Data analysis and visualisation report for Delivery Time Predictions

Actual Delivery Length:

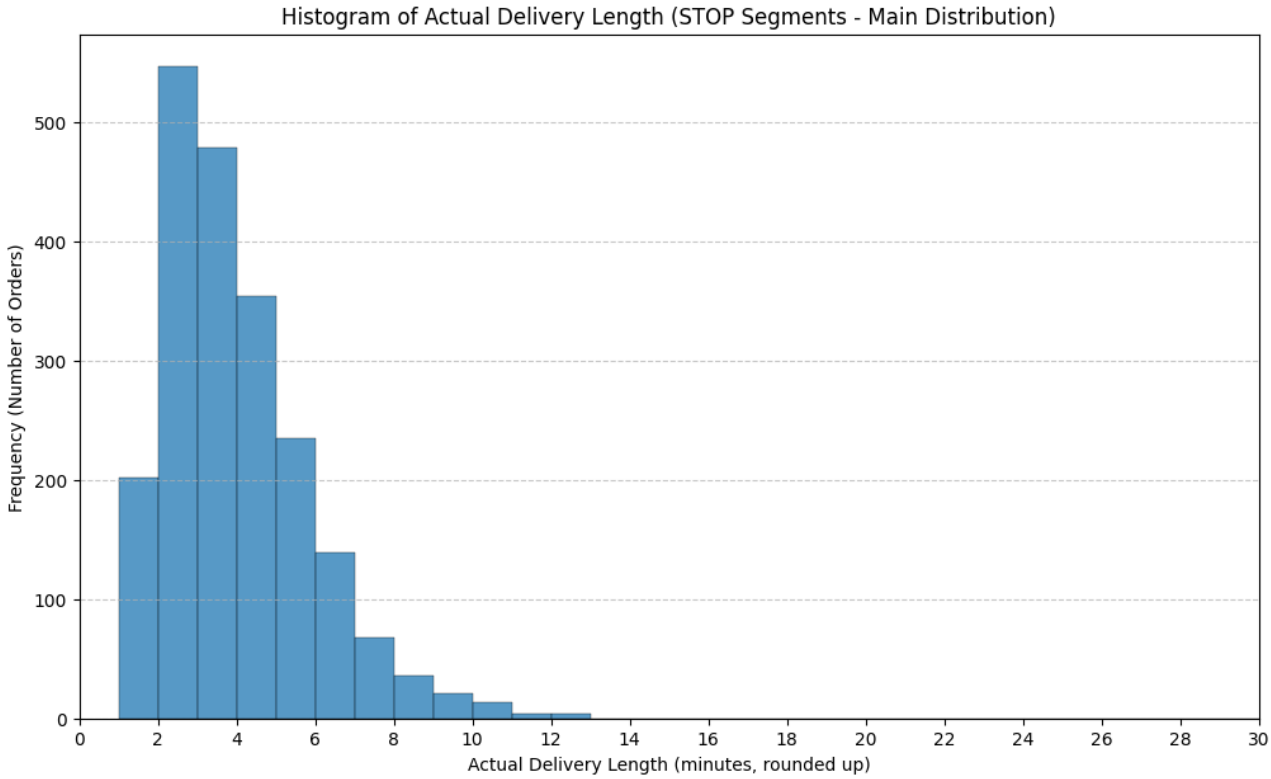
To make a sensible analysis I decided to start from visualisation.

Here, we can see data has extreme outliers (like the 250-minute delivery), and the histogram seems to be squished and hard to read. However, it is visible an additional 0 at the x-axis. It means there is a negative delivery time, which can not be true.

Therefore I decided to filter the data:

- Removed negative durations and zero-second durations
- Added checks for corrupted time_segment rows. It will prevent errors in duration calculation

Given how skewed the data is, I decided to Create a version of the histogram that focuses on the majority of the data by setting an x-axis limit.



Filtered Statistics:

- **count:** 2125: The sampling size after filtering non-positive delivery durations.
- **mean:** 6.20 minutes. The average delivery time
- **std:** 25.42 minutes: The standard deviation is very large compared to the mean, which is a strong indication of a skewed distribution and the presence of outliers (the long delivery times).
- **max:** 250.0 minutes: This is 250 / 60 \approx 4.17 hours, confirming the longest duration after rounding.

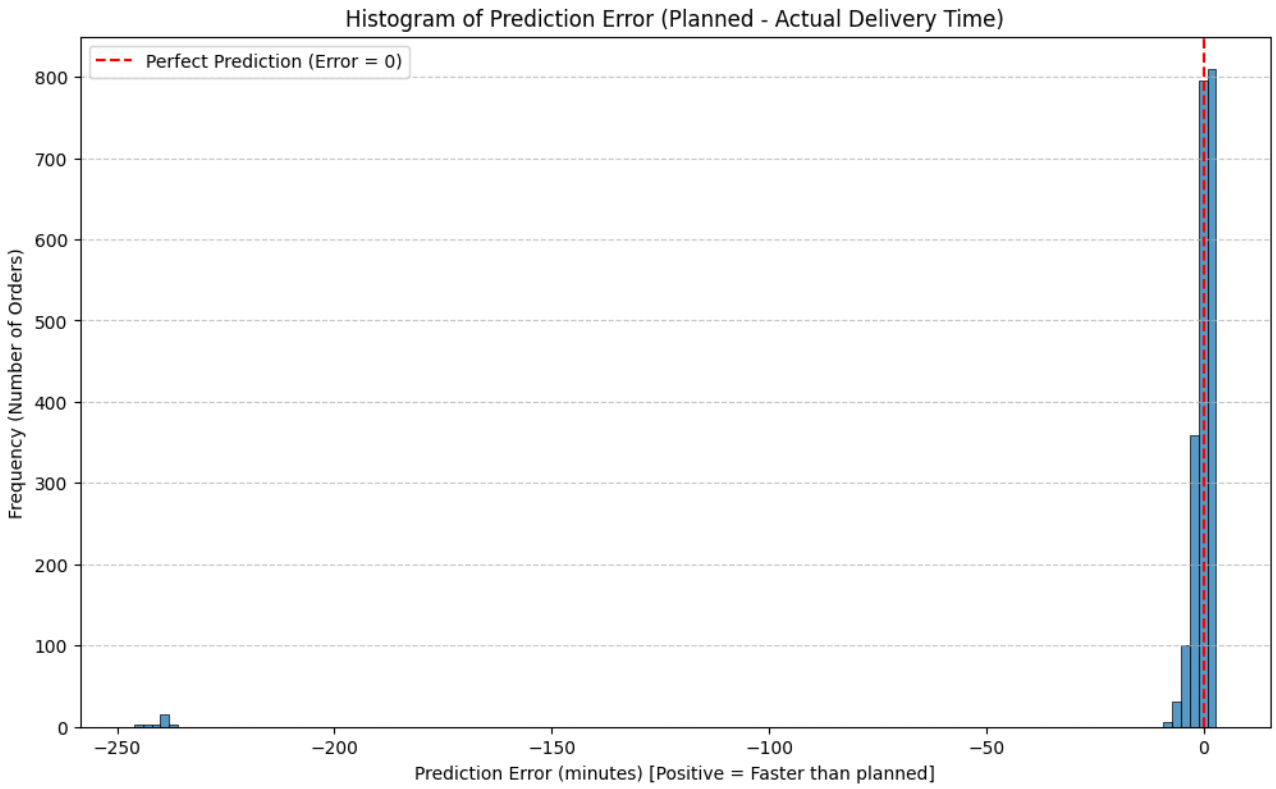
The histogram:

- The histogram visually confirms the statistics: most deliveries are very short (concentrated on the left, likely between 1-10 minutes).
- There's a tall bar at the very beginning, indicating a high frequency of short deliveries.
- There's a very long tail to the right due to the outliers, making the main distribution hard to see in detail. We can see a small bump near the 250-minute mark, representing those very long deliveries.

Prediction Error:

Now, we will analyse difference between planned and actual delivery times. The plan is to calculate the difference between planned delivery duration and actual delivery duration.

The planned delivery duration is taken from the “orders” table.



Statistics:

- mean:
  - Minutes: -2.76 minutes
  - A negative mean error indicates that, on average, the **actual delivery time is longer than the planned delivery time**. The current system is, on average, underestimating delivery times by about 2.76 minutes.
- std (Standard Deviation):
  - Minutes: 25.43 minutes
  - Similar to the actual delivery times, the standard deviation for the error is very large. It points to a significant variability.
- min:
  - Minutes: -246.10 minutes (approx. -4.1 hours)
  - This means the worst underestimation was an order that took about 4.1 hours *longer* than planned. This corresponds to one of the long actual deliveries combined with a much shorter planned time.
- max:
  - Minutes: 2.83 minutes

- The largest positive error means the biggest overestimation (delivery faster than planned) was by about 2.83 minutes. This is a very small value compared to the min error.

Percentiles (for minutes):

- **25%:** -1.13 minutes (25% of orders were delivered more than 1.13 minutes later than planned)
- **50% (median):** 0.37 minutes (The median error is slightly positive. This means **50% of orders were delivered up to 0.37 minutes *faster* than planned**. This is interesting because the mean is negative. The negative outliers are pulling the mean down).
- **75%:** 1.37 minutes (75% of orders were delivered no more than 1.37 minutes faster than planned).

The Histogram:

- **Shape:** The histogram is heavily skewed to the left (negative side), visually confirming the large negative errors (time underestimations).
- **Peak:** There's a very sharp peak slightly to the right of zero. This indicates that a large number of deliveries are actually completed a bit faster than planned or very close to the planned time.
- **Red Line (Error = 0):** This line shows that most of the "good" predictions are located around it.
- **Left Tail:** The long tail on the left (going down to -246 minutes) represents the significant underestimations where actual **delivery times were much longer than the simple mean prediction used.**
- **Right Tail:** The right tail is very short and truncated, indicating that **the system rarely overestimates delivery times by a large margin**. The maximum overestimation is only about 2.83 minutes.

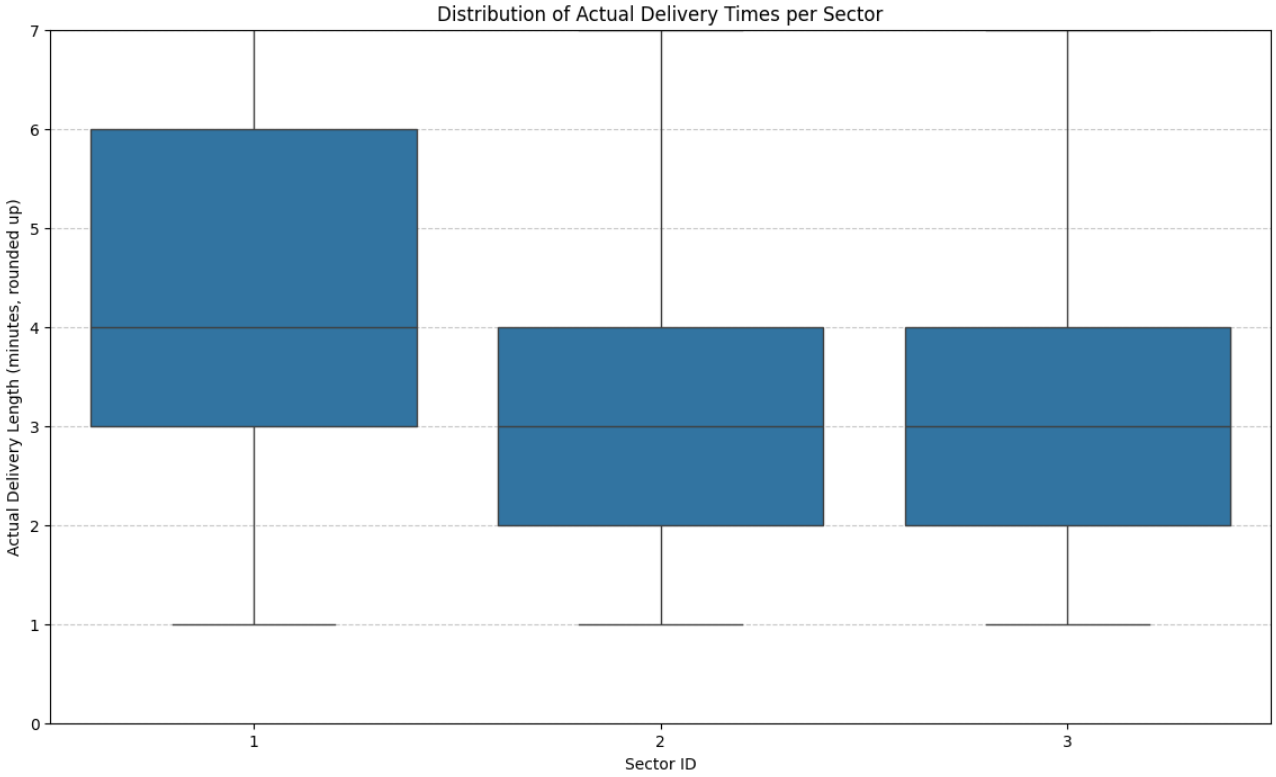
Summary analysis shows that while many deliveries are close to the planned time or a bit faster, when the system is wrong, it's often wrong by a large amount on the side of underestimation.

Abnormal sector theory:

"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."

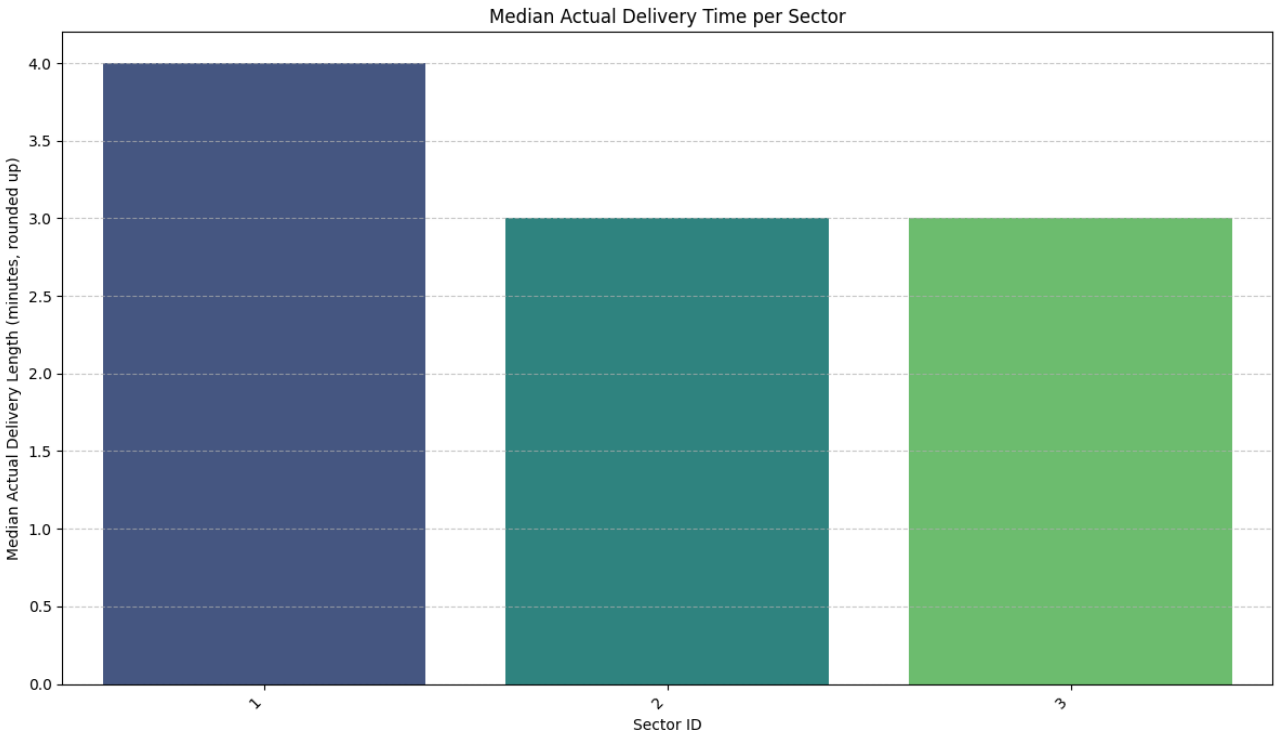
To visualize this, we need to:

1. Actual delivery times for each order. We already have it from previous step
2. Get the sector_id for each of these orders
3. Combine this information



Statistics per Sector:

- **Median Comparison:**
  - Sector 1 has a median delivery time of 4 minutes.
  - Sectors 2 and 3 have a median delivery time of 3 minutes.
  - This suggests that, based on the median, Sector 1 tends to have slightly longer delivery times than Sectors 2 and 3.
- **IQRs (Interquartile Ranges - the boxes):** The boxes for all sectors are quite low on the y-axis. This means that the bulk of the deliveries (the middle 50%) are completed in a short amount of time for all sectors.
- **Conclusion from Box Plot:** While Sector 1's median is slightly higher, the dominant feature is that *all* sectors suffer from extreme outliers, making their distributions highly skewed. The visual difference in the "bulk" of typical deliveries (the boxes) is not dramatically different when viewed on this scale.



Impact of Outliers: Crucially, the box plot reveals that all sectors experience very long delivery times (outliers up to ~250 minutes). While Sector 1's median is higher, the problem of extreme delays isn't unique to one sector. The *overall distribution* including these outliers makes it hard to say one is "dramatically" worse without further statistical testing, but the central tendency (median) is indeed higher for Sector 1.

Significance:

- Is a 1-minute difference in median "significant"? This depends on the business context. For a 3-4 minute typical delivery, a 1-minute increase is a 25-33% increase, which could be considered noteworthy.
- The driver's insight might be based on these slightly longer typical times in Sector 1, or perhaps they are more frequently exposed to the extreme outliers in that particular sector (though outliers appear in all).

Conclusion on Hypothesis:

- The data provides evidence that **Sector 1 tends to have slightly longer median and mean delivery times compared to Sectors 2 and 3.**
- However, all sectors exhibit significant variability with many extreme outliers. The issue of very long deliveries is widespread, not confined to a single sector, even if typical times vary slightly.