ONLINE FOOD DELIVERY DATA ANALYSIS

Author: Bharathi Jagadeesan

Domain: Data Analytics

Tools: Python, Pandas, NumPy, Matplotlib, Seaborn, MySQL, Power BI

Dataset Size: 1,00,000 records

1. INTRODUCTION

The rapid growth of online food delivery platforms has transformed how customers order food and how restaurants manage their operations. With increasing competition among delivery platforms, data-driven decision-making has become essential to improve customer satisfaction, operational efficiency, and profitability. This project focuses on analyzing an online food delivery dataset containing customer details, order information, restaurant attributes, delivery performance metrics, and financial data. By applying data cleaning, exploratory data analysis (EDA), feature engineering, and visualization techniques, meaningful insights are extracted to understand customer behavior, delivery efficiency, revenue trends, and business performance.

The objective of this project is to uncover hidden patterns and trends within the dataset that can help stakeholders such as food delivery companies, restaurants, and logistics teams make informed strategic decisions. The analysis emphasizes real-world business logic, ensuring that insights are practical and actionable.

Objectives of the Project

The key objectives of this project are:

- To clean and preprocess raw food delivery data
- To handle missing values, inconsistencies, and outliers
- To perform Exploratory Data Analysis (EDA) for identifying trends and patterns
- To analyze customer behavior, delivery efficiency, and revenue generation
- To evaluate the impact of discounts and peak hours on profit
- To store cleaned data in MySQL for querying
- To build an interactive Power BI dashboard for business insights

2. DATASET OVERVIEW

The dataset used in this project consists of approximately one lakh (100,000) records and multiple features covering different aspects of the food delivery ecosystem.

2.1 Data Categories

- **Customer Information**: Age, Gender, City, Area
- **Order Details**: Order ID, Order Date, Order Value, Discount
- **Restaurant Details**: Restaurant ID, Name, Cuisine Type, Ratings
- **Delivery Performance**: Delivery Time, Distance, Delivery Rating
- **Financial Metrics**: Cost, Profit Margin, Profit Percentage
- **Operational Data**: Payment Mode, Cancellation Status, Reasons

2.2 Data Understanding & Data Quality Check

Before analysis, the dataset was examined for:

- Missing values
- Incorrect data types
- Inconsistent categorical values
- Logical inconsistencies between columns

2.3 Key Issues Identified

- Missing values in age, delivery time, city, payment mode
- Cancelled orders having ratings
- Delivered orders having cancellation reasons
- Outliers in delivery time and order value

3. DATA CLEANING AND PREPROCESSING

Data cleaning is a critical step to ensure data quality and consistency before performing any analysis. The dataset contained missing values, inconsistent entries, invalid ratings, and columns with no meaningful information. Various preprocessing techniques were applied to handle these issues.

3.1 Handling Missing Values

Different strategies were used based on **data nature**:

- **Numerical Columns:**

  Median used for skewed data (delivery time, order value)

- **Categorical Columns:**

Filled with "Unknown" when inference was not possible

- **Logical Dependency Fixes:**

  Cancelled orders → delivery_rating set to NaN

  Delivered orders → cancellation_reason set to "Not Cancelled"

Data Type Correction

- Converted numeric fields from float to int where applicable
- Retained NaN for ratings (outcome-based metrics)
- Converted date fields to datetime

- Logical consistency checks were also performed. For example, delivery ratings were removed for cancelled orders because a cancelled order cannot logically have a delivery rating.
- Similarly, cancellation reasons were corrected based on order status. Columns that contained no useful information, such as order time with all zero values, were removed to avoid misleading analysis.
- Data types were validated and corrected, ensuring numeric fields were stored as integers or floats as appropriate, and date fields were converted to datetime format.
- This preprocessing ensured the dataset was reliable and analysis-ready.

3.2 Outlier Detection and Treatment

- Outliers are extreme values that deviate significantly from the majority of observations and can distort analysis results. In this project, outliers were examined for delivery time and order value. These variables are naturally right-skewed in business datasets, making traditional methods like Z-score less appropriate.
- The Interquartile Range (IQR) method was used to detect outliers. This approach is robust to skewed distributions and focuses on the middle 50% of the data.
- The decision to cap instead of remove outliers was based on the understanding that unusually high delivery times or order values can still be valid in real-world scenarios, such as long-distance deliveries or bulk orders.

Insight:

After cleaning, most values fell within acceptable business ranges, indicating improved data quality.

Standardization of Categorical Values

- Uniform casing (lowercase / title case)
- Removed extra spaces using strip()
- Unified labels like UPI, COD, Wallet, Card

4. EXPLORATORY DATA ANALYSIS (EDA)

EDA was performed to uncover hidden patterns, trends, and relationships.

4.1 Distribution Analysis

Order Value Distribution

- Right-skewed distribution
- Majority orders fall in mid-price range
- Few high-value orders drive revenue

Delivery Time Distribution

- Most deliveries occur within expected time
- Long-tail delays observed but minimal

Insight:

Business performance is stable with few extreme cases.

4.2 City-wise & Cuisine-wise Analysis

- Orders distributed evenly across major cities
- No single city dominates excessively
- Cuisine demand is balanced across types

Insight:

Platform has diversified customer demand across regions and cuisines.

4.3 Weekday vs Weekend Orders

- Weekdays: ~71% of orders
- Weekends: ~29% of orders

Insight:

Food delivery demand is significantly higher on weekdays, indicating workday dependency.

4.4 Distance vs Delivery Time

- Very weak correlation observed
- Distance alone does not determine delay

Insight:

Traffic, restaurant prep time, and peak hours play larger roles than distance.

4.5 Cancellation Reason Analysis

- Majority cancellations due to customer unavailability or delay
- Some cancellations lacked reasons → handled logically

Insight:

Improving delivery predictability can reduce cancellations.

4.6 Correlation Analysis

Heatmap analysis showed:

- Weak correlation between distance and delivery time
- Profit margin moderately affected by discount
- Order value strongly impacts revenue

Insight:

Discount strategy must be optimized carefully.

5. FEATURE ENGINEERING

- Feature engineering was applied to enhance the analytical value of the dataset. New derived columns were created to provide deeper insights.
- Customer age groups were created to analyze ordering behavior across demographic segments.
- Delivery performance categories such as "Fast", "On Time", and "Delayed" were derived from delivery time ranges to evaluate logistics efficiency.
- Profit margin percentage was calculated using cost and final amount to standardize profitability analysis across orders. Peak hour indicators were used to identify demand surges and assess their impact on delivery time and cancellations.
- These engineered features improved interpretability and allowed for more meaningful business insights.

Engineered Features:

- **Order Day Type:** Weekday / Weekend
- **Delivery Performance:** Fast / On-Time / Delayed
- **Profit Margin Percentage**
- **Customer Age Groups**
- **Peak Hour Indicator**

These features enabled deeper behavioral and operational analysis.

6. SQL ANALYSIS AND DATABASE INTEGRATION

- The cleaned dataset was stored in a MySQL database to enable structured querying and advanced analysis.

- SQL queries were used to analyze top-spending customers, age group versus order value, weekend versus weekday demand patterns, monthly revenue trends, and the impact of discounts on profitability.
- Grouping, aggregation, and conditional logic in SQL helped uncover insights such as high-revenue city–cuisine combinations, average delivery time by city, cancellation rates by restaurant, and payment mode preferences.
- Integrating SQL with Python notebooks allowed seamless transition between data processing, querying, and visualization.

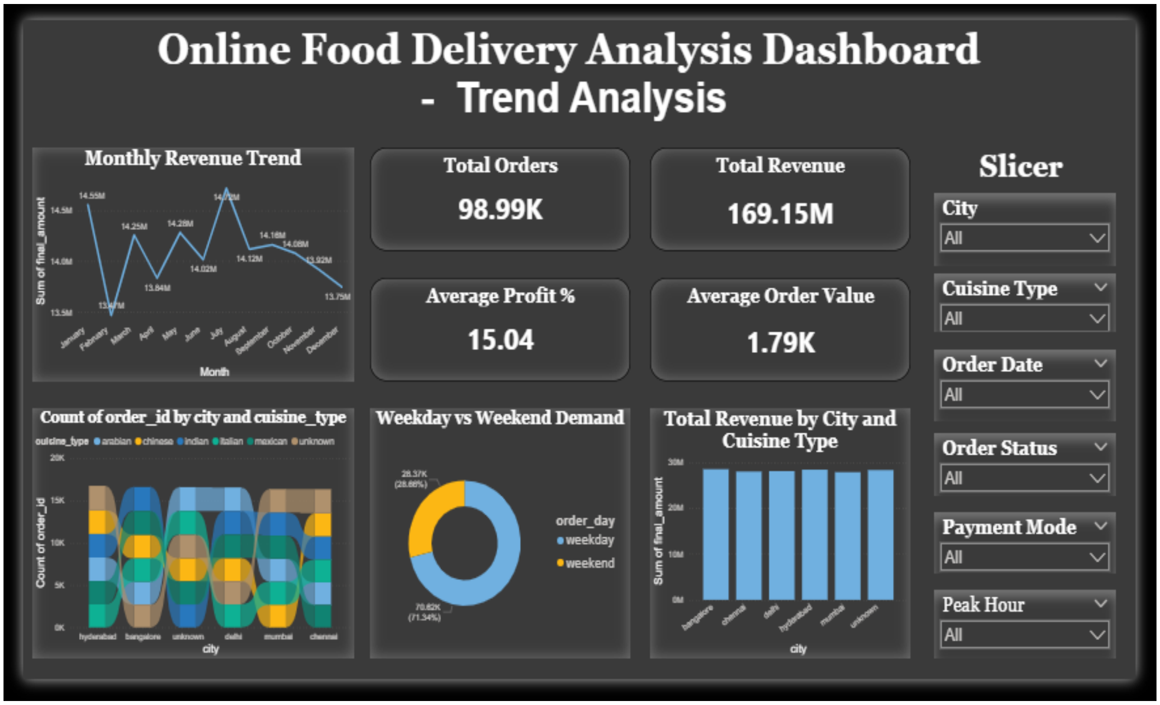
SQL Use Cases:

- Top-spending customers
- High-revenue cities and cuisines
- Discount impact on profit
- Average delivery time by city
- Cancellation rate by restaurant

Insight:

Combining SQL with Python allows scalable analytics and real-world querying.

7. POWER BI DASHBOARD



Online Food Delivery Analysis Dashboard - Trend Analysis

Total Orders
98.99K

Total Revenue
169.15M

Average Profit %
15.04

Average Order Value
1.79K

Slicer

City: All

Cuisine Type: All

Order Date: All

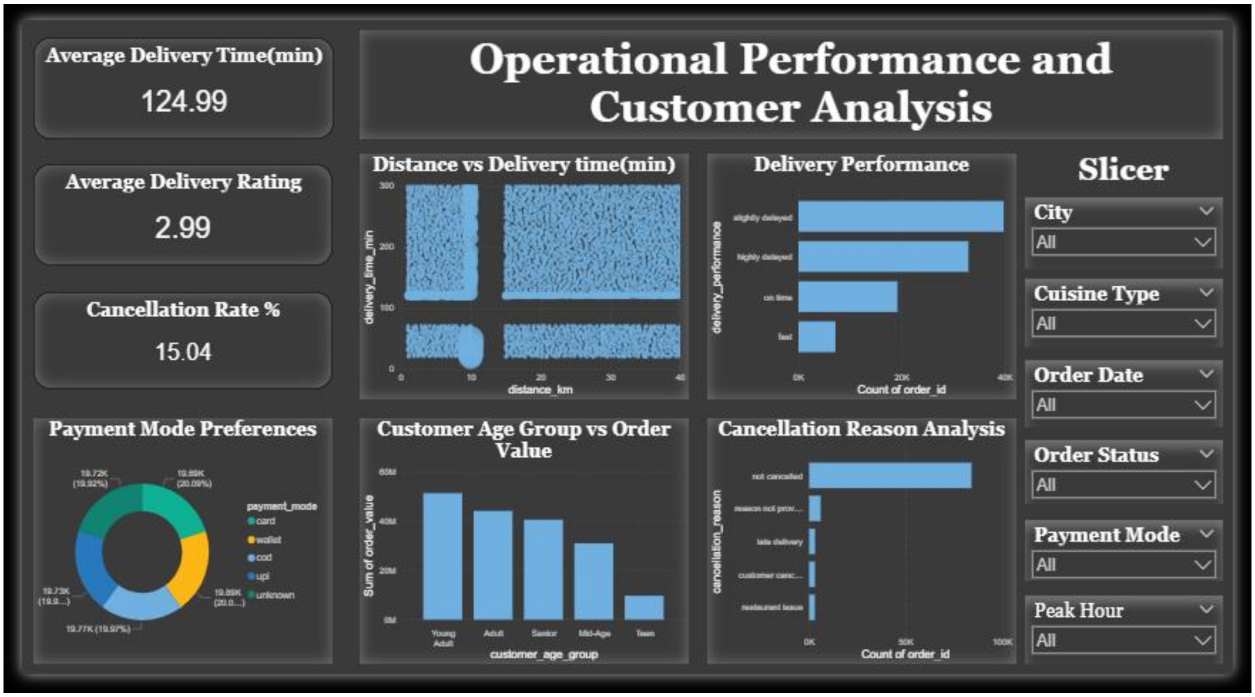
Order Status: All

Payment Mode: All

Peak Hour: All

Fig no 1.1: Power BI Dashboard for Online Food Delivery Analysis – Trend Analysis

- An interactive Power BI dashboard was developed to visualize key performance indicators (KPIs) and trends.
- The dashboard includes metrics such as total orders, total revenue, average order value, average delivery time, cancellation rate, average delivery rating, and profit margin percentage.
- Visual elements include bar charts, line charts, pie charts, and KPI cards.
- Filters for city, cuisine type, order date range, and order status enable dynamic exploration of the data.
- The dashboard was designed to fit within a single-page view to provide a comprehensive snapshot of business performance without requiring scrolling.



Operational Performance and Customer Analysis

Average Delivery Time(min)
124.99

Average Delivery Rating
2.99

Cancellation Rate %
15.04

Slicer

City: All

Cuisine Type: All

Order Date: All

Order Status: All

Payment Mode: All

Peak Hour: All

Fig no 1.2: Operational Performance and Customer Analysis

8. INSIGHTS AND BUSINESS IMPACT

- The analysis revealed several important insights. Weekday demand dominates overall order volume, emphasizing the need for strong weekday operations.
- Discounts increase order volume but do not always translate into higher profit margins, indicating the need for optimized discount strategies.
- Delivery delays are not strongly tied to distance, suggesting that operational improvements can significantly enhance delivery performance.
- Balanced city and cuisine demand indicate opportunities for uniform service quality improvements across regions.

9. FUTURE SCOPE AND ENHANCEMENTS

- This project can be extended in several ways. Predictive models can be built to forecast order demand, delivery delays, or cancellation probability using machine learning techniques.
- Customer segmentation models can be developed to identify high-value or churn-prone customers.
- Real-time dashboards can be implemented using live database connections for operational monitoring.
- Integration with external data sources such as weather and traffic data can further improve delivery performance analysis.
- Advanced recommendation systems can be developed to suggest optimal discounts, cuisines, or delivery routes. Additionally, sentiment analysis on customer reviews can provide qualitative insights into service quality.

10. CONCLUSION

This project demonstrates a comprehensive data analytics workflow, starting from raw data cleaning to advanced visualization and insights generation. By combining Python, SQL, and Power BI, the analysis provides a holistic understanding of online food delivery operations. The insights generated can help businesses optimize logistics, improve customer satisfaction, and enhance profitability. Overall, the project shows practical data analysis skills aligned with real-world business requirements.