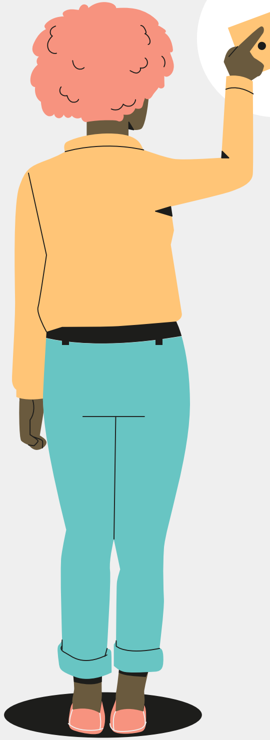


Airbnb Insights

Unveiling Host Experience

Group 06
Diogo Faria: up201907014
Sérgio Gama: up201906690
Tiago Rodrigues: up201906807
Valentina Wu: up201907483

Subject Description

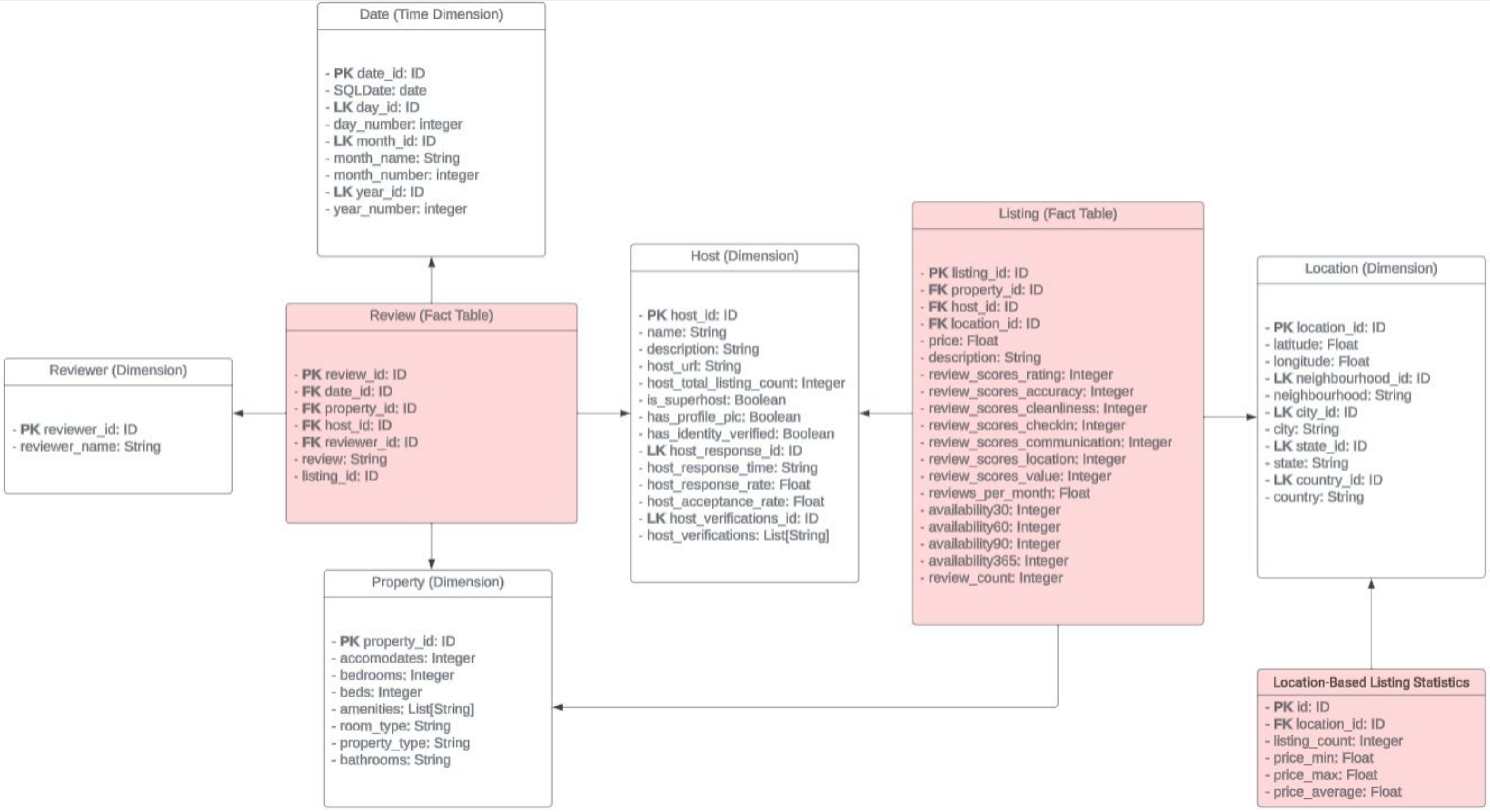
- Design and implement a data warehouse for Airbnb Reviews;
- Analysis of users' experience of a property within the Airbnb platform
- Dataset: City of Porto, Portugal
  - Necessary abundance of data;
  - Not become excessive;
- Dataset: Details
  - Around 745,000 facts related to reviews;
  - 12,818 listings
  - Details of property listed
  - Score assigned
  - Location

Planning

- **Dimensional Bus Matrix:**

Facts\Dimensions	Location	Date	Host	Property	Reviewer
Review		x	x	x	x
Listing	x		x	x	
Location-Based Listing Statistics	x				

Dimensional data model



Date (Time Dimension)

- **PK** date_id: ID
- SQLDate: date
- **LK** day_id: ID
- day_number: integer
- **LK** month_id: ID
- month_name: String
- month_number: integer
- **LK** year_id: ID
- year_number: integer

Reviewer (Dimension)

- **PK** reviewer_id: ID
- reviewer_name: String

Review (Fact Table)

- **PK** review_id: ID
- **FK** date_id: ID
- **FK** property_id: ID
- **FK** host_id: ID
- **FK** reviewer_id: ID
- review: String
- listing_id: ID

Property (Dimension)

- **PK** property_id: ID
- accomodates: Integer
- bedrooms: Integer
- beds: Integer
- amenities: List[String]
- room_type: String
- property_type: String
- bathrooms: String

Host (Dimension)

- **PK** host_id: ID
- name: String
- description: String
- host_url: String
- host_total_listing_count: Integer
- is_superhost: Boolean
- has_profile_pic: Boolean
- has_identity_verified: Boolean
- **LK** host_response_id: ID
- host_response_time: String
- host_response_rate: Float
- host_acceptance_rate: Float
- **LK** host_verifications_id: ID
- host_verifications: List[String]

Listing (Fact Table)

- **PK** listing_id: ID
- **FK** property_id: ID
- **FK** host_id: ID
- **FK** location_id: ID
- price: Float
- description: String
- review_scores_rating: Integer
- review_scores_accuracy: Integer
- review_scores_cleanliness: Integer
- review_scores_checkin: Integer
- review_scores_communication: Integer
- review_scores_location: Integer
- review_scores_value: Integer
- reviews_per_month: Float
- availability30: Integer
- availability60: Integer
- availability90: Integer
- availability365: Integer
- review_count: Integer

Location (Dimension)

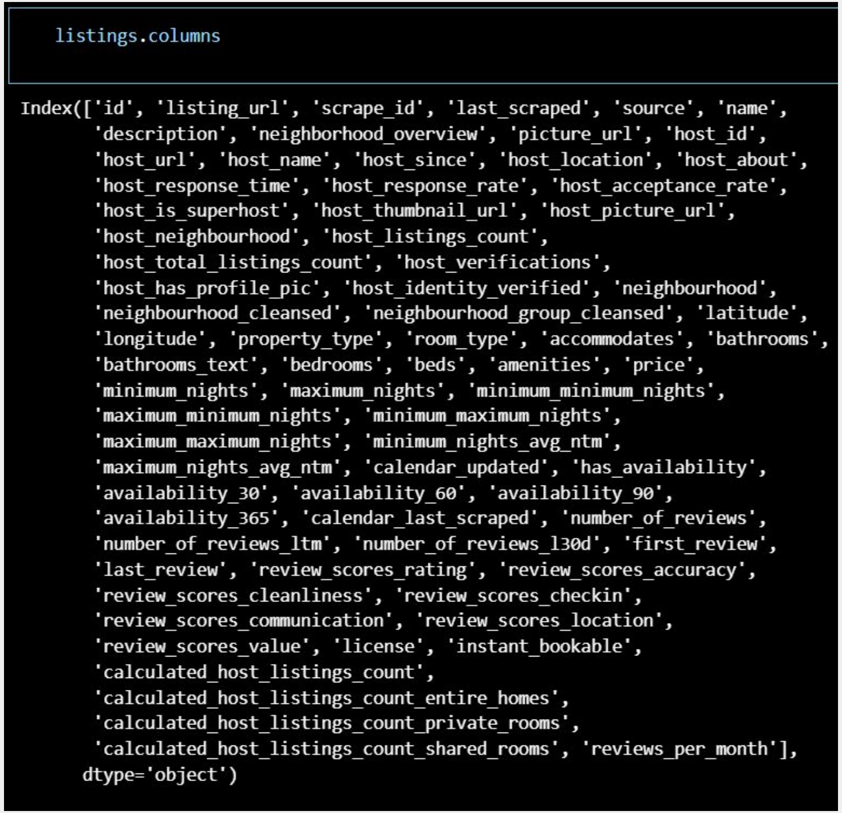
- **PK** location_id: ID
- latitude: Float
- longitude: Float
- **LK** neighbourhood_id: ID
- neighbourhood: String
- **LK** city_id: ID
- city: String
- **LK** state_id: ID
- state: String
- **LK** country_id: ID
- country: String

Location-Based Listing Statistics

- **PK** id: ID
- **FK** location_id: ID
- listing_count: Integer
- price_min: Float
- price_max: Float
- price_average: Float

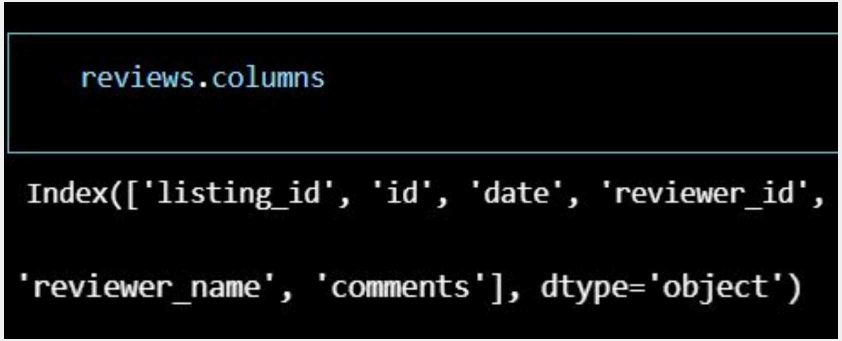
Dataset - Extraction

opendatasoft



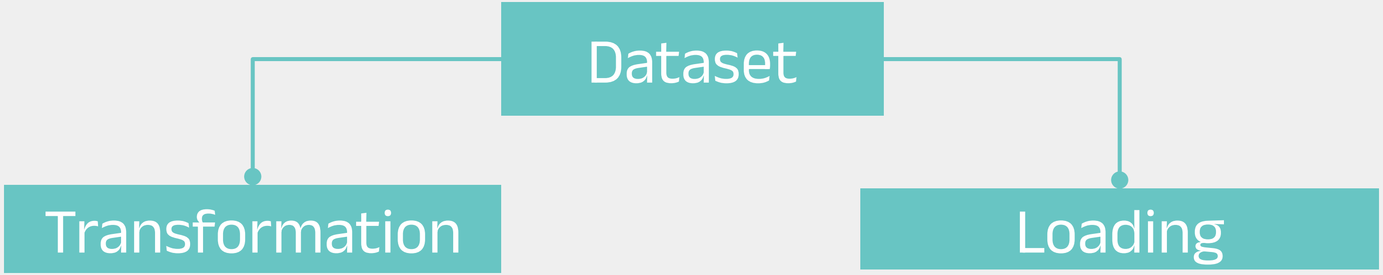
```
listings.columns
```

```
Index(['id', 'listing_url', 'scrape_id', 'last_scraped', 'source', 'name',
       'description', 'neighborhood_overview', 'picture_url', 'host_id',
       'host_url', 'host_name', 'host_since', 'host_location', 'host_about',
       'host_response_time', 'host_response_rate', 'host_acceptance_rate',
       'host_is_superhost', 'host_thumbnail_url', 'host_picture_url',
       'host_neighbourhood', 'host_listings_count',
       'host_total_listings_count', 'host_verifications',
       'host_has_profile_pic', 'host_identity_verified', 'neighbourhood',
       'neighbourhood_cleansed', 'neighbourhood_group_cleansed', 'latitude',
       'longitude', 'property_type', 'room_type', 'accommodates', 'bathrooms',
       'bathrooms_text', 'bedrooms', 'beds', 'amenities', 'price',
       'minimum_nights', 'maximum_nights', 'minimum_minimum_nights',
       'maximum_minimum_nights', 'minimum_maximum_nights',
       'maximum_maximum_nights', 'minimum_nights_avg_ntm',
       'maximum_nights_avg_ntm', 'calendar_updated', 'has_availability',
       'availability_30', 'availability_60', 'availability_90',
       'availability_365', 'calendar_last_scraped', 'number_of_reviews',
       'number_of_reviews_ltm', 'number_of_reviews_l30d', 'first_review',
       'last_review', 'review_scores_rating', 'review_scores_accuracy',
       'review_scores_cleanliness', 'review_scores_checkin',
       'review_scores_communication', 'review_scores_location',
       'review_scores_value', 'license', 'instant_bookable',
       'calculated_host_listings_count',
       'calculated_host_listings_count_entire_homes',
       'calculated_host_listings_count_private_rooms',
       'calculated_host_listings_count_shared_rooms', 'reviews_per_month'],
      dtype='object')
```



```
reviews.columns
```

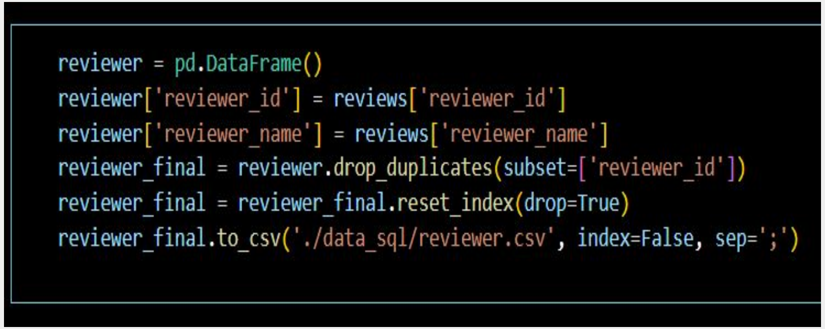
```
Index(['listing_id', 'id', 'date', 'reviewer_id',

'reviewer_name', 'comments'], dtype='object')
```



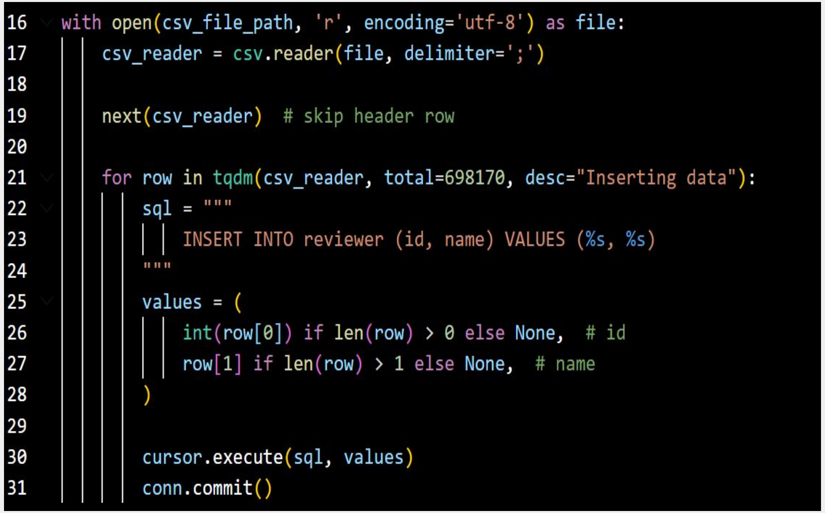
Dataset

Transformation

Loading



```python
reviewer = pd.DataFrame()
reviewer['reviewer_id'] = reviews['reviewer_id']
reviewer['reviewer_name'] = reviews['reviewer_name']
reviewer_final = reviewer.drop_duplicates(subset=['reviewer_id'])
reviewer_final = reviewer_final.reset_index(drop=True)
reviewer_final.to_csv('./data_sql/reviewer.csv', index=False, sep=';')
```



```python
with open(csv_file_path, 'r', encoding='utf-8') as file:
    csv_reader = csv.reader(file, delimiter=';')

    next(csv_reader)  # skip header row

    for row in tqdm(csv_reader, total=698170, desc="Inserting data"):
        sql = """
            INSERT INTO reviewer (id, name) VALUES (%s, %s)
        """
        values = (
            int(row[0]) if len(row) > 0 else None,  # id
            row[1] if len(row) > 1 else None,  # name
        )

        cursor.execute(sql, values)
        conn.commit()
```

Queries

01

Top Hosts

Find the top hosts based on the total number of listings they have

02

Review Count

Number of reviews for each combination of ‘property_id’, ‘host_id’ and ‘reviewer_id’

03

Most Reviews

Listings with the Most Reviews in each Neighbourhood

04

Review Rankings

Ranking listings by Review Scores within each neighbourhood

Queries

05

Price Analysis

Identifying Listings with Prices Above the Neighbourhood Average

06

Response Analysis

Analyze host response rates and their impact on listing popularity

07

Monthly Analysis

Analyze monthly review trends for each city, with subtotals for each year, city , state and grand total

08

Geographical Distribution

Explore the geographical distribution of listings across cities and countries

Queries

09

Top Amenities

Query to identify the top individuals amenities, considering each amenity as a separate entity in the list

10

Average Accommodates

Query to calculate the average accommodates for listings based on combinations of amenities

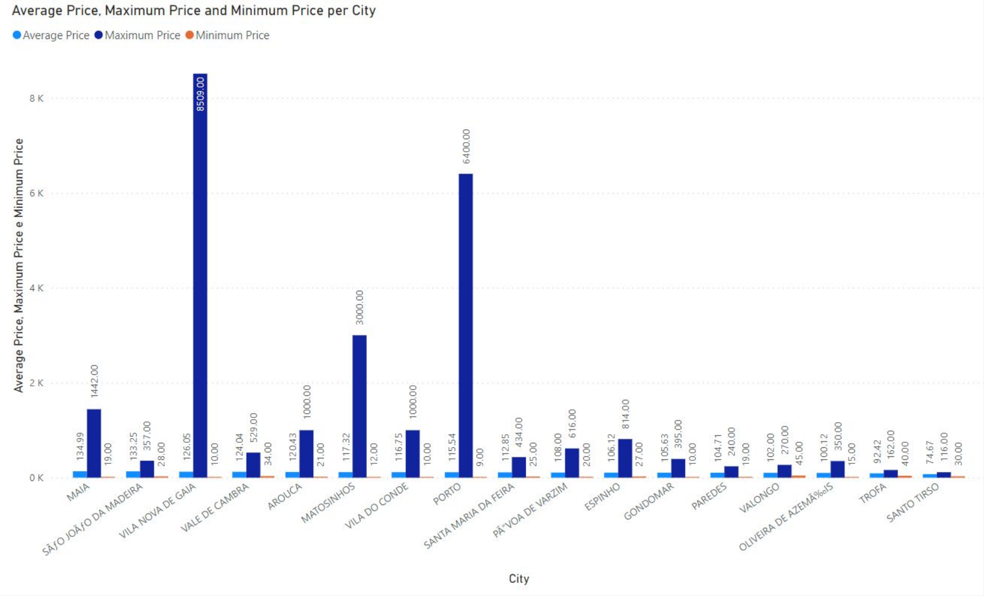
11

High Scores

Query to identify hosts whose listings consistently receive high review scores

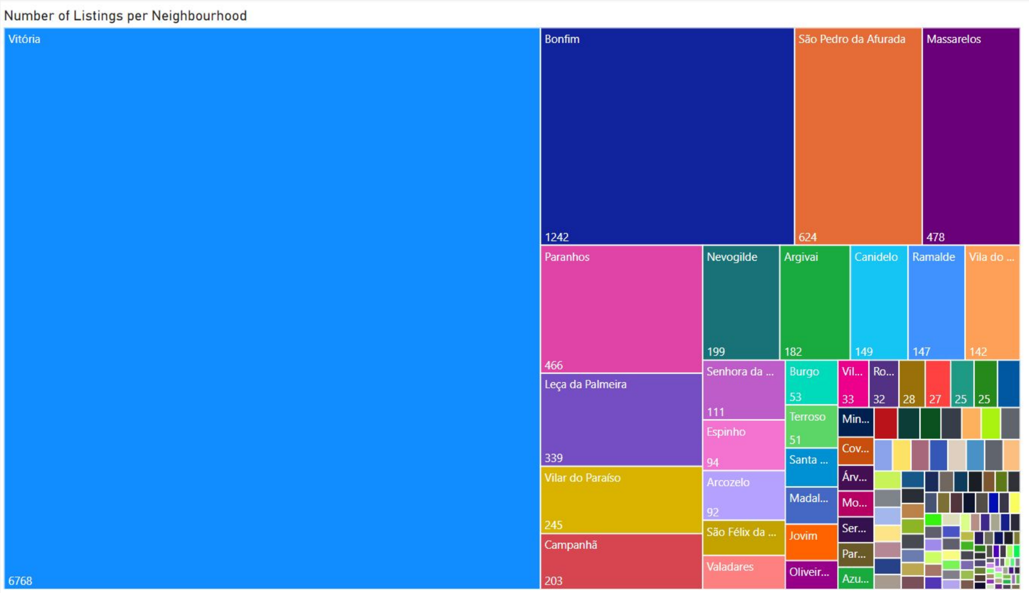
Data Analysis

- The price ranges in the cities our data has, coming directly from our aggregated fact table;
- City with the Highest Price: “Vila Nova de Gaia”;
- City with the Average Highest Price: “Maia”.



Data Analysis

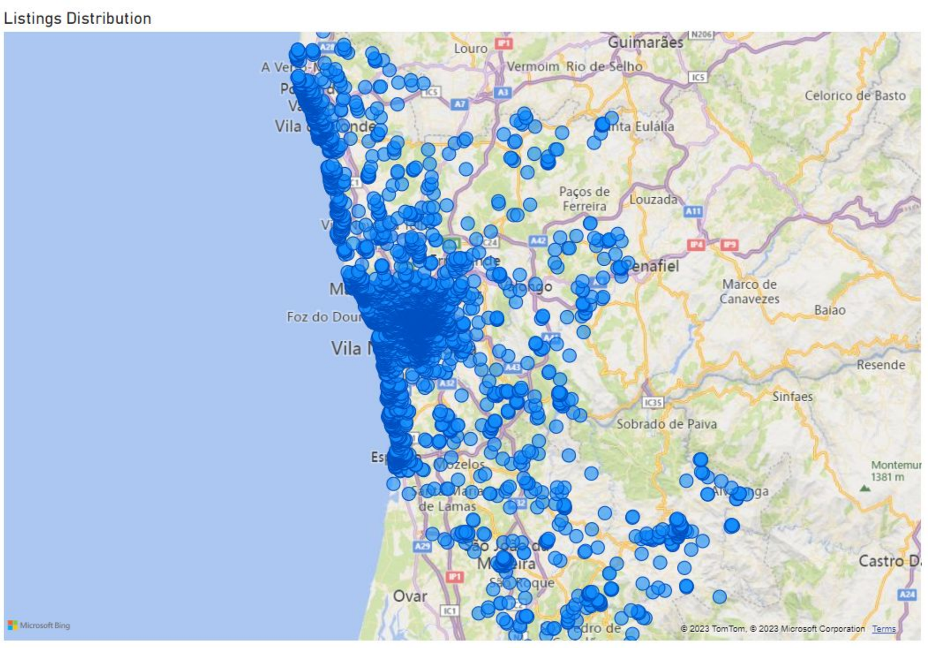
- The number of listings in the various neighbourhoods present in our data;
- “Vitória” seems to account for around half of our total Airbnb listings, probably indicating that it is a very enticing place for tourists.



Number of Listings per Neighbourhood

Data Analysis

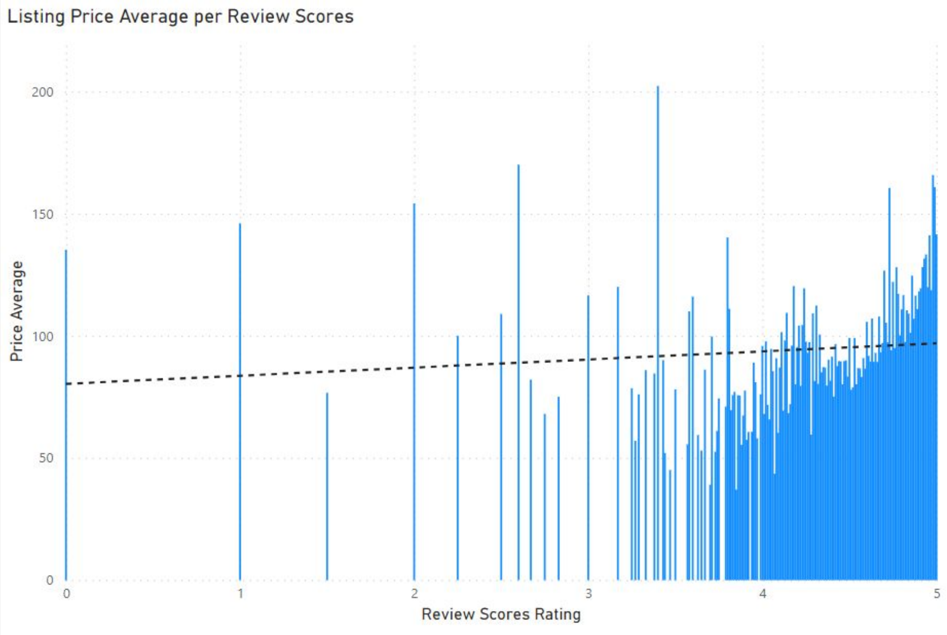
- The distributions of listings around the world, though our data set is limited to a smaller region;
- Certain zones have a much higher density of listings when compared to others.



Listings Distribution

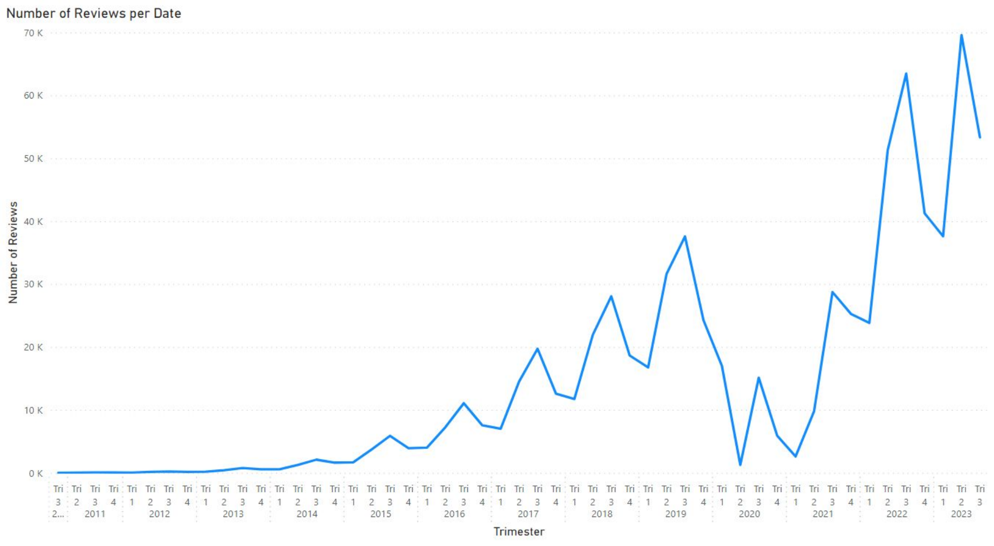
Data Analysis

- The average price of listings per review scores;
- Is there a correlation between prices and ratings?
- Yes, but probably not as much as one would expect.



Data Analysis

- A continuing trend of the peak of the year being around the 3rd quarter of the year (Summer);
- A couple of years the number of reviews went down instead of up, most likely due to COVID;
- The number of reviews has been steadily increasing (aside from COVID), meaning there are more listings and/or guests as the years have gone by.



Critical reflection about the advantages and shortcomes

Advantages

Detailed and Granular Analysis

Support in-depth exploration and analysis of the dataset, with a lot of dimensions.

Hierarchical Representation for Geospatial Insights

We use a hierarchical structure for comprehensive geospatial analysis.

Shortcomes

Denormalized Structure

Potential issues with data redundancy due to the denormalized model structure.

Limited Support for Real-time Updates

Inherent focus on analytical processing, limiting support for real-time and dynamic updates.

Thank You!
Any Question?