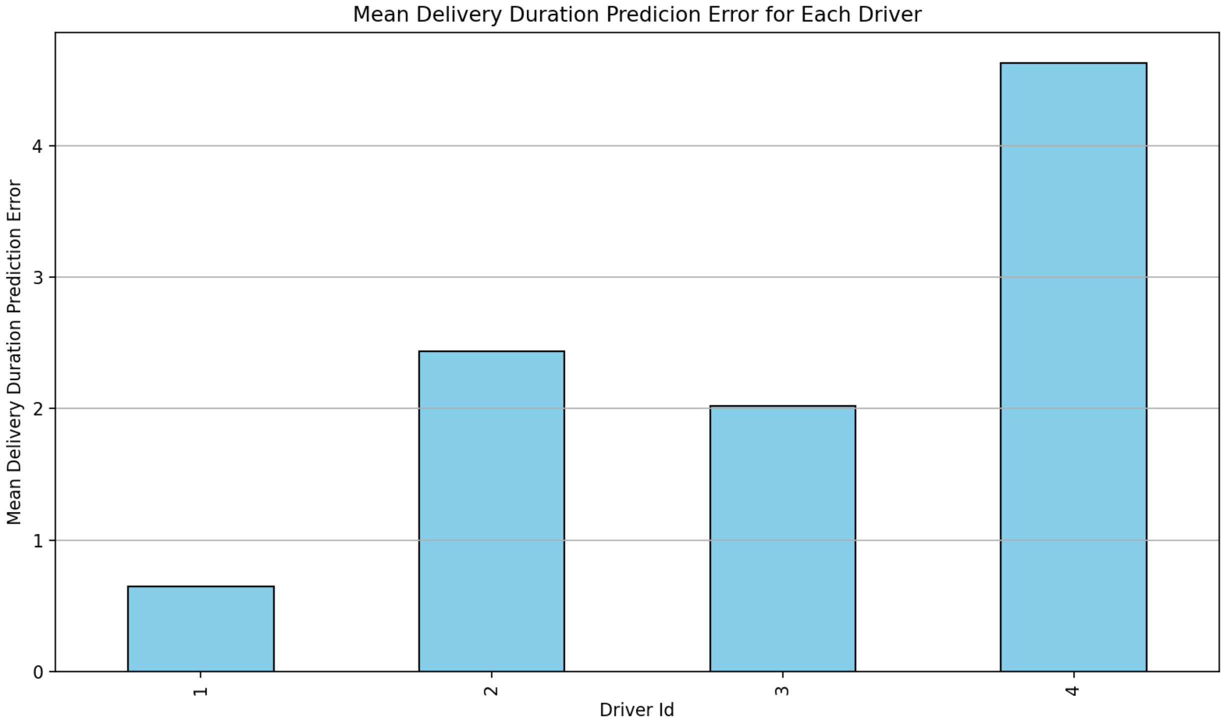
On the plots above we can see, that the deliveries that were started between 8 and 10 seem to be significantly longer than other ones. When it comes to end hours, orders delivered around midnight and midday take the most time. Moreover, also errors in predicted delivery times are the biggest in this hours. That leads to conclusion on this time of the day the deliveries are not only long, but also definitely longer than planned. The reason for that may be for example heavy traffic jam. This dependency should definitely be taken into account while planning delivery times.

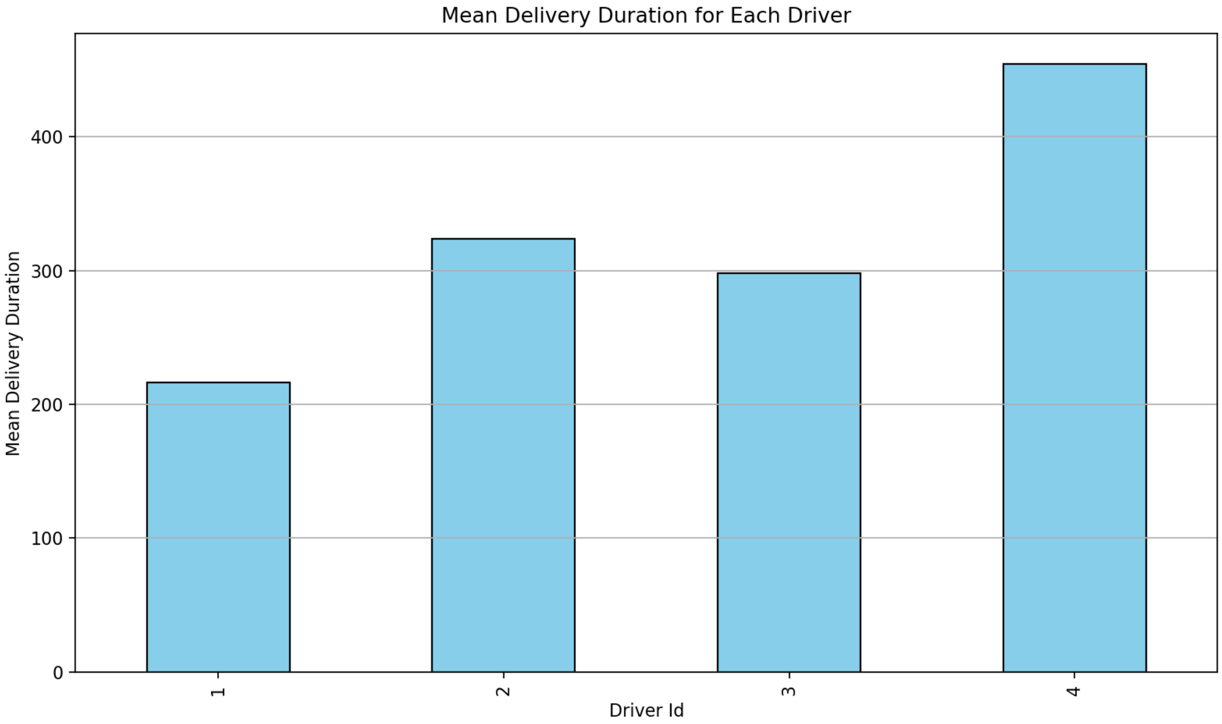
1.2.3 Segment and Driver

Another thing worth mentioning is the fact that some routes may take much more time than the others. The reason for that might be the length of the route, traffic jams and type of the road. Also drivers can impract the time of the delivery. Is there a driver that tends to drive slower? Let's visualise those hypotesis.

Firstly we can extract segments ids on which delivery usually lasts over 30 minutes. Then we extract the segments ids on which the error is usually bigger than 30 minutes. What is important is that both of these sets are exactly the same. Therefore, we can easily identify those more time consuming segments and take that into account while predictions. That approach will significantly improve our delivery times accuracy.

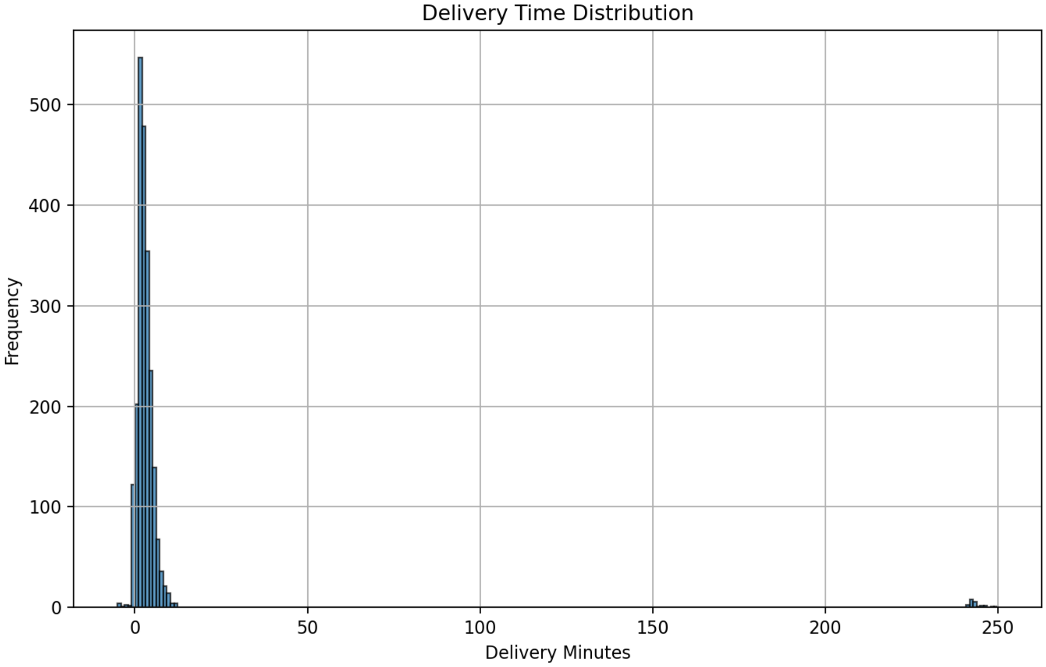
```
Index([ 400,  656,  949, 1055, 1169, 1772, 1778, 1907, 1919, 1964, 2192, 2248,
       2370, 2686, 2775, 2786, 3114, 3207, 3348, 3883, 4139, 4191, 4797, 4936],
      dtype='int64', name='segment_id')
```

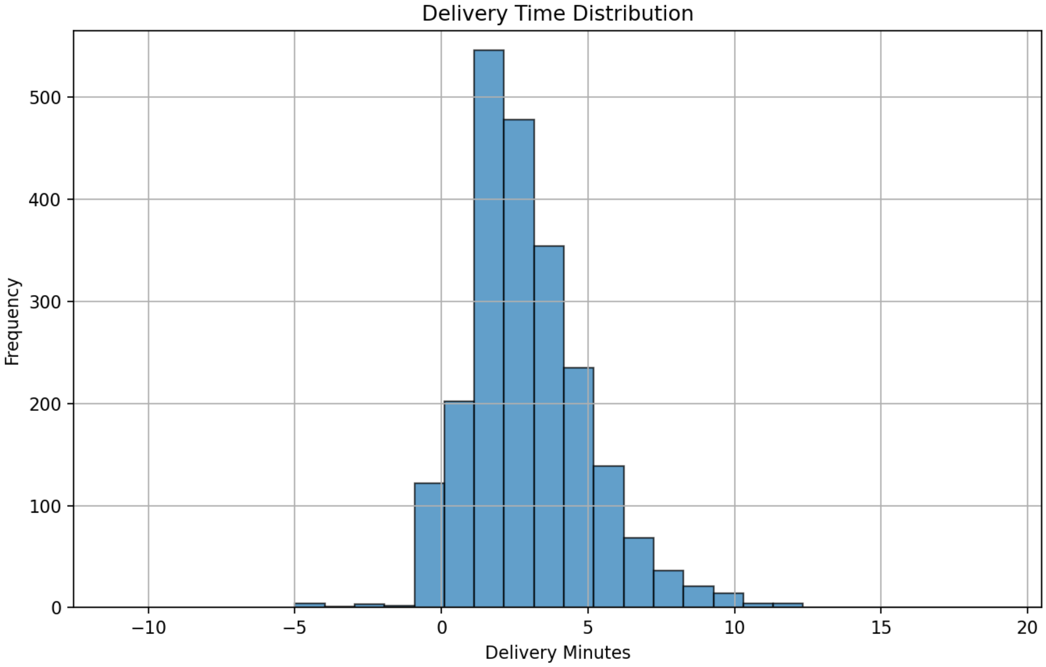


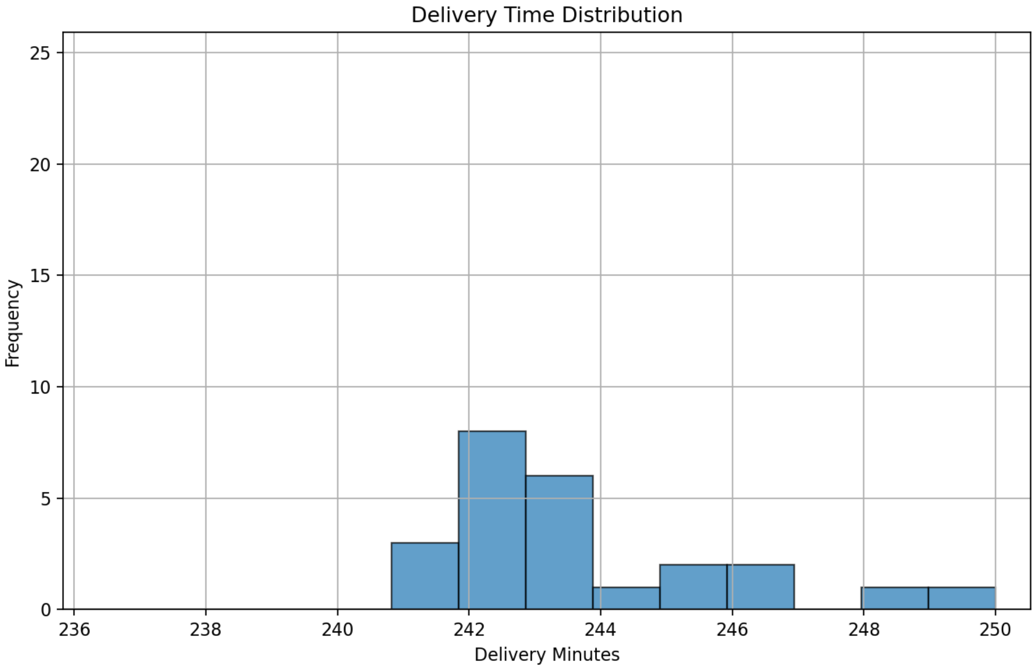


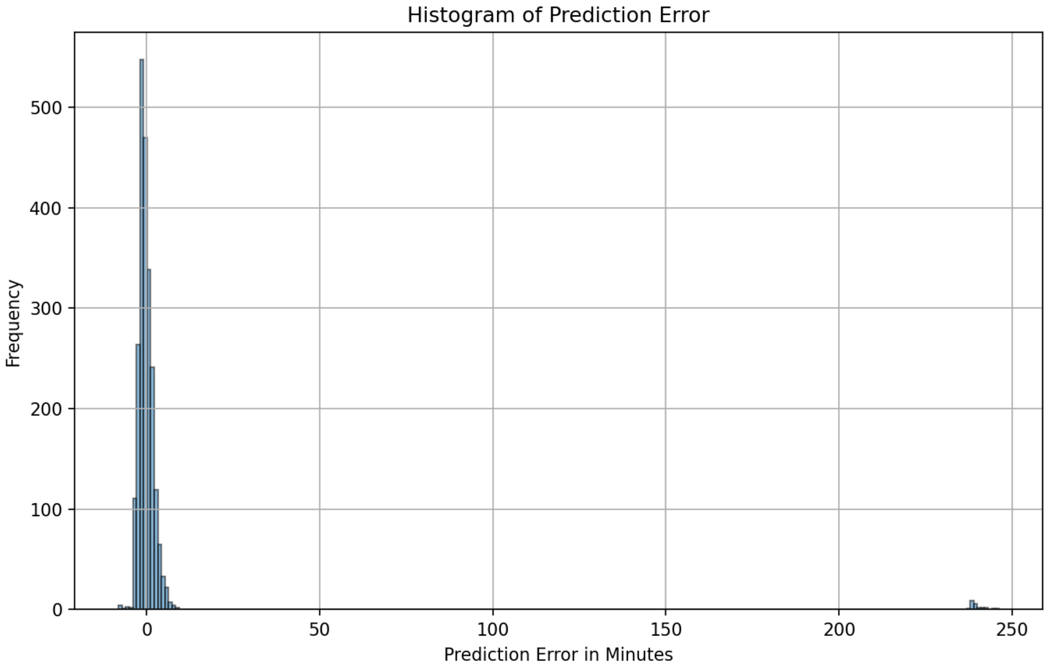
From the obtained data we can infer that not only routes but also drivers affect time of the delivery. Driver 4 seems to need much more time for his deliveries. Predictions for this driver are also frequently far from truth. Taking all of this into account, the the special time buffer should be provided for both mentioned segments and driver 4. Ideally, the predictions of delivery time should be done for each driver and segment individually.

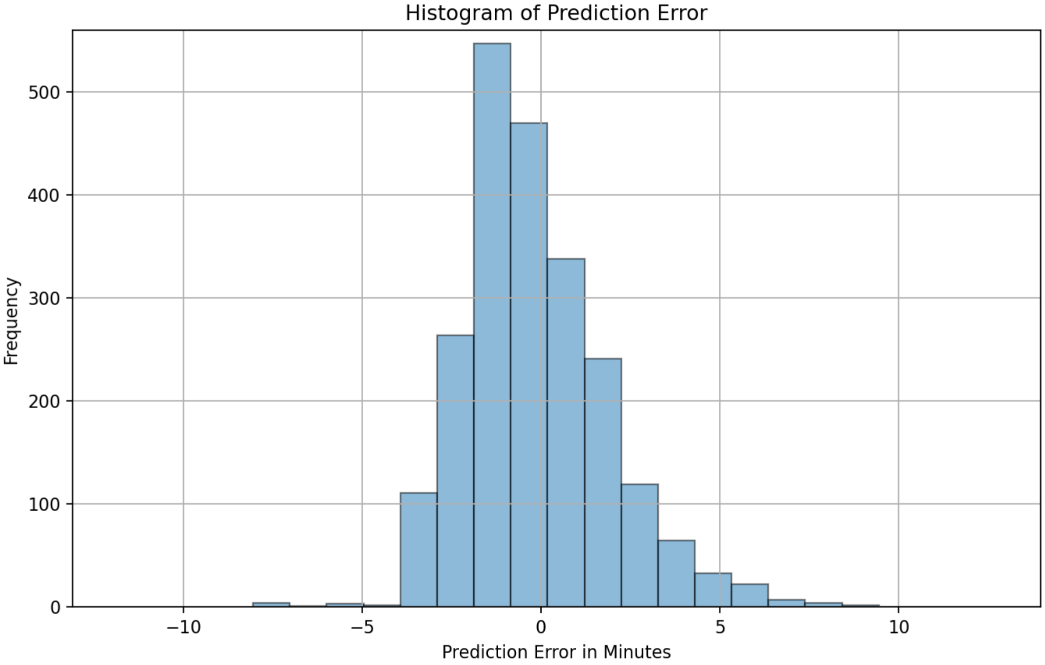
1.3 Delivery times distribution

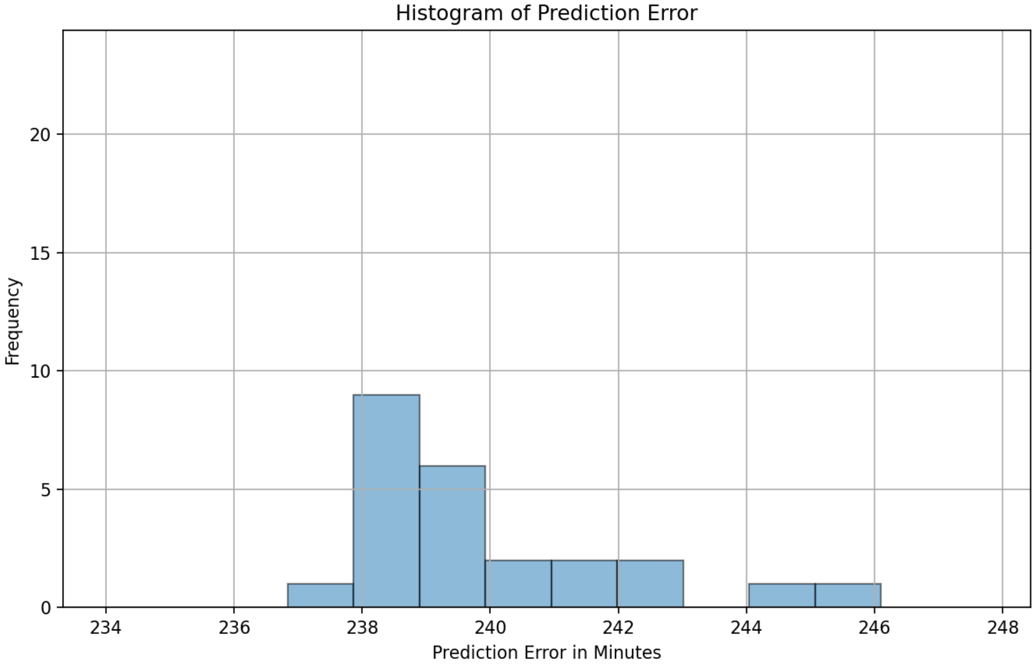





Delivery Time Distribution
Frequency
0
5
10
15
20
25
236
238
240
242
244
246
248
250
Delivery Minutes

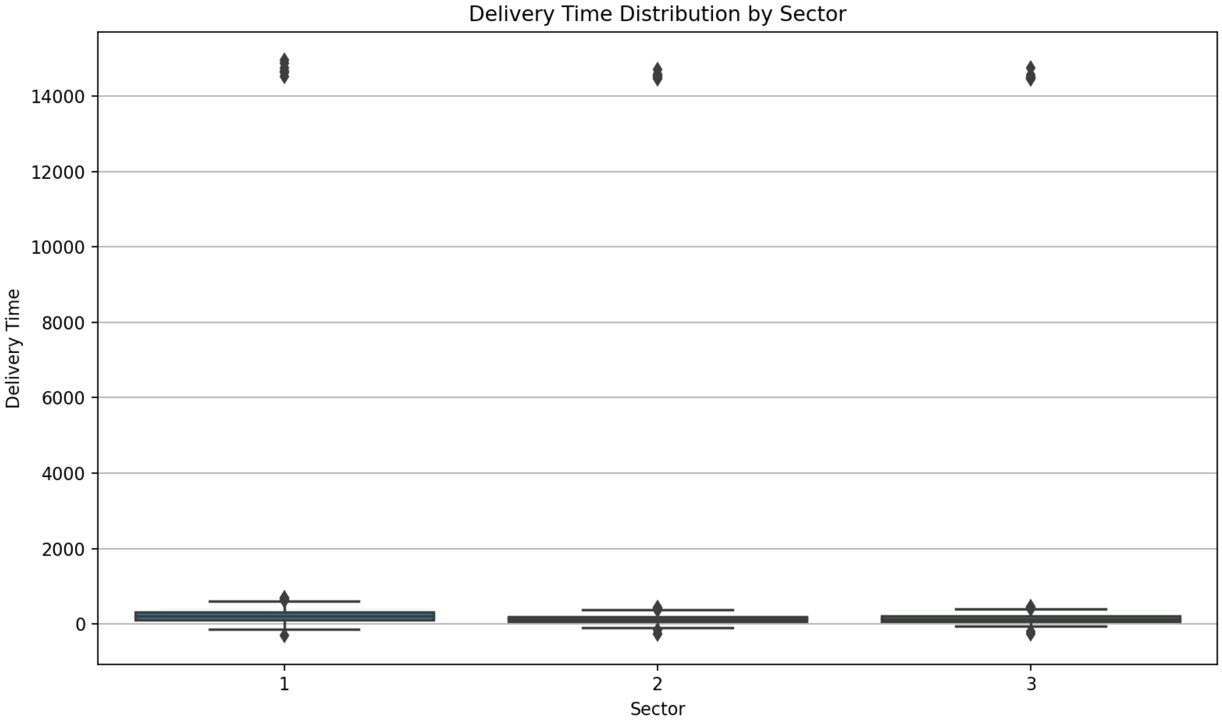

Histogram of Prediction Error
Frequency
0
100
200
300
400
500
0
50
100
150
200
250
Prediction Error in Minutes

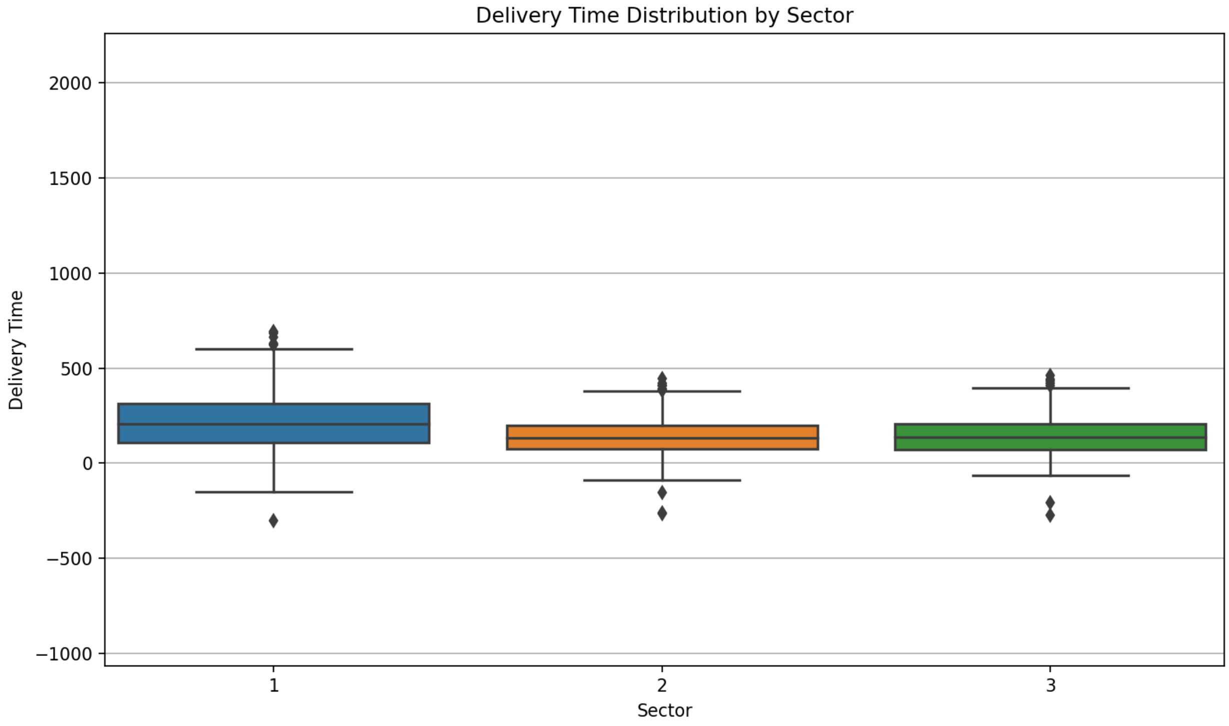

Histogram of Prediction Error
Frequency
0
100
200
300
400
500
−10
−5
0
5
10
Prediction Error in Minutes

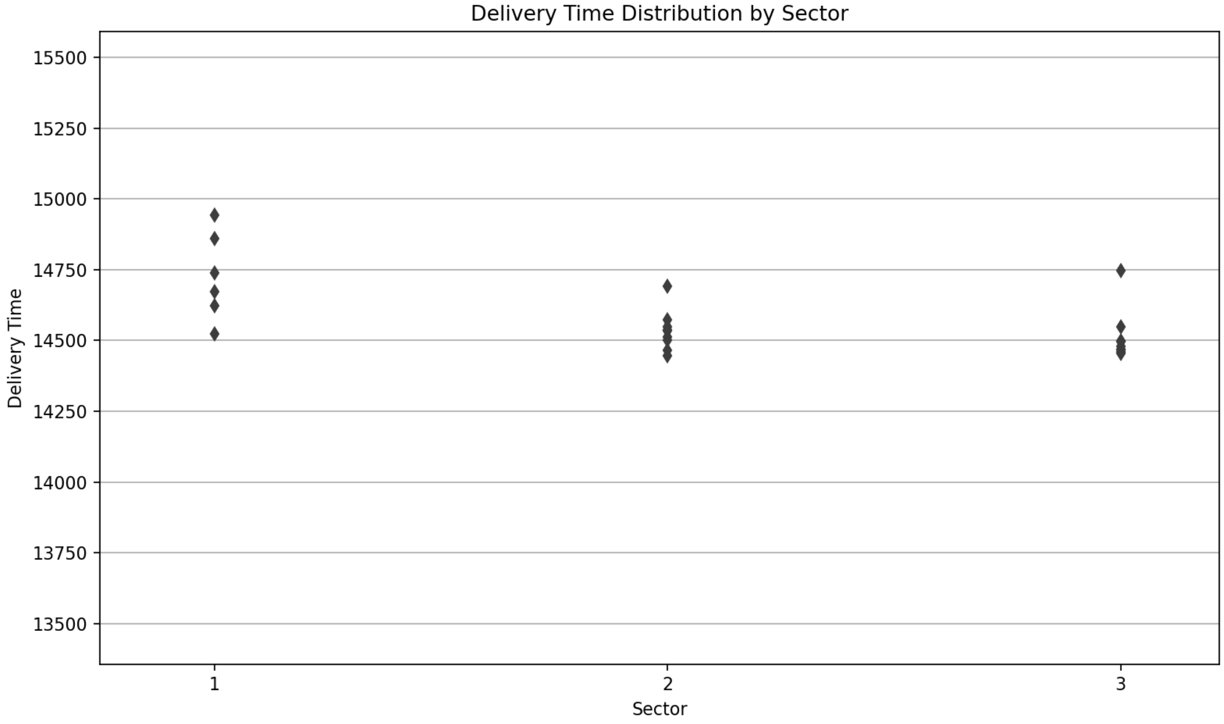

Histogram of Prediction Error
Frequency
0
5
10
15
20
234
236
238
240
242
244
246
248
Prediction Error in Minutes

From predictions errors distribution we can conclude that most of the predictions are slightly below 0. That means that the time predicted for the delivery was a little bit longer that the actual time, which is an optimal solution. However, about half of the predictions errors are positive - that means that the expected delivery time was shorter that the actual time. That may lead to dangerous situation when the minor delays overlap and create a huge delay. Therefore, it would be a good idea to slightly extend each predicted delivery time. Moreover, we can observe that the biggest errors happen for the longest deliveries. As each prediction is given as a mean of previous delivery times, it can't be suitable for routes that are simply much longer and therefore take more time. Again, specifying those long segments ids is crucial for accurate predictions.

1.4 Delivery time by sector




Delivery Time Distribution by Sector
Delivery Time
2000
1500
1000
500
0
−500
−1000
1
2
3
Sector


Delivery Time Distribution by Sector
Delivery Time
15500
15250
15000
14750
14500
14250
14000
13750
13500
1
2
3
Sector

In the obtained plots the difference in delivery times is not very vivid - the difference is about 2 minutes. However, when we take into account the very long deliveries (over 3 hours), we can see that most of them takes place in sector 3. Also the longest delivery happend in sector 3. Drivers bad experiences with unexpectedly long routes may have lead to false impression, that in sector 3 the delivery times are significantly longer. Finally, we should rather consider the segment id than the sector itself.