Data Analysis

Data Analysis

Introduction

Welcome to the report on the performance and prediction accuracy of my delivery service. In this report, I will explore delivery times, prediction errors, sector-specific durations, and potential connections in our data. It's important to note that there may be mistakes in our data. To ensure the reliability of my analysis, I took steps to clean the data.

Methodology

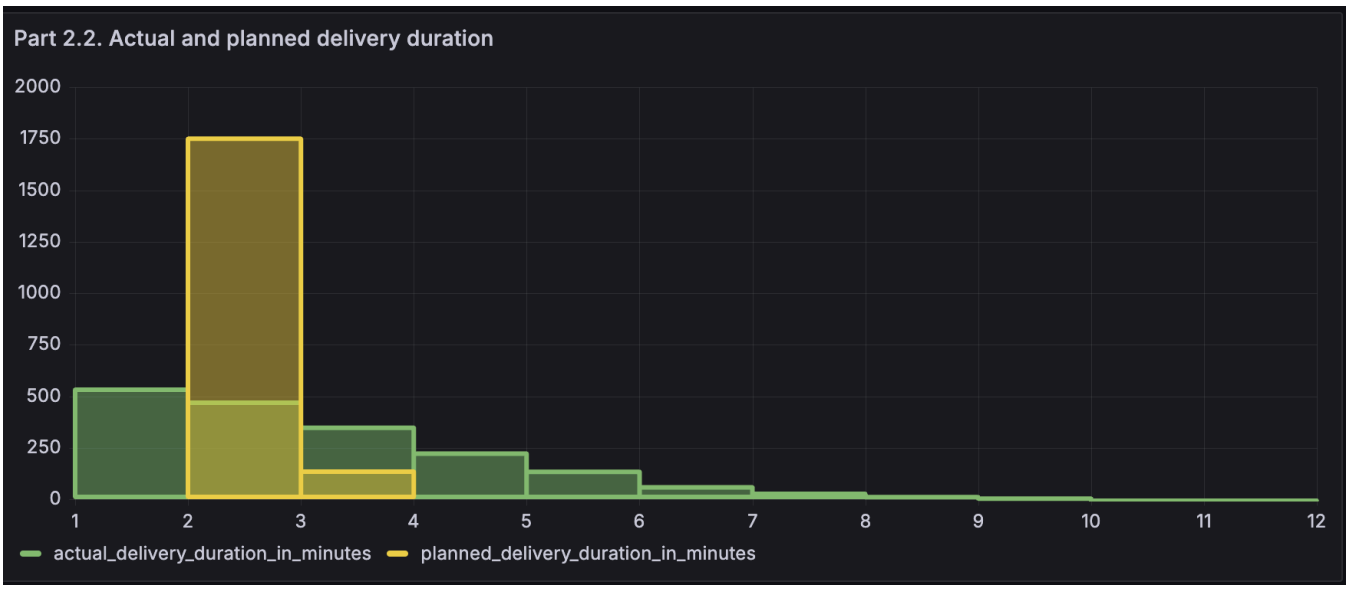
I carefully cleaned my dataset to address potential issues. First, I removed any data where deliveries seemed impossible, such as cases where the end time was before the start time. Additionally, I excluded deliveries taking more than two hours, as they represent only a small part of my data and could affect my analysis.

Part 2.1. Actual delivery

Part 2.1. Delivery Data Analysis

- After examining the data in route_segments, I found some deliveries with negative delivery times. To fix this, I filtered the data using the condition TIMESTAMPDIFF(MINUTE, segment_start_time, segment_end_time) > 0.
- For calculating delivery times, I used the TIMESTAMPDIFF function with the MINUTE parameter.
- I removed deliveries that took more than 200 minutes, as they are rare cases.
- Most deliveries finished within five minutes, but there is a significant group that took 10-15 minutes. I believe focusing on these deliveries could be beneficial.

```
SELECT
  TIMESTAMPDIFF(MINUTE, rs.segment_start_time, rs.segment_end_time) AS duration_in_minutes  -- Calculatin
g the time difference between segment_start_time and segment_end_time in minutes and naming the result 'd
uration_in_minutes'
FROM
  route_segments AS rs -- Selecting the 'route_segments' table and aliasing it as 'rs'
WHERE
  TIMESTAMPDIFF(MINUTE, rs.segment_start_time, rs.segment_end_time) > 0 -- Condition: the time difference
must be greater than 0, eliminates segments with zero or negative length
AND
  TIMESTAMPDIFF(MINUTE, rs.segment_start_time, rs.segment_end_time) < 200 -- Condition: the time differen
ce must be less than 200 minutes, optional constraint
```

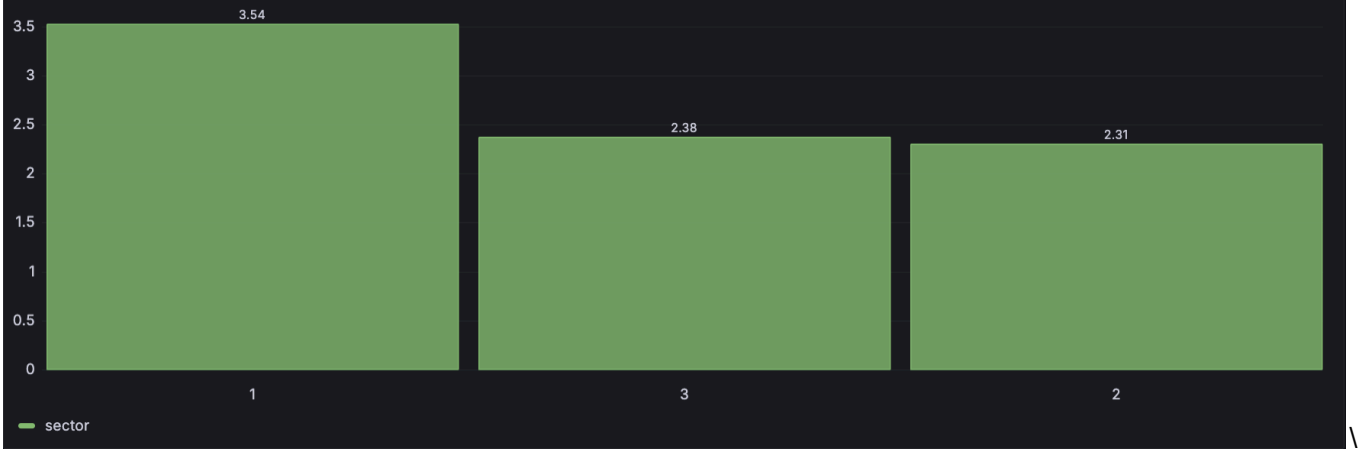


Part 2.2 Delivery and Planned duration comparison

- Upon careful observation, it became evident that the initial time estimations were excessively simplistic, frequently failing to accommodate diverse scenarios. Particularly, deliveries taking longer than five minutes often deviated from the planned delivery time.
- The decision to exclude deliveries surpassing a duration of 200 minutes was warranted, considering their rarity.
- It is worth noting that a substantial majority of deliveries are not effectively completed within the allocated time frame.

```
-- Selecting actual delivery duration in minutes and planned delivery duration in minutes
SELECT
  TIMESTAMPDIFF( -- Calculating the difference in minutes between two timestamps
    MINUTE, -- Specifying that we want the difference in minutes
    rs.segment_start_time, -- Start time of the delivery segment
    rs.segment_end_time -- End time of the delivery segment
  ) AS actual_delivery_duration_in_minutes, -- Assigning the calculated duration to a column alias

  o.planned_delivery_duration / 60 AS planned_delivery_duration_in_minutes -- Calculating planned delivery duration in minutes by dividing plann
ed duration by 60

FROM
  route_segments AS rs -- Selecting data from the 'route_segments' table and assigning an alias 'rs'
  JOIN orders AS o -- Joining with the 'orders' table and assigning an alias 'o'

WHERE
  TIMESTAMPDIFF( -- Ensuring the calculated difference is greater than 0 and less than 200 minutes
    MINUTE, -- Specifying that we want the difference in minutes
    rs.segment_start_time, -- Start time of the delivery segment
    rs.segment_end_time -- End time of the delivery segment
  ) > 0 -- Condition: actual delivery duration must be greater than 0
  AND TIMESTAMPDIFF( -- Additional condition: actual delivery duration must be less than 200 minutes
    MINUTE, -- Specifying that we want the difference in minutes
    rs.segment_start_time, -- Start time of the delivery segment
    rs.segment_end_time -- End time of the delivery segment
  ) < 200 -- Condition: actual delivery duration must be less than 200 minutes

  AND o.order_id = rs.order_id; -- Joining condition: Matching orders with their corresponding segments based on order IDs
```

Part 2.3. Average delivery duration per sector



\

Part 2.3. Delivery duration analyse per sector

- We needed to filter out the outlier deliveries because then we can't recognisse the main trend
- It looks like the avarage delivery time is higher for the sector number 1

```
SELECT
  o.sector_id AS sector, -- Select the sector identifier from the orders table and alias it as 'sector'
  avg( -- Calculate the average delivery time difference in minutes
    TIMESTAMPDIFF( -- Calculate the time difference in minutes between two timestamps
      MINUTE, -- Specify that we want the difference in minutes
      rs.segment_start_time, -- Start time of the delivery segment
      rs.segment_end_time -- End time of the delivery segment
    )
  ) AS actual_delivery_duration_in_minutes -- Assign the calculated average difference to a column named 'actual_delivery_duration_in_minutes'

FROM
  route_segments AS rs -- Select data from the 'route_segments' table and alias it as 'rs'
  JOIN orders AS o -- Join with the 'orders' table and alias it as 'o'

WHERE
  rs.order_id = o.order_id -- Match order IDs between the 'route_segments' and 'orders' tables
  AND TIMESTAMPDIFF( -- Add conditions to filter out invalid time differences
    MINUTE, -- Specify that we want the difference in minutes
    rs.segment_start_time, -- Start time of the delivery segment
    rs.segment_end_time -- End time of the delivery segment
  ) > 0 -- Condition: delivery time difference must be greater than 0
  AND TIMESTAMPDIFF( -- Additional condition: delivery time difference must be less than 200 minutes
    MINUTE, -- Specify that we want the difference in minutes
    rs.segment_start_time, -- Start time of the delivery segment
    rs.segment_end_time -- End time of the delivery segment
  ) < 200

GROUP BY
  sector; -- Group the results by sector identifier
```

Part 2.4. Average delivery duration per driver

Part 2.4. Delivery duration analyse per driver

- It looks like the avarage delivery time depends on the driver
- The driver 4 requires the most time for deliveries.

```
-- Selecting the average actual delivery duration in minutes per driver
SELECT
  rs.driver_id AS driver, -- Selecting the driver ID from the route_segments table and aliasing it as 'driver'
  avg( -- Calculating the average delivery duration in minutes
    TIMESTAMPDIFF( -- Calculating the time difference in minutes between two timestamps
      MINUTE, -- Specifying that we want the difference in minutes
      rs.segment_start_time, -- Start time of the delivery segment
      rs.segment_end_time -- End time of the delivery segment
    )
  ) AS actual_delivery_duration_in_minutes -- Assigning the calculated average duration to a column named 'actual_delivery_duration_in_minutes'

FROM
  route_segments AS rs -- Selecting data from the 'route_segments' table and aliasing it as 'rs'
  JOIN orders AS o -- Joining with the 'orders' table to get additional information
WHERE
  rs.order_id = o.order_id -- Matching order IDs between the 'route_segments' and 'orders' tables
  AND TIMESTAMPDIFF( -- Adding conditions to filter out invalid time differences
    MINUTE, -- Specifying that we want the difference in minutes
    rs.segment_start_time, -- Start time of the delivery segment
    rs.segment_end_time -- End time of the delivery segment
  ) > 0 -- Condition: delivery time difference must be greater than 0
  AND TIMESTAMPDIFF( -- Additional condition: delivery time difference must be less than 200 minutes
    MINUTE, -- Specifying that we want the difference in minutes
    rs.segment_start_time, -- Start time of the delivery segment
    rs.segment_end_time -- End time of the delivery segment
  ) < 200

GROUP BY
  driver -- Grouping the results by driver
ORDER BY
  actual_delivery_duration_in_minutes DESC; -- Ordering the results by actual delivery duration in minutes in descending order
```