Delivery Time Predictions

SZCZEPAN POLAK

TABLE OF CONTENTS

INTRODUCTION

Accurate delivery time predictions are crucial for efficient route planning, better customer experience, and optimal use of drivers' time.

The purpose of this analysis is to support the improvement of delivery time predictions for grocery orders fulfilled by our company. Currently, the prediction model is overly simplistic—it uses the historical average delivery time as a universal estimate for all future orders. While functional, this approach does not account for the variety of real-world factors that may influence delivery durations.

The aim of this project is to explore the available delivery data and identify trends or correlations that could be used to enhance the prediction algorithm. For example, a recent insight from drivers suggested that deliveries to apartment buildings may take longer than those to single-family homes, although the system lacks direct information about building types.

To perform the analysis, I used a relational database populated with historical delivery data. All required data was extracted using SQL queries, and the analysis itself was conducted in a Python environment using Jupyter Notebook. This approach enabled efficient data transformation, visualization, and statistical evaluation.

Data Overview

The analysis was based on historical delivery data stored in a relational database. The database consisted of the following four tables:

- **orders** – containing information about each order, such as order ID, customer ID, sector, and planned delivery time.
- **products** – including product IDs and their respective weights.
- **orders_products** – mapping products to orders with quantity information.
- **route_segments** – providing detailed delivery route data, including start and end times, segment types, and associated order IDs.

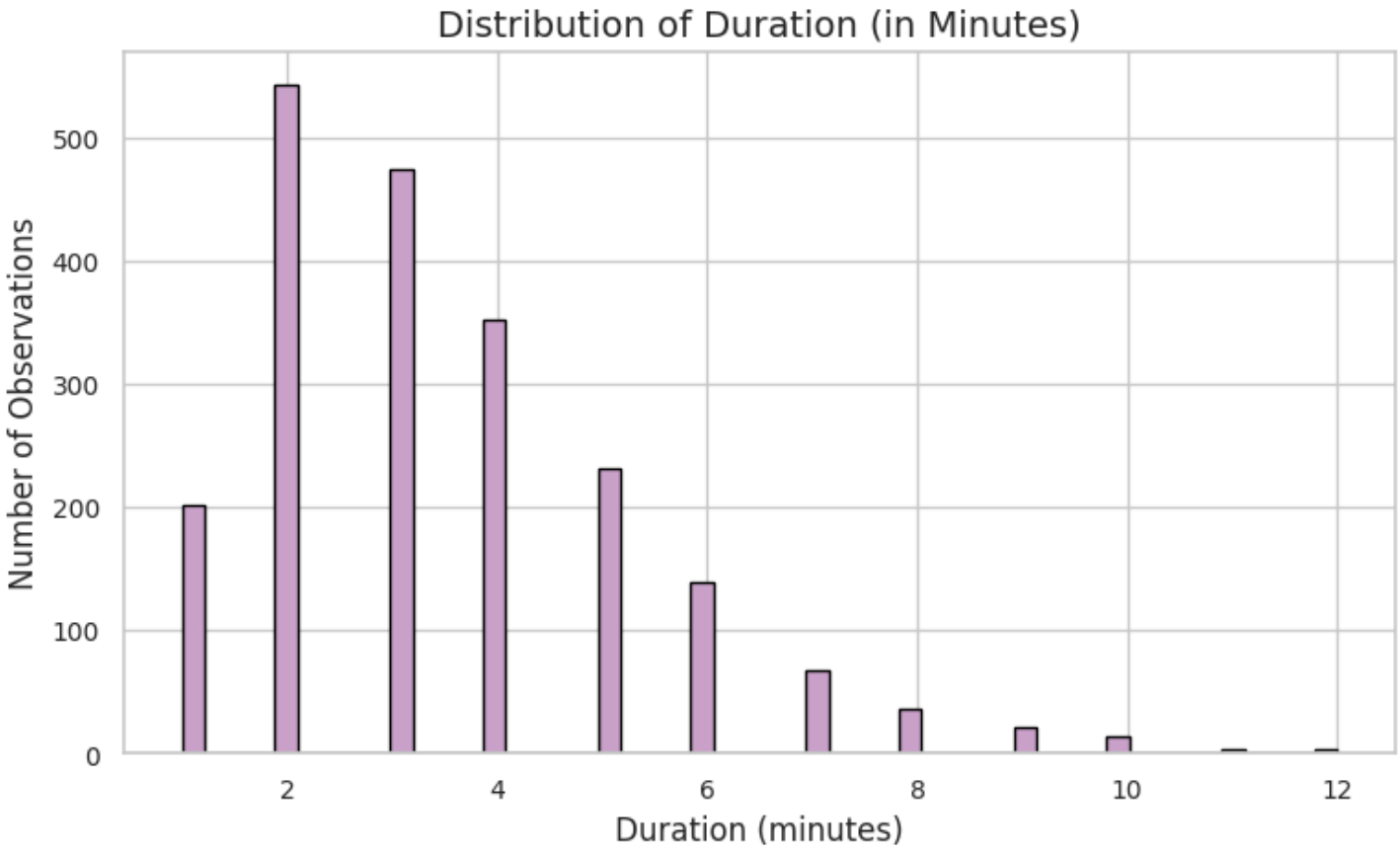
Before conducting the analysis, the data was carefully preprocessed. Only route segments with a valid order_id were included, as these were necessary for accurate delivery time calculations. Duplicate records were removed, and outliers such as negative delivery durations or unusually long deliveries (lasting several hours) were excluded. These cleaning steps were essential, as real-world data often contains erroneous or incomplete entries caused by technical or system-related issues.

Results and Visualizations

Actual Delivery Length Histogram

To analyze actual delivery durations, I calculated the time spent on each delivery by subtracting the segment_start_time from the segment_end_time for segments of type STOP, which represent actual deliveries. The resulting durations, initially in seconds, were converted to minutes by dividing by 60 and rounding up to the nearest whole number.

The histogram below shows the distribution of actual delivery durations, grouped by minute intervals.

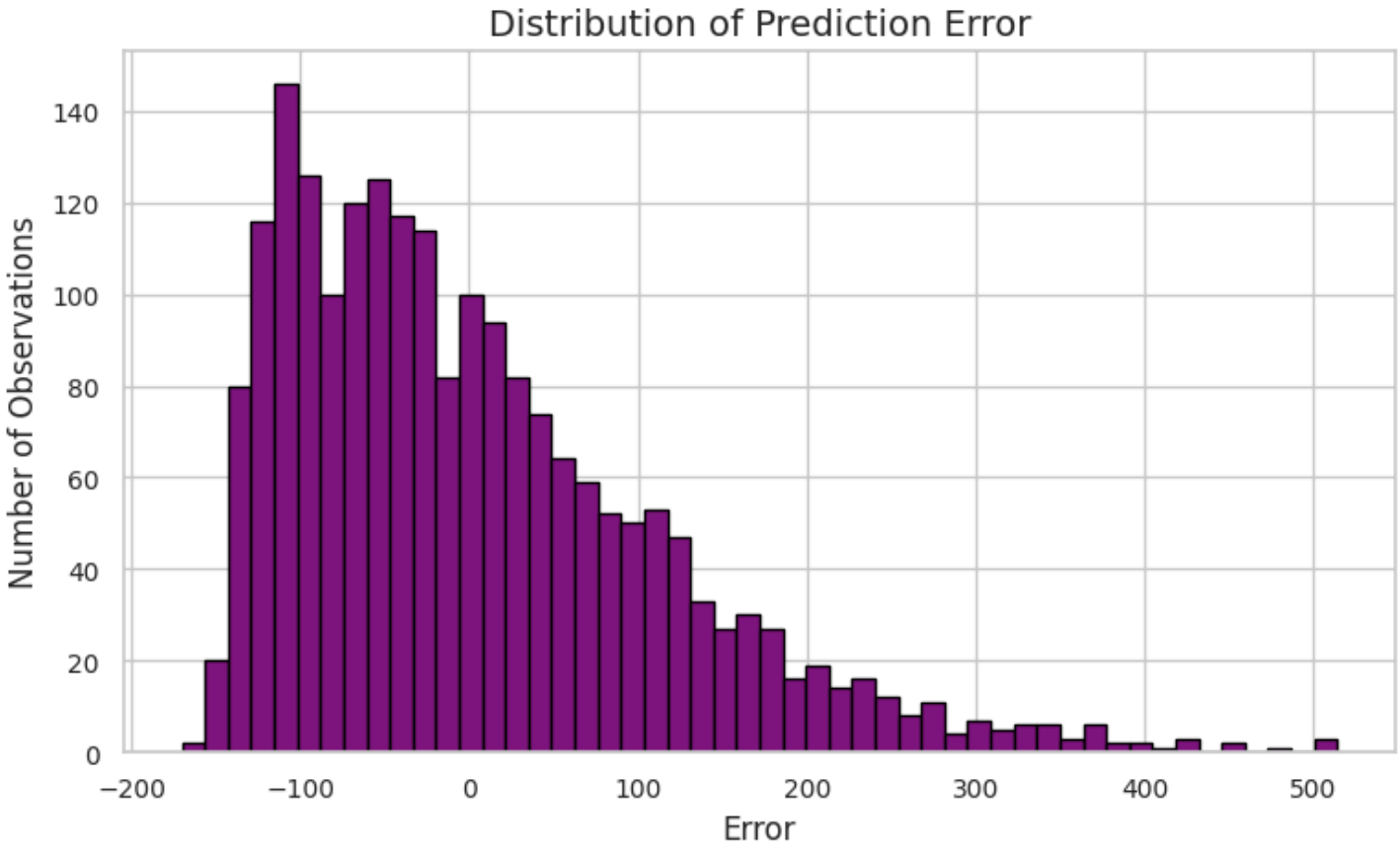


As shown in the plot, there is a small number of deliveries completed within a single minute, followed by a significant increase in deliveries taking exactly 2 minutes. After this peak, the number of deliveries gradually decreases as the duration increases. This pattern suggests that most deliveries are completed quickly, with a long tail of longer delivery times.

Prediction Error Histogram

To evaluate the accuracy of the current prediction model, I calculated the **prediction error** for each delivery by subtracting the predicted duration (planned_delivery_duration) from the actual delivery time (computed as segment_end_time - segment_start_time for STOP segments).

The resulting values represent the difference between actual and predicted times—positive values indicate underestimation (delivery took longer than predicted), while negative values indicate overestimation.

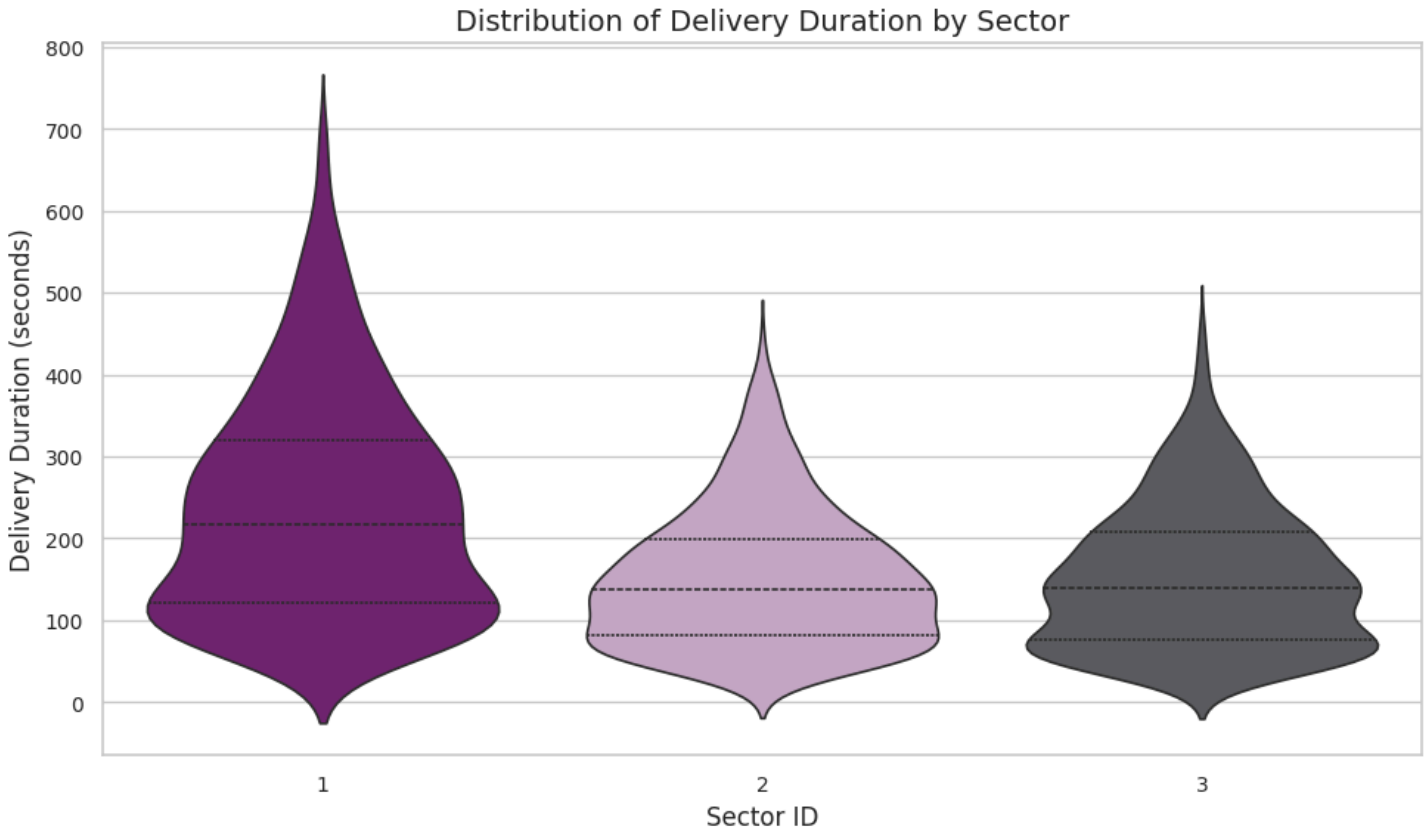


The histogram shows that the prediction model both overestimates and underestimates delivery times. There is a small peak in negative values, meaning some deliveries were quicker than predicted. However, the positive side (to the right of zero) has a longer tail, showing that many deliveries took longer than expected. This suggests that the model often underestimates the actual delivery time.

Sector-Based Delivery Duration

We wanted to check if calculating the average delivery time separately for each sector would be more accurate than using one global average for all deliveries. To do this, I grouped the data by sector_id and calculated the average delivery duration for each group.

To show the results, I used a **violin plot**. This type of chart shows how delivery times are spread out in each sector—it combines a boxplot with a shape that shows the distribution of values. Wider parts of the plot mean more deliveries with that time.



From the plot, we can see that sectors 2 and 3 have similar delivery time distributions. However, sector 1 looks different—it is more spread out and contains many deliveries that took longer. This confirms the idea that using separate averages for each sector, especially for sector 1, could improve prediction accuracy.

Additional Trends and Correlations

In this part, I explored the data further to find other patterns that could help improve delivery time predictions. I checked different factors like total weight, number of items in an order, and time of day to see if they had any visible impact on delivery duration.

Does the Driver Affect Delivery Duration?

To check if the person delivering the order has an impact on how long the delivery takes, I grouped the data by driver_id and plotted delivery times using a **violin plot**.

The chart shows that delivery times vary between drivers. Some drivers have many fast deliveries with a tight distribution, while others have a wider spread with more long deliveries. This suggests that individual drivers may influence delivery duration, and it could be useful to consider the driver in future prediction models.

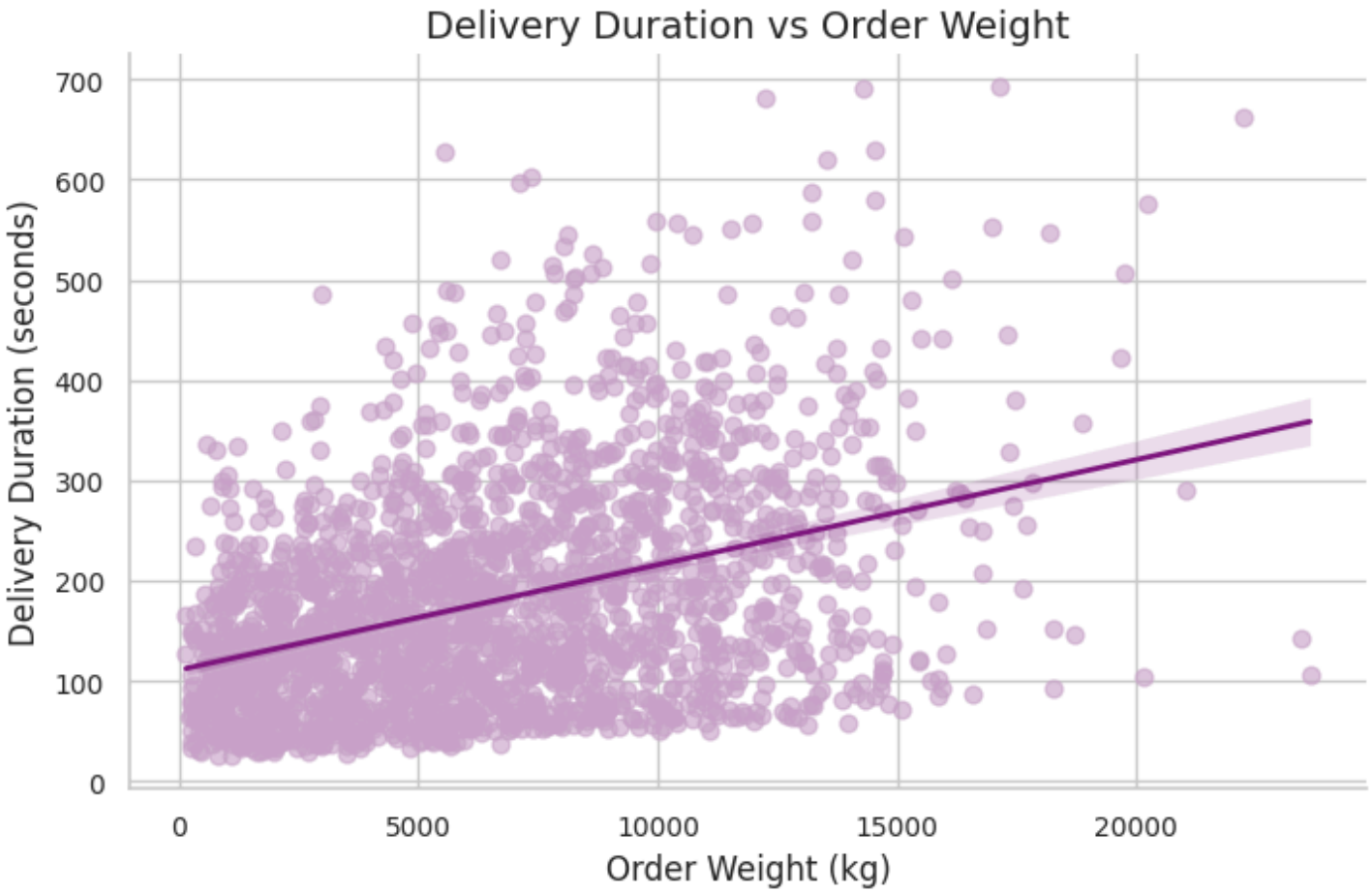
Does the Number of Products in an Order Affect Delivery Duration?

To explore whether the number of products in an order influences delivery time, I created **box plots** for different quantities of products in an order.

From the box plots, we can see noticeable differences in delivery durations based on the number of products in an order. Orders with more products tend to have a wider range of delivery times, while smaller orders have more consistent delivery durations. This observation suggests that the number of products might influence delivery time, and considering it in the prediction model could improve accuracy.

Does the Weight of an Order Affect Delivery Duration?

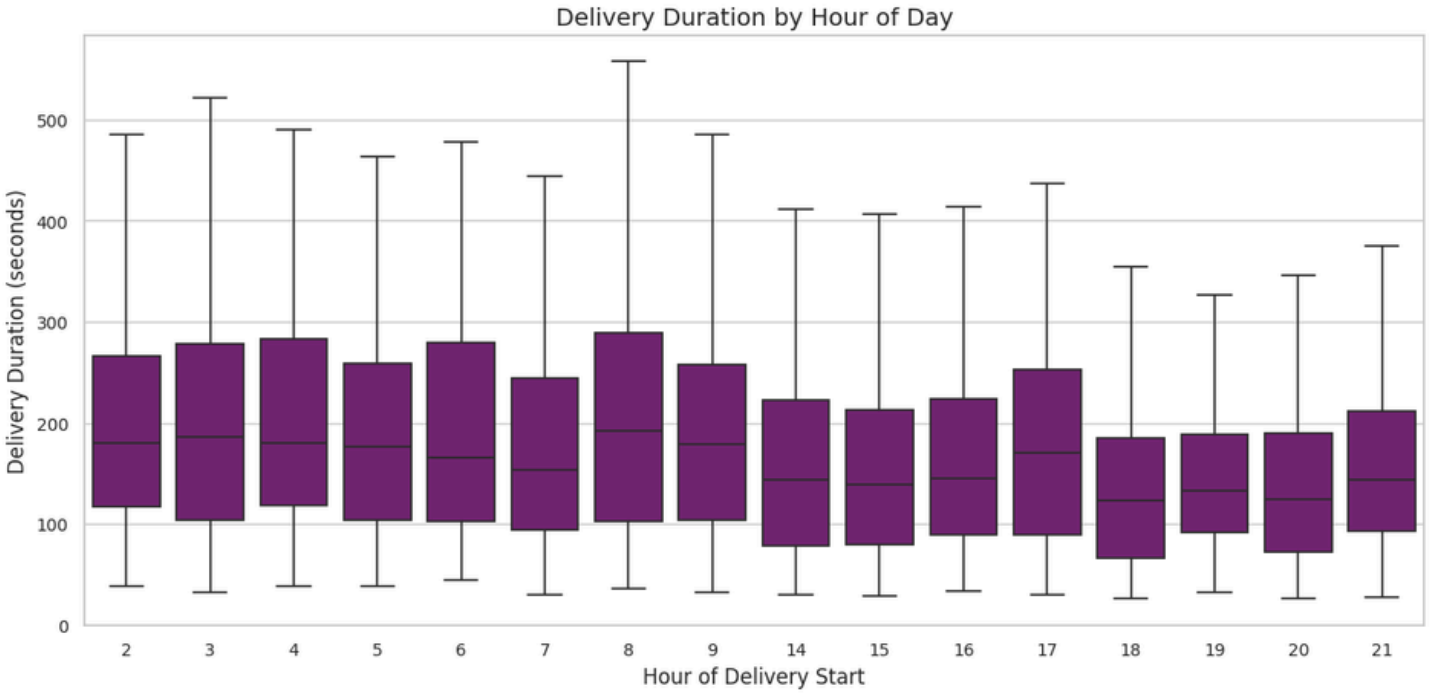
To investigate if the weight of an order impacts the delivery time, I used an **lmplot**, plotting the order weight on the x-axis and the delivery duration (in seconds) on the y-axis, along with a regression line.



The plot shows a slight upward trend, indicating that as the weight of an order increases, the delivery time also tends to increase. While the relationship is not very strong, it suggests that heavier orders might take longer to deliver, and incorporating weight as a factor in the prediction model could potentially improve its accuracy.

Does the Time of Day Affect Delivery Duration?

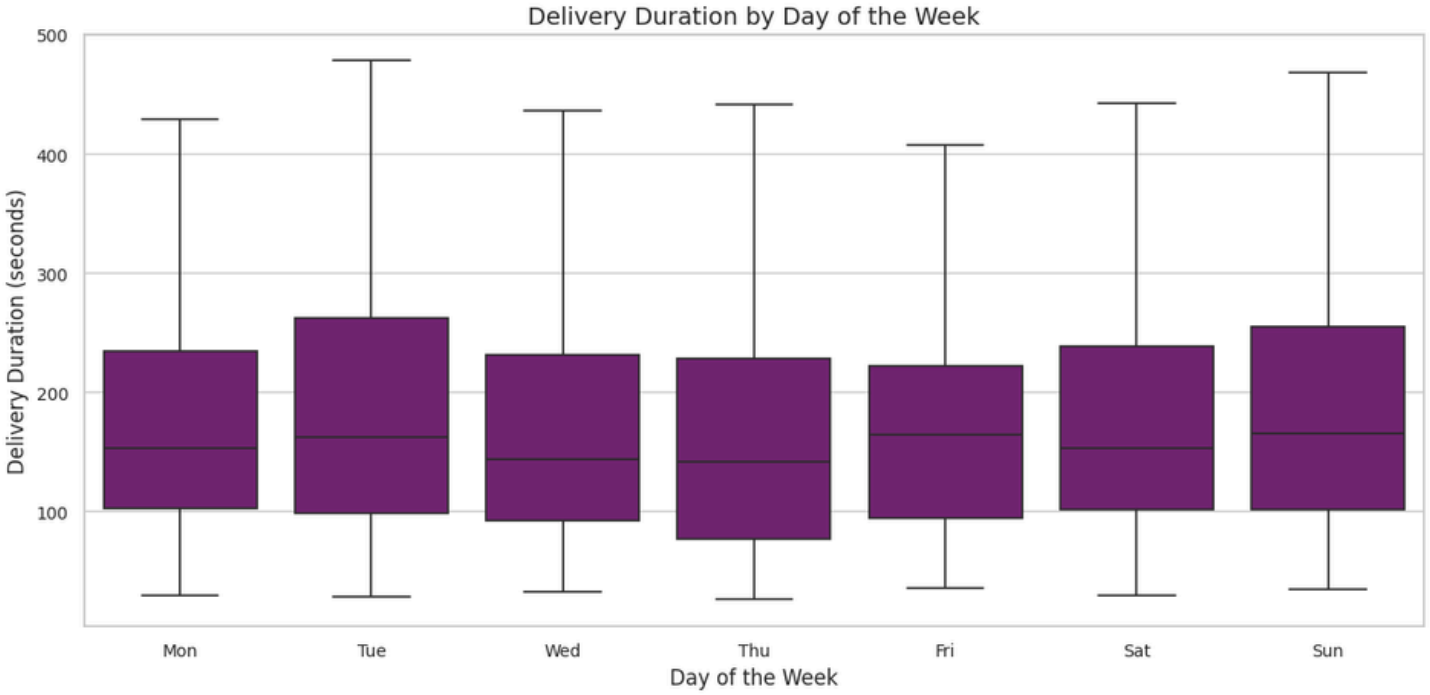
To check if the time of day has an impact on delivery times, I created **box plots** for different hours of day.



The box plots show some differences in delivery durations depending on the time of day, but the variations are relatively small. While there are slight changes, the data doesn't suggest a significant impact of time of day on delivery time. This could mean that, while time of day may have some influence, it may not be a strong predictor for delivery duration.

Does the Day of the Week Affect Delivery Duration?

To check if the day of the week affects delivery times, I created **box plots** for each day.



The box plots reveal some differences in delivery durations across different days of the week, but again, the variations are not very large. While certain days show slightly longer or shorter delivery times, the differences aren't significant enough to suggest a strong correlation between the day of the week and delivery duration.

Recommendations

This analysis provided valuable insights that can be further explored and tested in new predictive models. It appears that certain factors, such as the weight of an order, the number of products in the order, and potentially information about the driver, may influence delivery durations. These variables could be incorporated into future models to improve the accuracy of delivery time predictions. Further testing and refinement of these factors would help optimize the prediction algorithm and lead to better planning and resource allocation.

SZCZEPAN POLAK

2025