Data Analysis and Visualisation

Maciej Jamka

Introduction

The goal of this analysis is to evaluate the accuracy and efficiency of the delivery process by examining the relationship between planned and actual delivery times. Using the dataset provided, I will focus on the following objectives:

- Understanding how long deliveries actually take.
- Analyzing the accuracy of the delivery time predictions.
- Investigating whether there are significant differences in delivery times across various sectors.
- Exploring correlations and trends that may help improve the accuracy of future delivery time predictions.

Methodology

Data Cleaning and Preparation

Before conducting any analysis, I thoroughly cleaned the data to ensure that it was free of negative values in delivery times and any inconsistencies. This was achieved through filtering and grouping techniques. Specifically:

- **Checking for missing values:**
  I began by examining all relevant tables to identify any missing values in key fields. I used queries to count and confirm whether columns like order_id, product_id, sector_id, and others contained null or incomplete values. This process allowed me to ensure the integrity of the data. In the case of the route_segments table, I found missing values in the order_id field. I decided to filter out these incomplete records to avoid introducing errors in the analysis.
- **Handling negative values and outliers:**
  I then checked all numerical fields, such as delivery times, weights, and planned delivery durations, to ensure that there were no negative or unreasonable values (e.g., delivery times less than zero or weights equal to zero). Any such values were

removed from the dataset. To do this, I created views that filtered out records with problematic data. For instance:

  - I filtered out rows with negative actual_delivery_seconds and planned_delivery_duration.
  - I also ensured that weights were positive and non-zero, as negative or zero weights could distort the results.

- **Creating filtered views:**
  After cleaning the data, I created new views, such as clean_route_segments_no_negatives, which contained only valid records (e.g., no negative delivery times or missing values). These views became the foundation for all subsequent analysis.

This comprehensive data cleaning process ensured that the dataset used for analysis was robust and free from discrepancies that could impact the results. Once the data was properly prepared, I moved on to the analysis phase.

Histogram of Actual Delivery Length

Objective: To understand how long deliveries actually take.

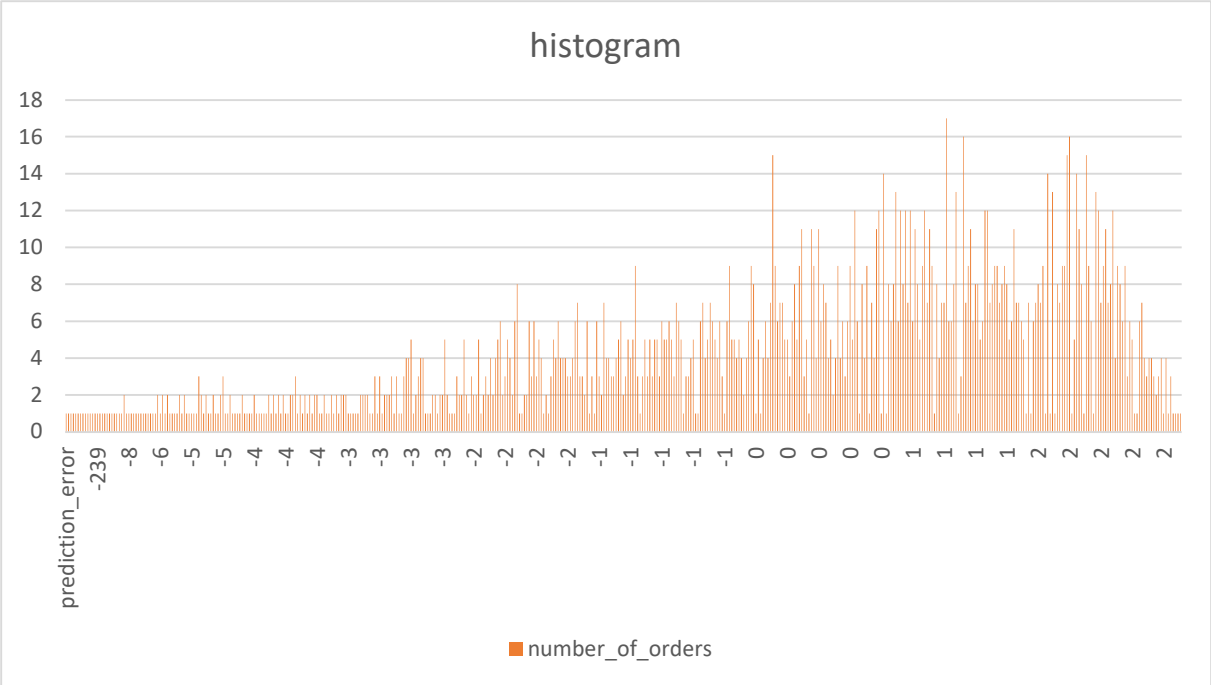
Methodology: We calculated the actual delivery time for each order by finding the difference between the segment_start_time and segment_end_time in the clean_route_segments_no_negatives view. The result was rounded up to the nearest minute, as requested. Then I created a histogram to visualize the distribution of these actual delivery times.

Finding: The histogram shows the distribution of actual delivery times in minutes. Most deliveries fall within a certain range, but there's a tail of longer delivery times. This suggests that while most deliveries are relatively quick, some take significantly longer. This could be due to various factors like traffic, parking, or customer issues.

As we see despite previous data wrangling, we can still face some outliers that we will investigate further.

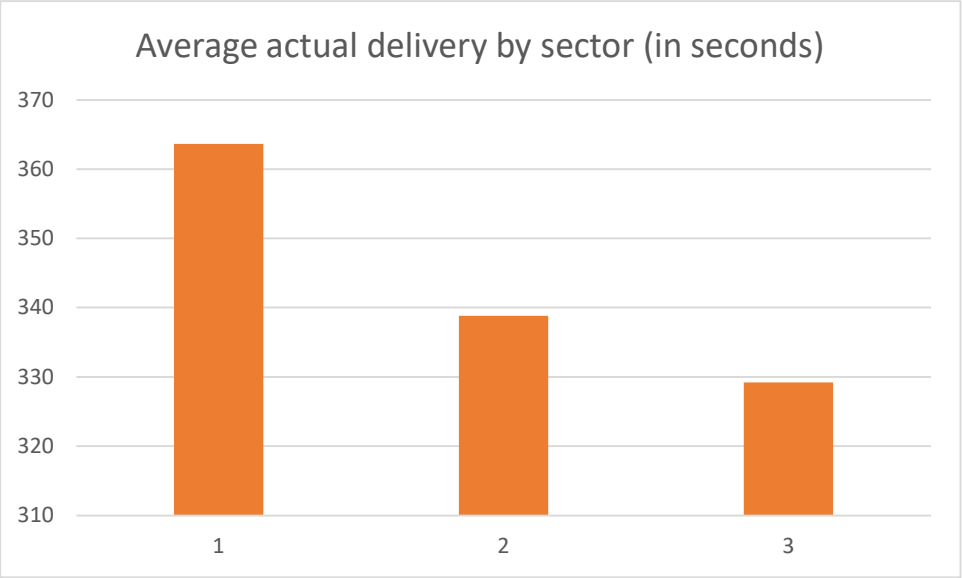
Histogram of Prediction Error

- **Objective:** To see how far off our current predictions are.
- **Methodology:** We calculated the difference between the planned delivery time and the actual delivery time for each order. A positive value means the planned time was longer than the actual time (overestimation), and a negative value means the planned time was shorter (underestimation). Then I created a histogram to visualize the distribution of these errors.
- **Finding:** The histogram of prediction errors shows how often our predictions are too high or too low. A wider spread indicates less accurate predictions.



Delivery Time by Sector

- **Objective:** To investigate the driver insight that deliveries in one sector take longer.
- **Methodology:** We calculated the average actual delivery time for each sector to compare delivery efficiency across different geographic areas.
- **Finding:** Average delivery times vary between sectors. Here's a summary of the average delivery times for the sectors in our data.



This data confirms that Sector 1 has a higher average delivery time compared to Sectors 2 and 3, supporting the drivers' observations that some sectors experience longer delivery times.