DATA SCIENCE PROJECT

Take Home Test

PRESENTED BY:
Rahma Anggana Rarastyasa

PRESENTED TO:
Nanda Rizkika Ruanawijaya

Introduction

Hi! I am Rara. A highly motivated Industrial Engineering graduate from Sebelas Maret University with 1 year of professional experience in port services and healthcare services. **Strongly interested in Data Science and Data Analytics**, with foundational skills in data processing, statistical analysis, and data visualization to support effective decision-making. Proficient in **Python, SQL**, and **Excel**, with experience using tools such as Pandas, Numpy, Matplotlib, and Looker Studio. Known for being **analytical, detail-oriented, and a fast learner**, with the ability to work both independently and in teams. Eager to contribute to data-driven projects and continuous improvement initiatives.

Working Experience

- PT Revolusi Kesehatan Indonesia

  Operation Team

  Nov, 2024 - Present

- PT Krakatau Bandar Samudera

  Port Area and Warehouse Intern

  Jan, 2023 - Feb, 2023

Educational Background

- Dibimbing.id

  Data Analyst & Data Science Bootcamp (Non-Degree)

  May, 2025 - Present

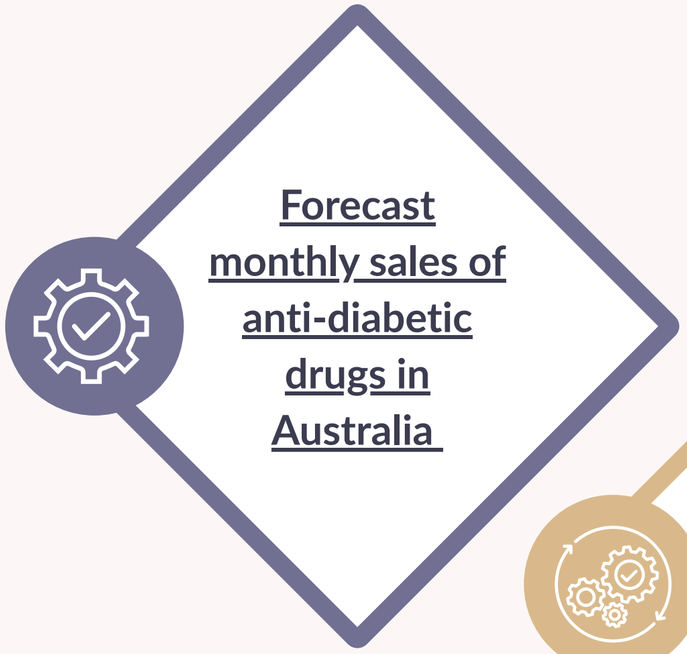
- Universitas Sebelas Maret

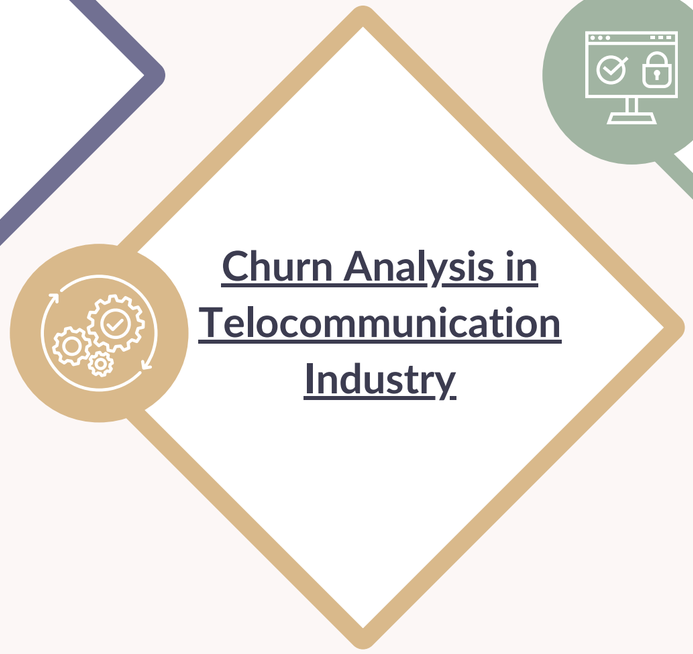
  Industrial Engineering (Bachelor of Engineering)

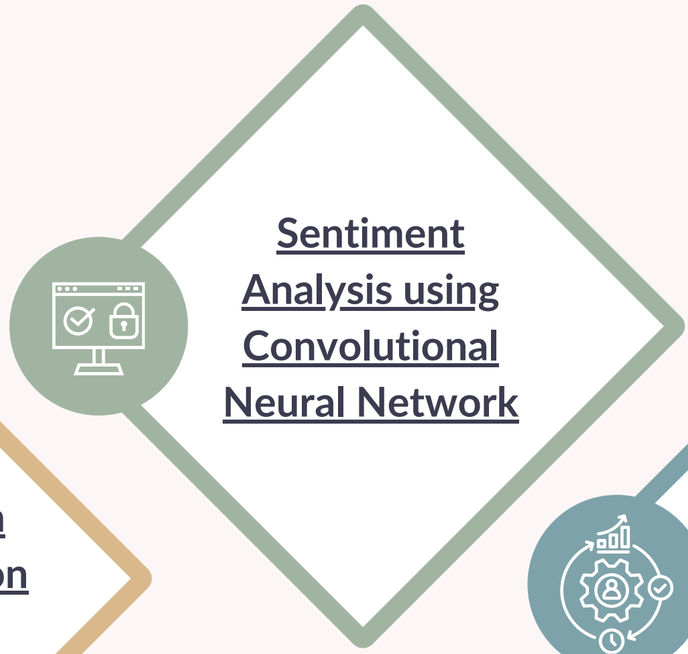
  Aug, 2020 - July, 2024

PREVIOUS PROJECT

there are some projects that have been build to make a data analytics portofolio







MAIN PROJECT

Optimizing Delivery Zones Using Geospatial Analytics

MAIN PROJECT CONTENTS

This project uses geospatial and operational data to group delivery areas into optimized clusters, understand the factors affecting delivery performance, and generate strategies that help companies deliver faster, safer, and more efficiently.

STEP 1

Bussiness Understanding

Including Project Background, Problem Statement, and Objectives

STEP 2

Data Understanding

Undertsand all about dataset

STEP 3

Data Preprocessing

Cleaning dataset to make a better exploration

STEP 4

Exploratory Data Analysis

Explore dataset to get more insight

STEP 5

Clustering Using K-Means

Making cluster for customer location

STEP 6

Deploy app using Streamlit

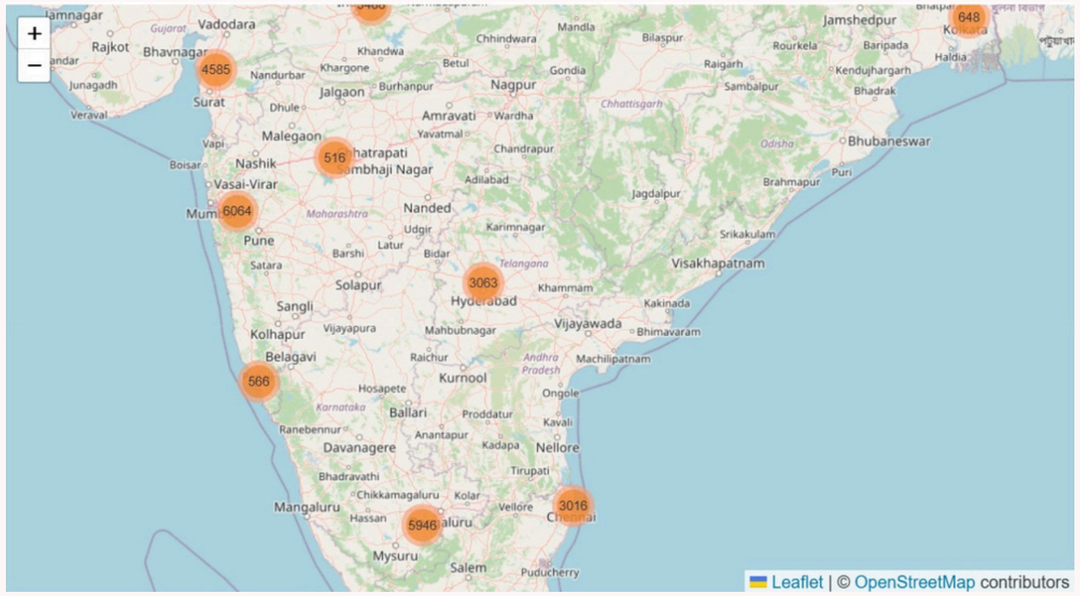
Deploy result of clustering using Streamlit

PROJECT BACKGROUND

Source: https://www.kaggle.com/datasets/saurabhbadole/zomato-delivery-operations-analytics-dataset

Efficient delivery operations are critical for Zomato to **ensure timely service** and **control costs**. When orders are geographically scattered, drivers travel longer distances, and delivery times become inconsistent. Using geospatial clustering, Zomato can **create optimized delivery zones to improve routing, reduce operational costs, and enhance delivery performance.**

BUSINESS PROBLEM



Source : Get map for customer location using folium

Longer delivery distances and higher costs

Drivers travel across wide areas, increasing fuel and labor expenses.

Inconsistent delivery times

Variability in routes and external factors like traffic or weather causes delays, reducing customer satisfaction.

Limited operational visibility

Without data-driven delivery zones, it is difficult to plan resources effectively and optimize routing.

The company needs a solution to define optimized delivery zones that minimize travel, reduce costs, and improve delivery consistency.

OBJECTIVES

The objective is to gain deeper insights to define optimized delivery zones that minimize travel, reduce costs, and improve delivery consistency.

Reduce delivery distances and operational costs

Cluster customer locations into optimized delivery zones to minimize driver travel.

Improve delivery consistency

Evaluate SLA per cluster and design zones that support predictable and on-time deliveries.

Enable data-driven decision making

Provide insights for assigning delivery zones and planning resources efficiently.

DATA UNDERSTANDING

This section outlines the dataset used in the project, describing its key features and how it reflects to make a delivery customer cluster.

01

Source Dataset

Kaggle - Zomato Delivery Operations Analytics Dataset

02

Total Data

45.584 rows and 20 columns

03

Data Informations

Provide delivery operations in India, including delivery person details, order timestamps, weather conditions, and more

04

Main Columns

Delivery_location_latitude and Delivery_location_longitude

DATA PREPROCESSING

the initial step in develop model aimed at cleaning and preparing the dataset for further exploration.

Duplicate Data Handling

No found data duplicate

Missing Value Handling

Drop NA > Time_Orderd
Fill NA median > Num columns
Fill NA mode > Cat columns

Outlier Handling

Focus on Lat and Long column: Filter only India lat 8-35 and India long 68-90

Changing Data Types

Column Order_Date and Time_Orderd from Obj to Date Time

RESULT

Initial data amount : 43853
Amount of clean data: 40188
Deleted data : 3665

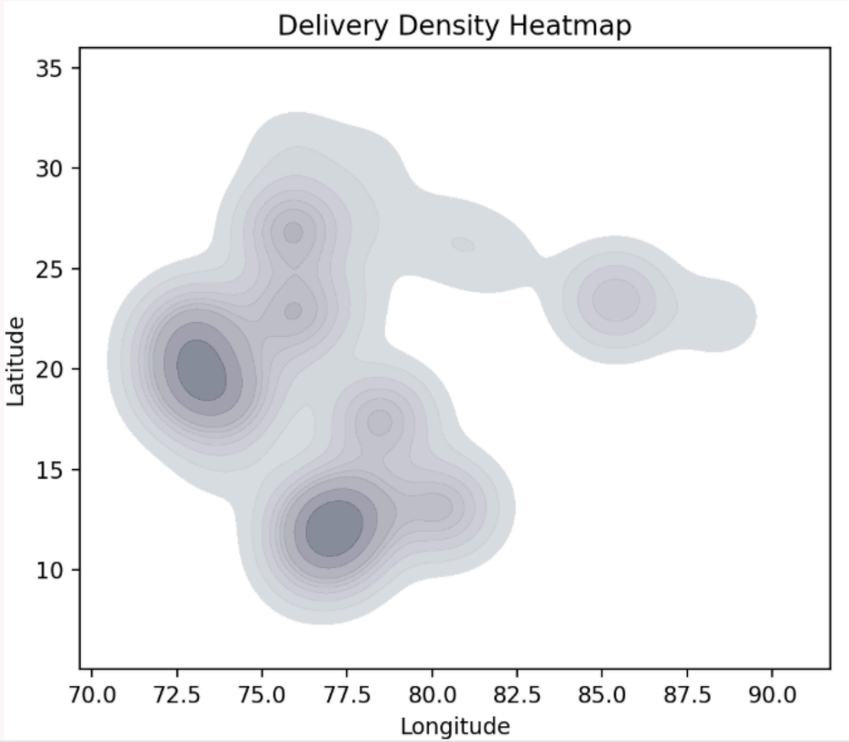
EXPLORATORY DATA ANALYSIS

To get more insight

GEOSPATIAL DISTRIBUTION ANALYSIS

Used to view point distribution, detect dense areas, location outliers, and understand the geographic structure that forms the basis for cluster formation.

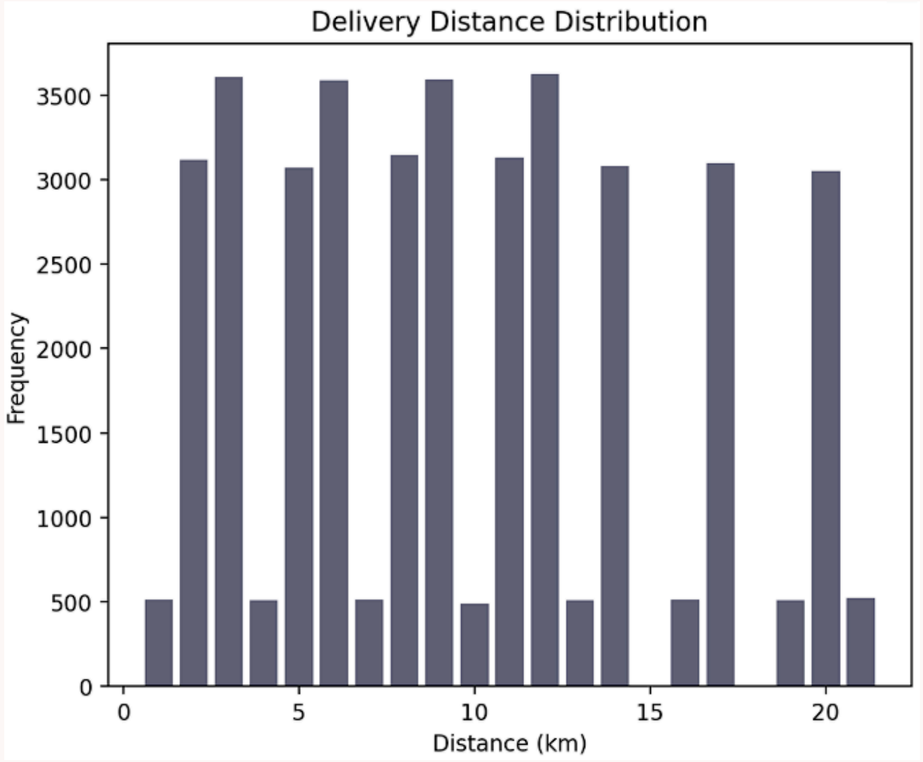
The scatter plot shows customer delivery points concentrated mainly around longitude 73–76 and latitude 18–22, forming several dense clusters.

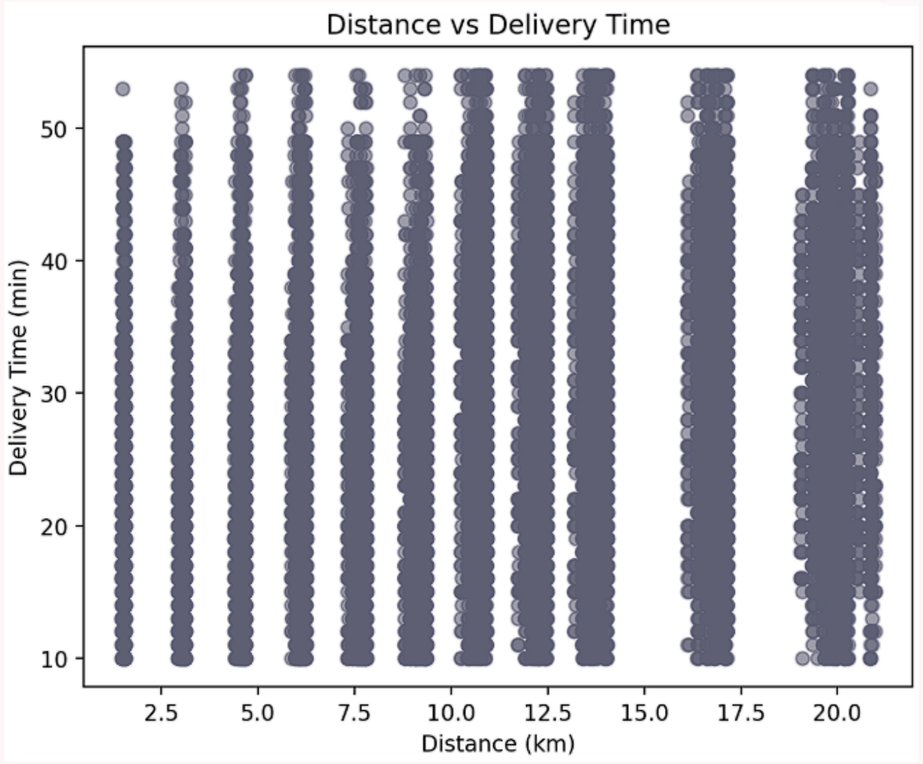


High-density areas appear around longitude 72–76 and latitude 18–22, indicating major customer hotspots.

DISTANCE ANALYSIS

This supports clustering because distance affects routes and dense clusters are usually close together





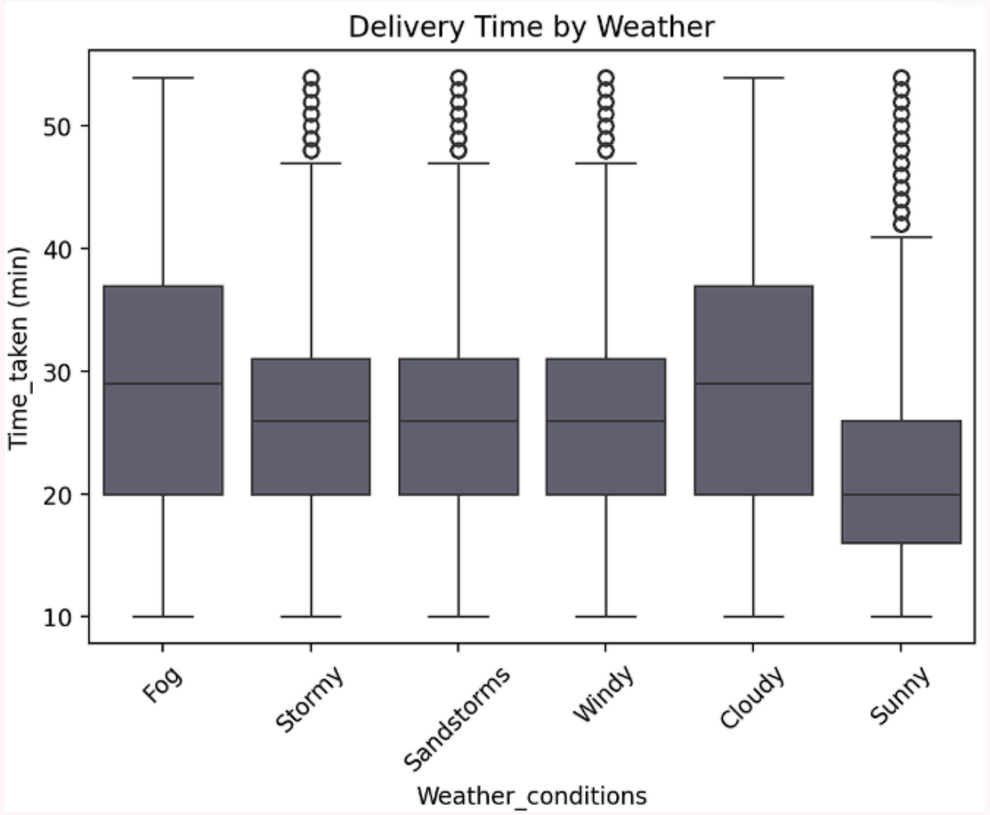
The barchart shows a fairly spread distribution of delivery distances, ranging from around 2 km to more than 20 km, with most deliveries falling within the 5–15 km range.

The scatter plot shows only a weak relationship between distance and delivery time. While delivery time slightly increases with distance, the data is widely scattered, suggesting other factors strongly influence the results.

DELIVERY TIME & DELAY ANALYSIS

Supports route optimization.

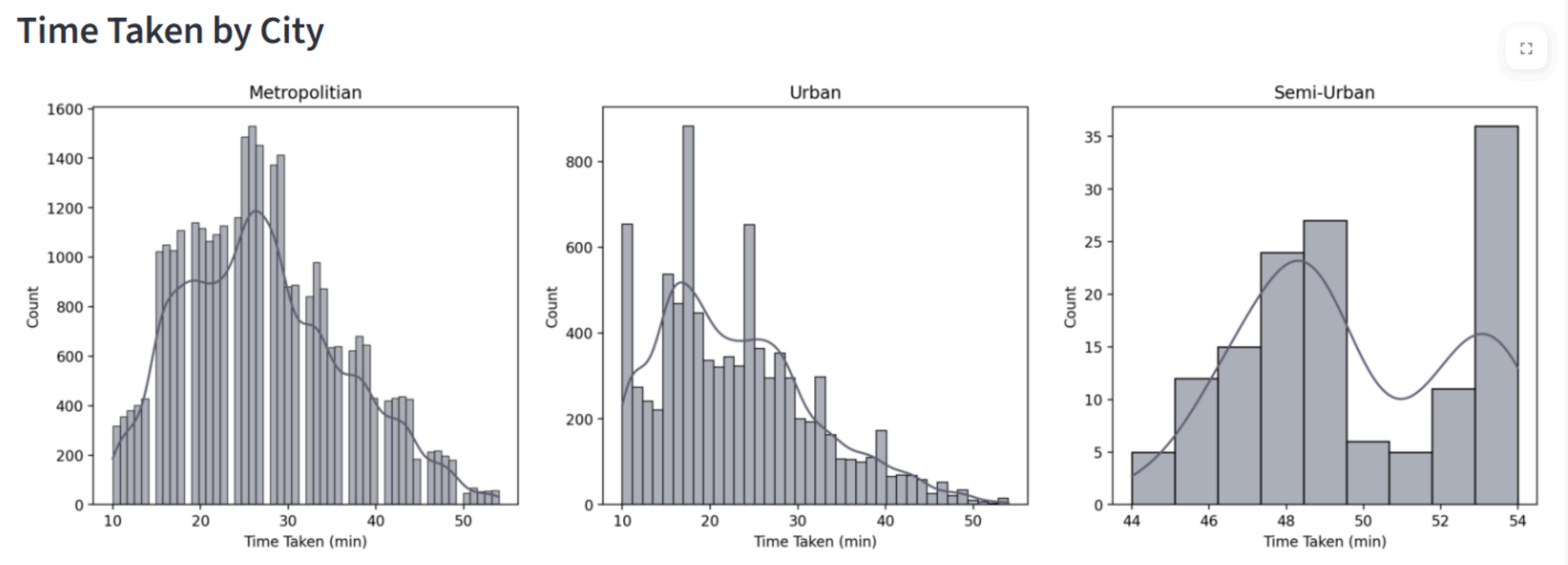
The bar chart shows the distribution of time taken for delivery. The range of time taken is between 10 and 60 minutes, with the majority of deliveries taking between 15 and 30 minutes.



Sunny weather has the fastest median delivery time (~18 minutes), while foggy and cloudy conditions have the longest (30–35 minutes). Stormy, sandstorm, and windy conditions show similar medians (25–30 minutes) but include outliers that indicate occasional severe delays.

DELIVERY TIME & DELAY ANALYSIS

Supports route optimization.

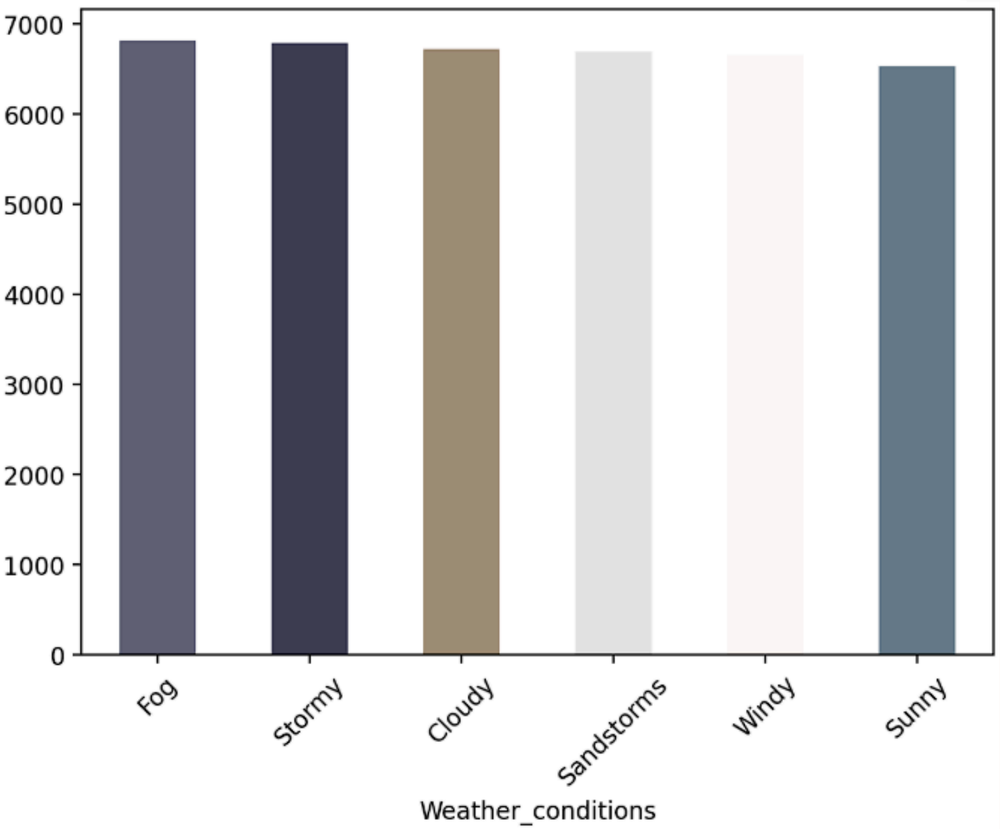


1. Metropolitan: Most deliveries 20–35 min, relatively stable despite high volume.
2. Urban: Faster deliveries 10–20 min, but long tail up to 50 min.
3. Semi-Urban: Slower 44–54 min but consistent, smaller dataset.

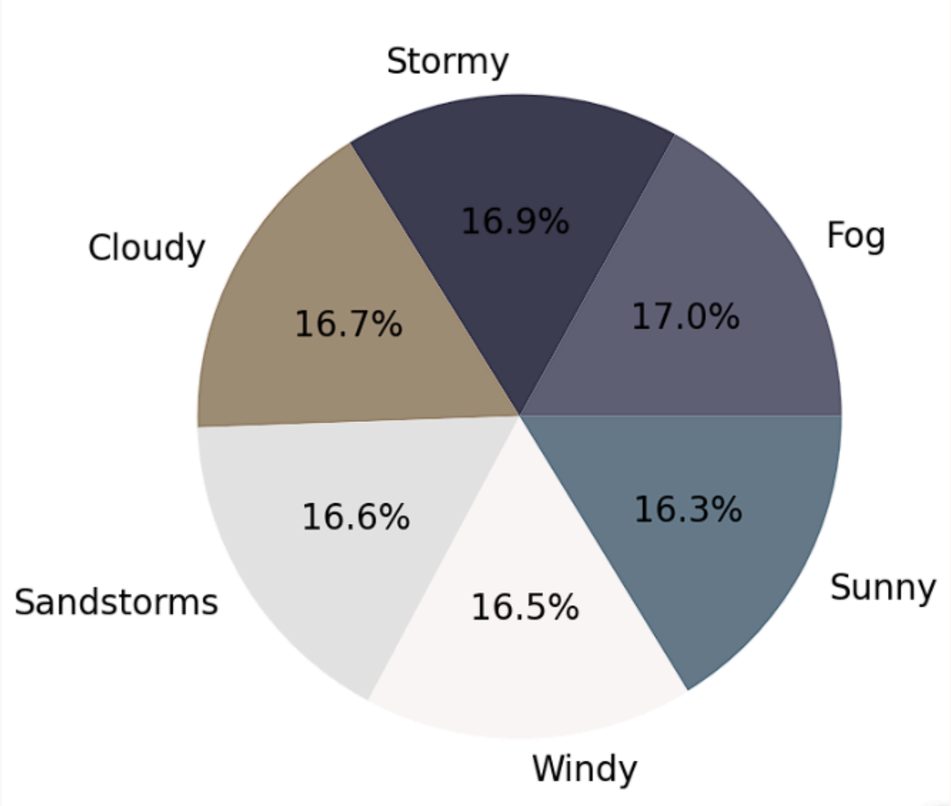
TRAFFIC & WEATHER PATTERN ANALYSIS

Used to read the characteristics of clusters after they are formed.

WEATHER CONDITIONS DISTRIBUTION



WEATHER CONDITIONS PERCENTAGE

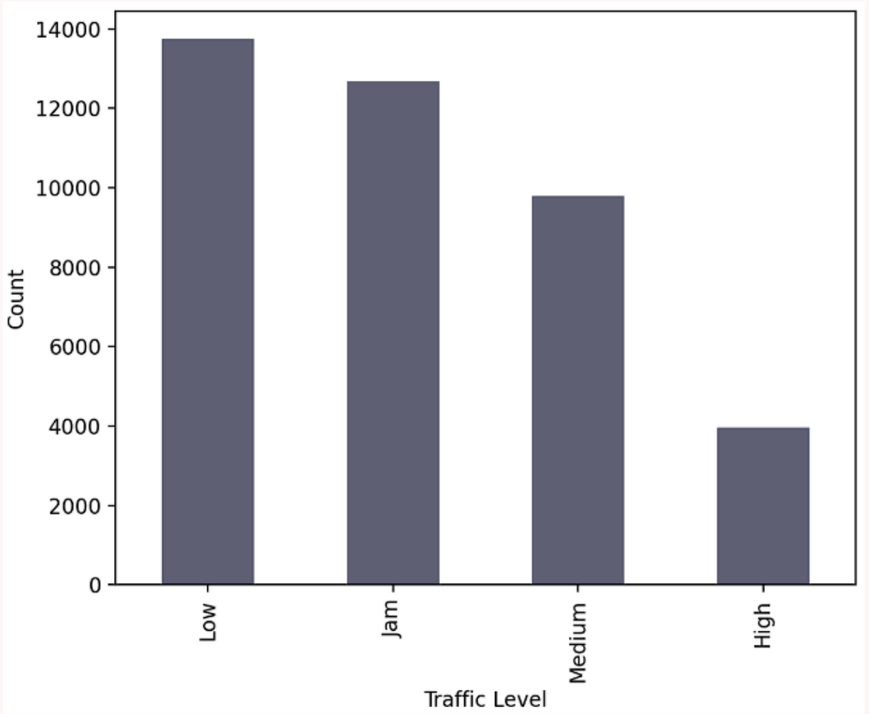


A very balanced number of observations for each weather condition. This means that the analysis of delivery times based on weather will not be biased due to the dominance of a particular weather condition in the dataset.

TRAFFIC & WEATHER PATTERN ANALYSIS

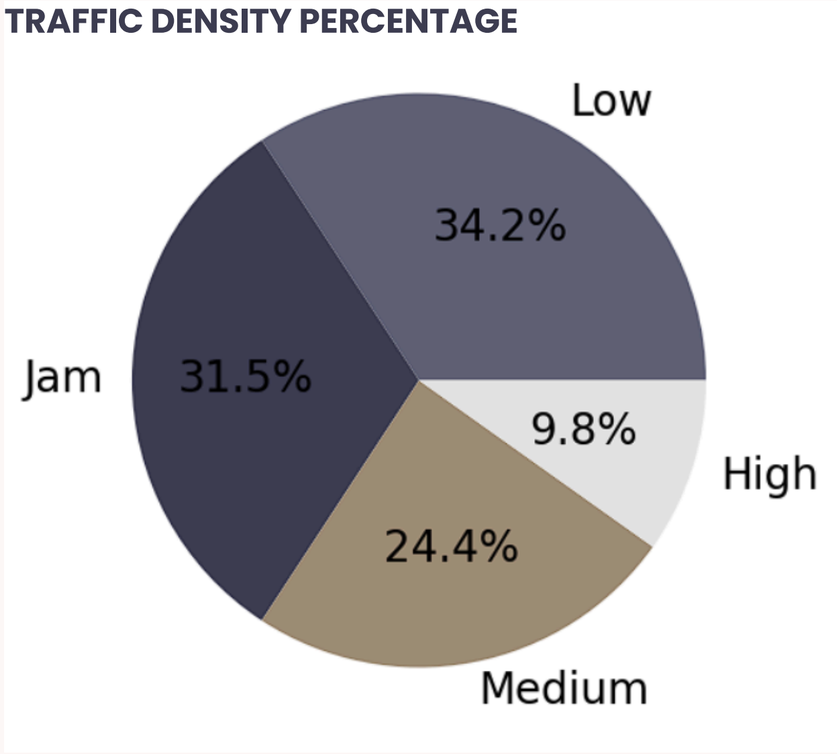
Used to read the characteristics of clusters after they are formed.

ROAD TRAFFIC DENSITY DISTRIBUTION



1. Jam: Peak 25–35 min, wide range up to 55 min.
2. High: Peak 25–30 min, range 20–40 min.
3. Medium: Peak 25–30 min, range 20–35 min.
4. Low: Peak 15–25 min, fastest and concentrated deliveries.

TRAFFIC DENSITY PERCENTAGE

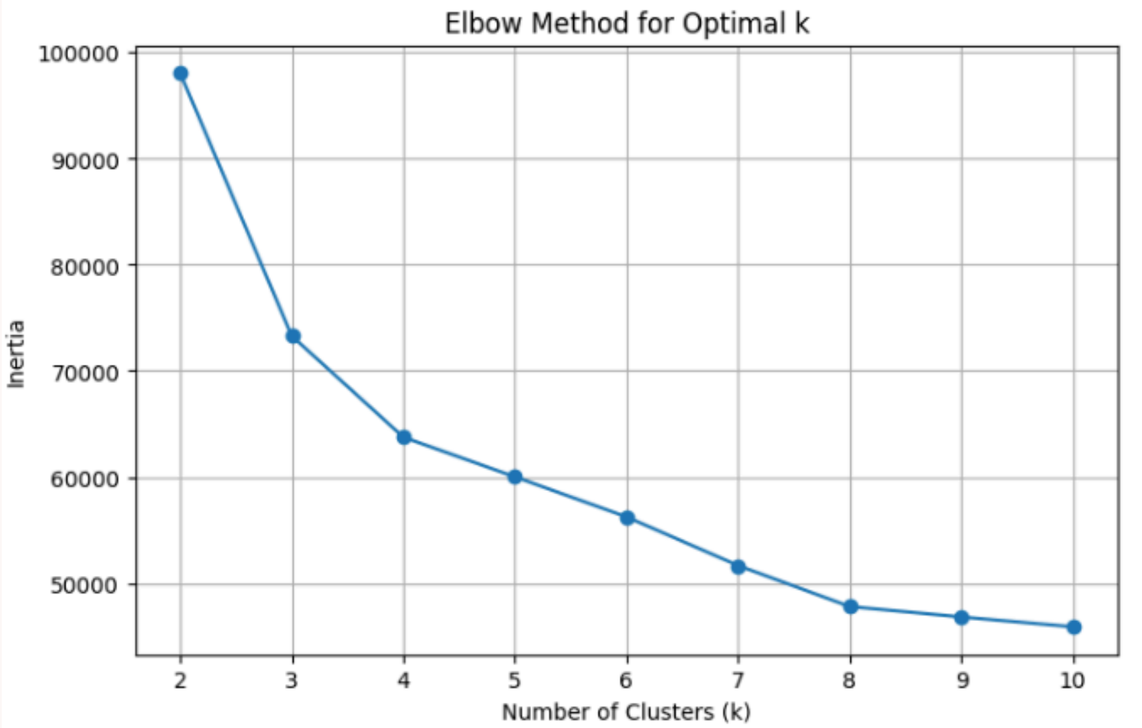


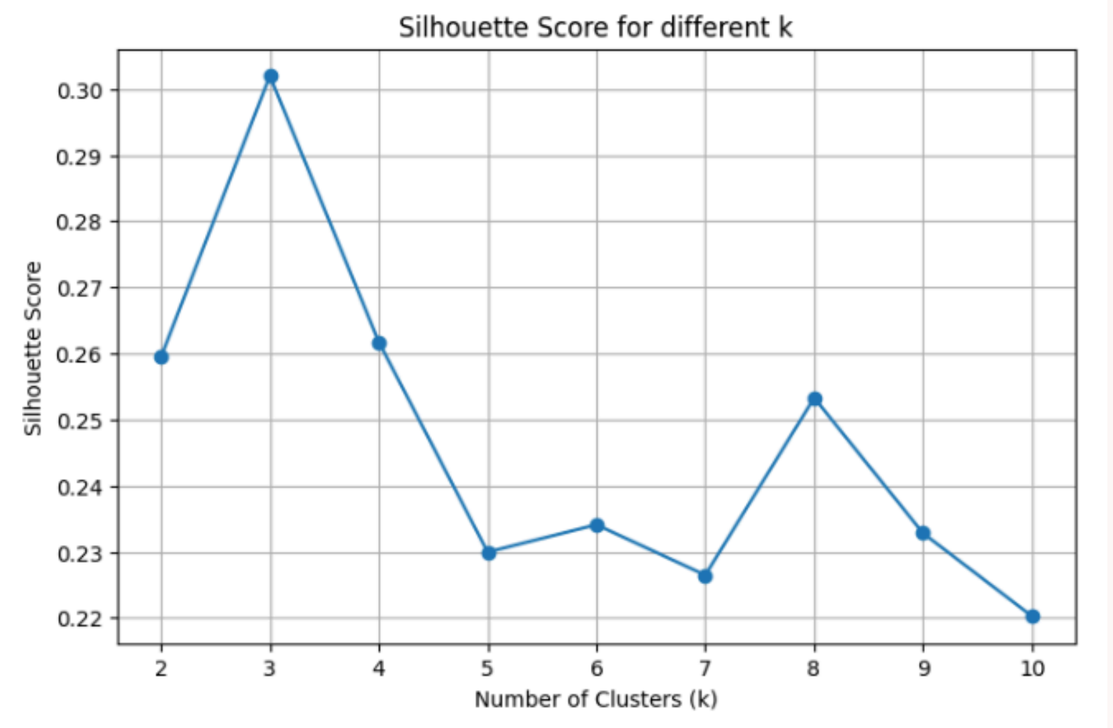
Low traffic density is the most common condition (~34.1%), while high traffic density is the least common (~9.8%). Traffic jams rank second (~31.6%), followed by medium traffic conditions (~24.5%).

CLUSTERING USING K-MEANS

FIND OPTIMAL K

"k" is the number of clusters we want to form.



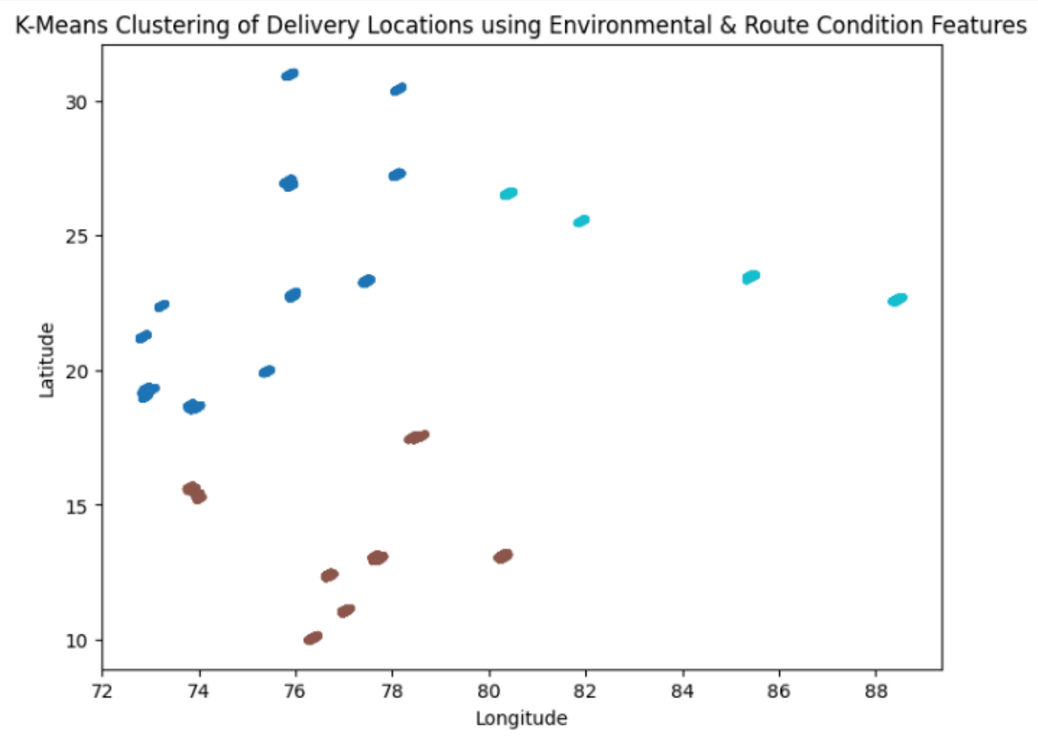


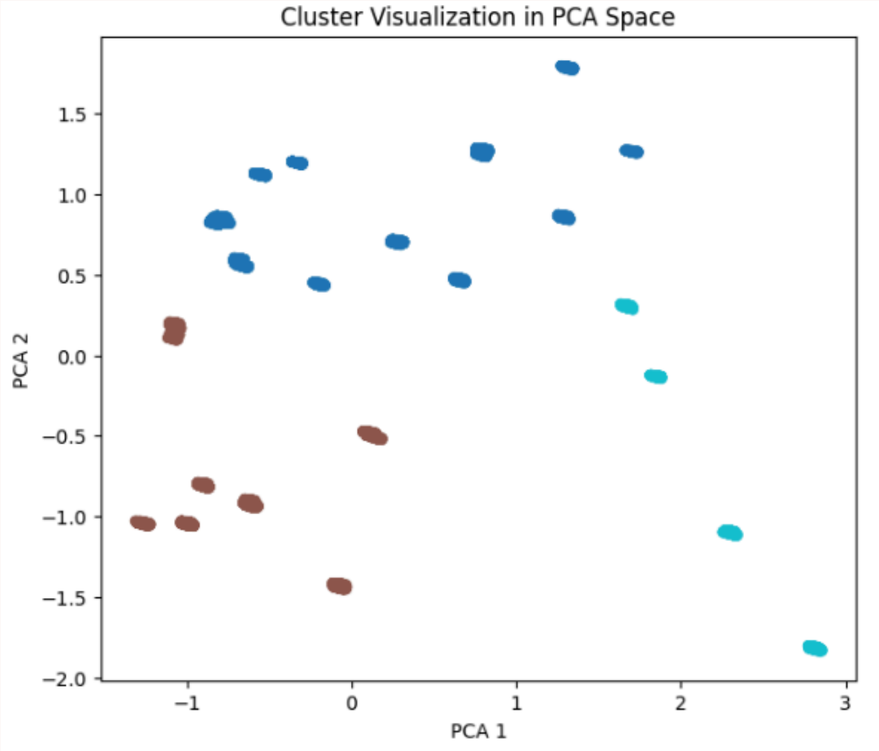
- The Elbow Method shows a sharp decrease in inertia at the beginning (K=3 to K=4), then begins to level off after K=4, indicating that adding more clusters after that point does not provide a significant improvement in cluster separation.
- The Silhouette Score graph shows that the highest value is at K = 3, which means that the clusters at K have the clearest separation and the most stable structure.

So, the optimal "k" based on 2 methods above is 3 and this result will be used on modelling

CLUSTERING RESULT

AIn the clustering modeling process, the features used include: numerical features “Delivery_location_latitude” and “Delivery_location_longitude,” categorical features ‘Road_traffic_density’ and “Weather_conditions.”



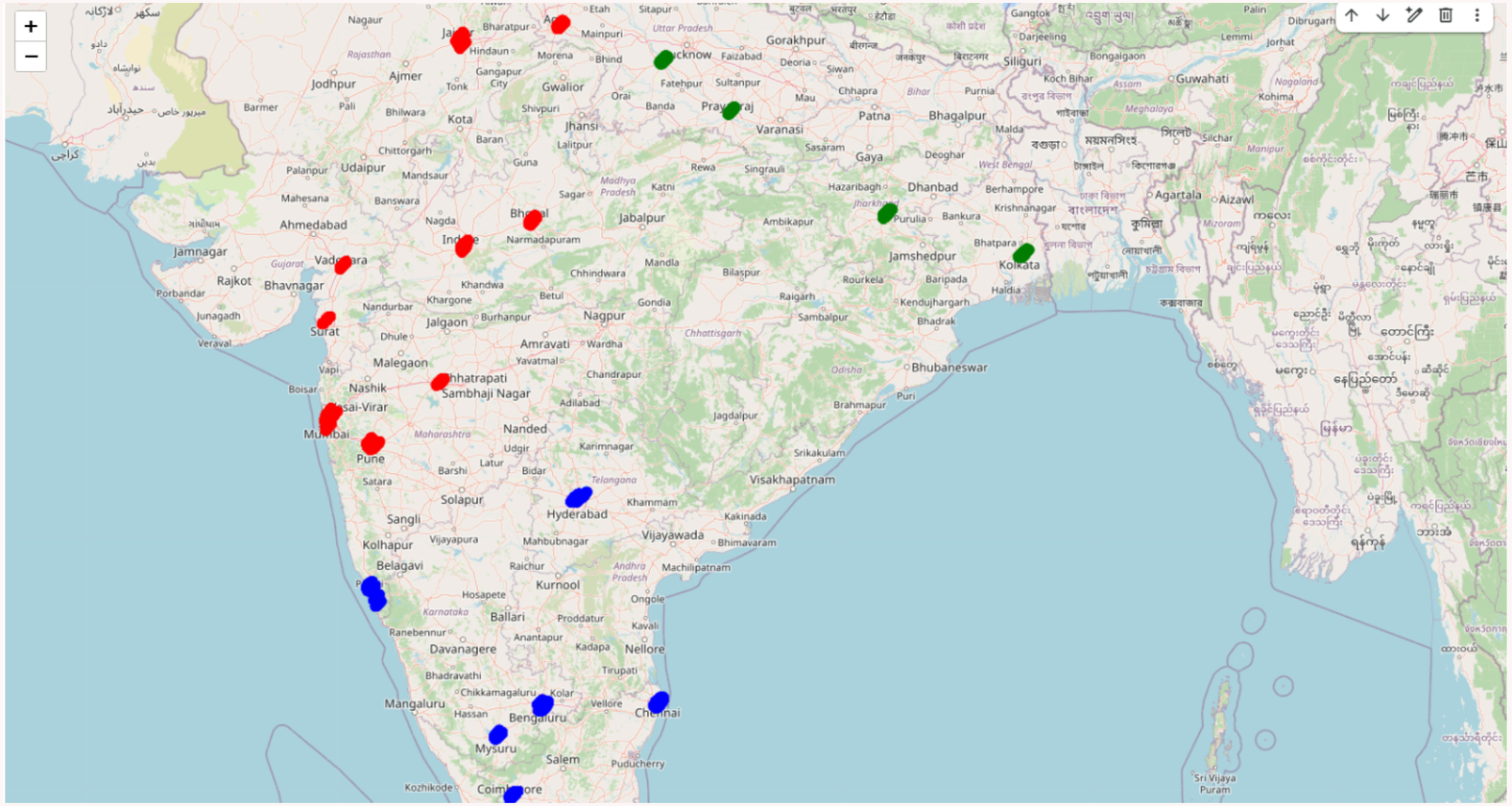


Cluster evaluation using Silhouette = 0.3022 >> Means modearte and usable for georgraphical dataset

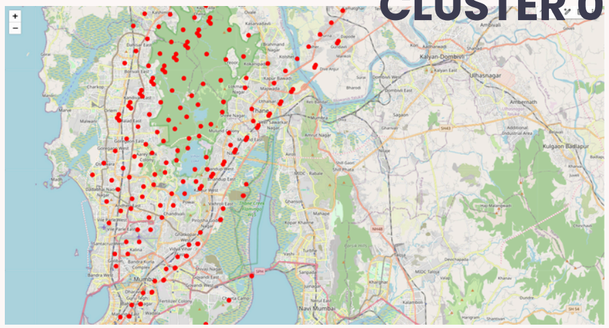
1. Cluster 0 — Northern Region with Very High Delivery Concentration: This cluster has a relatively high latitude value, indicating that deliveries occur mostly in the northern part of the operational area.
2. Cluster 1 — Southern Region with High Logistics Activity: This cluster is located at a lower latitude, which means it covers the southern region.
3. Cluster 2 — More Remote and Spread Out Eastern Region : This cluster has a much higher longitude value than other clusters, indicating a more easterly location.

DELIVERY ZONE MAP WITH CLUSTER

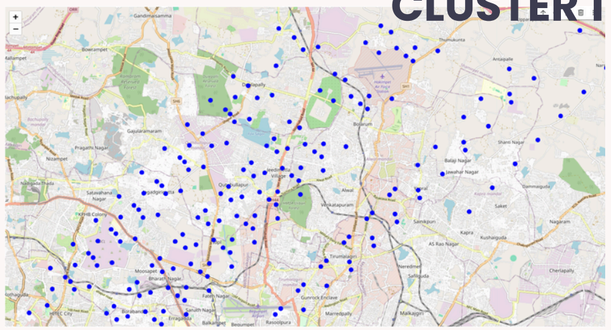
Here is a map of India marked with clustering results



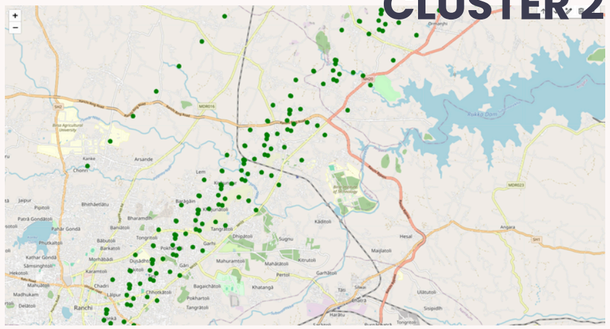
CLUSTER 0



CLUSTER 1



CLUSTER 2



CLUSTER SUMMARY

Cluster	Latitude mean	Longitude mean	Dominant Traffic	Dominant Weather	Order Total	Avg. Time	Std. Deviasi	SLA Time
0	22.523141	74.581245	Low	Cloudy	16110	26.34	9.33	35.67
1	13.359827	77.832133	Low	Stormy	13424	26.29	9.36	35.65
2	24.019721	84.686660	Low	Sandstorm	3533	26.57	9.50	36.07

SLA = promised delivery time to customers
Calculated per cluster because each zone has different characteristics (order volume, traffic, weather)
Function: Ensure orders are delivered on time, Monitor operational performance, and Assist with driver and route allocation

CONCLUSION

This is the summary of this project

- The clusters show that customers naturally group into specific zones. Assigning drivers to these clusters reduces the need to travel across wide areas.
- For example, Cluster 0 is concentrated around latitude 22.52 and longitude 74.58, allowing targeted routing instead of random assignment.

Delivery distances and operational costs can be reduced

- Average delivery times across clusters are similar (≈26–27 min), but standard deviations differ slightly (≈9–9.5 min), indicating some variability within zones.
- By monitoring SLA per cluster, Zomato can ensure predictable delivery times and address outliers.

Delivery consistency can be improved

Clusters provide a clear mapping of high-demand areas, traffic, and weather patterns, enabling data-driven decisions for resource allocation, driver scheduling, and contingency planning.

Operational visibility is enhanced

RECOMMENDATION

Here are some recommendations based on the Clustering using K-Means

Assign drivers per cluster

Allocate dedicated drivers to each delivery zone (cluster) to minimize travel distances and improve consistency exc: Clus 0 > 10-12, Clus 1 > 8-10, Clus 2 > 2-3

Optimize routes within each cluster

Use routing algorithms (e.g., Google OR-Tools or TSP solver) to define shortest/fastest paths within each cluster, considering traffic and weather conditions.

Monitor SLA and performance per cluster

Identify clusters with higher variability (e.g., Cluster 2 with std_dev ≈ 9.5 min) for targeted operational improvements.

Dynamic resource allocation for external factors

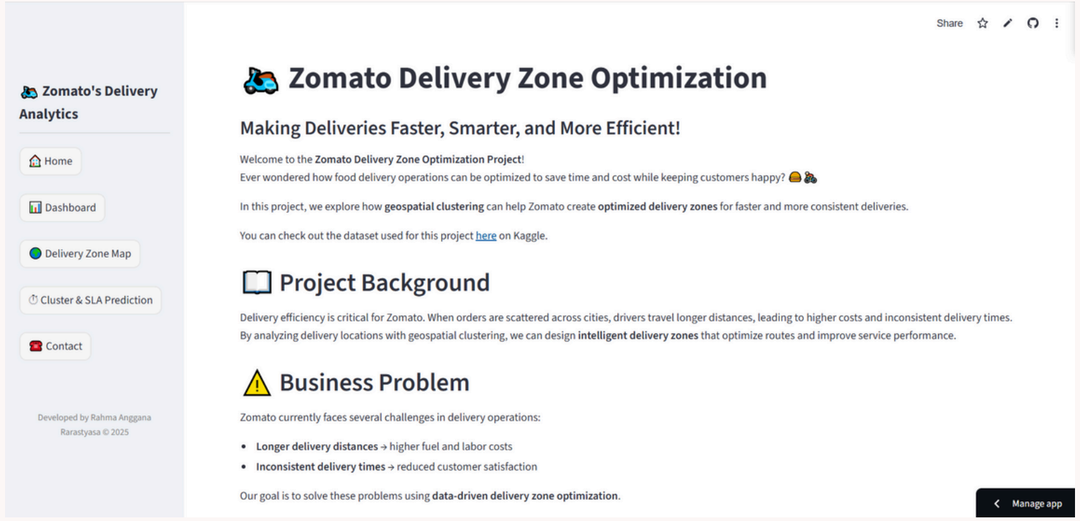
Adjust driver allocation in real-time based on dominant weather or traffic per cluster.

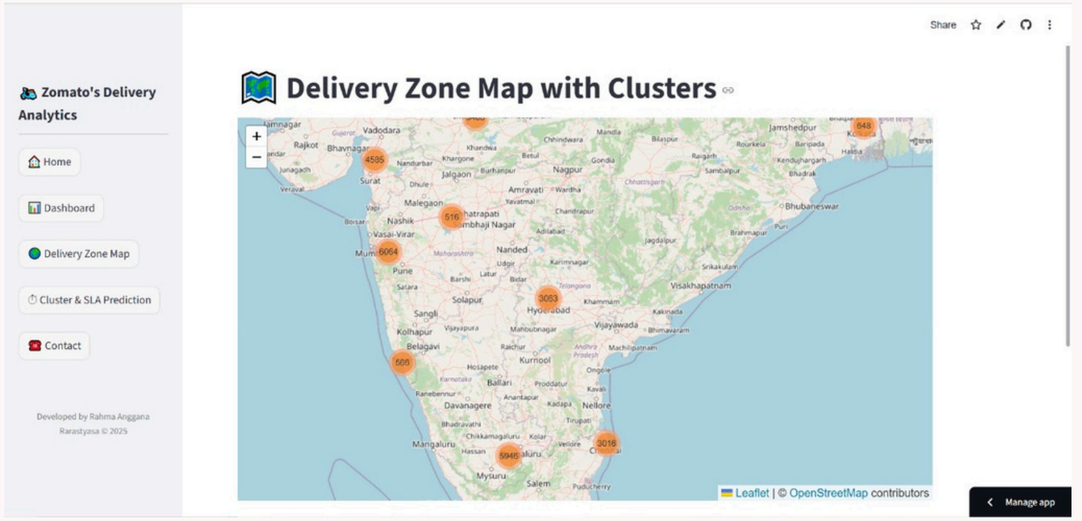
Data-driven zone adjustment

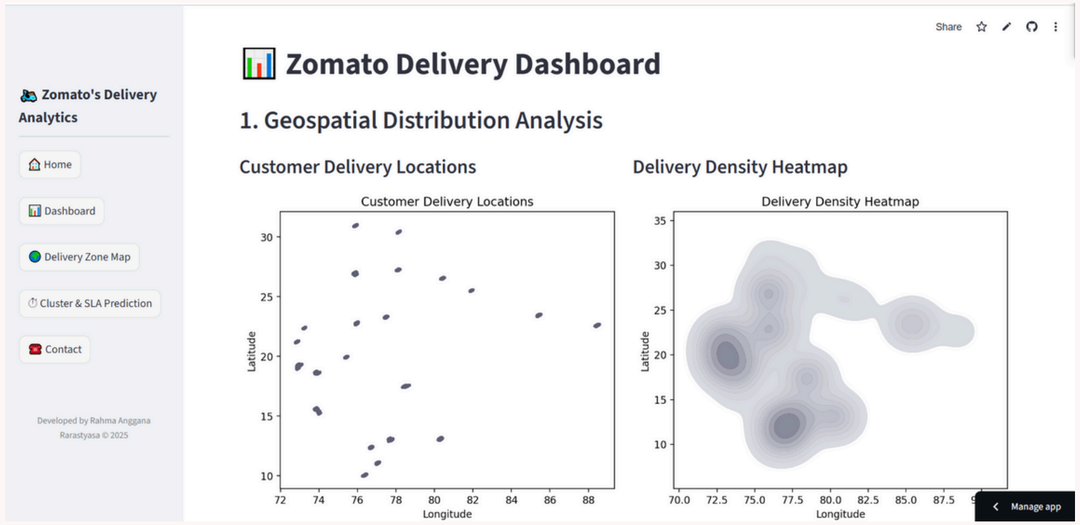
Re-run clustering periodically (e.g., monthly) to account for changing order patterns, seasonality, or new delivery areas.

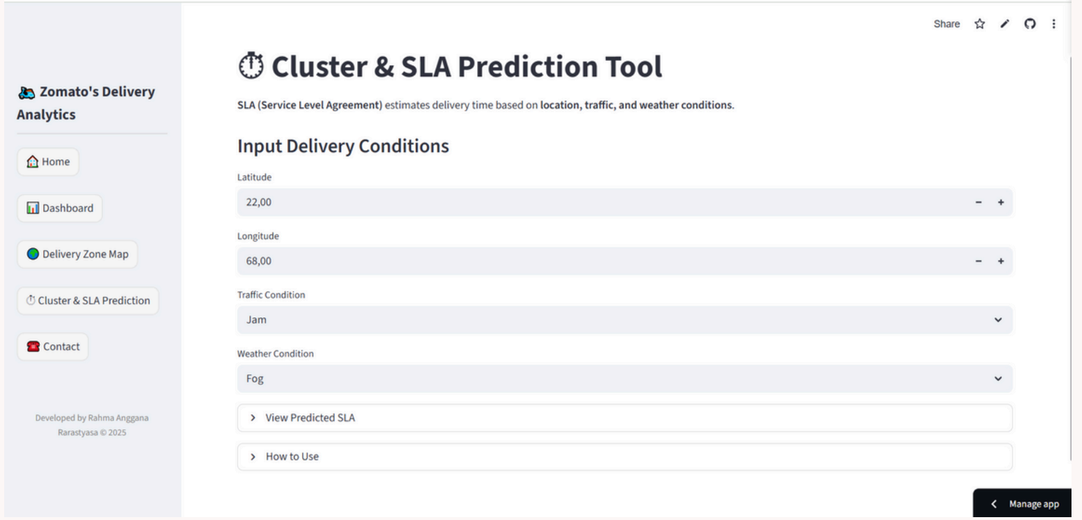
DASHBOARD

Developed using Streamlit. Get link here!

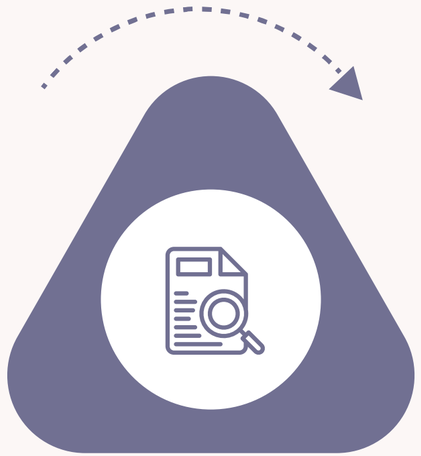








CONTACT ME ON

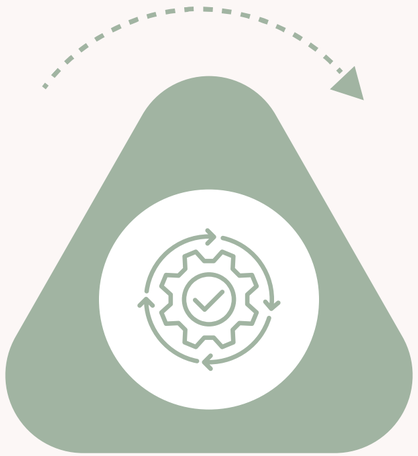


Linkedin

www.linkedin.com/in/rahma-anggana-rarastyasa

Email

rahmaanggana04@gmail.com

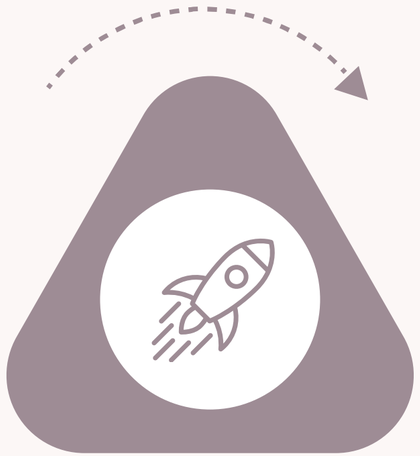


Github

https://github.com/rarastyasa/

Instagram

@rarastyasa_ra



Whatsapp

wa.me/6285257896356

THANK YOU