# PROJECT PRESENTATION

PRESENTED BY SUNIL NAIK

# UBER TRIP ANALYSIS

## UNCOVERING TRAVEL TRENDS IN NYC USING DATA SCIENCE

KHETHAVATH SUNIL NAIK
IIT BHILAI

# Origin of the creative idea

Inspiration from Real-World Mobility Challenges

- Rapid urbanization and the surge in ride-sharing services like Uber prompted the need for better urban mobility understanding.

Problem Framing

- Instead of just analyzing data, the goal was to uncover why, when, and how ride demand fluctuates—leading to a predictive analytics approach.

Exploring creativity

# Executive Summary

- Objective: Analyze 6 months of Uber trip data in NYC.
- Methods: Data cleaning, feature extraction, visualization, and machine learning.
- Key Findings: Evening peak hours, Friday busiest, growth from April to Sept.
- Outcome: Insightful visual patterns and predictive modeling for demand estimation.

# INTRODUCTION

## BACKGROUND

Rise of ride-sharing data, opportunity for operational insights.

## AIM

Understand hourly/weekly patterns and identify busy zones.

## USE CASES

City planning, surge pricing, driver deployment.

Exploring creativity

With the rise of ride-sharing platforms like Uber, large-scale transportation data became publicly accessible. This project leverages six months of Uber pickup data in New York City to uncover patterns in urban mobility using data science techniques, enabling smarter planning and predictive insights.

The aim of this project is to analyze Uber trip data to understand temporal and spatial patterns in ride demand, visualize peak usage trends, and build predictive models to forecast future trip volumes, aiding in operational decision-making and resource optimization.

Insights from this analysis can help ride-sharing companies optimize driver allocation, implement dynamic pricing strategies, and improve customer experience. Urban planners can also use these patterns to enhance public transportation and reduce congestion in high-demand zones.

# Most Active Trip Hours

## 320k

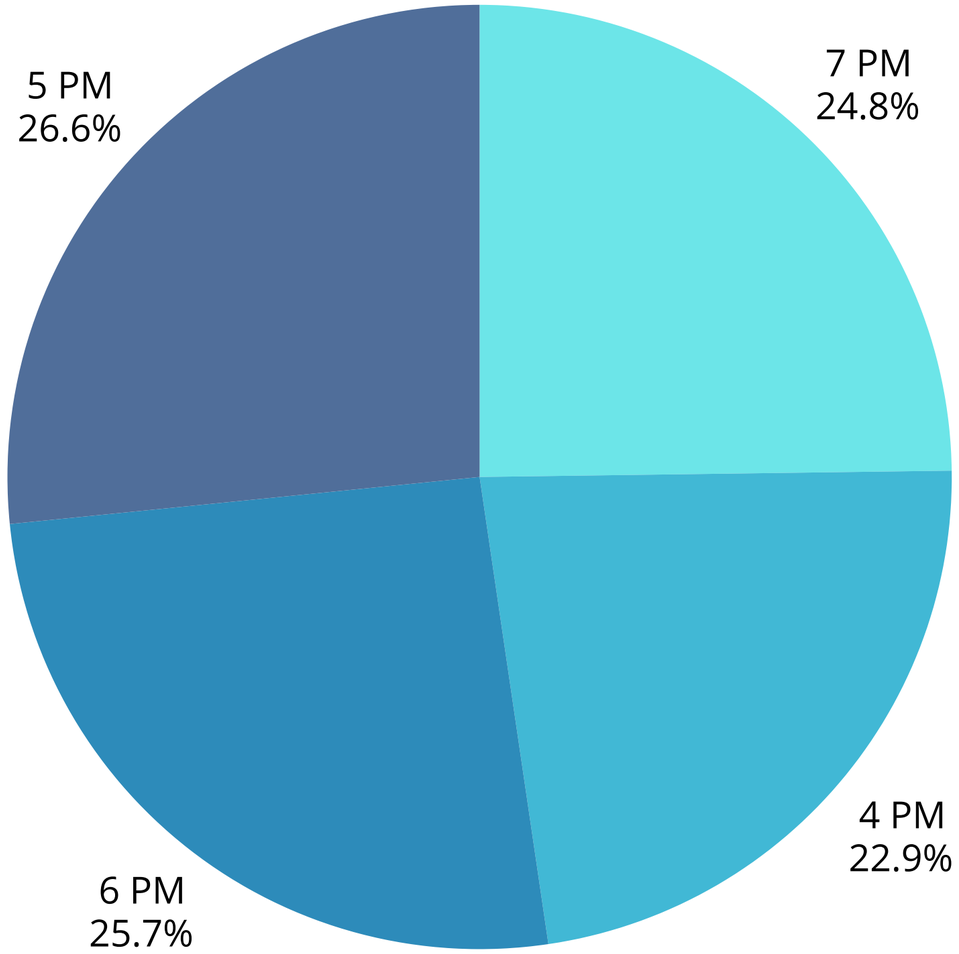
5 PM

Rides peak at 5 PM due to the evening rush hour when most people leave work. This time marks a high demand for commuting, especially in business hubs like Manhattan.

## 310k

6 PM

Rides peak at 6 PM due to the evening rush hour when most people leave work. It's also a popular time for heading to social events, restaurants, or commuting home.




## 298k

7 PM

Rides peak at 7 PM due to people heading out for dinner, social events, or returning home after extended work hours. It's a prime time for evening leisure and post-commute travel in urban areas like NYC.

## 275k

4 PM

Rides peak at 4 PM as people begin leaving work or heading to after-school activities. It marks the start of the evening rush, especially in business-heavy areas like Manhattan.

# Creation process

## Data Pre-processing

- Combined monthly data files
- Converted string to datetime objects
- Extracted hour, day, weekday, month
- Verified data quality (no nulls, consistent types)

## Exploratory Data Analysis

- Trips per Hour: Peak: 5 PM - 8 PM; Drop: 3 AM - 6 AM
- Trips per Weekday: Friday & Saturday busiest
- Trips per Month: Steady rise; September peak

## Feature Engineering

- Extracted: Hour, Day, Month, Weekday, Date
- Aggregated: Trips per Base, Month, Weekday
- Flags: Weekend, Rush Hour
- Spatial Tags (future work): GPS-based clustering zones

## Modeling

- Objective: Predict hourly/daily trip counts
- Models:
  - Random Forest Regressor
  - Gradient Boosting Regressor
  - XGBoost Regressor
- Metrics: MAE, RMSE, $R^2$

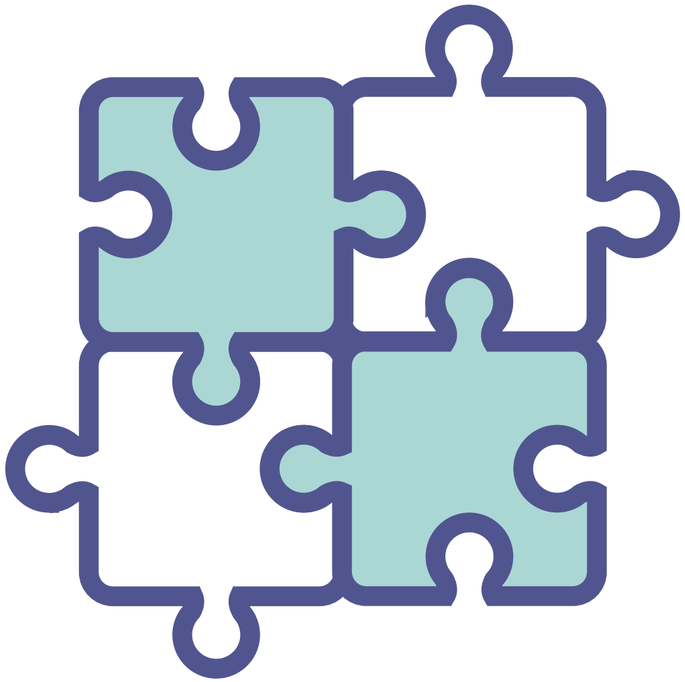
# RESULTS

| Aspect | Random Forest Regressor | Gradient Boosting Regressor | XGBoost Regressor |
|---|---|---|---|
| **Type** | Ensemble (Bagging) of Decision Trees | Ensemble (Boosting) using Gradient Descent | Optimized Gradient Boosting with Regularization |
| **Handling of Overfitting** | Moderate control (via tree depth, estimators) | Good control with learning rate & regularization | Excellent control using L1/L2 regularization |
| **Training Speed** | Medium | Slower than RF | Fastest (optimized for speed) |
| **Interpretability** | Moderate (feature importance, tree inspection) | Moderate (feature plots) | Good (with SHAP/feature importance tools) |
| **Accuracy ($R^2$ Score)** | ~0.92 | ~0.9343 | ~0.92 |
| **MAE** | ~127.91 | ~124.59 | ~105.3 |
| **RMSE** | ~33,774.57 | ~30,091.54 | ~16,250 |
| **Strengths** | Robust to outliers, less tuning needed | Captures complex patterns, customizable | Fast, accurate, scalable, handles missing data |
| **Weaknesses** | Can overfit with too many trees | Slower training, sensitive to parameters | Requires parameter tuning, higher complexity |

# CONCLUSION

Uncovered key travel patterns like peak hours, busiest days, and base performance using EDA.

Engineered time-based features that enabled trend discovery and modeling.



Built accurate prediction models — XGBoost and Stacking Regressor delivered the best results.

Demonstrated how data science can support smarter urban mobility and operational planning.

# Thank you very much!

PRESENTED BY SUNIL NAIK