Shyju S R

Order Delivery Time Prediction

Business Problem

Porter delivery refers to the logistics services offered by Porter, a platform that provides on-demand goods transportation solutions in India.

In order to improve customer experience, optimize operations, and increase efficiency, Porter needs to accurately predict delivery times. Inaccurate delivery estimates lead to customer dissatisfaction and operational challenges.

The key goals are:

- Predict the delivery time for an order based on multiple input features
- Improve delivery time predictions to optimise operational efficiency
- Understand the key factors influencing delivery time to enhance the model's accuracy

Approach

Created a data-driven prediction model through the use of linear regression, examining large number of delivery records that contained information about order details, restaurant location, delivery partner availability, and distance. EDA, feature engineering, data cleaning, and model optimisation were all part of our methodical approach.

The steps listed below are used to predict the delivery time.

1. Loading the data

Porter data is provided in csv format and it is loaded in Dataframe. There are no missing or null values in the dataset, and the data is nearly clean.

2. Data Preprocessing and Feature Engineering

Data Type Conversion

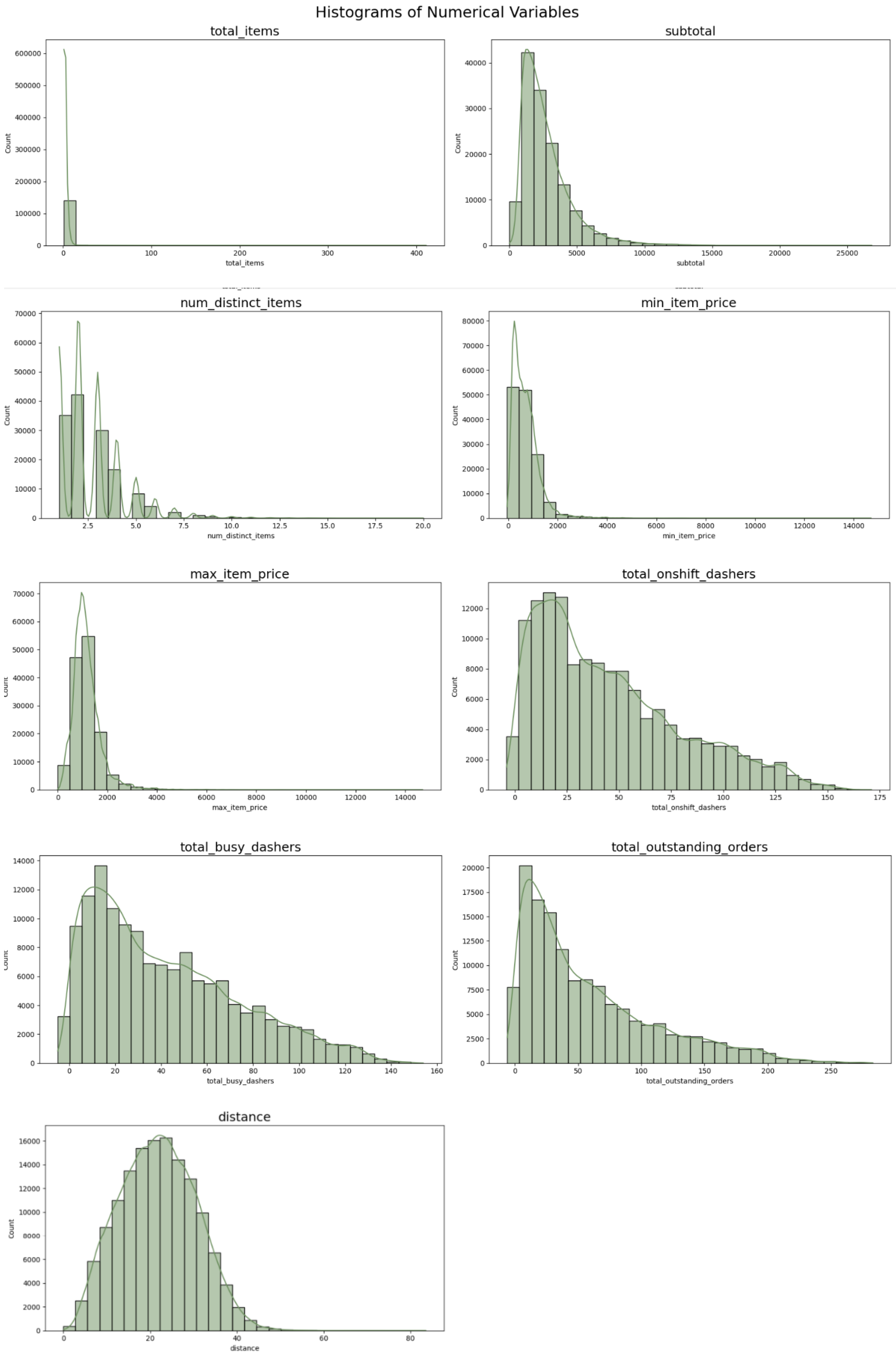
- Timestamps (created_at, actual_delivery_time) were converted to datetime format
- Categorical fields (market_id , store_primary_category and order_protocol) were converted to category

Feature Engineering

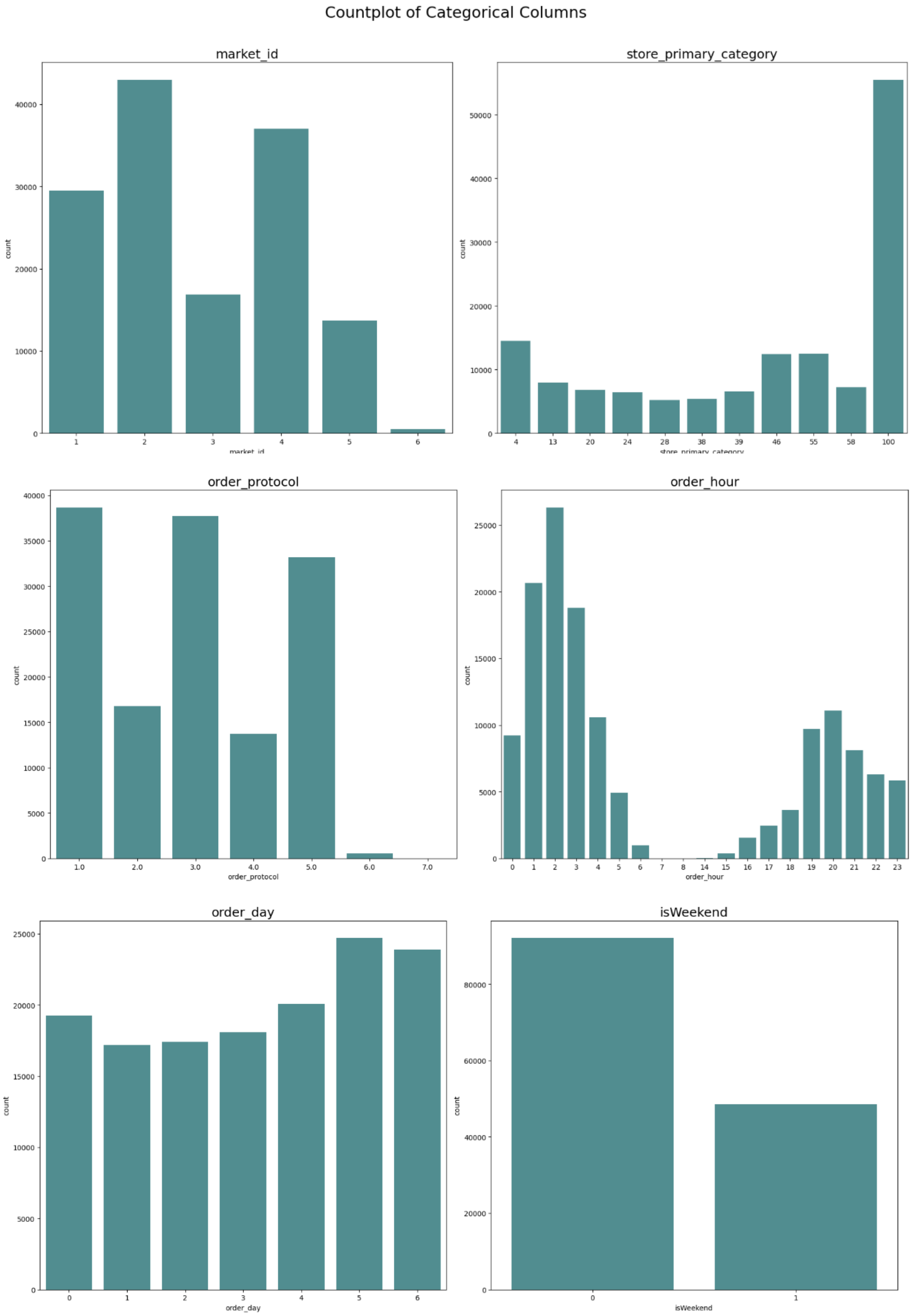
- By computing the difference between the creation and delivery timings, **time_taken** (in minutes) was created as the target variable.
- Order timestamps were used to extract the hour and **day_of _week**.
- To record weekend versus weekday trends, the **isWeekend** binary feature was added.
- Mapped day names to numerical values (0-6)

3. Exploratory Data Analysis on Training Data

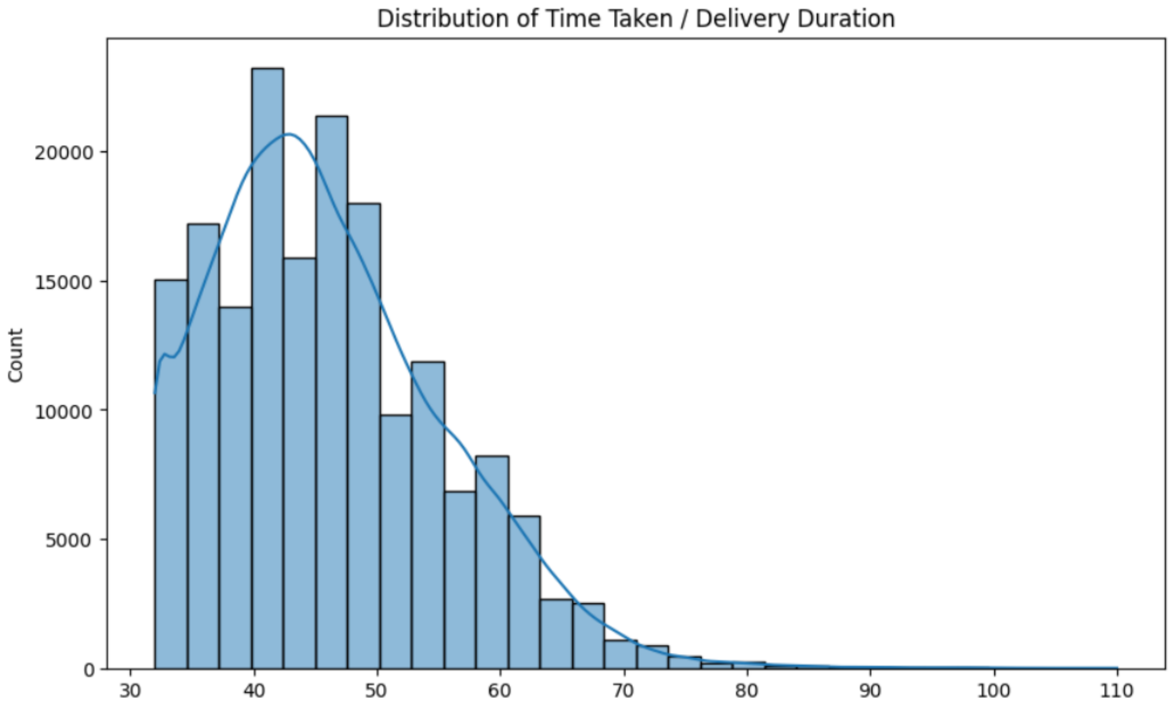
Distributions for numerical columns in the training set to understand their spread and any skewness.


Histograms of Numerical Variables
total_items
Count
total_items
subtotal
Count
subtotal
num_distinct_items
Count
num_distinct_items
min_item_price
Count
min_item_price
max_item_price
Count
max_item_price
total_onshift_dashers
Count
total_onshift_dashers
total_busy_dashers
Count
total_busy_dashers
total_outstanding_orders
Count
total_outstanding_orders
distance
Count
distance

Distribution of categorical features

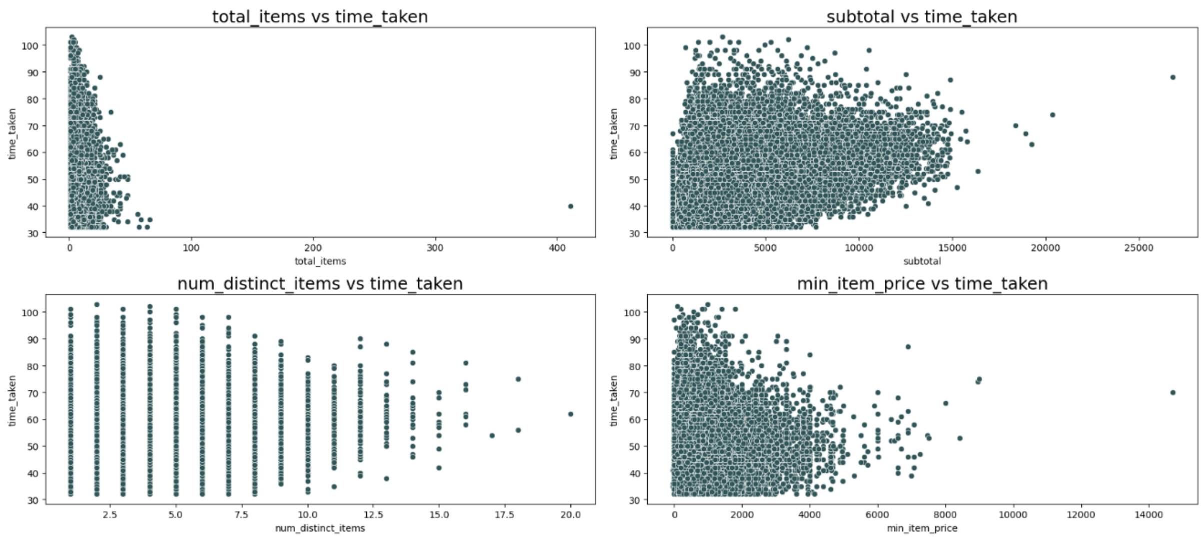


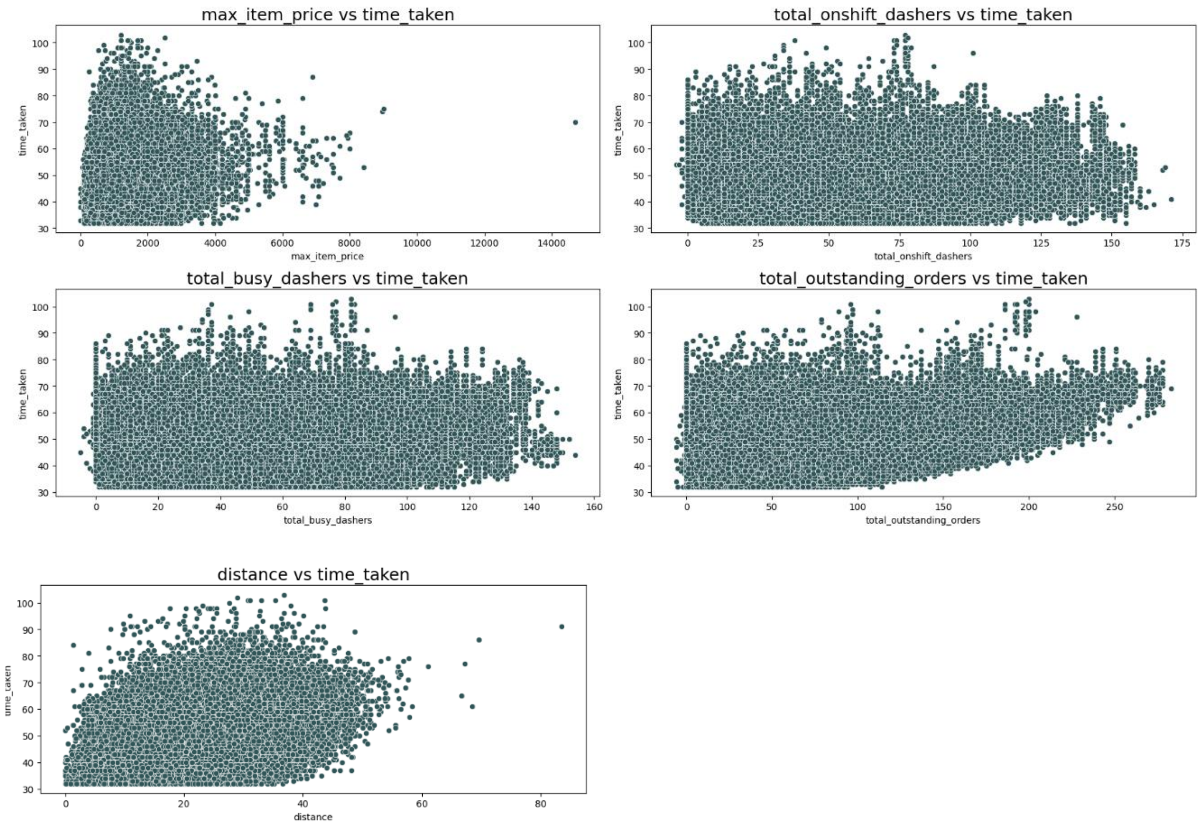
Distribution of the target variable to understand its spread and any skewness



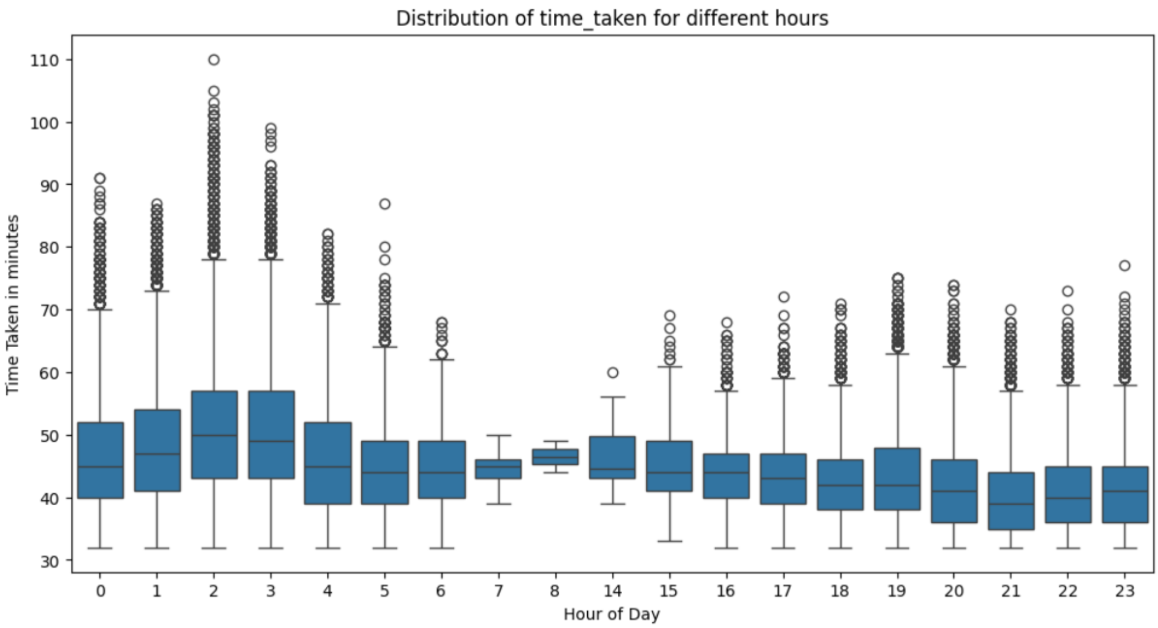
Relationships Between Features

Scatter plots for important numerical and categorical features to observe how they relate to **time_taken**

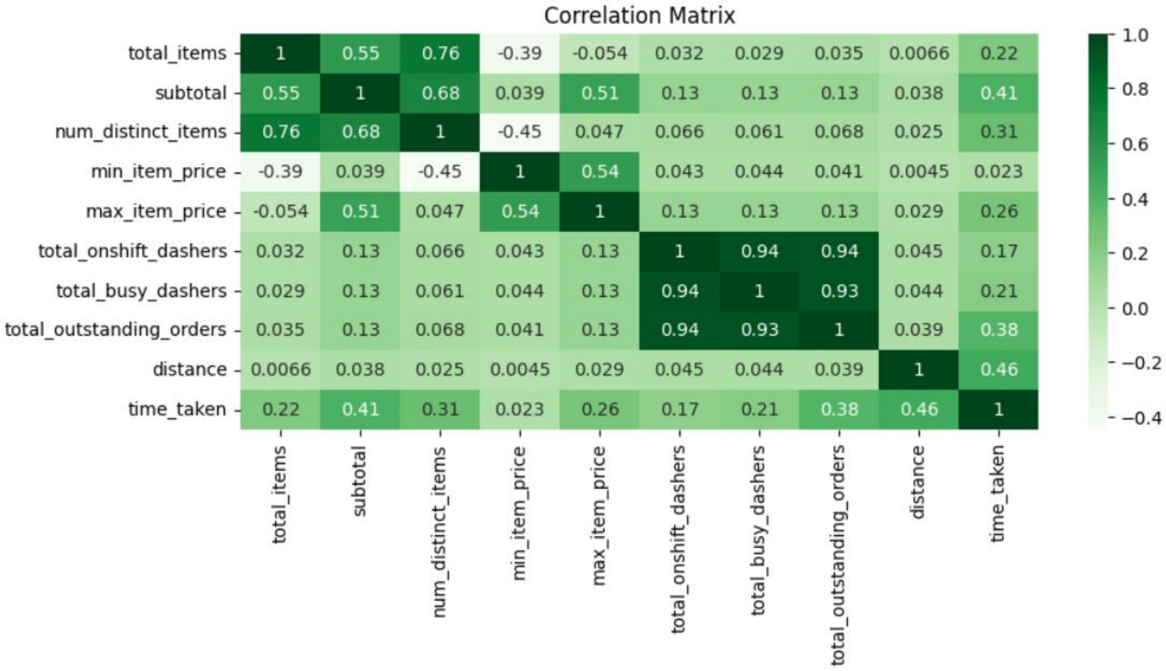




Distribution of time_taken for different hours



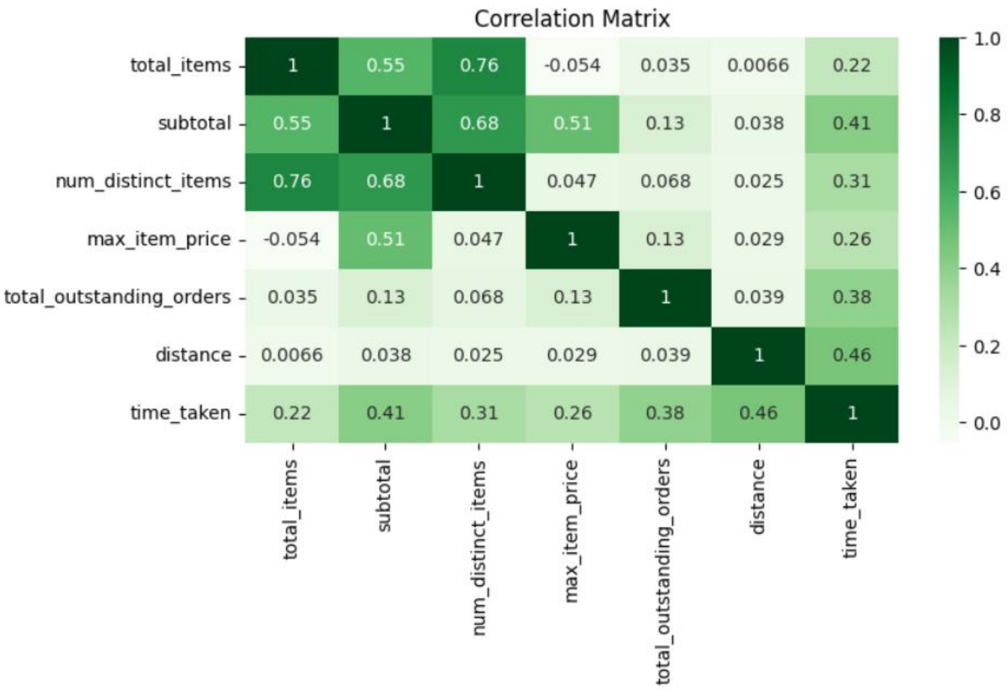
Correlation Matrix



Correlation Matrix after dropping columns with weak correlations with the target variable

Following columns are **dropped**

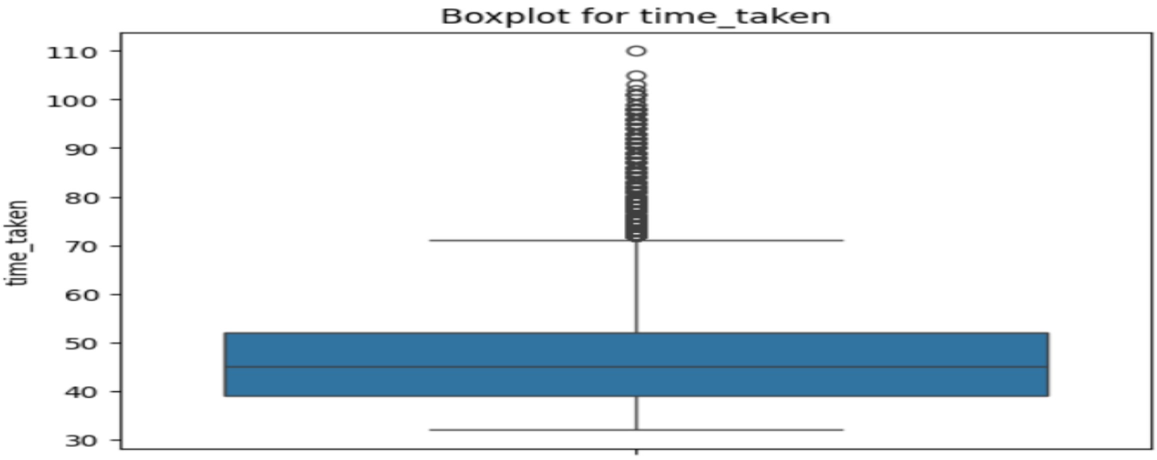
- min_item_price
- total_onshift_dashers
- total_busy_dashers

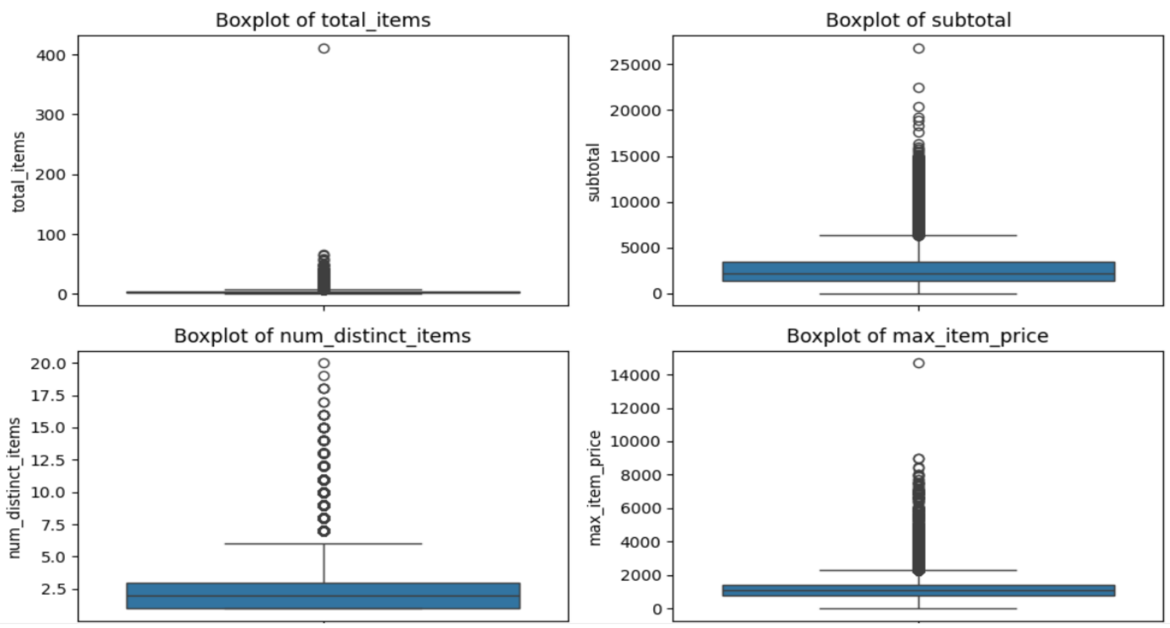


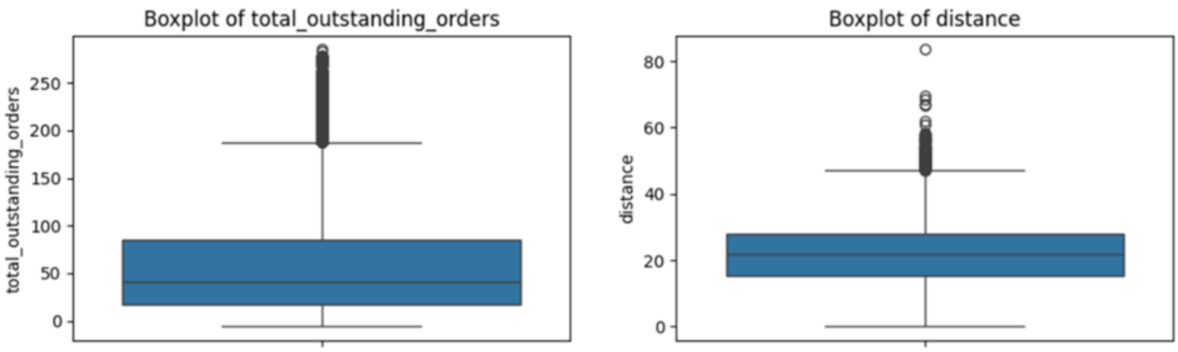
Handling the Outliers

Visualisation of potential outliers for the target variable and other numerical features using boxplots.

Box plot of target variable.







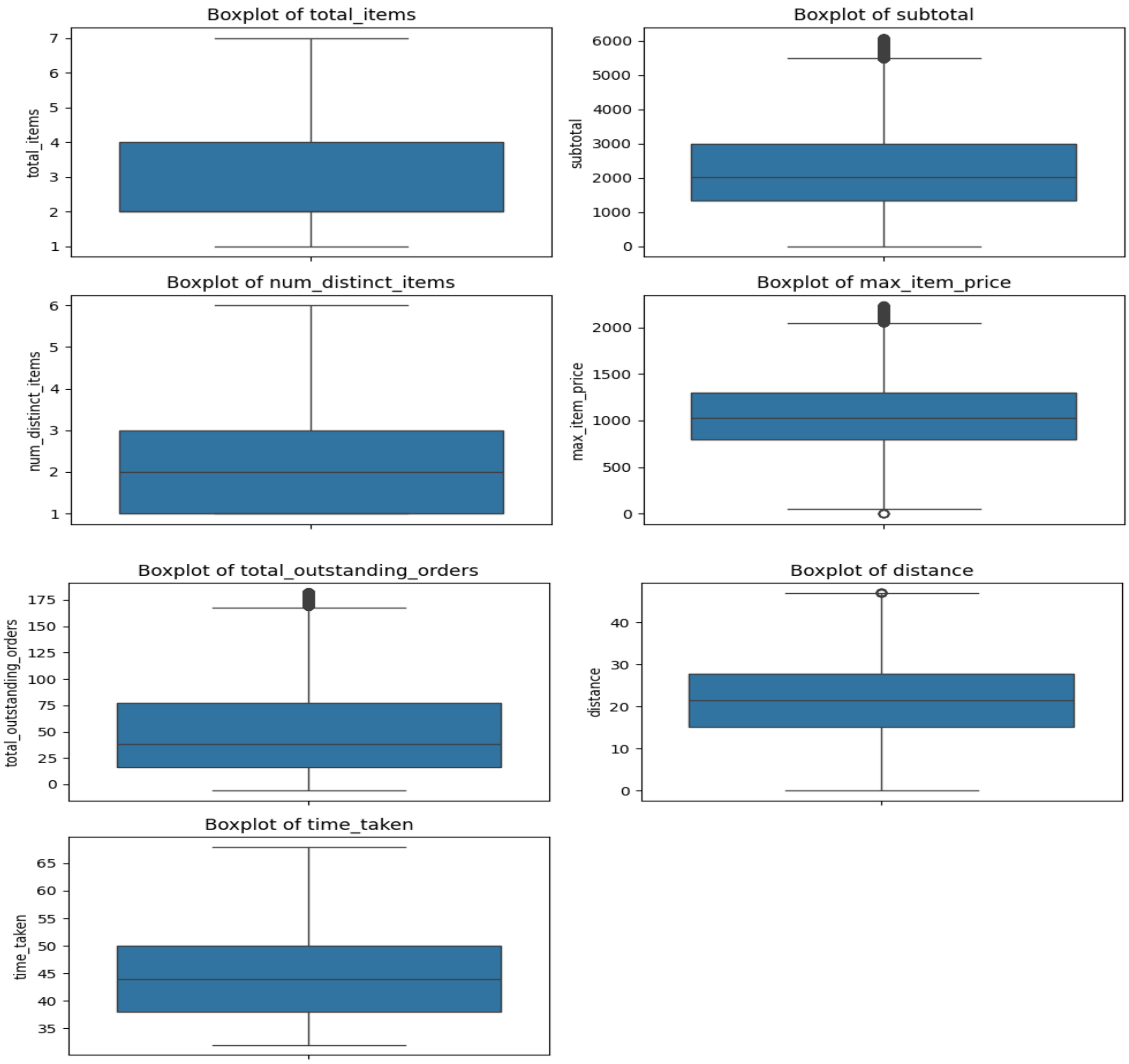
Removed outliers

- Removed outliers only from train_data
- test_data should represent real-world data, including outliers.
- Used Interquartile Range (IQR) method to remove outliers from the numerical columns.

  Here is the calculation.

  - **Q1** = quantile(0.25) = 25%
  - **Q3** = quantile(0.75) = 75%
  - **IQR** = Q3 - Q1 *(Standard calculation)*
  - **Lower Bound** = Q1 - 1.5 × IQR *(Standard calculation)*
  - **Upper Bound** = Q3 + 1.5 × IQR *(Standard calculation)* Only data falling within these ranges is taken for the training set.

Boxplot after remove outliers from train_data using IQR method



Separated cleaned training features and target

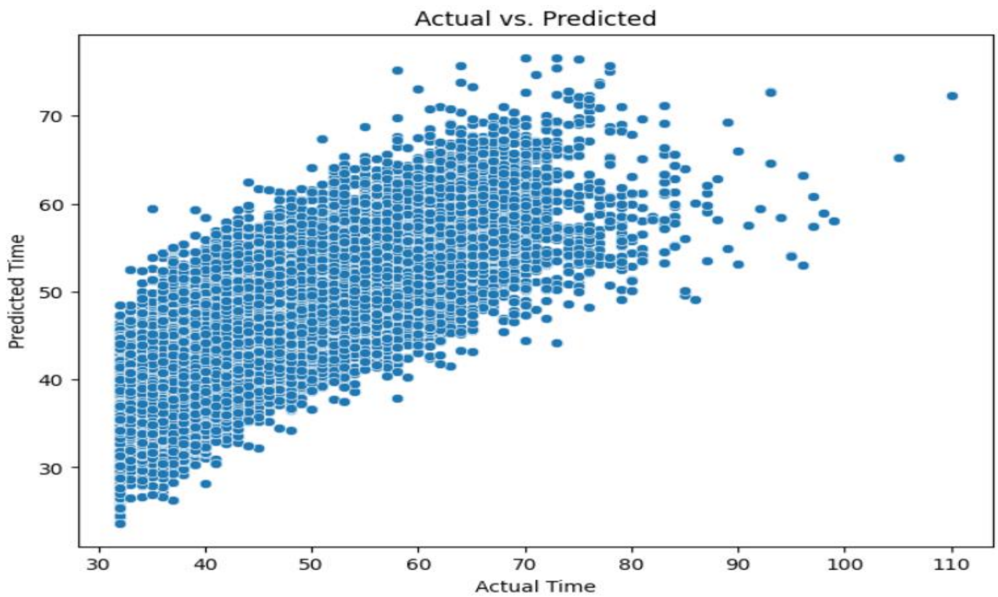
```
y_train_cleaned = cleaned_train_data['time_taken']
X_train_cleaned = cleaned_train_data.drop(columns=['time_taken'])
```

4. Exploratory Data Analysis on Validation Data

- Dropped the columns with weak correlations with the target variable

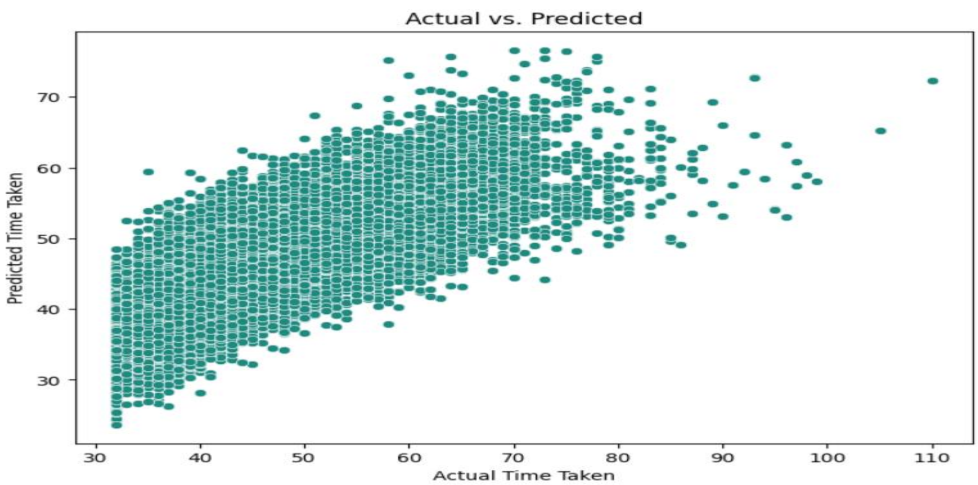
5. Model Building

- Performed **Feature scaling for numerical columns**
- Created **Dummies for Categorical columns**
- Build a linear regression mode using ‘statsmodels’ api’s.



Test data may contain outlyers

Linear Regression using using **sklearn**



5.3 Build the model and fit RFE to select the most important features

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             time_taken   R-squared:                       0.656
Model:                            OLS   Adj. R-squared:                  0.656
Method:                 Least Squares   F-statistic:                     5254.
Date:                Wed, 25 Jun 2025   Prob (F-statistic):               0.00
Time:                        22:26:15   Log-Likelihood:            -4.3846e+05
No. Observations:              140621   AIC:                         8.770e+05
Df Residuals:                  140569   BIC:                         8.775e+05
Df Model:                          51
Covariance Type:            nonrobust
================================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                       36.5656      0.105    348.905      0.000      36.360      36.771
total_items                -11.6815      3.508     -3.330      0.001     -18.557      -4.806
subtotal                    32.2744      0.406     79.478      0.000      31.479      33.070
num_distinct_items          10.3601      0.309     33.494      0.000       9.754      10.966
max_item_price               8.4545      0.535     15.798      0.000       7.406       9.503
total_outstanding_orders    24.3538      0.127    192.473      0.000      24.106      24.602
distance                    40.3139      0.140    287.484      0.000      40.039      40.589
market_id_2                 -8.9678      0.048   -186.758      0.000      -9.062      -8.874
market_id_3                 -4.4371      0.053    -83.417      0.000      -4.541      -4.333
market_id_4                 -7.1481      0.049   -144.842      0.000      -7.245      -7.051
market_id_5                 -4.0836      0.057    -71.461      0.000      -4.196      -3.972
market_id_6                 -4.9870      0.245    -20.318      0.000      -5.468      -4.506
store_primary_category_13    0.1058      0.080      1.331      0.183      -0.050       0.262
store_primary_category_20    0.0350      0.082      0.427      0.670      -0.126       0.196
store_primary_category_24    0.4534      0.083      5.460      0.000       0.291       0.616
store_primary_category_28    0.5820      0.103      5.668      0.000       0.381       0.783
store_primary_category_38    0.6473      0.089      7.315      0.000       0.474       0.821
store_primary_category_39    0.5703      0.083      6.887      0.000       0.408       0.733
store_primary_category_46    0.1300      0.068      1.906      0.057      -0.004       0.264
store_primary_category_55    0.5284      0.068      7.760      0.000       0.395       0.662
store_primary_category_58    0.4858      0.081      6.024      0.000       0.328       0.644
store_primary_category_100   0.3899      0.052      7.429      0.000       0.287       0.493
order_protocol_2            -0.8564      0.052    -16.338      0.000      -0.959      -0.754
order_protocol_3            -1.7587      0.042    -41.829      0.000      -1.841      -1.676
order_protocol_4            -2.2879      0.064    -35.598      0.000      -2.414      -2.162
order_protocol_5            -3.3828      0.043    -78.966      0.000      -3.467      -3.299
order_protocol_6            -1.4554      0.242     -6.022      0.000      -1.929      -0.982
order_protocol_7             1.0333      1.369      0.755      0.450      -1.650       3.716
order_hour_1                -2.1091      0.070    -29.947      0.000      -2.247      -1.971
order_hour_2                -1.2276      0.072    -16.936      0.000      -1.370      -1.086
order_hour_3                -0.9371      0.076    -12.410      0.000      -1.085      -0.789
order_hour_4                -1.7865      0.079    -22.485      0.000      -1.942      -1.631
order_hour_5                -0.1187      0.097     -1.223      0.221      -0.309       0.072
order_hour_6                 1.7439      0.185      9.430      0.000       1.381       2.106
order_hour_7                 2.5040      2.069      1.210      0.226      -1.550       6.559
order_hour_8                 7.4392      3.869      1.923      0.055      -0.144      15.022
order_hour_14                1.8490      1.018      1.816      0.069      -0.146       3.844
order_hour_15                1.8765      0.283      6.619      0.000       1.321       2.432
order_hour_16                1.9801      0.152     13.064      0.000       1.683       2.277
```

. . .

	Feature	VIF
0	const	48.577033
1	total_items	2.542091
2	subtotal	3.613663
3	num_distinct_items	3.286115
4	max_item_price	1.981981
5	total_outstanding_orders	2.499078
6	distance	1.014770
7	market_id_2	2.299047
8	market_id_3	1.405700
9	market_id_4	2.219878
10	market_id_5	1.350673
11	market_id_6	1.017007
12	store_primary_category_13	1.582853
13	store_primary_category_20	1.466716
14	store_primary_category_24	1.414174
15	store_primary_category_28	1.768798
16	store_primary_category_38	1.356834
17	store_primary_category_39	1.438837
18	store_primary_category_46	1.761599
19	store_primary_category_55	1.768300
20	store_primary_category_58	1.496240
21	store_primary_category_100	3.092024
22	order_protocol_2	1.357589
23	order_protocol_3	1.630440
24	order_protocol_4	1.708790
25	order_protocol_5	1.554546
26	order_protocol_6	1.030586
27	order_protocol_7	1.001784

. . .

1. R-squared : 0.656 indicates 65.6% of the variance in the target variable (time_taken) is explained by the model.

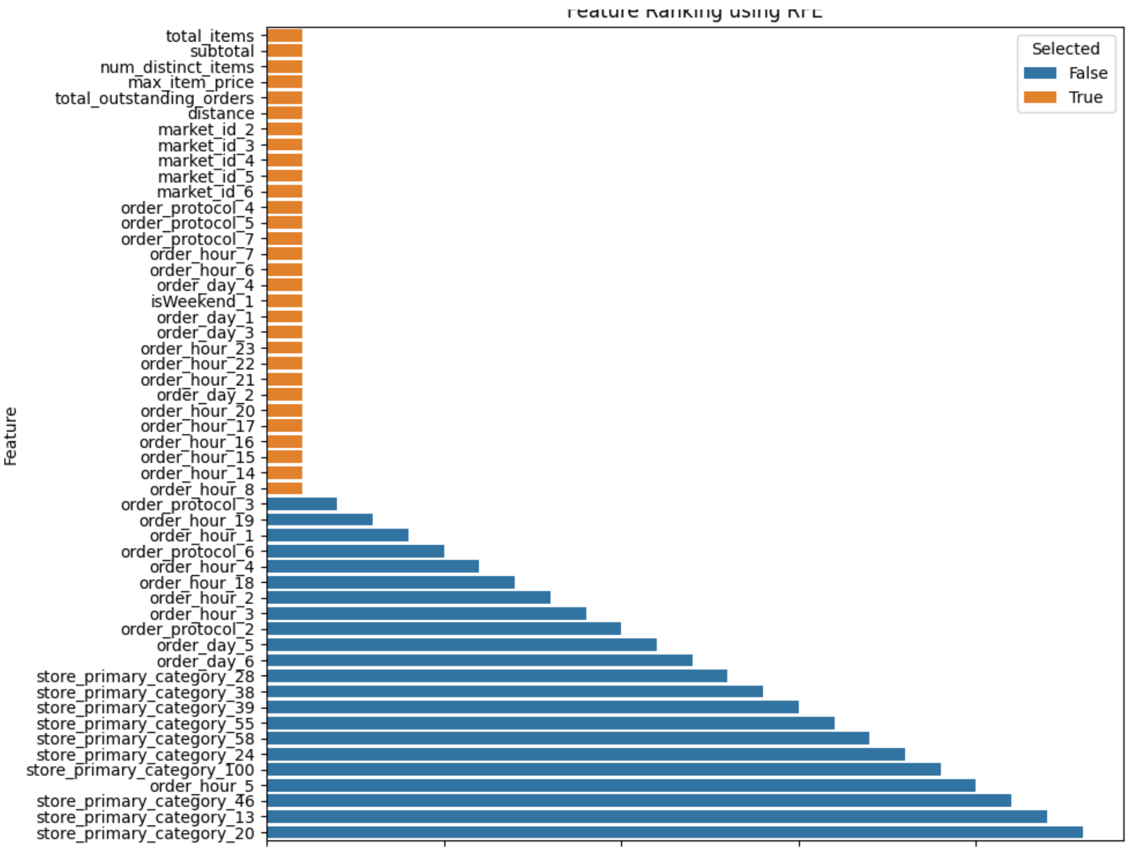
Decision for which variables has to be dropped can be taken based on

- Significance (p-value > 0.05 is higher) and VIF (VIF > 5 is not a good symbol)

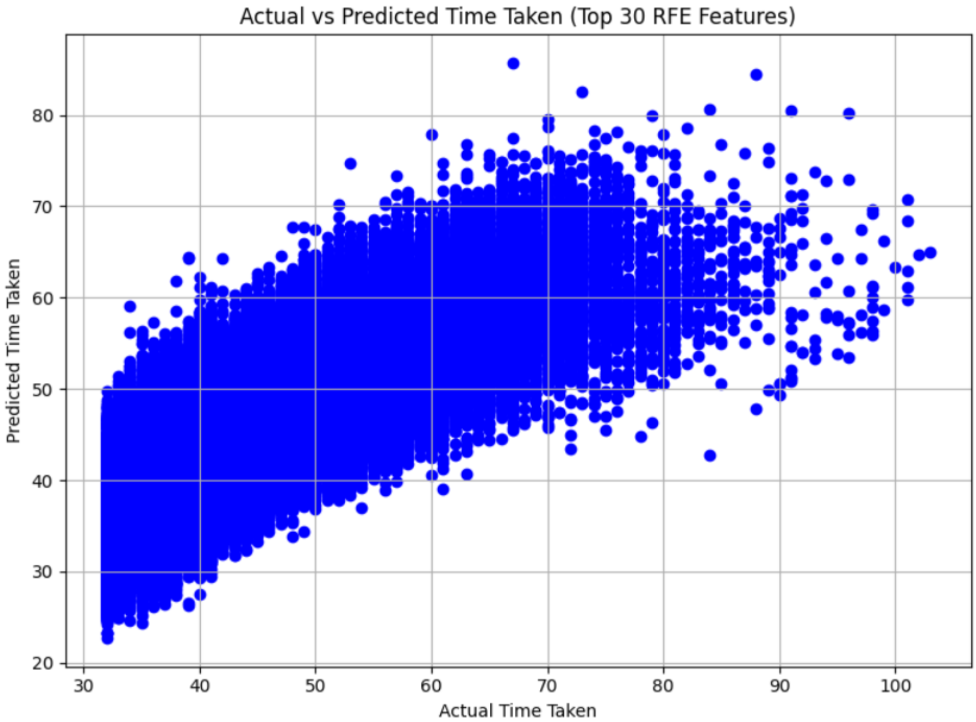
One of the approach is to delete one-by-one variable and check Significance and VIF again

Alternate Option to use RFE method

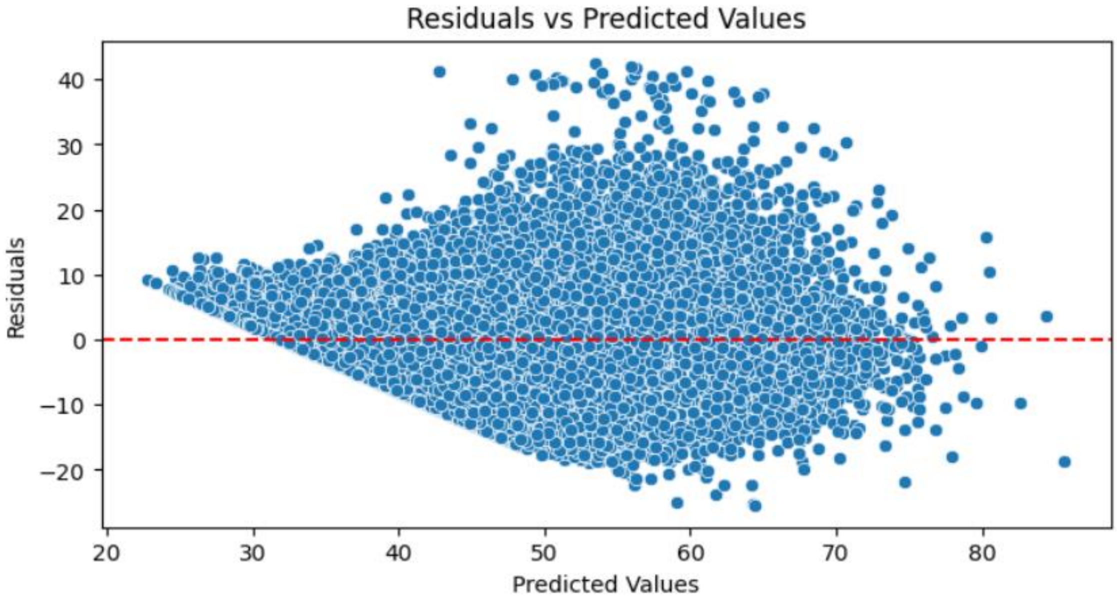
Shyju S R

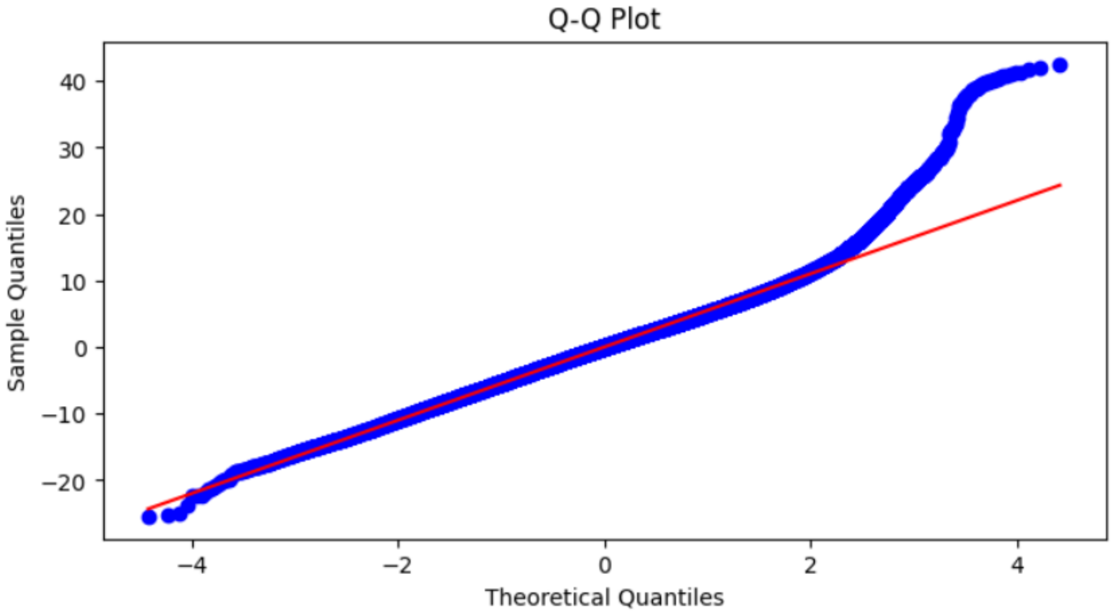


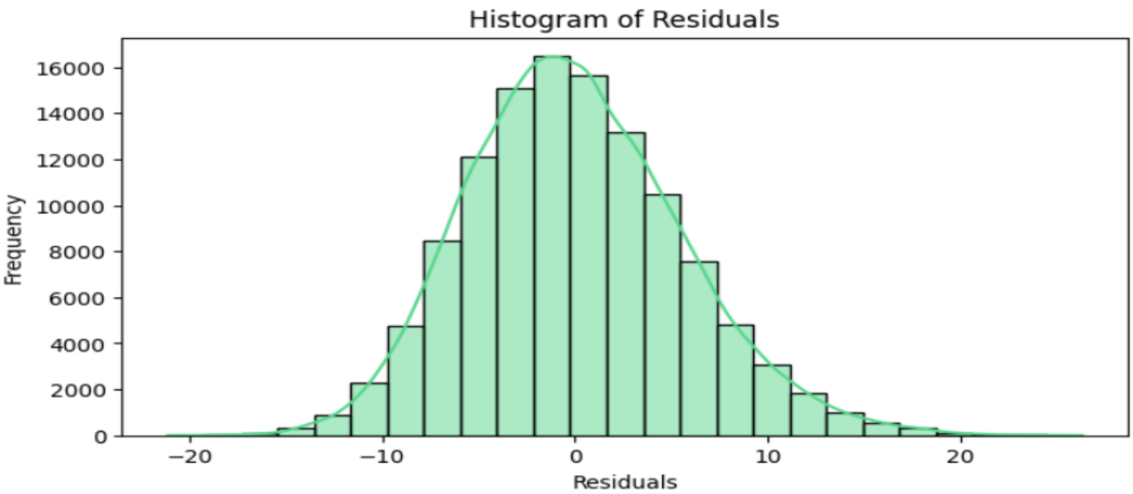
Feature Ranking using RFE
Selected
False
True
Feature
total_items
subtotal
num_distinct_items
max_item_price
total_outstanding_orders
distance
market_id_2
market_id_3
market_id_4
market_id_5
market_id_6
order_protocol_4
order_protocol_5
order_protocol_7
order_hour_7
order_hour_6
order_day_4
isWeekend_1
order_day_1
order_day_3
order_hour_23
order_hour_22
order_hour_21
order_day_2
order_hour_20
order_hour_17
order_hour_16
order_hour_15
order_hour_14
order_hour_8
order_protocol_3
order_hour_19
order_hour_1
order_protocol_6
order_hour_4
order_hour_18
order_hour_2
order_hour_3
order_protocol_2
order_day_5
order_day_6
store_primary_category_28
store_primary_category_38
store_primary_category_39
store_primary_category_55
store_primary_category_58
store_primary_category_24
store_primary_category_100
order_hour_5
store_primary_category_46
store_primary_category_13
store_primary_category_20



Actual vs Predicted Time Taken (Top 30 RFE Features)
Predicted Time Taken
Actual Time Taken
20
30
40
50
60
70
80
30
40
50
60
70
80
90
100


Residuals vs Predicted Values
Residuals
Predicted Values
40
30
20
10
0
−10
−20
20
30
40
50
60
70
80


Q-Q Plot
Sample Quantiles
Theoretical Quantiles
40
30
20
10
0
−10
−20
−4
−2
0
2
4


Histogram of Residuals
Frequency
Residuals
16000
14000
12000
10000
8000
6000
4000
2000
0
−20
−10
0
10
20

1. **Residuals vs Predicted Plot**

Residual = Actual time taken – Predicted time taken

Positive residuals (points above the red dashed 0-line) mean the model under-predicted: the real delivery took longer than forecast.

Negative residuals (points below the 0-line) mean the model over-predicted: it thought the delivery would take longer than it did.

The plot shows data points randomly scattered around the horizontal zero line without any clear patterns or trends. This confirms the linearity assumption of the regression model - the relationship between predictors and target variable is appropriately captured by our linear model

2. **Q-Q Plot of Residuals**

The data points follow the diagonal reference line fairly closely. The residuals approximate a normal distribution, satisfying another key regression assumption. This validates that our statistical inferences (p-values, confidence intervals) from the model are trustworthy

3. **Histogram of Residuals**

The distribution is approximately bell-shaped and cantered at zero. Confirms the normal distribution of errors with a mean of zero. Indicates our model is well-balanced in its predictions, not systematically biased in either direction

Overall The model performs OK on both training data ($R^2 \approx 0.658$)

These diagnostics suggest the model provides reliable predictions and that the coefficients can be confidently interpreted for business insights.

Based on the combined analysis of OLS regression summary and VIF values, the top 3 most significant features influencing the prediction of time_taken are:

- **total_items**
- **subtotal**
- **no_of_distinct_items**