Monika_Juzek_analysis

May 11, 2025

1 Delivery Time Analysis

Introduction

Predicting accurate delivery times is key to running our grocery service well. Our current system just uses one average time for everything. But, we think things like where we're delivering (houses vs. apartments) matter.

This report looks at our past delivery data (orders, products, routes) to find patterns that explain why delivery times change. I want to see if I can find trends similar to the house vs. apartment idea, even without that specific info.

What I learn here will help us build smarter prediction tools, so I can give better delivery estimates.

The full code used for this analysis is available in jupyter notebook in the github repository: https://github.com/mjuzek22/Monika_Juzek_Data_Krakow.

1.1 Data Overview

I am using four tables from our MySQL database:

- `orders`: Order details (ID, customer, sector, planned time).
- `products`: Product information (ID, weight).
- `orders_products`: Items in each order (order ID, product ID, quantity).
- `route_segments`: Delivery route steps (driver, type, order ID, start/end times).

1.2 Calculating Delivery Time

To find out how long deliveries really take, I am looking at the 'STOP' sections in the `route_segments` data. For each delivery (order ID), I will figure out the time spent at the stop(s). If an order has more than one stop listed, I will need to figure out the best way to combine those times to get the total delivery time. I will start by looking at orders with multiple stops to understand what is going on.

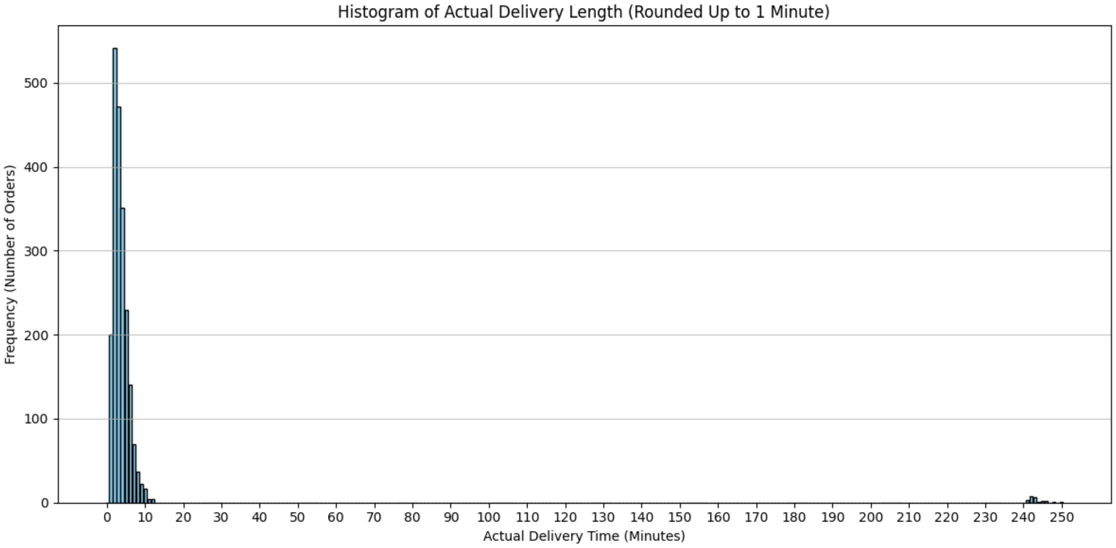
```
Number of orders with more than one STOP segment: 17
Order IDs with multiple STOP segments:
[158.0, 234.0, 244.0, 502.0, 513.0, 647.0, 788.0, 888.0, 989.0, 1297.0, 1529.0,
1551.0, 1699.0, 1822.0, 1896.0, 2029.0, 2112.0]

Let's look at a few examples to understand why this might be happening:
      segment_id  driver_id segment_type  order_id  segment_start_time  \
677          677          4         STOP     158.0 2024-02-24 14:44:13
```

```
680          680          4         STOP     158.0 2024-02-24 14:39:53
2620        2620          1         STOP     234.0 2024-02-16 14:19:29
2622        2622          1         STOP     234.0 2024-02-16 14:19:29
2628        2628          4         STOP     502.0 2024-02-11 14:56:47
2630        2630          4         STOP     502.0 2024-02-11 14:56:47
3181        3181          4         STOP     244.0 2024-02-11 14:59:58
3183        3183          4         STOP     244.0 2024-02-11 14:59:58
4468        4468          4         STOP     513.0 2024-02-26 19:19:47
4470        4470          4         STOP     513.0 2024-02-26 19:19:47

        segment_end_time  actual_delivery_duration
677  2024-02-24 14:39:53                    -260.0
680  2024-02-24 14:44:13                     260.0
2620 2024-02-16 14:20:25                      56.0
2622 2024-02-16 14:20:25                      56.0
2628 2024-02-11 15:01:40                     293.0
2630 2024-02-11 15:01:40                     293.0
3181 2024-02-11 15:02:38                     160.0
3183 2024-02-11 15:02:38                     160.0
4468 2024-02-26 19:24:40                     293.0
4470 2024-02-26 19:24:40                     293.0

First few rows of the orders data with calculated actual delivery duration:
   order_id  customer_id  sector_id  planned_delivery_duration  \
0         0          116          1                        176
1         1          160          1                        169
2         2           94          2                        177
3         3          165          3                        176
4         4           78          2                        177

   actual_delivery_duration
0                     309.0
1                      69.0
2                     201.0
3                      92.0
4                      81.0
```

My investigation into orders with multiple 'STOP' segments revealed data quality issues, including negative durations (likely due to timestamp errors) and segments with zero duration (possibly duplicates or very short stops). The revised plan to calculate the actual delivery duration is:

1. **Filter for 'STOP' segments.**
2. **Convert timestamps to datetime objects.**
3. **Calculate duration for each segment and discard those with non-positive durations.**
4. **For each order, sum the durations of all its remaining (positive) 'STOP' segments.**
5. **Merge this total duration with the `orders_df`.**

This approach aims to handle identified data errors while still assuming that all valid 'STOP' segments for an order contribute to the overall delivery time.



My analysis of actual delivery times revealed a small number of deliveries with exceptionally long durations (around 4 hours), significantly exceeding the typical delivery time of under 15 minutes.

These outliers could stem from:

- **Data Errors:** Incorrect timestamp recordings.
- **Unusual Circumstances:** Rare and extreme delivery events.
- **Misinterpretation of 'STOP' Segment:** The 'STOP' segment might have captured a much longer period for these specific orders.

Impact: These extreme values can distort my statistical analysis and negatively affect predictive models.

Next Steps:

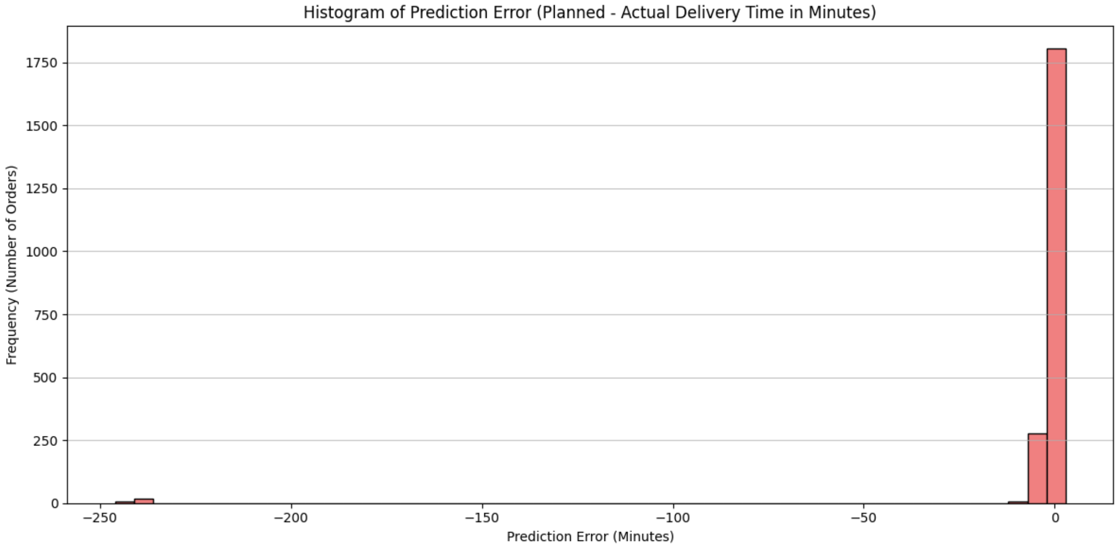
For my report and further analysis:

- **Acknowledge and Visualize:** I will mention and potentially visualize these outliers.
- **Consider Handling:** For the prediction error histogram, I might exclude them for clarity. For correlation analysis and modeling, I will need to consider their impact and choose appropriate techniques.

I will now proceed to generate the histogram of the prediction error, keeping in mind the potential influence of these delivery time outliers.

1.3 Distribution of Prediction Error

To understand how well our current prediction system performs, I calculated the prediction error for each order by subtracting the actual delivery duration from the planned delivery duration. I then generated a histogram to visualize the distribution of these errors.

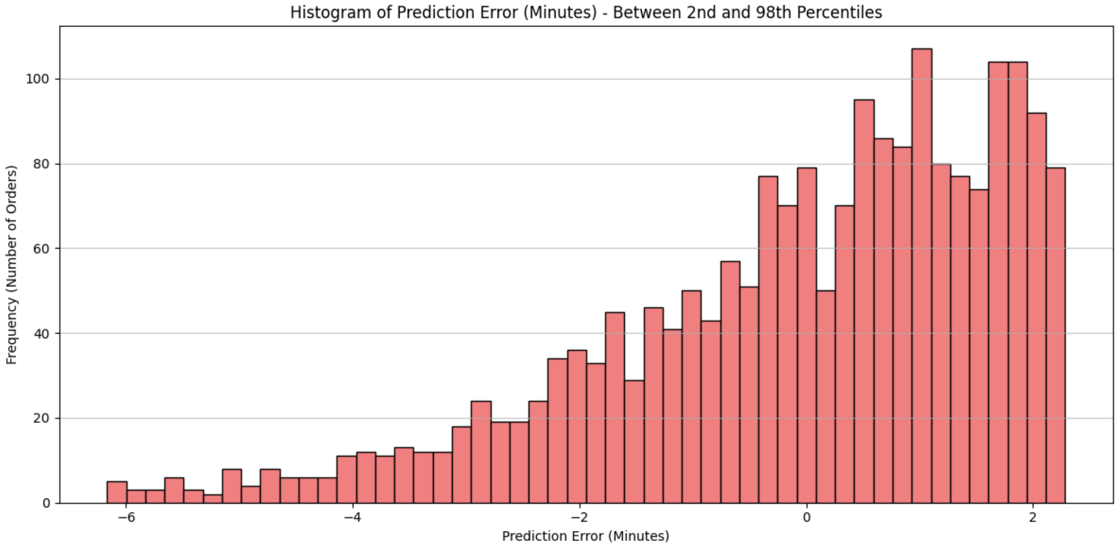


```
Descriptive Statistics of Prediction Error (Minutes):
count    2111.000000
mean       -2.797987
std        25.510344
min      -246.100000
25%        -1.150000
50%         0.350000
75%         1.366667
max         2.833333
Name: prediction_error_minutes, dtype: float64
```

The prediction error (planned - actual delivery time) for 2111 orders reveals that, on average, our current system overestimates delivery time by approximately 2.8 minutes. However, the high standard deviation (around 25.5 minutes) indicates significant variability and imprecision in these predictions.

Notably, the minimum prediction error is around -246 minutes, directly reflecting the exceptionally long actual delivery times I previously identified as outliers. These outliers skew the overall error distribution and statistics. The majority of the prediction errors are likely clustered closer to zero, but the large negative outliers pull the mean downwards.

To better visualize the typical prediction error, I will focus on the bulk of the data by setting limits based on percentiles (e.g., the 1st to 99th percentile) to exclude these extreme outliers from the histogram. This will provide a clearer picture of the prediction error distribution for the majority of our deliveries. The substantial mean error and high standard deviation confirm that the current single-average prediction method is not very effective.



```
Prediction error range for the histogram: [-6.22, 2.28] minutes
```

Looking at the prediction error for most deliveries (excluding outliers), I see:

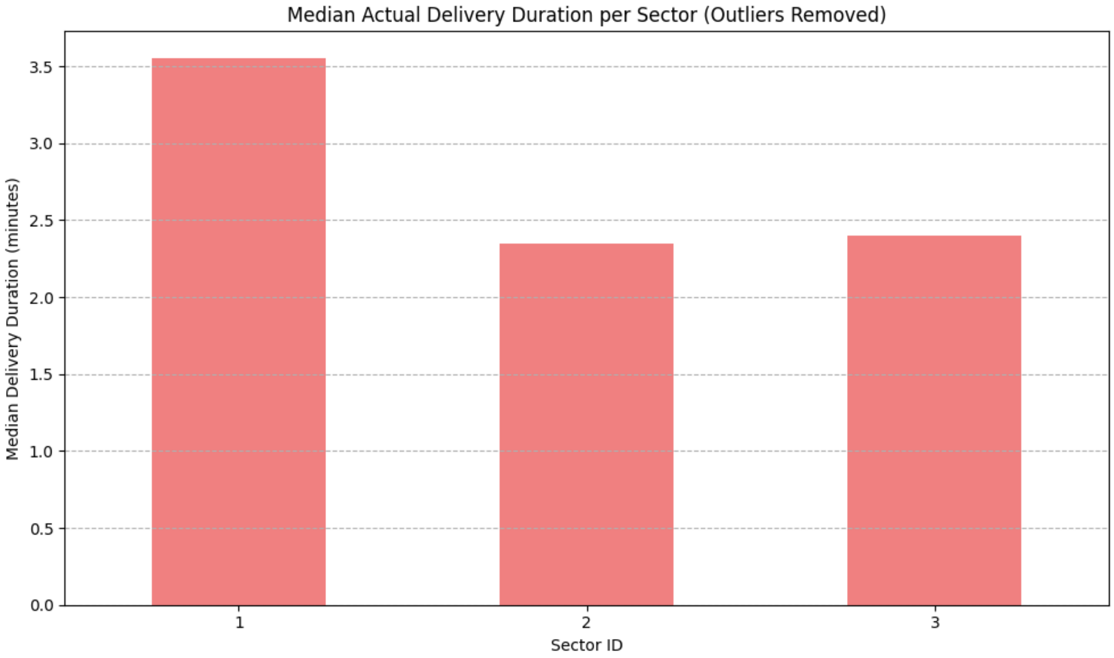
- **Slight Underestimation:** Planned times tend to be a bit shorter than actual times.
- **Still Variable:** Even for typical deliveries, the error can vary by several minutes.
- **More Often Under Than Over:** I am slightly more likely to underestimate.

This shows my current predictions aren't very accurate for most orders.

1.4 Delivery Time Differences Across Sectors

To investigate if delivery times vary by geographic area, I grouped the data by `sector_id` and calculated the median actual delivery duration for each sector. I then visualized these differences using a bar chart.

```
Number of valid deliveries before outlier filtering: 2111
Number of valid deliveries after outlier filtering (2nd to 98th percentile):
2026
```



```
Median and Standard Deviation of Actual Delivery Duration per Sector (Outliers
Removed):
            Median_Duration(min)  Std_Dev_Duration(min)
sector_id
1                           3.55               2.055833
2                           2.35               1.390543
3                           2.40               1.428831
```

Visualizing delivery times by sector shows clear differences:

- **Sector 1:** Takes longer on average, with some deliveries taking significantly longer than in other sectors.
- **Sector 2:** Generally has quicker deliveries with less variation in the typical delivery time.
- **Sector 3:** Also tends to be faster than Sector 1, with a similar spread to Sector 2.

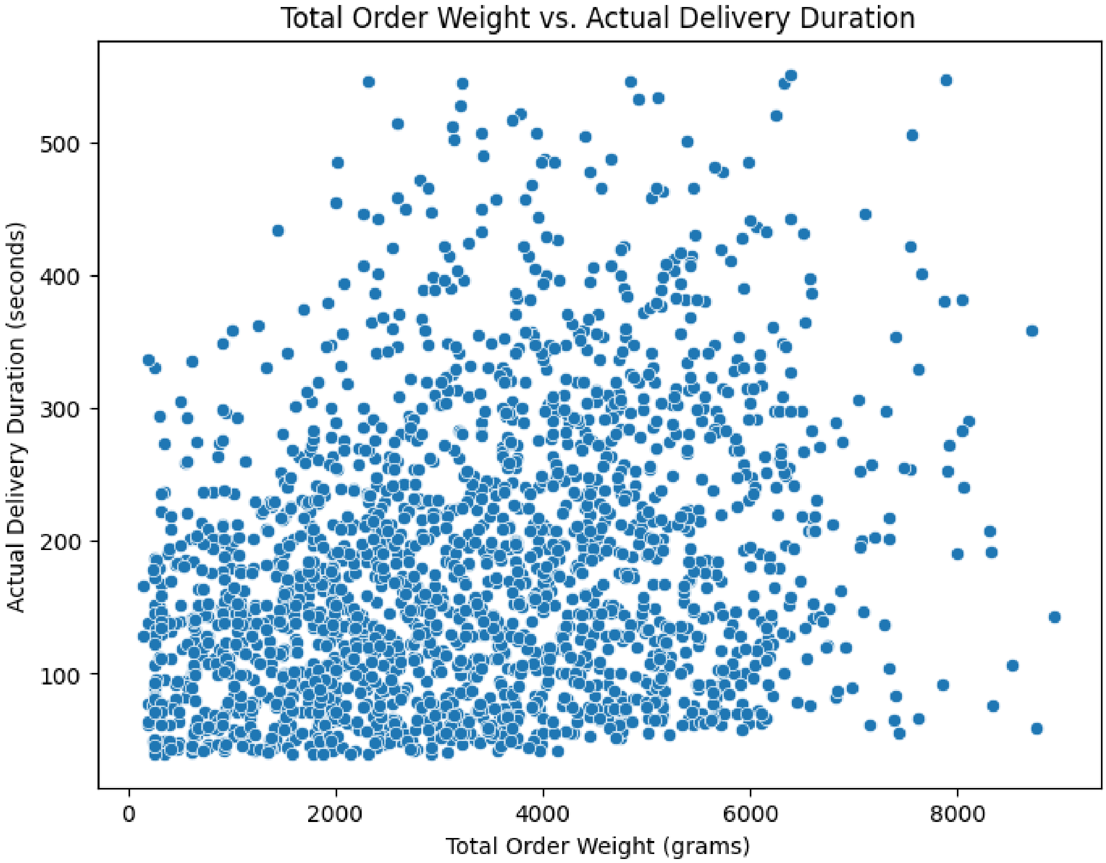
This suggests that location (sector) does indeed impact delivery time, supporting the drivers' observations that not all areas are the same.

1.5 What additional data would be worth collecting?

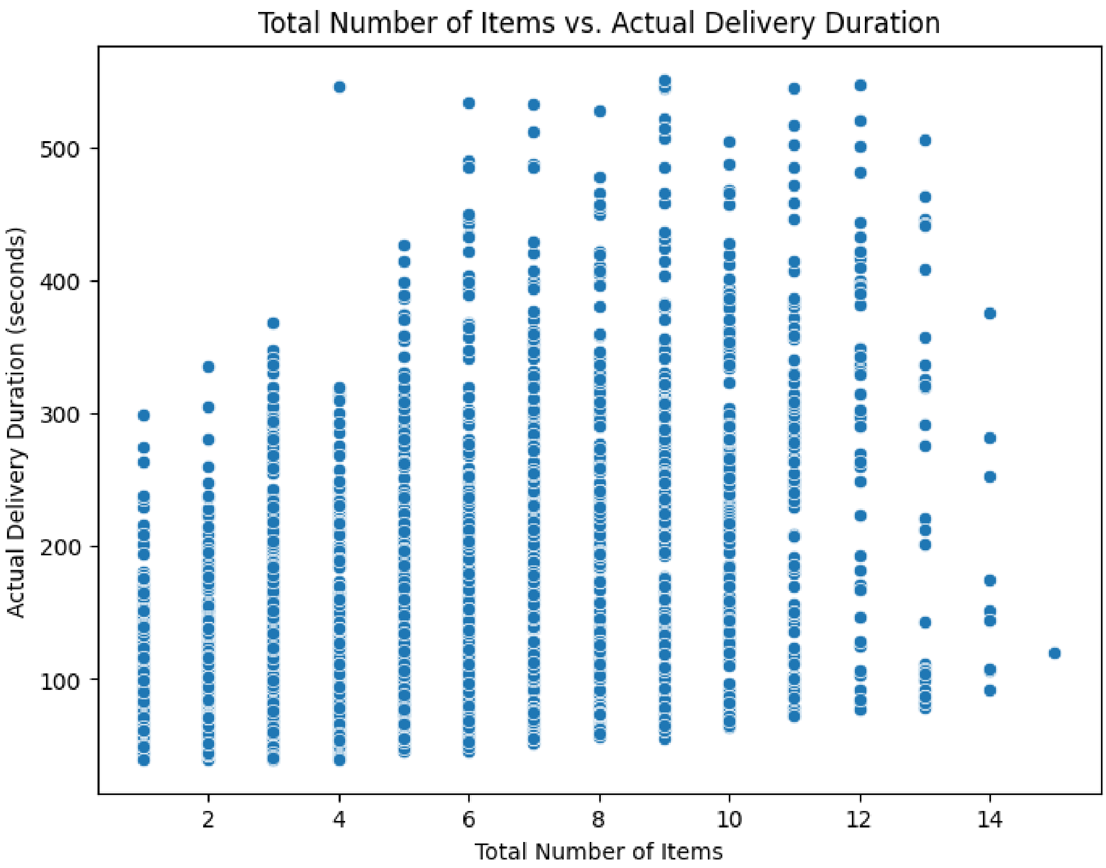
4.1. I can investigate if what's in the order affects delivery time. I'll look at:

- **Total weight:** How heavy the order is.
- **Total items:** How many items are in the order.
- **Unique products:** How many different types of items are in the order.

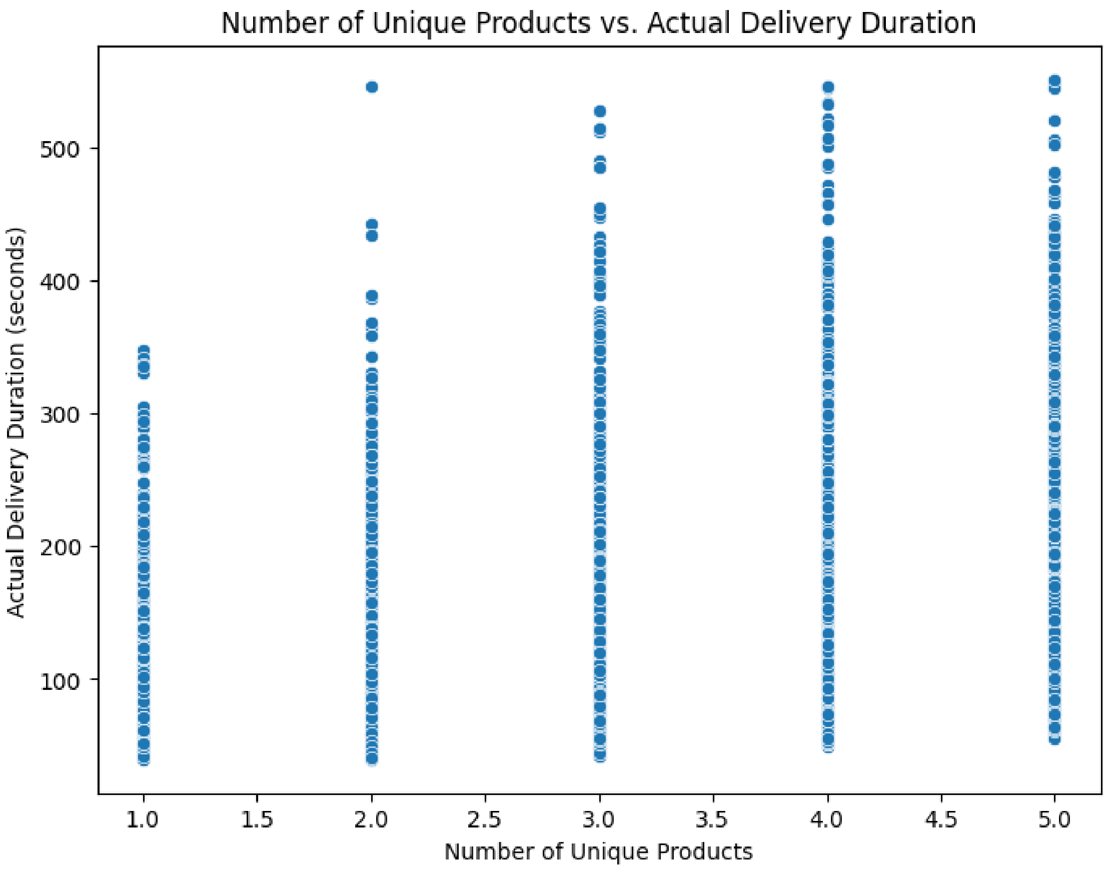
By combining order and product information, I can create charts (scatter plots or hexbin plots) to see if heavier orders or those with more items tend to take longer to deliver, likely due to handling.



```
Pearson correlation coefficient between Total Order Weight and Actual Delivery
Duration: 0.277
```



```
Pearson correlation coefficient between Total Number of Items and Actual
Delivery Duration: 0.380
```



```
Pearson correlation coefficient between Count of Distinct product_id and Actual
Delivery Duration: 0.321
```

My analysis of order weight, total number of items, and the number of unique products reveals only weak to moderate positive correlations with delivery time.

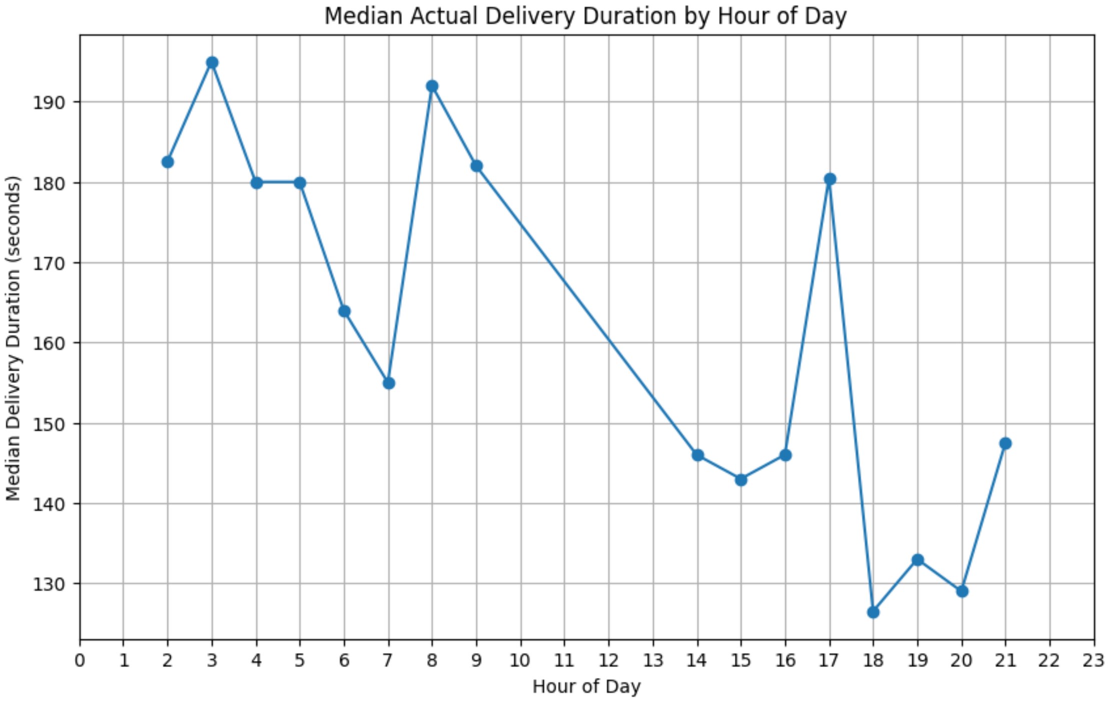
- **Weight:** Shows a negligible linear relationship with how long deliveries take. Heavier orders don't consistently mean longer deliveries.
- **Number of Items:** Has a slightly stronger positive link (correlation of 0.38), suggesting more items tend to lead to slightly longer deliveries, but it's not a strong predictor on its own.
- **Unique Products:** Also shows a weak positive correlation (0.321). Orders with more variety of items tend to take a bit longer, but this isn't a major factor.

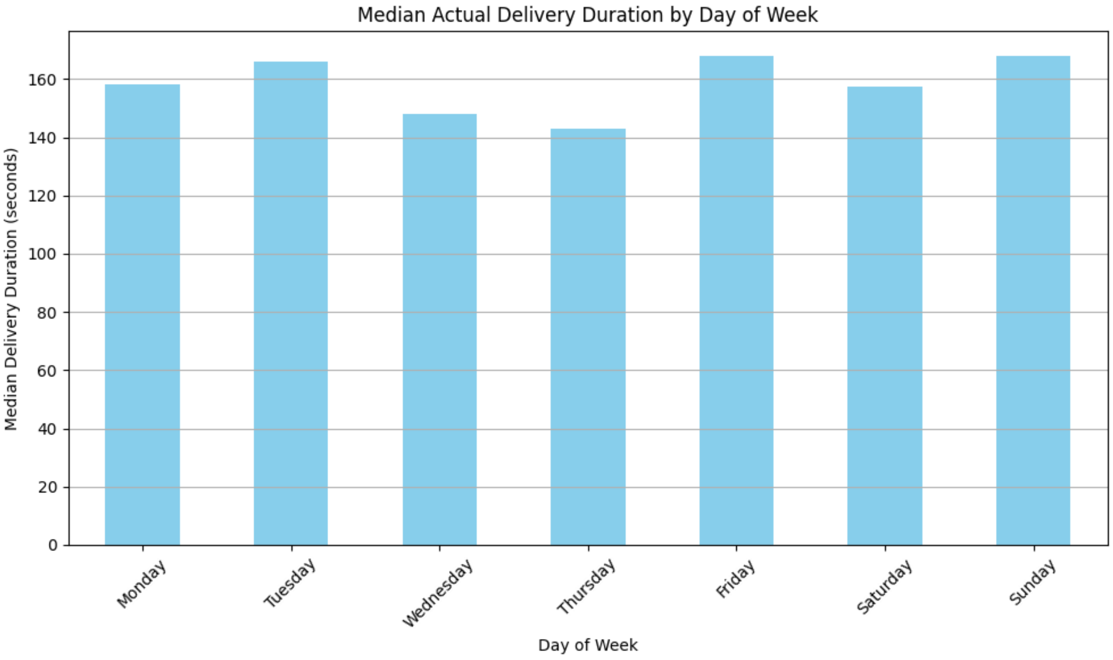
In short, what's in the order seems to have a limited direct impact on delivery time compared to other factors I've explored like location and time.

4.2. I also looked at whether *when* a delivery happens matters. I can analyze delivery times based on:

- **Hour of the day:** To see if rush hour or other times lead to longer deliveries.
- **Day of the week:** To check for weekday vs. weekend differences.

By looking at average or median delivery times for each hour and day, I can spot potential trends related to traffic or order volume.

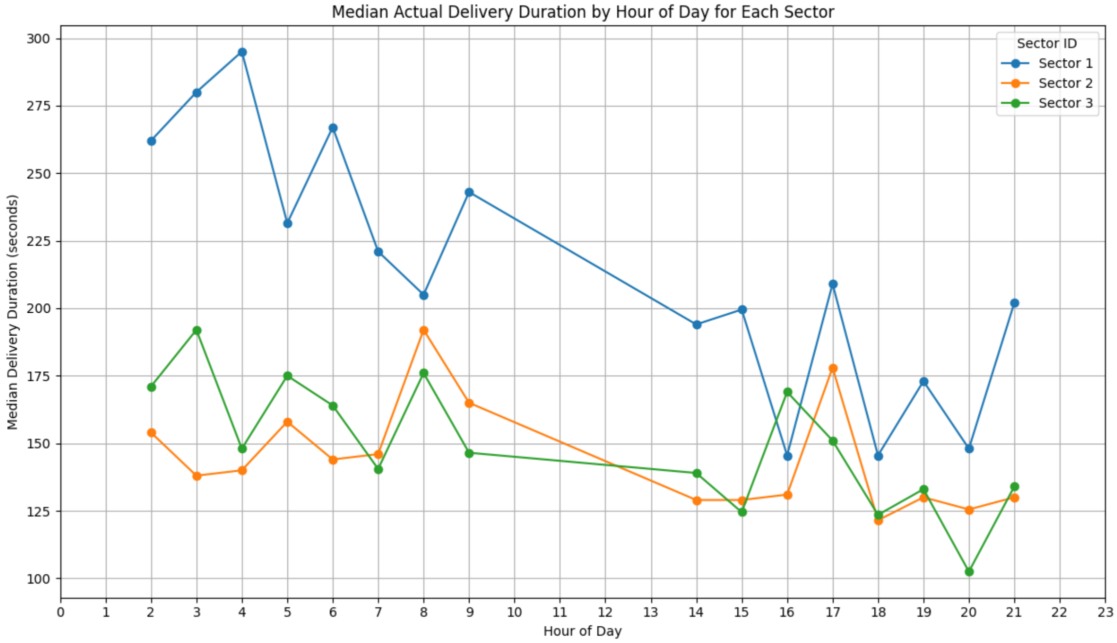




Looking at delivery times by the hour, I see a clear pattern:

- **Peaks:** Deliveries tend to take longer in the early morning and again in the late afternoon (around rush hour).
- **Dips:** Midday and late night/early morning see the fastest delivery times.

This strong link between the time of day and delivery duration means we should definitely consider the hour of delivery when making predictions. The relationship isn't simple, so I might need more advanced prediction methods to capture these ups and downs. I can explore further how these hourly trends interact with other factors such as the delivery sector or the day of the week.



When I look at how delivery times change throughout the day *within each sector*, I see different patterns:

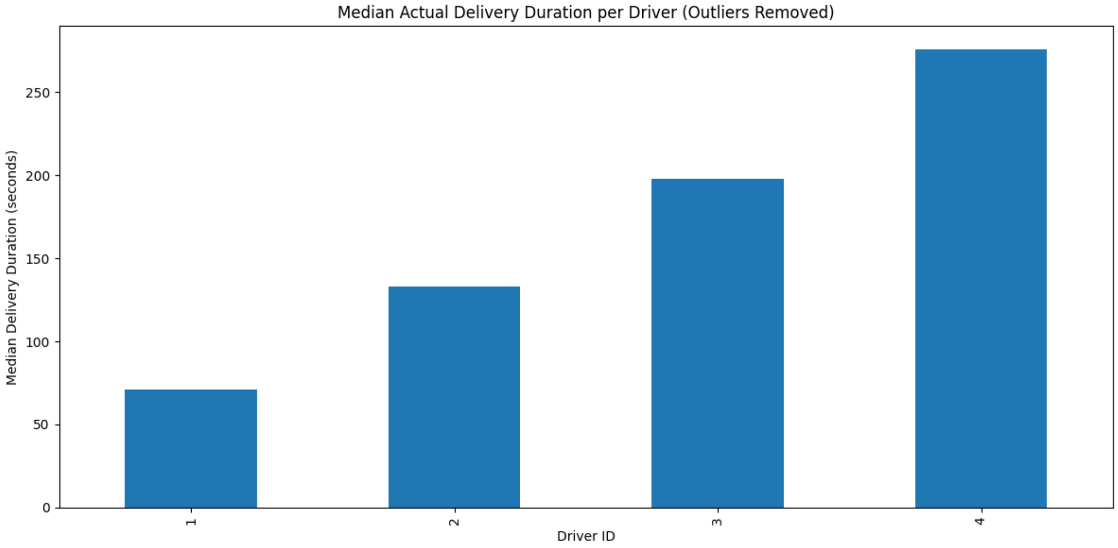
- **Sector 1:** Has big swings, with longer deliveries in the early morning and late afternoon/early evening.
- **Sector 2:** Is more stable, with a peak around morning commute time.
- **Sector 3:** Is slowest very early in the morning but generally the fastest during the day.

This shows that the best prediction will consider *both* the time of day and the delivery location (sector). For further analysis it would be beneficial to check causes of these unique patterns in each area (traffic, order volume, building types etc.).

4.3 Do Some Drivers Deliver Faster Than Others?

I also explored if individual drivers have different average delivery times. By looking at the median delivery time for each driver, I can see if some are consistently quicker or slower than others.

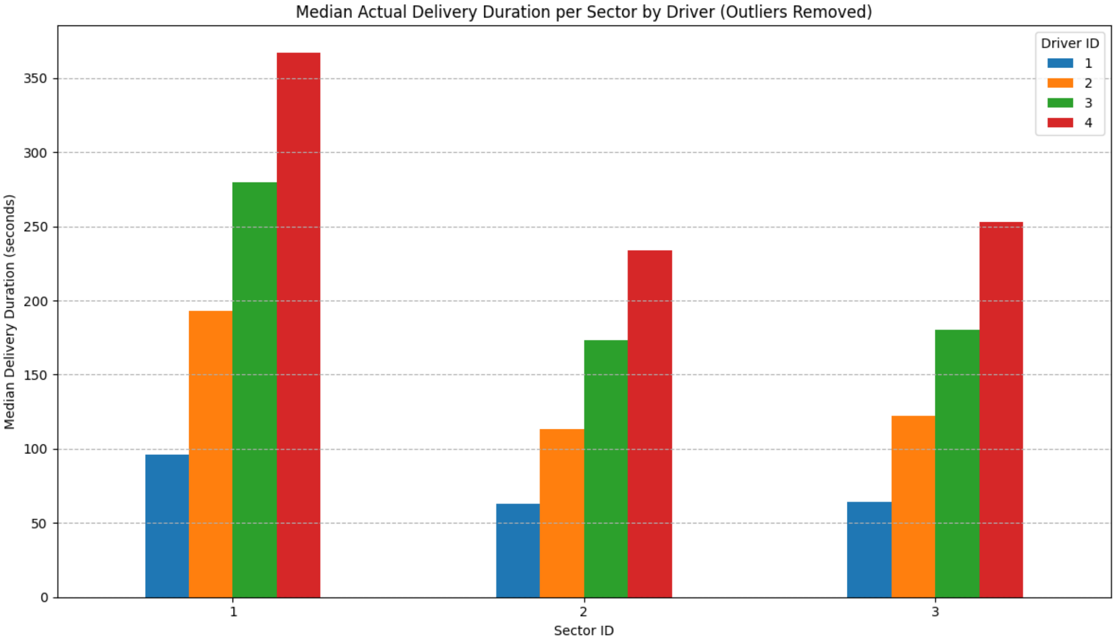
I need to keep in mind that differences in driver speeds might not just be about the driver themselves. They could be influenced by the areas they usually deliver to or the types of orders they handle most often.



Looking at the median delivery times for different drivers, I see clear differences. Some drivers consistently complete deliveries faster than others.

This suggests that *who* is making the delivery is another important factor in how long it takes. To make better predictions, I should consider the driver's ID. However, I need to be careful – faster or slower times could be due to the areas they usually work in or the types of deliveries they handle. Understanding these reasons could help me improve the predictions and even our operations.

```
Number of deliveries per driver per sector:
     driver_id  sector_id  delivery_count
0            1          1             190
1            1          2             133
2            1          3             173
3            2          1             167
4            2          2             171
5            2          3             156
6            3          1             167
7            3          2             185
8            3          3             193
9            4          1             149
10           4          2             178
11           4          3             179
```



When I compare driver speeds *within* each delivery area (sector), I see interesting patterns:

- **Consistent Fast/Slow Drivers:** Some drivers (like Driver 1) tend to be faster across all sectors, while others (like Driver 4) are consistently slower. This reinforces that the driver matters.
- **Sectors Have Different Difficulties:** Deliveries in Sector 1, as seen earlier, generally take longer for all drivers, suggesting this area might be inherently more challenging.
- **Driver Performance Varies by Sector:** While some drivers are generally faster, their advantage might be bigger in certain areas.

This means to really nail my predictions, I need to consider not just *who* is delivering, but *where* they are delivering. The prediction tools should probably learn these driver-location combinations.