Delivery Time Analysis Report

Natalia Klinik

Introduction

The purpose of this analysis was to evaluate how accurately current delivery time predictions reflect actual delivery durations and to identify factors that may influence them. The analysis was based on real delivery data, focusing specifically on delivery events — defined as route segments labeled *STOP*.

Data Loading and Preparation

To protect sensitive information, database connection details were stored in a secure *.env* file. Data was extracted from the *droptime* database using *mysql.connector* and saved into separate CSV files.

Relevant tables were loaded into dataframes and examined for structure and quality. The core of the analysis relied on a merged dataset, combining data from the *route_segments* and orders tables using the shared *order_id*.

The dataset contained several rows with missing values (marked as *NaN*). While no assumptions were to be made, this fact remained important for further data processing.

A key aspect was understanding what segments represent — essentially, parts of a driver's route. In this case, the focus was on *STOP* segments, which correspond to moments when the driver halts to deliver an order to the customer.

I checked for missing values in the *segment_start_time* and *segment_end_time* columns. Fortunately, all entries in these columns were complete. To enable proper time-based operations, their data types were converted to datetime.

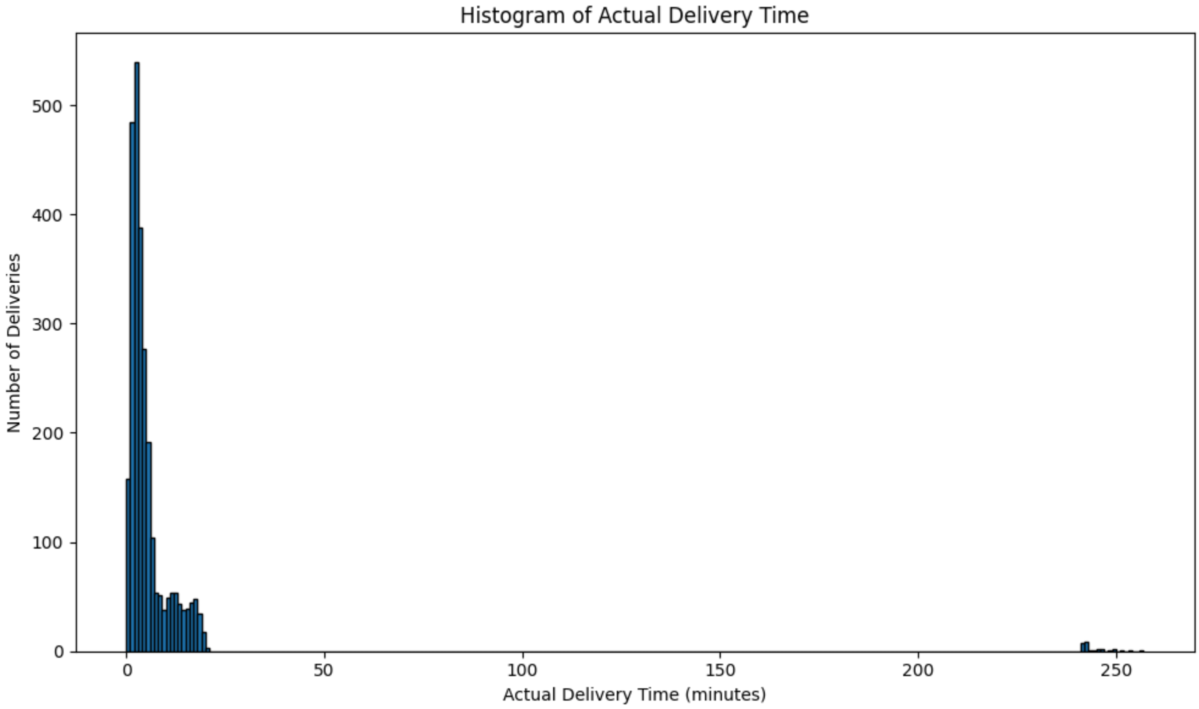
The analysis was limited to *STOP* segments, stored in a new dataframe called *stops_df*.

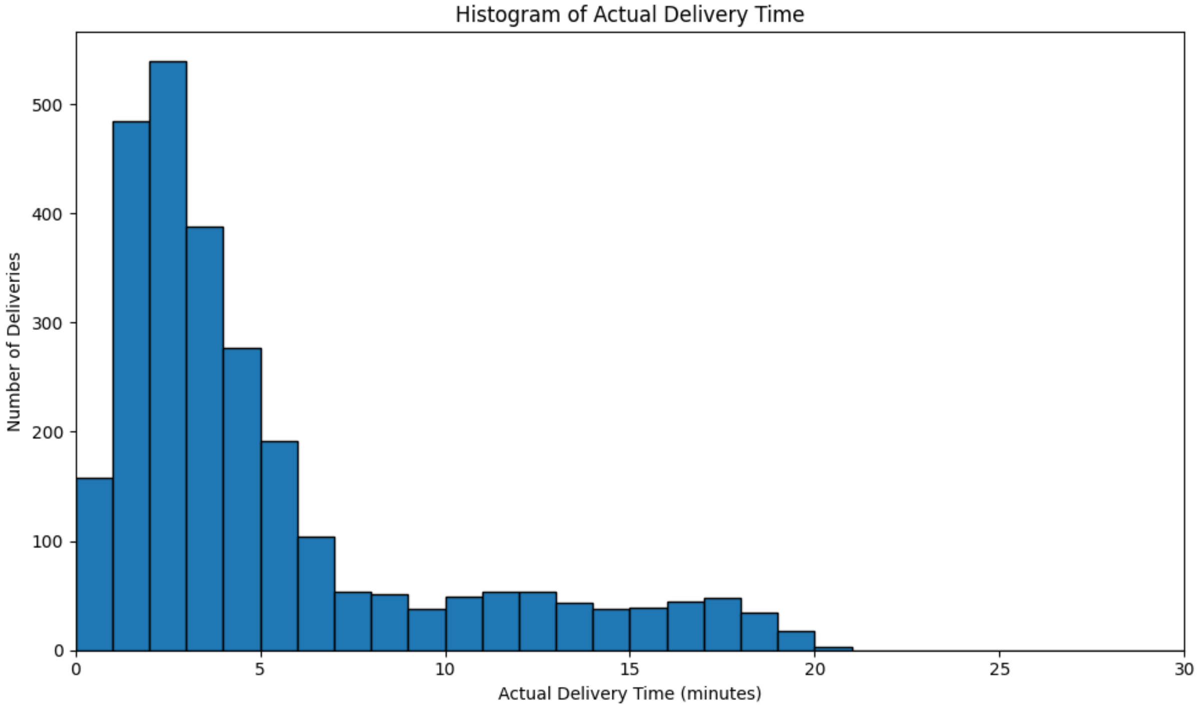
The main goal was to determine how much the actual delivery time deviated from the predicted delivery time. The *actual_delivery_time* column was created by subtracting *segment_start_time* from *segment_end_time*. This new column was added to the dataframe, with the results rounded to whole minutes.

Due to some missing values, it was crucial to process only non-null entries. In the original dataset, the predicted delivery time was provided in seconds, a format less intuitive for human interpretation, so the values were converted to minutes for clarity.

Visualizations and Findings

1. Actual Delivery Time

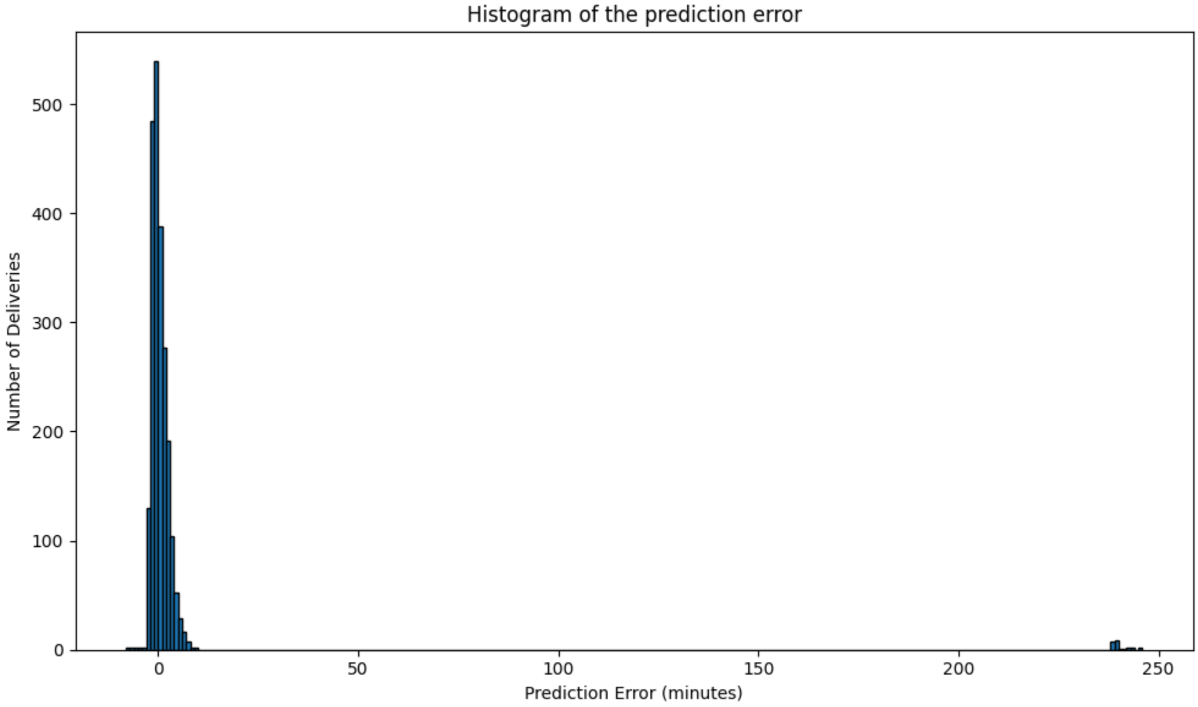


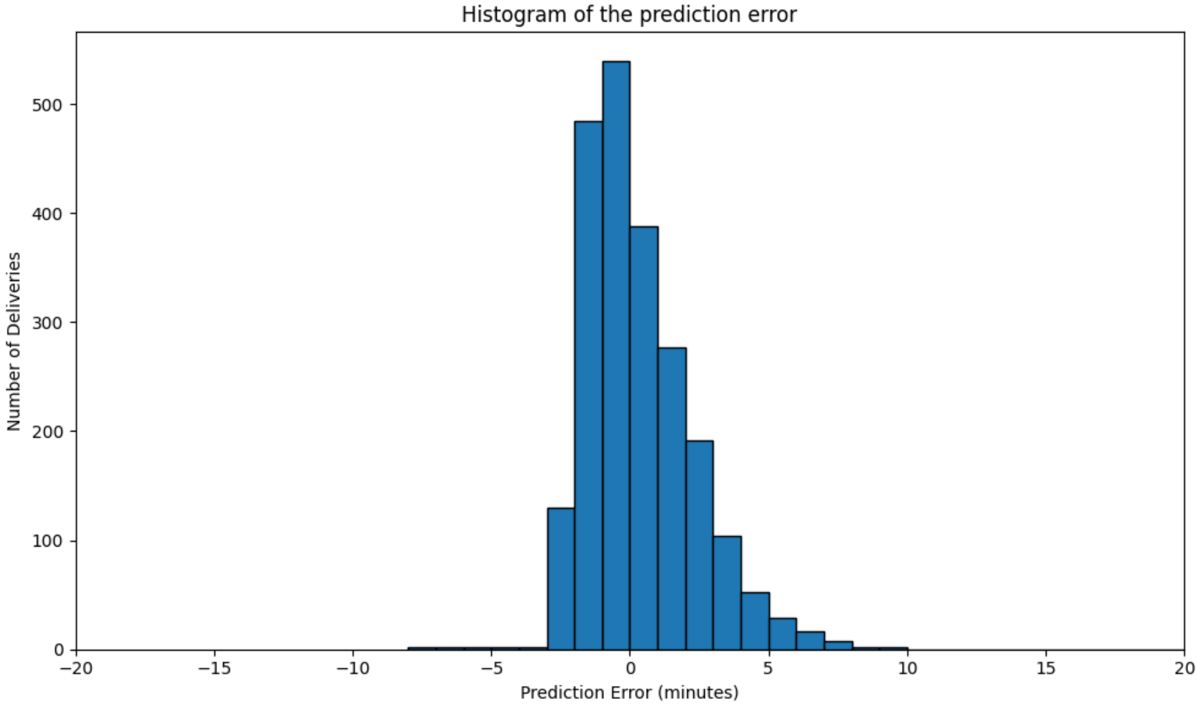


A histogram shows that the vast majority of deliveries are completed within 20 minutes. Although there are a few outliers (up to 250 minutes), they represent rare cases and are not necessarily errors.

Key takeaway: Most deliveries are quick and within expected timeframes.

2. Prediction Error



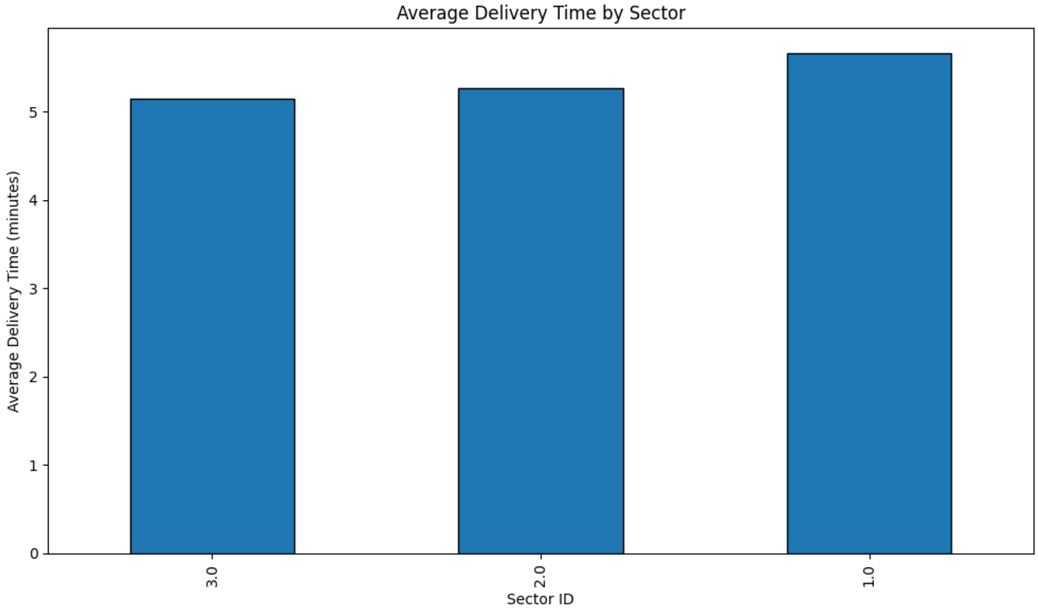


Another histogram shows the difference between predicted and actual delivery times. Most deliveries are within ±10 minutes of the prediction, with many even arriving earlier than expected.

Outliers exist but are uncommon. Limiting the graph to the -20 to +20 minute range helped visualize the bulk of the data more clearly.

Key takeaway: The current prediction system is mostly accurate, but sometimes underestimates delivery duration.

3. Average Delivery Time by Sector

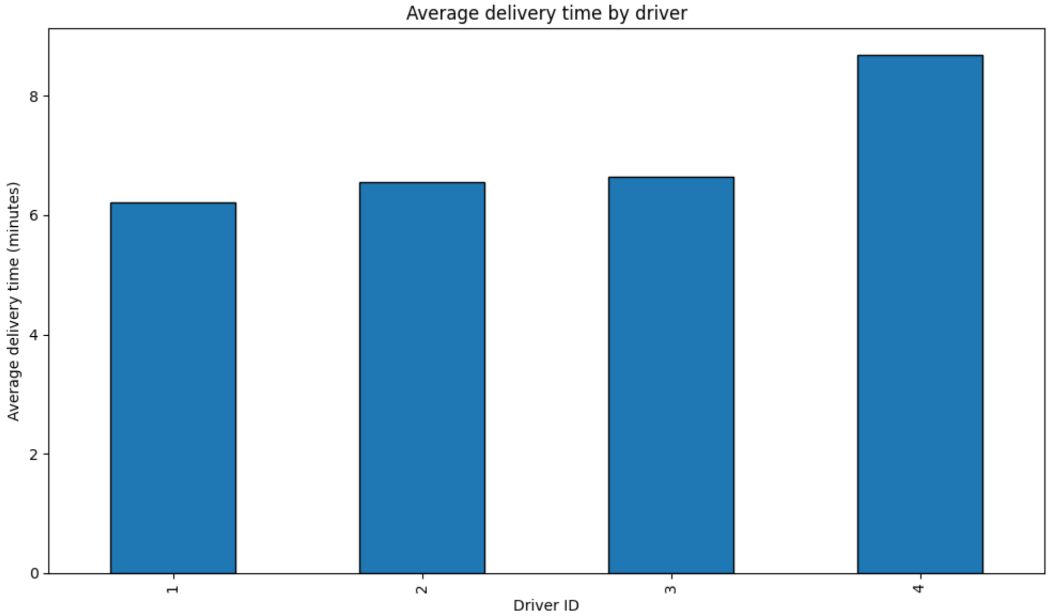


Delivery durations was compared across the three delivery sectors. Although Sector 3 appears slightly slower, the difference is minor — less than one minute on average.

Key takeaway: Sector has minimal influence on delivery duration.

4. Additional Analyses

4.1. Delivery Time by Driver



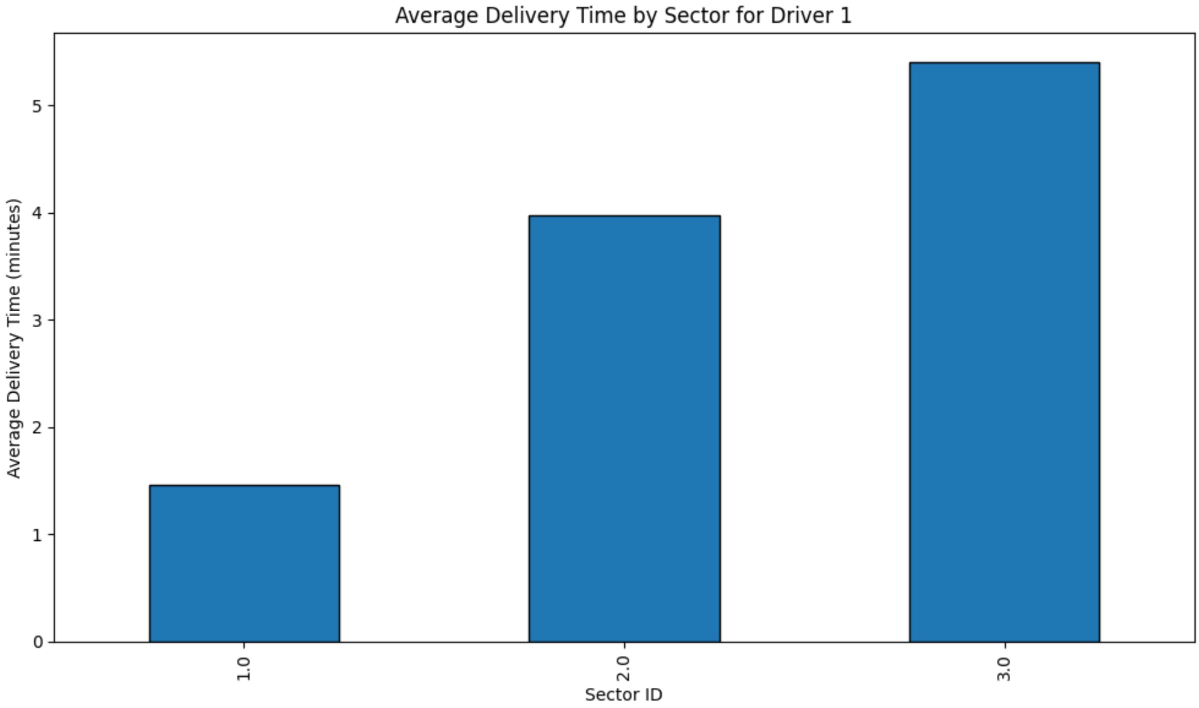
While plotting average delivery times according to drivers, it can be seen that Driver 4 takes significantly more time to deliver orders.

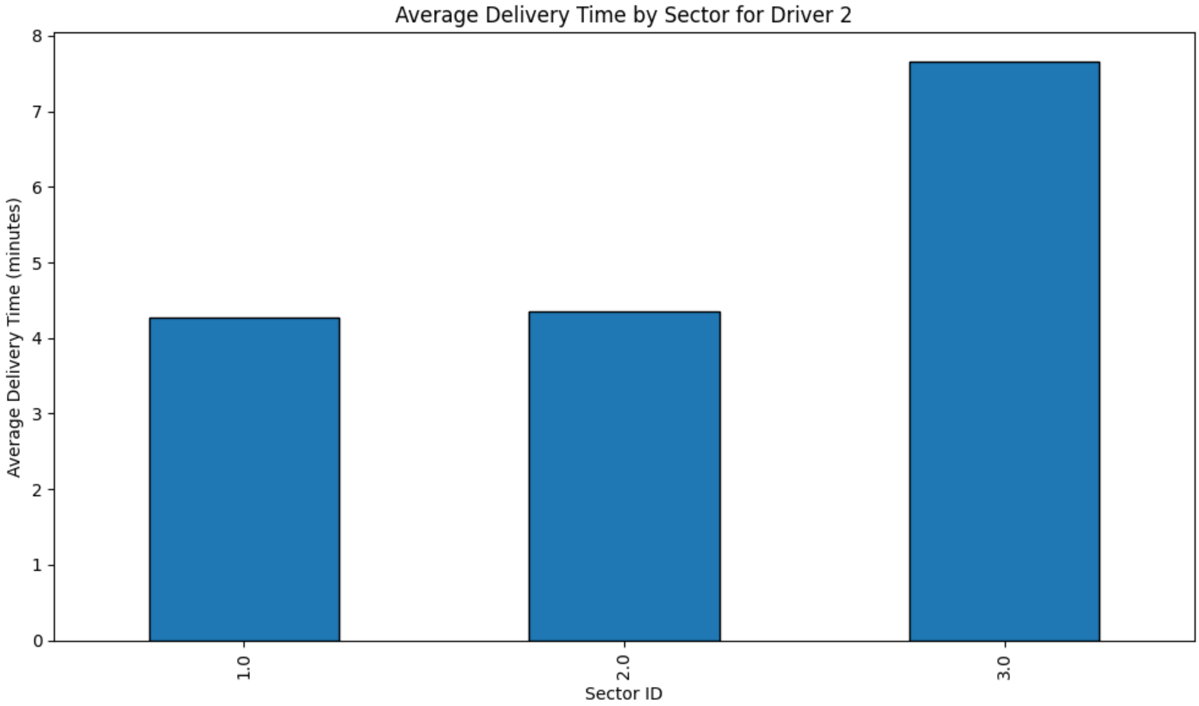
Key takeaway: Individual driver performance can affect delivery time.

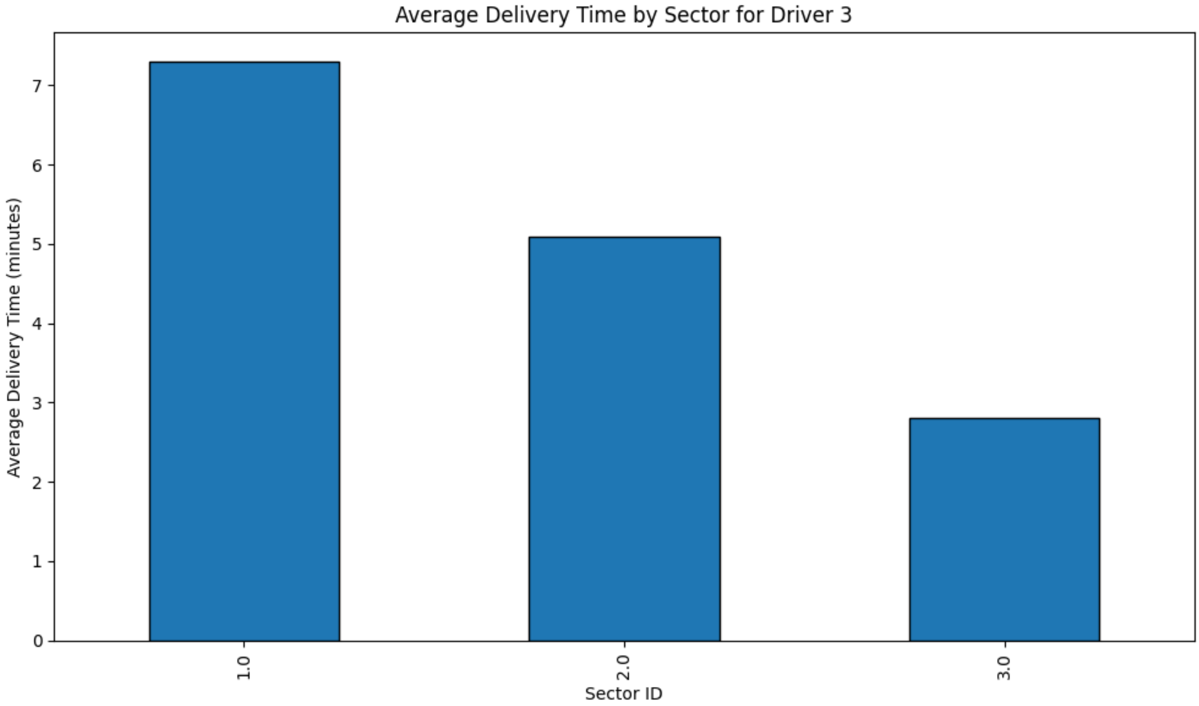
4.2. Driver + Sector Combination

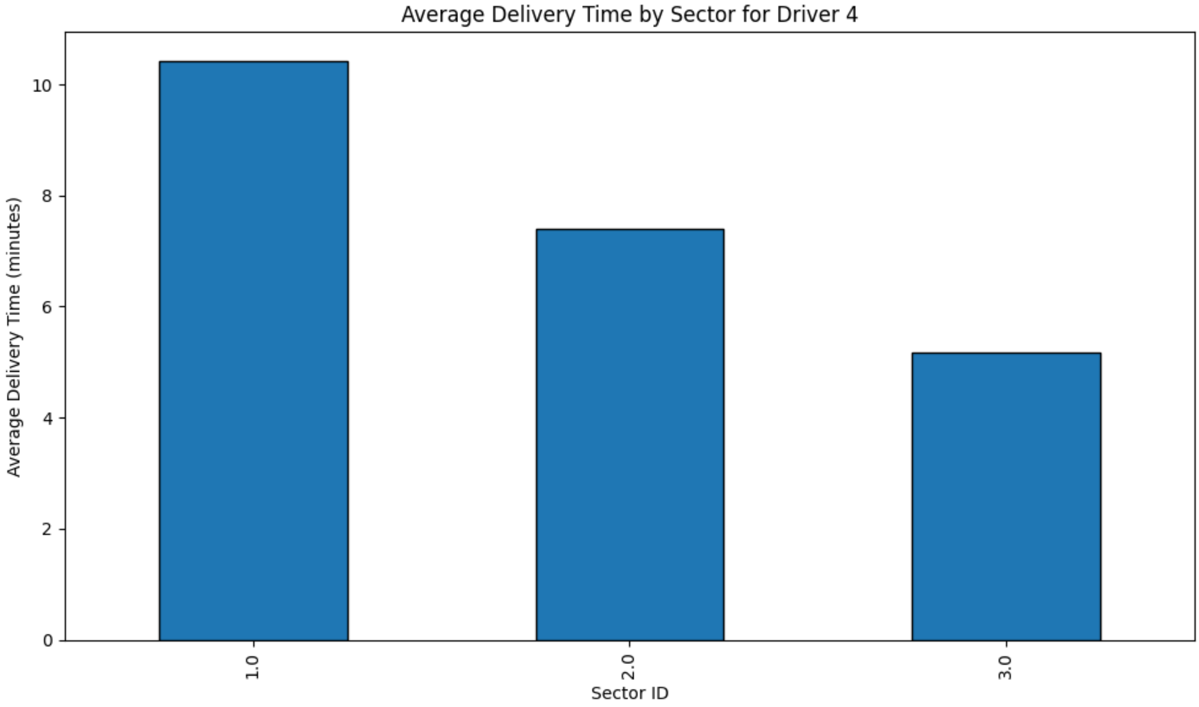
By analyzing driver performance within each sector, we noticed patterns:

- Drivers 1 and 2 are slower in Sector 3.
- Drivers 3 and 4 are slower in Sector 1.







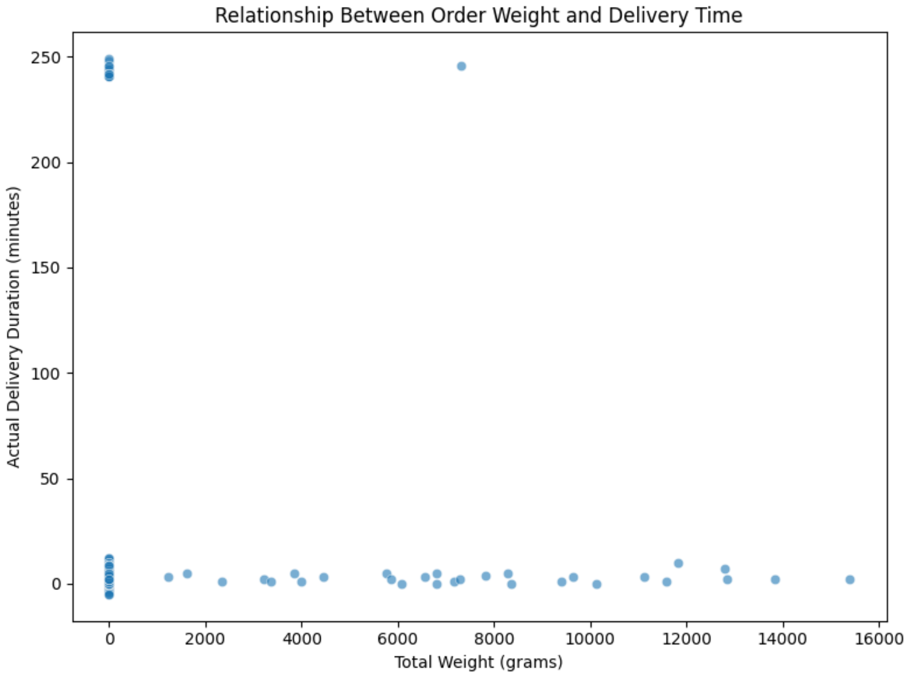


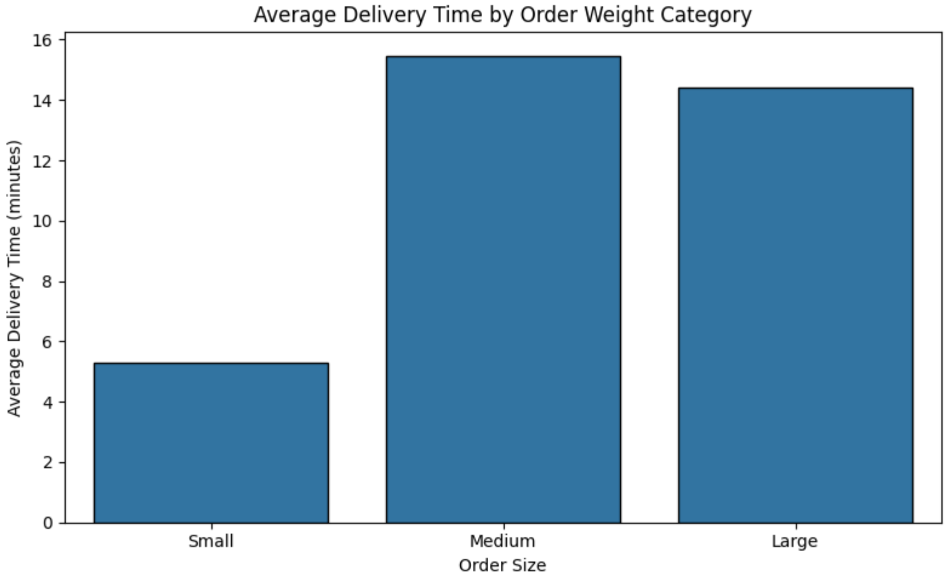
Key takeaway: Performance differences may depend on both driver and delivery area.

4.3. Delivery Time vs. Order Weight

Checking the distribution of *total_weight*, orders were categorized into:

- Small (<4kg)
- Medium (4–12kg)
- Large (>12kg)



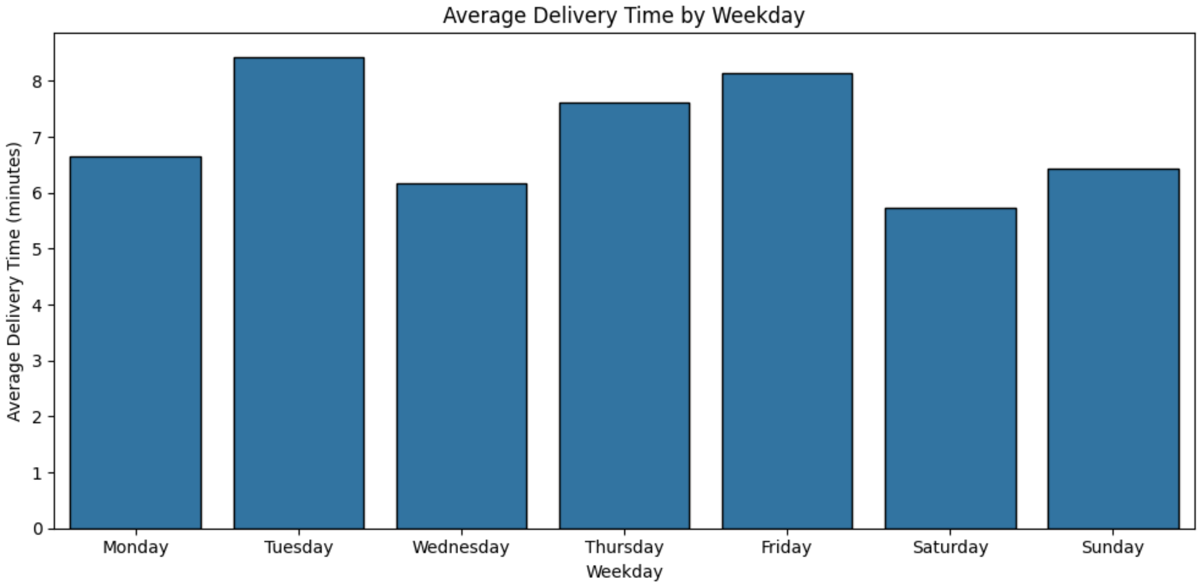


As expected, small orders took less to deliver. Medium and large orders took significantly more time.

Key takeaway: Heavier orders generally take longer, but not always — other variables matter too.

4.4. Delivery Time by Day of the Week

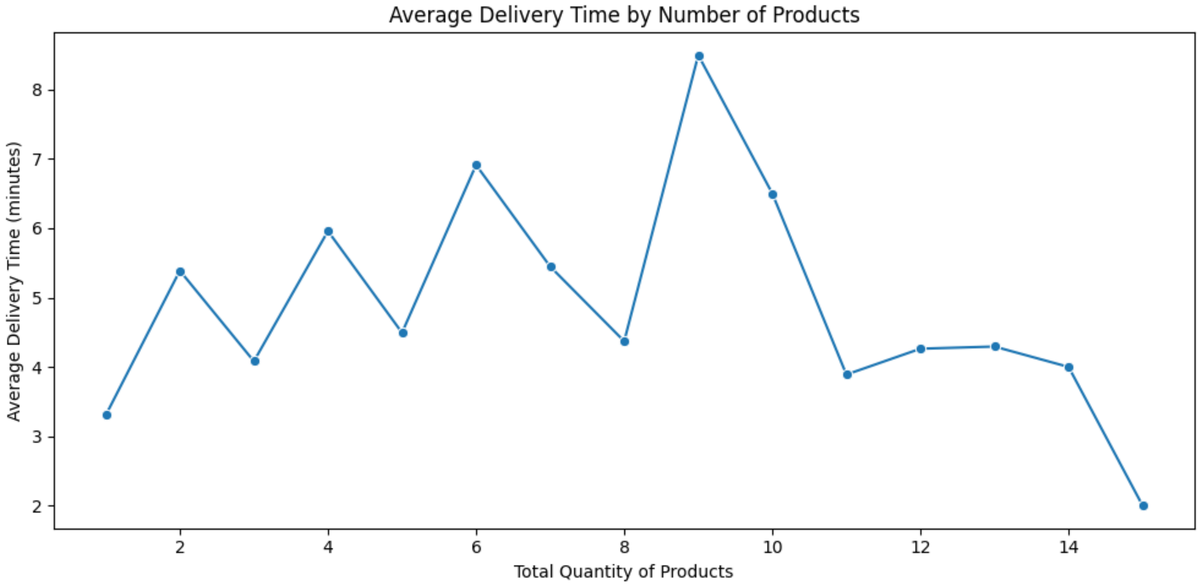
The deliveries were grouped by weekday, what showed slightly longer delivery times on Tuesdays. However, no consistent pattern was found across the week.



Key takeaway: Weekday does not significantly affect delivery time.

4.5. Delivery Time vs. Product Quantity

No strong correlation was found between how many items were in the order and how long the delivery took.



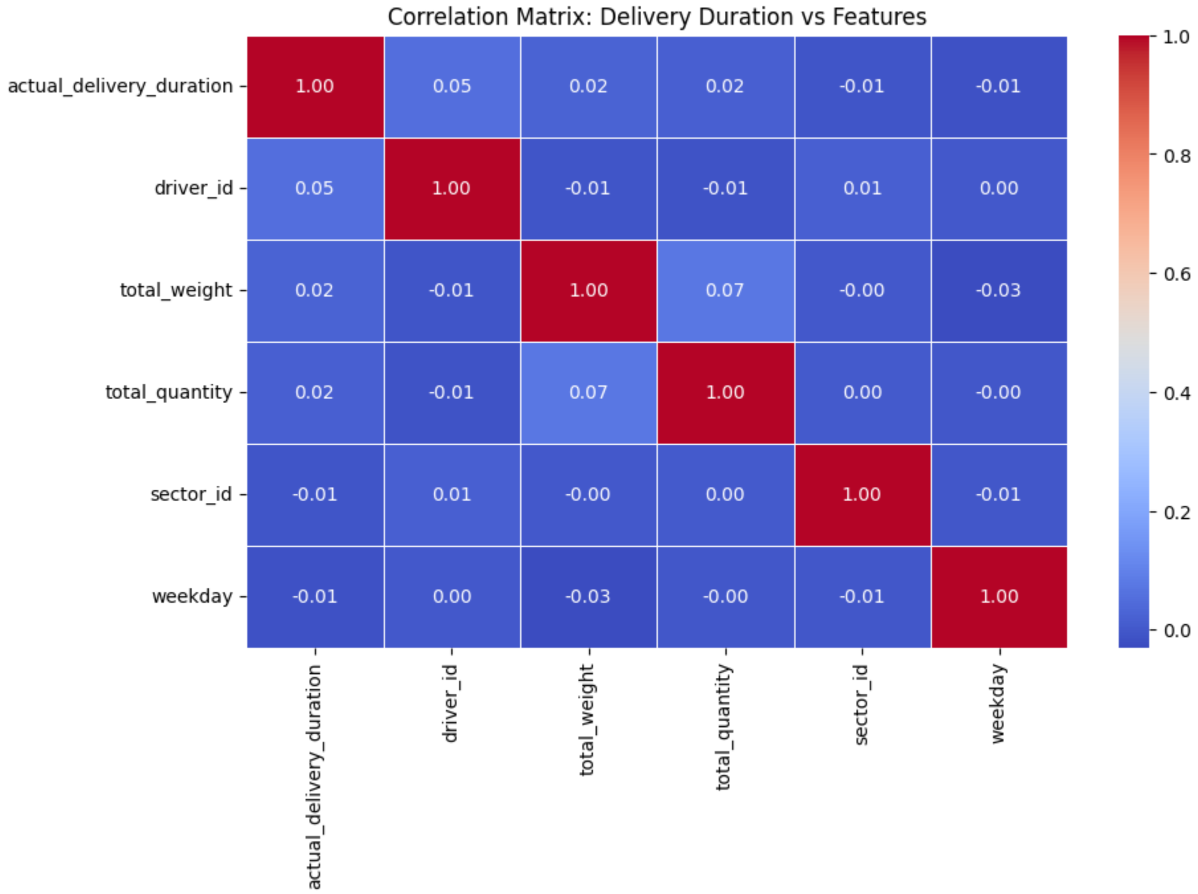
Key takeaway: Quantity of products does not strongly impact delivery duration.

4.6. Correlation Matrix

The correlation matrix was created to explore how delivery duration relates to:

- Driver ID
- Order weight
- Product quantity
- Sector
- Day of the week

However, o individual factor showed a strong correlation.



Key takeaway: Delivery time is influenced by a combination of factors rather than one dominant variable.

Conclusion

The current delivery time prediction system performs fairly well, with most predictions close to actual results. However, there's room for improvement — especially in edge cases involving specific drivers or heavier orders.

Key insights:

- Most deliveries are completed within 20 minutes.
- Prediction errors are usually small, but outliers exist.
- Driver behavior impacts delivery time more than location or weekday.
- Order weight plays a role, but not consistently.
- No single factor determines delivery time — a more complex prediction model could improve accuracy.

Note: For more details and technical aspects, the *part2.ipynb* file can be checked. It can be found in the *solutions/raw/* directory.