Part 2. Data Analysis and Visualization

1. Generate a histogram showing the actual delivery length with 1 minute granularity (rounded up).

The basic histogram looks like this:

I began by generating a histogram of actual delivery times, rounded up to the nearest minute. The initial visualization revealed a few questionable values—specifically, some delivery lengths below 0 and others exceeding 240 minutes. These seemed unrealistic, so I investigated further.

```
array([-16, -14,  -5,  -4,  -3,  -2,  -1, 241, 242, 243, 244, 245, 246,
       248, 250, 252, 253, 256])
```

There were no values between 20 and 240 minutes, suggesting 100 minutes was a reasonable upper cutoff.

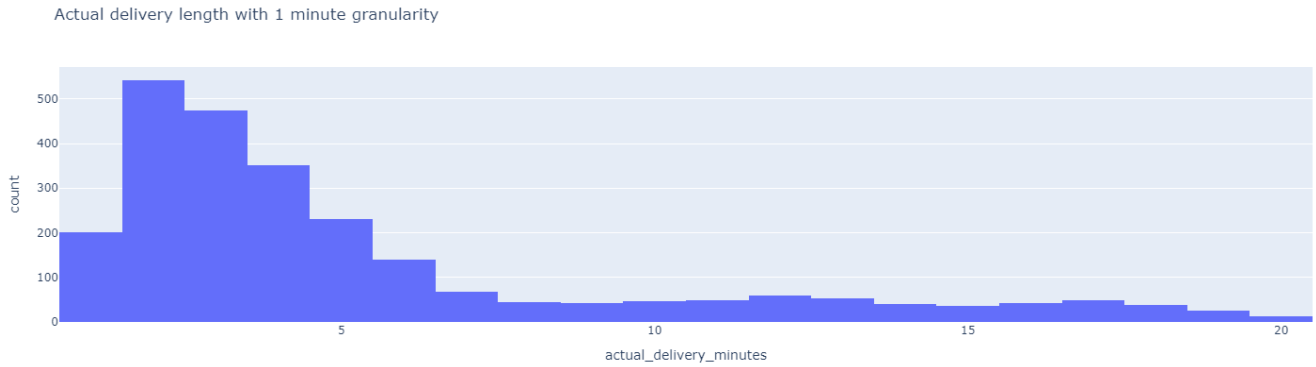
I also noticed some duplicate records—multiple segments with the same order ID. In a few cases, delivery times were negative or identical to the end time, which likely indicates data entry errors

Interestingly, 528 rows marked as "STOP" segments lacked an order ID. While they may be incomplete, I decided to keep them in the histogram as they might still hold valuable timing information.

	segment_id	driver_id	segment_type	order_id	segment_start_time	segment_end_time	actual_delivery_time	actual_delivery_minutes
6	6	1	STOP	NaN	2024-02-24 20:03:27	2024-02-24 20:20:43	0 days 00:17:16	18
9	9	2	STOP	NaN	2024-02-23 09:14:28	2024-02-23 09:31:22	0 days 00:16:54	17
14	14	1	STOP	NaN	2024-02-16 04:16:00	2024-02-16 04:28:51	0 days 00:12:51	13
29	29	2	STOP	NaN	2024-02-07 08:02:33	2024-02-07 08:15:27	0 days 00:12:54	13
56	56	3	STOP	NaN	2024-02-19 05:54:30	2024-02-19 06:08:22	0 days 00:13:52	14
...	...	...	...	...	...	...	...	...
4944	4944	2	STOP	NaN	2024-02-12 17:31:18	2024-02-12 17:45:59	0 days 00:14:41	15
4949	4949	1	STOP	NaN	2024-02-01 05:07:48	2024-02-01 05:19:08	0 days 00:11:20	12
4958	4958	2	STOP	NaN	2024-02-27 08:04:45	2024-02-27 08:24:22	0 days 00:19:37	20
4971	4971	3	STOP	NaN	2024-02-19 19:21:56	2024-02-19 19:37:57	0 days 00:16:01	17
4990	4990	2	STOP	NaN	2024-02-07 20:02:45	2024-02-07 20:02:45	0 days 00:00:00	0

528 rows × 8 columns

After filtering for delivery times between 1 and 240 minutes and removing clear duplicates, I produced a cleaned histogram:

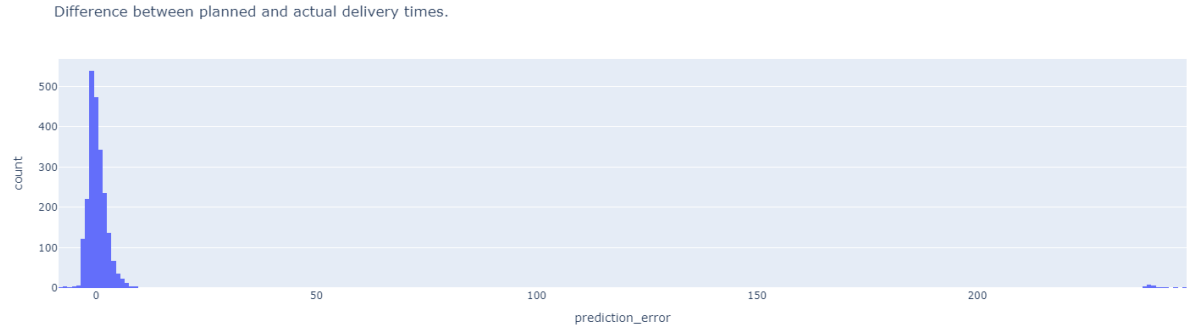


Key observations:

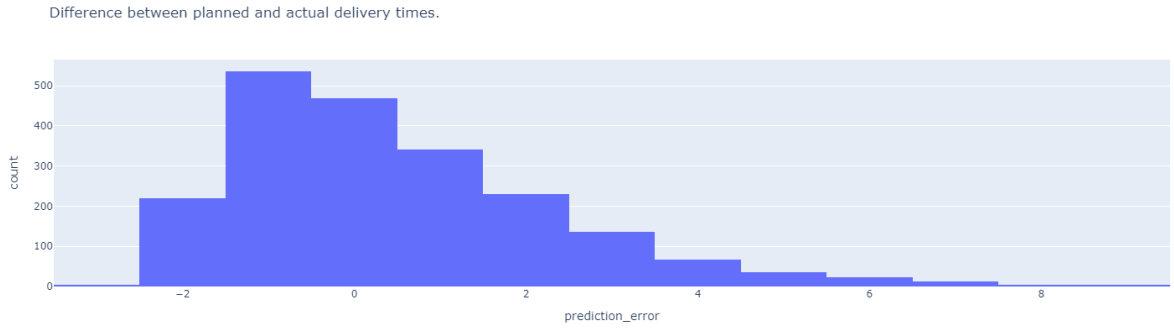
- Most deliveries are fast—typically between 1 and 7 minutes.
- The distribution is right-skewed, with a long tail extending beyond 15 minutes.
- Despite occasional delays, the overall delivery process seems efficient, with the majority completed in under 10 minutes.

2. Generate a histogram showing prediction error (difference between planned and actual delivery times).

Next, I created a histogram showing prediction error (planned minus actual delivery time).



After applying the same data cleaning as before, the updated histogram revealed the following:



We can see that most deliveries are on time or slightly early also the tail is short suggesting a few outliers. There is more delays than early deliveries, but the overall bias is small. Insights:

- Planned delivery times are generally accurate.
- We should investigate delays over 5 minutes to identify issues like elevators, big estates etc.

3. 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 test this, I grouped average delivery times by sector.

The data confirms the hypothesis:

- Sector 1 has an average delivery time of 4.45 minutes, noticeably higher than other sectors, which average closer to 3 minutes.

This suggests that environmental or logistical conditions in Sector 1 could be causing delays, and it may be worth analyzing this area further.

4. Play with the data by grouping, aggregating and remodelling it. Are you able to find any correlations or trends that could be valuable for prediction quality improvement? Describe briefly your findings and visualise them on charts.

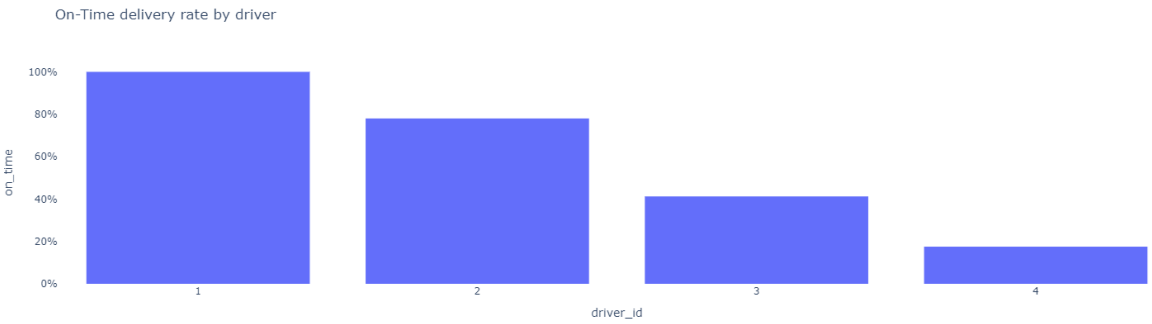
Driver Performance

I looked into whether specific drivers were more prone to delays. Here's what stood out:

- Driver 1: 100% - excellent.
- Driver 2: 80% - reliable.
- Driver 3: 41% - concerning.
- Driver 4: Only 17.5% - definitely worth reviewing.

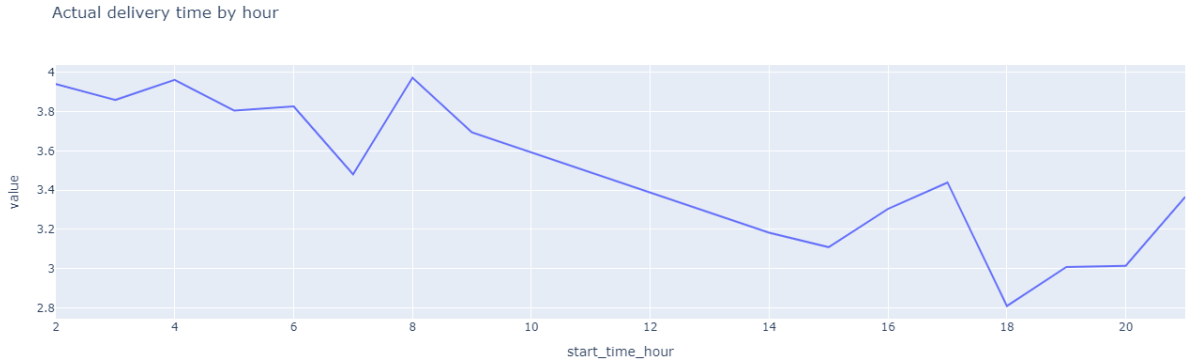
This insight could be useful for targeted training or for adjusting prediction models based on driver behavior.

.



Time of Day

Delivery times tend to increase during late hours compared to earlier in the day. This could be due to reduced accessibility, traffic, or staffing differences. Factoring in time-of-day into prediction models may improve accuracy.



Order features

I explored whether order attributes affect delivery time:

- **Order Weight**: Surprisingly, not a strong factor. Heavier orders didn't consistently take longer.
- **Number of Unique Products**: This **did** correlate with longer delivery times. More unique items likely mean more stops or longer handling time, which makes intuitive sense.