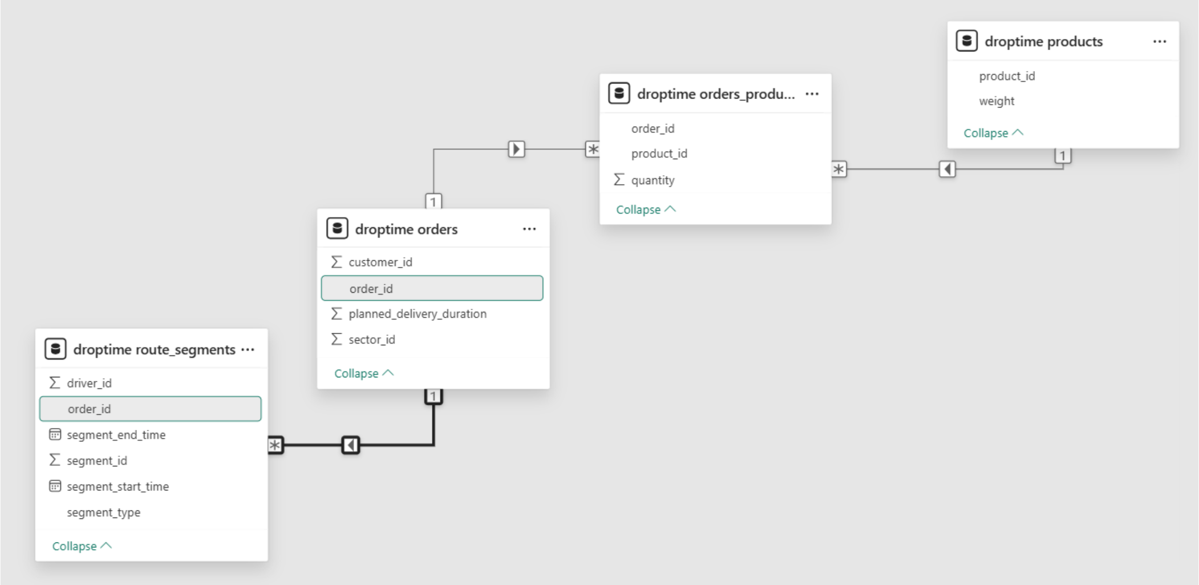
Part 2. Data analysis and visualisation

1. *Preparing the data*

As my tool for data analysis and visualization, I used Power BI. First, I connected to the MySQL database and imported four tables into the program

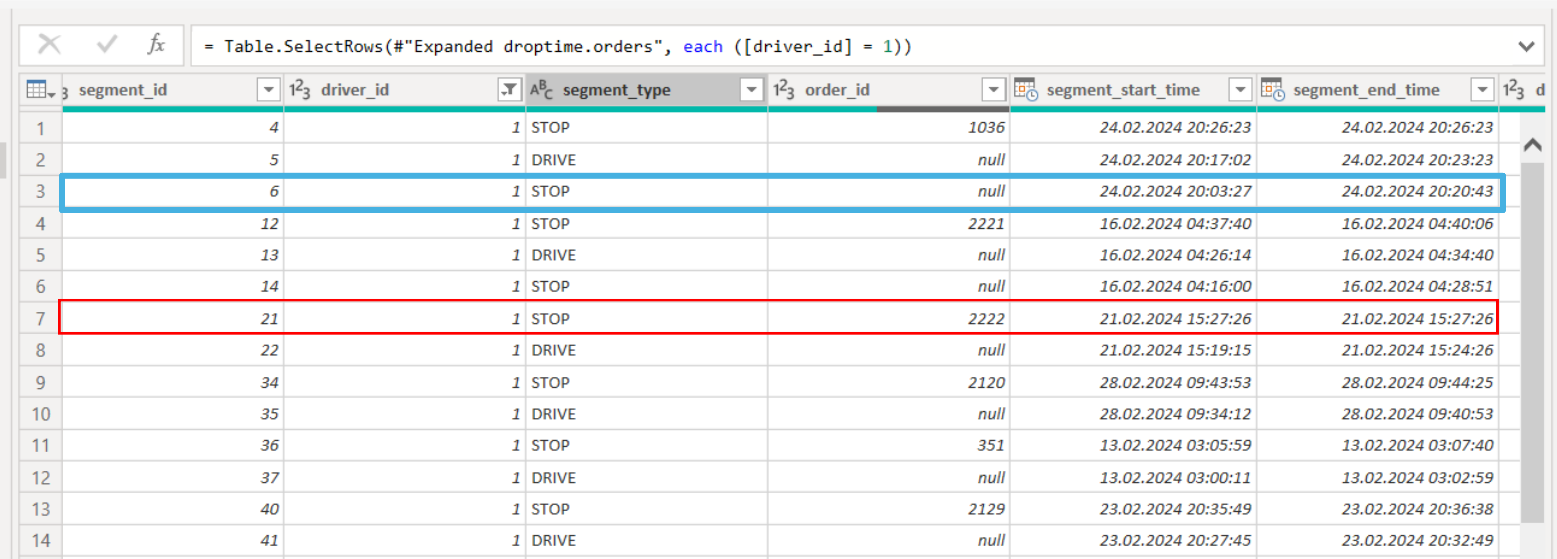
Next, I reviewed the model view - Power BI generated the relationships automatically after importing the data, and I verified that they were correct. The model includes one-to-many relationships between the tables, which is appropriate for the dataset



In the next step, I checked the data quality using Power Query. I noticed several issues:

- **missing order_id values for STOP segments** - in the route_segments table, the order_id column was sometimes null not only for DRIVE segments (as expected) but also for STOP segments (example highlighted in blue). This could indicate that the driver took a break, that other circumstances occurred, or simply that the data is incorrect
- **the same start and end time delivery** - this suggests a potential issue with how segment times are being recorded
- **end time earlier than start time** - the data in these rows was invalid

The problematic rows were excluded from further visualizations to ensure data accuracy.



```
= Table.SelectRows(#"Expanded droptime.orders", each ([driver_id] = 1))
```

	segment_id	driver_id	segment_type	order_id	segment_start_time	segment_end_time	d
1	4	1	STOP	1036	24.02.2024 20:26:23	24.02.2024 20:26:23	
2	5	1	DRIVE	null	24.02.2024 20:17:02	24.02.2024 20:23:23	
3	6	1	STOP	null	24.02.2024 20:03:27	24.02.2024 20:20:43	
4	12	1	STOP	2221	16.02.2024 04:37:40	16.02.2024 04:40:06	
5	13	1	DRIVE	null	16.02.2024 04:26:14	16.02.2024 04:34:40	
6	14	1	STOP	null	16.02.2024 04:16:00	16.02.2024 04:28:51	
7	21	1	STOP	2222	21.02.2024 15:27:26	21.02.2024 15:27:26	
8	22	1	DRIVE	null	21.02.2024 15:19:15	21.02.2024 15:24:26	
9	34	1	STOP	2120	28.02.2024 09:43:53	28.02.2024 09:44:25	
10	35	1	DRIVE	null	28.02.2024 09:34:12	28.02.2024 09:40:53	
11	36	1	STOP	351	13.02.2024 03:05:59	13.02.2024 03:07:40	
12	37	1	DRIVE	null	13.02.2024 03:00:11	13.02.2024 03:02:59	
13	40	1	STOP	2129	23.02.2024 20:35:49	23.02.2024 20:36:38	
14	41	1	DRIVE	null	23.02.2024 20:27:45	23.02.2024 20:32:49	

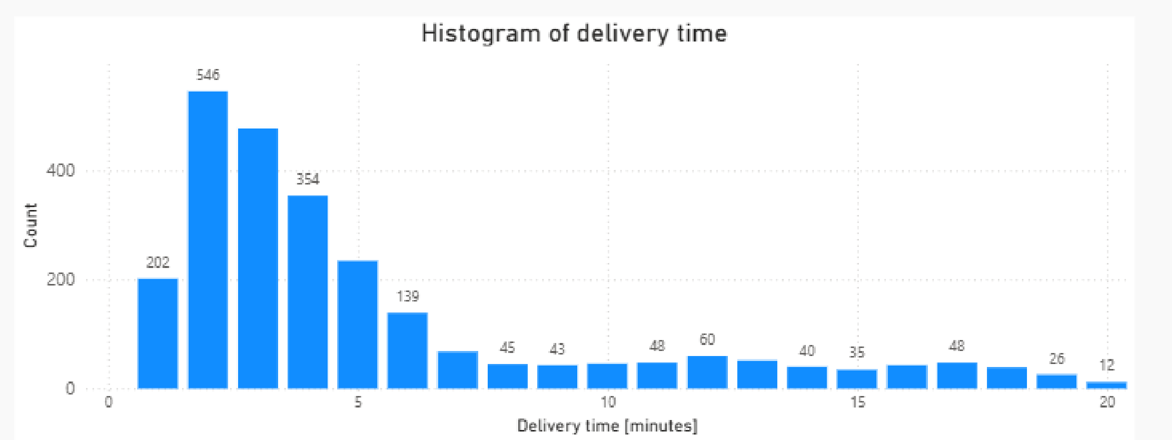
2. *Histogram with the actual delivery length with 1 minute granularity (rounded up).*

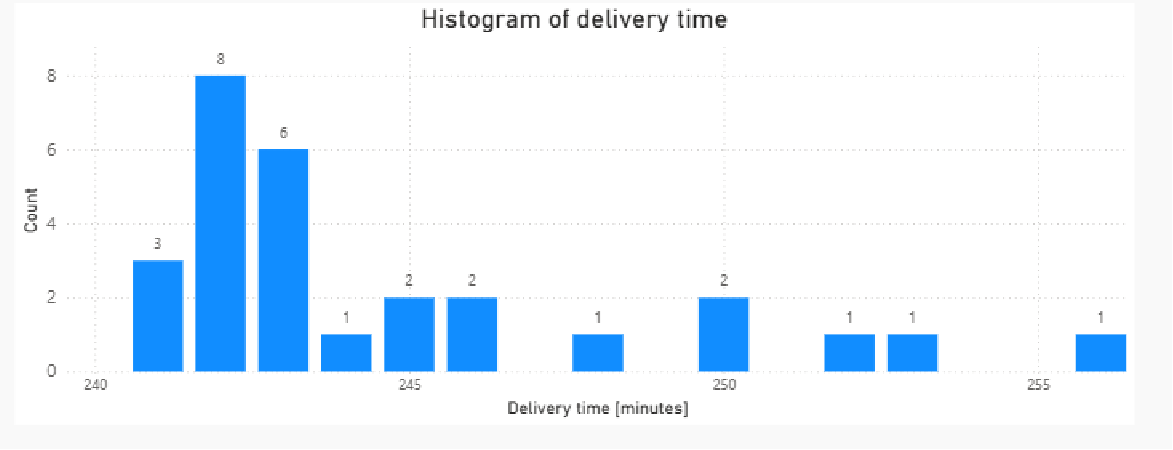
To begin, I created a new column that calculates the delivery time in minutes and rounds it up using the following formula:

Number.RoundUp(Duration.TotalMinutes([segment_end_time] - [segment_start_time]))

Next, I visualized the data using a histogram - most deliveries last between a few minutes and up to 20 minutes. However, there was a noticeable gap - a group of deliveries with durations between 240 and 260 minutes (over 4 hours), which is highly unusual. These records are likely incorrect or outliers, so I excluded them from the main (first) histogram.

After removing those values, the histogram clearly shows that most deliveries are completed within the first 5 minutes, with the peak occurring around 2nd minute.





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

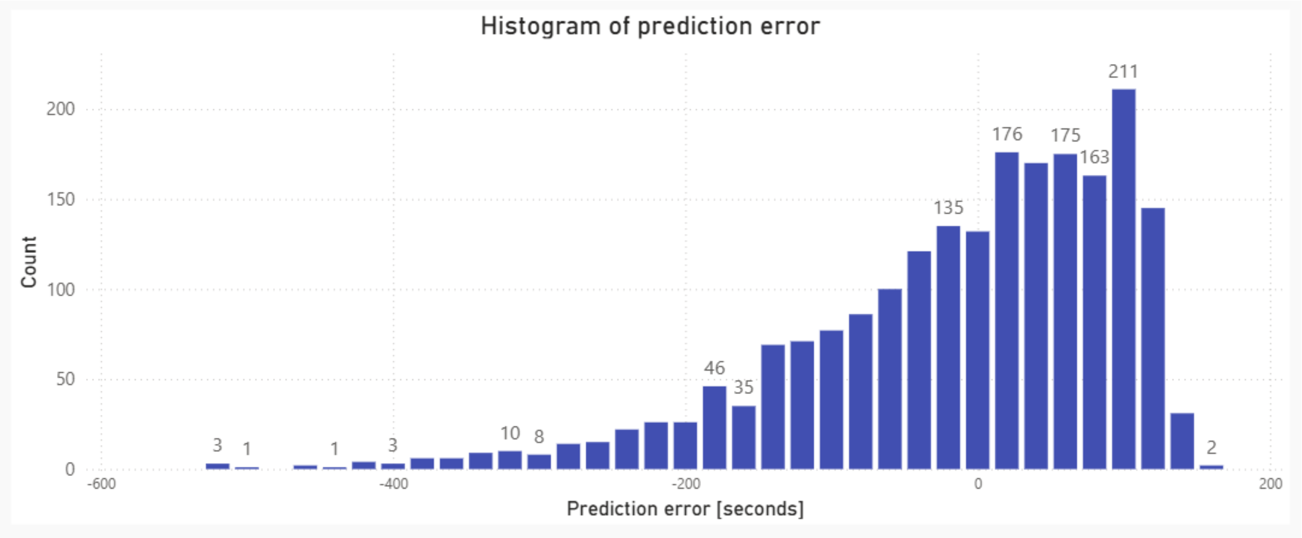
Before creating a chart, i calculated two columns :

- the first one calucating the real time delivery in seconds

```
if [segment_type] = "STOP" then Duration.TotalSeconds([segment_end_time] - [segment_start_time]) else null
```

- the second one calcuating the predicion error. To ensure that only the right data is calculated, i added some conditions

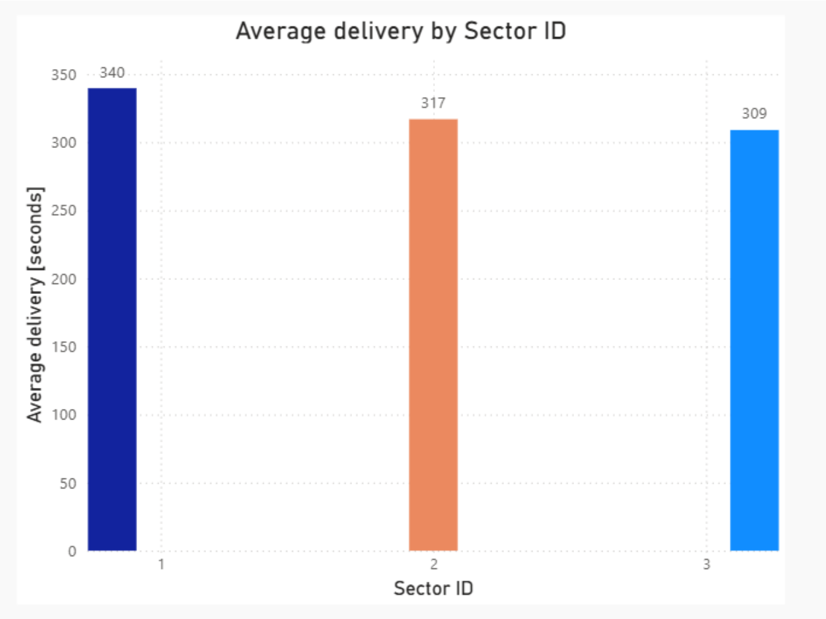
```
if [delivery_time_rounded] < 230        #  excluded wrong incorrect data

and [delivery_times_sec] <> null        # delivery time can't be a null value

and [delivery_times_sec] <> 0 and [delivery_times_sec] > 0      # value must be > than 0

then [droptime.orders.planned_delivery_duration]-[delivery_times_sec] else null
```



To create this chart, I aggregated the prediction error column. Since there were many unique values (seconds), each bar represents a 20-second interval. In most cases, the actual delivery was faster than planned - this is reflected in the fact that the most frequent bars show positive values.

4. *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.*

To visualize which sector takes longer to deliver packages, I used a bar plot. The y-axis shows the average delivery time for each sector. Sector 1 typically takes around 340 seconds per delivery, while sectors 2 and 3 take approximately 317 and 309 seconds - indicating that deliveries in sector 1 take longer.

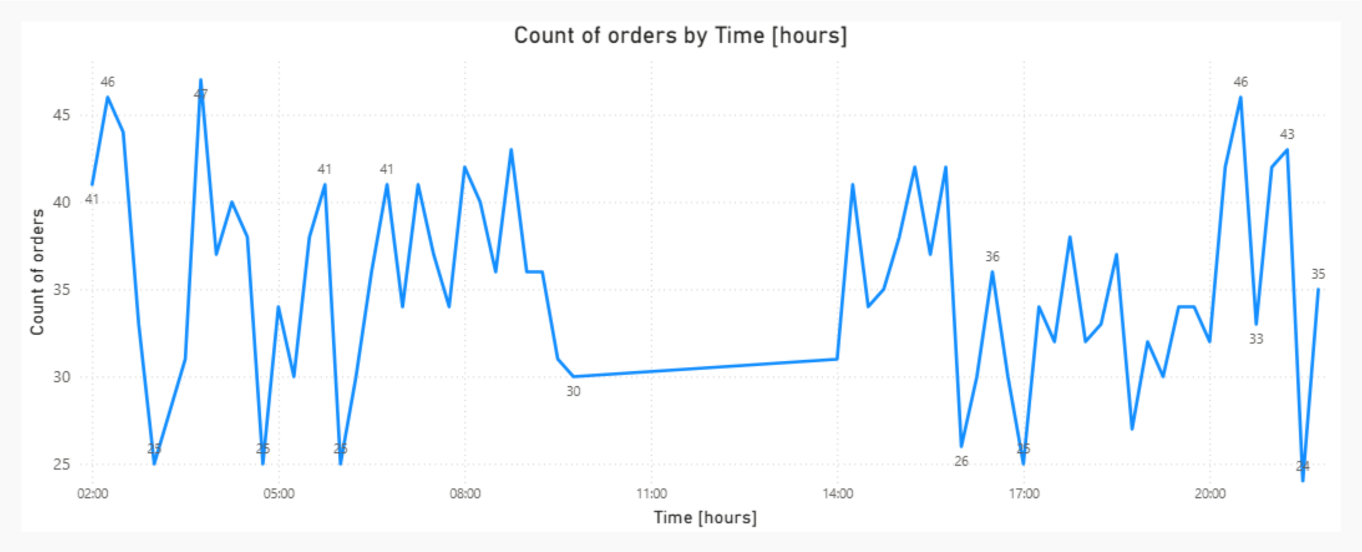


5. *Play with the data by grouping, aggregating and remodelling it. Are you able to find any correlations or trends that could be valuable for prediction quality improvement? Describe briefly your findings and visualise them on charts.*

a) number of orders delivered by time

I created a new column that extracts the delivery start time, using only valid data based on the previous analysis. Then, I aggregated the data by 15-minute intervals.

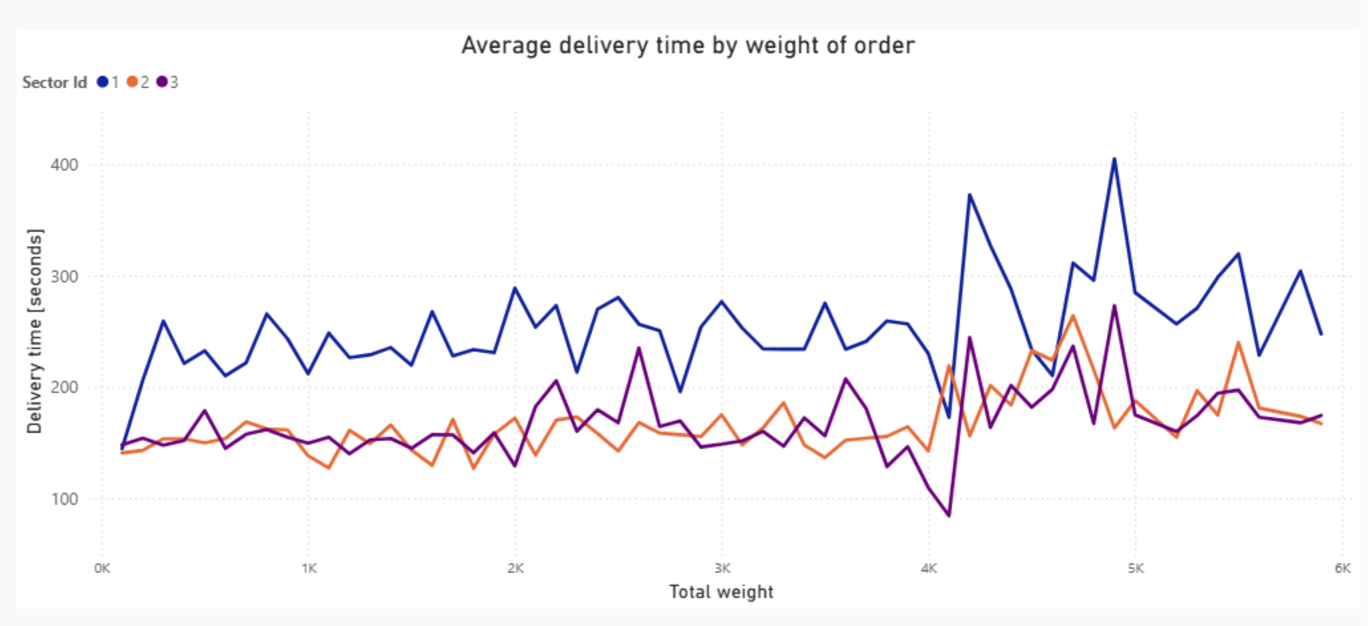
Next, I **created a line chart** showing the number of packages delivered at different times of the day. The chart indicates that deliveries start around 2 and end before 22. However, it also shows a surprisingly high number of deliveries taking place during the night, which is likely incorrect. This suggests issues with the timestamp data - for example, missing or incorrect dates - since deliveries are not typically made at night, and we would expect more activity in the afternoon.



b) average delivery time by weight of order and sector

I created a new column to calculate the total weight of each order. Using a line plot, I visualized the average delivery time (in seconds) depending on the order weight and sector. The weight was grouped in 100-gram intervals for better readability

The chart shows that packages in sector 1 take the longest to deliver overall. Additionally, order weight does not significantly correlate with delivery time until it exceeds 4 kg - after this threshold, delivery times increase across all sectors.


Average delivery time by weight of order
Sector Id ● 1 ● 2 ● 3
Delivery time [seconds]
0K
1K
2K
3K
4K
5K
6K
Total weight