🏢 Swiggy End-to-End Data Analytics & Churn Prediction Project

(SQL + Power BI + Python + Machine Learning + Cloud Deployment)

📘 1. Executive Summary

This project is a complete **end-to-end data analytics pipeline** built on a multi-table Swiggy-style food delivery dataset.

The objective was to:

- Understand revenue behavior
- Analyze customer purchasing patterns
- Identify churn behavior
- Build a churn prediction model
- Deploy the model as a production-ready web application

The project integrates:

- **SQL** → Data extraction & business queries
- **Power BI** → Interactive dashboarding & KPI visualization
- **Python (EDA + Stats)** → Advanced analytics & feature engineering
- **Machine Learning** → Churn prediction modeling
- **Flask + Docker + Render** → Cloud deployment

📊 2. Business Problem Statement

Swiggy operates in a highly competitive food delivery market.
The company faces:

- Revenue volatility
- High customer churn
- Uneven city-level performance
- Heavy dependence on high-value customers

Core Business Questions

1. Is revenue stable or volatile?
2. Which cities generate sustainable revenue?
3. Do ratings influence sales?
4. Are customers retained or churning?
5. What percentage of customers generate majority revenue?
6. Can we predict churn before it happens?

🗂️ 3. Dataset Architecture

Tables Used (5 Tables)

1. Users
2. Orders
3. Restaurants
4. Menu
5. Food

Data Relationships

Users (1) ─── (Many) Orders
Restaurants (1) ─── (Many) Orders
Restaurants (1) ─── (Many) Menu
Menu (1) ─── (Many) Food

Total Records: ~1.18M rows (after merge)

🛠️ 4. SQL Analysis Layer

SQL was used for:

- Data joins
- Revenue aggregation
- City-level grouping
- Customer segmentation
- Frequency calculations

📌 Key SQL Business Insights

1 Monthly Revenue Trend

Revenue fluctuates significantly.

- Peak: January 2018 (~₹5 Cr)
- Lowest: June 2020
- Volatility (Std Dev): ~₹5.38M
- Coefficient of Variation: 18.5%

🔎 Insight:
Revenue is not stable. Marketing and competition likely impact performance.

---

2 Revenue Distribution

Skewness: 18.36
Kurtosis: 569.97

🔎 Insight:
Revenue is extremely heavy-tailed.
Few large orders contribute disproportionately.

Business Risk:
Loss of high-value customers significantly impacts total revenue.

---

3 City-Level Revenue Performance

Premium Cities:

- Sirsi
- Tirupati
- Srikakulam

Stable Cities:

- Kovilpatti
- Orai
- Osmanabad

🔎 Strategic Decision:

- Retention focus in Premium cities

- Volume expansion in stable cities

---

4 Customer Lifetime Value (CLV)

Mean CLV: ₹12,484
Median CLV: ₹1,921

🔎 Insight:
Huge gap between mean & median confirms 80/20 rule.

20% customers generate ~80% revenue.

---

5 Age vs Spending

Correlation ≈ 0

Age is not a strong predictor of revenue.

---

6 Rating vs Revenue

Correlation = -0.0019

Higher ratings do NOT guarantee higher revenue.

Customers prioritize:

- Convenience
- Price
- Delivery speed

📊 5. Power BI Dashboard Layer

Power BI was used to:

- Build interactive KPI dashboard
- Create dynamic slicers

- Visualize churn patterns
- Display revenue segmentation
- Create city heatmaps

📈 Dashboard Features

KPI Cards

- Total Revenue
- Total Orders
- Average Order Value
- Active vs Churned Users

Revenue Trend

- Monthly revenue growth
- Revenue volatility visualization

City Performance Map

- Revenue by geography
- Stability comparison

Customer Segmentation View

- RFM breakdown
- Loyal vs At-Risk users

🎯 Power BI Business Insights

1. Revenue spikes are seasonal.
2. Customer churn is extremely high.
3. Few loyal customers sustain revenue.
4. Geographic segmentation is critical.

🧪 6. Python Advanced Analytics Layer

Python was used for:

- Deep statistical analysis
- Hypothesis testing
- Feature engineering
- RFM modeling
- Correlation testing

📌 EDA Findings

Order Frequency

Mean Orders: 2.4
Median Orders: 2

Most customers order once or twice and leave.

Recency vs Revenue

Correlation: -0.0673

Inactive customers tend to spend less.

Active vs Churned Comparison

Active Users:

- Avg Orders: 2.42
- Avg Revenue: ₹16,414

Churned Users:

- Avg Orders: 1.83
- Avg Revenue: ₹11,936

Statistical Test:
P-value = 0.0

Conclusion:
Difference is statistically significant.

📊 7. RFM Segmentation

Customers segmented using:

- Recency
- Frequency
- Monetary

Segments:

Segment	%
Champions	~15%
Loyal	~40%
Potential Loyalists	~28%
At Risk	~17%

Strategic Meaning

55% strong customer base
28% growth opportunity
17% reactivation target

🚨 8. Churn Analysis

Churn defined as 90 days inactivity.

Churn Rate: 87.76%

Active Users: 9,457
Churned Users: 67,778

Business Reality:
Swiggy is operating with an extremely high churn rate.

🤖 9. Machine Learning Layer

Models Tested:

Model	ROC AUC
Logistic Regression	0.72
Random Forest	0.74
XGBoost	0.76

Production Optimized Random Forest:
ROC AUC: 0.91

Feature Importance

1. Recency (Strongest predictor)
2. Total Orders
3. Avg Order Value
4. Total Revenue

🌐 10. Deployment Architecture

Stack Used:

Flask → Backend
Gunicorn → Production Server
Docker → Containerization
GitHub → CI/CD
Render → Cloud Deployment

Live URL:
https://swiggy-end-to-end-data-analytics.onrender.com/

🎯 11. Business Recommendations

🔴 High Risk Customers

- 30% discount coupon
- Free delivery
- Push notification
- Loyalty bonus

🟠 Medium Risk

- Personalized recommendations
- 15% discount
- Limited-time offers

🟢 Low Risk

- Promote Swiggy One
- Referral campaigns
- Reward points

🏆 12. Final Strategic Conclusion

This project proves:

1. Swiggy’s main problem is retention, not acquisition.
2. Revenue depends heavily on a small customer segment.
3. Predictive intervention can reduce churn significantly.
4. RFM segmentation combined with ML improves marketing precision.
5. A deployed churn model enables real-time business action.

Reducing churn by even 10% could:

- Stabilize revenue
- Increase lifetime value
- Improve profitability
- Reduce acquisition dependency

💼 13. Technical Skills Demonstrated

- ✔ SQL (Advanced joins, aggregation, analytics queries)
- ✔ Power BI (DAX, dashboards, KPIs)
- ✔ Python (EDA, Stats, Feature Engineering)
- ✔ Machine Learning (Classification, ROC AUC, Model Comparison)
- ✔ Deployment (Flask, Docker, CI/CD, Cloud Hosting)
- ✔ Business Strategy & Insight Generation