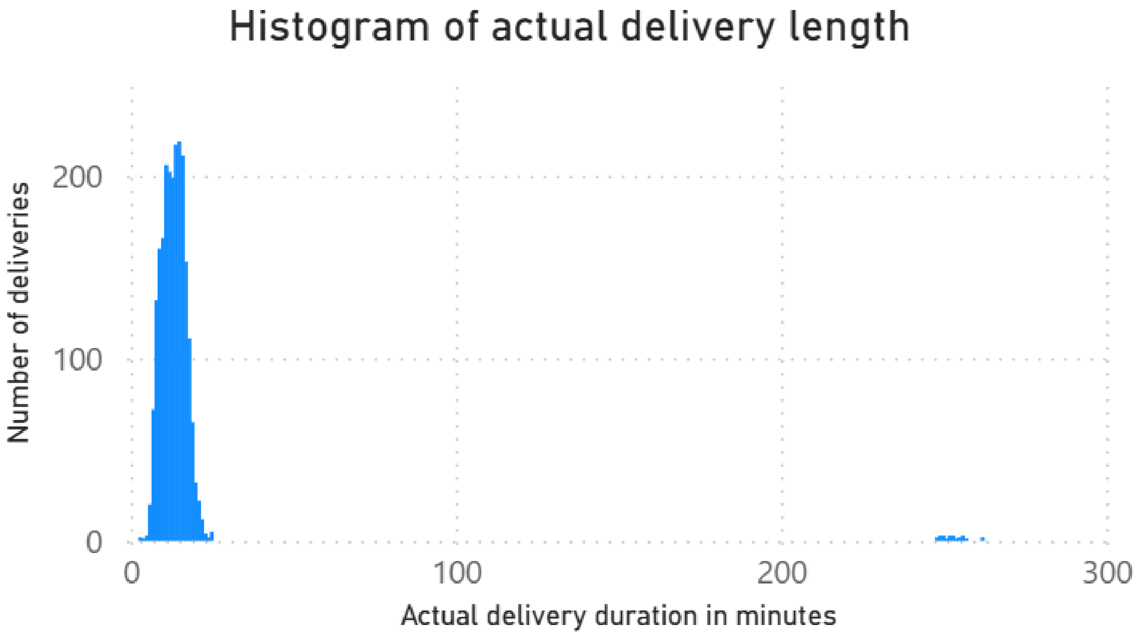
Analysis and visualisation based on delivery time data

1. Analysis of actual delivery length

The first step of the analysis was to generate a histogram showing the actual delivery length with 1 minute granularity.

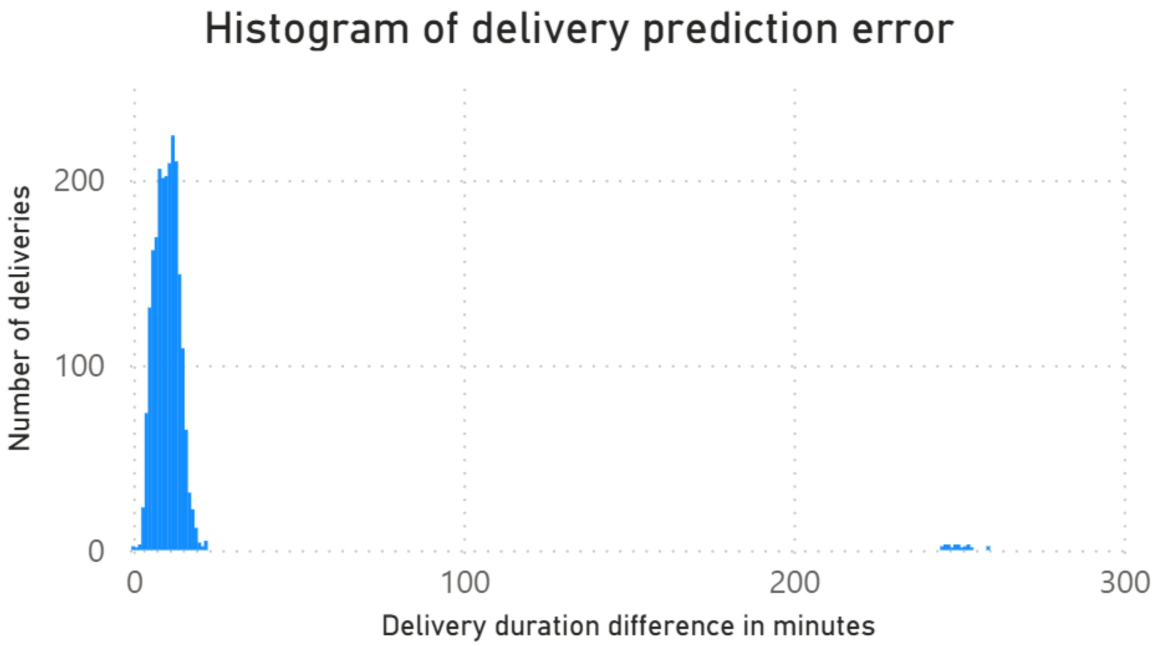
I started data modeling by filling missing values in the column contained Order IDs with values that had already appeared in delivered orders. In the next step, I found the earliest date of not delivered segments and the latest of delivered segments for every order individually. This allowed me to calculate actual delivery time. At the end, I counted how many times each delivery duration appeared and visualized the results on the histogram below.



On the histogram we can see that most deliveries took about 0-30 minutes. However, there are also a few of outliers with results of about 250 minutes. It could mean that some values are incorrect or there was exceptional delivery cases.

2. Analysis of prediction error

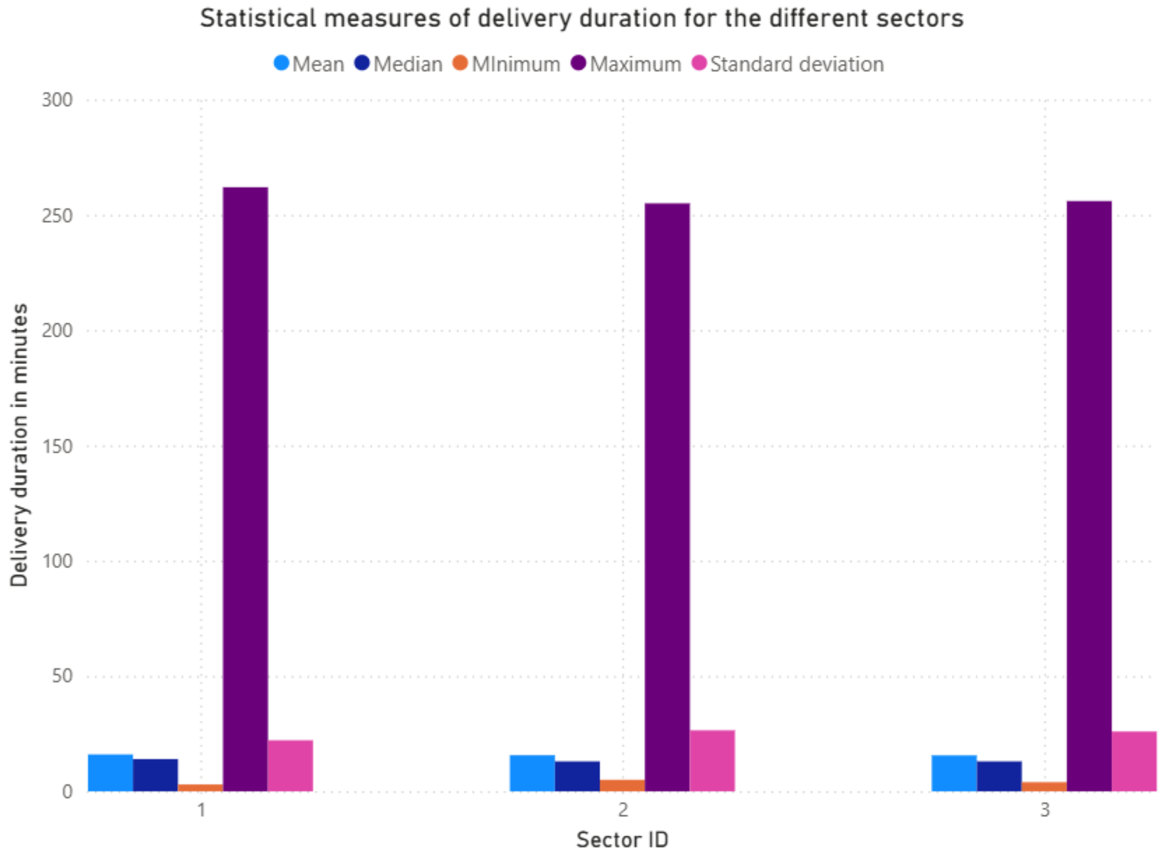
In this part of the analysis, at the start I used the data prepared in the previous step. Next, I converted the planned delivery duration from seconds to minutes. This allowed me calculate the difference between the actual and predicted delivery time. In the last step, I generated a histogram showing how many times error value appeared in the results.



As we can see, the distribution of the data looks similar to the first histogram. It might suggest that the analysis is correct. Histogram shows that most of the predictions were accurate or the prediction error was just a few minutes. However, here also some outliers and like earlier it could mean that some data is incorrect or there were some exceptional delivery cases.

3. Analysis of delivery duration in the different sectors

In case to visualize the hypothesis saying that delivering in one of the sectors takes significantly longer than in other sectors, I decided to use the actual delivery length calculated in the previous step. Next, I grouped the data by the sectors. I thought it could be valuable to use statistical measures to compare them. So I calculated the mean, median, minimum, maximum and standard deviation. Then I showed results on the chart below.



In my opinion, the visualization doesn't confirm the hypothesis. All of the measures are very similar for each sector. In every sector, there is a big difference between the minimum and maximum values, but the mean is around 15 minutes in all cases. The fact that the median is close to the means suggests that the extreme values don't have a big impact on the data distribution. Based on the results of this analysis we cannot confirm the insights given by the drivers.