Assignment :

LR Delivery Time Prediction

Ankit Bougal

Project Objective

- 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

Business Value

The growing demand for quick and efficient delivery in the logistics industry calls for the development of systems that can predict delivery times accurately. Porter, an intracity logistics marketplace, services millions of customers daily, and optimising delivery times is crucial for improving operational efficiency.

Data Preprocessing & Feature Engineering

- Converted timestamps to datetime
- Extracted created_hour, isWeekend
- Dropped unused fields: created_at, actual_delivery_time
- Converted categorical columns to appropriate type (market_id, order_protocol)

Train-Test Split

- Defined X and y:
  - y = delivery_time
  - X = all other features
- Used train_test_split(test_size=0.2, random_state=100) to create X_train, X_test, y_train, y_test

Distribution of Numeric Features in Training Set

(picked 3 of the most insightful features)

- **distance** (Bell-shaped, nearly normal) –

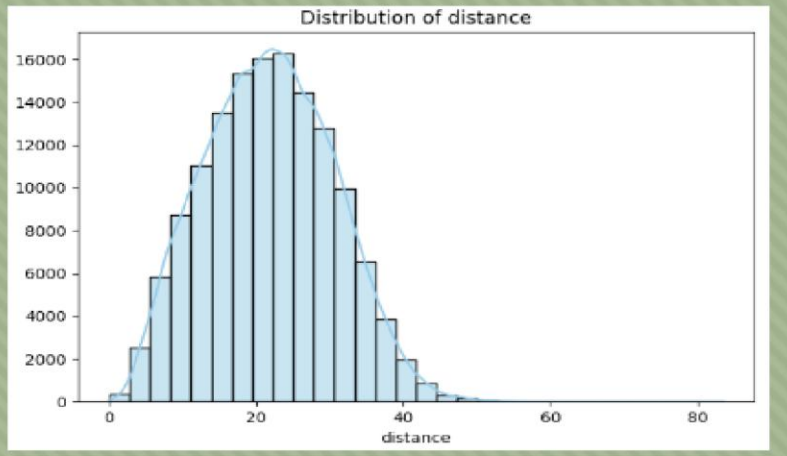
Unlike other features, distance follows a near-normal distribution, peaking around 25–30 km. This makes it statistically reliable and one of the strongest predictors of delivery time.

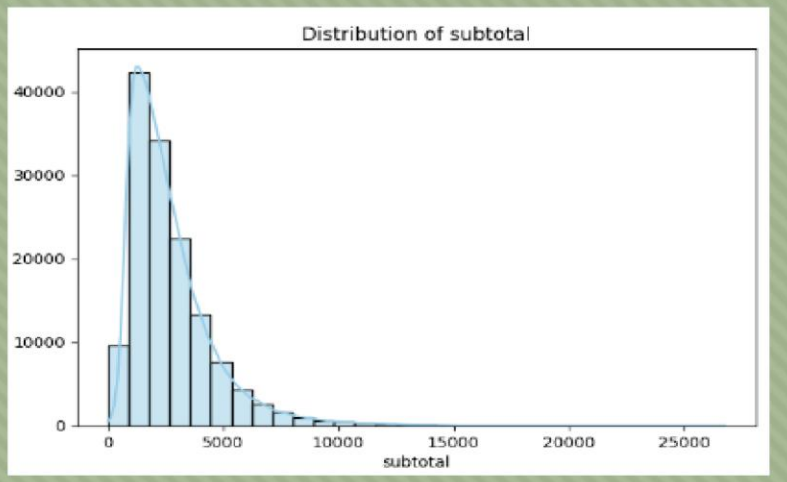
- **subtotal** (Highly right-skewed) –

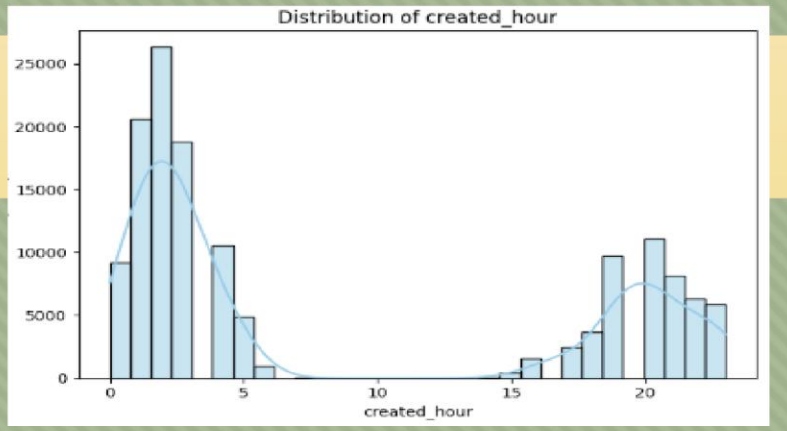
The subtotal feature is highly right-skewed, with most orders clustered below ₹5000. A small proportion of high-value orders extend the distribution to ₹20,000+, which justifies both outlier handling and feature scaling before modelling.

- **created_hour** (Bimodal distribution) –

The created_hour distribution is bimodal, with peaks around early morning and evening hours — likely reflecting breakfast and dinner rushes. This supports its role in explaining variation in delivery times

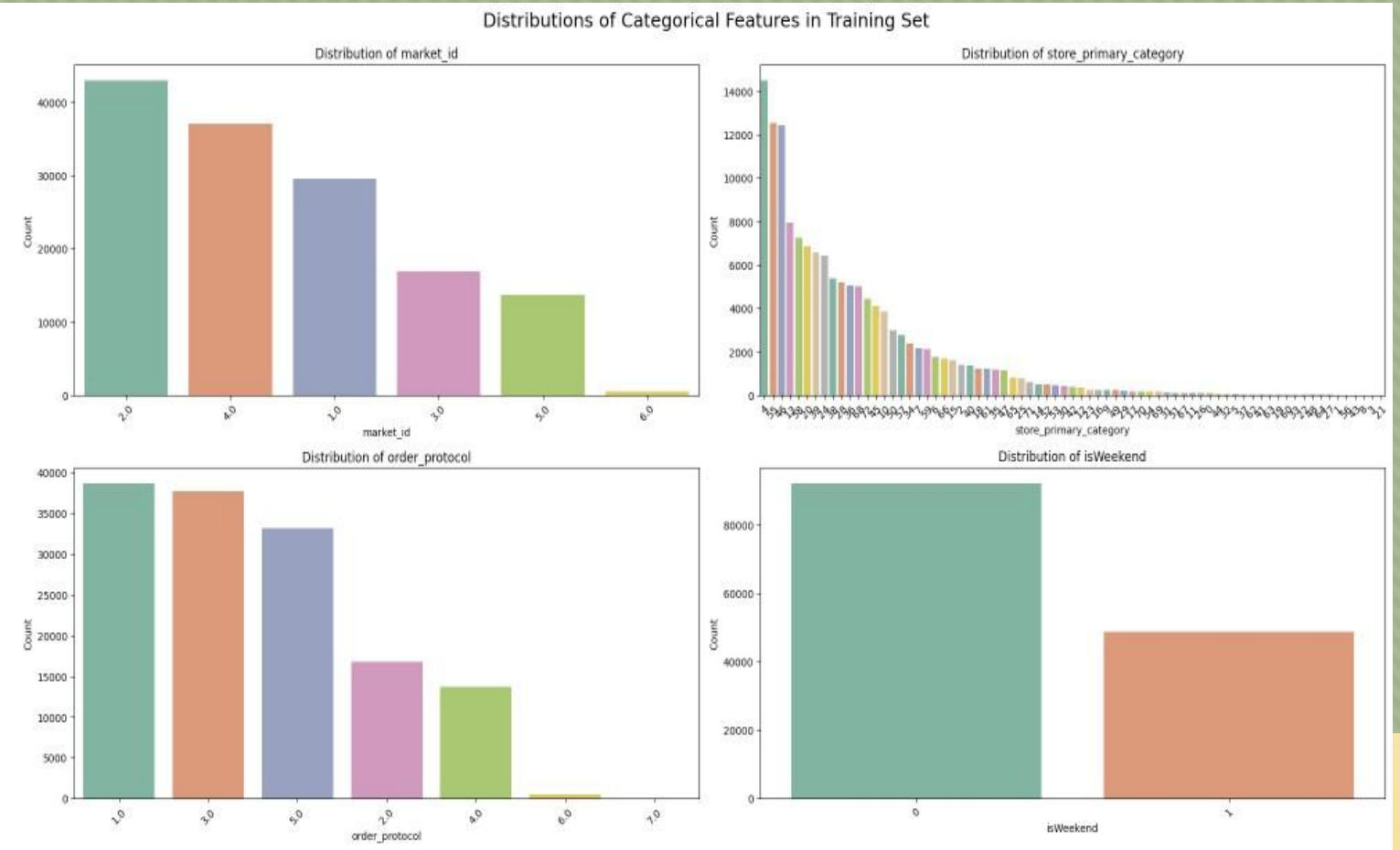






Distribution of Categorical Features in Training Set

- **market_id** - Most orders come from markets 2, 4, and 1.
- **store_primary_category** - Highly skewed distribution
- **order_protocol** - Most orders placed via protocols 1 to 3 (possibly app/web). Protocols 5+ are rare — minimal contribution.
- **isWeekend** - Around 65% of orders are placed on weekdays. Still a significant 35% placed on weekends, justifying its inclusion in the model.



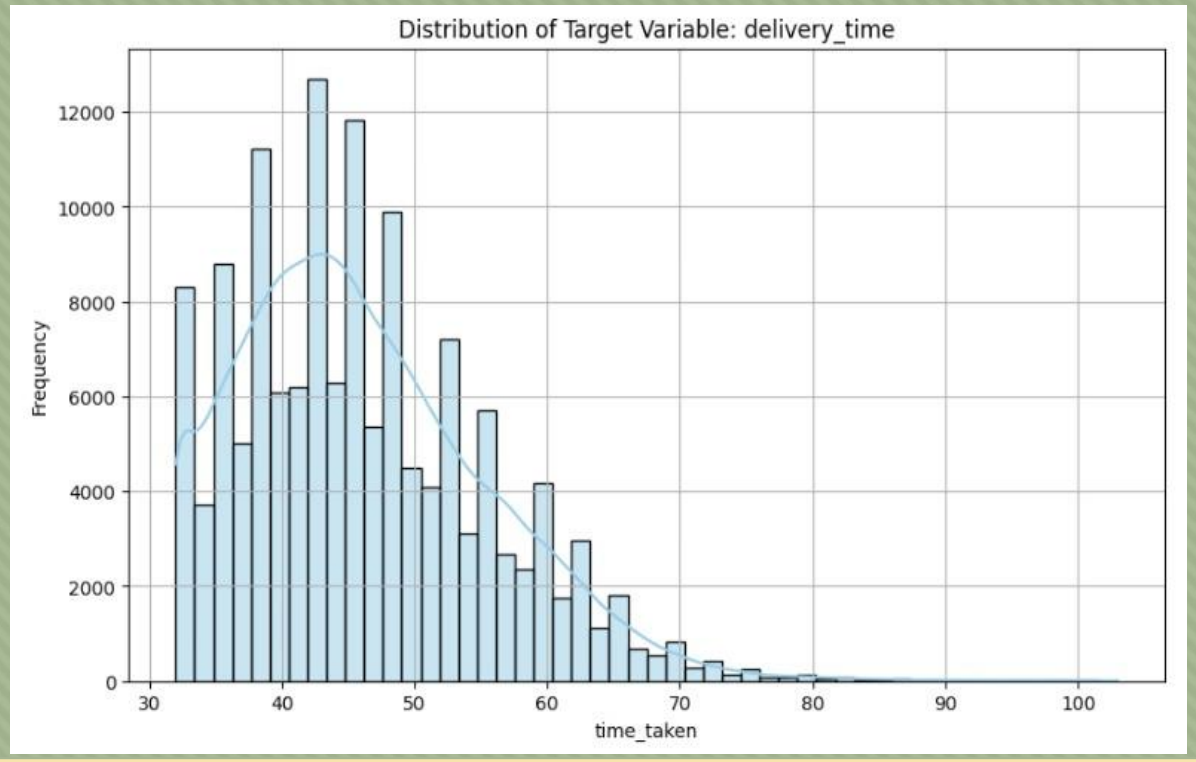
Target Variable Distribution

Histogram of delivery_time

- Slight right-skew, with some deliveries exceeding 100 minutes
- These are likely the extreme delays, and were addressed in your model by capping delivery_time at 72 minutes during outlier handling

Central Tendency

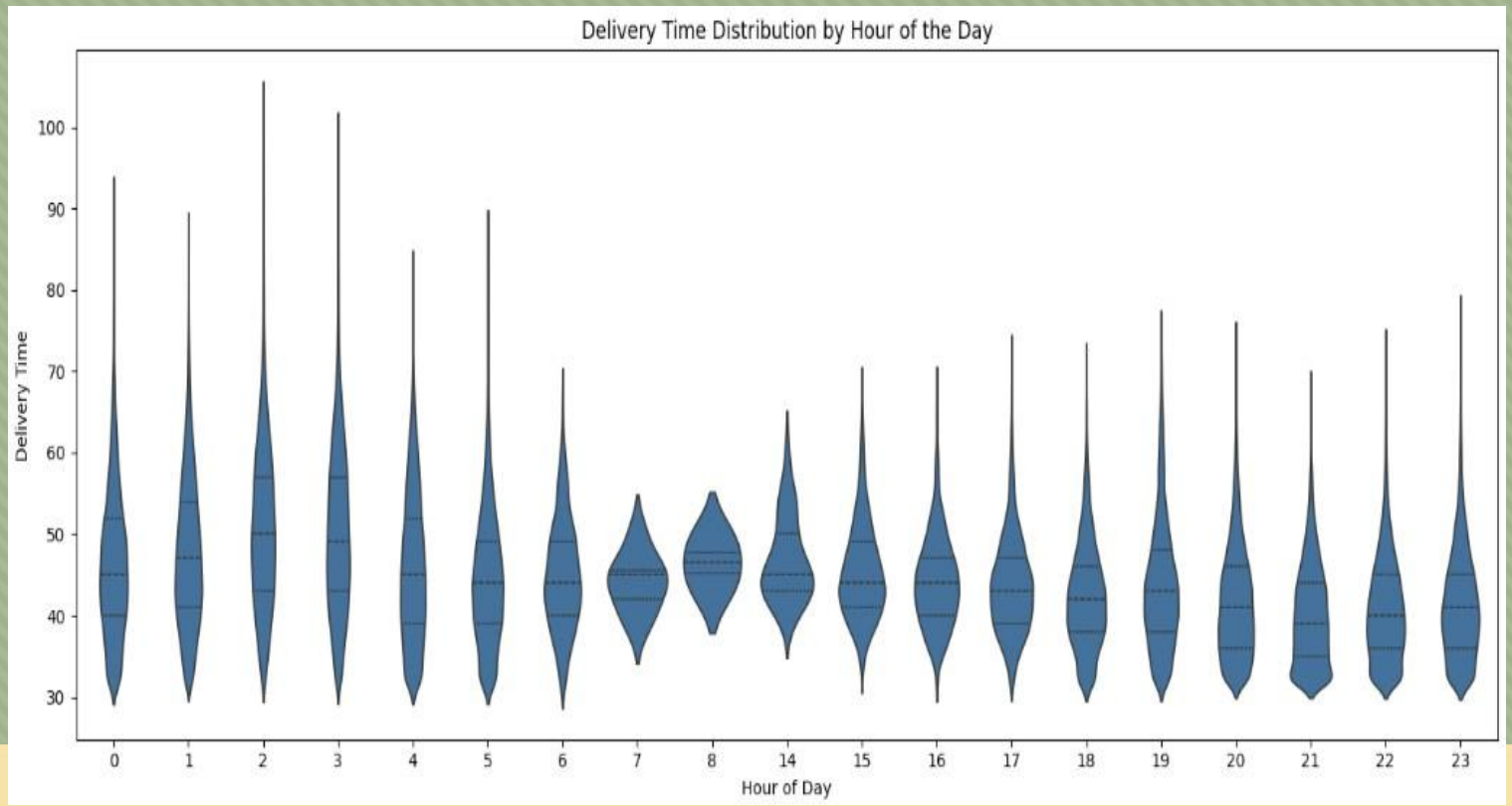
- The peak frequency (mode) appears around 40–45 minutes.
- This suggests that the typical delivery time is around 40–50 minutes, which aligns with operational expectations.



Distribution of time_taken for different hours

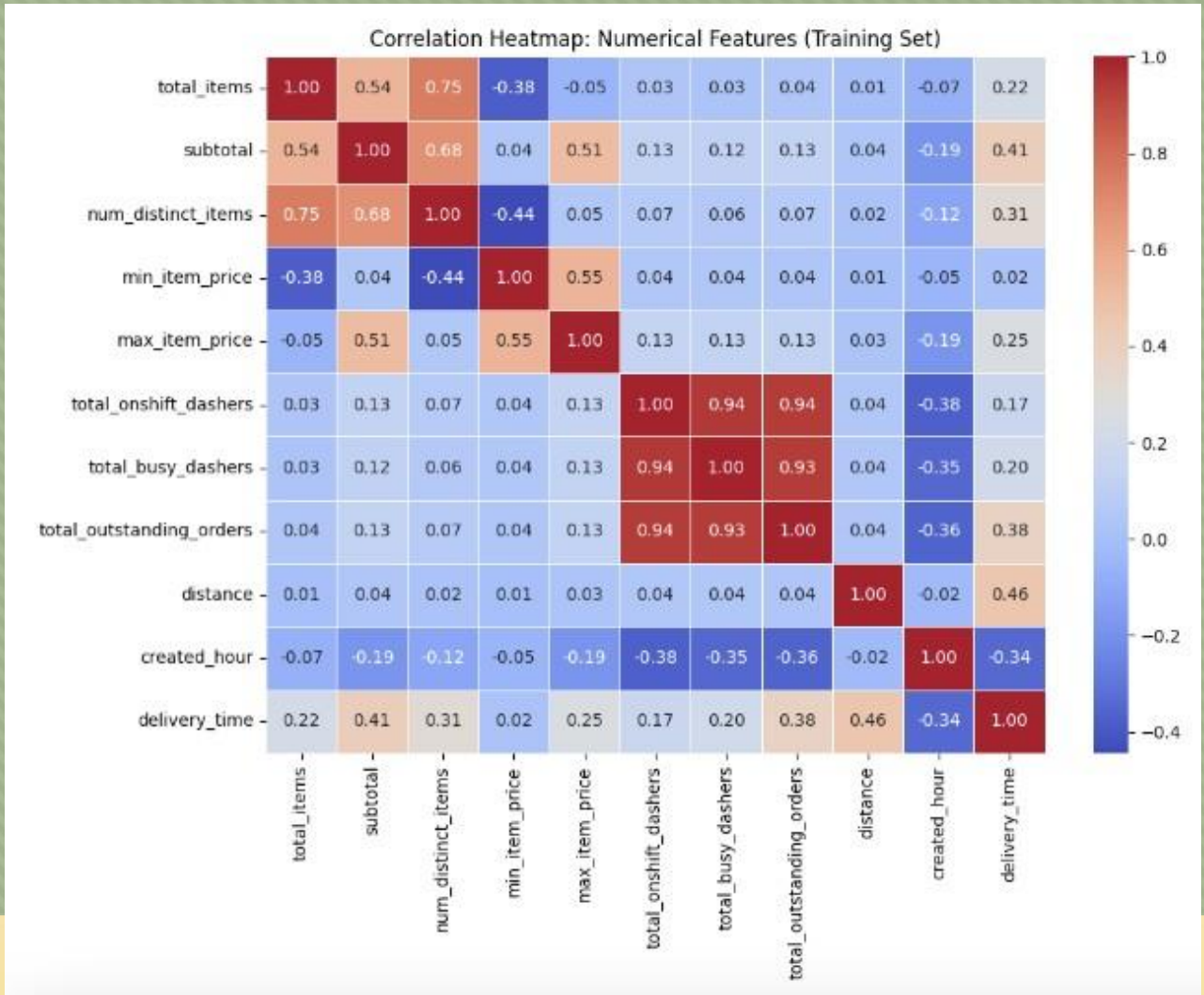
Delivery performance is **most efficient from 14:00 to 23:00**, where both median and spread are low.
Early hours (0–6) are **less predictable**, likely due to fewer available dashers or more variability in traffic/routes.

The violin plot shows that delivery times are higher and more variable during early morning hours (0–6 AM), and improve steadily throughout the day, with the most consistent performance seen between 14:00–23:00.
This validates the model's negative coefficient for created_hour, confirming that later orders tend to be delivered faster.



Correlation Analysis

- Heatmap created for all numerical features and delivery_time
- Top positively correlated features:
  - distance (0.46)
  - subtotal (0.41)
  - total_outstanding_orders (0.38)
- Top negatively correlated feature:
  - created_hour (-0.34)
- Dropped weakly correlated: min_item_price, total_busy_dashers, total_onshift_dashers



Outlier Detection and Removal

- Used boxplots to visualize and IQR method to remove outliers
- Cap applied:
  - delivery_time ≤ 72 minutes
- Removed outliers from all features using Q1 - 3*IQR to Q3 + 3*IQR

Feature Scaling

- Applied MinMaxScaler to numerical features after outlier removal
- Ensured comparability across variables during regression

Model 1 – All Features (Statsmodels)

Initial Model Summary (Before Feature Selection)

- Built using all cleaned and scaled features (11 predictors)
- R^2: 0.533 — indicates ~53% of the variance in delivery time is explained.
- All features statistically significant (p-values < 0.05)
- Durbin-Watson ~2: no autocorrelation in residuals

RFE was still needed, to deal with-

- Features with multicollinearity (e.g., subtotal, max_item_price)
- Redundant or correlated predictors (e.g., total_items vs. num_distinct_items)
- Less interpretable/operationally redundant variables

OLS Regression Results

Dep. Variable:	delivery_time	R-squared:	0.533
Model:	OLS	Adj. R-squared:	0.533
Method:	Least Squares	F-statistic:	1.403e+04
Date:	Tue, 06 May 2025	Prob (F-statistic):	0.00
Time:	06:02:56	Log-Likelihood:	-4.3280e+05
No. Observations:	135024	AIC:	8.656e+05
Df Residuals:	135012	BIC:	8.658e+05
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	33.9534	0.088	384.736	0.000	33.780	34.126
market_id	-0.6245	0.012	-50.853	0.000	-0.649	-0.600
store_primary_category	0.0024	0.001	3.035	0.002	0.001	0.004
order_protocol	-0.9287	0.011	-84.846	0.000	-0.950	-0.907
total_items	-0.4786	0.198	-2.416	0.016	-0.867	-0.090
subtotal	11.1640	0.196	57.096	0.000	10.781	11.547
num_distinct_items	4.4816	0.200	22.382	0.000	4.089	4.874
max_item_price	0.7770	0.164	4.734	0.000	0.455	1.099
total_outstanding_orders	13.5894	0.099	136.763	0.000	13.395	13.784
distance	26.2620	0.109	241.159	0.000	26.049	26.475
created_hour	-4.0615	0.047	-86.068	0.000	-4.154	-3.969
isWeekend	1.8714	0.035	54.243	0.000	1.804	1.939

Omnibus:	1977.088	Durbin-Watson:	2.002
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2114.496
Skew:	0.280	Prob(JB):	0.00
Kurtosis:	3.251	Cond. No.	720.

Final Model

Model Type: OLS (Ordinary Least Squares)

Number of Predictors: 8 (selected via Recursive Feature Elimination)

R-squared: **0.524**

- The model explains **52.4%** of the variance in delivery_time

Adj. R-squared: 0.524 (no overfitting from extra features)

F-statistic: **18,620** (Prob: 0.00) — very strong statistical significance

Durbin-Watson: **~2.00** — Residuals are independent (no autocorrelation)

OLS Regression Results

Dep. Variable:	delivery_time	R-squared:	0.524
Model:	OLS	Adj. R-squared:	0.524
Method:	Least Squares	F-statistic:	1.862e+04
Date:	Tue, 06 May 2025	Prob (F-statistic):	0.00
Time:	06:03:02	Log-Likelihood:	-4.3409e+05
No. Observations:	135024	AIC:	8.682e+05
Df Residuals:	135015	BIC:	8.683e+05
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	32.3545	0.078	417.407	0.000	32.203	32.506
order_protocol	-0.9101	0.011	-82.678	0.000	-0.932	-0.889
subtotal	11.1449	0.177	63.015	0.000	10.798	11.492
num_distinct_items	3.9618	0.139	28.557	0.000	3.690	4.234
max_item_price	0.8885	0.149	5.954	0.000	0.596	1.181
total_outstanding_orders	13.2555	0.100	132.712	0.000	13.060	13.451
distance	26.1476	0.110	237.897	0.000	25.932	26.363
created_hour	-4.1023	0.048	-86.124	0.000	-4.196	-4.009
isWeekend	1.8886	0.035	54.246	0.000	1.820	1.957

Omnibus:	3272.915	Durbin-Watson:	2.003
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3603.966
Skew:	0.366	Prob(JB):	0.00
Kurtosis:	3.323	Cond. No.	50.5

Residual Analysis

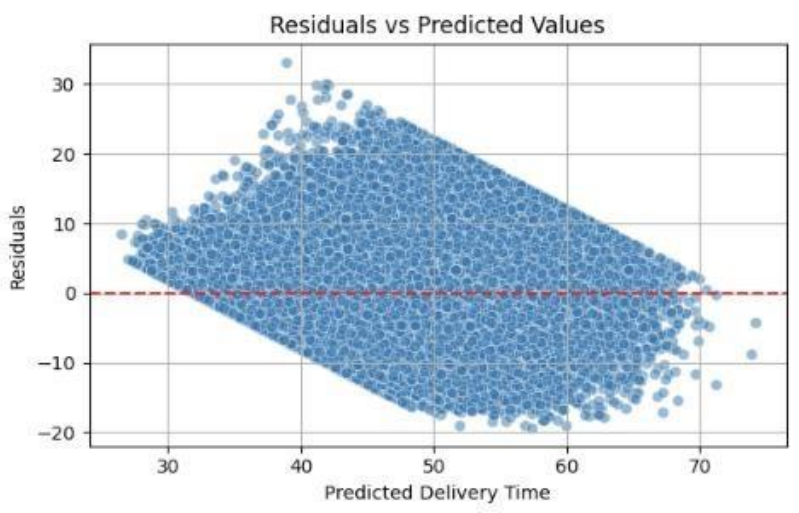
Residuals vs Predicted No major pattern

What we see:

- Residuals are centered around zero, which is good.
- However, there's a funnel pattern, suggesting some heteroscedasticity.

Interpretation:

- The linearity assumption is largely satisfied.
- Slight non-constant variance may exist



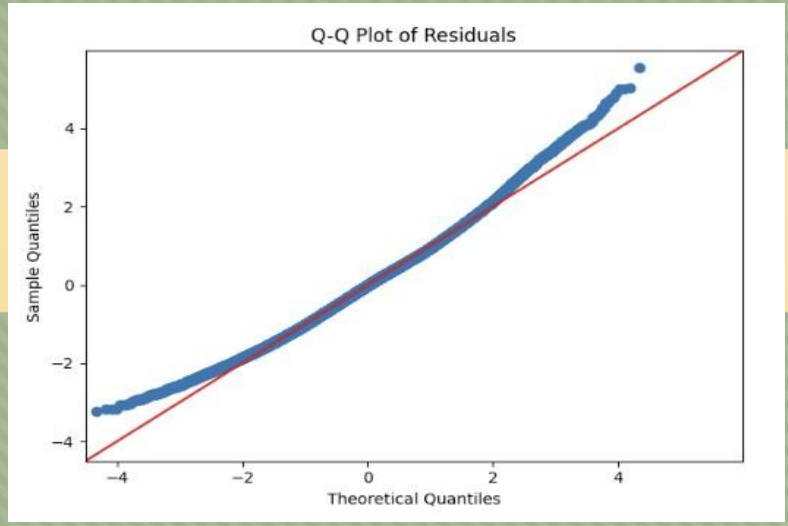
Q-Q Plot:holds normality

What we see:

- Most points lie on the 45° line → residuals follow a nearly normal distribution.
- Slight deviations at the tails (mild skew or outliers).

Interpretation:

- Normality assumption is reasonably satisfied.
- Minor tail deviations are expected in real-world data.



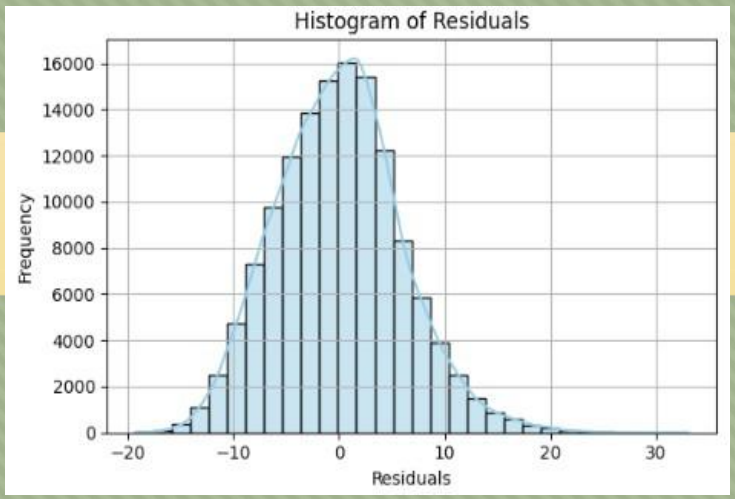
Histogram: Symmetric, bell-shaped residual distribution

What we see:

- Bell-shaped, symmetric distribution centered around zero.

Interpretation:

- Confirms that residuals are roughly normal.
- Supports the statistical validity of your model inference.



Final Insights for Selected features

- **distance** - Longer distances increase delivery time the most (~0.45 min/km)
- **total_outstanding_orders**- Higher backlog at the restaurant significantly delays deliveries
- **subtotal** - Expensive orders often take more time to prepare
- **num_distinct_items** - More unique items = longer prep = longer delivery
- **order_protocol** - Digital ordering channels (higher protocol value) lead to faster deliveries
- **created_hour** - Deliveries later in the day are slightly faster — likely off-peak hours
- **max_item_price** - Higher-priced items might indicate complex or premium orders
- **isWeekend** - Deliveries are slower on weekends due to traffic and higher order volume

Final Insights – Top indicators

- **distance** is the strongest drivers of delivery time. Each extra km adds ~0.45 minutes to delivery time
- **total_outstanding_orders** is second strongest driver of delivery time — operational decisions should prioritize minimizing these. Each additional pending order adds ~0.047 minutes (~2.8 sec) delay
- **subtotal -** contributes meaningfully to timing expectations. For every ₹100 increase in subtotal, delivery time increases by 0.13 minutes
- **order_protocol** plays a key role — encouraging digital channels may reduce delay. For each unit increase (e.g., phone → app), delivery time drops by ~0.18 minutes
- **Day of Week matters** — orders during weekdays and late hours tend to be quicker. Weekend deliveries take ~1.9 minutes longer (binary variable, not scaled)