

Perceived Touristic Attractiveness in the Porto Metropolitan Area Based on Google Places Reviews

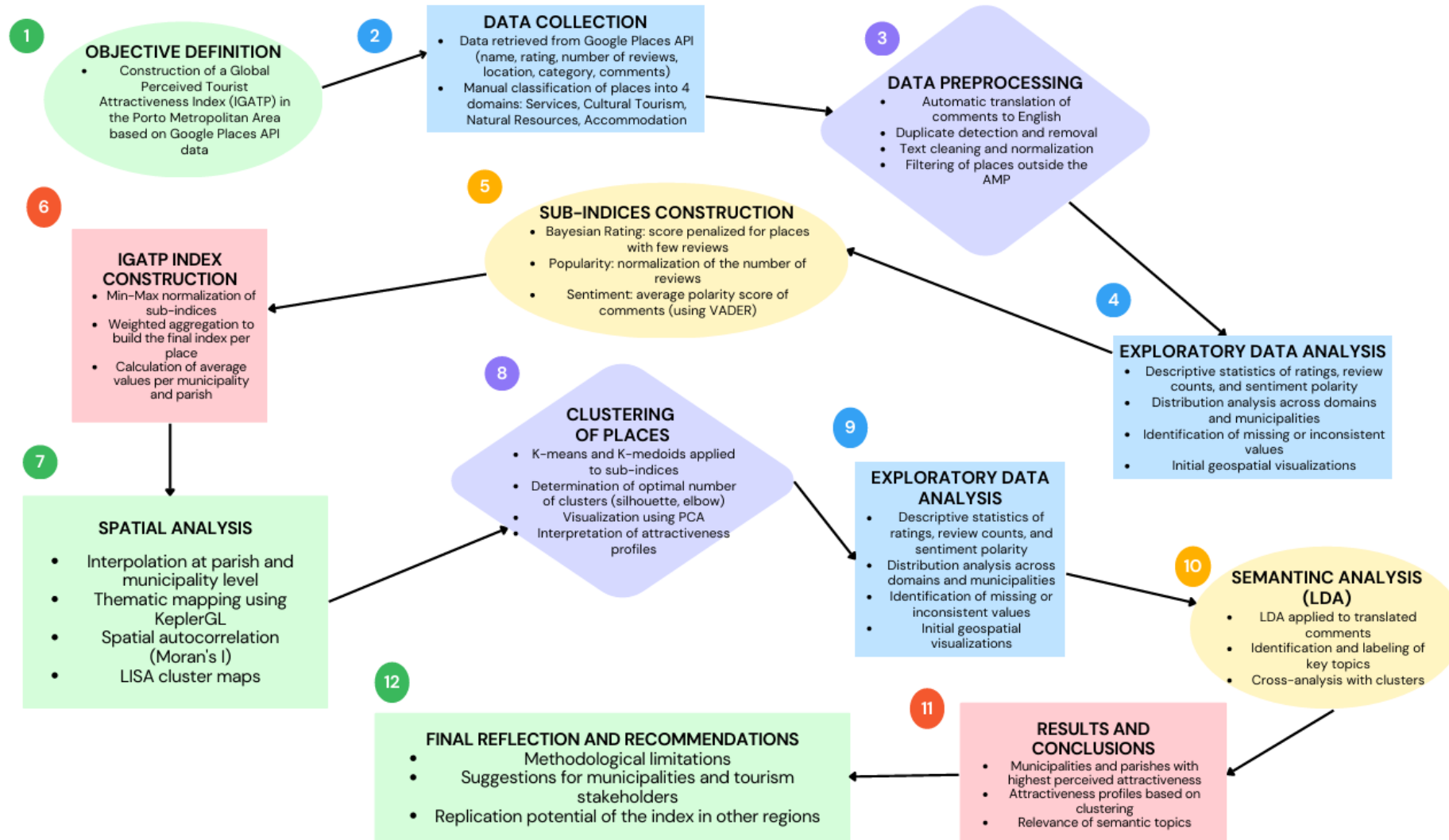
Authors:

Beatriz Santos, Bruno Rocha, Joana Guerreiro

Main Objective

- To build a **Global Index of Perceived Touristic Attractiveness (IGPTA)** using user ratings and reviews.
- To explore spatial patterns of perceived attractiveness in the Porto Metropolitan Area.

Flowchart



What is Touristic Attractiveness?



Definition (project-specific):

"Touristic attractiveness refers to the degree to which a location captures visitors' interest and preference, based on online perception, shared experiences, and qualitative evaluation."




Note:

- The index focuses on **perceived attractiveness**, not actual tourist flows.

Data Overview

 **Source:** Google Places API

 **Study Area:** 17 municipalities in the Porto Metropolitan Area (AMP)

 **Datasets:**

- google_places_AMP_with_coordinates.csv: place information + average rating
- comments_google_maps_AMP.csv: **up to 5 reviews per place**

Data Overview

How Google selects the 5 reviews shown via the Places API:

Although we did not find official documentation from Google explicitly detailing how the 5 reviews returned by the Places API are selected, we believe that the selection is likely based on internal criteria such as overall relevance (i.e., how useful the review is to other users), engagement (e.g., number of likes), the reviewer's profile (such as being an active Local Guide), recency (with newer reviews more likely to appear), and language preferences (according to the request's language or regional settings). In our view, Google seems to aim for a representative sample of opinions, rather than showing only the most positive or the most negative reviews.

Data Overview



Variable Dictionary – Ratings (google_places_AMP_with_coordinates.csv)

Variable	Description
City	Municipality where the point of interest is located
Category	Category assigned to the place (restaurant, museum, hotel, bar, tourist attraction, cafe, church, park, natural feature, viewpoint, trail, lodging)
Name	Name of the establishment or point of interest
Rating	Average rating given by users for the place
Address	Full address of the place
Types	List of categories assigned by the Google API
Latitude	Geographic latitude coordinate of the place
Longitude	Geographic longitude coordinate of the place
Total Reviews	Total number of reviews received by the place



Variable Dictionary – Comments (comments_google_maps_AMP.csv)

Variable	Description
City	Municipality where the point of interest is located
Category	Category assigned to the place (e.g., restaurant, hotel, park)
Place Name	Name of the establishment or point of interest
Author	Name of the user who wrote the comment
Text	Original text of the comment published by the user
Date	Relative date of the comment (e.g., 'a year ago')
Rating	Rating assigned in the comment (1 to 5 stars)

Database Structure

Ratings (google_places_AMP_with_coordinates.csv)

```
avaliacoes.head()
```

✓ 0.0s

	Cidade	Categoria	Nome	Rating	Endereço	Tipos	Latitude	Longitude	Total_Reviews
0	Arouca	restaurant	Tasquinha da Quinta	4.6	R. 1º de Maio 3, 4540-121 Arouca, Portugal	restaurant, food, point_of_interest, establish...	40.929109	-8.245191	2154.0
1	Arouca	restaurant	A Assembleia	4.5	Tv. da Ribeira 11, 4540-102 Arouca, Portugal	restaurant, bar, food, point_of_interest, esta...	40.928766	-8.247588	1788.0
2	Arouca	restaurant	Parlamento	4.6	Tv. da Ribeira 2, 4540-148 Arouca, Portugal	restaurant, food, point_of_interest, establish...	40.929011	-8.247392	2469.0
3	Arouca	restaurant	Casa Testinha	4.5	R. 1º de Maio 4, 4540-113 Arouca, Portugal	restaurant, food, point_of_interest, establish...	40.928876	-8.245147	623.0
4	Arouca	restaurant	Pedestre 142	4.4	R. Dr. Ângelo Miranda 142 RC, 4540-140 Arouca,...	restaurant, food, point_of_interest, establish...	40.930185	-8.253210	1047.0

Comments (comments_google_maps_AMP.csv)

```
comentarios.head()
```

✓ 0.2s

	Cidade	Categoria	Nome_Local	Autor	Texto	Data	Rating
0	Arouca	restaurante	Tasquinha da Quinta	IC	I came here as part of a tour, and this restau...	10 months ago	5
1	Arouca	restaurante	Tasquinha da Quinta	Preetam Nath	Has to be the most delicious veal I've had in ...	a year ago	5
2	Arouca	restaurante	Tasquinha da Quinta	Roya MJ	Came here as part of a tour and very much appr...	a year ago	5
3	Arouca	restaurante	Tasquinha da Quinta	Jonathan lugo	Very good food and an excellent place to eat! ...	a year ago	5
4	Arouca	restaurante	Tasquinha da Quinta	Benjamim Nande	Great place for a good typical Portuguese food...	4 months ago	5

Pre Processing - Ratings (google_places_AMP_with_coordinates.csv)

- Delete rows where rating value is null
- Replace null values with 0 in the “Total_Reviews” column
- Grouped places by thematic category (each place was classified into one of four categories): **Services, Cultural Tourism, Natural Resources, or Lodging**



based on its Google Places types (**Services:** restaurants, cafés, bars | **Cultural Tourism:** museums, attractions, churches | **Natural Resources:** parks, viewpoints, trails | **Lodging:** hotels, informal lodging)



💡 This categorization is used for map filters and exploratory analysis, not for the composite index.

Pre Processing - Ratings (google_places_AMP_with_coordinates.csv)

- A unique identifier was created for each combination of place name and address, in order to ensure the distinction between truly different establishments and to avoid double-counting the same location.

```
endereco_repetido = avaliacoes["Endereço"][[avaliacoes["Endereço"].duplicated()].iloc[0]]  
avaliacoes[avaliacoes["Endereço"] == endereco_repetido]
```

✓ 0.0s

	Cidade	Categoria	Nome	Rating	Endereço	Típos	Latitude	Longitude	Total_Reviews
7	Arouca	restaurant	Café Arouquense	4.3	Av. 25 de Abril, 4540-102 Arouca, Portugal	cafe, restaurant, food, point_of_interest, est...	40.928469	-8.245599	901.0
20	Arouca	restaurant	Sabores da serra	4.1	Av. 25 de Abril, 4540-102 Arouca, Portugal	restaurant, food, point_of_interest, establish...	40.928561	-8.244947	16.0
94	Arouca	bar	BOM COPO - restaurante & bar	4.8	Av. 25 de Abril, 4540-102 Arouca, Portugal	restaurant, food, point_of_interest, establish...	40.928572	-8.244460	11.0
119	Arouca	cafe	Café Arouquense	4.3	Av. 25 de Abril, 4540-102 Arouca, Portugal	cafe, restaurant, food, point_of_interest, est...	40.928469	-8.245599	901.0
133	Arouca	cafe	Bakery Village II	4.4	Av. 25 de Abril, 4540-102 Arouca, Portugal	cafe, store, food, point_of_interest, establis...	40.927949	-8.251804	378.0
151	Arouca	cafe	Sabores da serra	4.1	Av. 25 de Abril, 4540-102 Arouca, Portugal	restaurant, food, point_of_interest, establish...	40.928561	-8.244947	16.0

Pre Processing - Ratings (google_places_AMP_with_coordinates.csv)

- New variable: **Nearby_Similar_Places**

↪ Counts places of the same category within 100 meters (using geodesic distance)



However, due to the structure of the dataset—where the same location may be assigned to multiple categories (e.g., restaurant, café, bar)—multiple entries may exist for a single place. To avoid duplication and ensure a representative measure of thematic competition, the variable was aggregated using the maximum value recorded per place (ID). This decision aims to capture the most competitive context a location may face. For example, if a place is classified simultaneously as a restaurant and a bar, and there are 5 restaurants but only 1 bar nearby, the variable should reflect the value 5, acknowledging that the location exists in a highly competitive thematic environment, rather than smoothing it to an average of 3.



This variable is used to contextualize popularity — distinguishing between places that are popular despite nearby alternatives and those that benefit from being the only option. It was not included in the Global Tourist Attractiveness Index because it does not directly measure the attractiveness of a location itself, but rather its surrounding competitive environment. This variable enriches the exploratory analysis and interactive visualizations by helping to identify clustering patterns and contextualize tourist behavior, and is also displayed in the popup when a user clicks on a point of interest.

Data Overview



Variable Dictionary – Ratings (ratings_clean.csv)

Variable	Description
City	Municipality where the point of interest is located
Category	Category assigned to the place (restaurant, museum, hotel, bar, tourist attraction, cafe, church, park, natural feature, viewpoint, trail, lodging)
Name	Name of the establishment or point of interest
Rating	Average rating given by users for the place
Address	Full address of the place
Types	List of categories assigned by the Google API
Latitude	Geographic latitude coordinate of the place
Longitude	Geographic longitude coordinate of the place
Total Reviews	Total number of reviews received by the place
Thematic Group	Thematic group in which the place was classified (e.g., services, cultural tourism)
Similar Places Nearby	Number of places in the same category within 100 meters radius
Is Positive	Binary variable based on polarity (1 if positive, 0 otherwise)

Database Structure

Ratings (ratings_clean.csv)

```
avaliacoes.head(5)
```

✓ 3.3s Python

	Cidade	Categoria	Nome	Rating	Endereço	Tipos	Latitude	Longitude	Total_Reviews	id_unico	Grupo_Tematico	Locais_Semelhantes_Perto
0	Arouca	restaurant	Tasquinha da Quinta	4.6	R. 1º de Maio 3, 4540-121 Arouca, Portugal	restaurant, food, point_of_interest, establish...	40.929109	-8.245191	2154	d111a3c6	Serviços	6.0
1	Arouca	restaurant	A Assembleia	4.5	Tv. da Ribeira 11, 4540-102 Arouca, Portugal	restaurant, bar, food, point_of_interest, esta...	40.928766	-8.247588	1788	54221336	Serviços	6.0
2	Arouca	restaurant	Parlamento	4.6	Tv. da Ribeira 2, 4540-148 Arouca, Portugal	restaurant, food, point_of_interest, establish...	40.929011	-8.247392	2469	ca96ed20	Serviços	7.0
3	Arouca	restaurant	Casa Testinha	4.5	R. 1º de Maio 4, 4540-113 Arouca, Portugal	restaurant, food, point_of_interest, establish...	40.928876	-8.245147	623	348acace	Serviços	6.0
4	Arouca	restaurant	Pedestre 142	4.4	R. Dr. Ângelo Miranda 142 RC, 4540-140 Arouca,...	restaurant, food, point_of_interest, establish...	40.930185	-8.253210	1047	55a3ab0d	Serviços	0.0

Pre Processing - Comments (comments_google_maps_AMP.csv)

- Removal of missing comments.

```
comentarios[comentarios['Texto'].isnull()].head()
```

	Cidade	Categoria	Nome_Local	Autor	Texto	Data	Rating
186	Arouca	restaurante	O Canastro	Cláudia Martins	NaN	2 months ago	5
321	Arouca	hotel	Arouca Guest House 2	Chus	NaN	10 months ago	5
334	Arouca	hotel	MS Collection Arouca - Mosteiro de Arouca	Vítor Neves	NaN	a week ago	5
335	Arouca	hotel	MS Collection Arouca - Mosteiro de Arouca	Alexandra Barreiros	NaN	a year ago	5
590	Arouca	praia	Casa do Tanque - Arouca	Hugo Cunha	NaN	2 years ago	5

- Whitespace at the beginning and end of each comment was removed (strip), and the comment column was converted to string format to ensure consistency.
- Conversion of relative dates – The original Date column in the comments contained relative expressions (e.g., “4 months ago”, “a year ago”). To enable accurate chronological analysis, the dateparser library was used to interpret these expressions and convert them into absolute dates (Data_Convertida), based on the system’s current date.

for the average
polarity
monthly
trend chart

NLP - Comments

💡 Text processing steps:

- **Language detection and translation:** Each comment's language was detected using the langdetect library (with exception handling via LangDetectException). If the comment was not originally in English, it was translated using GoogleTranslator from the deep_translator library. The final English version was stored in a new column named translated_text.
 - **Text normalization:** Lowercasing, removal of punctuation and extra whitespace. A customized stopwords removal was applied — keeping semantically relevant terms like "not" while discarding only function words (articles, prepositions, etc.).
 - **Tokenization and lemmatization:** The spaCy library was used to tokenize and reduce words to their canonical form (lemmas), improving the consistency of semantic analysis.
 - **Sentiment analysis (TextBlob and VADER):** Both sentiment tools were applied to each cleaned comment. The resulting polarity score (scale -1 to 1) was stored, and the average polarity per place was computed to support the attractiveness index.
- to then apply topic modeling
- to perform one of the sub indexes

Data Overview



Variable Dictionary – Comments (comments_clean.csv)

Variable	Description
City	Municipality where the point of interest is located
Category	Category assigned to the place (e.g., restaurant, hotel, park)
Place Name	Name of the establishment or point of interest
Author	Name of the user who wrote the comment
Text	Original text of the comment published by the user
Date	Relative date of the comment (e.g., 'a year ago')
Rating	Rating assigned in the comment (1 to 5 stars)
Language	Original language of the comment
Translated Text	Comment translated into English (if applicable)
Normalized Text	Normalized version of the comment (lowercase, no punctuation/accents)
Lemmatized Text	Lemmatized version of the comment (base forms of words)
Polarity	Sentiment polarity score of the comment (continuous value from -1 to 1)

Database Structure

Comments (comments_clean.csv)

comentarios.head(5)

✓ 23.7s

Python

	Cidade	Categoria	Nome_Local	Autor	Texto	Data	Rating	Data_Convertida	Idioma	translated_text	Texto_Normalizado	Texto_Lematizado	Polaridade
0	Arouca	restaurante	Tasquinha da Quinta	IC	I came here as part of a tour, and this restau...	10 months ago	5	2024-06-21	en	I came here as part of a tour, and this restau...	i came here as part of a tour and this restaur...	come tour restaurant save tour goat meat pot...	0.000000
1	Arouca	restaurante	Tasquinha da Quinta	Preetam Nath	Has to be the most delicious veal I've had in ...	a year ago	5	2024-04-21	en	Has to be the most delicious veal I've had in ...	has to be the most delicious veal ive had in m...	delicious veal ve life probably good meat ve l...	0.583333
2	Arouca	restaurante	Tasquinha da Quinta	Roya MJ	Came here as part of a tour and very much appr...	a year ago	5	2024-04-21	en	Came here as part of a tour and very much appr...	came here as part of a tour and very much appr...	come tour appreciate enjoy food vibe service \...	0.240000
3	Arouca	restaurante	Tasquinha da Quinta	Jonathan lugo	Very good food and an excellent place to eat! ...	a year ago	5	2024-04-21	en	Very good food and an excellent place to eat! ...	very good food and an excellent place to eat w...	good food excellent place eat go group food ta...	0.850000
4	Arouca	restaurante	Tasquinha da Quinta	Benjamim Nande	Great place for a good typical Portuguese food...	4 months ago	5	2024-12-21	en	Great place for a good typical Portuguese food...	great place for a good typical portuguese food...	great place good typical portuguese food good ...	0.446667

Descriptive Exploratory Analysis

- **Number of tourist locations: 3498**

```
# Ver número de locais únicos
n_locais = ratings_df['id_unico'].nunique()
print(f"Número de locais turísticos: {n_locais}")
```

Número de locais turísticos: 3498

- **Total number of comments collected: 9063**

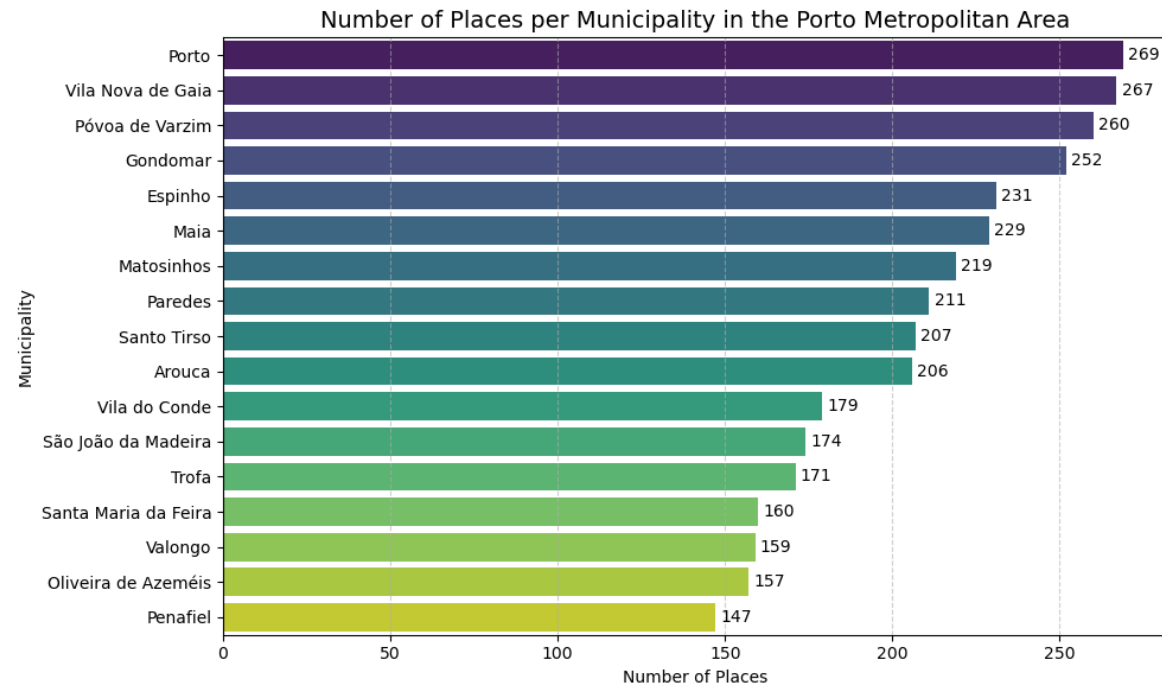
```
# Ver número total de comentários
n_comentarios = comments_df.shape[0]
print(f"Número total de comentários: {n_comentarios}")
```

Número total de comentários: 9063

Descriptive Exploratory Analysis

Locations by Municipality:

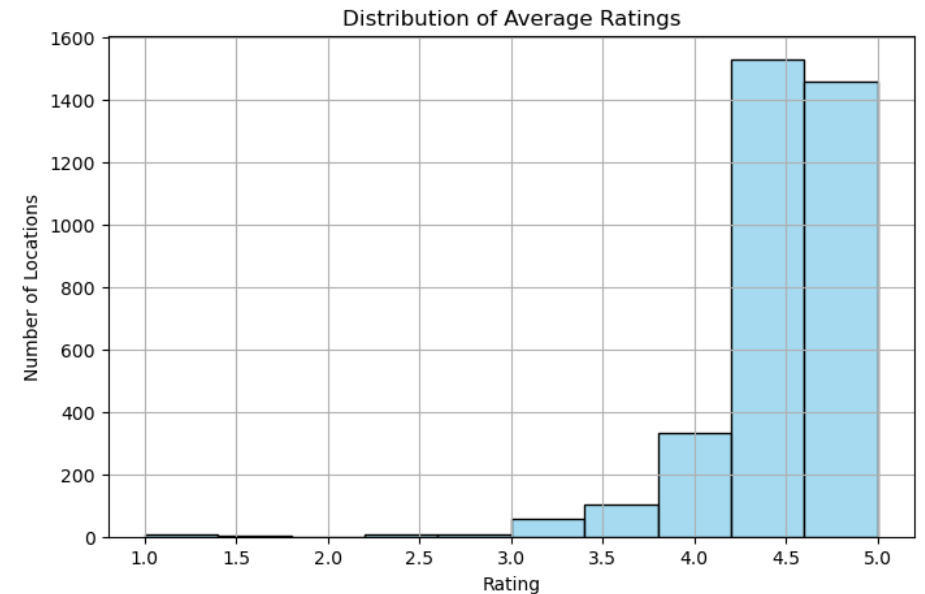
- The distribution of tourist sites across the Porto Metropolitan Area shows a notably higher concentration in the larger urban centres—Porto (269 sites) and Vila Nova de Gaia (267)—followed by Póvoa de Varzim (260), Gondomar (252) and Espinho (231). This pattern reflects not only population density and economic dynamism but also the digital visibility of these attractions on review platforms. Nonetheless, the gap between the municipality with the most sites (Porto) and the one with the fewest (Penafiel, with 147) remains relatively moderate, indicating a balanced territorial coverage well-suited for comparative analysis of perceived tourist attractiveness.



Descriptive Exploratory Analysis

Descriptive Statistics for “Rating”:

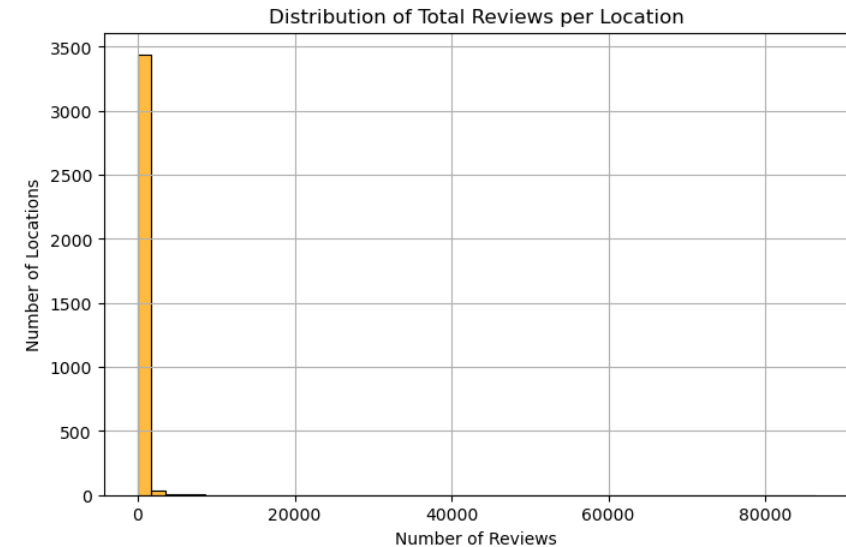
- The Ratings exhibit a strongly positive overall perception of the assessed sites, with a mean of 4.46 on a 1–5 scale. The median is 4.5, and 75% of locations have ratings of 4.3 or higher, indicating a pronounced skew toward favorable evaluations. Nonetheless, there is some variability (standard deviation = 0.40), and the minimum value of 1.0 confirms the presence of a small number of negative experiences.

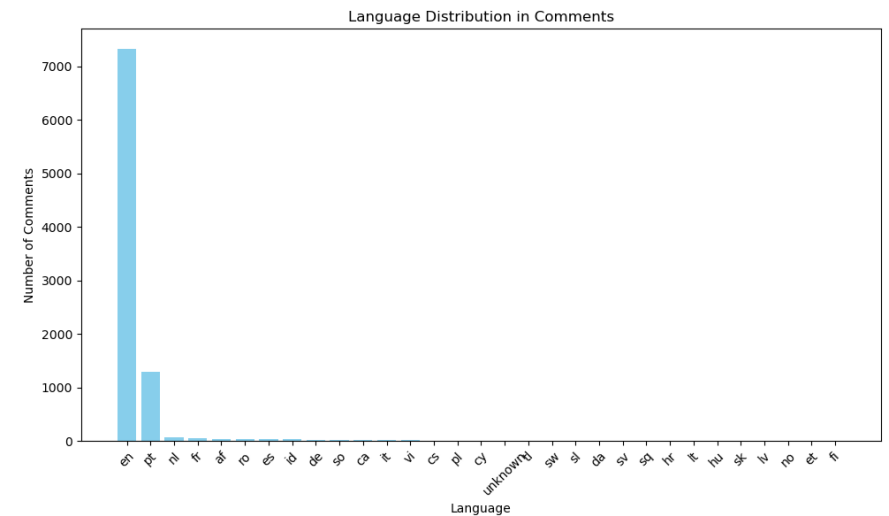


Descriptive Exploratory Analysis

Descriptive Statistics for “Total_Reviews”:

- The Total Reviews variable is highly skewed: among 3,498 sites, the median is 0 (half of all POIs have received no reviews, and this remains true up to the 75th percentile), whereas the mean (≈ 166) and standard deviation ($\approx 1,780$) are driven upward by a handful of extremely popular locations. Indeed, the maximum value of 86,331 reviews underlines a long right tail, where few sites concentrate the vast majority of feedback. Consequently, when constructing the Popularity sub-index, it is advisable to apply an appropriate transformation (e.g., logarithm) or robust normalization technique to mitigate the influence of these outliers and enable fairer comparisons across all locations.

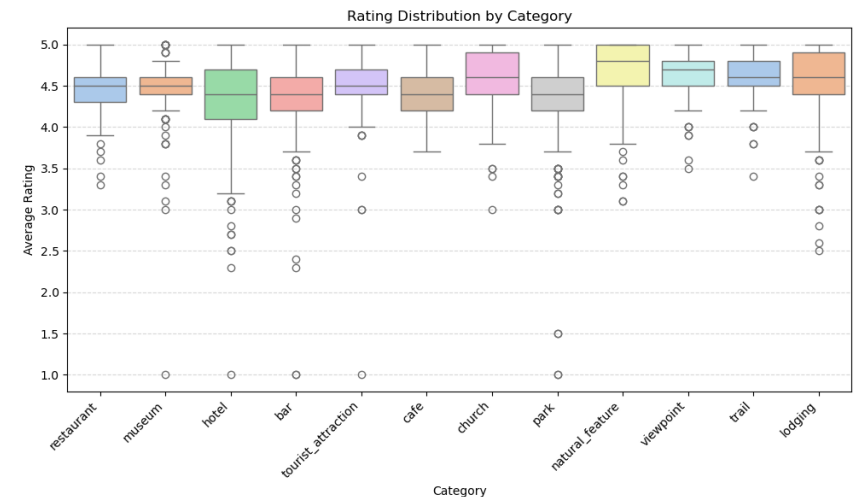




Descriptive Exploratory Analysis

Rating Distribution by Category

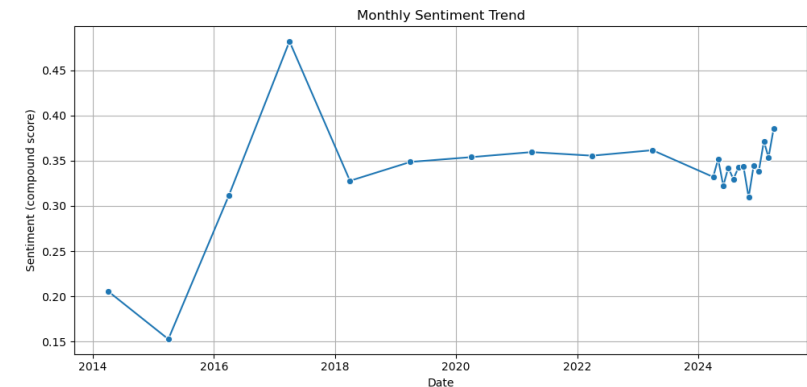
- The category-wise rating distributions exhibit distinct patterns in both central tendency and variability. “Natural_feature,” “viewpoint,” and “church” show medians around 4.8–5.0★ with narrow interquartile ranges, indicating uniformly high and consistent ratings. In contrast, “hotel,” “bar,” and “park” have lower medians (approximately 4.2–4.4★) and wider boxes, reflecting more diverse visitor experiences; these categories also include a noticeable number of low-end outliers (1.0–2.0★). “Cafe” and “tourist_attraction” share similar medians (≈ 4.4 – 4.5 ★) with moderate spread, while “restaurant” and “museum” occupy intermediate positions, featuring medians near 4.5★ and relatively compact distributions. Overall, natural sites and viewpoints enjoy the strongest consensus of positive ratings, whereas accommodations, bars, and parks display more polarized feedback.



Descriptive Exploratory Analysis

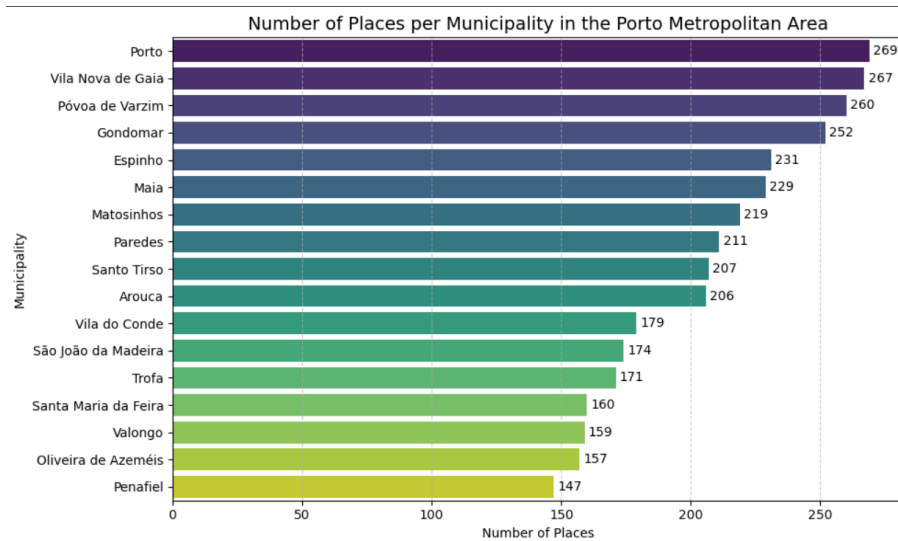
Monthly Trend of Average Polarity

- The monthly average polarity trend chart displays several notable peaks and troughs—most prominently a sharp rise between 2017 and 2018—which may reflect specific events (such as attraction renovations, promotional campaigns, or seasonal factors). From 2019 through mid-2024, polarity stabilizes around 0.33–0.36, with a modest uptick toward the end of 2024 and early 2025, indicating a slight increase in positivity in the most recent comments. Although based on up to five reviews per location and relative-date conversions, this longitudinal pattern underscores a generally consistent positive perception over time, warranting future validation with more comprehensive historical data.



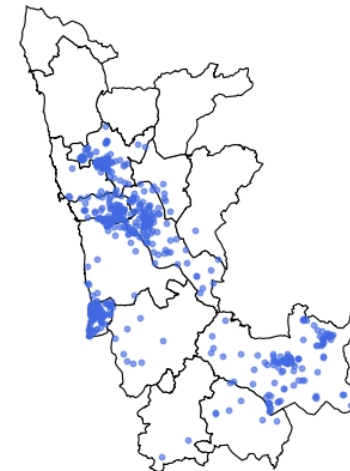
Spatial Exploratory Analysis

Based on the CSV data:



However, when we plot the spatial distribution using the geographic coordinates, there are several gaps in coverage, indicating that the coordinates obtained from the Google Places API are not fully reliable.

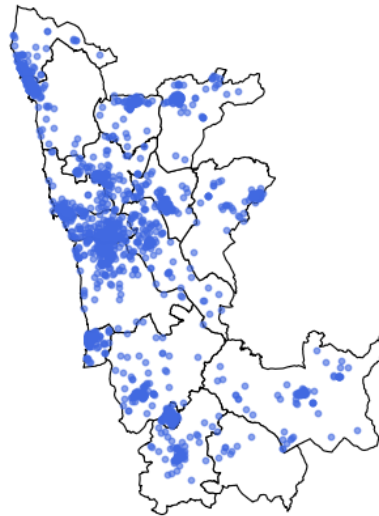
Figure 1: Geographic Distribution of Tourist Locations in the Porto Metropolitan Area



Spatial Exploratory Analysis

- Given that we had the address column, we attempted geocoding to retrieve missing coordinates. The result improved overall coverage, although some gaps remain due to addresses that could not be resolved via geocoding.
- The outcome was as follows:

Figure 1: Geographic Distribution of Tourist Locations in the Porto Metropolitan Area



Spatial Exploratory Analysis

- Nonetheless, even after geocoding some locations still lack coordinates, but this gap has been substantially reduced.

```
# Locais cuja Cidade é "Porto"
porto_csv = ratings_df[ratings_df['Cidade'] == 'Porto']

# Locais que caem espacialmente dentro do município de Porto (via shapefile)
porto_spatial = joined[joined['Cidade'] == 'Porto']

# Diferença entre os dois conjuntos
print(f"Nº locais marcados como 'Porto' no CSV: {len(porto_csv)}")
print(f"Nº locais em 'Porto' segundo o shapefile: {len(porto_spatial)}")
```

Nº locais marcados como 'Porto' no CSV: 269
Nº locais em 'Porto' segundo o shapefile: 220

```
# Obter os ID únicos dos que estão realmente no Porto (via shapefile)
ids_porto_espacial = set(porto_spatial['id_unico'])

# Filtrar os que têm 'Porto' no CSV mas NÃO estão no polígono do Porto
porto_fora_poligono = porto_csv[~porto_csv['id_unico'].isin(ids_porto_espacial)]

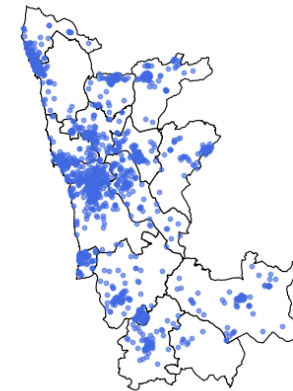
# Ver alguns exemplos
porto_fora_poligono[['Nome', 'Endereço', 'Categoria', 'Latitude_Nova', 'Longitude_Nova']].head()
```

	Nome	Endereço	Categoria	Latitude_Nova	Longitude_Nova
1652	The Door	Rua das Taipas 94 a 96, 4050-598 Porto, Portugal	restaurant	NaN	NaN
1653	Tapabento S.Bento	Estação de S. Bento, ao lado da Linha 1, R. da...	restaurant	NaN	NaN
1668	Culto ao Bacalhau	Piso Galeria, R. Formosa 322 Loja R8, 4000-248...	restaurant	NaN	NaN
1671	Cozinha das Flores	Lgo de S. Domingos 62, 4050-545 Porto, Portugal	restaurant	NaN	NaN
1673	Do Norte Café by Hungry Biker Brunch & Break...	Rua do Almada 57/59 4000, 4050-036 Porto, Port...	restaurant	NaN	NaN

Spatial Exploratory Analysis

- The geographic distribution of points of interest shows a strong concentration in the urban core of the Porto Metropolitan Area, particularly along the coastal corridor encompassing Porto, Vila Nova de Gaia, and Matosinhos. Secondary clusters appear in Póvoa de Varzim and Vila do Conde, as well as more scattered groups in inland municipalities such as Penafiel and Santa Maria da Feira. Despite gaps in coordinates for some records—which may underrepresent the most central areas—it is evident that perceived attractiveness is not confined to a single hub: there are both densely reviewed historic centers and coastal tourism poles, along with occasional clusters in peripheral municipalities. This pattern suggests a complementarity between established urban areas and natural or cultural destinations in the surrounding region.

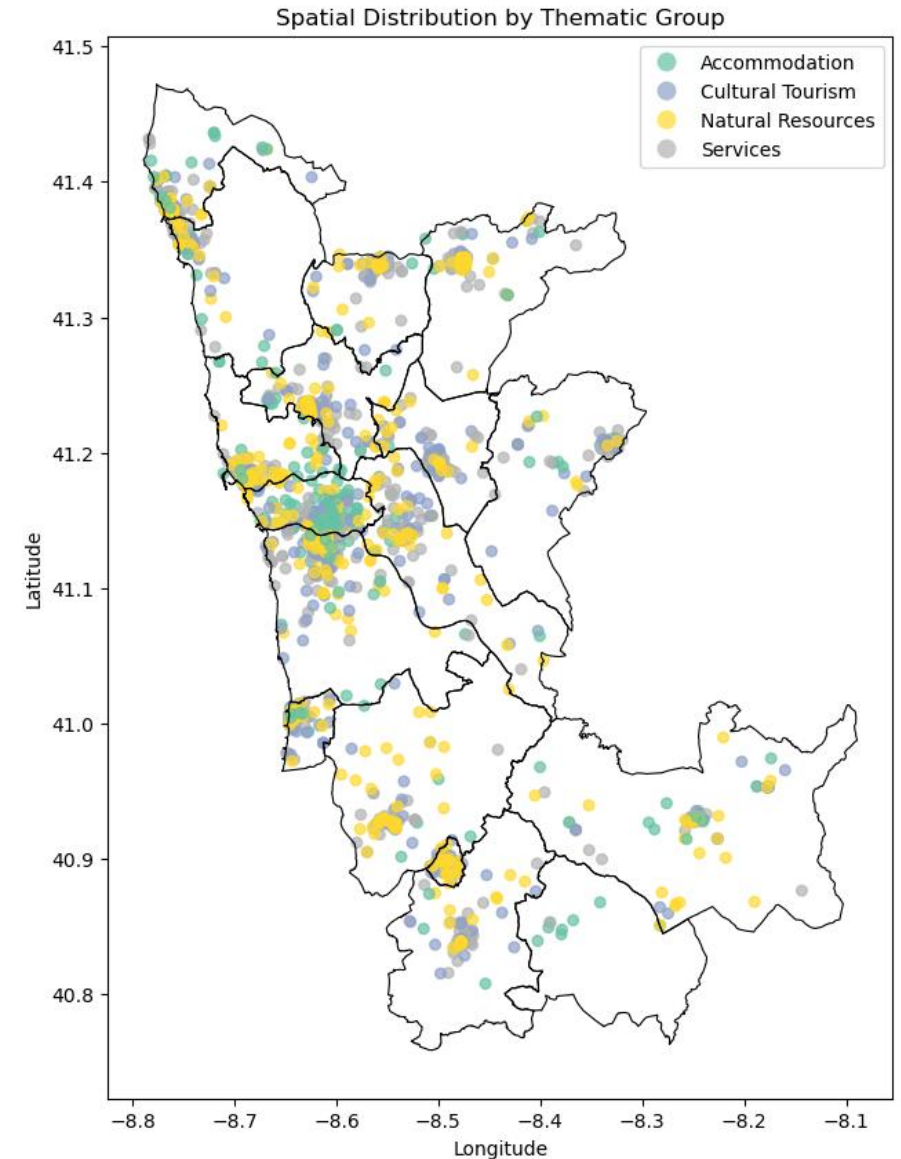
Figure 1: Geographic Distribution of Tourist Locations in the Porto Metropolitan Area



Spatial Analysis

Spatial Distribution of Tourist Locations by Thematic Group

- This map displays the spatial distribution of tourist-related locations in the Porto Metropolitan Area (AMP), classified by thematic group.
- Thematic categories shown:
 - Accommodation: hotels, hostels, guesthouses.
 - Cultural Tourism: museums, historical sites, churches.
 - Natural Resources: parks, viewpoints, nature trails.
 - Services: cafés, restaurants, bars.
- A higher concentration of diverse locations is observed in urban centers such as Porto, Vila Nova de Gaia, and Matosinhos.
- In more peripheral municipalities, natural and cultural points of interest tend to dominate the landscape.
- This spatial breakdown confirms the territorial reach and thematic diversity of the data, serving as a baseline for further spatial or cluster-based analysis.



Global Index of Perceived Tourist Attractiveness

Purpose

- The IGATP synthesizes three distinct dimensions of online visitor perception—star ratings, review volume and textual sentiment—into a single metric on [0, 1].

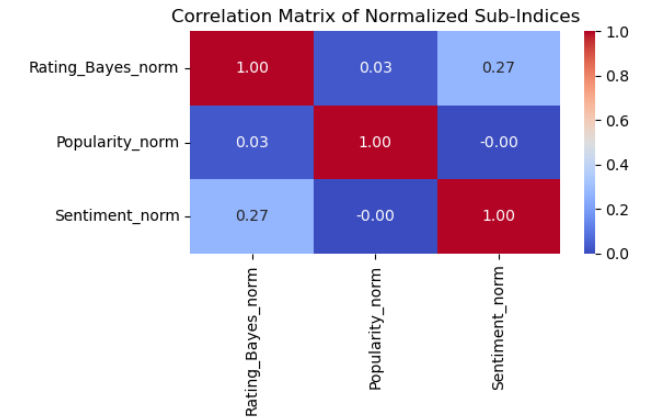
Sub-indices

- (1) **Rating_Bayes_Norm_i**: Bayesian-shrunk average star rating
- (2) **Popularity_Norm_i**: Min–Max normalized total number of reviews
- (3) **Sentiment_Norm_i**: Linear rescaling of mean text sentiment (compound polarity)

$$IGATP_i = \frac{1}{3} \textit{Rating_Bayes_Norm}_i + \frac{1}{3} \textit{Popularity_Norm}_i + \frac{1}{3} \textit{Sentiment_Norm}_i.$$

Weights

The correlation matrix between the three normalized sub-indices — Bayesian rating, popularity, and sentiment — reveals weak or negligible correlations ($r \approx 0.00$ to 0.27). This indicates that each sub-index captures a distinct dimension of perceived tourist attractiveness. Therefore, assigning equal weights to the components when computing the IGATP is statistically justified, as it ensures a balanced contribution from each factor without redundancy or dominance.



Global Index of Perceived Tourist Attractiveness

Empirical Bayes Adjustment for Ratings

- Prior distribution:

$$\mu_0 \approx 4,456 \quad \sigma^2 \approx 0,161 \quad \tau^2 \approx 0,000$$

- Shrinkage factor

$$\lambda_i = \frac{\tau^2}{\tau^2 + \frac{\sigma^2}{n_i}}$$

- Posterior (adjusted) rating

$$\hat{r}_i = \lambda_i * \bar{r}_i + (1 - \lambda_i) * \mu_0$$

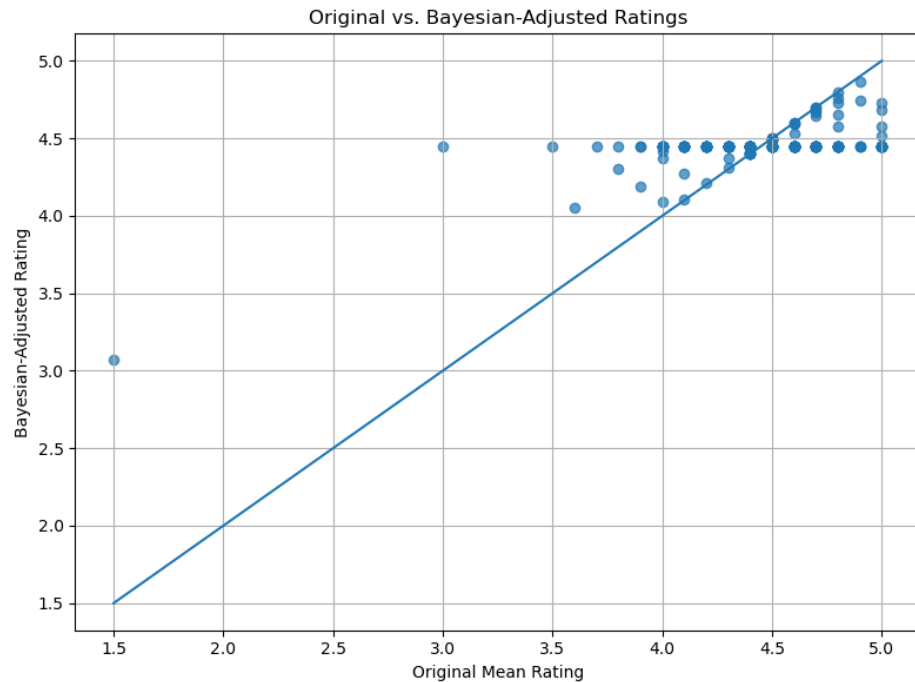
Global Index of Perceived Tourist Attractiveness

Empirical Bayes Adjustment for Ratings

- Interpretation:
 - If n_i (number of reviews) is small $\Rightarrow \lambda_i \approx 0 \Rightarrow \hat{r}_i \rightarrow \mu_0$.
 - If n_i is large $\Rightarrow \lambda_i \approx 1 \Rightarrow \hat{r}_i \approx \bar{r}_i$

Global Index of Perceived Tourist Attractiveness

Empirical Bayes Adjustment for Ratings

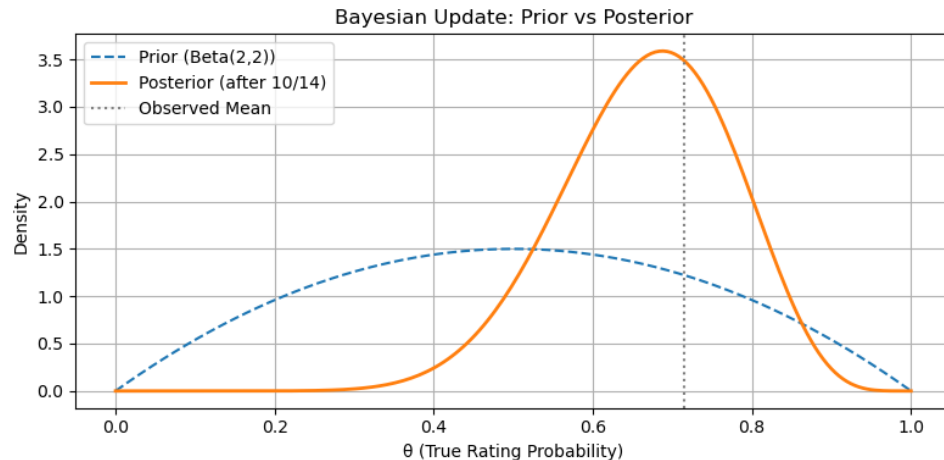


The scatter-and-identity-line plot clearly illustrates the “shrinkage” effect of the Bayesian adjustment. Venues with unusually low original ratings tend to have their adjusted scores pulled upward toward the global mean, appearing above the diagonal line. Conversely, venues with very high original ratings see their adjusted scores pulled downward, falling below the identity line. Points closer to the diagonal correspond to venues with more stable and well-supported ratings, which are only minimally adjusted. Overall, the Empirical Bayes adjustment works to compress extreme values, stabilizing ratings and offering a more reliable and comparable measure of perceived attractiveness across venues.

Global Index of Perceived Tourist Attractiveness

Empirical Bayes Adjustment for Ratings

- Interpretation of the Bayesian Update Plot:



Prior Distribution (Beta(2,2))

- Represents initial belief before seeing any data.
- It is symmetric around 0.5, reflecting neutral expectations (neither optimistic nor pessimistic).
- Equivalent to imagining 2 positive and 2 negative ratings (weak prior).

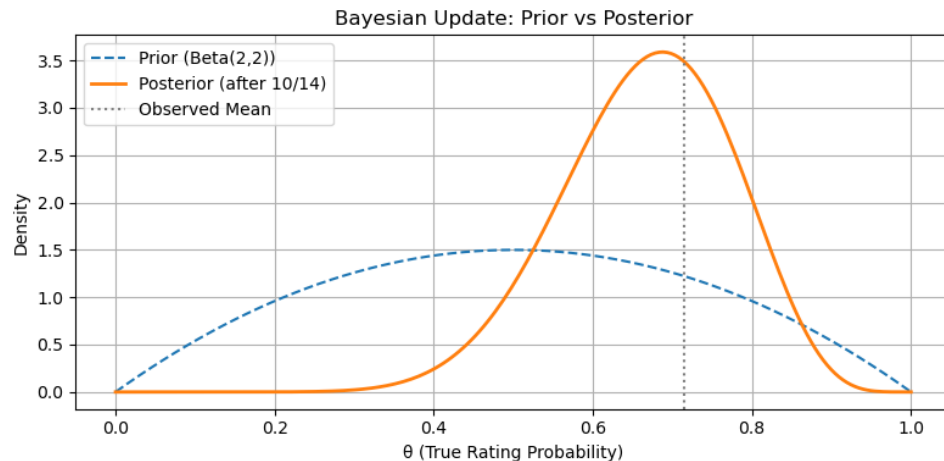
Observed Data

- 10 out of 14 reviews are positive \rightarrow empirical mean ≈ 0.714 .
- This provides the evidence to update the prior.

Global Index of Perceived Tourist Attractiveness

Empirical Bayes Adjustment for Ratings

- Interpretation of the Bayesian Update Plot:



Posterior Distribution (Beta(12,6))

- Combines the prior with the observed data.
- The curve shifts right toward the observed mean, but does not fully reach it.
- Reflects increased confidence in higher true rating probability, while still incorporating uncertainty.

Vertical Line (Observed Mean)

- Marks the empirical proportion of positive reviews.
- The posterior centers close to this line, but is smoothed due to the prior.

Key Insight

- Bayesian updating balances prior knowledge and new data.
- It avoids overconfidence in small samples and stabilizes comparisons across locations with different numbers of reviews.

Global Index of Perceived Tourist Attractiveness

Average sub-indices:

- **(1) Rating_Bayes_Norm** ≈ 0.81
- **(2) Popularity_Norm** ≈ 0.0022
- **(3) Sentiment_Norm** ≈ 0.6548

Global Index Range: [0.095, 0.873]

Global Index of Perceived Tourist Attractiveness

Top 5 locations by IGATP:

Rank	Location	IGATP
1	Luís I Bridge	0.873
2	Jardim do Morro	0.644
3	Abadessa Restaurante e Petiscos	0.635
4	Parque Nascente Shopping	0.622
5	La Bocca Dolce	0.613

Global Index of Perceived Tourist Attractiveness

Data was exported to `composite_index.csv`

Columns: 'Cidade', 'Categoria', 'Nome', 'Rating', 'Endereço', 'Tipos',
'Latitude', 'Longitude', 'Total_Reviews', 'id_unico', 'Grupo_Tematico',
'Locais_Semelhantes_Perto', 'Latitude_Nova', 'Longitude_Nova',
'Endereço_Limpo', 'shrinkage', 'Rating_Bayes', 'Nome_Local',
'Avg_Polarity', 'Rating_Bayes_norm', 'Popularity_norm',
'Sentiment_norm', 'IGATP'

Clustering Methodology

Why we applied clustering techniques:

To better understand the diversity of touristic locations in the Porto Metropolitan Area (AMP), we applied clustering techniques as an unsupervised learning approach to group places based on similarities in their perceived attractiveness profiles. By combining normalized sub-indices—Bayesian-adjusted rating, popularity, and sentiment polarity—we aimed to uncover latent patterns and classify locations into distinct segments. This segmentation enables a more interpretable analysis of the territory, identifying, for example, "boutique gems" or "mainstream attractions", and supports more targeted insights for stakeholders interested in tourism planning, territorial marketing, or resource allocation.

Clustering Methodology

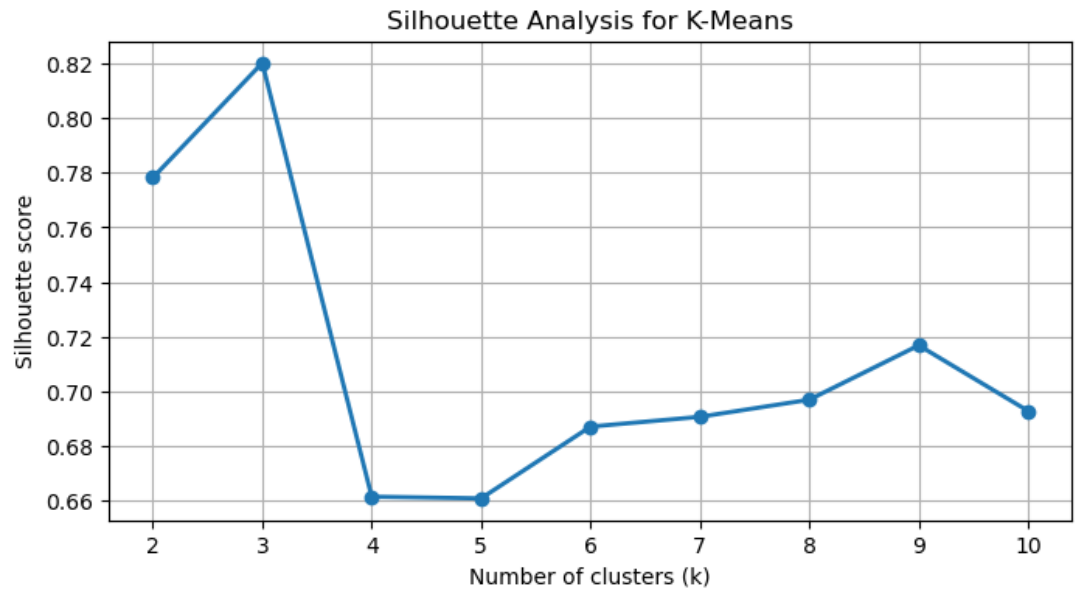
Silhouette Analysis (K-Means)

- Loaded the composite index dataset (composite_index.csv) containing the three normalized sub-indices:
 - **(1) Rating_Bayes_norm**
 - **(2) Popularity_norm**
 - **(3) Sentiment_norm**
- Extracted these three columns into a feature matrix for clustering.
- Applied mean imputation to replace any missing values with the column mean.
- Standardized each feature to zero mean and unit variance (z-score normalization).
- Performed K-Means clustering for $k = 2 \dots 10$ and computed the silhouette score for each k .
- Plotted the silhouette scores versus k to identify the optimal number of clusters.

Clustering Methodology

Silhouette Analysis (K-Means)

The silhouette analysis reveals a clear peak at $k = 3$ (≈ 0.82), indicating strong cohesion and separation in a simplified segmentation. However, almost all observations fall into a single cluster, limiting interpretability. Beyond $k = 3$, scores decrease but stabilize, remaining in the 0.66–0.70 range. Among these, $k = 6$ offers a valuable compromise: it segments the data into multiple interpretable groups while maintaining a silhouette score close to the plateau. In practical terms, $k = 6$ is selected to balance segmentation richness and clustering quality.



Clustering Methodology

Imbalanced Cluster Sizes

At $k = 6$, K-Means yields a well-distributed segmentation:

- One large cluster ($n = 1983$)
- Three moderately sized clusters ($n = 246, 145, 118$)
- One small but non-negligible group ($n = 27$)
- One singleton ($n = 1$), likely an outlier

This configuration avoids the micro-cluster inflation observed at $k = 8$, while still enhancing granularity compared to $k = 3$.

With a silhouette score of 0.68, this structure offers improved interpretability without introducing excessive fragmentation.

Cluster Size Distribution
at $k=6$:

Cluster	Count of Points
0	1983
4	246
3	145
5	118
2	27
1	1

Clustering Methodology

Imbalanced Cluster Sizes

We compared K-Means configurations for $k = 3, 6, 7$, and 8 :

- $k = 3$ yields strong silhouette but lacks segmentation (one cluster holds 97% of data).
- $k = 6$ introduces three well-sized groups and a small meaningful segment, with minimal noise.
- $k = 7$ and $k = 8$ bring micro-clusters ($n = 1-2$), suggesting over-fragmentation.

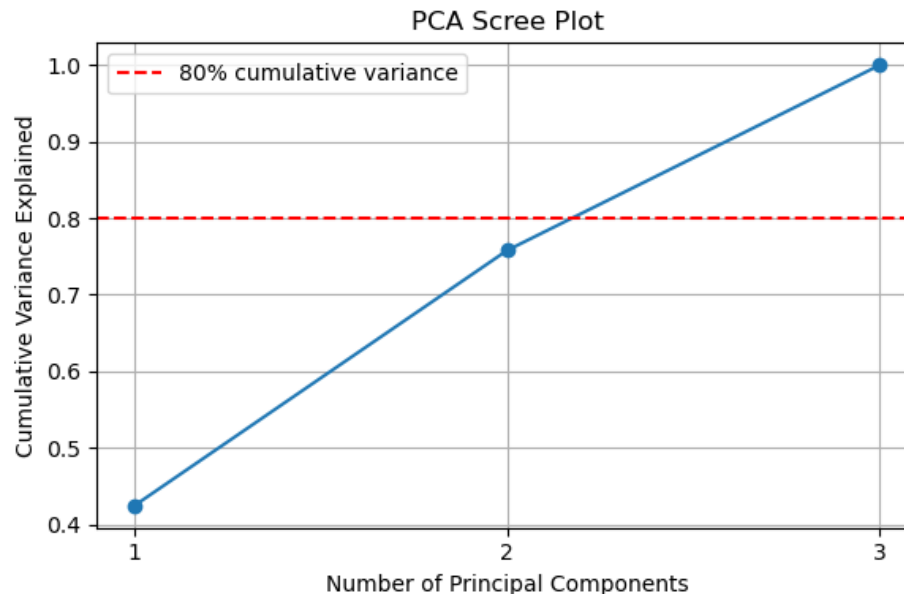
Decision: We proceed with $k = 6$, which balances structural insight, cluster size distribution, and silhouette stability. The singleton is treated as an outlier.

k	Cluster Sizes (descending)
3	2471, 48, 1
6	1983, 246, 145, 118, 27, 1
7	1983, 246, 145, 116, 27, 2, 1
8	1947, 243, 144, 140, 31, 12, 2, 1

Clustering Methodology

PCA Visualization (K-Means): Choosing the Number of Principal Components

- Cumulative Variance Criterion
 - PC1 explains 41.7 % of variance; PC1 + PC2 bring total to 75.0 %.
 - While 80 % is a common rule-of-thumb, 75 % already captures the majority of structure in our three-dimensional feature space.
- Elbow Method & Parsimony
 - The scree curve shows a marked “elbow” after the second component, indicating diminishing returns from additional PCs.
 - Retaining only PC1 and PC2 balances model simplicity and information retention, avoiding overfitting or over-compression.



Clustering Methodology

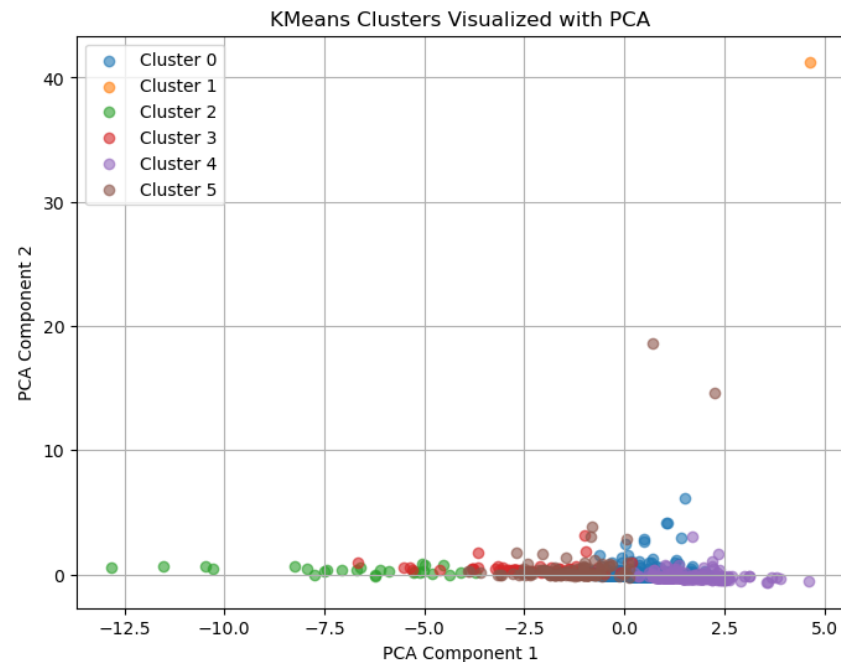
PCA Visualization (K-Means): Choosing the Number of Principal Components

- Interpretability & Visualization
 - Two components facilitate clear 2D visualizations (scatter plots, cluster overlays) for stakeholder communication.
 - Simplifies the interpretation of feature contributions (loadings) without substantial loss of explanatory power.
- Practical Trade-Off
 - Using two components reduces computational complexity for downstream tasks (e.g. clustering, mapping) while preserving the core signal.
 - A third component could be added if $> 80\%$ variance is strictly required; however, PC1–2 suffice for our exploratory and segmentation objectives.

Clustering Methodology

PCA Visualization (K-Means)

- Applied PCA to project scaled data into 2 dimensions (PC1, PC2)
- Colored each point by its K-Means cluster label for visual inspection



Feature	PC1 Loading	PC2 Loading
Rating_Bayes_norm	0.798	0.003
Popularity_norm	0.081	0.995
Sentiment_norm	0.794	-0.105

PCA Interpretation:

- PC1 (≈ 0.80 Rating + 0.79 Sentiment): captures combined “quality” and “emotional intensity”
- PC2 (≈ 0.995 Popularity): isolates “visibility/popularity” dimension

Clustering Methodology

Clustering Results & Component Loadings (k=6)

Using the 6-cluster K-Means solution and the 2D PCA projection:

- PC1 is interpreted as a proxy for quality/sentiment,
- PC2 reflects popularity/visibility.

The PCA scatterplot reveals six distinguishable clusters:

- **Cluster 0 (Mainstream Core):** The largest group, centered near the origin — reflects mid-tier, balanced locations.
- **Cluster 5 (Flagship Venues):** High quality and visibility — prominent, top-rated points.
- **Cluster 4 (Boutique / Niche):** High sentiment but low visibility — hidden gems or less-frequented highlights.
- **Cluster 2 (Underperformers):** Low quality and low popularity — likely less attractive options.
- **Cluster 3 (Hidden Popular):** Higher visibility despite below-average sentiment — potentially overrated or controversial spots.
- **Cluster 1 (Extreme Outlier):** Isolated point with extreme values in both dimensions — possibly an anomaly or very unique venue.

Cluster	Size	PC1	PC2	Profile
0	1983	≈ 0	≈ 0	Mainstream Core
5	118	↑	↑	Flagship Venues
4	246	↑	↓	Boutique / Niche
2	27	↓	↓	Underperformers
3	145	↓	↑	Hidden Popular
1	1	↑↑	↑↑	Extreme Outlier

Clustering Methodology

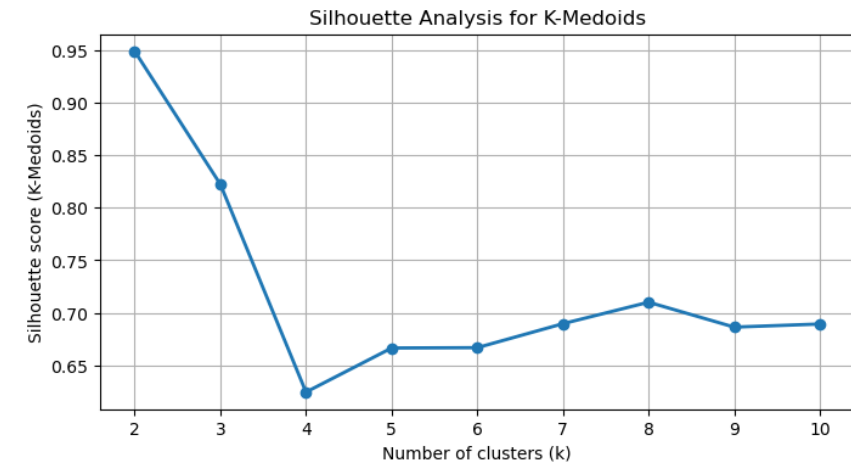
Silhouette Analysis (K-Medoids)

Decision Rationale:

- $k = 2-3$ shows high silhouette but overly broad segmentation.
- $k = 4$ exhibits a sharp drop in silhouette, suggesting weak structure.
- $k = 5-8$ yields stable silhouette scores ($\sim 0.66-0.71$), with improved interpretability.
- $k = 6$ offers a balance between cohesion and segmentation richness:
 - Clear structure without over-fragmentation.
 - Compatible with the K-Means solution for consistency.

→ **We adopt $k = 6$ for our final K-Medoids clustering, as it:**

- Provides interpretable and balanced cluster sizes
- Ensures robust cohesion and separation
- Aligns with prior K-Means results to enable cross-method comparison



Clustering Methodology

Clustering Performance Comparison (K-Means vs K-Medoids)

Adjusted Rand Index (ARI)

- ARI (KMeans vs KMedoids) for $k=6$: 0.861
⇒ Indicates very high agreement, with both algorithms recovering similar segmentations despite methodological differences.

Silhouette Scores

- K-Means ($k=6$) 0.687
- K-Medoids ($k=7$): 0.667
 - ⇒ Both > 0.6 → strong cluster cohesion and separation
 - ⇒ K-Means only marginally higher (+0.020)

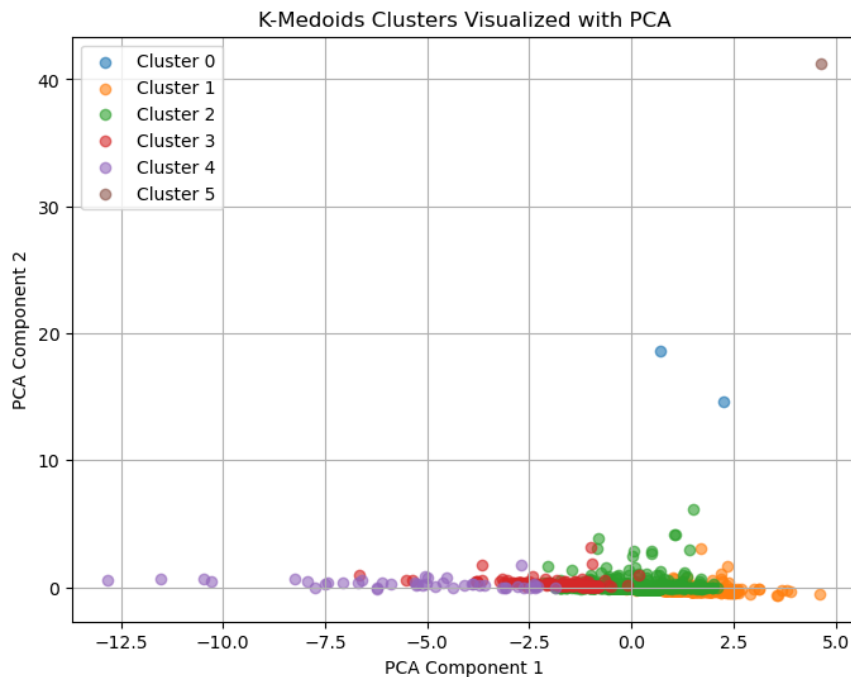
Interpretation & Choice

- Clusters are **comparably well-defined** (silhouette difference < 0.02).
- **K-Medoids preferred** for practical interpretation:
 - **Real-world exemplars:** medoids correspond to actual POIs, enhancing interpretability.
 - **Outlier resistance:** more robust against extreme or noisy values.
- ➤ Despite the slight drop in silhouette, K-Medoids provides representative, actionable tourism clusters.

Clustering Methodology

PCA Visualization (K-Medoids)

- Applied PCA to project scaled data into 2 dimensions (PC1, PC2)
- Colored each point by its K-Medoids cluster label for visual inspection



Feature	PC1 Loading	PC2 Loading
Rating_Bayes_norm	0.798	0.003
Popularity_norm	0.081	0.995
Sentiment_norm	0.794	-0.105

PCA Interpretation:

- PC1 (≈ 0.80 Rating + 0.79 Sentiment): captures combined “quality” and “emotional intensity”
- PC2 (≈ 0.995 Popularity): isolates “visibility/popularity” dimension

Clustering Methodology

Clustering Results & Component Loadings (k=6)

Using the 6-cluster K-Medoids solution and the 2D PCA projection:

- PC1 is interpreted as a proxy for quality/sentiment,
- PC2 reflects popularity/visibility.

The PCA scatterplot reveals six distinguishable clusters:

- **Cluster 2 (Mainstream Core):** The dominant segment (2070 points), positioned near the center — average locations with balanced sentiment and visibility.
- **Cluster 3 (Flagship Venues):** Highly rated and popular locations (148 points), situated in the upper-right quadrant of the PCA space.
- **Cluster 1 (Hidden Popular):** 254 points with low sentiment but higher popularity — possibly overrated or trending places.
- **Cluster 4 (Underperformers):** Small cluster (45 points) with low quality and visibility — potentially weak or outdated attractions.
- **Cluster 0 (Boutique / Niche):** Only 2 observations, high sentiment but limited exposure — hidden gems or specialized locations.
- **Cluster 5 (Extreme Outlier):** A single observation with extreme values in both dimensions — likely an anomalous or unique spot.

Cluster	Size	PC1	PC2	Profile
2	2070	≈ 0	≈ 0	Mainstream Core
3	148	↑	↑	Flagship Venues
1	254	↓	↑	Hidden Popular
4	45	↓	↓	Underperformers
0	2	↑	↓	Boutique / Niche
5	1	↑↑	↑↑	Extreme Outlier

Clustering Methodology

Exporting Clustered Dataset

We exported the full composite index with all cluster assignments to a CSV file for reproducibility and downstream analysis.

File saved as **composite_index_with_clusters.csv**, containing:

- Original POI attributes (ratings, sentiment, popularity, PCA scores, etc.)
- K-Means labels for k=2,3,6, 7, 8 and 9 (cluster_k2, (...), cluster_k9)
- K-Medoids labels for k=6 and k=7 (cluster_k6_pam, cluster_k7_pam)

This consolidated file ensures easy sharing and integration with mapping or spatial-statistical tools.

Clustering Methodology

Data was exported to `composite_index_with_clusters.csv`

Columns: 'Cidade', 'Categoria', 'Nome', 'Rating', 'Endereço', 'Tipos',
'Latitude', 'Longitude', 'Total_Reviews', 'id_unico', 'Grupo_Tematico',
'Locais_Semelhantes_Perto', 'Latitude_Nova', 'Longitude_Nova',
'Endereço_Limpo', 'shrinkage', 'Rating_Bayes', 'Nome_Local',
'Avg_Polarity', 'Rating_Bayes_norm', 'Popularity_norm',
'Sentiment_norm', 'IGATP', 'cluster_k2', 'cluster_k3', 'cluster_k6',
'cluster_k7', 'cluster_k8', 'cluster_k9', 'cluster_k7_pam',
'cluster_k6_pam'

Topic Modeling

Why Perform Topic Modeling?

- **Uncover Latent Themes**
 - Automatically extract recurring topics from thousands of visitor comments
 - Reveal what aspects (e.g. food, service, views) drive perceptions
- **Complement Quantitative Indices**
 - Go beyond star ratings and sentiment scores to understand why attractions are loved (or criticized)
 - Link thematic insights to our Bayesian-adjusted ratings for richer interpretation
- **Inform Strategic Actions**
 - Identify strengths (e.g. “wine tasting,” “scenic overlooks”) and pain points (e.g. “long waits,” “high prices”)
 - Guide targeted improvements and marketing messages based on actual visitor language
- **Enhance Segmentation & Personalization**
 - Combine topic profiles with cluster membership to describe each tourist-attraction segment in human-readable terms
 - Support tailored recommendations and experience design for different visitor interests

Topic Modeling

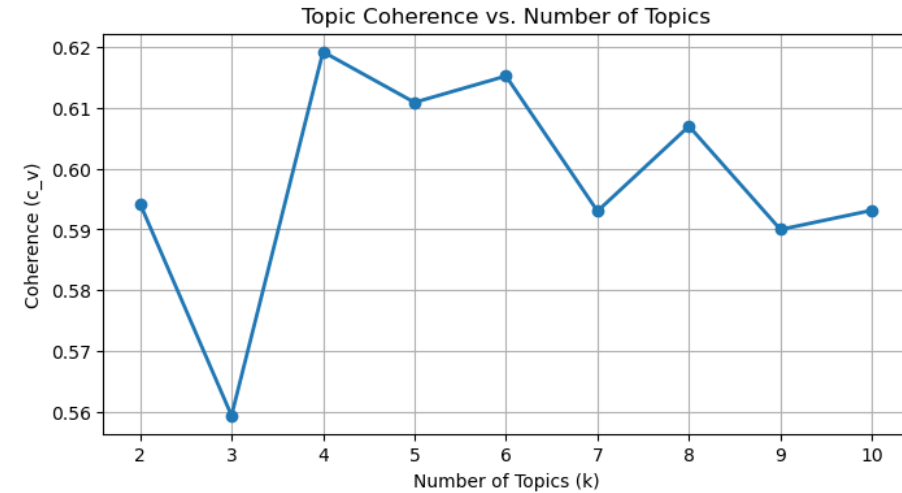
Choosing the Number of LDA Topics:

Interpretability & Parsimony:

- $k = 4$ provides a semantic sweet spot:
 - Captures clear, coherent, and distinct themes.
 - Avoids fragmentation while ensuring each topic remains interpretable.
 - Easier to label and communicate to non-technical stakeholders.

Final Choice:

- We select $k = 4$ topics for our LDA model — balancing high coherence (≈ 0.620) with clarity, conciseness, and actionability in topic interpretation.



Topic Modeling

Interpretation:

Topic 0 – Staff & Overall Experience

Emphasizes staff friendliness and overall satisfaction — terms like “stay”, “recommend”, “friendly”, “excellent”, and “clean” highlight positive general experiences with people and places.

Topic 1 – Gastronomy & Dining Quality

Focuses on food-related experiences — “food”, “restaurant”, “service”, “wine”, “delicious”, and “price” point to culinary satisfaction and value.

Topic 2 – Beach & Cultural Exploration

Combines coastal enjoyment and sightseeing — “beach”, “walk”, “museum”, “beautiful”, and “visit” suggest nature and culture-based tourism.

Topic 3 – Accommodation & Comfort

Highlights lodging quality — “room”, “hotel”, “clean”, “bed”, and “breakfast” focus on comfort, cleanliness, and sleep quality.

Topic	Top 10 Terms
0	stay , great , staff, place , recommend , friendly, good , clean, room, excellent
1	food, good , service, restaurant, great , place , price, wine, delicious, recommend
2	beach, place , nice , beautiful, walk, visit, museum, good, great , water
3	room, hotel, clean, good , stay , breakfast, night, staff, bed, nice

Stopwords (semantic):

- **good**: a vague evaluative adjective frequently used across topics, offering little thematic differentiation.
- **great**: similar to “good”; common in positive sentiment but lacks content-specific value.
- **nice**: a generic descriptor often applied in multiple contexts without semantic depth.
- **recommend**: a universal term of approval; signals satisfaction but does not specify the theme.
- **stay**: semantically broad — can refer to lodging, duration, or general experience.
- **place**: an ambiguous noun occurring in all contexts, offering no topic-level discrimination.

Topic Modeling

Interpretation (after stop-word removal):

Topic 0 – Scenic Beauty & Leisure

Highlights visual and emotional appreciation of surroundings — “beautiful”, “view”, “walk”, “enjoy”, and “perfect” suggest a focus on aesthetic pleasure and relaxation.

Topic 1 – Accommodation & Location

Centers on lodging experiences — “room”, “hotel”, “clean”, “bed”, and “location” reflect comfort, hygiene, and proximity, with some overlap into coastal contexts.

Topic 2 – Cultural & Museum Visits

Describes cultural exploration — “museum”, “visit”, “experience”, “interesting”, and “history” indicate educational and enriching activities.

Topic 3 – Dining & Hospitality

Focuses on the food experience — “food”, “restaurant”, “friendly”, “delicious”, “wine” and “service” reflect culinary satisfaction and host interactions.

Topic	Top 10 Terms
0	beautiful, beach, view, walk, amazing, porto , perfect , enjoy, area , thank
1	room, hotel, clean, beach, breakfast, staff, night, bed, location , walk
2	museum, visit, time , experience, worth, go , hour , interesting, history, small
3	food, service, restaurant, friendly, staff, delicious, price, excellent , wine, amazing

Stopwords (semantic):

- **porto**: a location-specific term that skews topic distribution; not representative of the broader metropolitan area under study.
- **thank**: a generic expression of gratitude lacking thematic content.
- **area**: vague spatial reference with low discriminative power across contexts.
- **location**: a general-purpose noun appearing in multiple topics without thematic specificity.
- **time**: a temporal word commonly used but rarely tied to a specific theme.
- **hour**: similar to “time”; frequently used in visit descriptions but with low semantic contribution.
- **go**: a high-frequency verb with little value in distinguishing topic themes.
- **amazing**: a non-specific evaluative adjective that signals sentiment but lacks content.
- **perfect**: emotionally charged but not thematically informative.
- **excellent**: common in praise across topics but does not anchor specific content.

Topic Modeling

Interpretation (after stop-word removal):

Topic 0 – Lodging & Cultural Impressions

Reflects a combination of accommodation-related experiences and museum or indoor visits — terms such as “room”, “museum”, “hotel”, “night”, and “house” suggest overnight stays and cultural exploration. Previously dominant vague or negative terms (e.g., “bad”, “work”) have been removed to clarify the theme.

Topic 1 – Accommodation & Guest Comfort

Captures positive lodging attributes — “clean”, “comfortable”, “quiet”, “city”, and “walk” reflect location accessibility and overall guest satisfaction. The removal of “helpful” improves topic cohesion.

Topic 2 – Coastal Leisure & Scenic Elements

Emphasizes outdoor, seaside enjoyment — terms like “beach”, “sand”, “wave”, “sea”, and “view” point to contact with nature and coastal walking.

Topic 3 – Gastronomy & Culinary Satisfaction

Dedicated to food quality and restaurant service — “food”, “restaurant”, “wine”, “dish”, and “service” convey a clear focus on dining experience.

Topic	Top 10 Terms
0	room, museum, hotel, bad , night, visit, like , work , small , house
1	room, clean, hotel, breakfast, staff, walk, comfortable, quiet, helpful , city
2	beach, walk, water, view, sand, visit, people , wave, long, sea
3	food, service, restaurant, delicious, price, staff, wine, dish, quality, eat

Stopwords (semantic):

- **like**: a filler verb used broadly in expressions of preference or similarity, lacking thematic relevance.
- **work**: a vague term that may refer to functionality or employment, offering low topic specificity.
- **small**: a generic size descriptor with little semantic contribution to topic differentiation.
- **bad**: a subjective evaluation common across complaint narratives, lacking thematic precision.
- **people**: an overly broad noun that appears across contexts without adding content clarity.
- **helpful**: a sentiment-laden adjective with limited thematic value across diverse contexts.

Topic Modeling

Interpretation (after stop-word removal):

Topic 0 – Coastal & Cultural Attractions

Captures seaside and museum-related experiences — terms such as “beach”, “walk”, “museum”, “water”, and “view” reflect outdoor leisure and cultural visits. “parking” and “worth” suggest practical aspects and perceived value.

Topic 1 – Accommodation & Guest Comfort

Focuses on positive lodging experiences — “room”, “clean”, “comfortable”, “bed”, and “quiet” indicate rest and hygiene, while “pool” and “breakfast” highlight amenities.

Topic 2 – Gastronomy & Culinary Experience

Centers on dining quality and satisfaction — “food”, “restaurant”, “delicious”, “dish”, and “wine” express pleasure, while “price” and “staff” address value and service.

Topic 3 – Booking & Service Frustrations

Describes issues related to reservations and customer service — “book”, “leave”, “pay”, “find”, “tell”, and “ask” suggest communication problems or dissatisfaction with booking procedures or stay.

Topic	Top 10 Terms
0	beach, walk, visit, museum, water, view, sand, wave, parking, worth
1	room, hotel, clean, breakfast, staff, comfortable, night, bed, pool, quiet
2	food, service, restaurant, delicious, price, staff, wine, dish, quality, eat
3	room, night, book, hotel, leave , pay, find , tell , ask , center

Stopwords (semantic):

- **worth:** a generic evaluative term (“worth it”) conveying subjective value but lacking thematic specificity.
- **leave:** a broad verb that may refer to checkout, dissatisfaction, or abandonment, without clear topic anchoring.
- **tell:** a vague communication verb with minimal semantic contribution in thematic modeling.
- **ask:** a frequent action verb that reflects interaction but lacks discriminative power.
- **find:** a common verb expressing discovery or confusion, semantically broad and unspecific.
- **center:** an ambiguous spatial noun that may refer to location or urban context, but does not define a coherent topic.

Topic Modeling

Interpretation (after stop-word removal):

Topic 0 – Lodging & Guest Comfort

Focused on accommodation quality and guest experience — terms like “room”, “clean”, “comfortable”, “bed”, and “bathroom” reflect attention to cleanliness and comfort, while “walk” and “night” suggest convenience and overnight stays.

Topic 1 – Culinary Experience

Captures gastronomic satisfaction and service — “food”, “restaurant”, “delicious”, and “dish” indicate enjoyment, while “price” and “service” point to value and interaction quality.

Topic 2 – Museum & Cultural Visits

Describes visits to museums and historical sites — “museum”, “history”, “house”, and “tour” evoke indoor cultural experiences. Verbs like “say”, “pay”, and “book” reflect logistical and interpretative aspects.

Topic 3 – Coastal Leisure

Reflects outdoor and seaside experiences — “beach”, “sand”, “sea”, and “wave” indicate contact with nature, while “walk”, “view”, and “parking” add practical and aesthetic dimensions.

Topic	Top 10 Terms
0	room, hotel, clean, breakfast, staff, night, bed, comfortable, bathroom, walk
1	food, service, restaurant, delicious, staff, price, wine, dish, eat, quality
2	visit, museum, history, interesting, say , house, pay, tour, book, english
3	beach, walk, water, view, sand, wave, parking, sea, lovely , clean

Stopwords (semantic):

- **say**: a vague communication verb frequently used in quoted opinions or indirect references, offering low thematic value.
- **lovely**: a sentiment-driven adjective that signals positive emotion but lacks specific contextual meaning, reducing topic clarity.

Topic Modeling

Interpretation (after stop-word removal):

Topic 0 – Seaside Leisure & Accessibility

Focused on coastal enjoyment and practical surroundings — “beach”, “sea”, “sand”, “wave” and “view” reflect natural features, while “walk”, “parking”, and “restaurant” suggest convenience and amenities near the coast.

Topic 1 – Culinary Satisfaction & Dining Quality

Highlights gastronomic experience — “food”, “restaurant”, “delicious”, and “wine” convey positive perception of meals, while “price”, “service”, and “staff” refer to practical service dimensions.

Topic 2 – Lodging Comfort & Facilities

Emphasizes hotel quality and comfort — “room”, “clean”, “bed”, “shower”, “bathroom” and “comfortable” indicate satisfaction with accommodation and hygiene.

Topic 3 – Personal Welcome & Emotional Connection

Distinctively captures emotional warmth and host interaction — “love”, “feel”, “welcome”, “family”, “super” and “host” suggest affective language tied to homestays or guest relations.

Topic	Top 10 Terms
0	beach, walk, water, view, sand, clean, restaurant, parking, sea, wave
1	food, service, restaurant, delicious, price, staff, wine, dish, eat, quality
2	room, hotel, clean, staff, breakfast, night, bed, bathroom, comfortable, shower
3	visit, museum, house, love , space, welcome , feel , family, super , host

Stopwords (semantic):

- **super**: an informal intensifier used in casual praise (e.g., “super nice”, “super host”), lacking thematic precision.
- **love**: an emotion-laden verb expressing strong sentiment but offering no topic-specific content.
- **feel**: a subjective verb related to personal impressions, weak in thematic discrimination.
- **welcome**: a positive reception term that conveys friendliness but lacks contextual depth.

Topic Modeling

Interpretation (after stop-word removal):

Topic 0 – Coastal & Outdoor Leisure

Focused on seaside and open-air experiences — “beach”, “sand”, “wave”, “view”, and “walk” highlight nature and landscape enjoyment, while “restaurant”, “quiet”, and “parking” reflect convenience and amenities.

Topic 1 – Accommodation & Guest Facilities

Centers on lodging and comfort — “room”, “clean”, “bed”, “shower”, and “comfortable” describe physical conditions, while “staff”, “breakfast”, and “night” capture service and overnight experience.

Topic 2 – Cultural & Historical Visits

Emphasizes museum-related and heritage tourism — “museum”, “history”, “tour”, “portuguese”, and “space” suggest educational or historical interest. “visit” and “host” may indicate structured experiences or guided interactions.

Topic 3 – Gastronomic Quality & Dining Service

Strongly reflects food-related satisfaction — “food”, “restaurant”, “delicious”, “wine”, and “dish” point to enjoyment, while “price”, “service”, and “staff” convey aspects of value and attention.

Topic	Top 10 Terms
0	beach, walk, view, water, sand, parking, clean, restaurant, quiet, wave
1	room, hotel, clean, breakfast, staff, night, bed, comfortable, bathroom, shower
2	visit, museum, house, history, interesting, portuguese, space, tour, family, host
3	food, service, restaurant, delicious, price, staff, wine, dish, eat, quality

**No additional stop-words jump out—
this set suffices before moving on to
the next analysis step.**

Topic Modeling – Cross-tab with Clusters

Linking Comment Topics to POI Clusters

- **1. Infer Topic Distributions**
 - Apply the final LDA model to every comment
 - Obtain a probability vector of length k for each document
- **2. Label Each Comment by Its Dominant**
 - ThemelIdentify the single topic with the highest probability
 - Assign that topic as the comment’s “dominant topic”
- **3. Combine with POI Cluster Assignments**
 - Bring in each point-of-interest’s k-medoids cluster label
 - Align comments to their originating POI and its cluster
- **4. Analyze Theme vs. Cluster Frequencies**
 - Create a matrix counting how many comments of each dominant topic appear in each POI cluster
 - Reveals which thematic concerns (e.g., dining, beaches, history) are most prominent per cluster

Dominant Topic	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Topic 0	158	360	143	16
Topic 1	193	656	375	215
Topic 2	203	370	160	16
Topic 3	877	978	210	28

Outcome:

This mapping uncovers the dominant visitor feedback themes within each POI segment, guiding targeted improvements and insights.

Although clusters 0 and 5 contain real places within the study area — with valid ratings and review counts — these locations do not include textual comments. As such, they were retained in the overall attractiveness index computation (via rating and popularity), but excluded from topic-based analyses, due to the absence of sentiment and thematic information.

Topic Modeling – Cross-tab with Clusters

Mainstream Core (Cluster 2)

- Largest POI segment.
- Covers balanced experiences between comfort (Topic 1) and gastronomy (Topic 3).
- Represents mid-tier venues with steady service and average visibility.

Flagship Venues (Cluster 3)

- Strongly dominated by food-related content (Topic 3).
- Popular, top-rated locations praised for their dining quality and overall experience.
- Hidden Popular (Cluster 1)
- Despite lower visibility, shows rich feedback in cultural visits (Topic 2) and comfort (Topic 1).
- Possibly under-the-radar attractions or trending places with niche appeal.

Underperformers (Cluster 4)

- Most comments highlight mediocre or weak experiences (especially comfort and service).
- Low prominence in leisure or cultural themes.
- Boutique / Niche (Cluster 0)
- Only 2 POIs, with no associated comments.
- Retained for index calculation only — excluded from topic analysis.

Extreme Outlier (Cluster 5)

- Single POI with no text reviews.
- Excluded from thematic analysis due to missing sentiment and content data, but kept in the final index.

Dominant Topic	Hidden Popular (Cluster 1)	Mainstream Core (Cluster 2)	Flagship Venues (Cluster 3)	Underperformers (Cluster 4)
Topic 0 (Beach & Outdoor Leisure)	143	360	158	16
Topic 1 (Accommodation & Comfort)	375	656	193	215
Topic 2 (Cultural & Historical Visits)	160	370	203	16
Topic 3 (Gastronomy & Dining)	210	978	877	28

Topic Modeling

Exporting Topic Modeling Results

Export to CSV

- We compiled each POI's dominant topic alongside its rating and sentiment metrics.
- Exported the combined data to ratings_polarity_lda_topics.csv for easy sharing and reproducibility

Columns: 'Cidade', 'Categoria', 'Nome_Local', 'Autor', 'Texto', 'Data', 'Rating',
'Idioma', 'Data_Convertida', 'translated_text', 'Texto_Normalizado',
'Texto_Lematizado', 'Polaridade', 'Polaridade_Média', 'Topic_0',
'Topic_1', 'Topic_2', 'Topic_3', 'dominant_topic',
'cluster_k6_pam'

Validation

Correlation Between Sentiment and Bayesian Rating (Sub-Indices)

Method: Pearson Correlation (cleaned data)

Result: $r = 0.27$, $p\text{-value} = 0.000$

Interpretation:

- Weak but significant positive correlation.
- Suggests partial overlap, but each metric captures distinct aspects of perceived attractiveness.

Implication:

- Justifies keeping both sub-indices in the global index (IGATP).
- Enhances multidimensionality and avoids redundancy.

Validation

Sensitivity Analysis: IGATP Weighting Scenarios

- Objective:
 - To evaluate the stability of the IGATP when adjusting the relative weight of its three components:
 - Rating_Bayes_norm
 - Popularity_norm
 - Sentiment_norm
- Tested Scenarios:

Scenario	Weights (Rating / Popularity / Sentiment)
Equal Weights	1/3 – 1/3 – 1/3 (reference)
Rating-Heavy	0.50 – 0.25 – 0.25
Popularity-Heavy	0.25 – 0.50 – 0.25
Sentiment-Heavy	0.25 – 0.25 – 0.50

Validation

Sensitivity Analysis: IGATP Weighting Scenarios

- Results (Pearson correlation with the equal-weight IGATP):
 - Rating-Heavy: $r = 0.968$
 - Popularity-Heavy: $r = 0.980$
 - Sentiment-Heavy: $r = 0.968$
- Conclusion: All alternative weighting scenarios show very high correlation ($r > 0.96$) with the original IGATP. This indicates that the index is highly robust and not sensitive to moderate changes in weight distribution.

Validation

IGATP Consistency by Review Count

- Comparison:
 - We tested whether the IGATP values change significantly when calculated on two review-count-based subsets:
 - Locations with ≥ 10 reviews
 - Locations with < 10 reviews
- Pearson correlations with full dataset:
 - ≥ 10 reviews: $r = 1.000$, $p < 0.001$
 - < 10 reviews: $r = 1.000$, $p < 0.001$
- Conclusion:
 - The IGATP index is fully consistent across subsets, confirming its internal reliability regardless of the number of reviews per location.

Validation

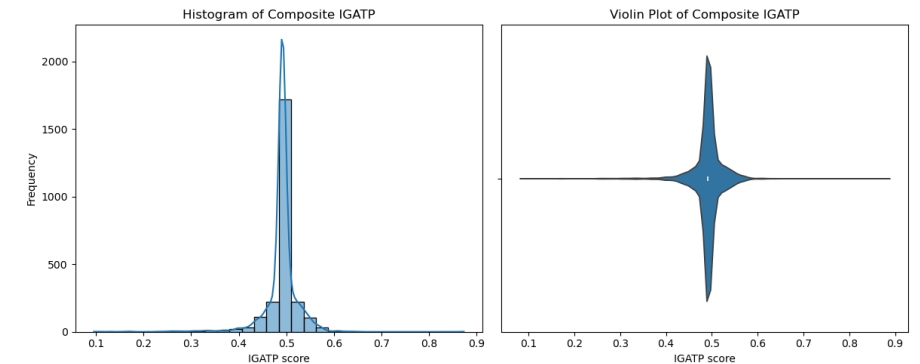
Distribution of Composite Index (IGATP)

Objective:

To examine the shape, central tendency, and dispersion of the IGATP scores.

Visualizations Used:

- **Histogram with KDE (left):**
 - Shows a strongly peaked distribution, slightly right-skewed, with most values concentrated around 0.48–0.52. There are a few mild outliers above 0.6 and below 0.4, indicating limited tail behavior.
- **Violin Plot (right):**
 - Reinforces the narrow concentration near the center and the low spread, confirming that the index is tightly centered with few extreme values.



Validation

Bootstrap Confidence Interval for Pearson's r

Objective: To quantify the uncertainty around the observed Pearson correlation between Rating_Bayes_norm and Sentiment_norm.

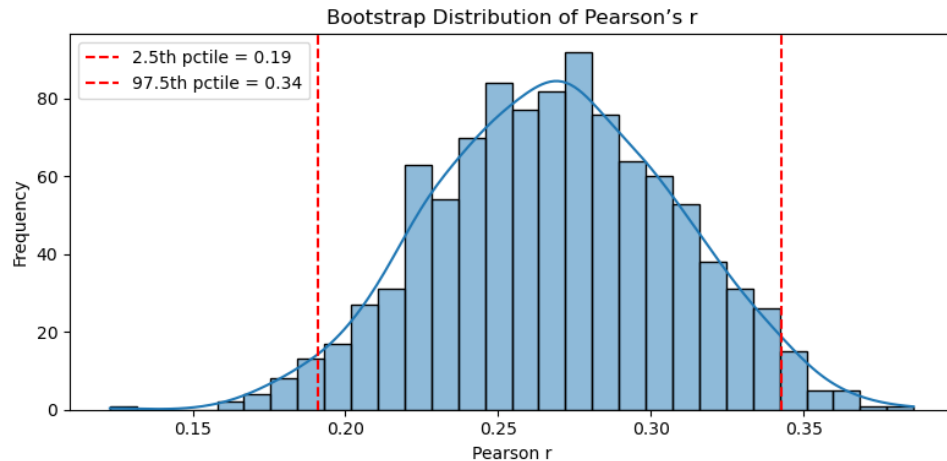
Methodology:

1,000 bootstrap resamples (with replacement)

Computed Pearson's r in each resample

Extracted 2.5th and 97.5th percentiles as non-parametric CI

Results: Observed r : 0.2795 Bootstrap CI: [0.19, 0.34]



Conclusion: The correlation is statistically significant and robust, but remains moderate in strength. Confirms that the two sub-indices capture complementary dimensions of perceived attractiveness.

Validation

Validation of Clustering Structure

Objective:

- To evaluate the quality of the clustering solution ($k = 7$, K-Medoids) based on the normalized sub-indices of the IGATP.

Methodology:

- Used Silhouette Score on feature matrix: Rating_Bayes_norm, Popularity_norm, Sentiment_norm
- Labels: cluster_k7_pam (K-Medoids, $k = 7$)

Result:

- Silhouette Score = 0.667

Interpretation:

- Score well above 0.6 \rightarrow indicates strong cluster cohesion and separation.
- Confirms the validity and interpretability of the 7-cluster K-Medoids structure.

Validation

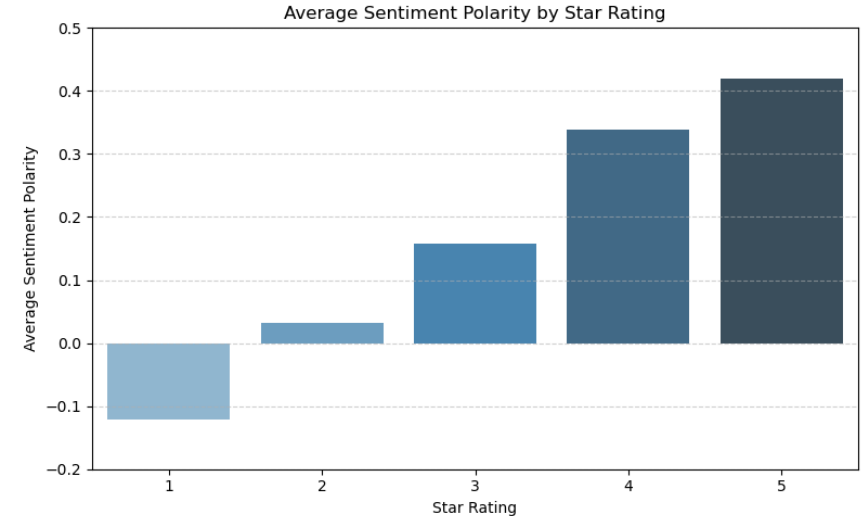
Objective:

To verify the alignment between NLP-based sentiment polarity and explicit user ratings (1★–5★).

Findings from the Plot:

- 1★ → clearly negative polarity (≈ -0.1)
- 2★ → near-neutral sentiment
- 3★ → moderate positivity
- 4★ → clearly positive
- 5★ → strongly positive ($\approx +0.42$)

Conclusion: The monotonic increase confirms that the sentiment polarity scale is consistent with user satisfaction. Validates sentiment polarity as a trustworthy component of the IGATP.



Spatial Analysis

Spatial Analysis of Points of Interest – Methodology

Objective:

To visualise the spatial distribution of the IGATP and its sub-indices (Rating_Bayes, Popularity, and Sentiment) within the Porto Metropolitan Area (AMP), using georeferenced points of interest.

Procedures:

- Removal of observations without coordinates (Latitude_Nova, Longitude_Nova).
- Conversion to a GeoDataFrame with point geometries (EPSG:4326).
- Application of a spatial join with the AMP municipal polygons to retain only valid points.
- Individual visualisation of each indicator based on value-driven colour mapping (cmap = "viridis") over municipal boundaries.

Spatial Analysis

Spatial Analysis of Points of Interest – Results

Limitations of Point-Based Visualization:

The high density of points in urban areas leads to visual overlap, making it difficult to perceive clear spatial patterns. Since many points are very close or nearly overlapping, point-based maps are not sufficiently informative to reveal significant territorial differences.



Spatial Analysis

Aggregated Spatial Analysis by Municipality – Methodology

Objective:

Evaluate the average distribution of the IGATP and its sub-indices (Rating_Bayes, Popularity, Sentiment) aggregated by municipality within the Porto Metropolitan Area (AMP).

Procedures:

- Performed a spatial join between each point of interest and its corresponding municipality (DICO_right).
- Calculated the mean score per municipality for each indicator.
- Merged results with the municipal shapefile (gdf_mun) using a left join.
- Created thematic maps with municipal-level averages, coloured using the viridis scale for visual contrast.

Spatial Analysis

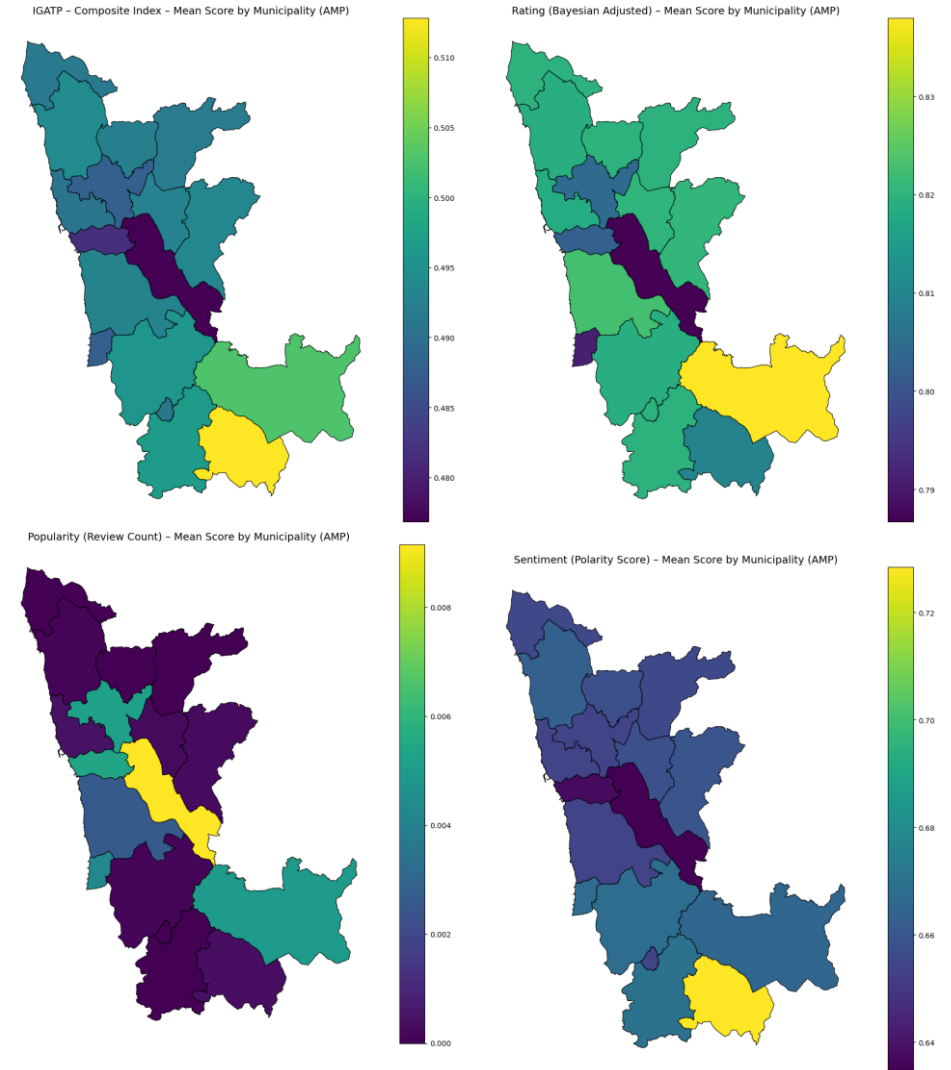
Aggregated Spatial Analysis by Municipality – Results

IGATP – Composite Index:

- Municipal averages show very little variation (≈ 0.48 – 0.51).
- Suggests a regionally balanced level of perceived attractiveness, with minor peaks in Arouca and Vale de Cambra.

Rating (Bayesian Adjusted):

- Also highly uniform across municipalities (≈ 0.79 – 0.83), reinforcing consistent user satisfaction.



Spatial Analysis

Aggregated Spatial Analysis by Municipality – Results

Popularity (Review Count):

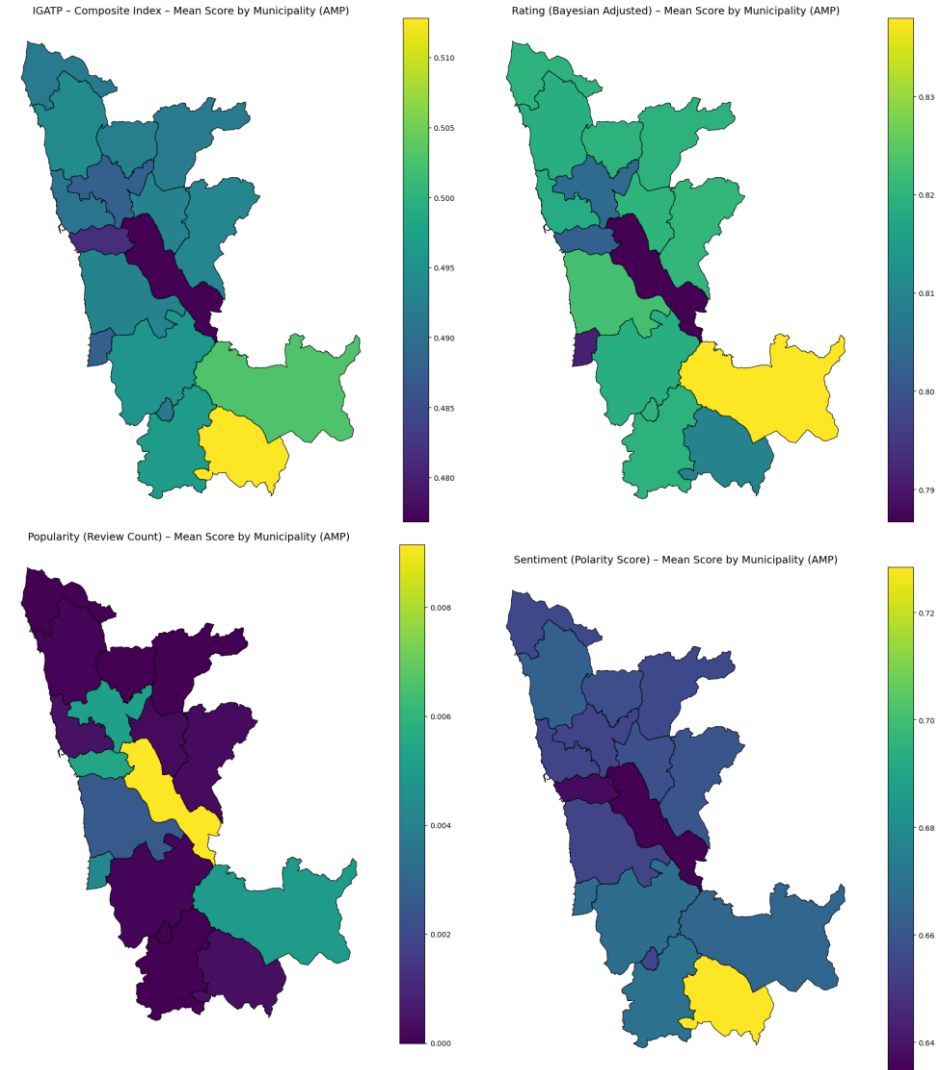
- Despite low absolute variation (≈ 0.000 – 0.009), the distribution is highly unequal:
 - A few central municipalities (e.g. Porto, Gaia) concentrate most reviews;
 - Many others show near-zero visibility.
- The map accentuates this relative disparity due to colour scaling.

Sentiment (Average Polarity Score):

- Slightly wider range than rating (≈ 0.64 – 0.72).
- More positive sentiment observed in the southern and inland municipalities.

Conclusion:

While most indicators show limited absolute variation, the Popularity sub-index reveals substantial disparities in digital engagement. This municipal-level analysis provides a baseline for the upcoming finer-grained analysis at the parish level.



Spatial Analysis

Aggregated Spatial Analysis by Parish – Methodology

Objective:

To analyse the spatial distribution of the IGATP and its sub-indices (Rating_Bayes, Popularity, and Sentiment) at the parish level within the Porto Metropolitan Area (AMP), allowing for finer-grained detection of local patterns not visible at the municipal scale.

Procedures:

- A spatial join was performed between the georeferenced points (gdf_points) and the parish polygons (gdf_freg) using the within predicate.
- For each parish (DICOFRE_le), the mean value of each index was calculated.
- These averages were then merged with the parish shapefile for spatial visualisation.
- Thematic maps were generated using a continuous colour scale (viridis) to facilitate comparative interpretation across parishes.

Note: Some parishes do not contain any georeferenced points, resulting in missing values (NaNs) for one or more indices. These gaps reflect either low digital representation or absence of reviewed points of interest.

Spatial Analysis

Aggregated Spatial Analysis by Parish – Results

IGATP – Global Index:

- Mean values range between ≈ 0.30 and 0.55.
- Greater heterogeneity is observed compared to the municipal scale, with some inland parishes and those in southern Arouca standing out positively.

Rating (Bayesian Adjusted):

- Generally positive distribution, with some more inland parishes showing average values above 0.90.
- Two parishes exhibit atypically low values (below 0.60).



Spatial Analysis

Aggregated Spatial Analysis by Parish – Results

Popularity (Review Count):

- Despite a narrow absolute range, the distribution is highly unequal:
 - A few parishes (notably in Porto and Vila Nova de Gaia) concentrate the majority of reviews.
 - Most parishes have near-zero values, reflecting very limited digital visibility.

Sentiment (Polarity Score):

- Wider range (≈ 0.40 to 0.80), enabling the identification of more extreme perceptions.
- Several southern parishes stand out for their positive emotional tone in reviews.

Conclusion:

Parish-level analysis allows for the detection of more detailed spatial patterns and highlights local inequalities in perceived attractiveness. The IGATP, when disaggregated, proves sensitive to intra-municipal diversity, which reinforces the relevance of using a finer-scale approach.



Spatial Analysis

Top Municipalities by Perceived Attractiveness (IGATP)

Interpretation:

- The highest-ranked municipalities are mostly peripheral within the AMP.
- The strong performance of Vale de Cambra and Arouca likely reflects a combination of good ratings and more positive perceived sentiment.
- These results suggest that perceived attractiveness is not concentrated in major urban centers, but may instead emerge in less central territories.

Top 5 Municipalities – Mean Global Attractiveness Index (IGATP):

Municipality	Global Index Mean
Vale de Cambra	0.513
Arouca	0.503
Oliveira de Azeméis	0.497
Santa Maria da Feira	0.496
Vila do Conde	0.494

Spatial Analysis

Bottom Municipalities by Perceived Attractiveness

Interpretation:

- Larger and denser municipalities like Porto, Gondomar, and Matosinhos appear among the lowest-ranked.
- Despite hosting many points of interest, these areas may be affected by overexposure, lower sentiment, or greater rating variability.
- The difference between top and bottom performers is modest (≈ 0.04), reinforcing the overall consistency of the index at the municipal level.

Bottom 5 Municipalities – Mean IGATP:

Municipality	Global Index Mean
Matosinhos	0.491
Maia	0.488
Espinho	0.488
Porto	0.482
Gondomar	0.477

Spatial Analysis

Parish-Level Ranking: Local Disparities

Interpretation:

- Lower-scoring parishes include dense urban areas such as Bonfim (Porto) and suburban zones like Pedrouços (Maia).
- Pedrouços, with the lowest index (0.273), represents an extreme case, possibly associated with a small number of reviews, poor sentiment, or low digital presence.
- The intra-municipal disparities observed at the parish level highlight the importance of spatially disaggregated analysis, where more granular patterns of perceived attractiveness emerge.

Top 5 Parishes – Mean IGATP:

Parish	Global Index Mean
Chave	0.564
Urrô	0.549
Folgosa	0.542
São Roque	0.538
Alvarenga	0.535

Bottom 5 Parishes:

Parish	Global Index Mean
Bonfim	0.450
União das Freguesias de Alvarelhos e Guidões	0.446
União das Freguesias de Pedroso e Seixezelo	0.418
Aguiar de Sousa	0.401
Pedrouços	0.273

Spatial Analysis

Global Spatial Autocorrelation – Moran's I (Methodology)

Objective:

To evaluate the presence of global spatial autocorrelation patterns in the IGATP (Global Perceived Attractiveness Index), at both the parish and municipality levels.

Steps Performed:

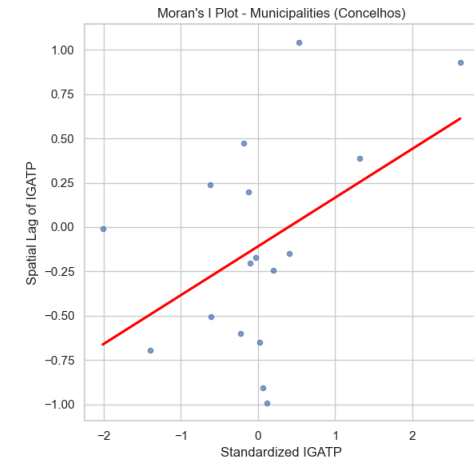
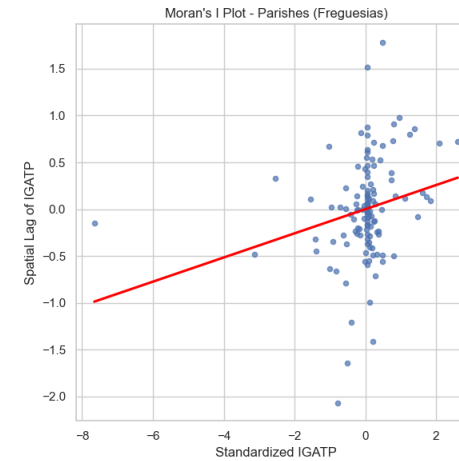
- **Data Cleaning:**
 - Observations with missing IGATP_mean values were removed.
- **Spatial Weights Matrix:**
 - A Queen contiguity matrix was created (`Queen.from_dataframe()`), considering neighboring units that share either borders or vertices.
- **Moran's I Calculation:**
 - The global Moran's I coefficient and significance (p-value) were computed using the `Moran()` function from the `esda` module, based on random permutations.
- **Variable Standardization:**
 - IGATP values were standardized (z-scores) to allow scatterplot representation.
- **Visualization:**
 - A Moran Scatterplot was built to visualize the relationship between standardized IGATP values and their spatial lag (average value among neighbors).
 - A red regression line indicates the strength and direction of the spatial autocorrelation.

Spatial Analysis

Spatial Autocorrelation Analysis – Moran's I

Interpretation:

- The results confirm the presence of a spatial structure in the IGATP:
 - At the parish level, the spatial dependence is weaker, yet still statistically significant;
 - At the municipal level, the pattern is clearer and more pronounced.
- This suggests that neighboring territories tend to share similar levels of perceived attractiveness, validating the use of spatial approaches in this type of analysis.
- The upward trend in both Moran's scatterplots (red line) visually confirms the presence of positive spatial dependence.



Spatial Autocorrelation Results:

Level	Moran's I	p-value	Interpretation
Parishes	0.1301	0.0200	Weak but statistically significant positive spatial autocorrelation
Municipalities	0.2751	0.0110	Moderate positive spatial autocorrelation , also significant

Spatial Analysis

Spatial Cluster Mapping – LISA Methodology

Objective:

- To identify significant local clusters of perceived attractiveness (IGATP) using Local Indicators of Spatial Association – Local Moran (LISA).

Procedures:

- A Queen contiguity spatial weights matrix was created;
- The Local Moran's I was applied to the average IGATP at the parish and municipality levels;
- Statistical significance was assessed using 999 random permutations ($\alpha = 0.05$);
- Spatial units were classified into five categories:

Category	Description
High-High	High value surrounded by high values (positive cluster)
Low-Low	Low value surrounded by low values (negative cluster)
High-Low	High value surrounded by low values (positive spatial outlier)
Low-High	Low value surrounded by high values (negative spatial outlier)
Not Significant	No statistically significant spatial autocorrelation

Spatial Analysis


LISA Results – Interpretation of Clusters

Parishes:

- High-High clusters are concentrated in the southern part of the AMP, indicating areas with consistently high attractiveness;
- Low-Low clusters are found in more peripheral zones and some urban centers;
- The presence of High-Low and Low-High configurations reveals spatial outliers, where attractiveness levels contrast with surrounding areas;
- Most parishes show no significant spatial autocorrelation, reinforcing the intra-municipal heterogeneity of the index.

Municipalities:

- High-High: detected in southern AMP, highlighting regions with spatially contiguous positive attractiveness;
- High-Low: municipalities like Vila Nova de Gaia, where high IGATP values contrast with neighboring low values — suggesting local spatial polarization;
- Low-High: municipalities with lower attractiveness surrounded by higher-scoring areas — indicating potential targets for policy intervention (Porto, Maia);
- The presence of clear spatial patterns suggests that territorial promotion strategies should take into account regional dynamics rather than isolated contexts.

 *Note: Some identified clusters, particularly in peripheral areas, may be partially influenced by limited data availability, especially due to the scarcity of user-generated reviews or content in those regions. Interpretation should therefore account for potential sampling bias.*

Visualization

Purpose and Technologies Used

What is IGATP?

We developed an interactive dashboard to compute and visualize the IGATP – Global Index of Perceived Touristic Attractiveness in the Porto Metropolitan Area (AMP).

The index combines three **dimensions** from Google Maps:

- Bayesian Rating
- Popularity (number of reviews)
- Sentiment (review polarity)

Technologies used:

- Python (pandas, geopandas, scikit-learn, Altair)
- LDA Topic Modeling and PCA + K-Medoids clustering
- Kepler.gl and Streamlit for geospatial visualization and interactive interface



Visualization

Point Map – Locations with IGATP

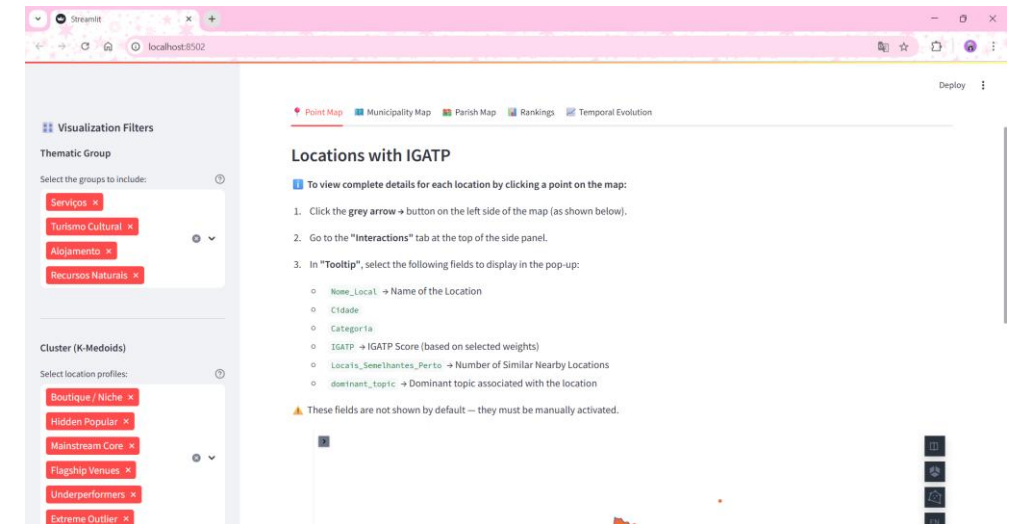
Displays all georeferenced touristic locations in the AMP.

Points can be filtered by thematic group and tourist profile (cluster).

Users can click the left-side arrow on the map to activate pop-ups showing:

- Name, category, city, IGATP score
- Number of similar nearby locations
- Dominant topic for each place

Filters and index weights are fully adjustable in the sidebar.



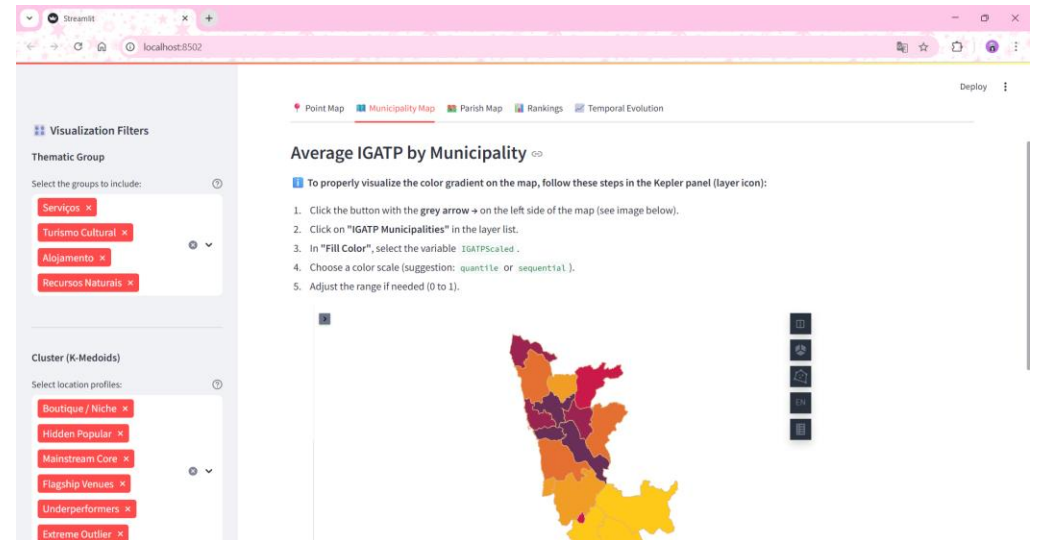
Visualization

Average IGATP by Municipality

Heatmap by municipality, showing the average IGATP based on the locations within each municipality.

Users can activate the layer in Kepler and customize the color gradient (e.g., quantile, sequential).

Helps identify municipalities with higher or lower perceived attractiveness.



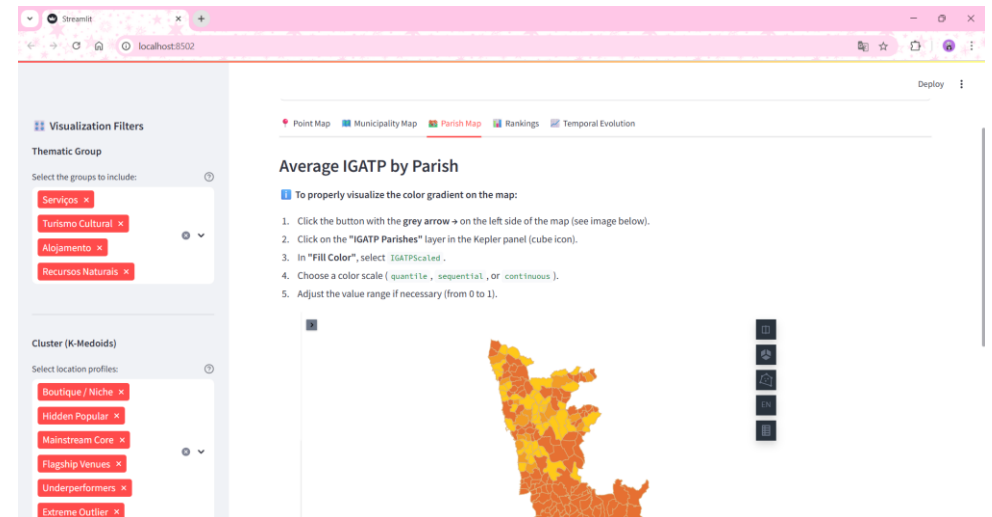
Visualization

Average IGATP by Parish

More detailed view of the index at the parish level.

Based on a pre-processed dataset with average IGATP values per parish.

Ideal for assessing the local variation in perceived attractiveness, relevant for local planning and public policy.



Visualization

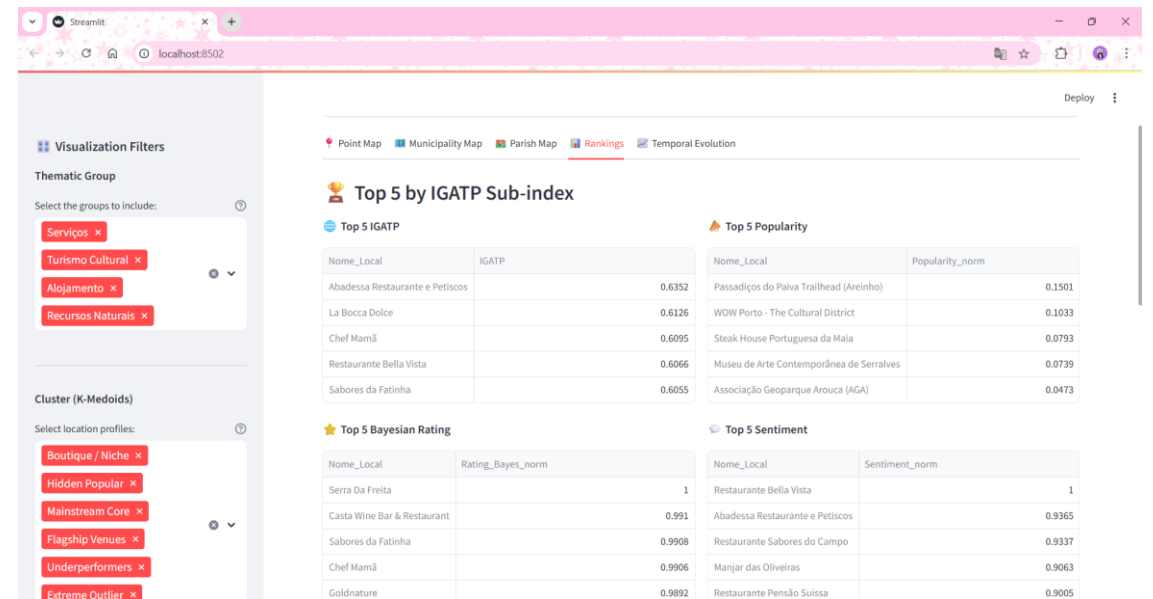
Touristic Attractiveness Rankings

Displays the Top 5 locations for each sub-index and for the overall IGATP.

Also includes:

- Municipalities with highest and lowest average IGATP
- Parishes with highest and lowest average IGATP

Provides a quick overview of highlights and low-performing areas in the region.



Visualization

Sentiment Trend Over Time

Line chart showing the monthly evolution of average review sentiment.

Helps identify:

- Whether user perception is improving or worsening
- Seasonal trends or event-related variations

Serves as a starting point for longitudinal analysis or strategic planning.

