Demo 1

CMPT-3835 Machine Learning Work Integrated Project II

AGENDA

- 1. INTRODUCTION
- 2. DATASET OVERVIEW
- 3. DATA CLEANING & PREPROCESSING
- 4. EXPLORATORY DATA ANALYSIS (EDA)
- 5. FEATURE ENGINEERING
- 6. CHALLENGES & SOLUTIONS
- 9. Q&A SESSION

PROBLEM STATEMENT

Islamic Family aims to optimize food hamper distribution in Edmonton by identifying areas with high or low demand. By analyzing historical data and socio-economic factors, we can pinpoint regions requiring targeted outreach or mobile distribution points..

PROJECT OBJECTIVE

- Predict geographic areas with increasing or decreasing demand for food hampers.
- Use data-driven insights to assist Islamic Family in efficient resource allocation.
- Incorporate socio-economic factors to refine predictions and improve outreach strategies.

ABOUT DATA SET

Clients Dataset: (25505,44)

- Includes details about individuals receiving food hampers, such as age, address, family status, language preferences, and client status.
- Features like preferred contact methods and socio-economic indicators help in understanding the demographics and needs of recipients.

Food Hampers Dataset: (16605,39)

- Records each food hamper distribution event, including appointment details, pickup locations, scheduled dates, quantity of hampers distributed, and organization handling the distribution.
- Contains geospatial and scheduling data, helping to track demand trends and optimize future allocations.

ABOUT DATA SET

Number of Null values..

client_list= 119

Pickup_date = 9580

unique_client = 119

How did we fixed it ?

Mode imputation for client_list & unique_client

Forward Fill (ffill) or Backward Fill (bfill) &

Mean/Median Imputation for pickup_date, datetime_from, datetime_to, and related_scheduler

DATA CLEANING & TRANSFORMATION

1. Dropped Unnecessary Columns: Removed qrcode, meeting_link, and others to reduce noise.

2. Converted Data Types:

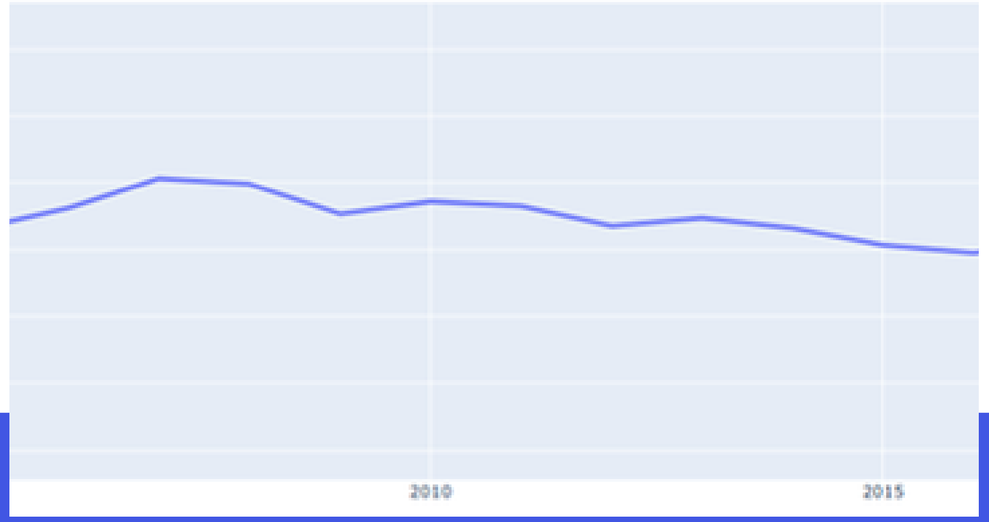
- Changed pickup_date, datetime_from, and datetime_to to proper datetime format.
- Used frequency encoding for categorical variables (client_list, hamper_type, etc.).

FEATURE ENGINEERING

date-related features from the collect_scheduled_date column.

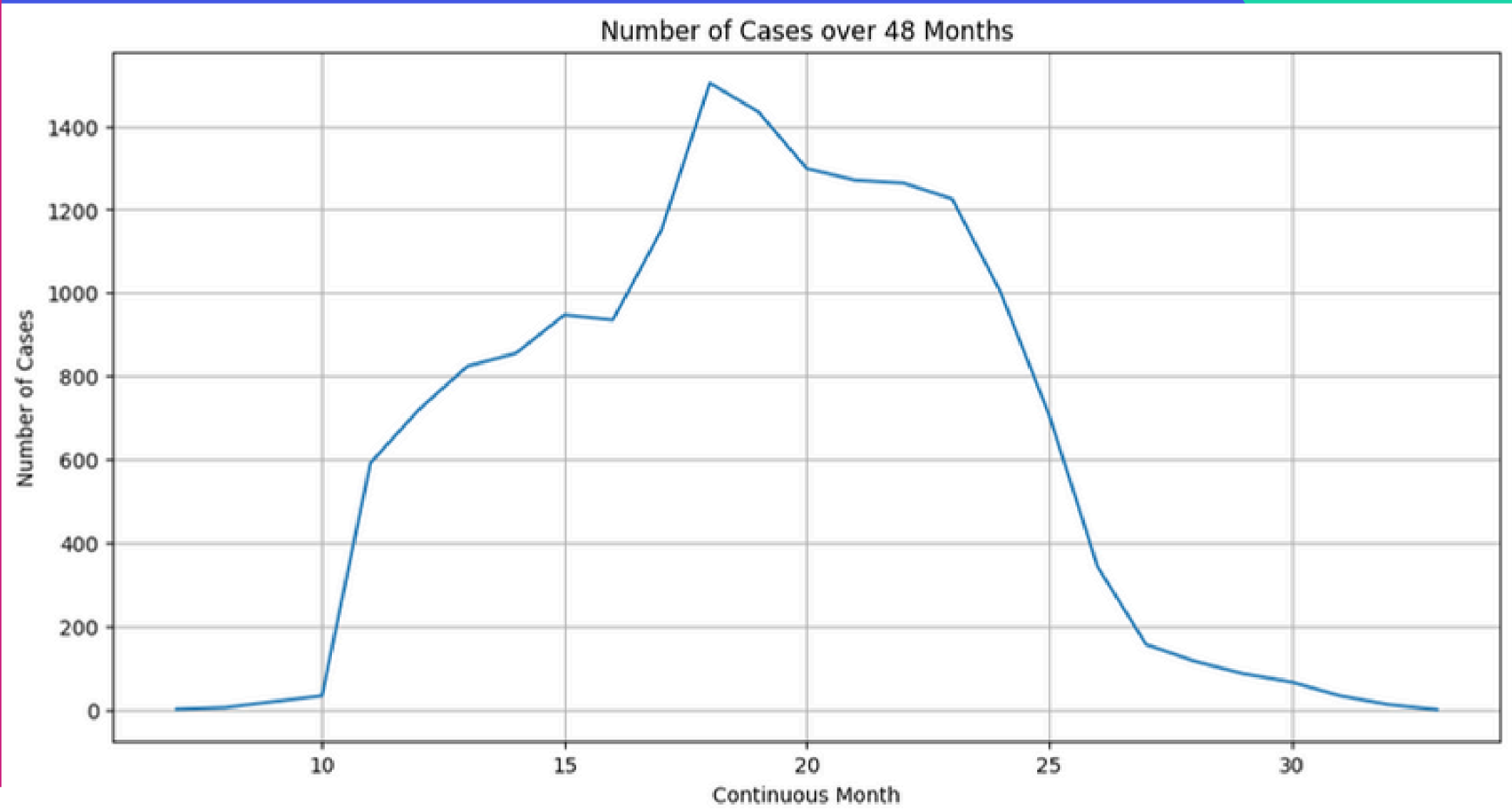
Feature Name	Description	Why?
`year`	Extracts the **year** from the date	Helps in analyzing yearly trends
`month`	Extracts the **month** (1-12)	Helps in identifying seasonal patterns
`day`	Extracts the **day of the month** (1-31)	Useful for daily trend analysis
`day_of_week`	Extracts the **day of the week** (0=Monday, 6=Sunday)	Helps in weekday/weekend analysis
`week_of_year`	Extracts the **week number** of the year (1-52)	Useful for tracking weekly trends

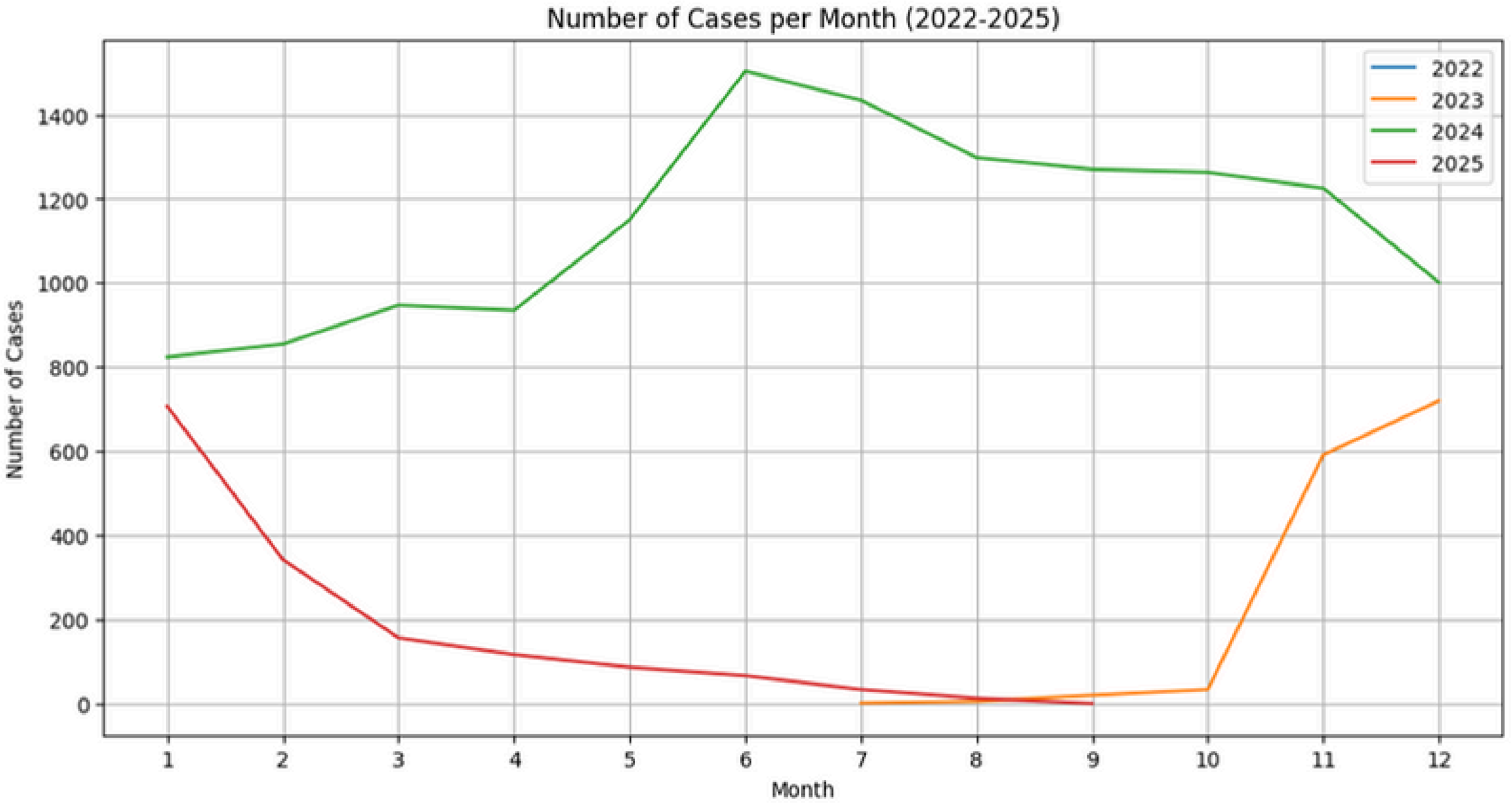
FREQUENCY ENCODING

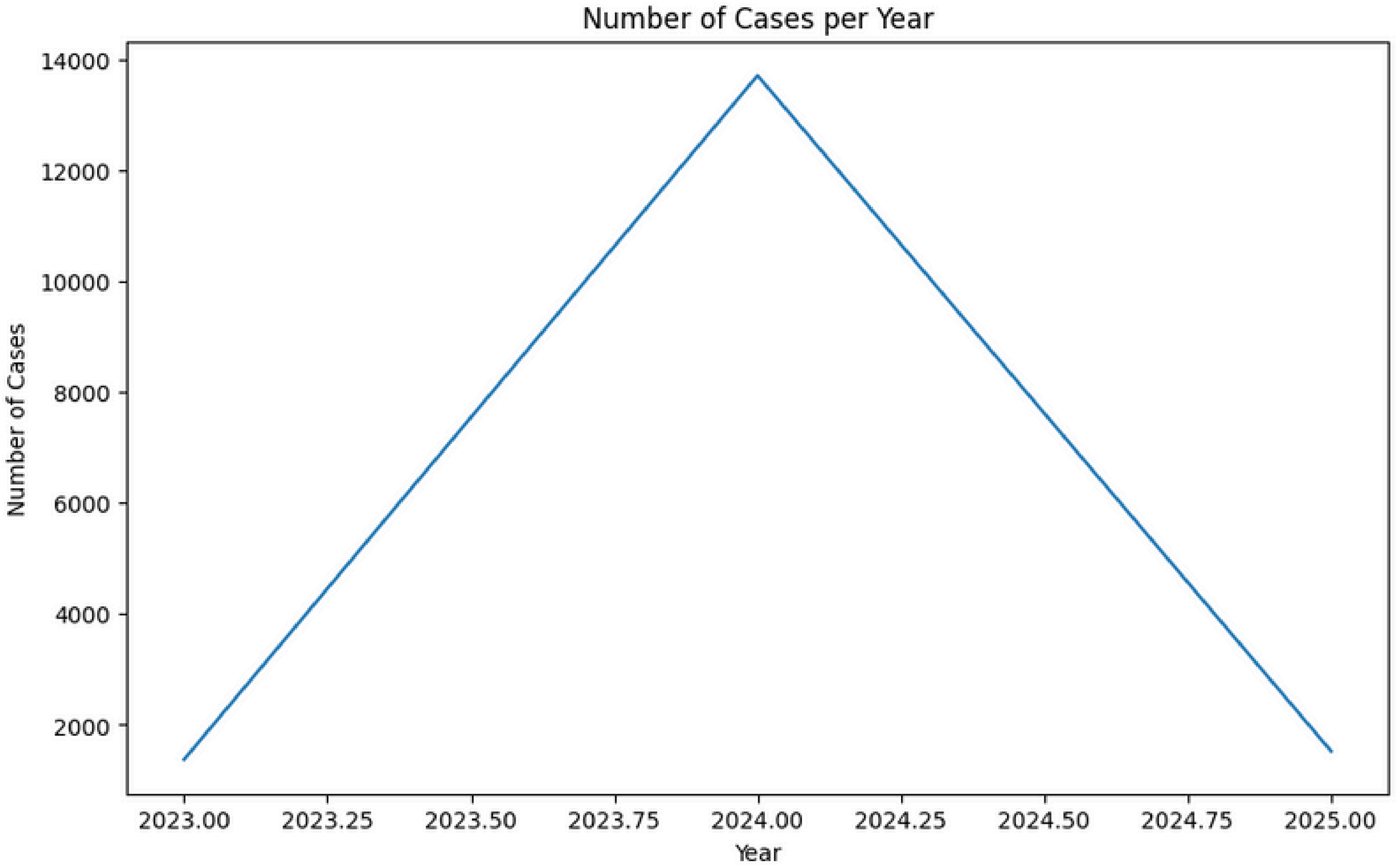


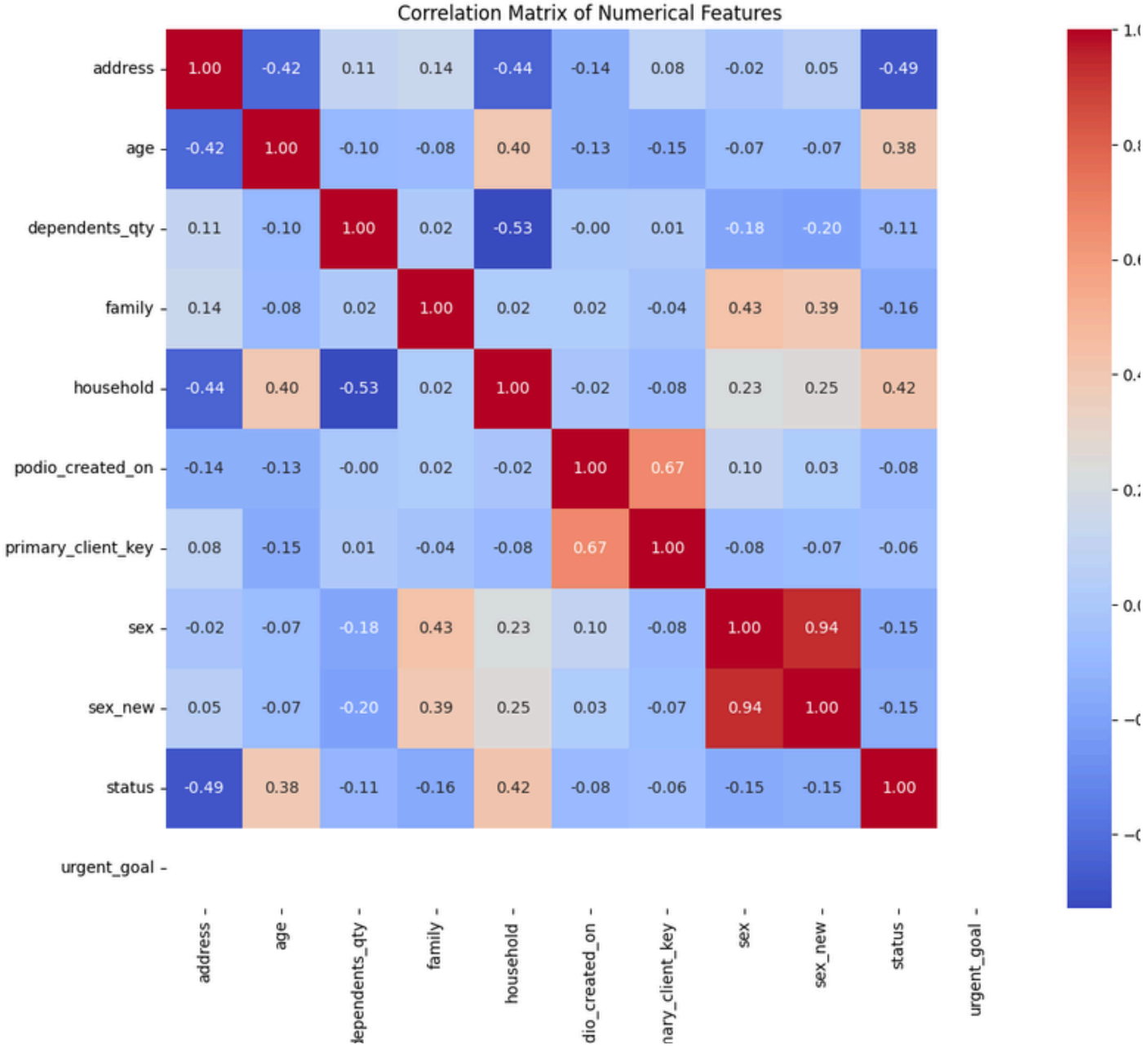
Transforms categorical variables into numerical values by frequency encoding
WHY? To handle unique Id without losing potential information.

GROUPS DATA PER MONTH AND COUNTS OCCURRENCES (CASES).




Number of Cases per Month (2022-2025)
2022
2023
2024
2025
Number of Cases
0
200
400
600
800
1000
1200
1400
Month
1
2
3
4
5
6
7
8
9
10
11
12


Number of Cases per Year
14000
12000
10000
8000
6000
4000
2000
Number of Cases
2023.00
2023.25
2023.50
2023.75
2024.00
2024.25
2024.50
2024.75
2025.00
Year


Correlation Matrix of Numerical Features
address
age
dependents_qty
family
household
podio_created_on
primary_client_key
sex
sex_new
status
urgent_goal
1.00 -0.42 0.11 0.14 -0.44 -0.14 0.08 -0.02 0.05 -0.49
-0.42 1.00 -0.10 -0.08 0.40 -0.13 -0.15 -0.07 -0.07 0.38
0.11 -0.10 1.00 0.02 -0.53 -0.00 0.01 -0.18 -0.20 -0.11
0.14 -0.08 0.02 1.00 0.02 0.02 -0.04 0.43 0.39 -0.16
-0.44 0.40 -0.53 0.02 1.00 -0.02 -0.08 0.23 0.25 0.42
-0.14 -0.13 -0.00 0.02 -0.02 1.00 0.67 0.10 0.03 -0.08
0.08 -0.15 0.01 -0.04 -0.08 0.67 1.00 -0.08 -0.07 -0.06
-0.02 -0.07 -0.18 0.43 0.23 0.10 -0.08 1.00 0.94 -0.15
0.05 -0.07 -0.20 0.39 0.25 0.03 -0.07 0.94 1.00 -0.15
-0.49 0.38 -0.11 -0.16 0.42 -0.08 -0.06 -0.15 -0.15 1.00

MERGING DATA SET

Merging food hamper and Clients Raw data set results increase in size to 31481 rows

Null values

address	24069
address_complement	28039
address_text	25096
age	471
bio	31476
...	...
where	31481
Modified Date_y	25470
Slug_y	31481
Creator_y	25470
unique id_y	25470

82 rows × 1 columns

CHALLENGES WE FACED

Handling Missing Values

Mismatched Column Names

Duplicate Entries After Merging

Handling Index Issues

Null values

address	24069
address_complement	28039
address_text	25096
age	471
bio	31476
...	...
where	31481
Modified Date_y	25470
Slug_y	31481
Creator_y	25470
unique id_y	25470

82 rows × 1 columns

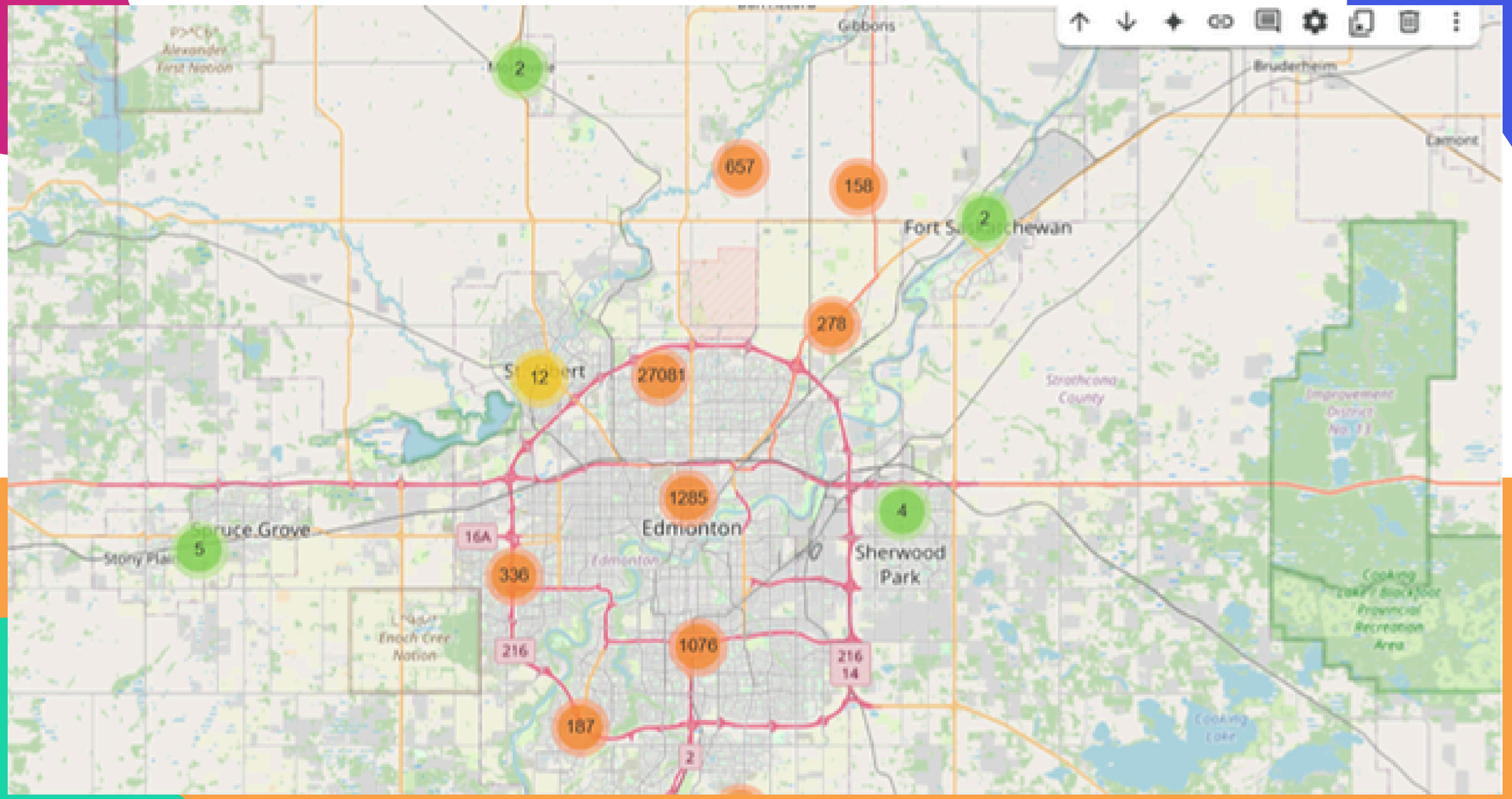
**Generate Spatial coordinates(Latitude and Longitude) using "zz_address_txt" *

Extract Postal Codes using regex

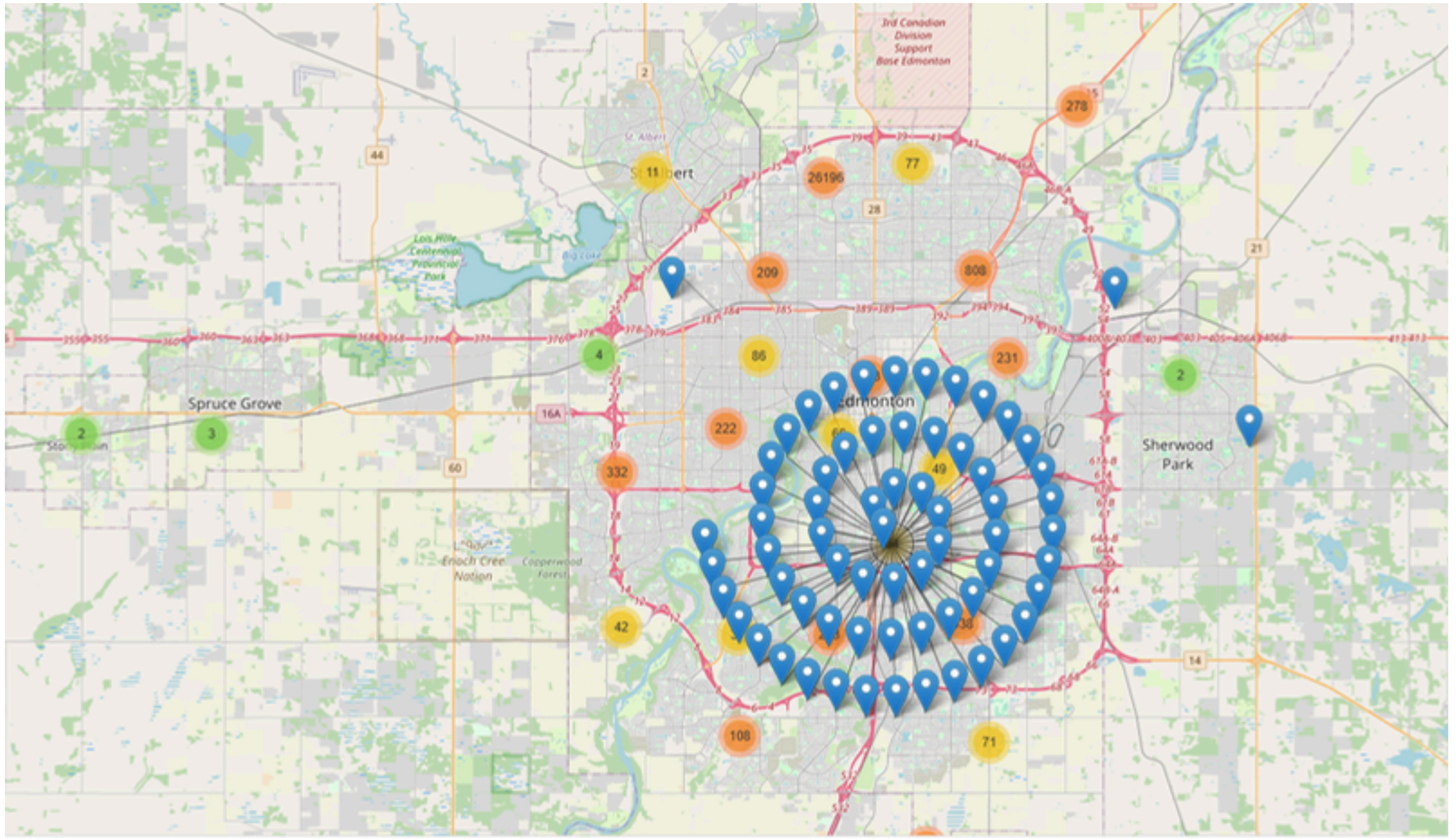
```
import re

# Extract Postal Codes using regex

#Geocoding
!pip install pgeocode
import pgeocode
nomi = pgeocode.Nominatim('ca')
```


Gibbons
Bruderheim
Fort Saskatchewan
657
158
2
278
27081
12
Strathcona County
1285
Edmonton
4
Sherwood Park
Spruce Grove
5
16A
336
216
1078
216
14
187
2

Create MarkerCluster to group the nearby points(for instance,MarkerCluster #64)



Explanation to the MVP:

https://drive.google.com/file/d/11Mc4kW_l50pkjv0AQtp-wAYDOmlJ9DM8/view?usp=sharing

By Sahil :

Link to the MVP:

https://fooddemand-yg3xzlfgfu3bpf66zzvtg4.streamlit.app/

THANK YOU!