Order Delivery Time Prediction: A Linear Regression Analysis

Order Delivery Time Prediction: A Linear Regression Analysis

1. Project Objective

The primary goal of this project is to develop a robust regression model to accurately predict the delivery time for orders placed through Porter. By analyzing a wide range of features—including order specifics, restaurant details, and the availability of delivery partners—the model aims to uncover the key drivers of delivery duration. The insights from this model are intended to optimize logistical efficiency, improve customer satisfaction by setting accurate delivery expectations, and enhance the overall delivery service.

2. Data Understanding and Pipeline

The analysis was performed on the `porter_data_1.csv` dataset, which contains transactional order data. The project followed a structured data science pipeline:

1. Data Loading: Importing the raw dataset.
2. Data Preprocessing & Feature Engineering: Cleaning the data and creating new, more informative features.
3. Exploratory Data Analysis (EDA): Visually inspecting the data to uncover patterns and relationships.

4. Model Building: Developing the predictive model using selected features.
5. Model Evaluation & Inference: Assessing the model's performance and saving it for future use.

3. Data Preprocessing and Feature Engineering

A series of transformations were applied to prepare the dataset for modeling. These steps were crucial for handling inconsistencies and extracting maximum value from the raw data.

- Target Variable Creation: The core target variable, `time_taken` (in minutes), was engineered by calculating the difference between `actual_delivery_time` and `created_at`.
- Temporal Feature Extraction: To capture time-based patterns, the `created_at` timestamp was used to derive new features:
    - `order_placed_hour`: The hour of the day the order was placed.
    - `order_placed_day`: The day of the week.
    - `isWeekend`: A binary flag indicating if the order was placed on a weekend.
- Handling Missing Values: Rows with missing values in key operational fields like `total_onshift_dashers`, `total_busy_dashers`, and `store_primary_category` were removed to ensure data quality.
- Categorical Encoding: The `store_primary_category` feature was converted into numerical format using one-hot encoding, creating separate binary columns for each category.
- Numerical Scaling: All numerical features were standardized using `StandardScaler`. This process scales the data to have a mean of 0 and a standard deviation of 1, preventing features

with larger scales from unduly influencing the model.

- Feature Dropping: Redundant or unnecessary columns, including original timestamps and IDs like `market_id`, were dropped to simplify the model.

4. Exploratory Data Analysis (EDA)

EDA revealed several important characteristics of the dataset:

- Feature Distributions: The distributions of numerical features like `total_items` and `subtotal` were found to be right-skewed, indicating that most orders contain fewer items and have lower subtotals, with a smaller number of large orders.
- Time-Based Trends: Analysis showed that delivery times fluctuate significantly based on the hour of the day, with noticeable peaks during typical meal times.
- Correlations: A correlation matrix was used to examine relationships between variables, helping to identify potential multicollinearity before model building.

5. Model Development

A Linear Regression model was selected for its interpretability and efficiency.

- Feature Selection: To build a parsimonious and effective model, Recursive Feature Elimination (RFE) was employed. RFE iteratively removes the least important features, and based on this analysis, the top 9 features were selected. These included `total_items`, `subtotal`, and specific one-hot encoded hour features like `order_placed_hour_5` (5 AM) and `order_placed_hour_16` (4 PM).

- Model Training: The dataset was split into a training set (80%) and a testing set (20%). The Linear Regression model **was** trained on the standardized training data using the 9 features identified by RFE.

6. Results and Model Evaluation

The model's performance was assessed on the unseen test data.

- Performance Metric: The model achieved an R-squared (R^2) score of 0.1267. This means the model explains approximately 12.7% of the variance in delivery time. While this is a modest score, it indicates the selected features have some predictive power. The remaining variance is likely due to factors not included in the dataset, such as real-time traffic, weather, or specific order complexities.
- Residual Analysis:
    - Residuals vs. Predicted Plot: A plot of the model's residuals (errors) against the predicted values showed a random scatter of points around the zero line. This is a positive diagnostic result, suggesting that the assumption of linearity is appropriate and there is no significant heteroscedasticity (uneven error variance).
    - Histogram and Q-Q Plot: These plots confirmed that the residuals were approximately normally distributed, satisfying another key assumption of linear regression.

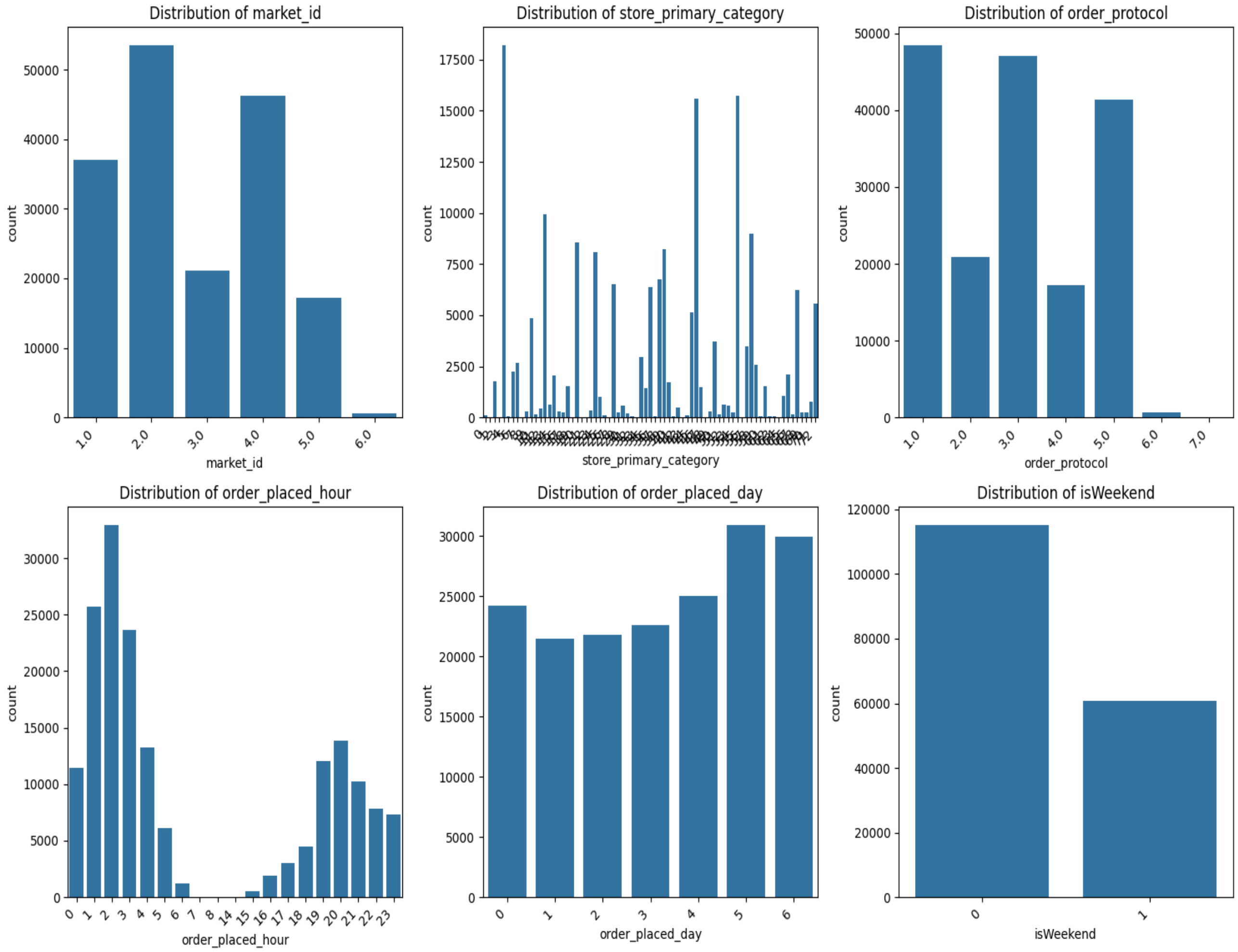
7. Insights and Business Outcomes

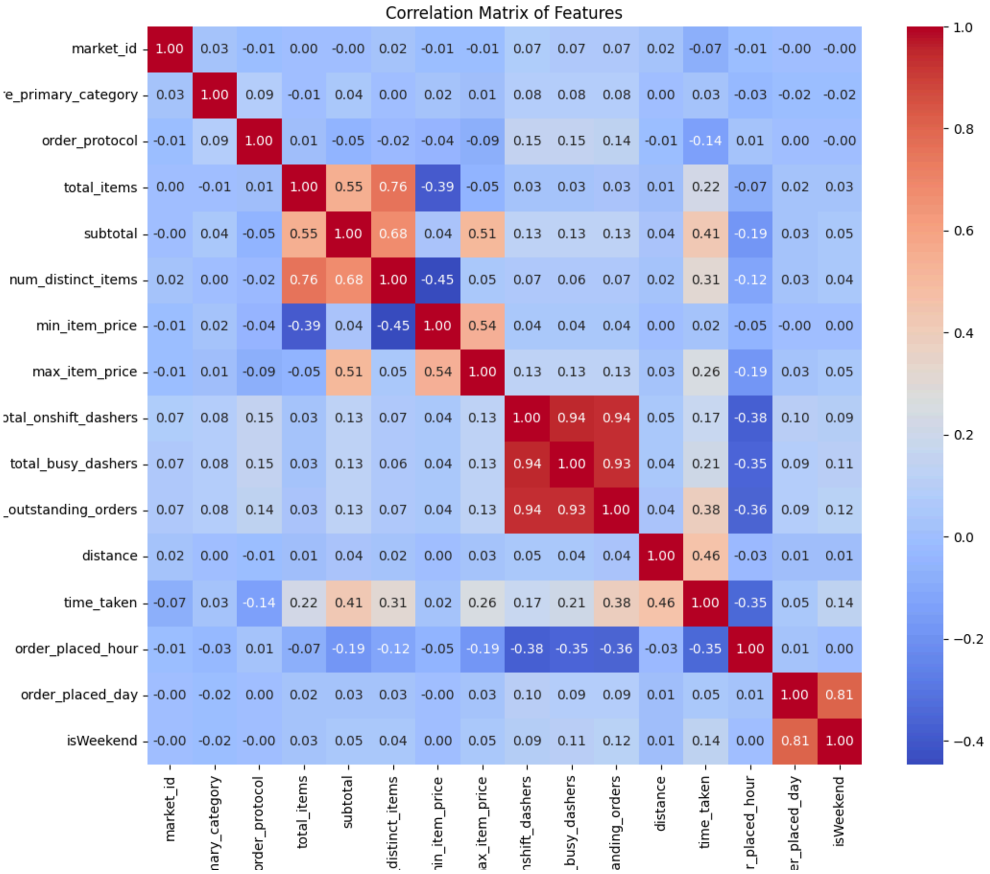
The analysis and resulting model provide several actionable insights:

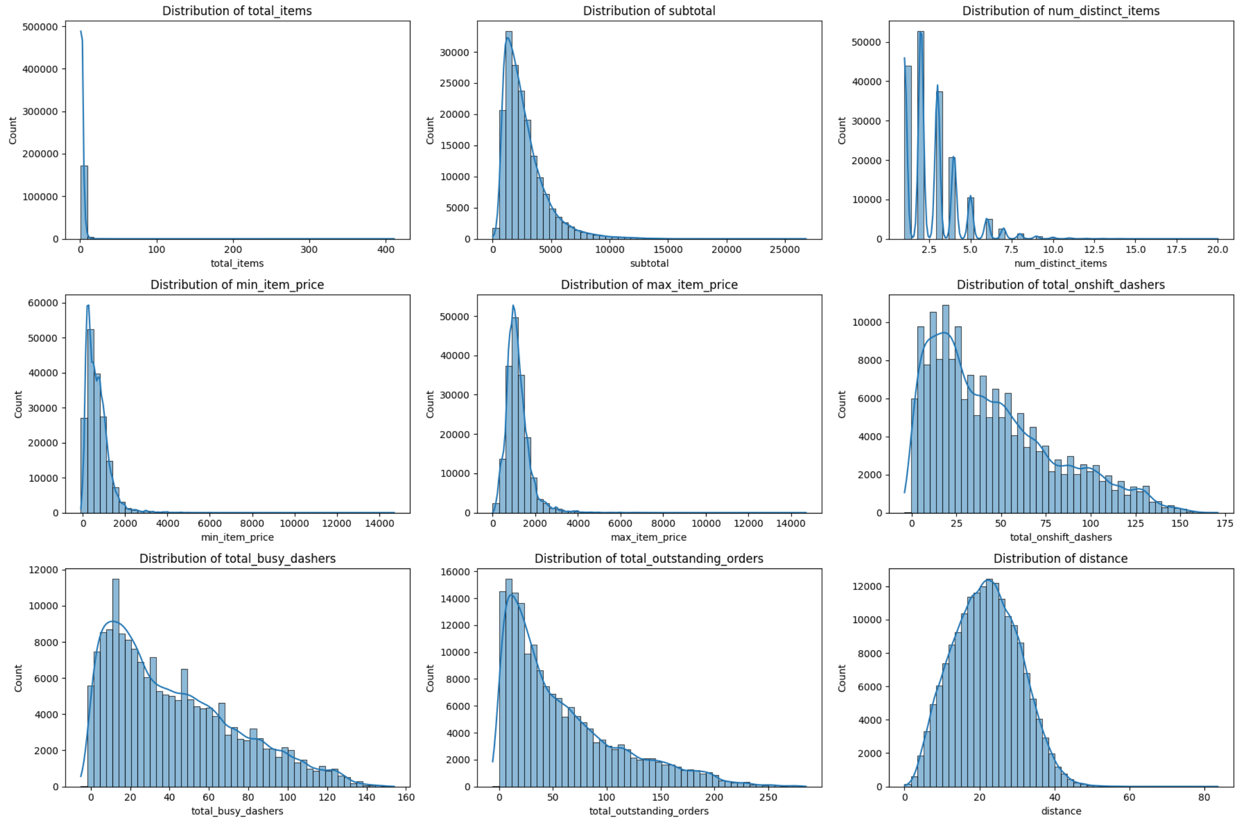
- Key Drivers of Delivery Time: The most influential predictors include the number of items, the order subtotal, and the specific time of day (e.g., 4 PM and 5 AM). This suggests that order complexity and operational conditions at specific hours are critical.

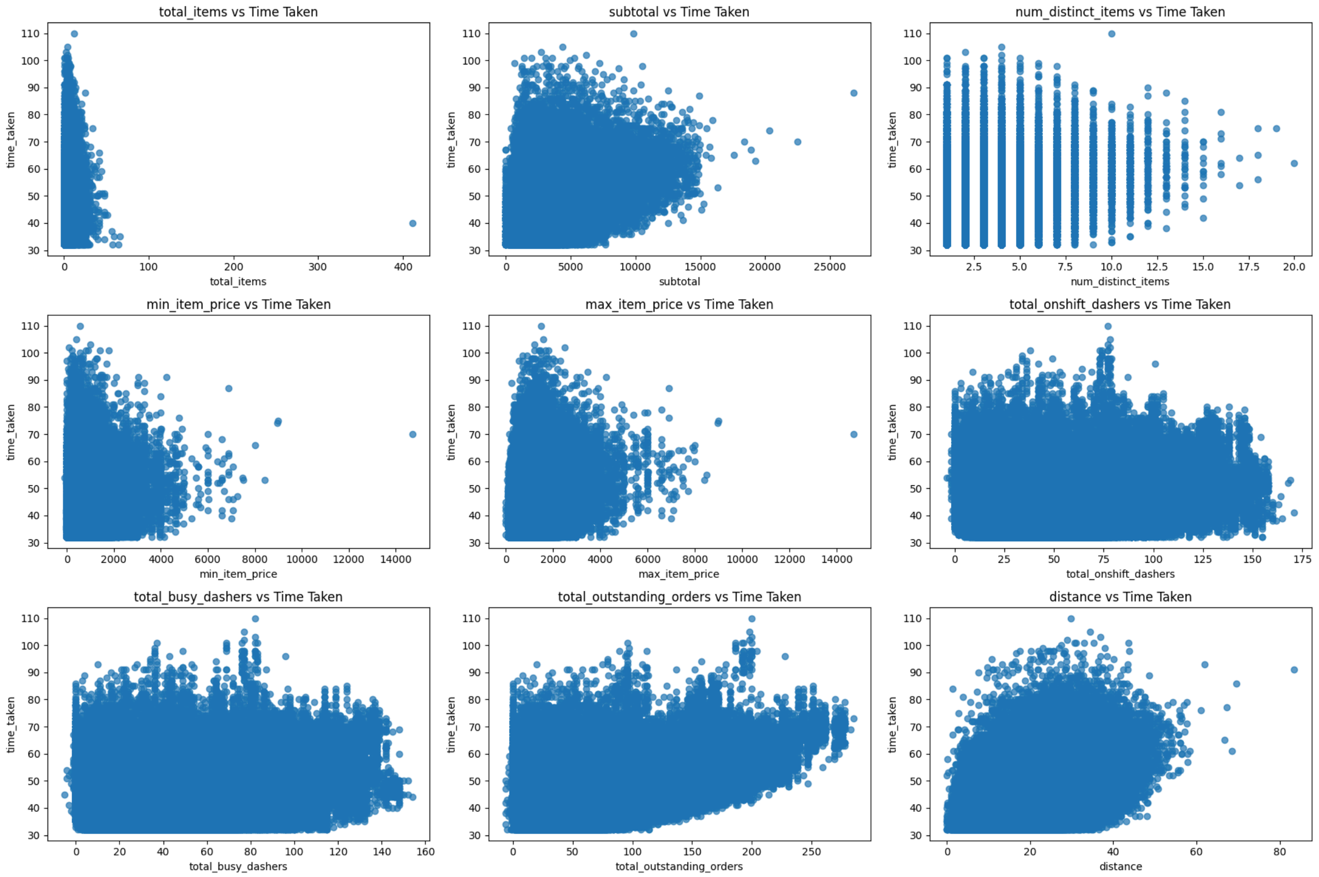
- Operational Optimization: By understanding that certain hours are associated with longer delivery times, the business can proactively adjust staffing levels of delivery partners to meet demand and mitigate delays.
- Customer Experience: While the model is not perfectly predictive, it provides a baseline for more realistic delivery time estimates, which can be used to manage customer expectations and improve satisfaction.

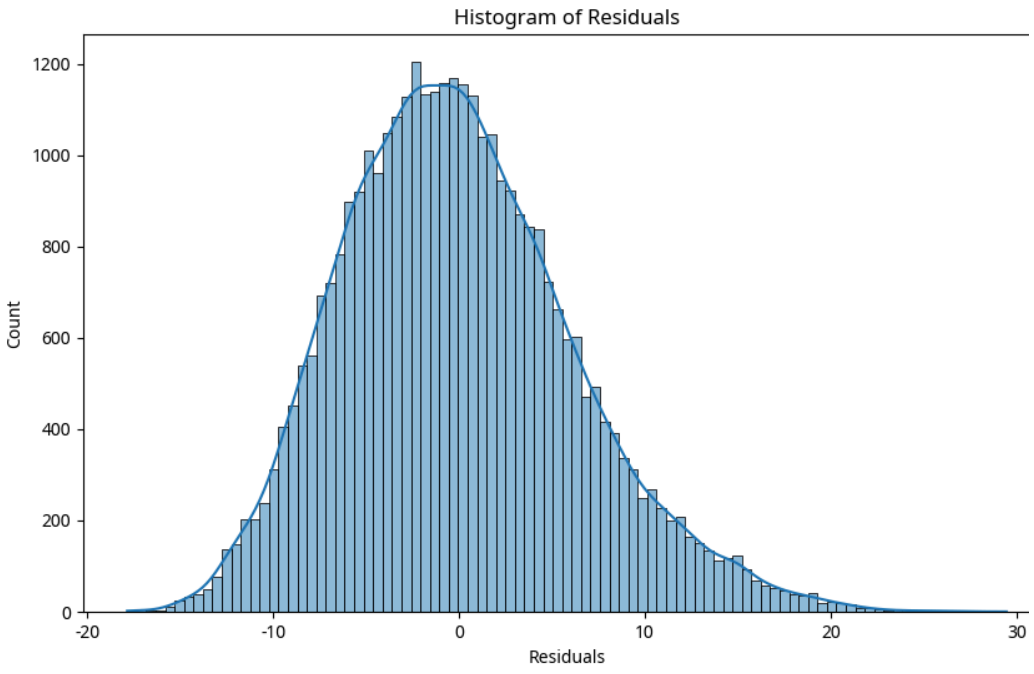
8. Visualizations

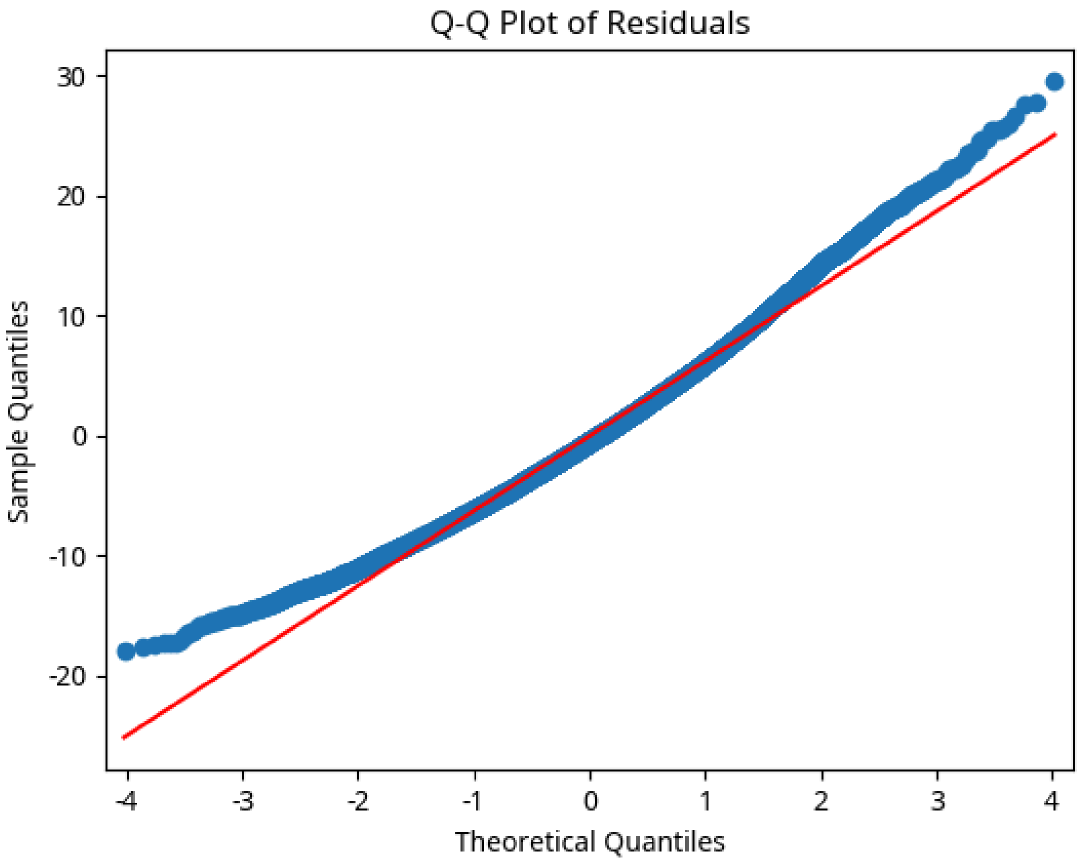


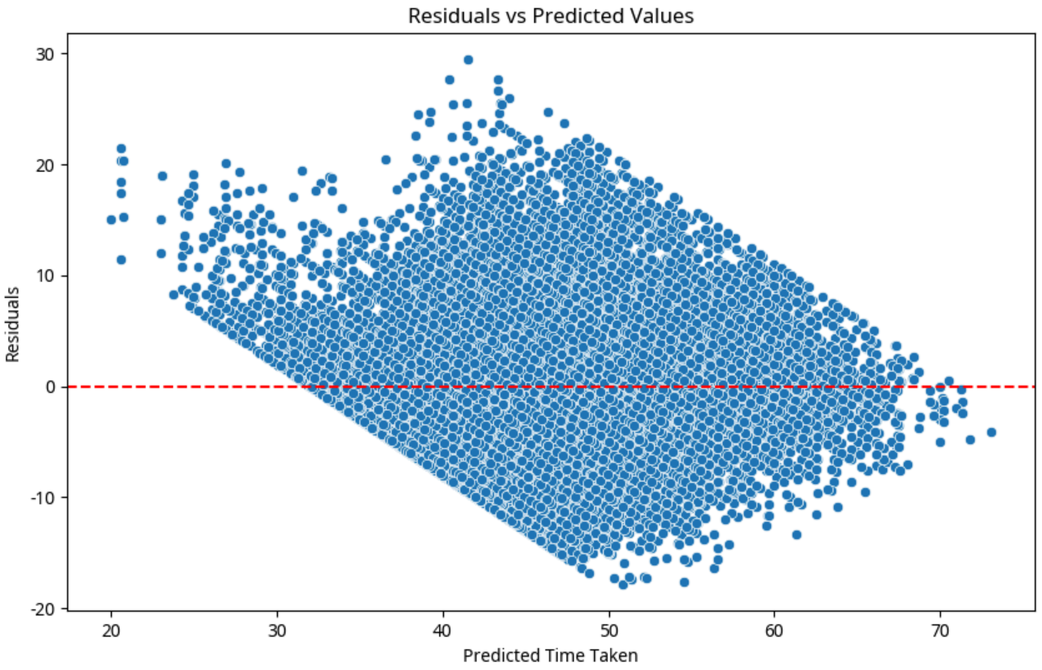

Correlation Matrix of Features
market_id 1.00 0.03 -0.01 0.00 -0.00 0.02 -0.01 -0.01 0.07 0.07 0.07 0.02 -0.07 -0.01 -0.00 -0.00
'e_primary_category 0.03 1.00 0.09 -0.01 0.04 0.00 0.02 0.01 0.08 0.08 0.08 0.00 0.03 -0.03 -0.02 -0.02
order_protocol -0.01 0.09 1.00 0.01 -0.05 -0.02 -0.04 -0.09 0.15 0.15 0.14 -0.01 -0.14 0.01 0.00 -0.00
total_items 0.00 -0.01 0.01 1.00 0.55 0.76 -0.39 -0.05 0.03 0.03 0.03 0.01 0.22 -0.07 0.02 0.03
subtotal -0.00 0.04 -0.05 0.55 1.00 0.68 0.04 0.51 0.13 0.13 0.13 0.04 0.41 -0.19 0.03 0.05
num_distinct_items 0.02 0.00 -0.02 0.76 0.68 1.00 -0.45 0.05 0.07 0.06 0.07 0.02 0.31 -0.12 0.03 0.04
min_item_price -0.01 0.02 -0.04 -0.39 0.04 -0.45 1.00 0.54 0.04 0.04 0.04 0.00 0.02 -0.05 -0.00 0.00
max_item_price -0.01 0.01 -0.09 -0.05 0.51 0.05 0.54 1.00 0.13 0.13 0.13 0.03 0.26 -0.19 0.03 0.05
otal_onshift_dashers 0.07 0.08 0.15 0.03 0.13 0.07 0.04 0.13 1.00 0.94 0.94 0.05 0.17 -0.38 0.10 0.09
total_busy_dashers 0.07 0.08 0.15 0.03 0.13 0.06 0.04 0.13 0.94 1.00 0.93 0.04 0.21 -0.35 0.09 0.11
_outstanding_orders 0.07 0.08 0.14 0.03 0.13 0.07 0.04 0.13 0.94 0.93 1.00 0.04 0.38 -0.36 0.09 0.12
distance 0.02 0.00 -0.01 0.01 0.04 0.02 0.00 0.03 0.05 0.04 0.04 1.00 0.46 -0.03 0.01 0.01
time_taken -0.07 0.03 -0.14 0.22 0.41 0.31 0.02 0.26 0.17 0.21 0.38 0.46 1.00 -0.35 0.05 0.14
order_placed_hour -0.01 -0.03 0.01 -0.07 -0.19 -0.12 -0.05 -0.19 -0.38 -0.35 -0.36 -0.03 -0.35 1.00 0.01 0.00
order_placed_day -0.00 -0.02 0.00 0.02 0.03 0.03 -0.00 0.03 0.10 0.09 0.09 0.01 0.05 0.01 1.00 0.81
isWeekend -0.00 -0.02 -0.00 0.03 0.05 0.04 0.00 0.05 0.09 0.11 0.12 0.01 0.14 0.00 0.81 1.00
market_id
nary_category
order_protocol
total_items
subtotal
distinct_items
in_item_price
ax_item_price
nshift_dashers
busy_dashers
anding_orders
distance
time_taken
r_placed_hour
er_placed_day
isWeekend
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

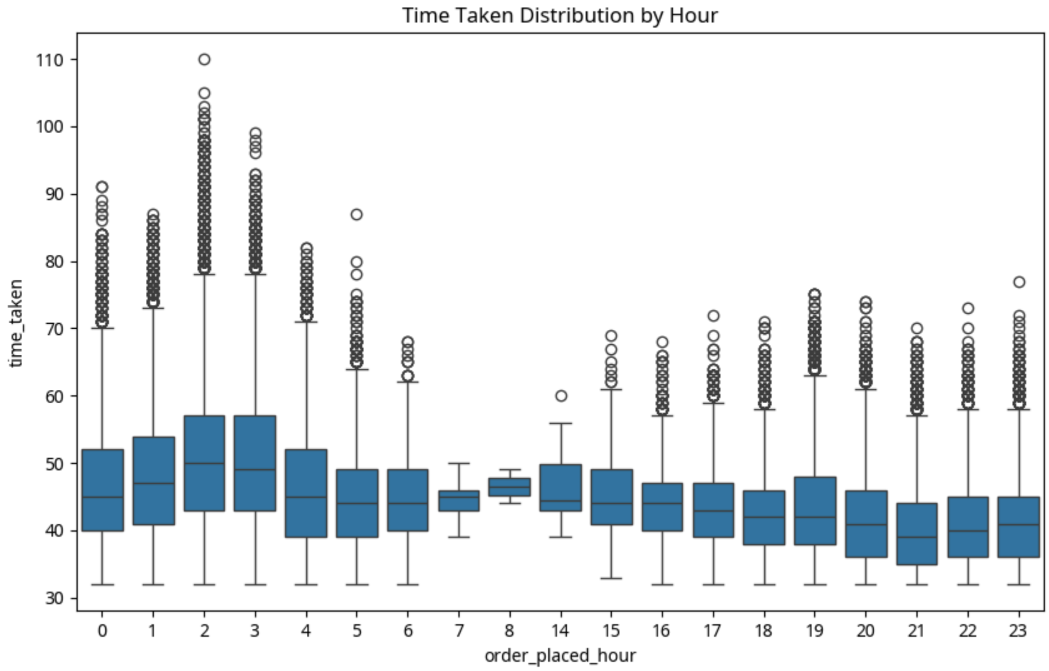

Distribution of total_items
Count
0 100000 200000 300000 400000 500000
0 100 200 300 400
total_items
Distribution of subtotal
Count
0 5000 10000 15000 20000 25000 30000
0 5000 10000 15000 20000 25000
subtotal
Distribution of num_distinct_items
Count
0 10000 20000 30000 40000 50000
2.5 5.0 7.5 10.0 12.5 15.0 17.5 20.0
num_distinct_items
Distribution of min_item_price
Count
0 10000 20000 30000 40000 50000 60000
0 2000 4000 6000 8000 10000 12000 14000
min_item_price
Distribution of max_item_price
Count
0 10000 20000 30000 40000 50000
0 2000 4000 6000 8000 10000 12000 14000
max_item_price
Distribution of total_onshift_dashers
Count
0 2000 4000 6000 8000 10000
0 25 50 75 100 125 150 175
total_onshift_dashers
Distribution of total_busy_dashers
Count
0 2000 4000 6000 8000 10000 12000
0 20 40 60 80 100 120 140 160
total_busy_dashers
Distribution of total_outstanding_orders
Count
0 2000 4000 6000 8000 10000 12000 14000 16000
0 50 100 150 200 250
total_outstanding_orders
Distribution of distance
Count
0 2000 4000 6000 8000 10000 12000
0 20 40 60 80
distance

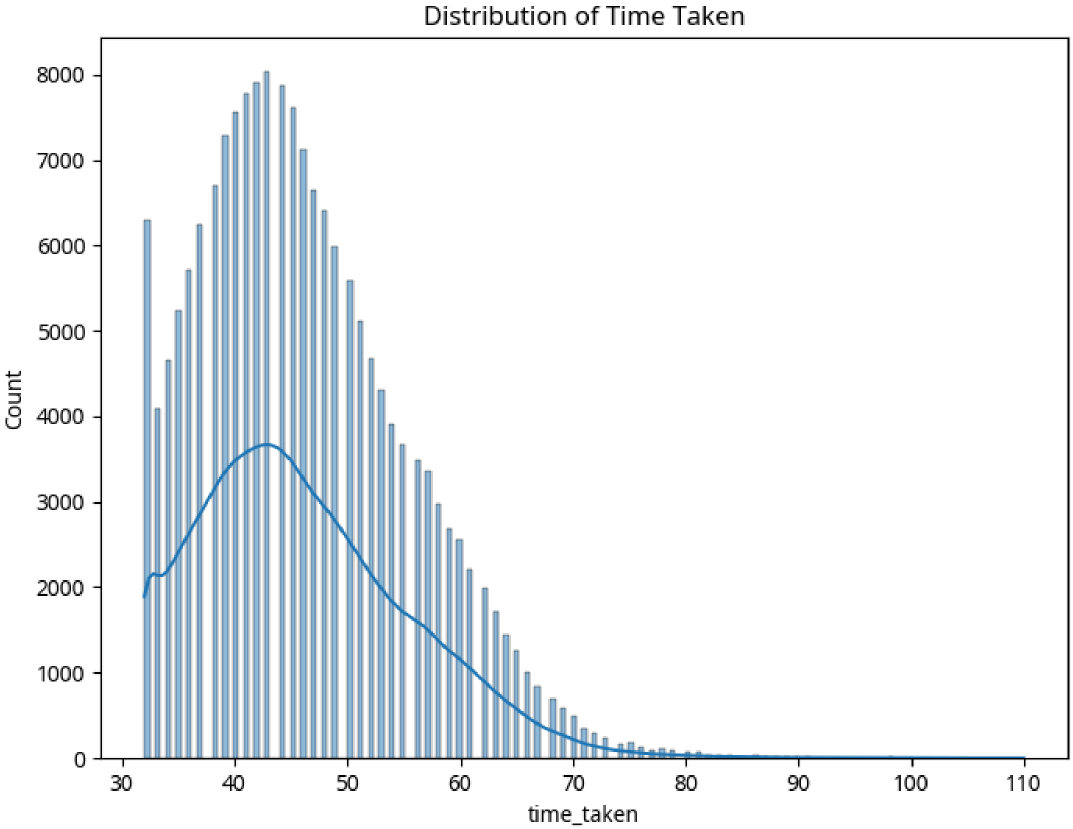

total_items vs Time Taken
time_taken
110
100
90
80
70
60
50
40
30
0
100
200
300
400
total_items
subtotal vs Time Taken
0
5000
10000
15000
20000
25000
subtotal
num_distinct_items vs Time Taken
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
num_distinct_items
min_item_price vs Time Taken
0
2000
4000
6000
8000
10000
12000
14000
min_item_price
max_item_price vs Time Taken
max_item_price
total_onshift_dashers vs Time Taken
0
25
50
75
100
125
150
175
total_onshift_dashers
total_busy_dashers vs Time Taken
0
20
40
60
80
100
120
140
160
total_busy_dashers
total_outstanding_orders vs Time Taken
0
50
100
150
200
250
total_outstanding_orders
distance vs Time Taken
0
20
40
60
80
distance


Histogram of Residuals
Count
1200
1000
800
600
400
200
0
-20
-10
0
10
20
30
Residuals


Q-Q Plot of Residuals
Sample Quantiles
30
20
10
0
-10
-20
Theoretical Quantiles
-4
-3
-2
-1
0
1
2
3
4


Residuals vs Predicted Values
Residuals
30
20
10
0
-10
-20
Predicted Time Taken
20
30
40
50
60
70


Time Taken Distribution by Hour
time_taken
110
100
90
80
70
60
50
40
30
order_placed_hour
0
1
2
3
4
5
6
7
8
14
15
16
17
18
19
20
21
22
23


Distribution of Time Taken
Count
8000
7000
6000
5000
4000
3000
2000
1000
0
time_taken
30
40
50
60
70
80
90
100
110

8. Conclusion and Future Work

This project successfully developed a linear regression model that provides valuable, interpretable insights into the factors affecting order delivery time. The model serves as a strong foundation for a more sophisticated prediction system.

Future work could focus on:

- Incorporating External Data: Enhancing the model by adding features like real-time traffic data, weather conditions, and local events.
- Advanced Modeling: Experimenting with more complex, non-linear models (e.g., Gradient Boosting, Random Forest) that may capture more intricate patterns in the data and improve the R^2 score.

9. Conclusion

This project successfully developed a linear regression model to predict order delivery times. While the model's predictive power is moderate, it provides valuable insights into the factors influencing delivery duration. The findings from this analysis can be used to inform business decisions, improve operational efficiency, and enhance the customer experience.