# Practical Data Science

## Assignment A3

Giagkos Stylianos
f3352410

# Summary

**Section A**

"Sidewalk Obstructions  Annotation on Athens Images "

3/36

**Section B**

Data Mining

7/36

**Section C**

Unsupervised Learning

20/36

**Section D**

End Application

30/36

**Q&A**

Practical Data Science Assignment 3

35/36

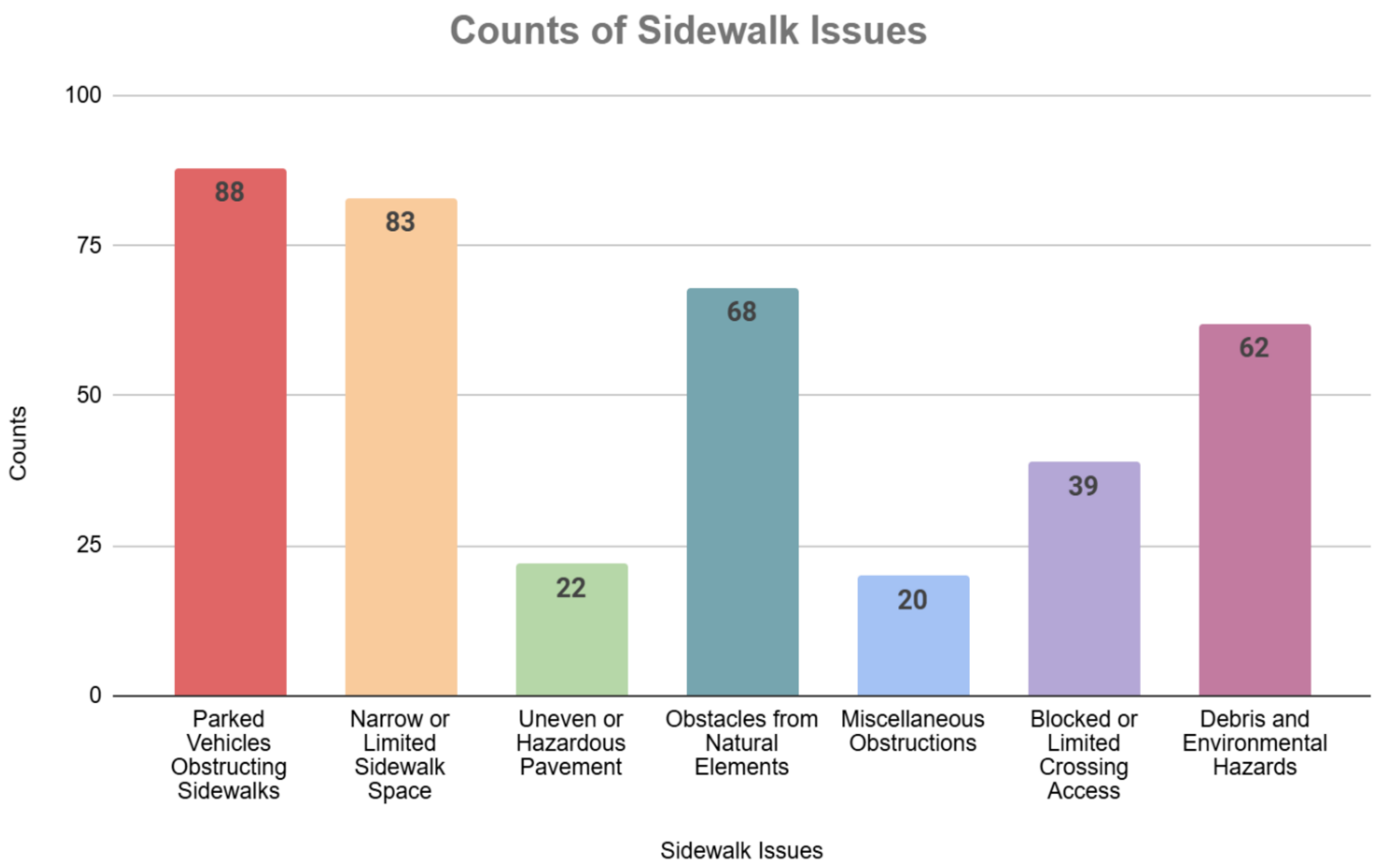
# Section A

“Sidewalk Obstructions  Annotation on Athens Images ”

# Counts of Sidewalk Issues

# Counts of Sidewalk Issues

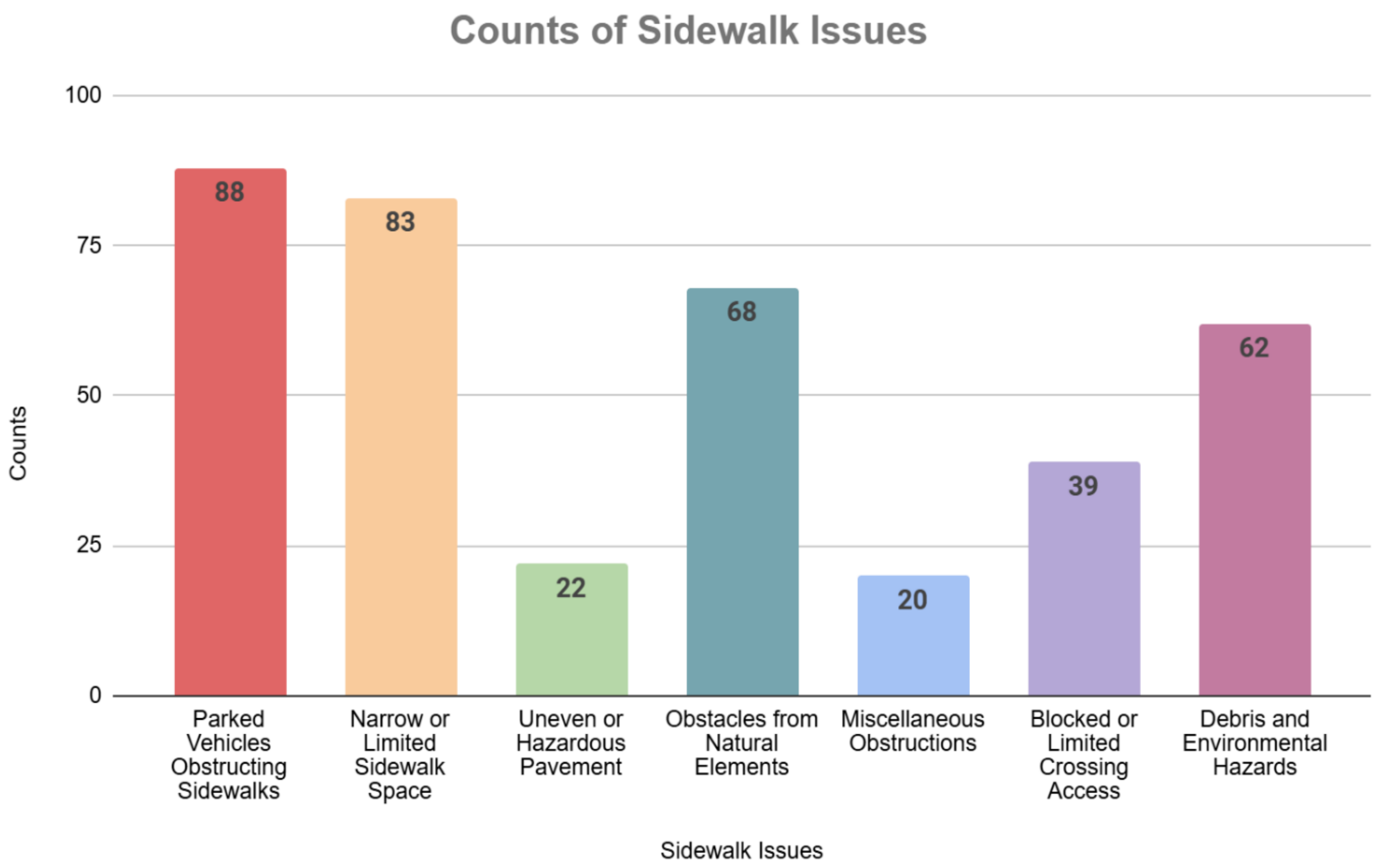
- Parked Vehicles Obstructing Sidewalks
  - Frequency: 88 instances (**23.04% of issues**).
- Narrow or Limited Sidewalk Space
  - Frequency: 83 instances (**21.73% of issues**).
- Uneven or Hazardous Pavement
  - Frequency: 22 instances (**5.76% of issues**).
- Obstacles from Natural Elements
  - Frequency: 68 instances (**17.80% of issues**).
- Miscellaneous Obstructions
  - Frequency: 20 instances (**5.24% of issues**).
- Blocked or Limited Crossing Access
  - Frequency: 39 instances (**10% of issues**).
- Debris and Environmental Hazards
  - Frequency: 62 instances (**16.23%** of issues).



# Counts of Sidewalk Issues

*All seven categories are well-represented, with Parked Vehicles Obstructing Sidewalks and Narrow Sidewalk Space being the most frequent.*

*Many issues overlap, compounding pedestrian challenges. The "Other" category was unnecessary as all observations fit predefined categories.*



# Section B

Data Mining

# Initial Dataset Preprocessing

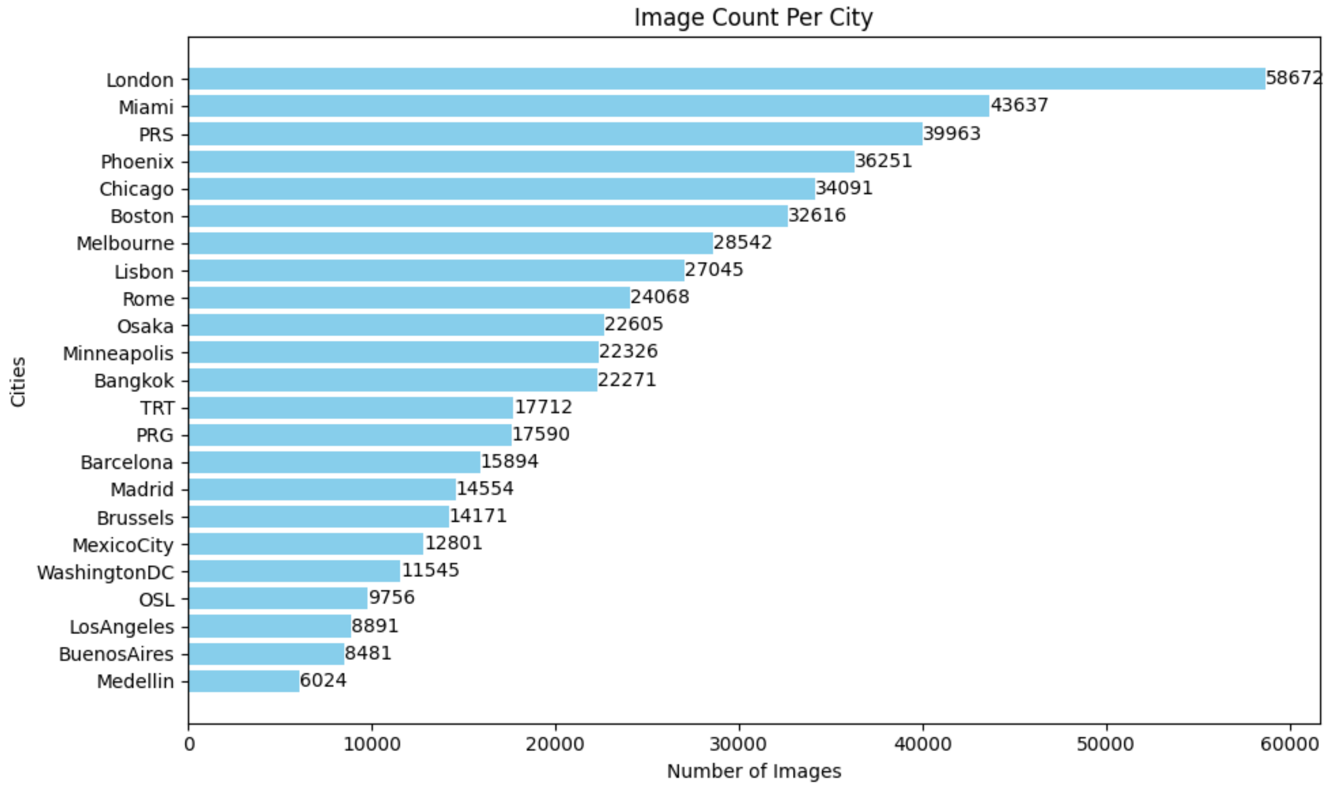
# Exploring the **"GSV Cities"** dataset

Focusing on acquisition, exploration, image analysis, and preprocessing.

- Dataset Details:
  - **Source:** Kaggle dataset with city images and metadata.
  - Metadata Includes:
    - place_id, year, month, northdeg
    - city_id, lat, lon, panoid
- **Images:** Organized in city-specific folders, covering diverse geographical areas.

# Key Steps in the Process

- **Downloading the Dataset**:
  - The dataset is downloaded from Kaggle and stored in a specific directory.
- **Exploring the Dataset**:
  - Listing directories and files.
  - Inspecting metadata for missing values and structure.
- **Image Count per City:**
  - Visualizing image counts with a horizontal bar chart.
  - Cities like **London**, **Miami**, and **Phoenix** have the most images, while **Medellin** and **Buenos Aires** have fewer.



# Key Steps in the Process

- **Sampling Images:**
  - Random sampling of images per city to manage memory size.
  - Storing samples and metadata in Google Drive and pickle files.
- **Merging DataFrames:**
  - Merging individual city DataFrames into one consolidated DataFrame for analysis.
  - Saving the merged dataset as a CSV.
- **Removing Duplicates:**
  - Identifying and removing duplicate metadata rows and image files.
- **Preprocessing Images for K-means:**
  - Resizing images to 512x512, converting to RGB, and normalizing pixel values to the range [0,1] for K-means clustering.

# EDA on sampled Data

# Final Sampled Dataframe

Columns:
- **Place ID**: Unique identifier for places.
- **Year**: Year the data was recorded.
- **Month**: Month the data was recorded.
- **Northdeg**: Measurement indicating northern direction (degree).
- **Lat**: Latitude of the location.
- **Lon**: Longitude of the location.

| | place_id | year | month | northdeg | city_id | lat | lon | panoid | image_name | image_path |
|---|---|---|---|---|---|---|---|---|---|---|
| 360027 | 4538 | 2009 | 10 | 58 | Minneapolis | 44.987275 | -93.221127 | I2CfJnG67tnSAsiA3NMm1g | Minneapolis_0060027_2009_10_058_44.98727479016... | /content/drive/My Drive/PDS_A3/PDS_A3_Sampled_... |
| 350694 | 2916 | 2011 | 9 | 263 | Minneapolis | 44.969870 | -93.295165 | IjRpYVa8kN2fCd2PMNFmug | Minneapolis_0050694_2011_09_263_44.96987026561... | /content/drive/My Drive/PDS_A3/PDS_A3_Sampled_... |
| 350270 | 2742 | 2011 | 6 | 283 | Minneapolis | 44.968334 | -93.287806 | HkEPC9_d6P640aisJYngbA | Minneapolis_0050270_2011_06_283_44.96833429689... | /content/drive/My Drive/PDS_A3/PDS_A3_Sampled_... |
| 359023 | 3223 | 2016 | 9 | 225 | Minneapolis | 44.973403 | -93.261019 | TWBksn5jDCQzK80nhrCD5Q | Minneapolis_0059023_2016_09_225_44.97340315544... | /content/drive/My Drive/PDS_A3/PDS_A3_Sampled_... |
| 355953 | 136 | 2019 | 7 | 244 | Minneapolis | 44.939543 | -93.254126 | 5Vy75AKPixWUC0ScEC72Fg | Minneapolis_0055953_2019_07_244_44.93954344166... | /content/drive/My Drive/PDS_A3/PDS_A3_Sampled_... |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 113512 | 514 | 2014 | 9 | 348 | PRS | 48.809480 | 2.257262 | Q0CCsTy1wpXe1kgNBm3UIQ | PRS_0013512_2014_09_348_48.80947970218288_2.25... | /content/drive/My Drive/PDS_A3/PDS_A3_Sampled_... |
| 109063 | 1521 | 2016 | 6 | 554 | PRS | 48.831663 | 2.365907 | uVMEQA9EV1e-2q1ZH5dV0w | PRS_0009063_2016_06_554_48.83166304649817_2.36... | /content/drive/My Drive/PDS_A3/PDS_A3_Sampled_... |
| 134267 | 1573 | 2019 | 6 | 499 | PRS | 48.833256 | 2.317289 | mzxQyVLDLwlBHFH8ZWU3Bw | PRS_0034267_2019_06_499_48.83325630517728_2.31... | /content/drive/My Drive/PDS_A3/PDS_A3_Sampled_... |
| 116443 | 2719 | 2015 | 6 | 78 | PRS | 48.857426 | 2.399494 | 1jUSTVj_-zVEbt2mi5sC0g | PRS_0016443_2015_06_078_48.85742623906796_2.39... | /content/drive/My Drive/PDS_A3/PDS_A3_Sampled_... |
| 136262 | 1753 | 2014 | 7 | 686 | PRS | 48.837333 | 2.345425 | Z3iNkSW__jwOIFZOV4ueOA | PRS_0036262_2014_07_686_48.83733317560306_2.34... | /content/drive/My Drive/PDS_A3/PDS_A3_Sampled_... |

11309 rows × 10 columns

11309 rows × 10 columns

# Exploration

| | year | month | northdeg | lat | lon |
|---|---|---|---|---|---|
| count | 11309.000000 | 11309.000000 | 11309.000000 | 11309.000000 | 11309.000000 |
| mean | 2014.801574 | 6.781059 | 274.080821 | 31.893728 | -23.158901 |
| std | 3.576798 | 2.707434 | 157.099201 | 24.131281 | 73.269631 |
| min | 2007.000000 | 1.000000 | -4.000000 | -37.854864 | -118.269349 |
| 25% | 2013.000000 | 5.000000 | 153.000000 | 25.782030 | -80.217042 |
| 50% | 2015.000000 | 7.000000 | 267.000000 | 40.417174 | -9.167805 |
| 75% | 2018.000000 | 9.000000 | 365.000000 | 44.957103 | 10.701730 |
| max | 2021.000000 | 12.000000 | 726.000000 | 59.958047 | 145.024955 |

1. Year

•Spans **2007–2021**, with most data concentrated in **2013–2018** (IQR).

2. Month

•Higher density in **May–September**, possibly reflecting **seasonal activities** or **favorable conditions**.

3. Northdeg (Direction Measurement)

•Range: **-4 to 726**, IQR: **153–365**.

•Indicates diverse terrains or varying directional orientations.

4. Latitude

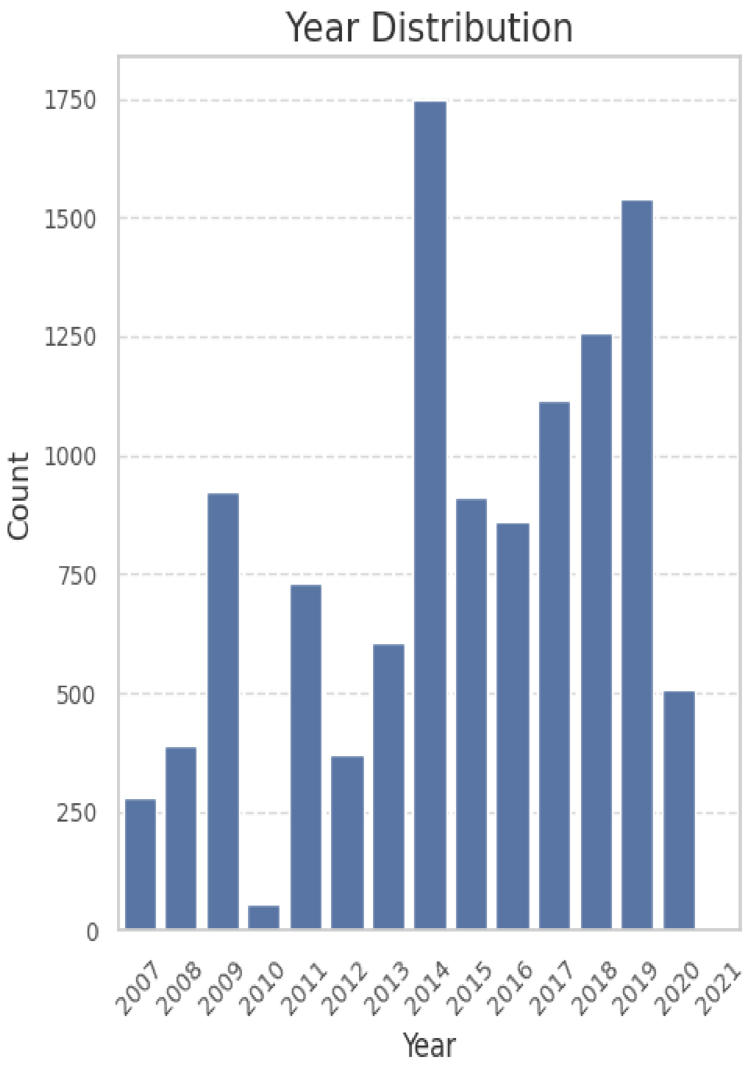
•Range: **-37.85° to 59.96°**, IQR: **25.78°–44.96°**.

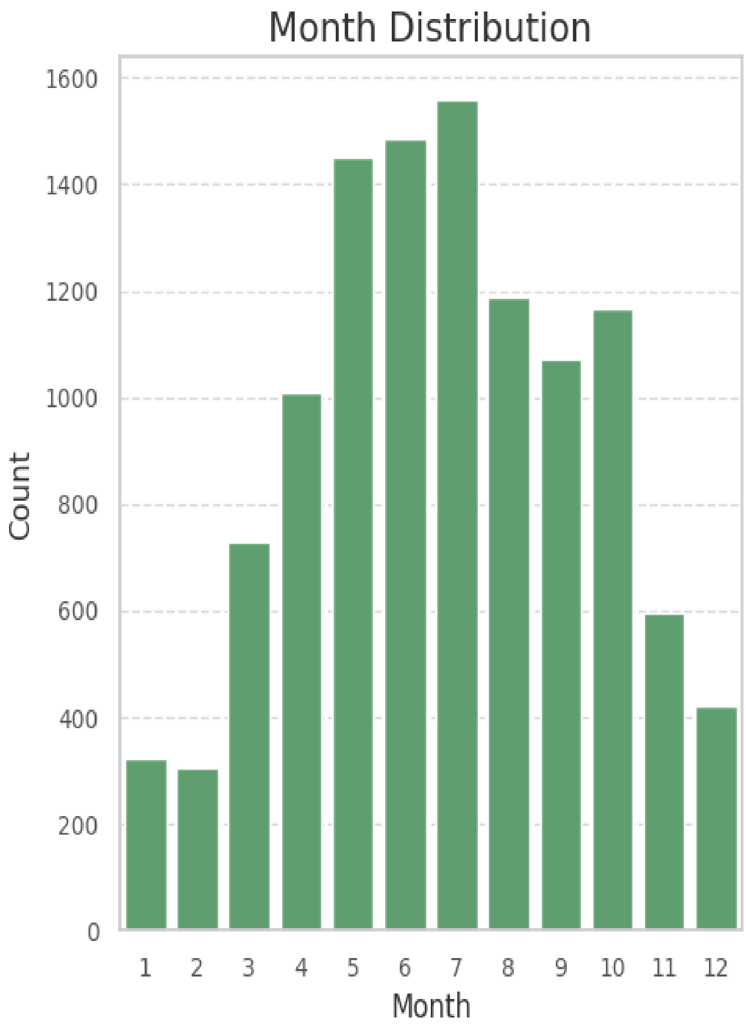
•Focuses on **mid-latitude regions**, spanning subtropical to temperate zones.

5. Longitude

•Range: **-118.27° to 145.02°**

•Most data points fall within **-80.22° to 10.70°**, indicating concentrations in **Americas, Europe, and parts of Africa or Asia**.



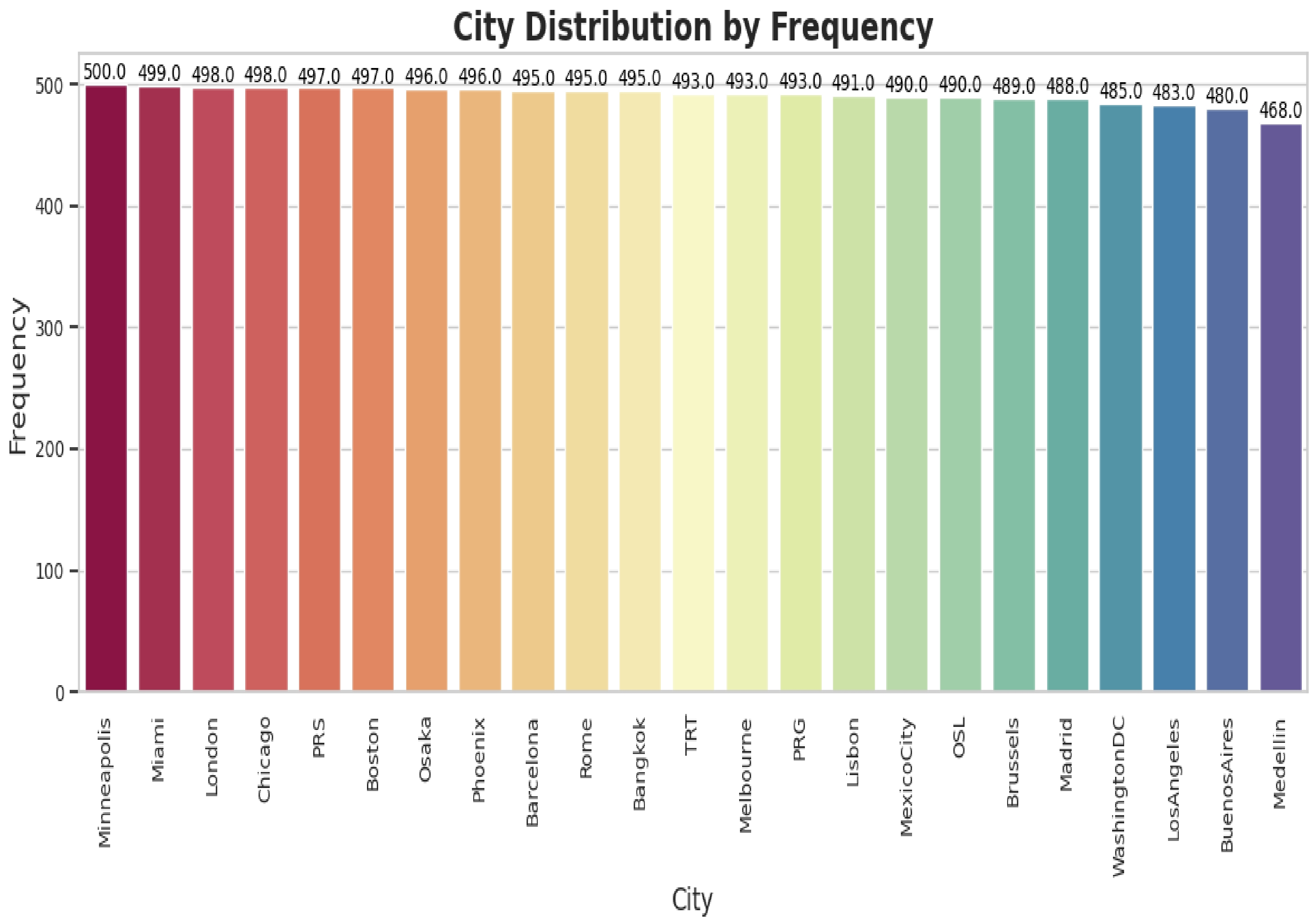




## Year Distribution

- **Key Trends**:
  - Significant data increase from **2014**, peaking that year.
  - Decline after 2014, lowest in **2021**.
  - Minimal data before **2007**.
- **Implications**:
  - Potential **bias** due to concentration around **2014–2018**.
  - Drop in **2021** may reflect **external constraints**.

## Month Distribution

- **Key Trends**:
  - Higher data in **May–September**, peaking in **June–July**.
  - Lower data in **December–February**.
- **Implications**:
  - Seasonal data variation likely due to **environmental factors**




1. **City Distribution**:
    1. Shows unique record counts per city after duplicate removal.
    2. Cities like **London**, **Miami**, and **Medellin** have frequencies near 500 but not exactly 500.
2. **City Variety**:
    1. Represents a wide range of cities, including **Osaka**, **Rome**, **Melbourne**, and **Mexico City**.
3. **Duplicate Removal**:
    1. Duplicates were removed, so no city reaches exactly 500.
4. **Color Coding**:
    1. Each city is assigned a unique color for easy identification.

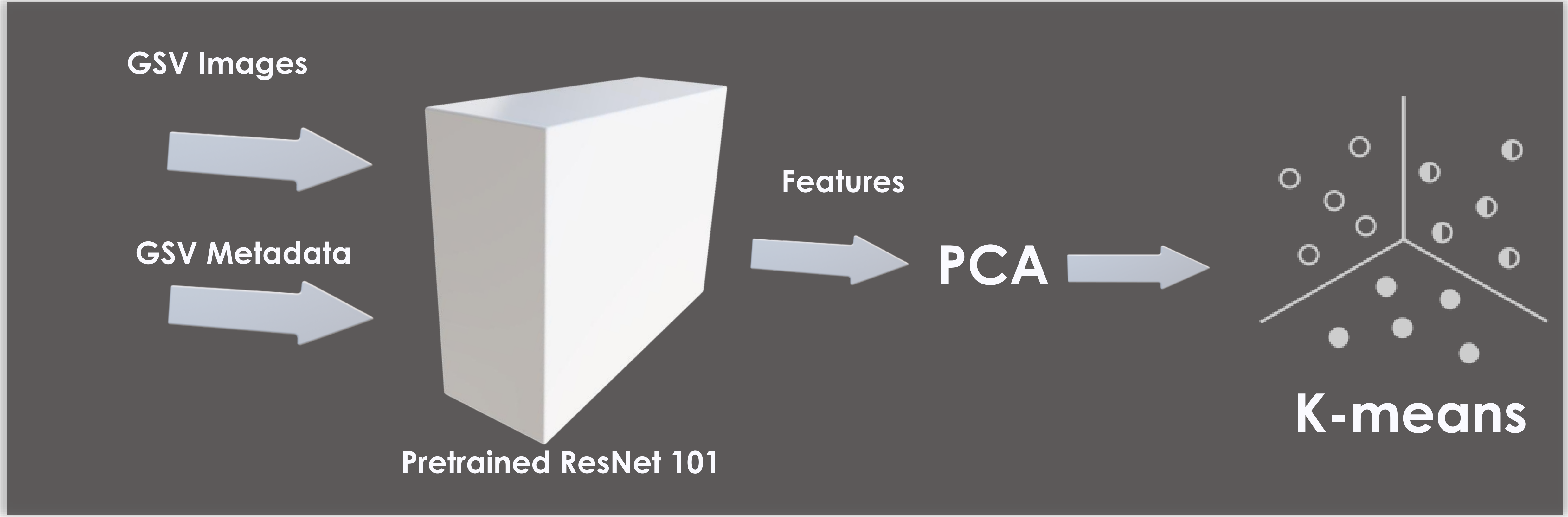
# Geolocation Visualisation

# Section C

Unsupervised Learning

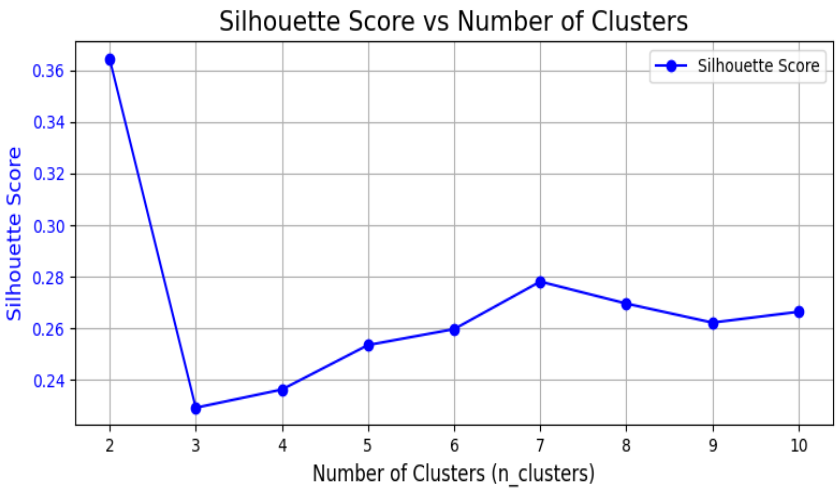
# Feature Extraction

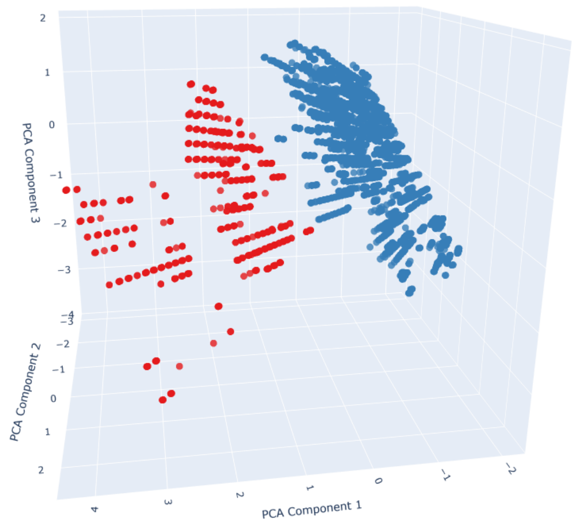
# Feature Extraction



# Methods for Clustering Evaluation and Visualization

# Silhouette Score





- **Definition:**

Measures how similar an object is to its own cluster compared to other clusters.

- **Range of Values:**
- **Close to 1:** Well-separated clusters.
- **Close to 0:** Overlapping clusters.
- **Negative values:** Incorrect clustering.
- **Usage:**

The number of clusters with the **highest silhouette score** indicates the **optimal clustering solution**.

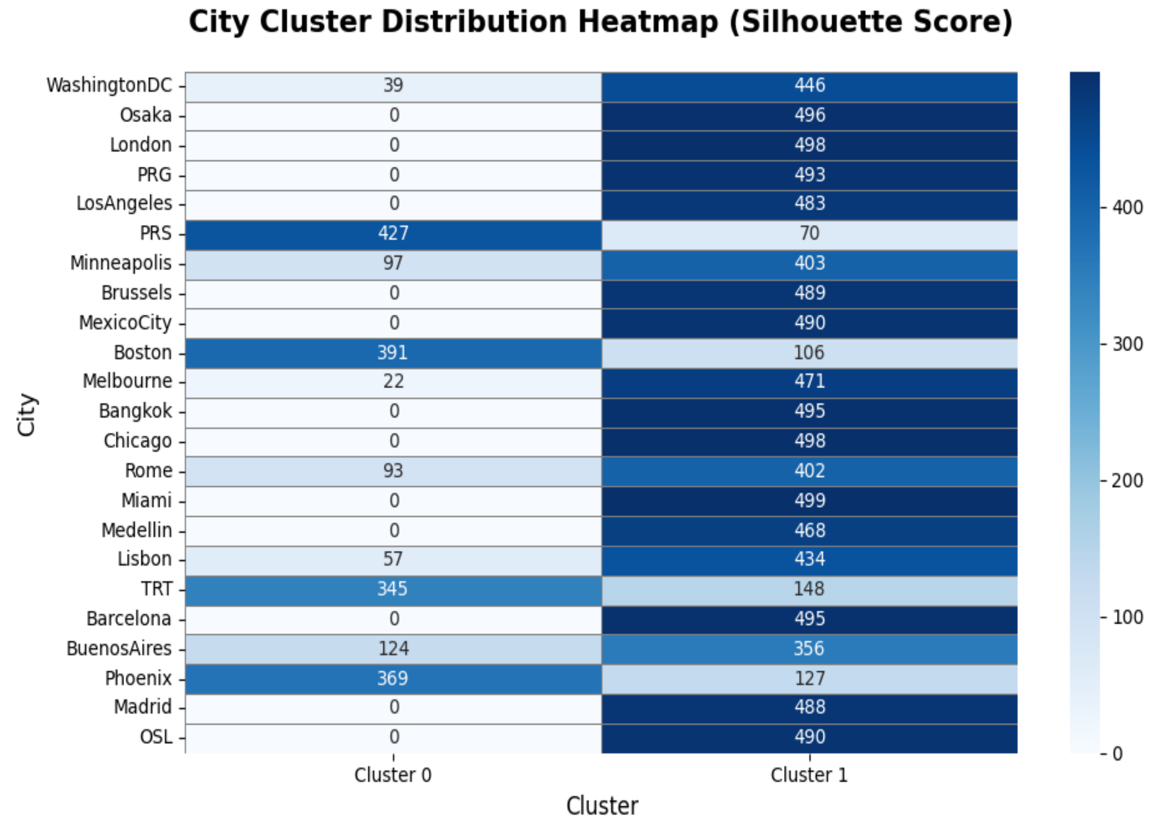
**Cluster 0: "Suburban Area"**

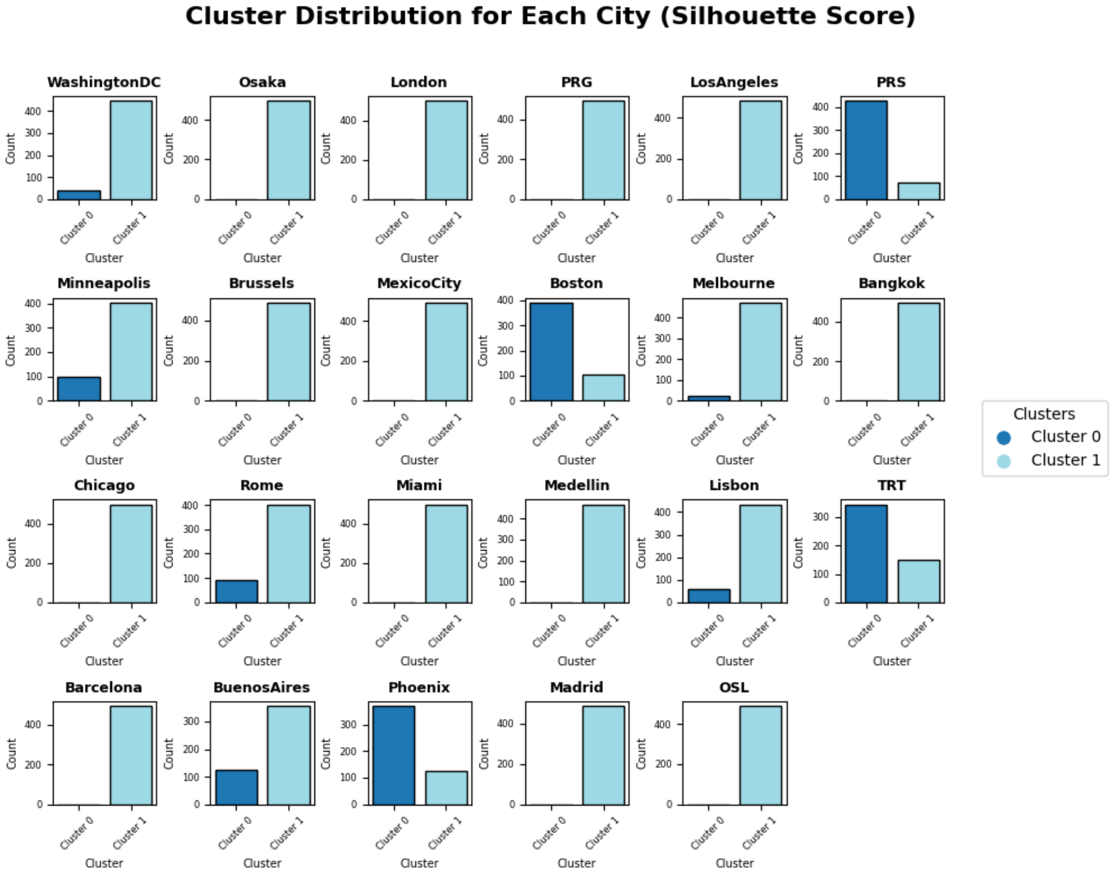
- **Cities:** PRS, Boston, TRT, Phoenix, Buenos Aires
- **Characteristics:** Wide streets, residential buildings, open spaces, low-rise buildings, less commercial activity.

**Cluster 1: "Urban Area"**

- **Cities:** Washington DC, London
- **Characteristics:** Narrow streets, commercial buildings, compact areas, high population density, less greenery.

# Silhouette Score




**Cities Grouped by Silhouette Score Method Clustering**

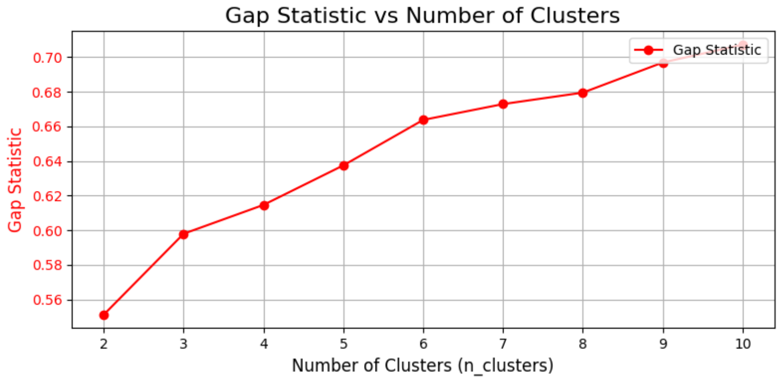
**Cluster 0:**

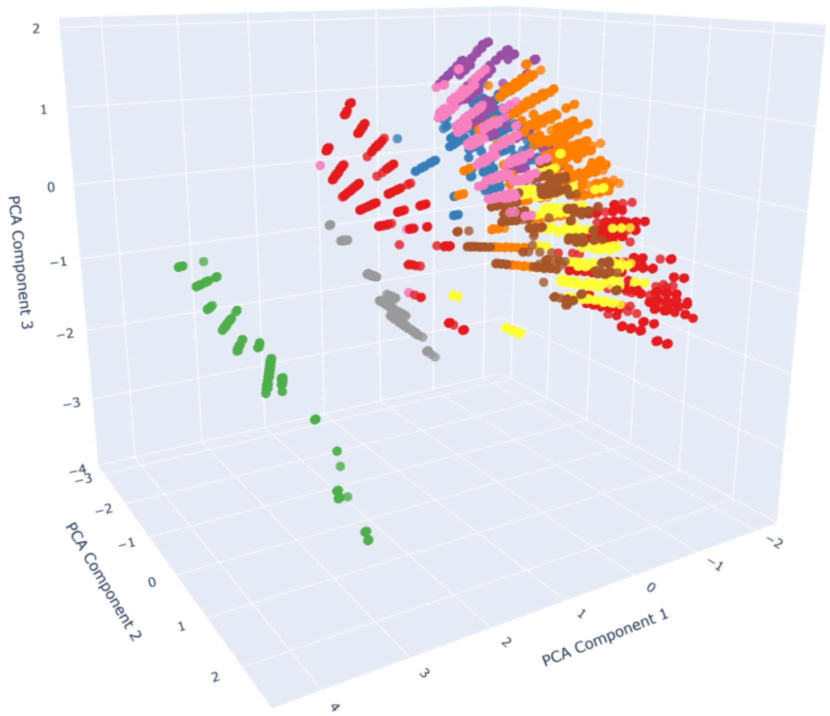
Cities most similar in cluster distribution: Phoenix, Boston, TRT, PRS.

**Cluster 1:**

Cities most similar in cluster distribution: Osaka, London, PRG, Los Angeles, Minneapolis, Brussels, Mexico City, Melbourne, Bangkok, Chicago, Lisbon, OSL.

# Gap Statistic



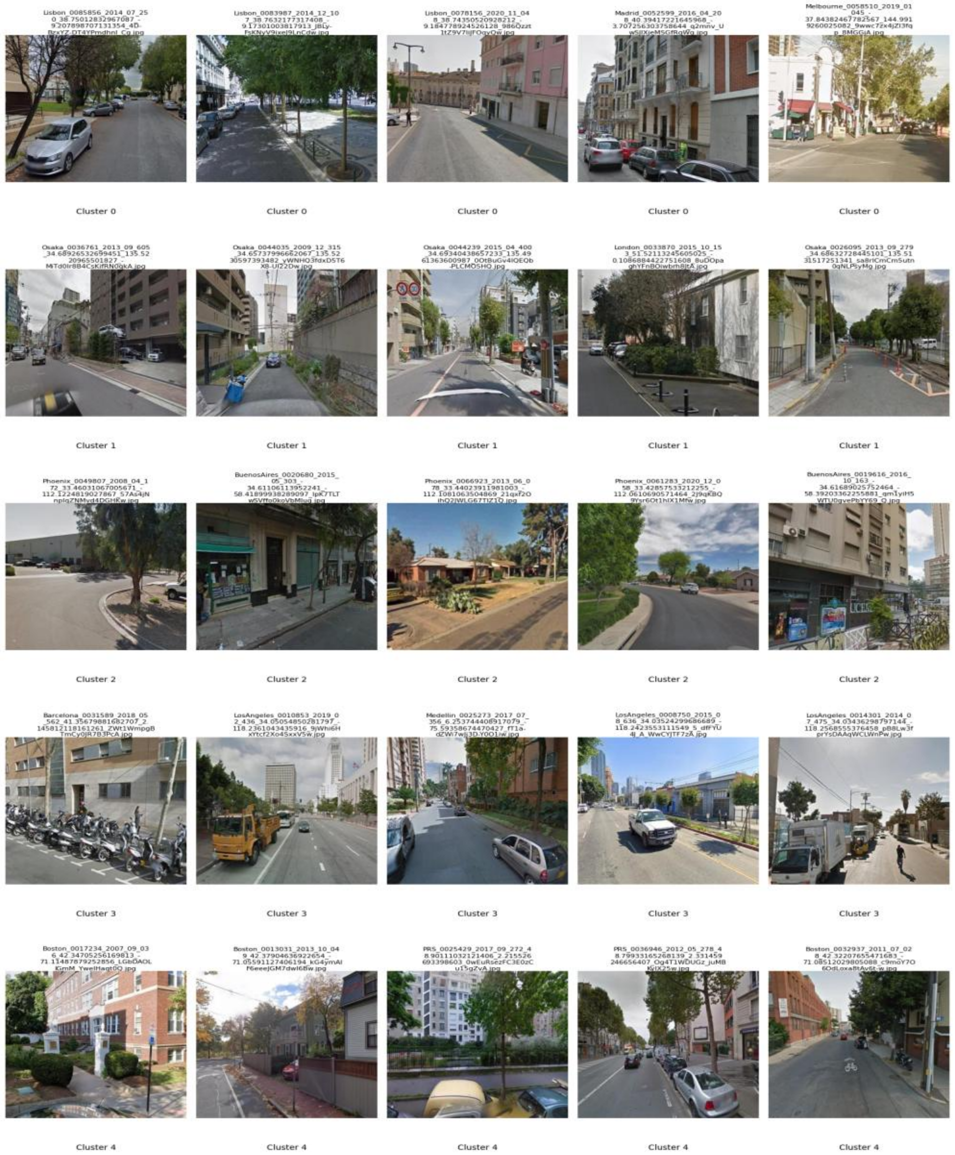

**Definition:**
Compares the total intra-cluster variation for different numbers of clusters with that expected under a **reference distribution**. This reference distribution is typically a random distribution of points, used to assess whether the observed clustering is significantly better than random.

**Range of Values:**

- **Higher Gap Statistic**: Indicates more distinct clustering.
- **Lower Gap Statistic**: Suggests less distinct clustering.

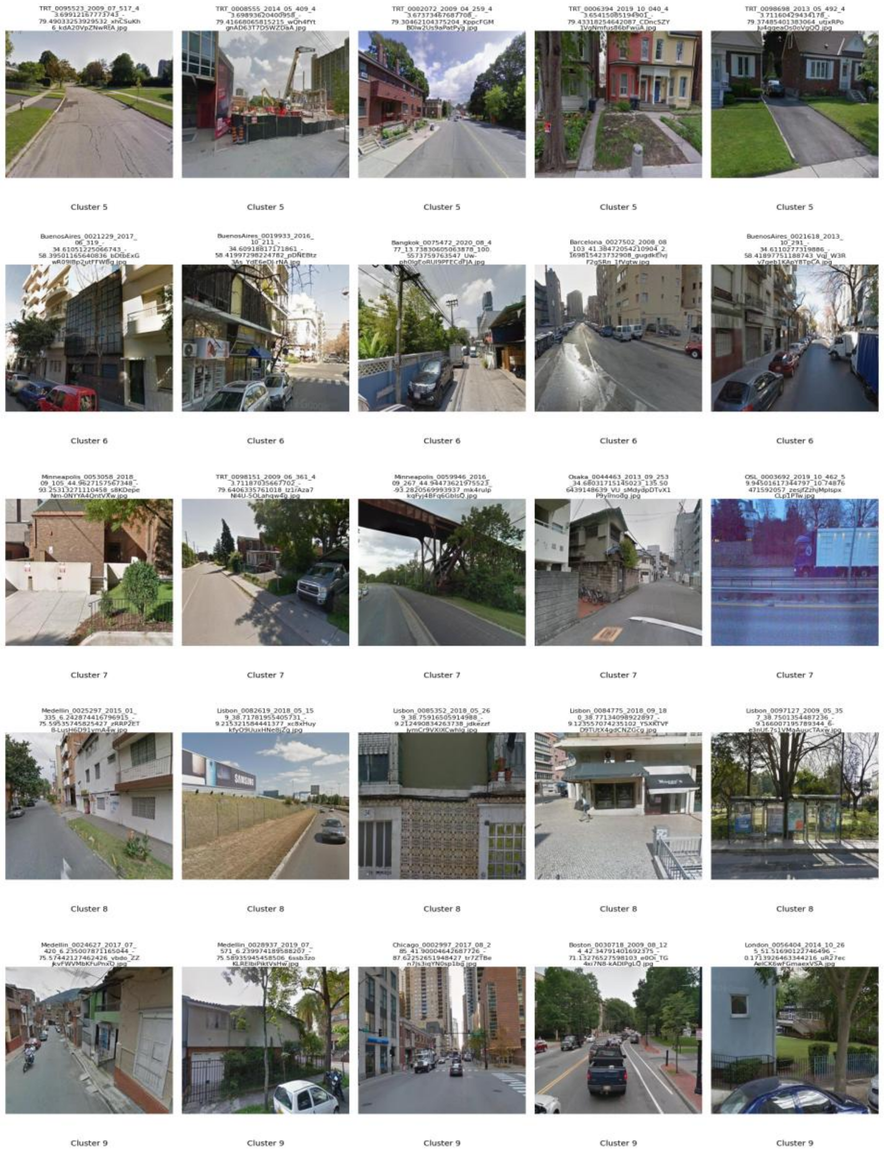
**Usage:**
The optimal number of clusters is where the **gap statistic is maximized**, providing a robust estimate for the number of clusters.

# Gap Statistic



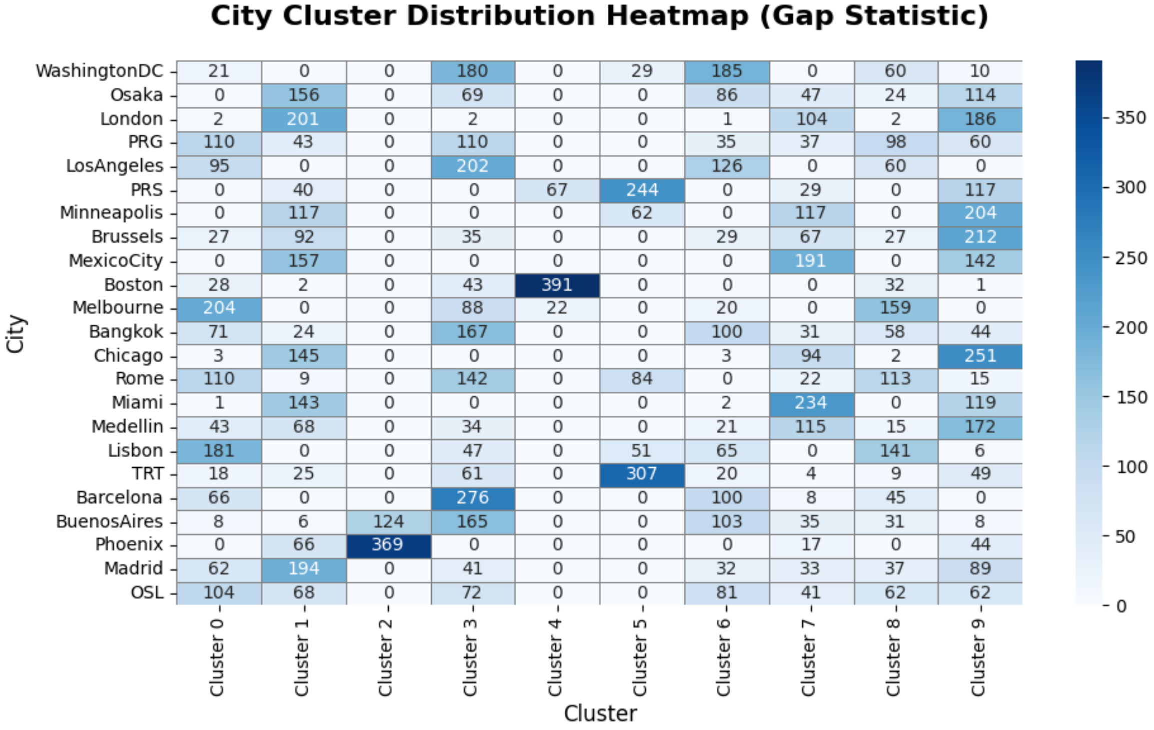

Cluster 0
Tree-lined urban streets, residential and commercial mix, moderate traffic.
**Feature:** Tree-lined streets, residential-commercial balance.
Cluster 1
Narrow streets, mixed buildings, pedestrian-friendly, few vehicles.
**Feature:** Narrow streets, defined lanes.
Cluster 2
Suburban streets, residential homes, green spaces, wide roads.
**Feature:** Single-story homes, greenery.
Cluster 3
Wide lanes, high traffic, commercial and residential mix.
**Feature:** Wide roads, busy urban setting.
Cluster 4
Residential streets, low-rise homes, green spaces, quiet.
**Feature:** Single-family homes, greenery.

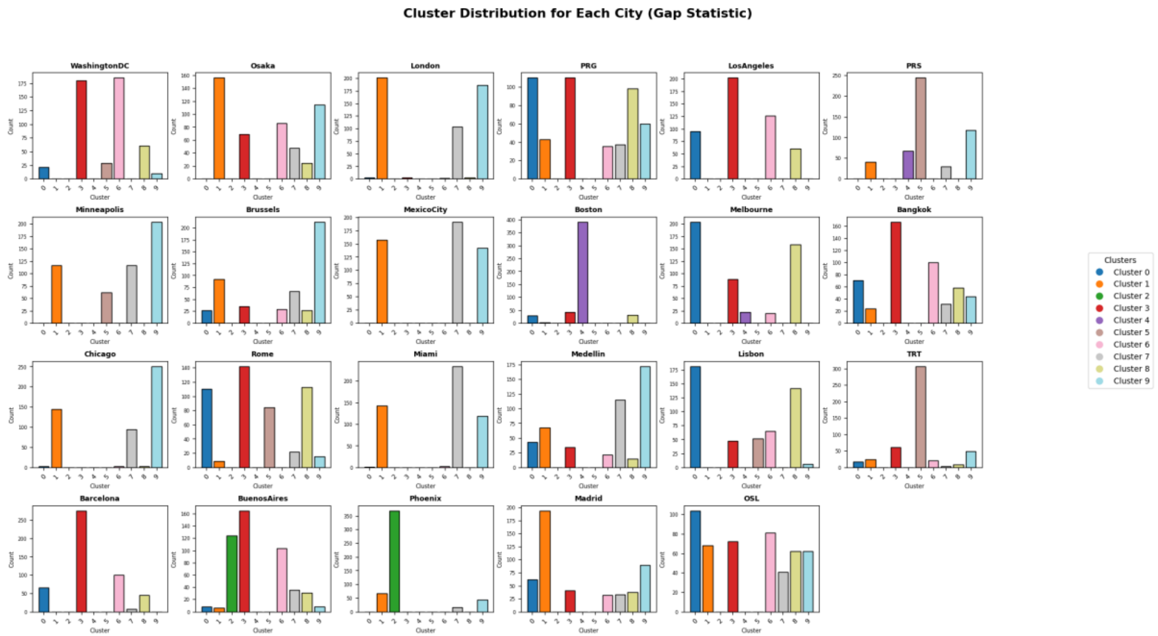
# Gap Statistic




**Cluster 5**
Suburban streets, homes under renovation, green spaces.
**Feature:** Renovating homes, clean streets.
**Cluster 6**
Commercial and residential mix, narrow roads, parked cars.
**Feature:** Narrow streets, active urban vibe.
**Cluster 7**
Wide highways, minimal traffic, residential and industrial mix.
**Feature:** Wide highways, open spaces.
**Cluster 8**
Urban streets, home facades, storefronts, billboards.
**Feature:** Home facades, commercial storefronts.
**Cluster 9**
Wide streets, active traffic, residential and urban mix.
**Feature:** Wide roads, active traffic.

# Gap Statistic




**Cities Grouped by Gap Statistic Method Clustering**

**Cluster 0:** Prague, Melbourne, Lisbon, Oslo

**Cluster 1:** Osaka, London, Mexico City, Madrid

**Cluster 2:** Phoenix

**Cluster 3:** Washington DC, Prague, Barcelona, Buenos Aires

**Cluster 4:** Boston

**Cluster 5:** Prague, TRT

**Cluster 6:** Oslo, Bangkok

**Cluster 7:** Mexico City, Miami, Medellin

**Cluster 8:** Melbourne, Lisbon, Oslo, Rome, Medellin, Brussels, London, Mexico City

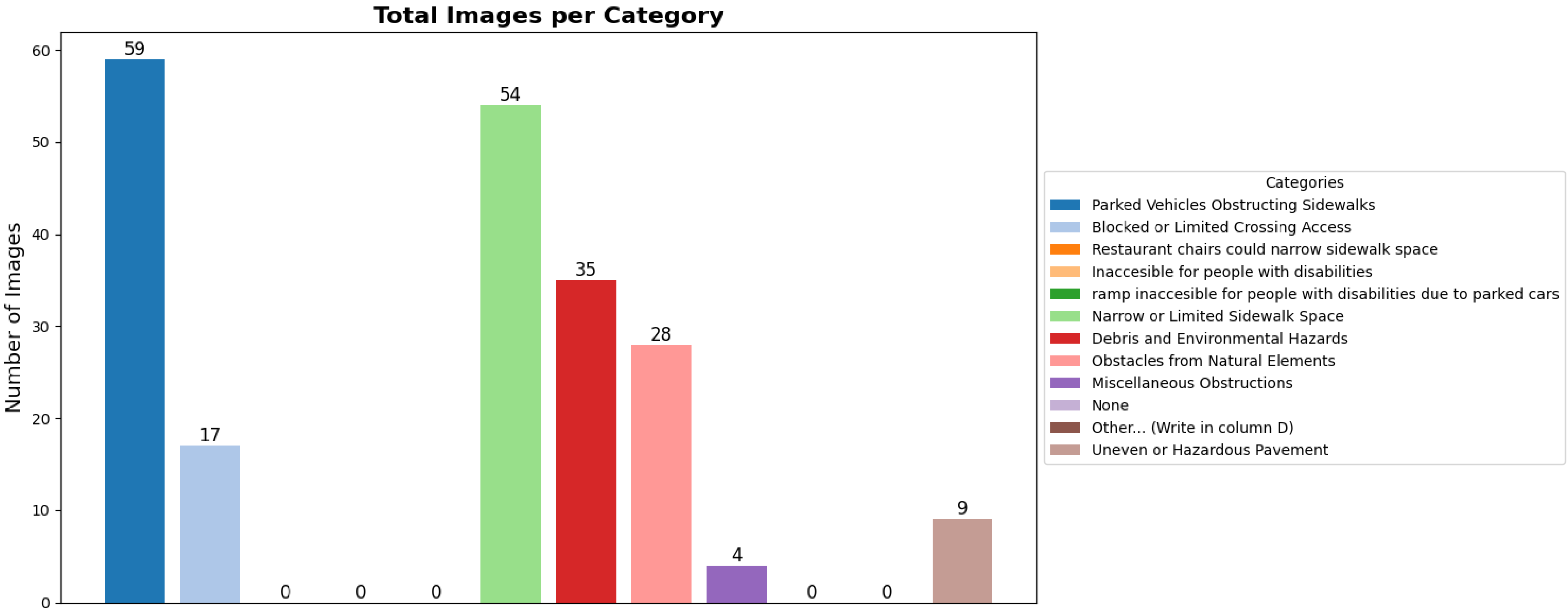
**Cluster 9:** Minneapolis, Medellin, Brussels

# Section D

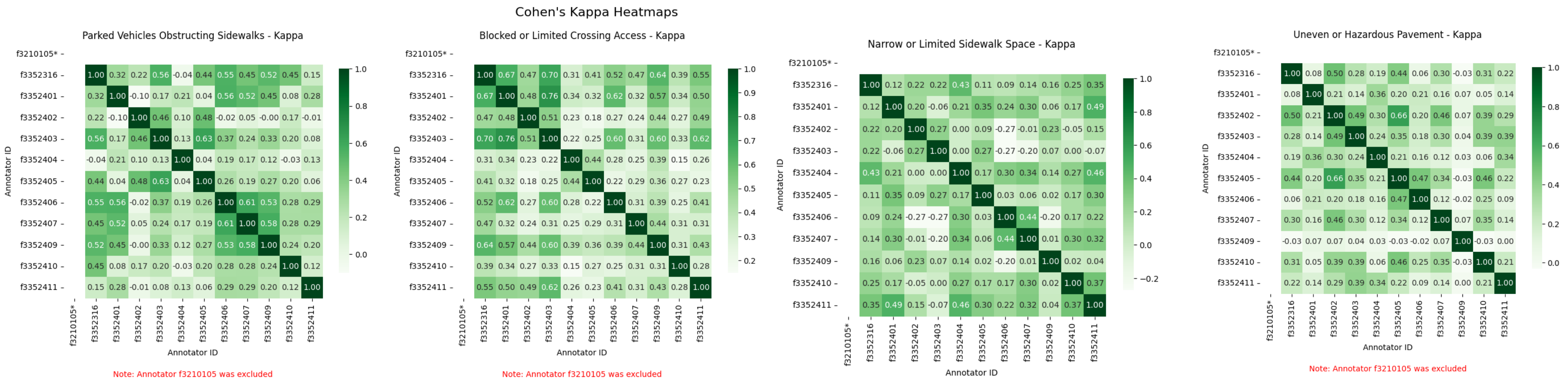
End Application

# End Application

# Distribution based on Ground Truth

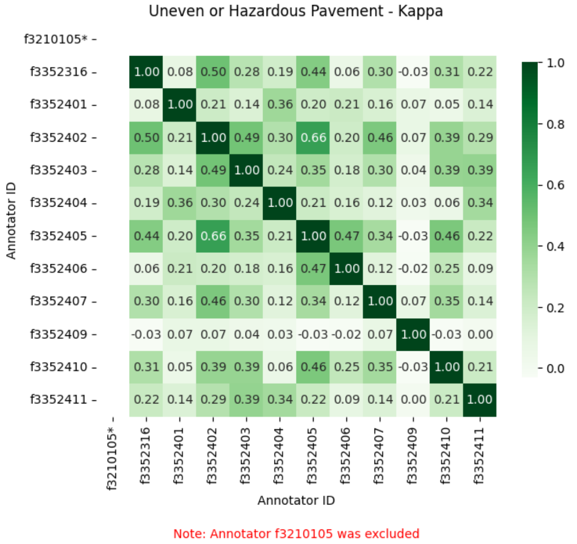


# Cohen’s Kappa

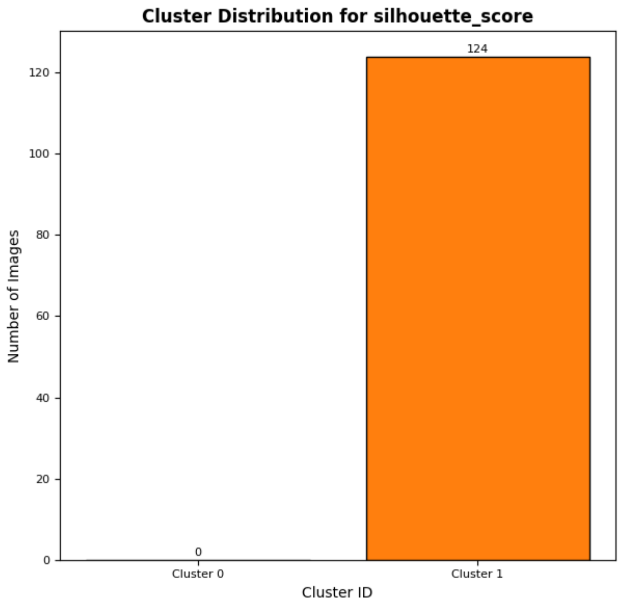


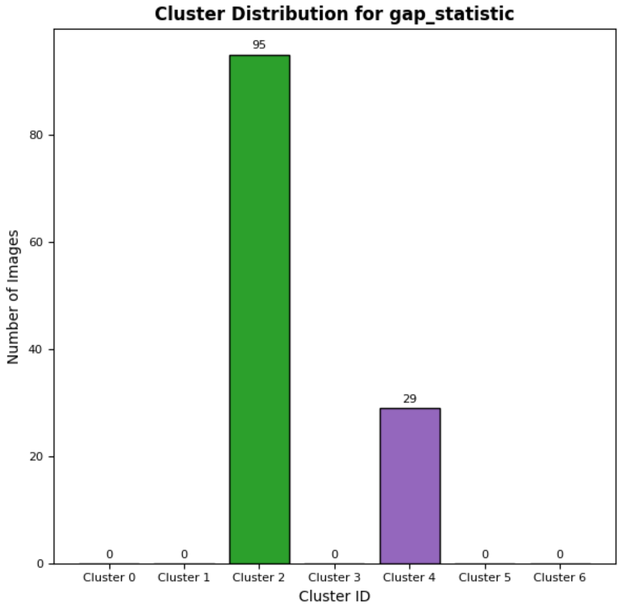

**Moderate Agreement among Annotators**

- Blocked or Limited Crossing Access
- Debris and Environmental Hazards
- Obstacles from Natural Elements
- Narrow or Limited Sidewalk Space
- Parked Vehicles Obstructing Sidewalks
- Uneven or Hazardous Pavement
- Miscellaneous Obstructions



# Athens Cluster Distribution Insights




## Silhouette Score

- Athens images are similar to **Cluster 1**: Osaka, London, Prague, Los Angeles, Oslo.
- Urban, dense cities with higher population, narrower streets, and less greenery.

## Gap Statistic Method

- Athens images belong to **Cluster 2**: Minneapolis, Brussels and **Cluster 4**: Barcelona, Buenos Aires.
  - **Cluster 2**: Wide streets, active traffic, residential and urban blend.
  - **Cluster 4**: Low-rise homes, green spaces, tree-lined streets.
  - Green spaces and parked cars may explain **Cluster 4** similarity.

# Q&A

# Thank You!