Assignments 2-3

Data Science

Evangelia P. Panourgia

Department of Informatics
Athens University of Economics and Business

March 31, 2025

Assignment 2: Exploratory Analysis

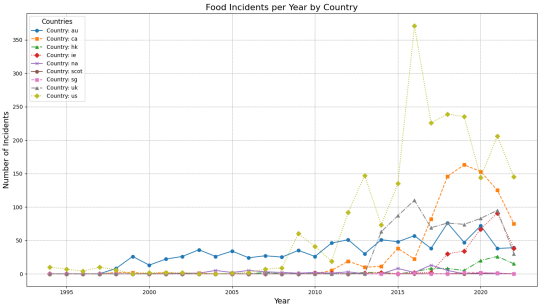


Figure: Food Incidents per Year per Country

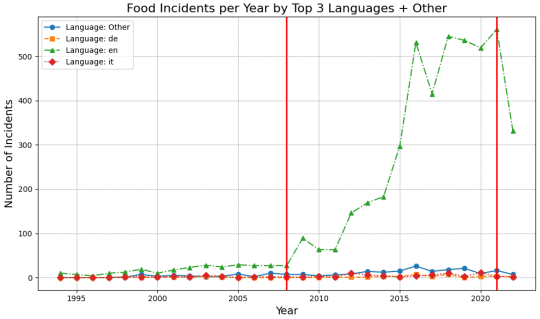
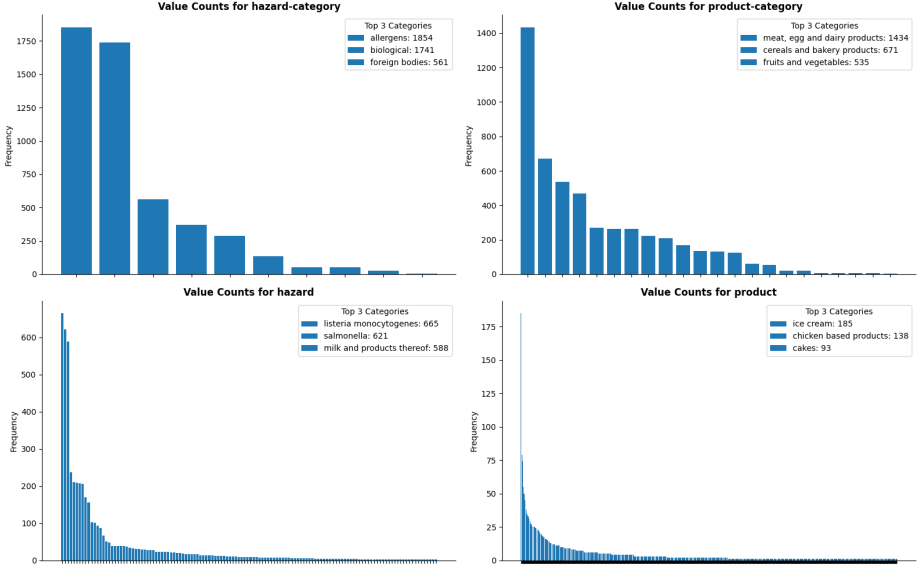


Figure: Food Incidents per Year by Top 3 Languages

Assignment 2: Imbalance of Y Labels



Assignment 2: Benchmark Analysis Based on Augmented Data

Table: Model Performance Comparison

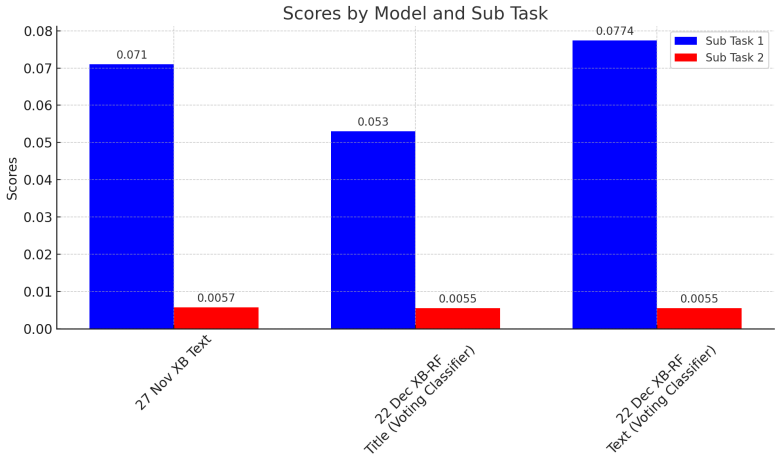
Model	Sub Task 1	Sub Task 2
Random Baseline	0.057	0.003
Majority Baseline	0.031	0.001
LogisticRegression Title	0.690	0.425
Random Forest Title	**0.760**	**0.721**
X-Boost Title	0.741	0.647
LogisticRegression Text	0.695	0.427
Random Forest Text	0.784	0.758
X-Boost Text	**0.814**	**0.759**
X-Boost Text (tuned)	**0.815**	**0.762**

Assignment 2: Benchmark Analysis Based on Initial Data

Table: Model Performance Comparison

Model	Sub Task 1	Sub Task 2
Random Baseline	0.051	0.002
Majority Baseline	0.039	0.002
LogisticRegression Title	0.39	0.13
Random Forest Title	0.50	0.32
X-Boost Title	0.54	0.31
Voting Classifier Title (XB+RF)	**0.56**	**0.34**
LogisticRegression Text	0.36	0.11
Random Forest Text	0.42	0.26
X-Boost Text	**0.51**	**0.33**
Voting Classifier Text (XB+RF)	0.49	0.33

Assignment 2: Competition Results

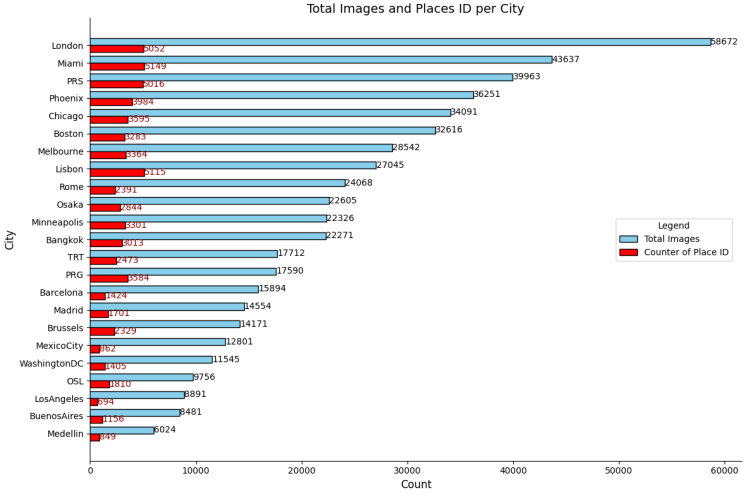


Assignment 3: Data Acquisition and Preprocessing - Challenges

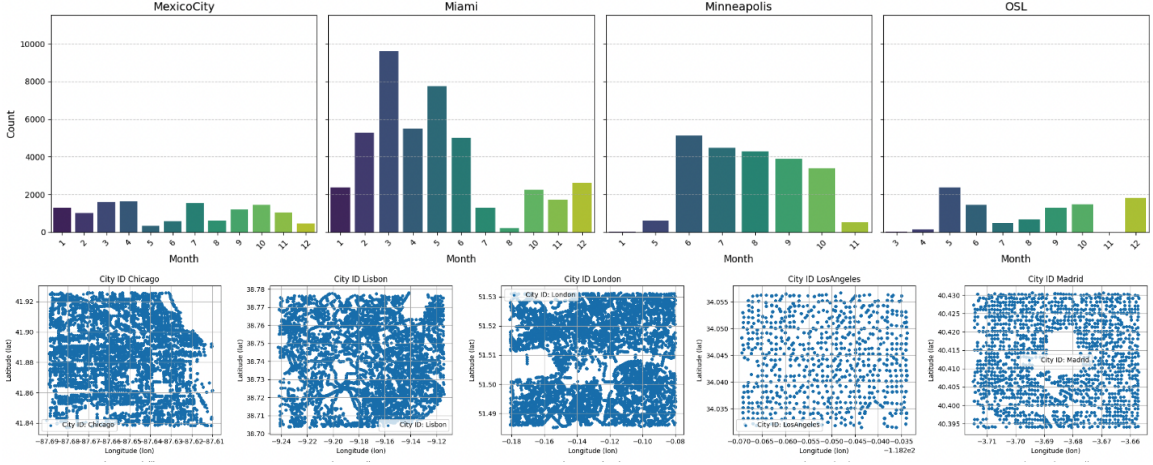
Table: Challenges and Solutions in Image Processing

#	**Challenge**	**Description**	**Solution**
1	Imbalance of images per city	Some cities, like London, have more images compared to others.	Equal number of random samples will be pulled from each city.
2	Resize of images	Some images may have better quality by default due to their original size and resolution.	All images will be resized to a standard dimension for consistency.
3	Hardware limitations memory analysis	Limitations in hardware memory can affect the processing of large image datasets.	We will limit our analysis to a sample of 100 images per city (2300 total number of images).

Assignment 3: Challenge 1 - Imbalance of Images per City



Assignment 3: Challenge 1 - Imbalance of Images per Month and PlacedID



Assignment 3: Challenge 2 - Image Size

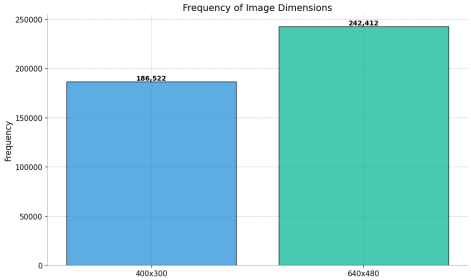


Figure: Distribution of Dimension (width x height) of Images

Assignment 3: Challenge 3 - Hard Ware Memory Limitations

```
 # Get the total physical memory (RAM) available on the system
total_memory = psutil.virtual_memory().total

# Assume a safety factor of 70% to avoid exhausting all system
    memory
memory_safety_factor = 0.7
available_memory = total_memory * memory_safety_factor

# Function to calculate memory usage per image (all images are
    colored)
def calculate_memory_usage(row):
    pixel_size = 3  # RGB images use 3 bytes per pixel
    return row['image_width'] * row['image_height'] * pixel_size
```

Assignment 3: Challenge 3 - Hard Ware Memory Limitations

```
# Add memory usage per image to the DataFrame (ensure safe
    assignment)
df_integrated = df_integrated.copy()  # Avoid SettingWithCopyWarning
df_integrated.loc[:, 'memory_usage_bytes'] = df_integrated.apply(
    calculate_memory_usage, axis=1)

# Calculate total memory usage and the maximum number of images
total_memory_usage = df_integrated['memory_usage_bytes'].sum()
average_memory_per_image = df_integrated['memory_usage_bytes'].mean
    ()
max_images_possible = int(available_memory /
    average_memory_per_image)
```

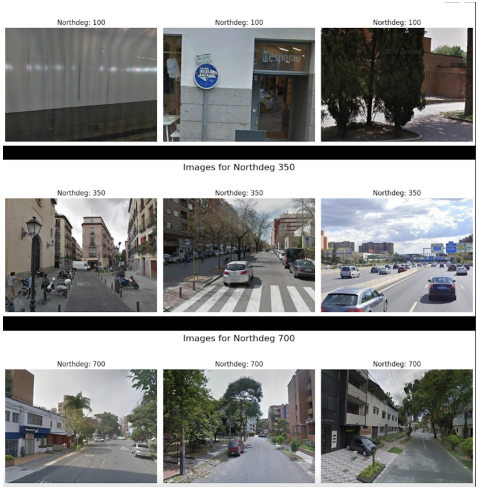
Assignment 3: Challenge 3 - Hardware Memory Limitations

Table: Memory Specifications and Utilization

Specification	Value
Total Physical Memory	8.00 GB
Available Memory for Clustering	5.60 GB
Total Memory Required for All Images	270.60 GB
Average Memory Usage per Image	0.65 MB
Maximum Number of Images that Can Be Handled	8876
Estimated Images per City	**385** (8876 / 23)

Important Note: Due to the features that we will create in the section on Exploratory Data Analysis (EDA), we selected **100** images per city. Some features, e.g., the array stemming from **ResNet**, contain many features. Despite applying PCA during training, we encountered memory problems.

Assignment 3: Feature - Observation - Northdeg



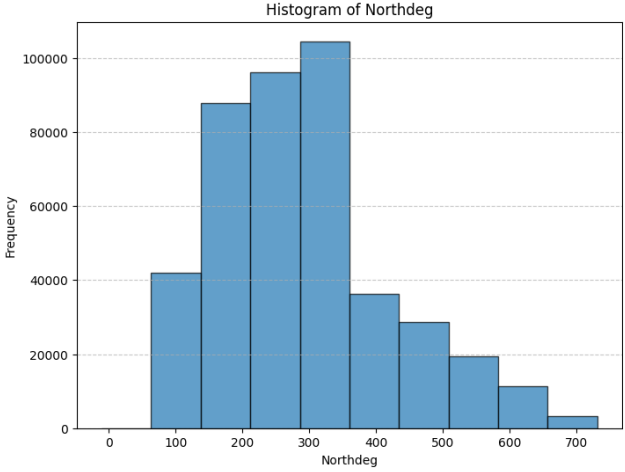


Figure: Understanding of feature Northdeg.

Assignment 3: Creation of Feature based on Preprocess Images

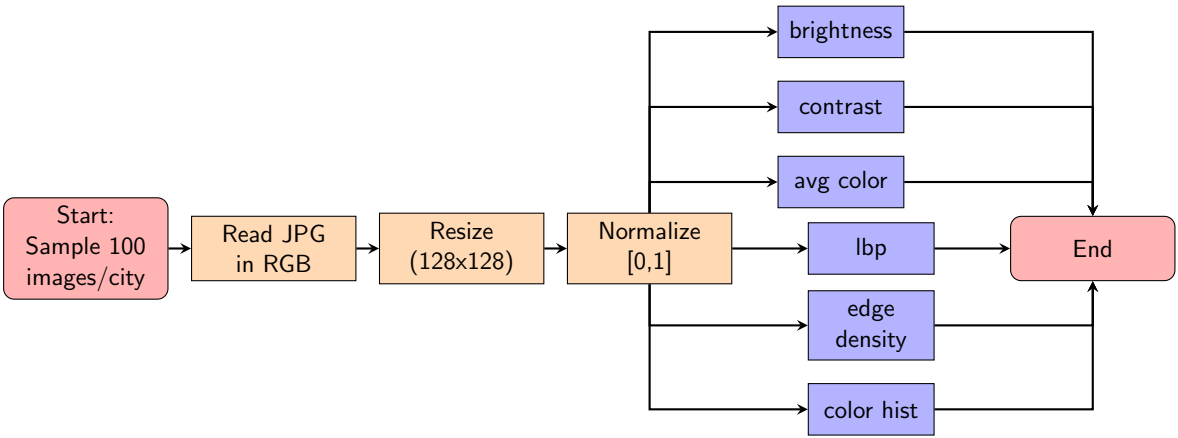


Figure: Pipeline for Image Processing with Parallel Feature Extraction

Assignment 3: EDA: Overall Uni-variate Analysis

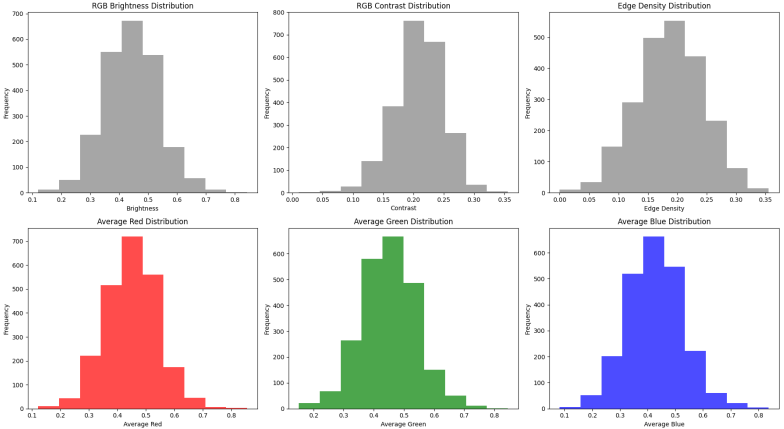
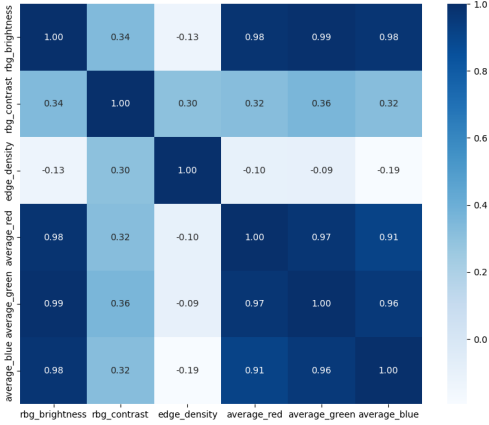


Figure: Distribution of Features

Assignment 3: EDA: Overall Bi-variate Analysis



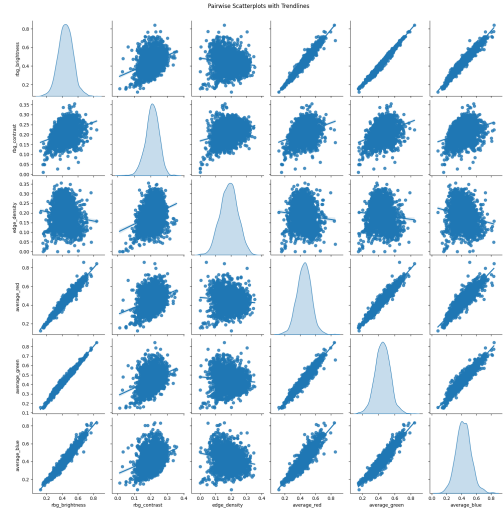


Figure: Bi variate Analysis based on Corelation and Pairwise Scatter plots.

Assignment 3: EDA: Per-city Analysis, Edge - Contrast - Brightness

Table: ANOVA Test Results for Image Features

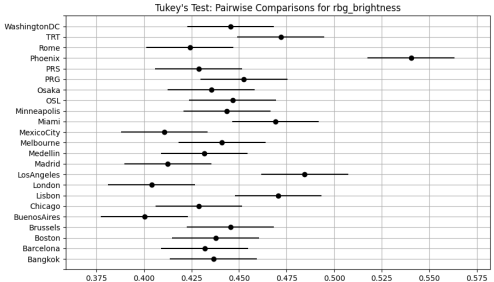
Feature	**F-statistic**	**P-value**
Brightness	11.45	0.0000
Contrast	2.01	0.0035
Edge Density	11.14	0.0000

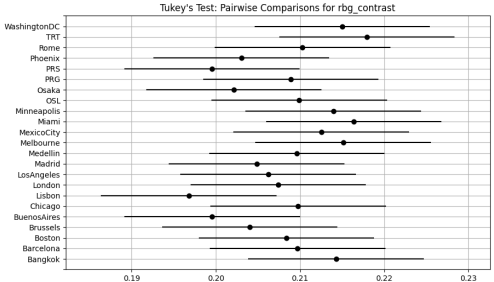
Summary:

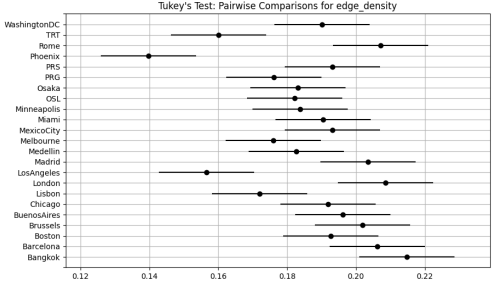
Brightness and **Edge Density** show strongly significant and pronounced differences, highlighting their importance in capturing city-specific features.

Contrast, while significant, exhibits less variability, suggesting it may play a secondary role in distinguishing cities.

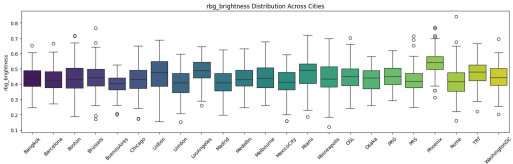
Assignment 3: EDA: Edge - Contrast - Brightness

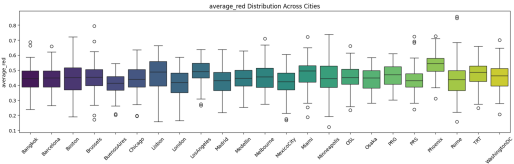


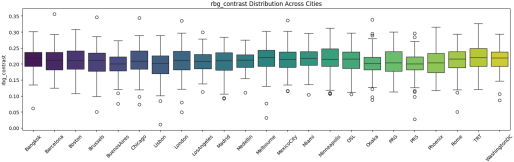


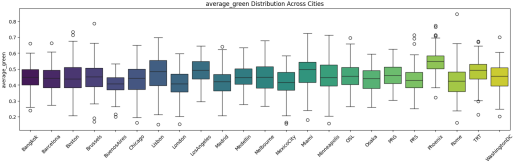


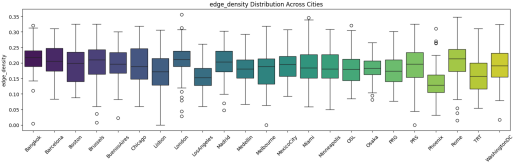
Assignment 3: EDA: Per-City Analysis, Edge - Contrast - Brightness

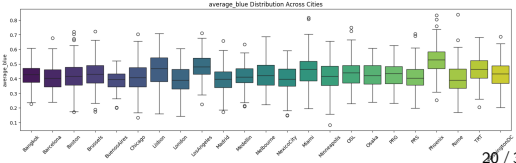












Cluster Analysis Strategy Overview

Feature Selection Approach: Begin with a comprehensive set including promising features like edge density, brightness, and contrast. Refine through empirical testing to find the most influential.

Initial Cluster Count: Start with k=3 clusters to explore broad groupings and facilitate distinct group detection within the dataset.

Cluster Evaluation: Use silhouette scores for assessing cluster quality and supplement with visual inspection of images near cluster centers for intuitive validation and leveraging PCA plots.

Focused Analysis: Direct efforts towards clusters with clear, interpretable patterns and statistical significance.

Integration of Approaches: Combine empirical and intuitive methods to ensure relevance and strategic alignment of the cluster analysis.

Empirical Evaluation K-Means for Separate Combinations

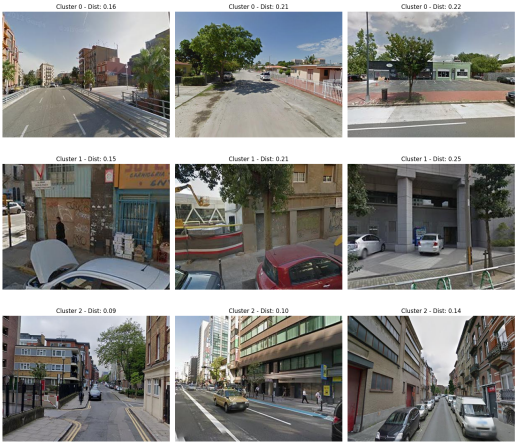
Table: Evaluation of K-means Runs

feature combination	**k**	**silhouette**	**centroid img**	**PCA**
brightness, contrast, edge density	4	0.252	66%	33%
brightness, contrast, edge density	3	0.272	100%	66%
brightness, contrast, edge density, RGB, color histogram	4	0.247	50%	66%
brightness, contrast, edge density, RGB, color histogram	3	0.279	33%	66%
all features	3	0.170	0%	0%

centroid img: The proportion of centroids where the images can be distinctly labeled, relative to the total number of clusters.

PCA: Examine the proportions of separation and overlap between clusters, as well as the density and shape within each cluster.

Top 3 Images for Each Cluster (Lowest Distance to Cluster Center)



Top 3 Images for Each Cluster (Lowest Distance to Cluster Center)

PCA Promising (3 features) vs Non-Promising Combination (all)

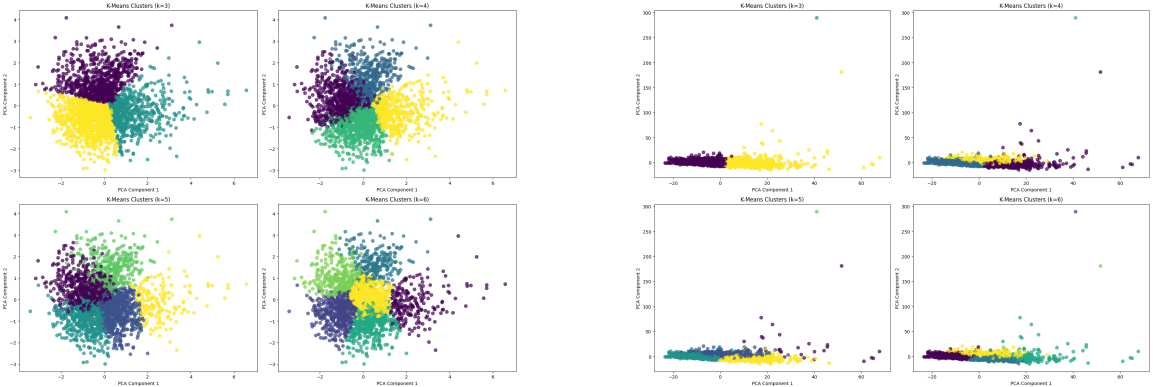


Figure: PCA Comparison Based on Number of Features.

Radar Plot for the promising combination of features

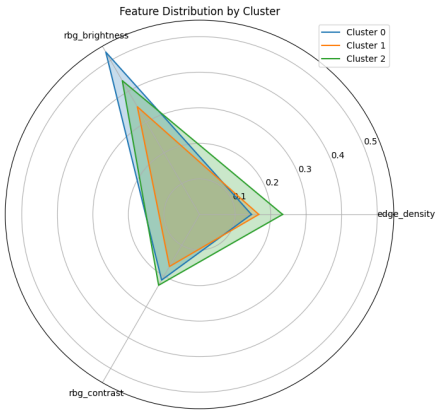


Figure: Radar Plot for each feature of the most "promising" combination.

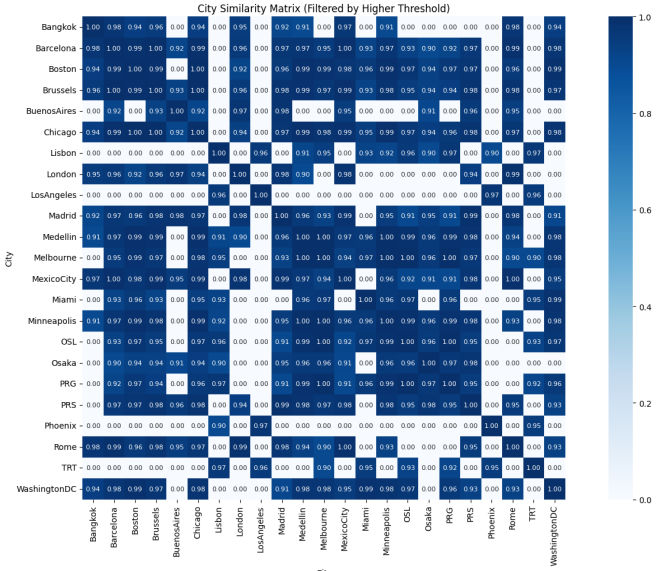
Cluster 0: Represents well-lit, open environments such as wide streets or bright outdoor areas. (High brightness, Moderate contrast, Low edge density)

Cluster 1: Captures urban settings with moderate activity and lighting, often shaded or dimly lit. (Lowest brightness, Moderate contrast, Moderate edge density)

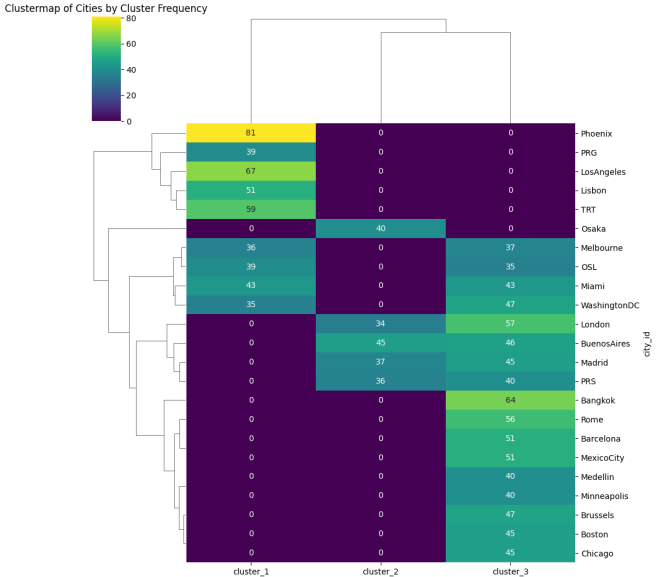
Cluster 2: Represents dense, urban environments with busy streets, characterized by high detail and sharp contrasts in lighting. (Moderate to High brightness, Highest contrast, High edge density)

Note due to Overal in Individual Features: Definitely there is room for improvement via trying other combination of features and re-evaluating.

Overall Similarity



Similarity Per Cluster, Zoom In



Future Work

1. Increase image resolution (e.g. 224 x 224) to improve feature extraction and utilize models optimized for higher resolutions.

2. Explore more combinations of already features, particularly those related to color, to refine image clustering, and try to apply PCA too before feeding to k-means (check variability of each).

3. Implement sub-clustering on ambiguous clusters to gain more nuanced insights.

4. Develop **additional image features** such as those from **edge detection** techniques to enhance clustering differentiation.

5. Improve evaluation methodology via using more visual techniques like **T-SNE** additionally for evaluating clusters (non-linear relations), manually **evaluating random images** based on distance from centroids consider outliers, too and **radar plots** for overlaps (**extend the array of evaluation with additional dimensions**).

6. Evaluate cluster **robustness** by using **Monte Carlo** bootstrap sampling to apply your clustering algorithm on various data samples and assess the stability of the resulting clusters.

Application - Annotators' Agreement - Reasons for Kappa

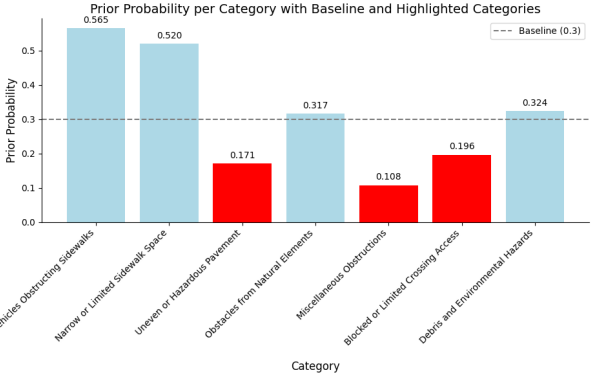
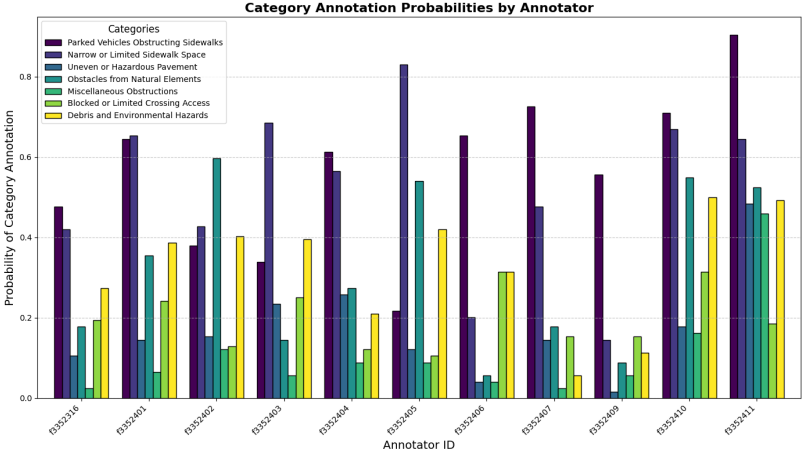


Figure: Randomness and Imbalance factors.

Application - Annotators' Agreement - Reasons for Kappa



Annotators's Agreement - Kappa

9% – **Miscellaneous Obstructions**: Percentage of coverage having more than 65% values in each cell.

9% – **Parked Vehicles Obstructing Sidewalks**: Percentage of coverage having more than 65% values in each cell.

9% – **Narrow or Limited Sidewalk Space**: Percentage of coverage having more than 65% values in each cell.

10% – **Uneven or Hazardous Pavement**: Percentage of coverage having more than 65% values in each cell.

12% – **Obstacles from Natural Elements**: Percentage of coverage having more than 65% values in each cell.

12% – **Blocked or Limited Crossing Access**: Percentage of coverage having more than 65% values in each cell.

16% – **Debris and Environmental Hazards**: Percentage of coverage having more than 65% values in each cell.

Reasons for Disagreement

Subjectivity and Perspective:

Each annotator's personal experiences or perceptions may influence their interpretation of images, causing variability.

Assumptions due to image zoom or focus can lead to inconsistent labels.

Complexity of Multiple Categories:

Annotating multiple categories in a single image can be challenging and lead to oversights.

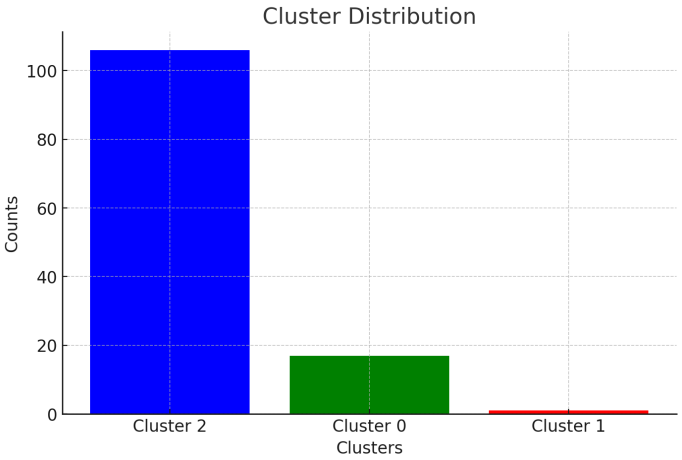
A focused, category-specific review process could improve accuracy.

Need for a Strategic Methodology:

Clear, structured guidelines could enhance consistency and accuracy.

A systematic approach to image review and labeling is recommended.

Athens' Images Distribution per Clusters

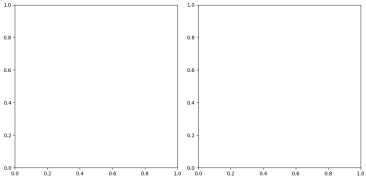


Cluster 0: Open Space

Cluster 1 Street, Cars, Walls

Cluster 2: City Blocks

Close Clustered Images of Athens



Unseen Images

Different Compositions

Over fitted trained Cluster