Universiti Teknikal Malaysia Melaka
Jalan Hang Tuah Jaya
76100 Durian Tunggal
Melaka

11 SUSTAINABLE CITIES AND COMMUNITIES

13 CLIMATE ACTION

Melaka Go

Presenter

Syahmirul Afiq bin Gafar
Student

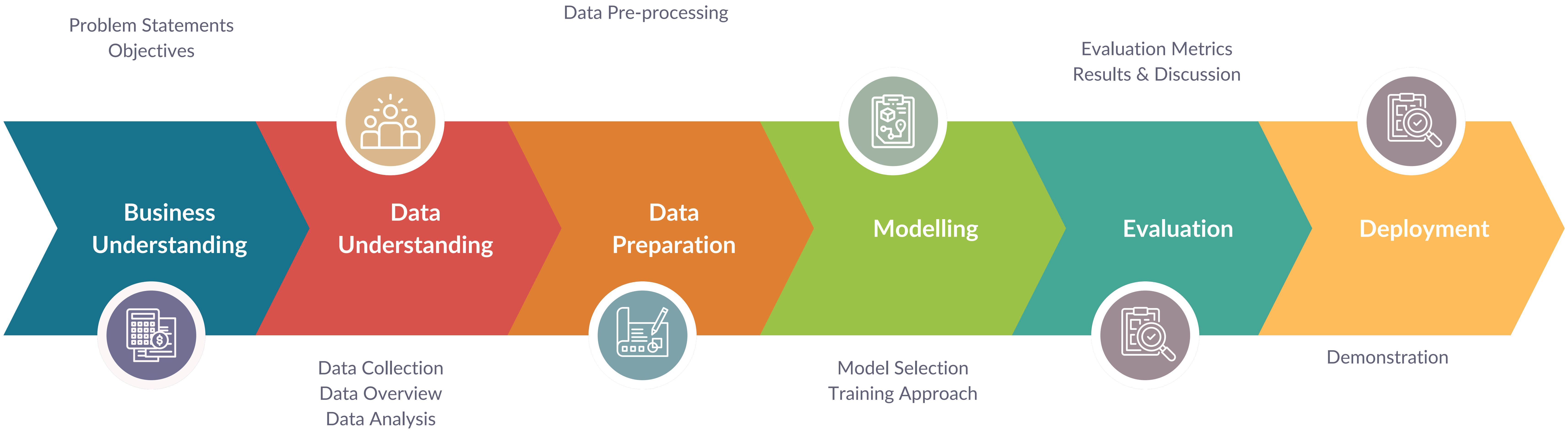
Muhammad Syahmi bin Mohd Alwi
Student

Muhammad Aqil Irsyad bin Mohd Iskandar
Student

Abby Iddin Harith bin Bahrin
Student

Agenda

We'll follow
CRISP-DM: The Cross Industry Standard Process for Data Mining



Problem Statement

01

Real-Time Systems Only

Designed to monitor and display current traffic congestion only, without forward-looking insights.

02

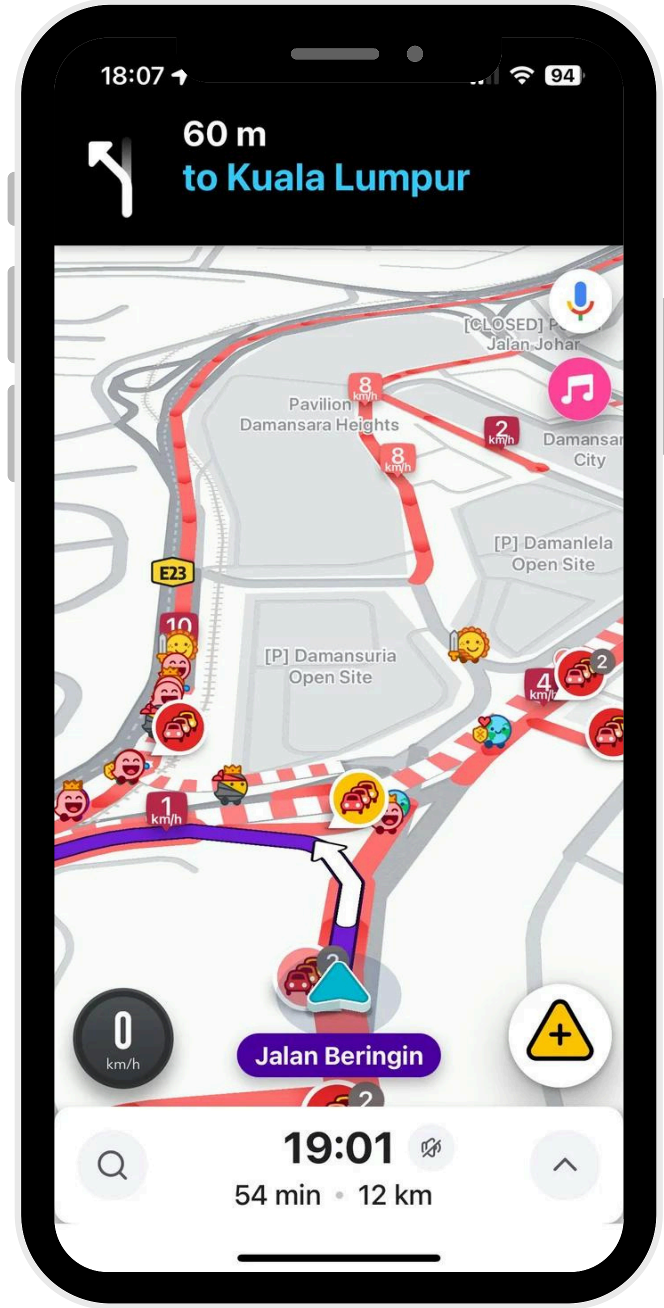
Difficulty Choosing Vehicles

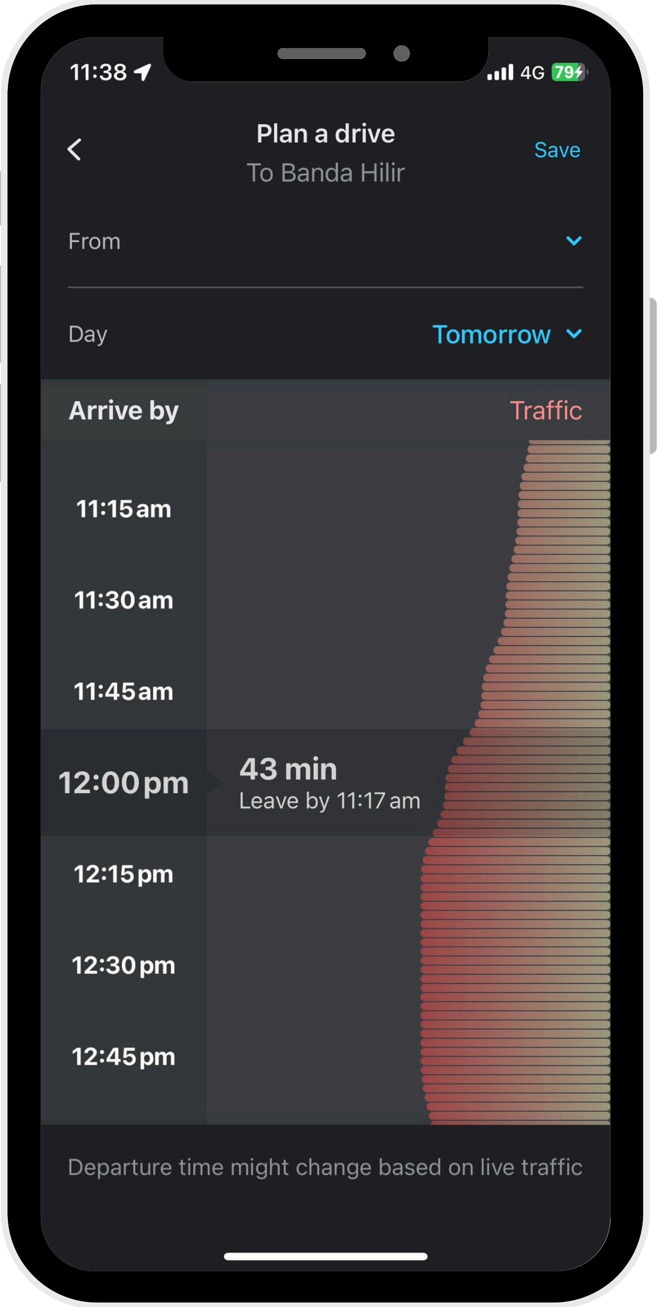
Commuters struggle to pick the right vehicle due to unpredictable conditions.

03

Limited Traffic Effecting Att.

Current systems focus only on traffic, ignoring weather data.





Objectives

01

Attributes Impact on Traffic

Study how different conditions affect traffic flow and congestion levels.

02

Vehicle Suitability Based on Conditions

Determine the suitable vehicle considering combined factors.

03

Peak and Non-Peak Hours

Analyze historical data to accurately identify actual peak and non-peak hours.

Stakeholders

MAJLIS BANDARAYA MELAKA BERSEJARAH

MET Malaysia

Role

serves as the primary data provider for this project,

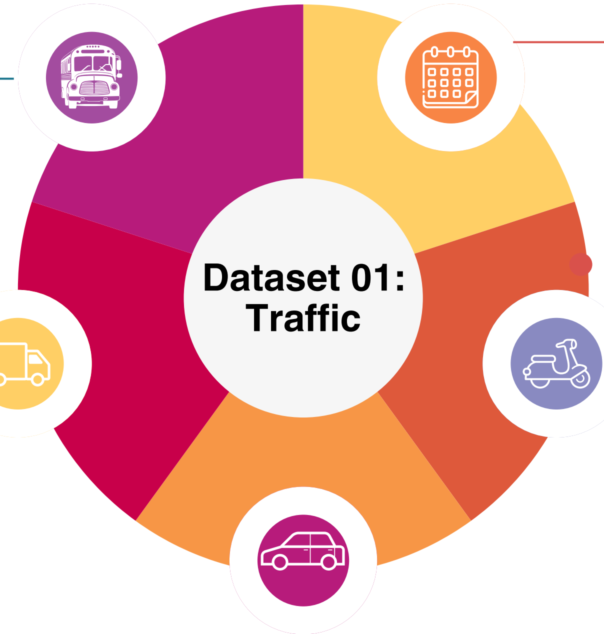
MAJLIS BANDARAYA MELAKA BERSEJARAH

PANORAMA Melaka

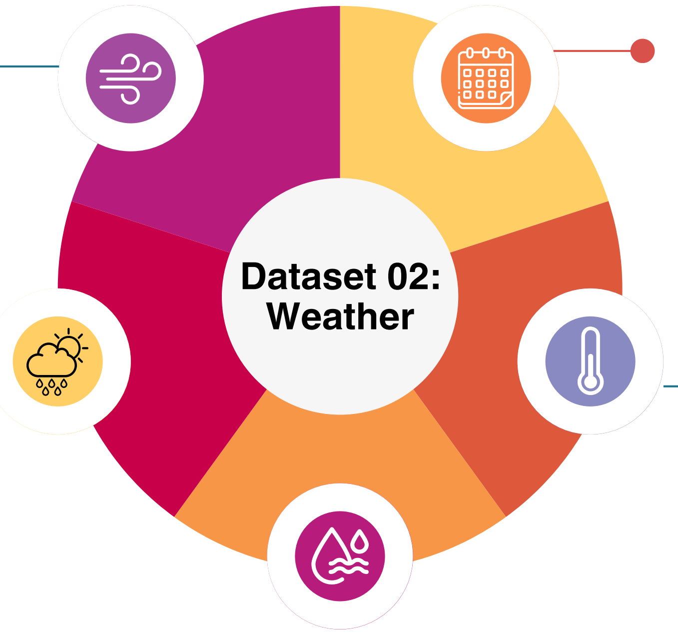
Benefits

Gains access to detailed analyses of peak and non-peak traffic hours derived from historical data.

4 Data Overview



1. **Date & Day** formatted as YYYY-MM-DD.
2. **Motorcycle**
3. **Car / MPV**
4. **Truck / Lorry**
5. **Bus**

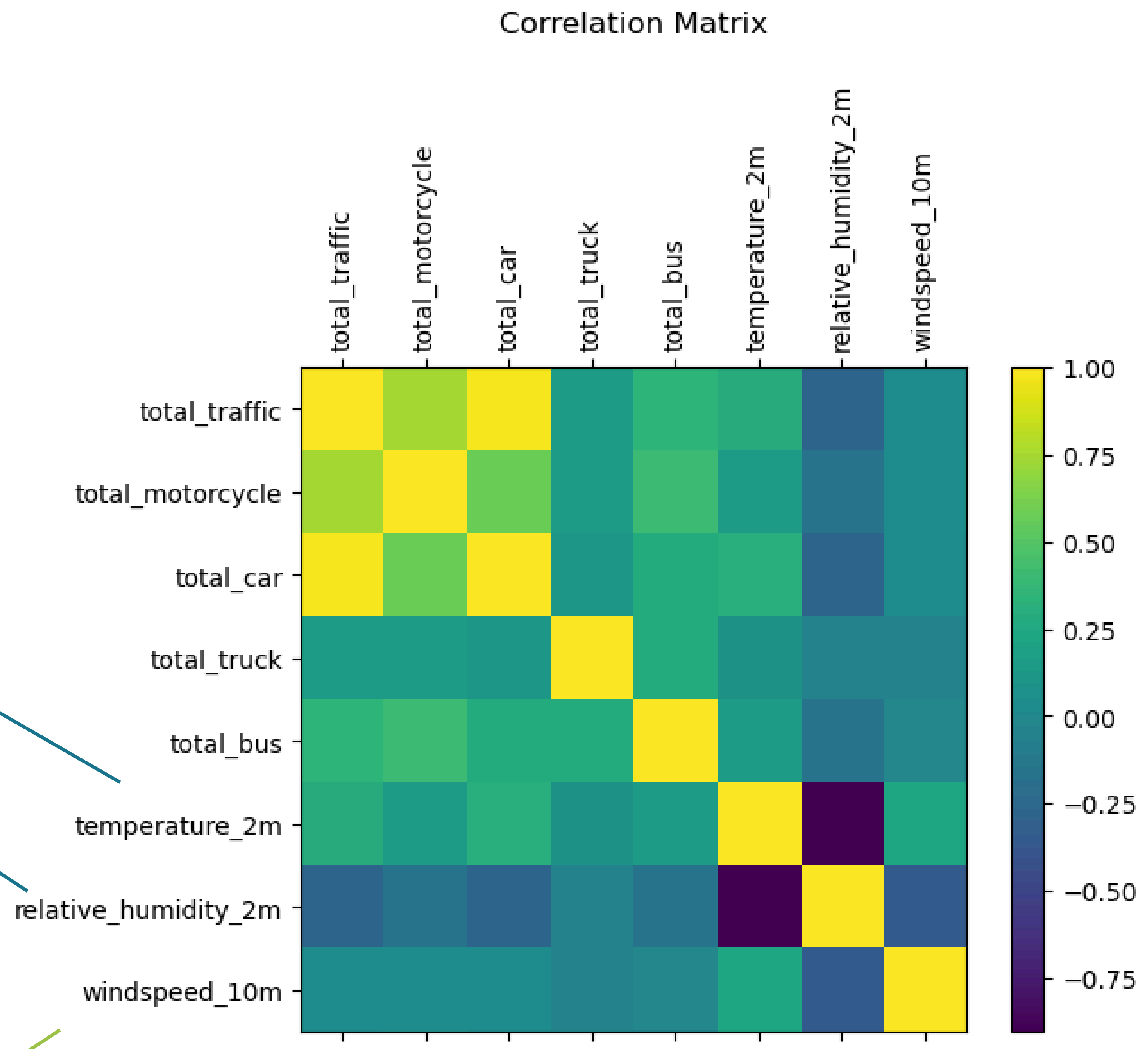


1. **datetime** formatted as YYYY-MM-DD HH:MM:SS.
2. **temperature** shows the air temperature in Celcius.
3. **relative_humidity_2m** indicates the moisture in the air.
4. **weathercode** description of the weather.
5. **windspeed_10m** measures speed of the wind.

01 Data Extraction

02 Data Cleaning

03 Exploratory Data Analysis



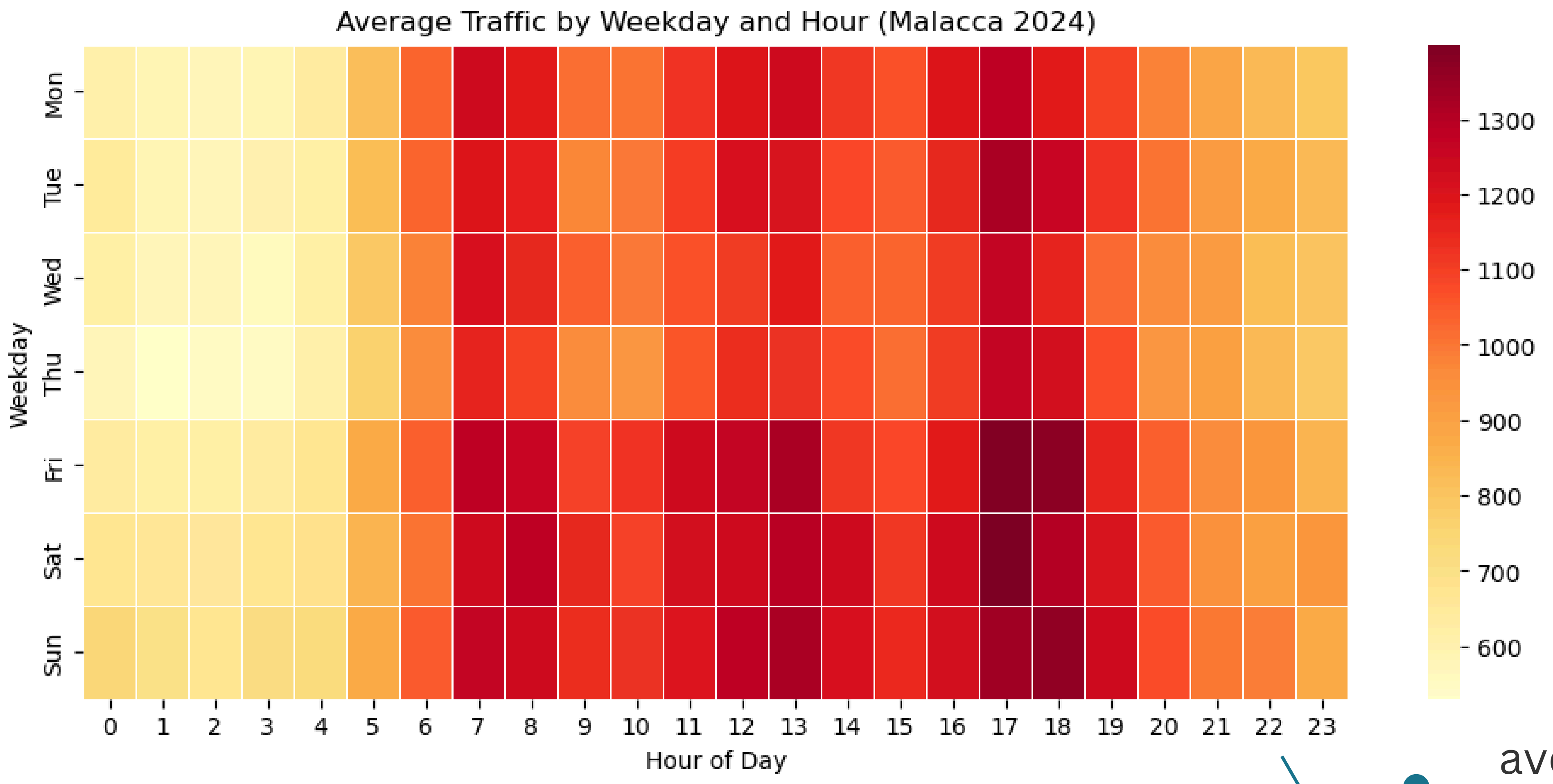
Higher temperatures are linked to **more overall vehicle activity.** Due to great condition for outdoor activities

Higher humidity often accompanies less favorable weather like rain or fog, which **tends to reduce traffic volumes.**

Higher winds make driving more challenging and less comfortable, and **reducing overall vehicle activity.**

Exploratory Data Analysis

Analyze the patterns, trends, and relationships in the data.



6

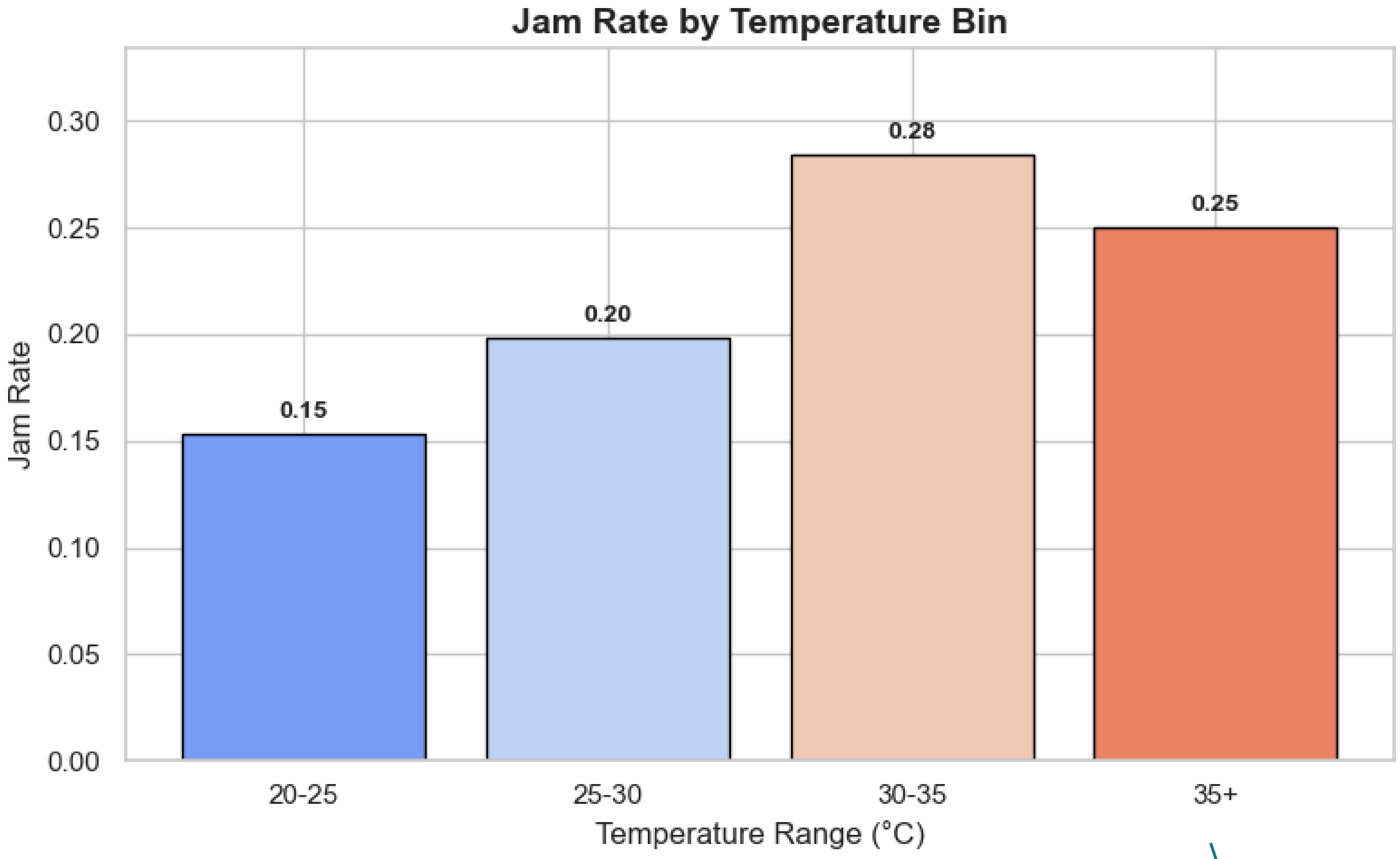
average traffic from **8 PM to 11 PM** on **weekends appears higher.**

morning **rush hours**, roughly from **7 AM to 9 AM.**

most intense traffic occurs during the evening rush hour, from around **5 PM to 7 PM**

Exploratory Data Analysis

Analyze the patterns, trends, and relationships in the data.



7

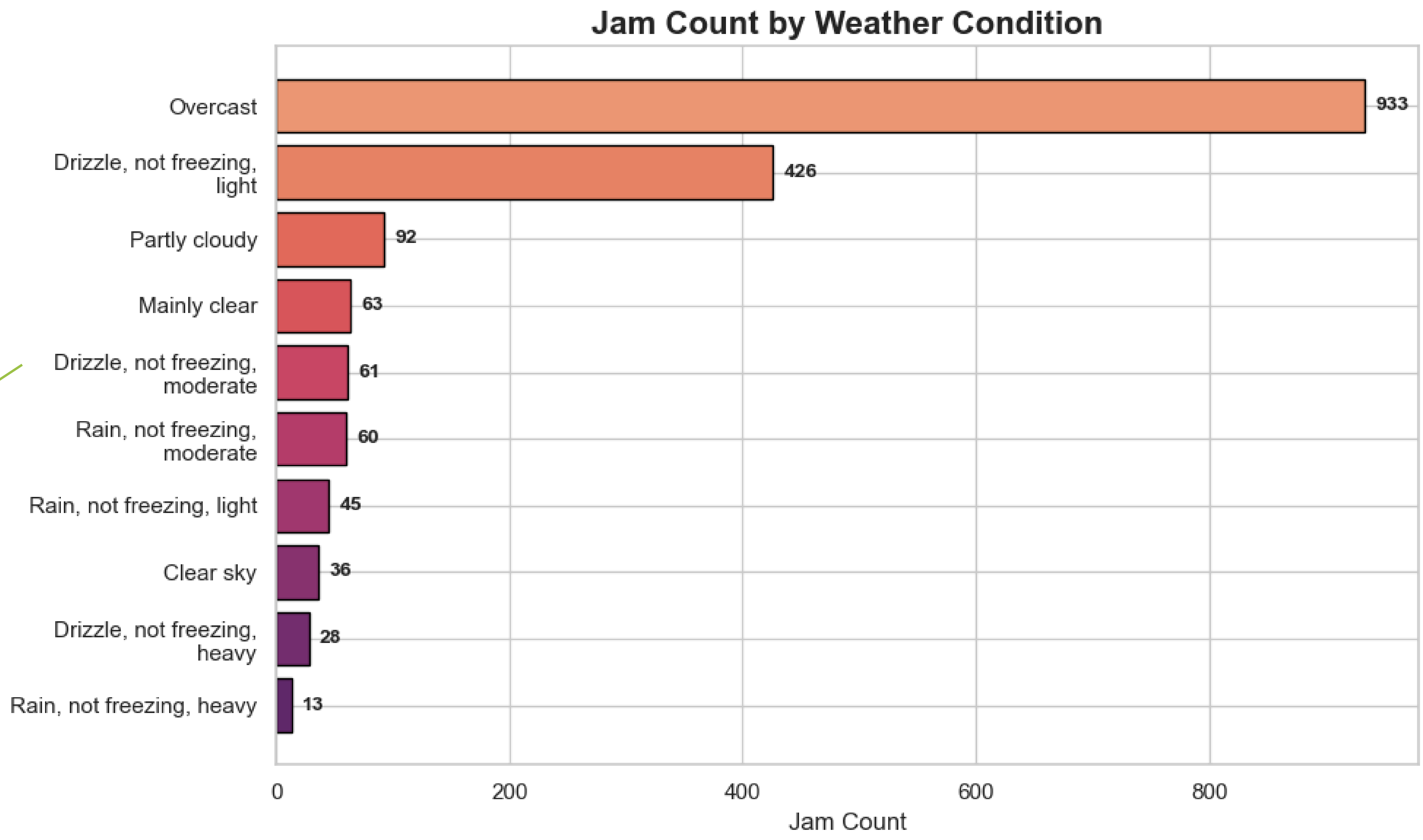
Higher jam rate due to outdoor activities in **pleasant weather**, leading to **more vehicles** on the road.

due to **extreme heat** made outdoor activities unpleasant, leading **lower vehicle volume**

Exploratory Data Analysis

Analyze the patterns, trends, and relationships in the data.

8



Drizzle / rain decrease travel, thereby reducing **traffic volume** that could potentially get jammed (by accident).

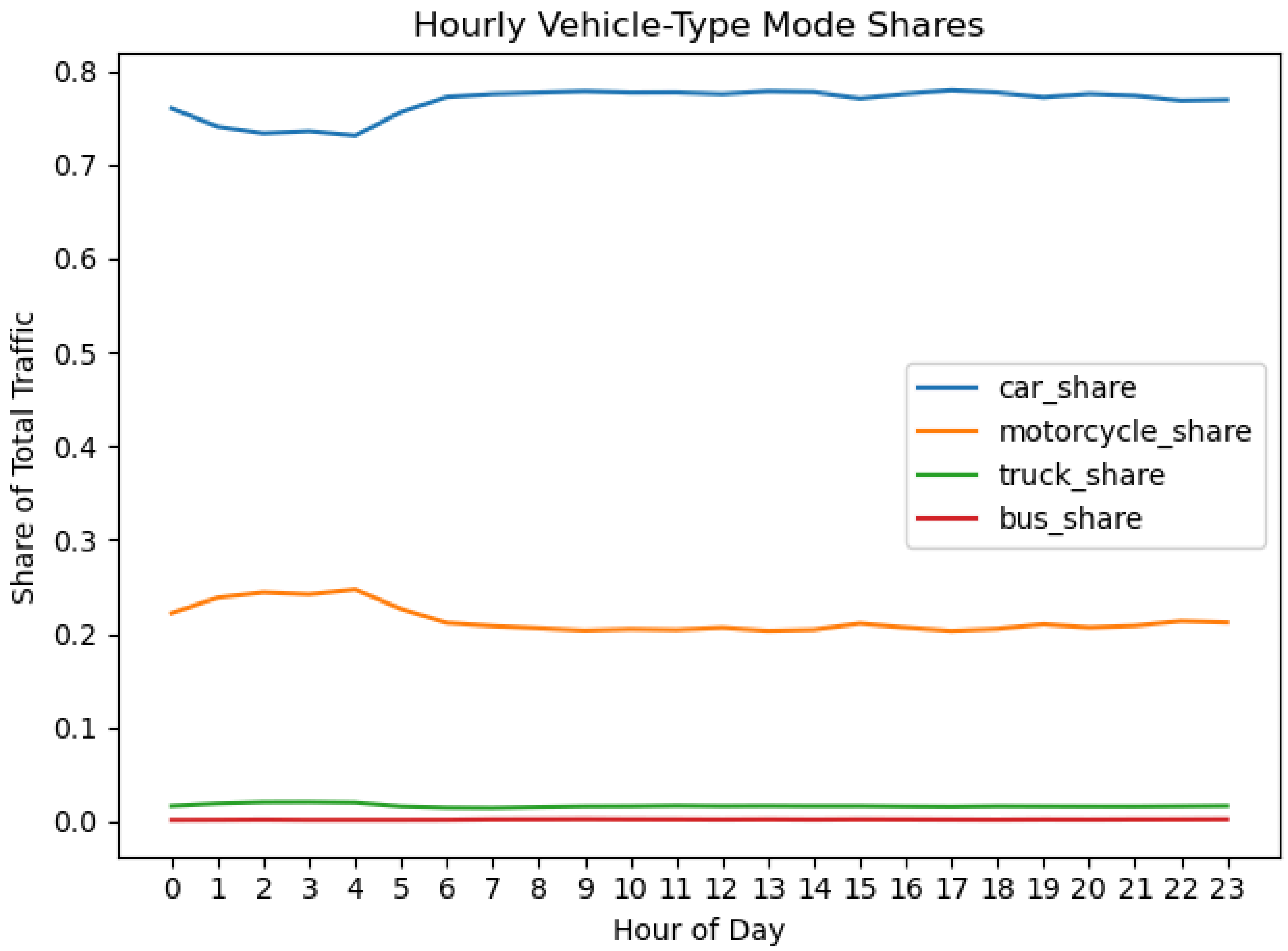
Exploratory Data Analysis

Analyze the patterns, trends, and relationships in the data.

The **car share** consistently forms the **largest proportion** of total traffic across all hours

The **motorcycle share** is the **second largest**, generally ranging from **20% to 25%**

Truck and bus shares are very **low**, both staying below **5%**



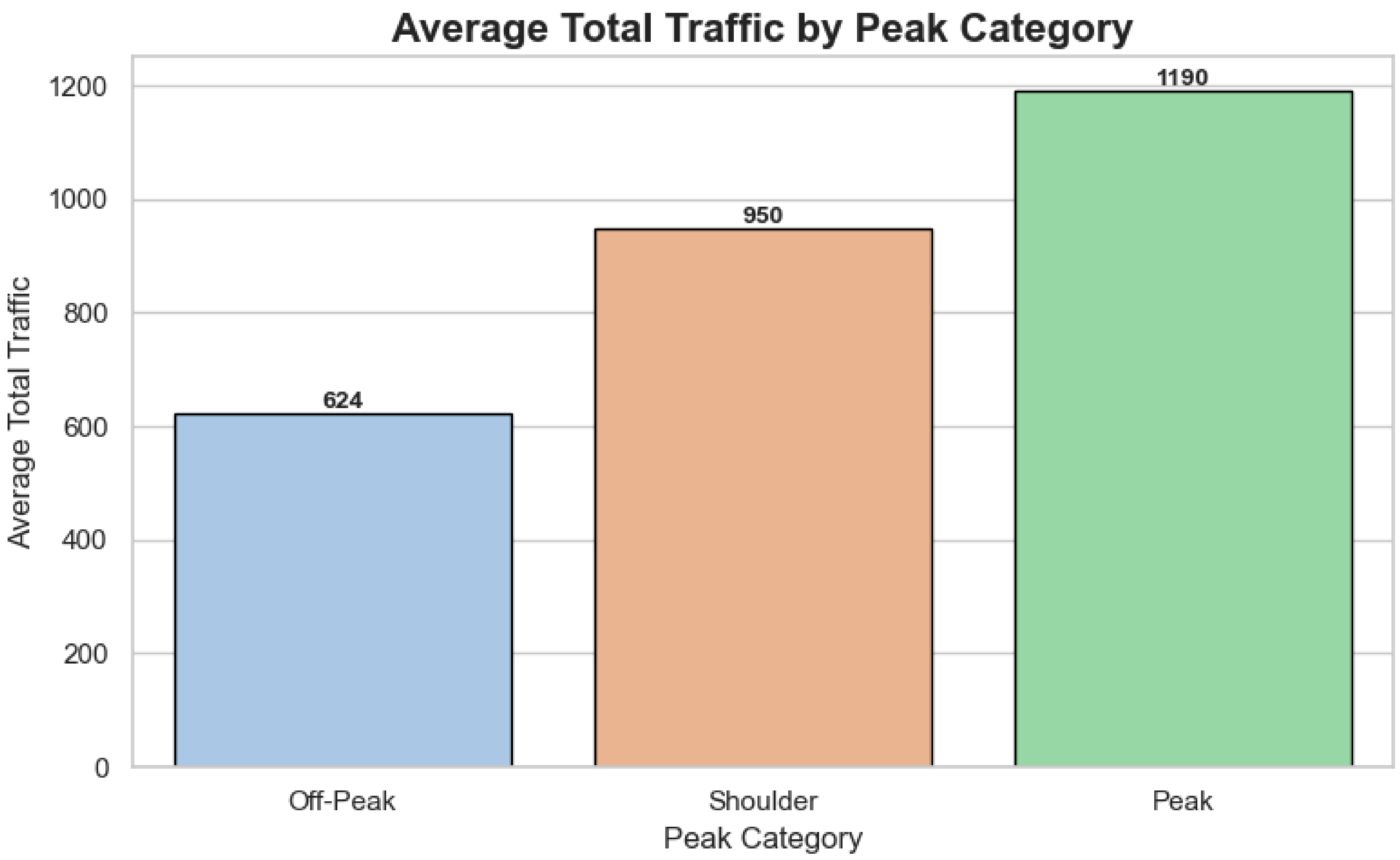
Exploratory Data Analysis

Analyze the patterns, trends, and relationships in the data.

Peak hours are the busiest, with nearly **double** the traffic volume compared to off-peak hours.

Shoulder hours act as **transitional** periods

Off-peak hours have the **lightest** traffic



Exploratory Data Analysis

Analyze the patterns, trends, and relationships in the data.

Modelling

Two-Model Approach

The Congestion System

Goal: Predict is_jam (Yes/No)

Purpose: Provide clear, high-stakes risk warnings. Answers "Will I get stuck?"

The Traffic Indicator

Goal: Predict peak_category (Peak / Shoulder / Off-Peak)

Purpose: Provide nuanced context on traffic levels. Answers "How busy will it be?"

Feature Selection

From the preprocessed data,

We have decided to pick these features as we think it would benefit the model training greatly.

Features not included in the model training:

- Total Traffic
- Total Cars
- Total Motorcycle

Reason: Overfitting, model will be dependent to said features. More towards logic rather than predictive

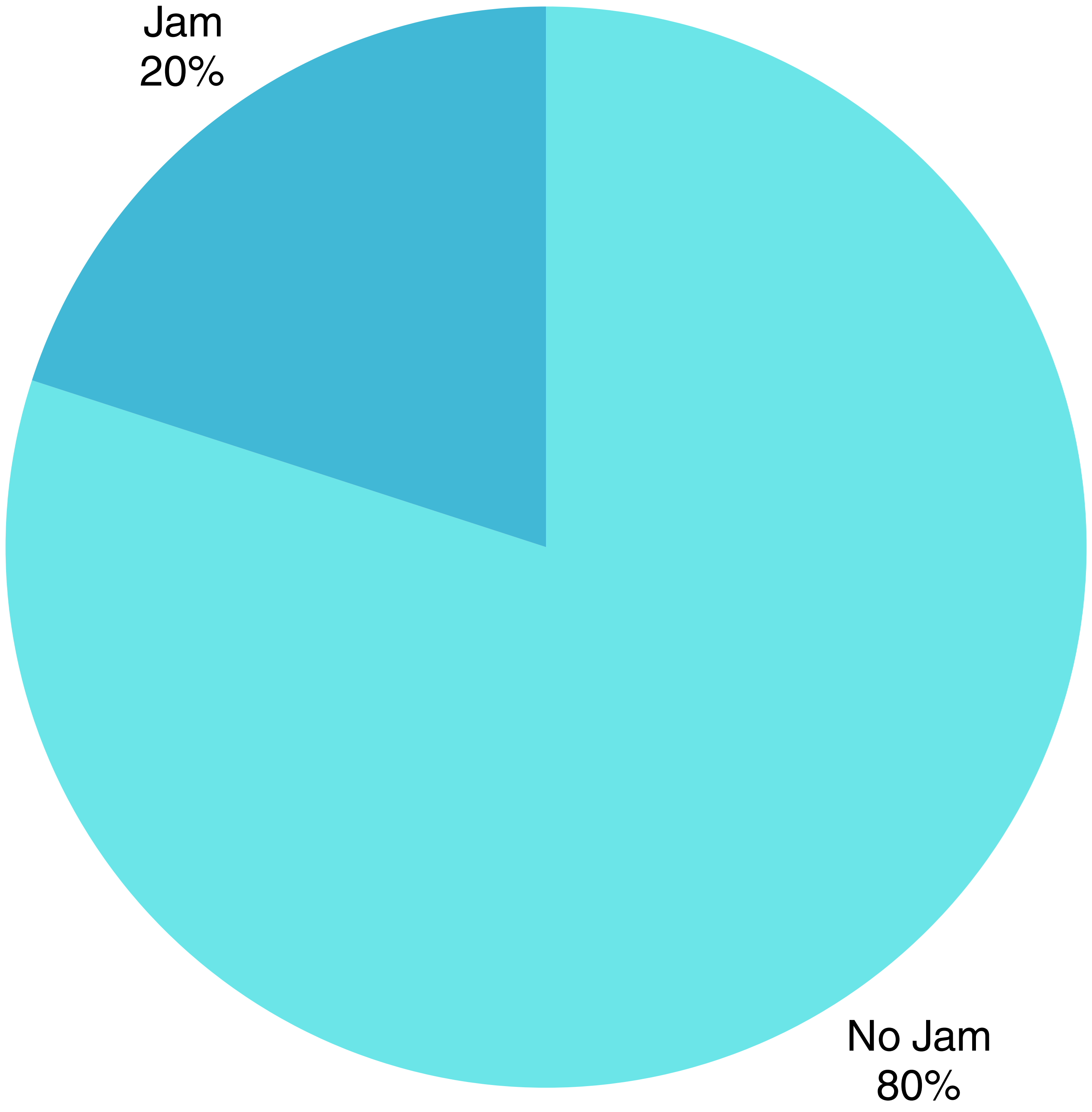
```
features_to_use = [
    'temperature_2m',
    'relative_humidity_2m',
    'weathercode',
    'windspeed_10m',
    'is_weekend',
    'hour',
    'is_holiday_mlk',
    'day_of_week',
    'month'
]
```

Challenges

Our data is imbalanced: Far more "No Jam" hours than "Jam" hours.

The Problem:
A basic model could achieve 80% accuracy by always predicting "No Jam," making it useless.

The Solution:
We used a technique called SMOTE (Synthetic Minority Over-sampling Technique).



What is SMOTE

Problem:

Imbalanced datasets cause machine learning models to perform poorly on minority classes (e.g., jam).

Solution (SMOTE):

- Generates new synthetic samples for the minority class.
- Improves model performance by balancing the class distribution.

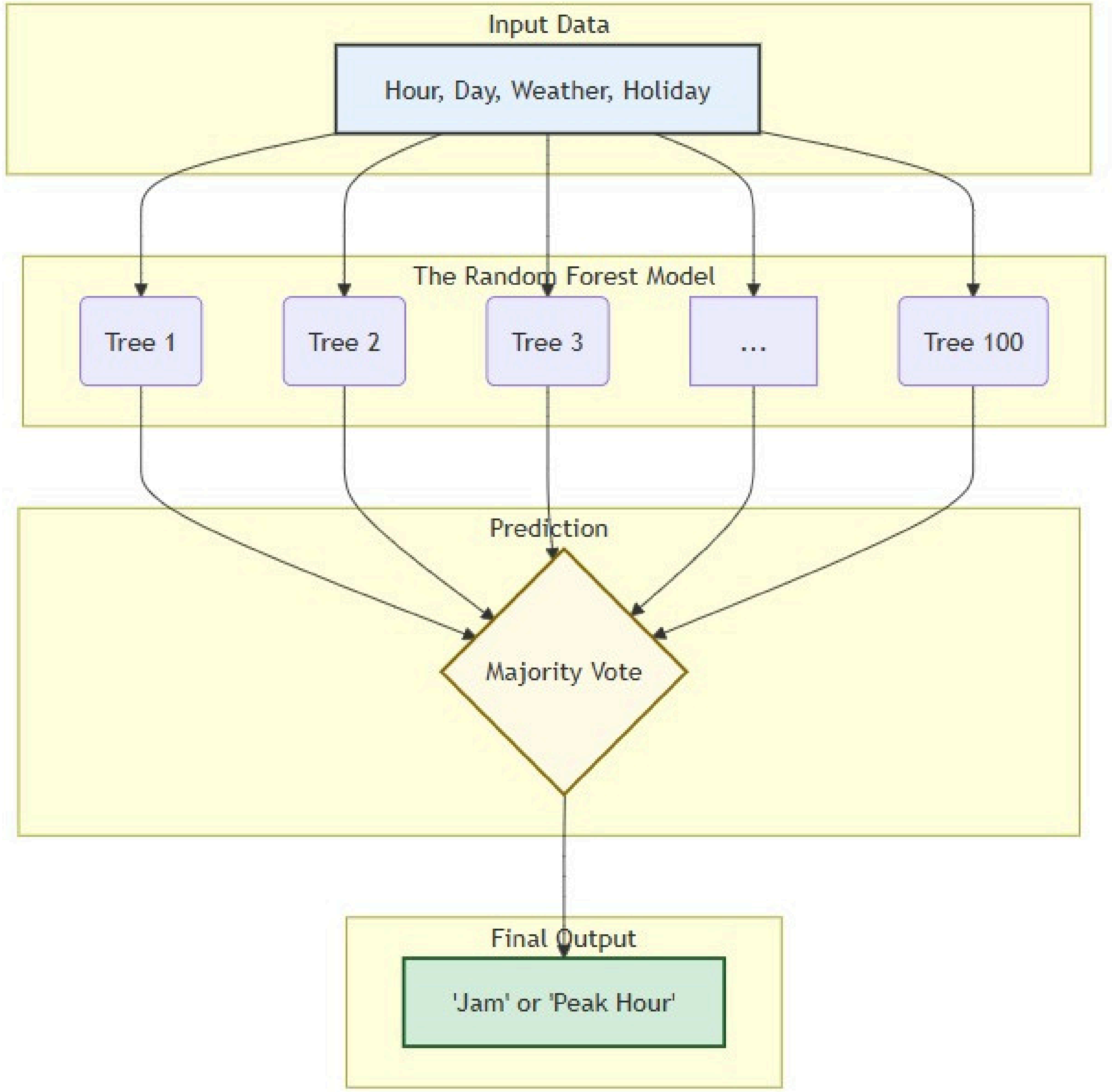
How SMOTE Works:

- Select a sample from the minority class.
- Find its k-nearest neighbors.
- Randomly pick a neighbor.
- Interpolate between the two to create a new sample.

Benefits:

- Reduces bias toward majority class.
- Improves recall and F1-score.
- Avoids overfitting (better than random oversampling).

How it works



Random Forest is Used

Features like Hour, Day, Weather, Holiday are used as inputs.

The input data is passed to multiple decision trees (Tree 1 to Tree 100).

Each tree makes an independent prediction (e.g., "jam" or "no jam", "off-peak", "shoulder", or "peak").

All predictions are collected and combined using a majority voting system.

The most common prediction among the trees becomes the final result, either "Jam" or "Not Jam" for model 1 and "Off-Peak", "Shoulder", or "Peak" for model 2.

Evaluation

Confusion Matrix (metrics)

Suits best for classification model.

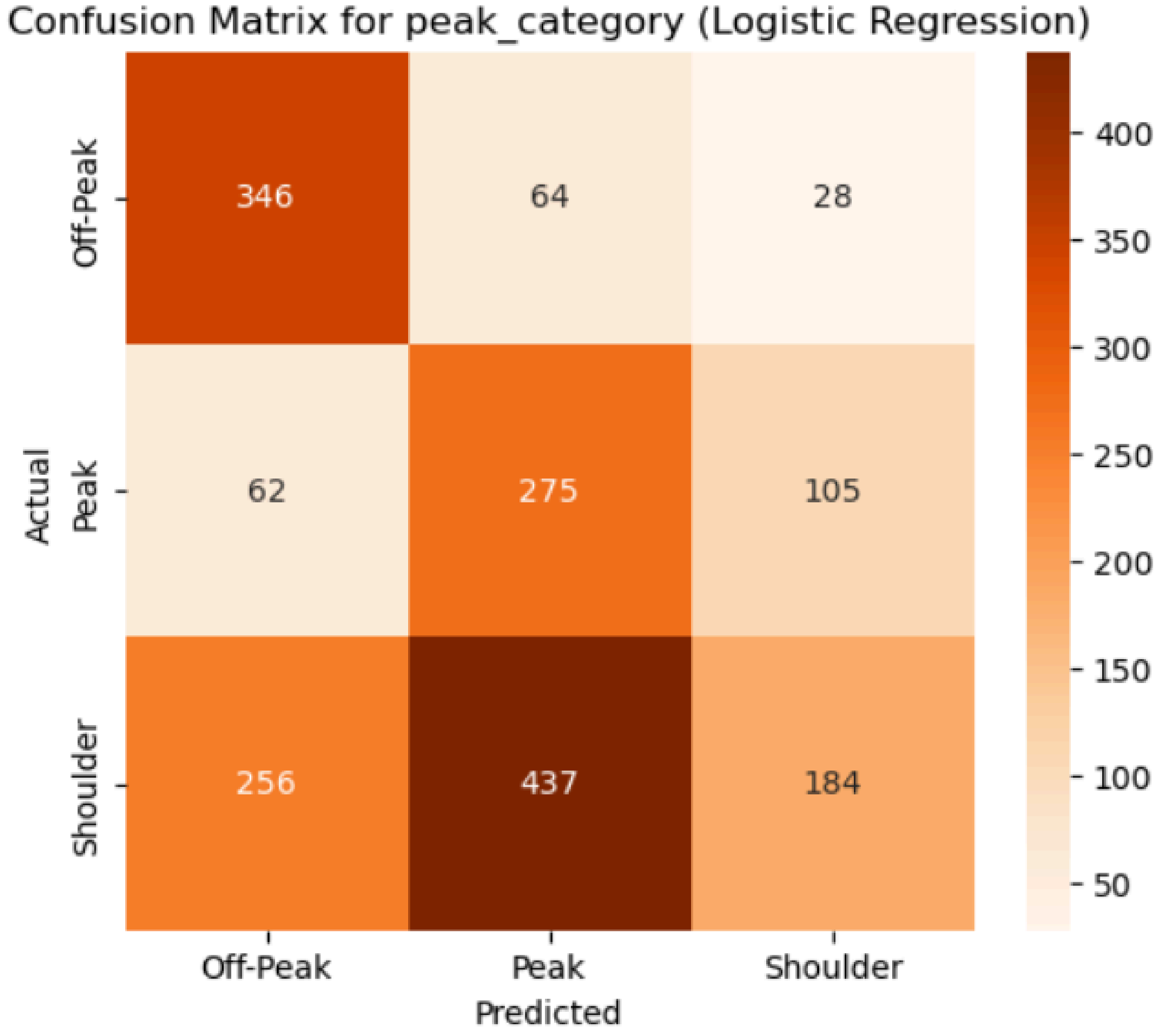
Purpose

Summarizes number of correct and incorrect predictions made, comparing against the actual outcomes.

Key Components (binary)

- True Positive (TP): Correctly predicted positive.
- True Negative (TN): Correctly predicted negative.
- False Positive (FP): Incorrectly predicted positive (Type I error).
- False Negative (FN): Incorrectly predicted negative (Type II error).

From a Simple Baseline to a Powerful Model

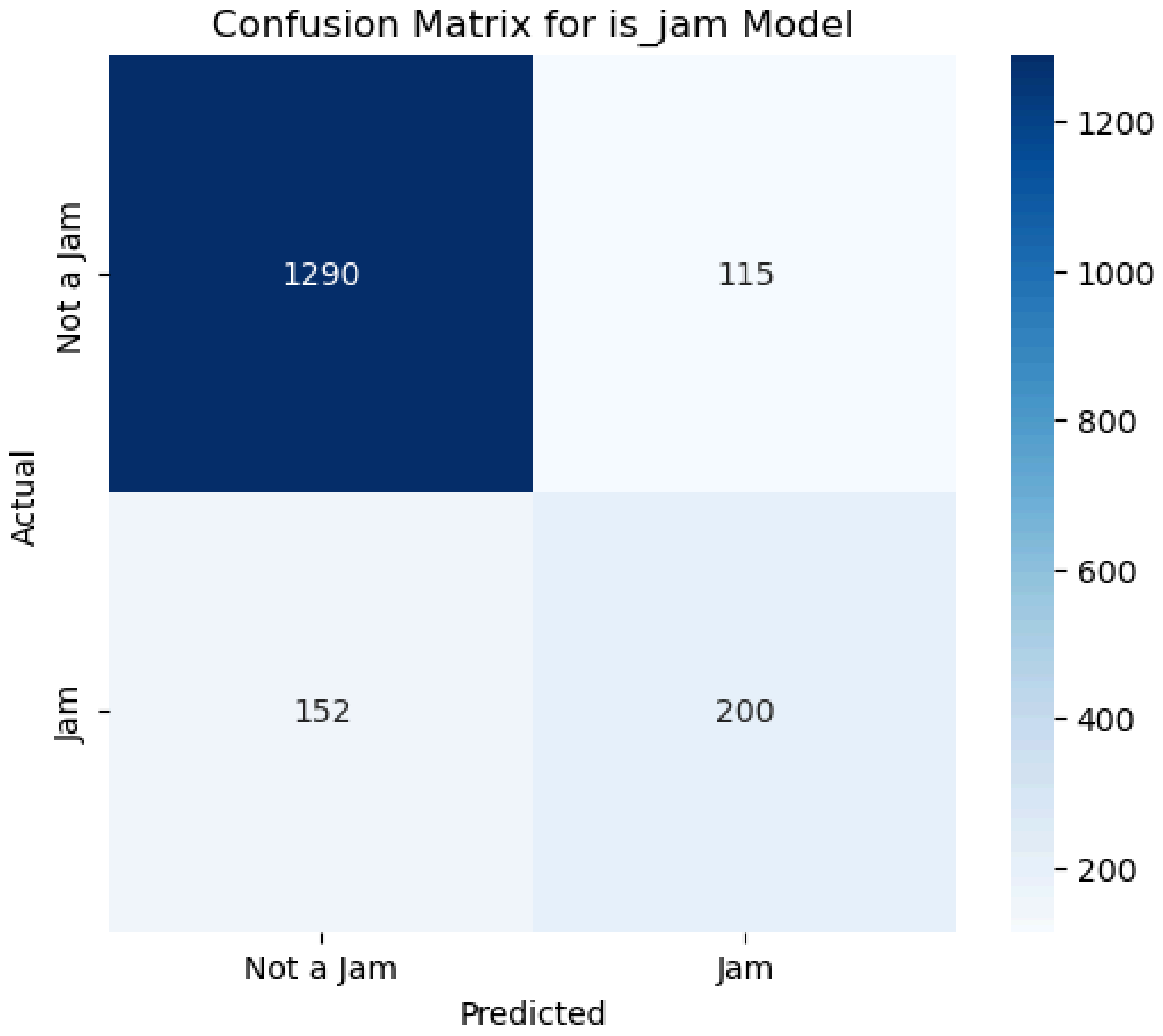


Model 2 (peak_category) trained using Logistic Regression Results:

- Accuracy: 45.80%

Key Insight: A linear model is insufficient. It fails to learn the non-linear relationships in the data.

Final Model Evaluation



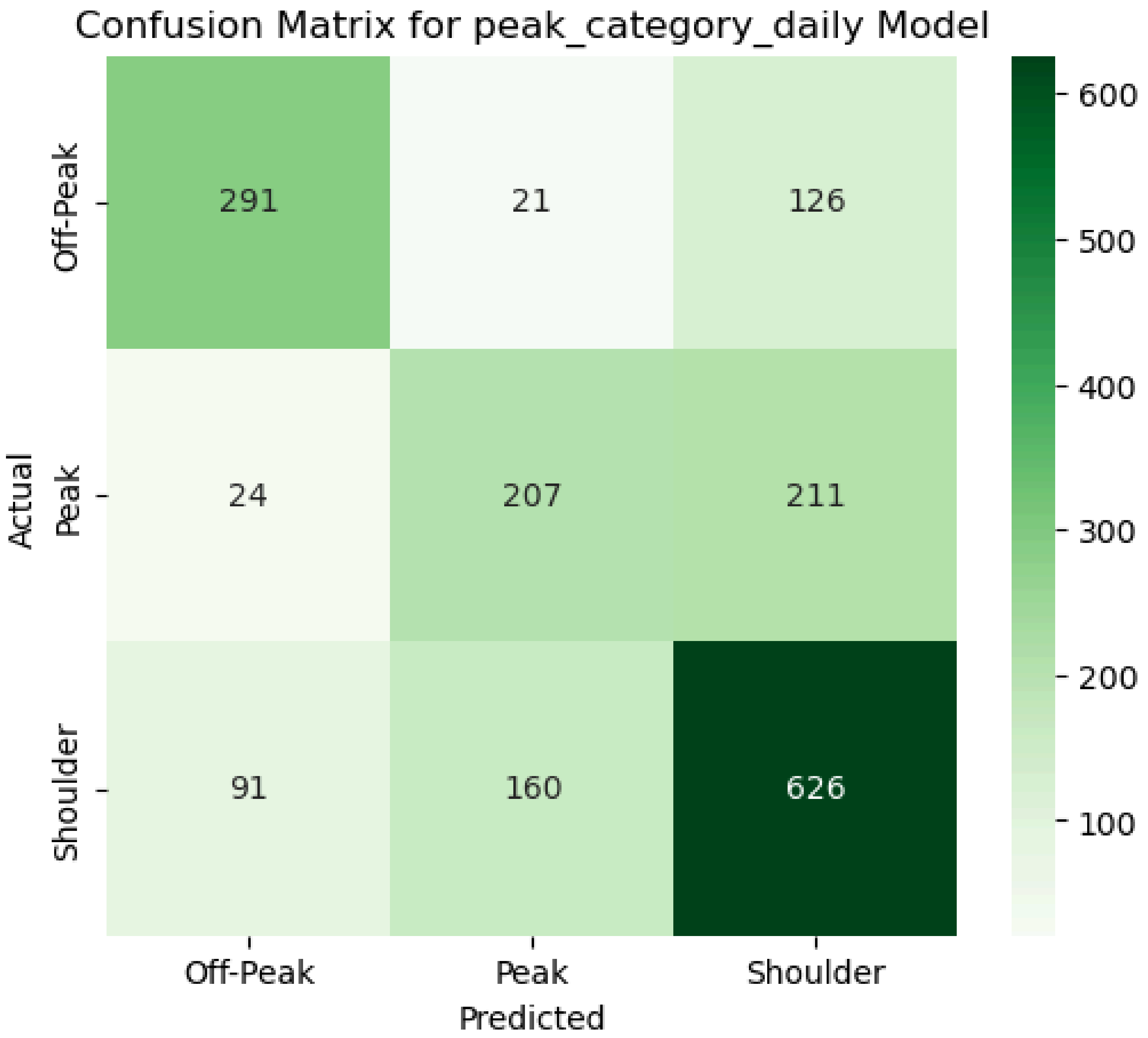
Model 1 (is_jam) Results:

- Accuracy: 85%
- Key Metric (Recall): "Successfully identifies 57% of all actual traffic jams."

Performance & Results

Model 2 (peak_category) trained on Random Forest Results:

- Accuracy: 64%
- Key Insight: "Effectively distinguishes between different traffic levels, especially Off-Peak and Shoulder hours."

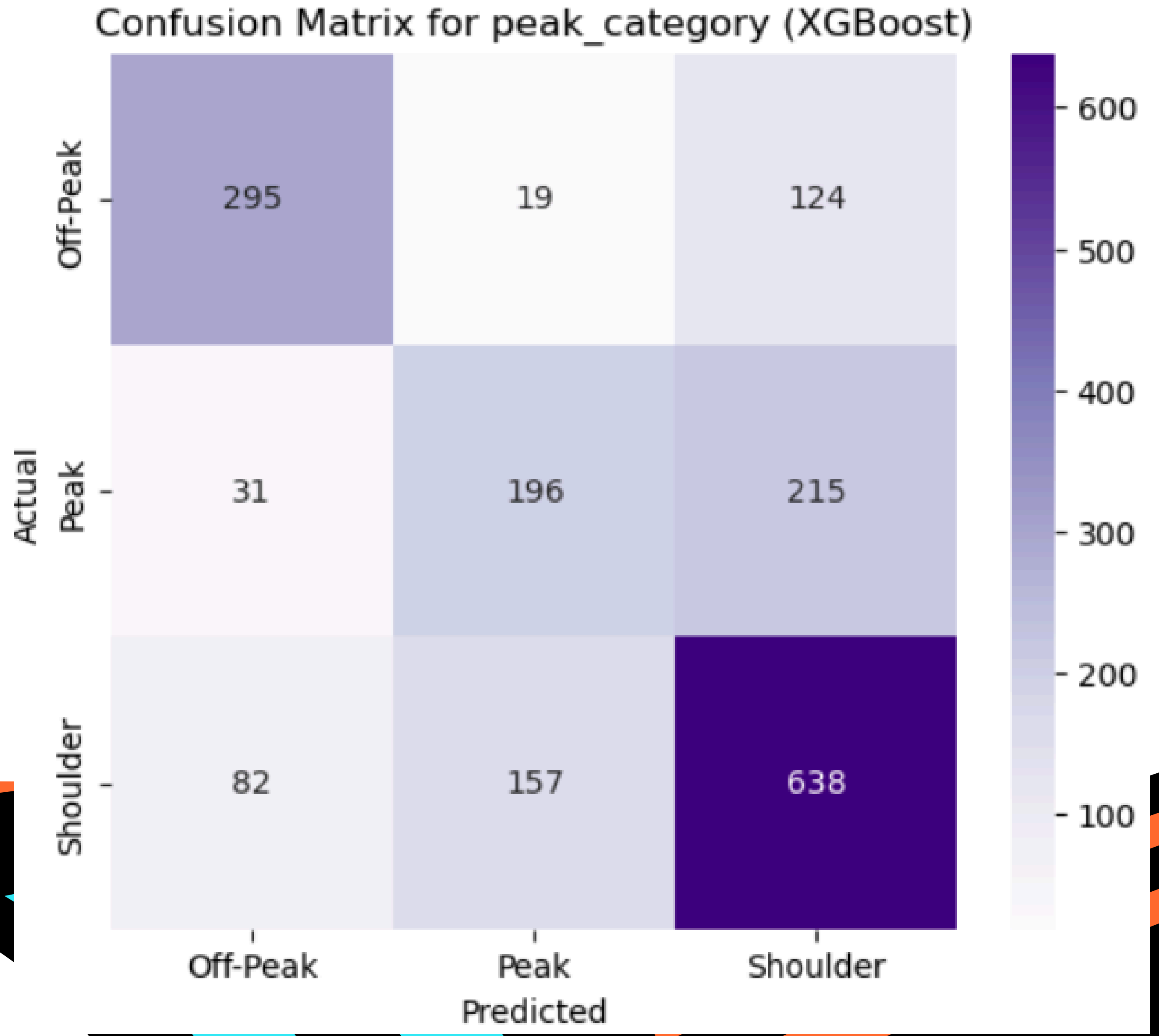


Alternative Model

Model 2 (peak_category) trained using XGBoost Results:

- **Accuracy: 64.3%**

Key Insight: Similar high performance, but Random Forest offers better interpretability for our needs.



Comparison

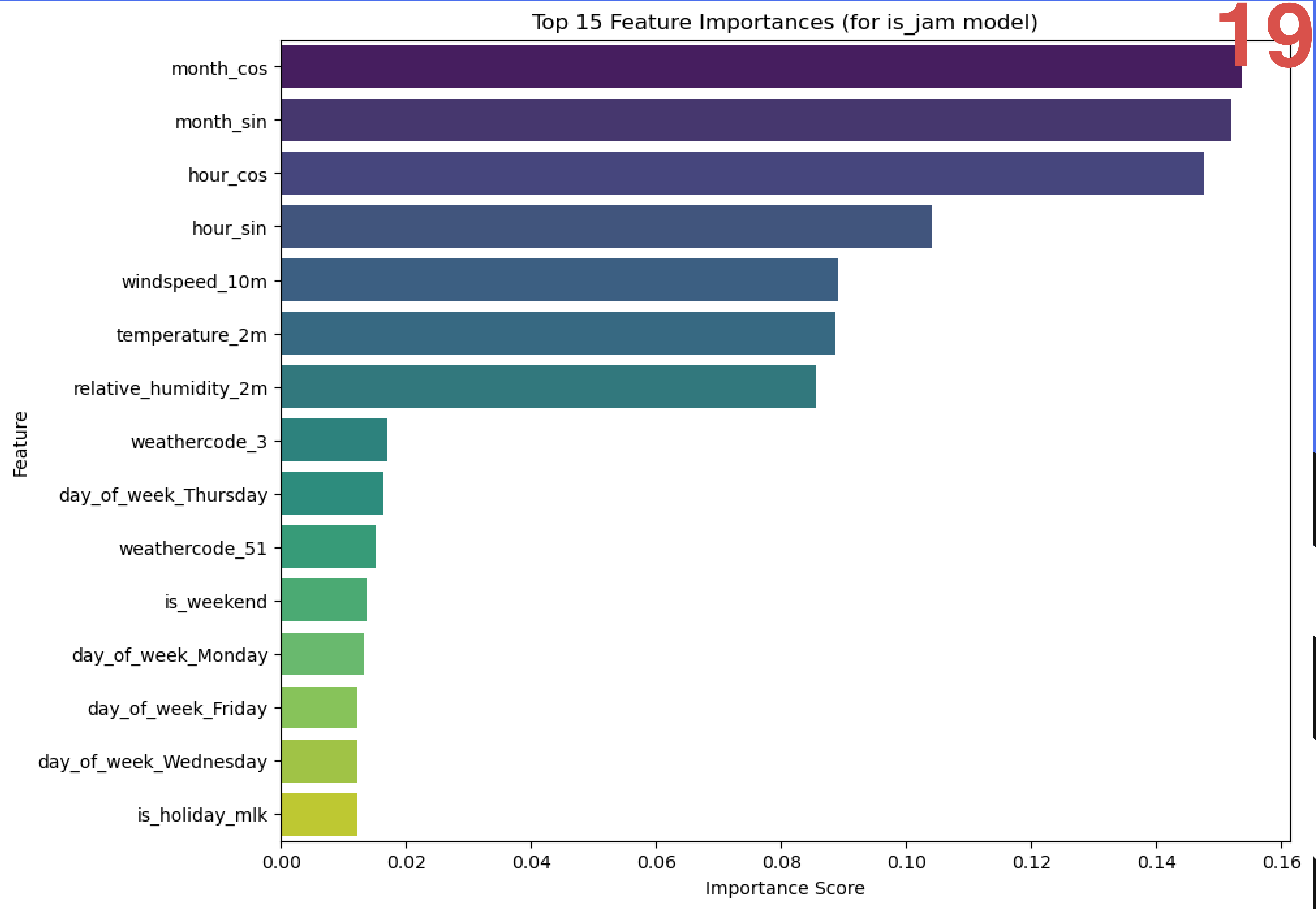
Model	Algorithm Type	Overall Accuracy	Key Strengths	Key Weaknesses
Logistic Regression	Linear Model	45.8%	- Simple and fast - Good for establishing a baseline	- Very poor performance - Fails to capture non-linear patterns - Extremely low recall (21%) for the 'Shoulder' class
Random Forest	Ensemble (Bagging)	64.0%	- Strong, balanced performance across all classes - Good interpretability (clear Feature Importance) - Robust and less prone to overfitting	- Slightly lower overall accuracy than XGBoost
XGBoost	Ensemble (Boosting)	64.3%	- Highest overall accuracy - Excels at predicting 'Shoulder' and 'Off-Peak' categories	- Slightly lower F1-score for the 'Peak' category compared to Random Forest

Why we choose Random Forest?

Why Random Forest is a "Balanced" Model:

- It builds hundreds of decision trees on random subsets of the data.
- It makes a final prediction by averaging the "votes" from all the trees.
- The Result: This averaging process makes the model very robust and stable. It's less likely to be thrown off by unusual data points and performs well without extensive fine-tuning.

Feature Importance bar chart



Demonstration

Thank you!

تريما كاسيه

谢谢

நன்றி

Question & Answer Session