Introduction

In this report, processing and analysis of delivery of grocery orders placed online was carried out to improve delivery time predictions. In order to perform this task, historical supply data stored in four separated tables (Figure 1) were used and maintained using MySQL. Preprocessing and all visualizations were created in RStudio environment and Excel.

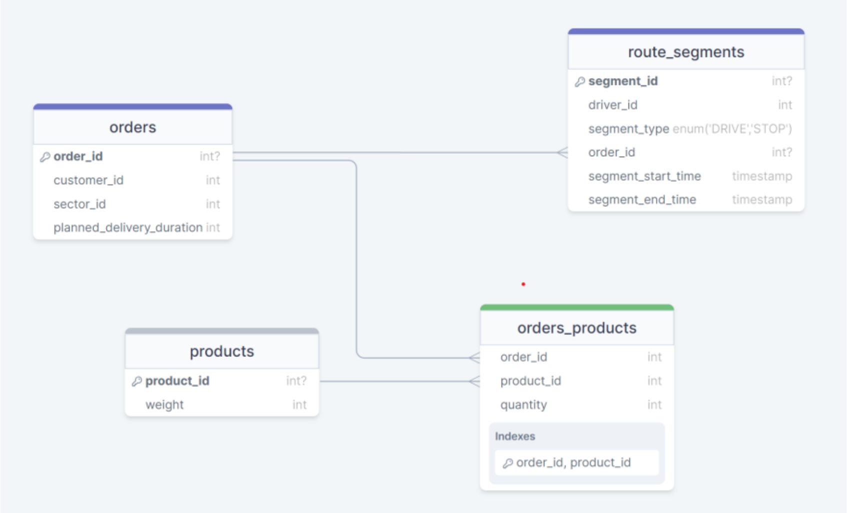


Figure 1. Data of historical deliveries in the system stored in 4 separate tables. Connected to each other using foreign keys (ex. Order_id, product_id)

Actual delivery length

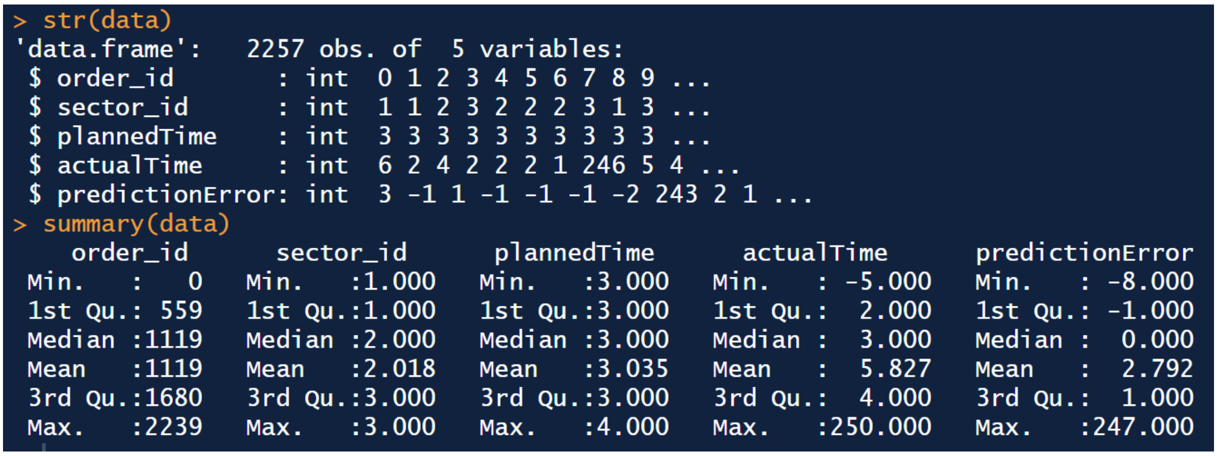
For the purpose of actual delivery time calculation, segament_start_time and segment_end_time data stored in table ‘route_segments’ were used. Only segment_type with “STOP” value was used for this analysis. It is due to the fact that only this type of segment contains information that are useful for the delivery time calculation. However, it was noticed that for some filtered lines order_id value is missing and therefore it was decided to remove such data from further analysis as they were not completed or damaged and could potentially return misleading results. Furthermore, it was identified that for some of the records start and end timestamps were incorrectly populated which result in negative value of actual deliver length, which is obviously incorrect. The first mentioned adjustment to the data was made directly in MySQL while the second one, related to negative actual delivery time, later in RStudio.

The below presented table (Figure 2), was created to store information about order id, sector id, planned and actual delivery time, both in minutes rounded upwards together with prediction error which was difference between actual and planned delivery times. By doing so, preprocessing and visualization could be done.

order_id	sector_id	plannedTime	actualTime	predictionError
0	1	3	6	3
1	1	3	2	-1
2	2	3	4	1
3	3	3	2	-1

Figure 2. Sample lines from table.

At first, the structure and statistics of the descriptive data had to be checked.



```
> str(data)
'data.frame':   2257 obs. of  5 variables:
 $ order_id       : int  0 1 2 3 4 5 6 7 8 9 ...
 $ sector_id      : int  1 1 2 3 2 2 2 3 1 3 ...
 $ plannedTime    : int  3 3 3 3 3 3 3 3 3 3 ...
 $ actualTime     : int  6 2 4 2 2 2 1 246 5 4 ...
 $ predictionError: int  3 -1 1 -1 -1 -1 -2 243 2 1 ...
> summary(data)
    order_id      sector_id      plannedTime      actualTime      predictionError
 Min.   :   0   Min.   :1.000   Min.   :3.000   Min.   : -5.000   Min.   : -8.000
 1st Qu.: 559   1st Qu.:1.000   1st Qu.:3.000   1st Qu.:  2.000   1st Qu.: -1.000
 Median :1119   Median :2.000   Median :3.000   Median :  3.000   Median :  0.000
 Mean   :1119   Mean   :2.018   Mean   :3.035   Mean   :  5.827   Mean   :  2.792
 3rd Qu.:1680   3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:  4.000   3rd Qu.:  1.000
 Max.   :2239   Max.   :3.000   Max.   :4.000   Max.   :250.000   Max.   :247.000
```

The variables are numerical and count 2257 observations each. In case of actual time the minimum and maximum values deviate significantly from the mean or median, and the mean is greater than the median and three quartiles, which may indicate the presence of extreme values. To confirm these hypotheses, it was decided to create a boxplot (Figure 3) and histogram (Figure 4).

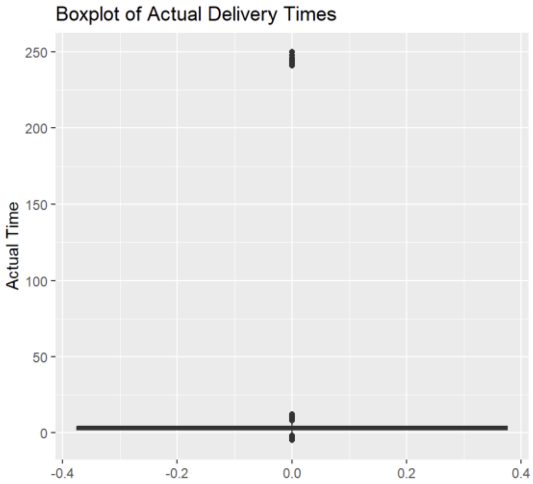


Figure 3. Boxplot of actual delivery times before preprocessing.

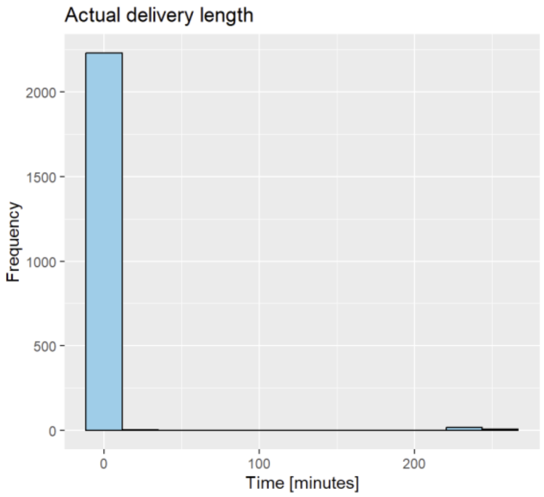


Figure 4. Histogram of actual delivery length before preprocessing.

Based on the box plot, it can be identified that the vast majority of values fall within a very small range, but single, strongly outlier observations contribute to the large scatter in the data. The distribution of the data on the histogram is disturbed - values in the 0-10 range (the beginning of the scale), were recorded more than 2,000 times, representing significant part of observations. In addition to this one high bar, two more very small ones can be seen, which coincide with the single outliers in the boxplot. Therefore, it was decided to replace the extreme values which are the values above 200 and as mentioned before those below 1 because time cannot be negative or zero. Moreover, coarse errors detected by the three sigma test were replaced. By doing this, the boxplot (Figure 5) and histogram (Figure 6) were repeated.

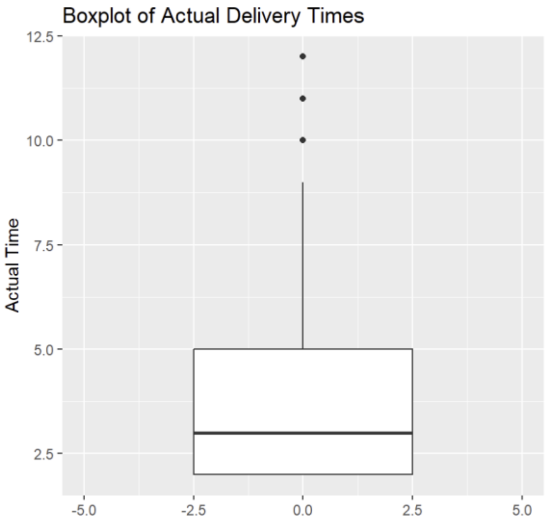


Figure 5. Boxplot of actual delivery times after preprocessing.

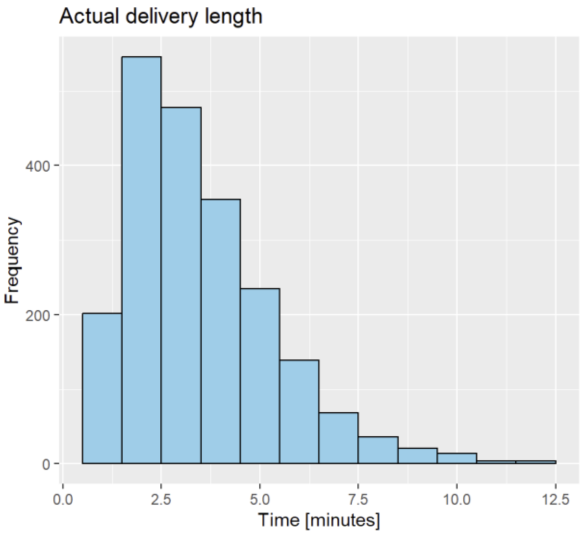
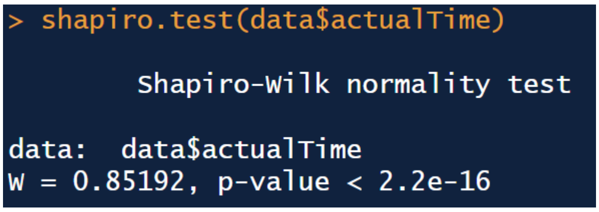


Figure 6. Histogram of actual delivery length after preprocessing.

After preprocessing, the scatter of the data decreased significantly. The 3 stronger outliers relative to the others above the upper whisker have been isolated, so that the distribution may manifest right-handedness. The median is a bit below the middle of the box. Histogram manifests right-handed asymmetry, and based on the Shapiro-Wilk test, the hypothesis of a normal distribution of the studied trait can also be rejected (significance level p-value=0.05).



```
> shapiro.test(data$actualTime)

        Shapiro-Wilk normality test

data:  data$actualTime
W = 0.85192, p-value < 2.2e-16
```

Observing the histogram, one can see that most deliveries were made within 1 minute, and the longer delivery time is, the smaller number of orders is.

Prediction error

In the Figure 2 table presented above, the last column was dedicated to reflect difference between actual and planned delivery time. What is worth to mention, calculated numbers were not converted to present absolute value of error, as the actual sign of difference can have significant impact on further analysis. Once again the structure and statistics of the descriptive data had to be checked and it also turned out that the minimum and maximum values od prediction error deviate significantly from the mean or median, and the mean is greater than the median and three quartiles, which may indicate the presence of extreme values. For this reason, boxplot (Figure 7) and histogram (Figure 8) were made.

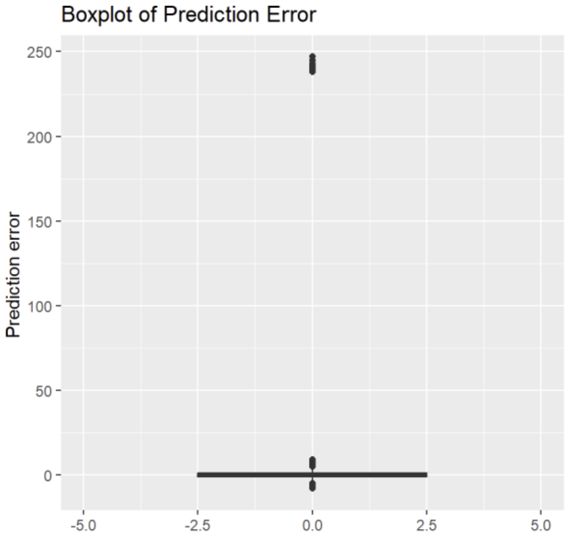


Figure 7. Boxplot of prediction error before preprocessing.

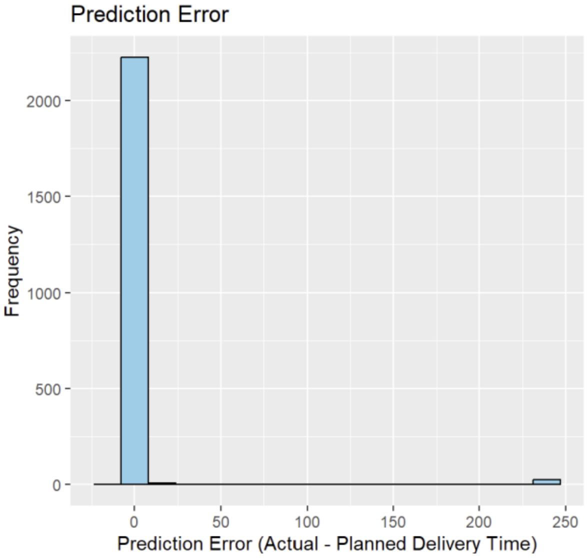


Figure 8. Histogram of prediction error before preprocessing.

Just like in a Figure 3 and Figure 4 the overhelming majority of values cluster tightly within a narrow range, yet individual extreme outliers significantly contribute to the broad dispersion observed in the data. On the histogram, values in the -5-5 range (the beginning of the scale), were recorded more than 2,000 times, definitely dominate. Also, one very small column could be seen by the end of plot. On balance, it was decided to replace the extreme values which are the values above 200 and coarse errors detected by the three sigma test. After the preprocessing, the boxplot (Figure 9) and histogram (Figure 10) were created once again.

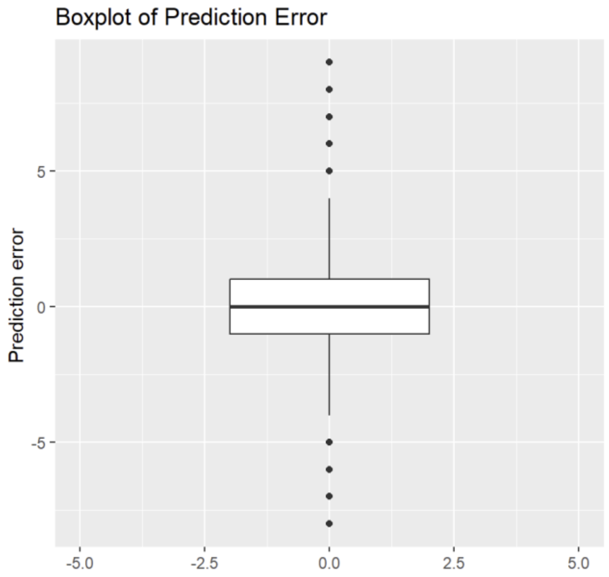


Figure 9. Boxplot of prediction error after preprocessing.

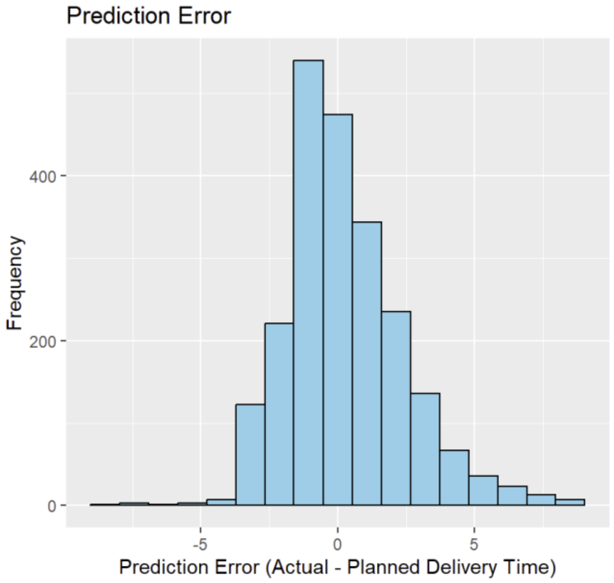
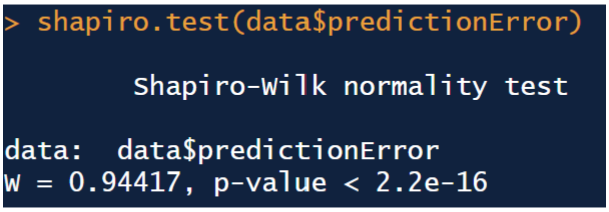


Figure 10. Histogram of prediction error after preprocessing.

Now the scatter in the data has narrowed considerably. Five outliers above the upper whisker and four below the lower whisker have been isolated, making the distribution appear normal. The median in this case is in the middle of the box. According to the Shapiro-Wilk test, the hypothesis of a normal distribution is almost fulfilled (significance level p-value=0.05).



```
> shapiro.test(data$predictionError)

        Shapiro-Wilk normality test

data:  data$predictionError
W = 0.94417, p-value < 2.2e-16
```

Based on the above charts, once can say that although the prediction error seems to be definitely more normalised than the before analysed actual delivery time, extreme values are definitely more common for the positive error, which means that more often actual delivery time was significantly longer than the planned one, compared to the opposite scenario. However, it is important to notice that exactly half of deliveries were successfully done before planned time.

Delivers in the sectors

In order to visualize the difference in delivery time between sectors (Figure 11), (Figure 12) the average for each was calculated separately. The average was: 4.44 minutes for sector

1, 3.02 minutes for sector 2 and 3.04 minutes for sector 3. The longest averaged time to deliver the product turned out in the first sector. The second and the third sector were characterized by similar timing. The average difference in actual time of delivery could be caused by type of buildings characteristics for sector, weather conditions, absence of the client or code-operated gates. Median for sector 1 was 4 minutes and for both sector 2 and 3 was 3 minutes which can be seen on a box plot.

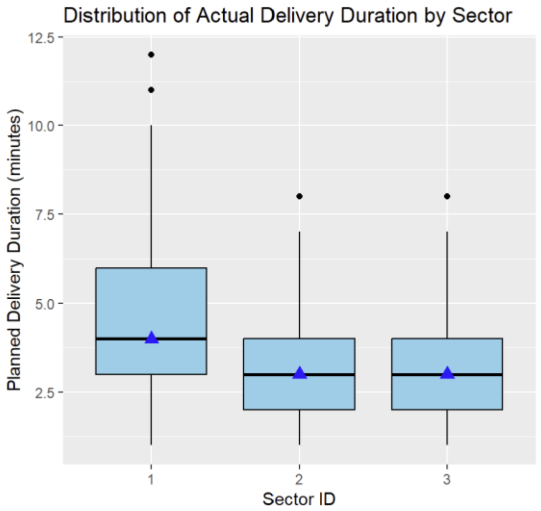


Figure 11. Boxplots of actual delivery by sector.

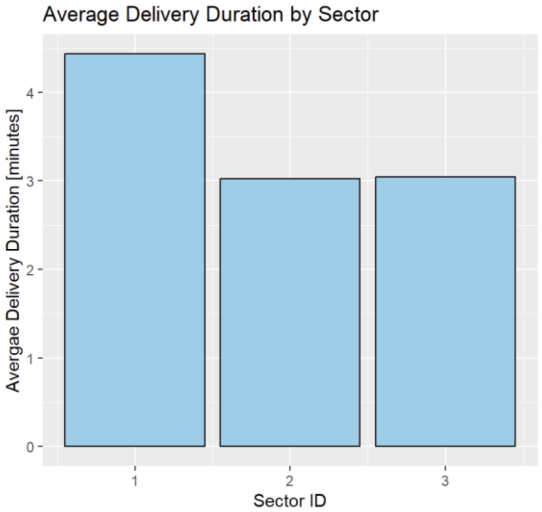


Figure 12. Chart of actual delivery by sector.

Trends and correlations

As part of the analysis of the dependence of the actual delivery time on various parameters, those elements that were assessed as potentially significant were designated first. They included the total weight of the order, the number of products in the order (product id multiplied by quantity for each order id), and the performance of the driver.

As can be seen in a chart below (Figure 13) the dependence of delivery time on the number of products in an order increase. It can be observed that for the range of the number of products between 1-14 there is a fairly steady increase in the average delivery time. Values for 15 and more products are single orders, so they should not be taken into account. Therefore, you can see the relationship between the number of items in an order and their delivery time.

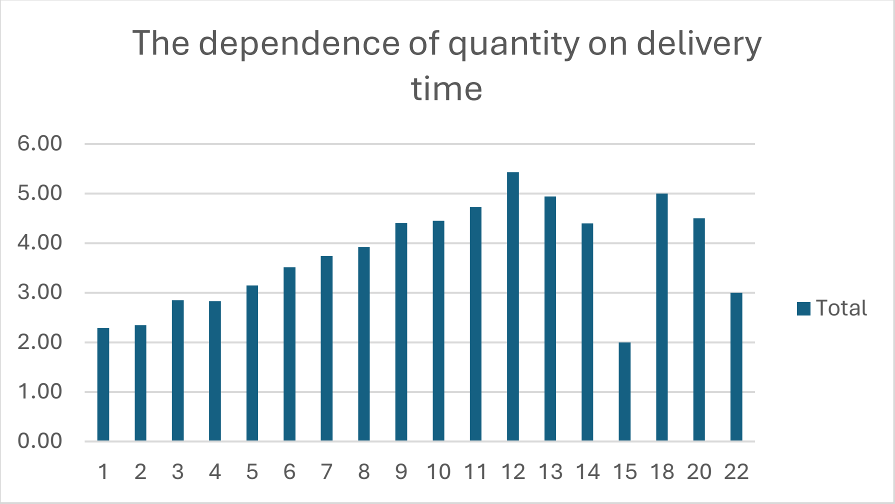


Figure 13. The dependence of quantity on actual delivery times.

Another analyzed variable was the weight of the order (Figure 14). The mindset in this case was quite simple - the greater the weight of the product, the longer the courier needs to deliver the order to the door. As can be seen in the chart below, the length of delivery time increased with the total weight of the order.

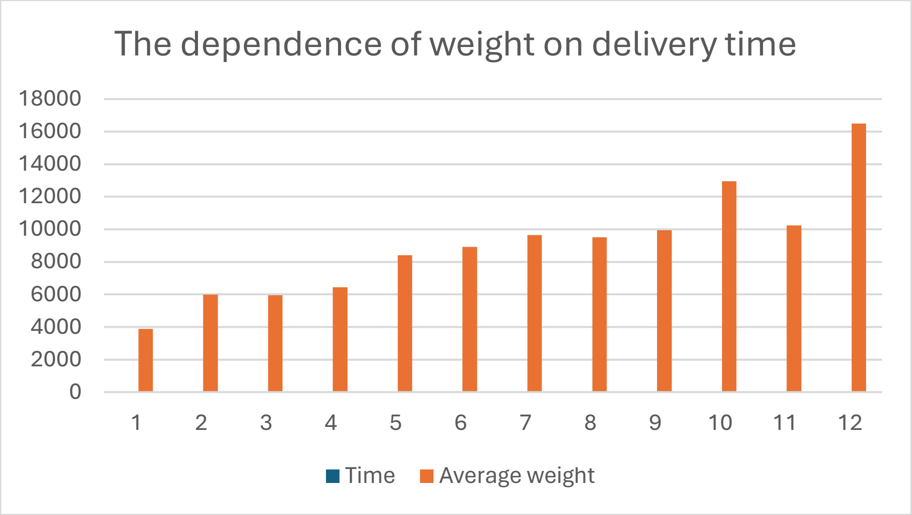


Figure 14. The dependence of quantity on actual delivery times.

An additional element that proved worthy of verification is the pereformance of individual drivers (Figure 15). At first, the average time each driver takes to deliver a product was analyzed. Then, the sector parameter was also added, in order to more objectively assess the efficiency of drivers in the same environment - after all, it may turn out that a given driver drives mainly in that sector, which by its nature implies a longer delivery time. As you can see in the chart below, driver number one has the best performance regardless of the sector. In addition, drivers 2,3 and 4, respectively, have increasingly worse performance.

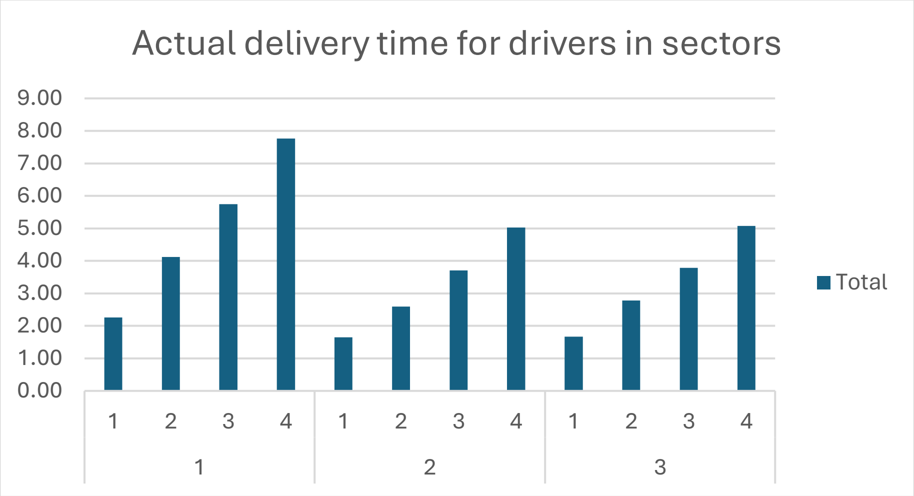


Figure 15. Actual delivery time for each driver in sectors.

Conclusion

This report examines true delivery time and analyzes the factors affecting it. Focusing on relevant segments and removing incomplete data, correlations between delivery time and variables such as order size, weight and driver performance were identified. Additionally, prediction error was measured and considered what it might result from. These insights provide valuable hints for optimizing delivery processes and ultimately improving customer satisfaction. Going forward, further research can improve predictive models and further improve delivery performance.