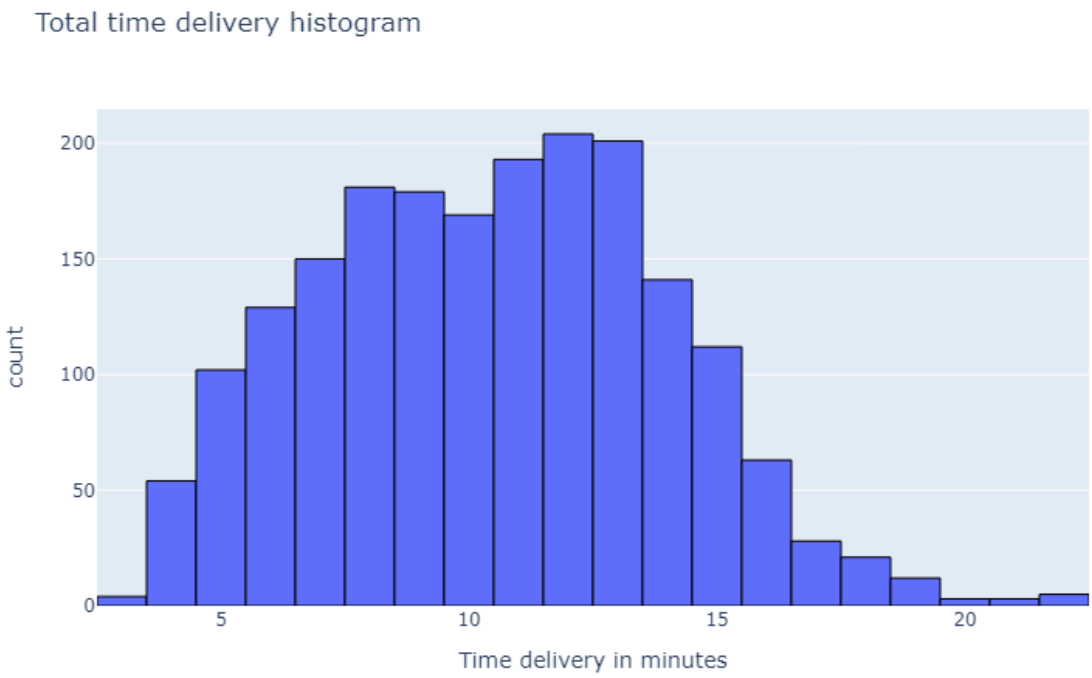
Data analysis and visualization

Szczepan Polak

Actual delivery histogram



The histogram indicates that delivery times are in the range of 3 to 22 minutes. However, the majority are delivered within several minutes. With a median around 10.

The preparation of the histogram can be divided into several parts:

- Extracting the data from the database
- Cleaning the data
- Creation of the histogram

The data was extracted using an SQL query, from the route_segments table. In addition, the times of each segment were counted as part of the data extraction.

The next step was to understand the data, I recognized that one delivery consists of a run and a stop time segments.

The data appeared to be contaminated. So I decided to remove the questionable records:

- Duplicates

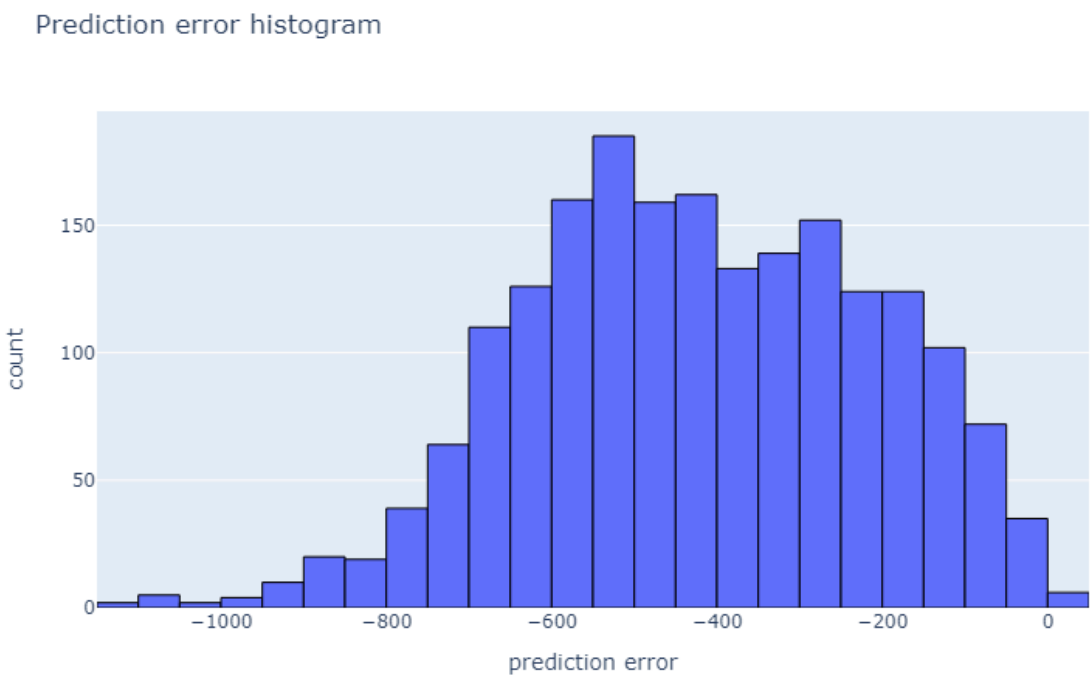
- outliers (journey times less than or equal to 0 or significantly large)

The next step was to calculate the total delivery times stop time + drive time.The problem in this case was the lack of an order id in the run times.

To address this, I created a pipeline in Python that checked that the drive + stop sequence matched before counting the time. I have assumed that if the previous record has an id different from one and the segment type is the opposite and the driver's id is the same then it is a single delivery sequence.

To create the histogram, I used the plotly library and divided the previously calculated run times by 60 to get the minute results.

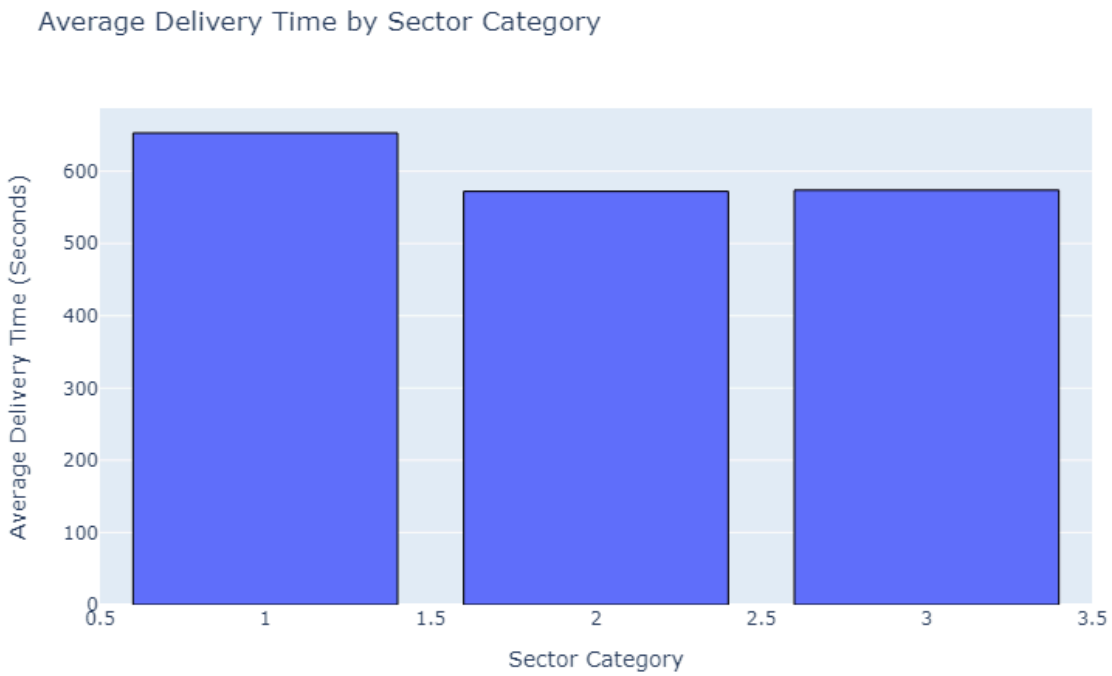
Prediction error histogram



The histogram shows that the vast majority of predictions are underestimated. Real delivery times are much longer. Median around 400 s.

To count prediction errors. I extracted data from a database containing predictions of travel times for deliveries. And I merged them with the previous table using the order ID. So that for existing orders, the corresponding predictions are assigned.

Sectors hypothesis



The hypothesis is confirmed. Sector 1 has, on average, longer delivery times than the other sectors.

To test the hypothesis, I grouped the previously prepared data by sector_id and then counted the average delivery times for each category.

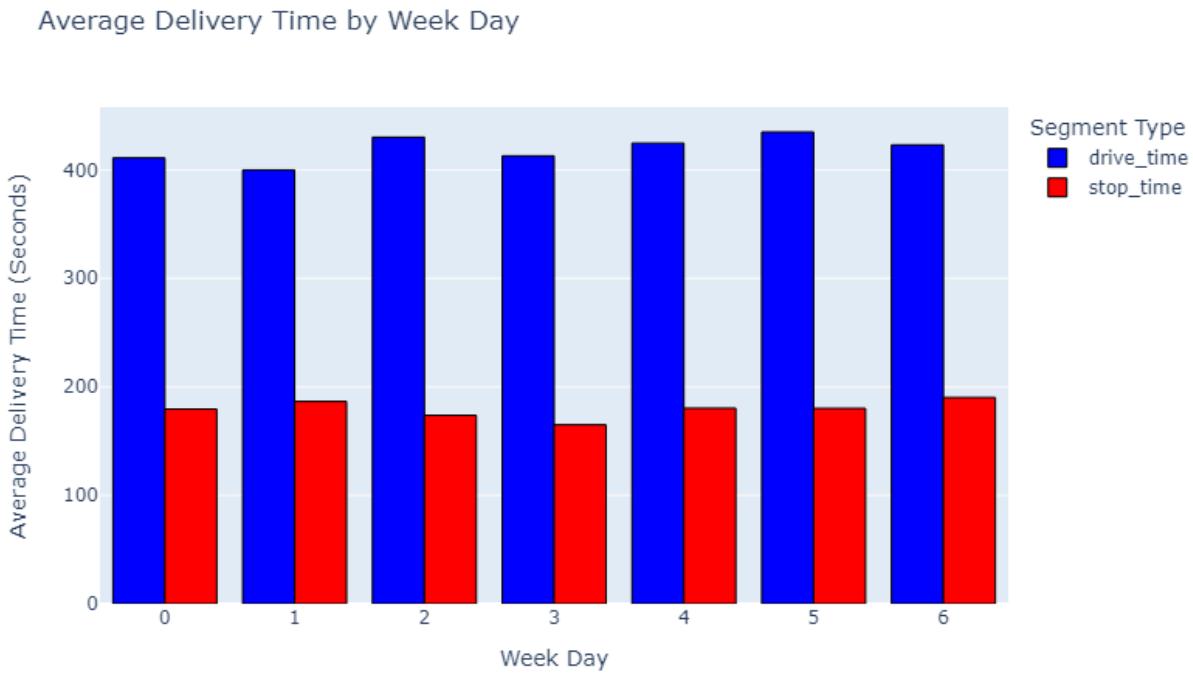
Play with the data

I decided to test a few scenarios.

- What does the delivery time look like on different days of the week?
- What does the delivery time look like at different times of the day?
- Is there a relationship between delivery time and the weight or quantity of products in an order?
- Do drivers have similar delivery times?
- Driving time vs. stopping time

To do this, I added some new columns by pulling them from the database (total quantity and total weight of each order), and created new features based on time (day of the week, hour)

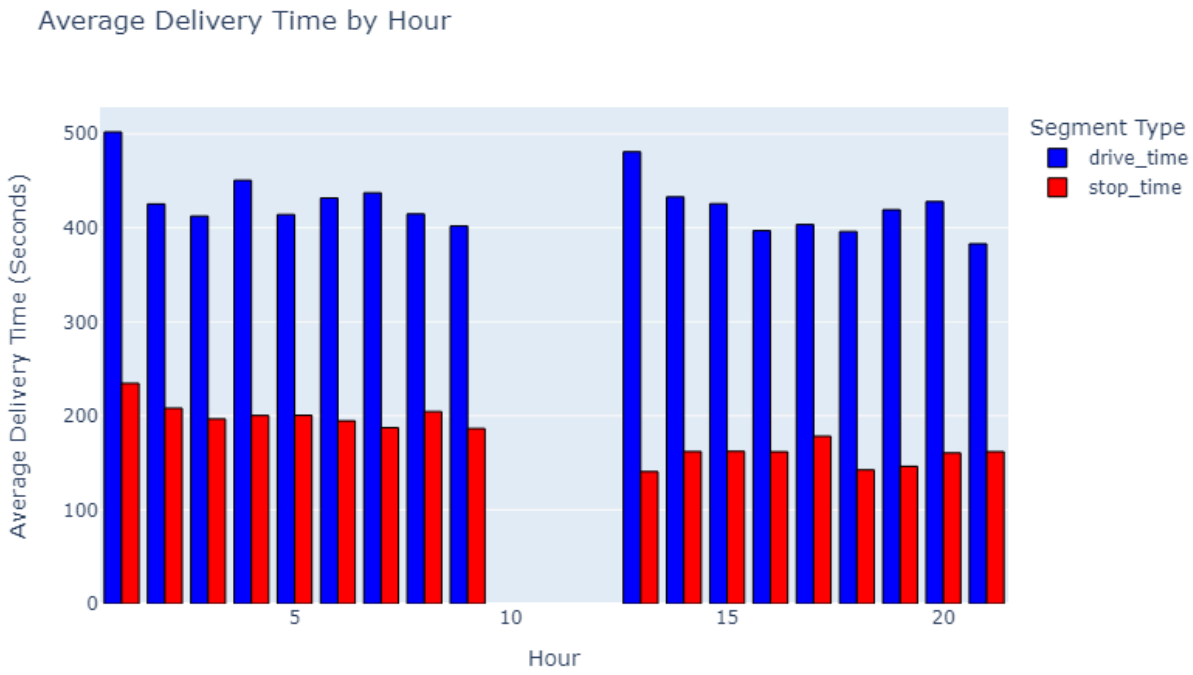
Delivery by week day



The graph shows no relationship.

The data was grouped by day of the week and the average travel time for each category was counted.

Delivery by hour



The first thing that strikes you is that there are two breaks of several hours per day. Around midday and midnight. Despite this, the times are rather similar, with longer deliveries immediately after the break and slightly falling off until the break.

The data was grouped by hour and the average travel time for each category was counted.

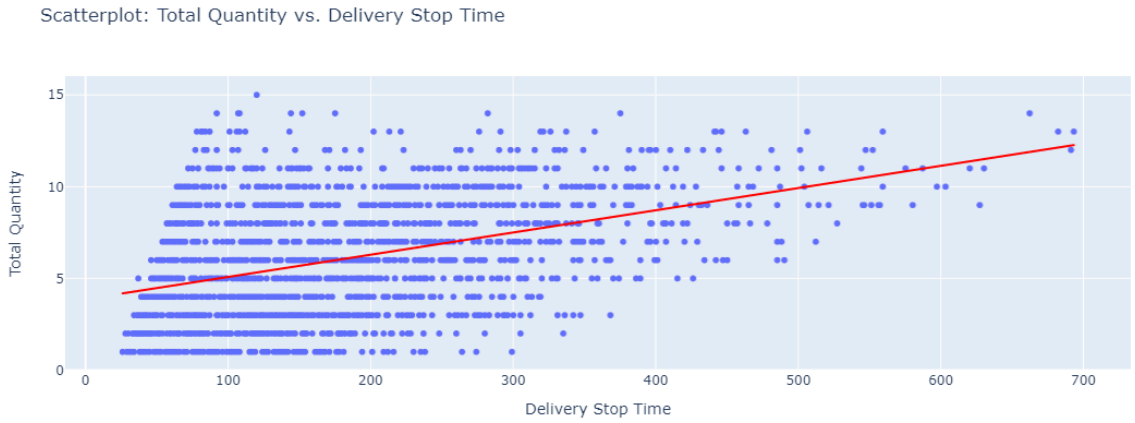
Weight vs time

Despite the ability to unload even heavy loads quickly. Cases with longer standstill times are generally characterized by heavier loads.

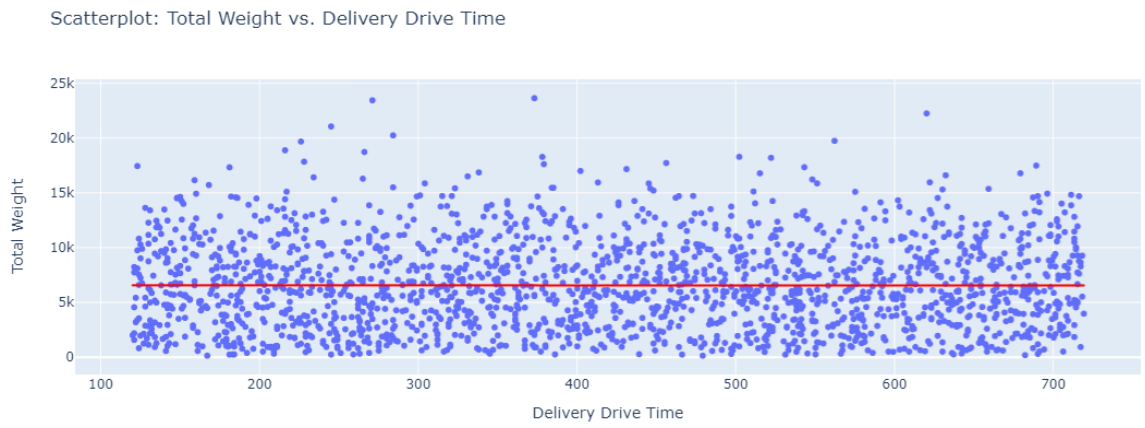
This type of graphs on the oy axis has weight data and on the ox axis has delivery time in seconds.

The chart shows that there is no relationship between delivery time and order weight.

Quantity vs time

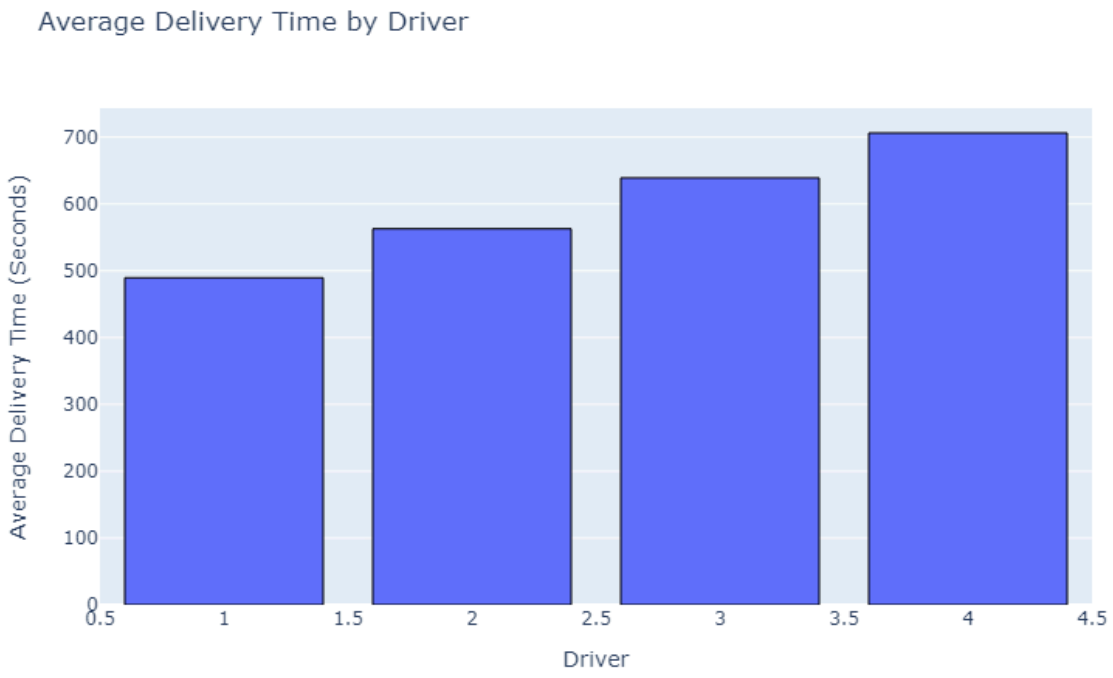


Despite the ability to quickly unload, loads with large volumes of products. Cases with longer standstill times tend to be characterized by higher quantities.



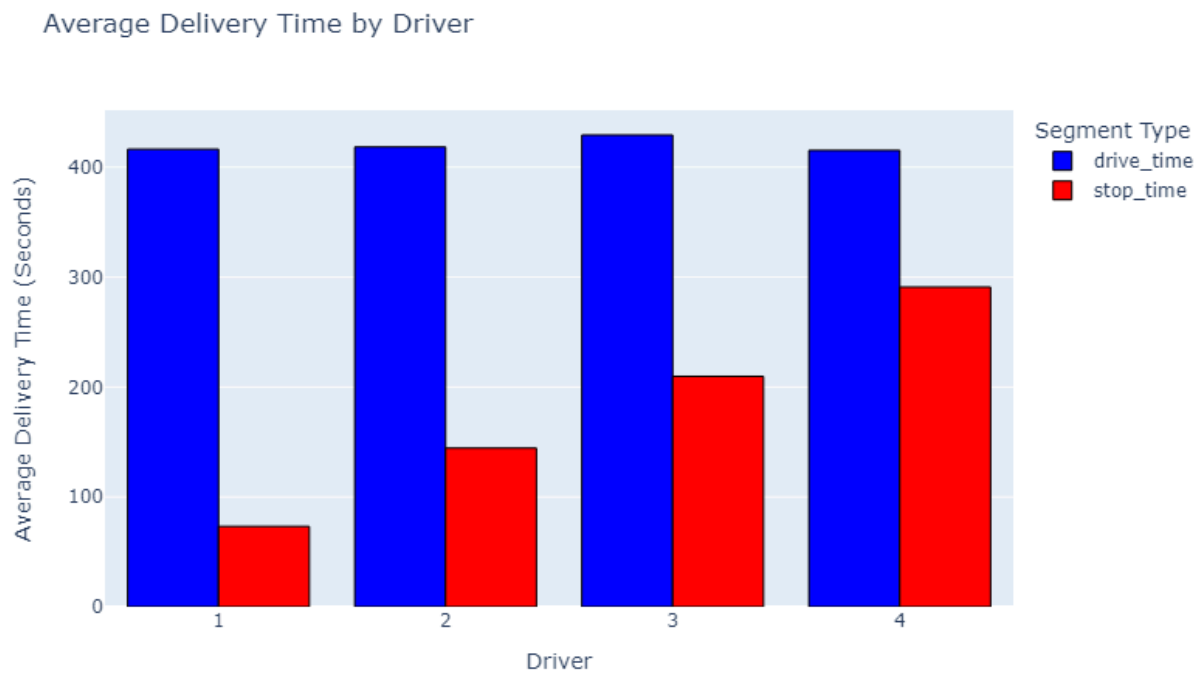
The graph shows no relationship.

Drivers



The graph shows that drivers have different average journey times. It is worth checking if anything affects this.

The data was grouped by driver id and the average travel time for each category was counted.



It is clearly seen that the difference is due to stop time/unpacking time.

Now the previous analyses make more sense.