Order Delivery Time Prediction

Objective

To build a predictive model that estimates the delivery time of orders based on various features, using exploratory data analysis and linear regression techniques.

Data Overview & Cleaning

Dataset Insights:

The dataset contains fields such as delivery distances, order created date etc.

Data Cleaning:

There are no missing values. Applied transformations like converting categorical variables, normalizing features where needed. Created dummy variables say isWeekend for categorical inputs.

Dropped order protocol the integer code indicating how the order was placed (e.g., viaPorter or call to restaurant. store\_primary\_category, order\_protocol, total\_items, num\_distinct\_items etc are irrelevant for predicting delivery time

Exploratory Data Analysis (EDA)

1. Key Visualizations:

- Histograms: Distribution of delivery duration.

Plot distributions for all numerical columns

A group of blue and black graphs

AI-generated content may be incorrect.

Frequency represents delivery duration time

1. For market id : There are clear spikes at integer values like 1, 2, 3, 4, and 5.

The distribution is **non-uniform**, suggesting some markets handle significantly more deliveries than others.

1. For distribution of subtotal :

The histogram is **right-skewed** (positively skewed).

1. For distribution of distance:

Appears well-behaved; likely a strong predictor of delivery time.

For categorical values :

A graph with a green and orange square

AI-generated content may be incorrect.

The majority of deliveries happen on **weekdays**, nearly **double** the count compared to weekends.

# Distribution of time\_taken

A graph of a distribution of delivery duration

AI-generated content may be incorrect.

The distribution is **right-skewed** (positively skewed). The **peak (mode)** appears around **45 minutes**, suggesting this is the most common delivery time.

**Boxplots:** To observe outliers in features like delivery duration and hour of day.

Show the distribution of time\_taken for different hours

A graph with blue squares and black lines

AI-generated content may be incorrect.

Wider interquartile ranges (IQR) and a large number of outliers indicate inconsistent delivery performance around early morning and operations **ramp up smoothly** during the early business hours. Afternoon to Evening (12–20) is the **most efficient window**, likely due to full operational capacity. Late evening median still remains moderate, but outliers start showing again—could relate to **end-of-day deliveries** or **reduced traffic**.

**Correlation Matrix Heatmap:** Identified multicollinearity.Check correlations between numerical features to identify which variables are strongly related to time taken

A screenshot of a graph

AI-generated content may be incorrect.

|  |  |  |
| --- | --- | --- |
| Feature | Correlation | Description |
| Distance | 0.46 | Moderate positive correlation: **Longer distances take more time**, as expected |
| Subtotal | 0.41 | Moderate positive correlation: **Higher order value might reflect complexity or larger orders**, increasing delivery time. |
| isWeekend | 0.13 | Weak positive correlation: Slightly longer deliveries on weekends. |
| Order\_hour | -0.34 | **Order Hour** has a **negative correlation**, meaning deliveries are quicker later in the day. |

order\_dayofweek, market\_id are Very weak correlation

**2. Key Insights:**

Delivery time tends to increase with longer distances. Certain weather and traffic conditions also contributed to longer delivery times.

**Delivery Duration Distribution**

The delivery duration is right-skewed.

Most deliveries happen within 35–55 minutes.

A few outliers stretch delivery time to 90+ minutes.

**Top Influencing Features** are

**Distance** and **Subtotal** are moderately positively correlated with delivery\_duration

**Order Hour** has a **negative correlation**, meaning deliveries are quicker later in the day.

Distance distribution is bell ideal for linear modeling.

Model Building

Algorithm

Linear Regression via sklearn and statsmodels.

Model Evaluation

Train-test split was performed (80:20 ratio).

**In training data:**

Intercept: 46.192656225872256  
This is the **baseline delivery time in minutes** when both distance and subtotal are zero. It reflects fixed time like order preparation, pickup time, etc.

Coefficients: [3.71011389 4.16270964]

For **each additional unit of distance**, the delivery time increases by **~3.71 minutes**, assuming subtotal remains constant.

**Subtotal Coefficient (4.16)**:  
For **each unit increase in subtotal**, the delivery time increases by **~4.16 minutes**, assuming distance remains constant.

**Performance metrics included:**Top 3 features are distance, subtotal and order\_hour.

| **Metric** | **Value** | **Interpretation** |
| --- | --- | --- |
| **R² Score** | **0.367** | ~36.7% of the variance in delivery time is explained by the model. Not perfect, but a decent baseline for linear models. |
| **Mean Absolute Error** (MAE) | **5.68 minutes** | On average, the model's predictions are off by about 5.7 minutes. |
| **Root Mean Squared Error** (RMSE) | **7.41 minutes** | Indicates the typical size of the prediction error. Larger errors are penalized more than MAE. |

The model performs **reasonably well**, especially given the simplicity of a linear approach and limited features.

**Model diagnostics:**  
- Residual plot

A diagram of a delivery duration

AI-generated content may be incorrect.

Residuals should be **randomly scattered** around zero with no visible pattern. The residual tends to show **heteroscedasticity** (non-constant variance), which **violates a key assumption** of linear regression.

A graph with orange lines

AI-generated content may be incorrect.

In Histogram of Residuals aren’t perfectly normal.It does limit the precision of confidence intervals and p-values.

A graph with a line going up

AI-generated content may be incorrect.  
In Q-Q Plot (Quantile-Quantile Plot) compares actual residual distribution with a theoretical normal distribution. We can see significant **deviation from the diagonal** line, especially in the tails. Here Residuals are not normally distributed.

Conclusion

A Linear Regression model was built to estimate delivery duration using key features like distance, subtotal, and order\_hour.

The model achieved an R² score of 0.367, explaining roughly 37% of the variance in delivery time.

Both distance and subtotal were positively correlated with delivery time, while order\_hour showed a negative correlation, indicating faster deliveries later in the day.

Model coefficients were interpretable and aligned with real-world expectations (e.g., longer distance = longer time).

However, residual analysis revealed non-normality and heteroscedasticity, violating linear regression assumptions.