

**airbnb**



|  |  |  |
| --- | --- | --- |
|  |  |  |
| **Sameer Afzal** | **Harsh Mehta** | **Asad Aslam** |
| **1000053142** | **1000055155** | **1000053142** |

**Presented To**

**Prof. Orazio Tomarchio**

**Data Base & Big Data Analytics**

**Contents**

[**1.** **Introduction** 3](#_Toc136968760)

[**2.** **Starting Dataset** 3](#_Toc136968761)

[**3.** **Business Questions** 4](#_Toc136968762)

[**4.** **Dimensional Fact Model** 4](#_Toc136968763)

[**5.** **ETL Process** 6](#_Toc136968764)

[**6.** **Dashboard** 7](#_Toc136968765)

[**6.1.** **Main Dashboard** 7](#_Toc136968766)

[**6.2.** **Price – Host – City Dashboard** 8](#_Toc136968767)

[**6.3.** **Room Type Analysis.** 9](#_Toc136968768)

[**6.4.** **Timelines.** 10](#_Toc136968769)

[**6.5.** **Listings in Sicily.** 11](#_Toc136968770)

[**7.** **Conclusion** 12](#_Toc136968771)

[**8.** **Color blindness test** 15](#_Toc136968772)

[**8.1.** **Red Weak test** 15](#_Toc136968773)

[**8.2.** **Red Blind test** 16](#_Toc136968774)

[**8.3.** **Blue Weak test** 16](#_Toc136968775)

[**8.4.** **Blue Blind test** 17](#_Toc136968776)

[**8.5.** **Green Weak test** 18](#_Toc136968777)

[**8.6.** **Green Blind test** 19](#_Toc136968778)

[**8.7.** **Monochromic Test** 19](#_Toc136968779)

# **Introduction**

Airbnb has revolutionized the way people travel and find accommodations worldwide, and the beautiful island of Sicily in Italy is no exception. With its stunning landscapes, rich history, and vibrant culture, Sicily has become an increasingly popular tourist destination, attracting diverse travelers. As a result, Airbnb has emerged as a prominent platform for finding unique and affordable lodging options across the island. This dataset encompasses a wide range of information, including property listings, Cities, guest reviews, and more. By exploring this data, we can delve into various aspects of the Airbnb ecosystem in Sicily, such as popular destinations, and pricing dynamics.

# **Starting Dataset**

The report is based on the inside Airbnb dataset for the Sicily region. The data is extracted from [Insideairbnb](http://insideairbnb.com/sicily). We have extracted four tables:

**Listing Table**



**Calendar Table**



**Reviews Table**



**Neighborhood Table**



# **Business Questions**

The report aims to analyses the trend between listings, reviews, and calendars of the properties listed in the region of Sicily, Italy. The business questions, which the report aims to answer, are:

* How does the average price & count of review differ for different room types within a specific city?
* What is the distribution of room type in different cities with respect to average price & reviews?
* Which month experience the highest demand booked listing for each room type?
* What are the seasonal trends in booking behavior?
* Where does one or a few cities dominate the market in terms of listing percentage?
* What is the trend in the number of reviews over the past 10 years and how has it been impacted by covid-19 pandemic?

# **Dimensional Fact Model**

The dimensional fact model (DFM) is an ad hoc and graphical formalism specifically devised to support the conceptual modelling phase in a DW project. DFM is extremely intuitive and can be used by analysts and non-technical users as well. Thus, the DFM is a graphical conceptual model, specifically. devised for multidimensional design, to:

• lend effective support to conceptual design

• create an environment in which user queries may be formulated intuitively

• make communication possible between designers and end users with the goal of formalizing requirement specifications

• build a stable platform for the logical design

• provide clear and expressive design documentation

The conceptual representation generated by the DFM consists of a set of fact schemata. Fact schemata model facts, measures, dimensions, and hierarchies. Besides these basic elements, the DFM includes a large set of constructs for expressing the multitude of conceptual nuances that characterize actual. Modeling scenarios in projects of small to large complexity. A fact is a concept relevant to decision-making processes and represents the core of the information to analyze. In the considered dataset, we identified a fact:

**Listing:** The numerical aspects of the fact, which are relevant to the analysis, or measures, are id, name, and price. Its dimensions, namely the analysis axes of the fact in the multidimensional paradigm, are Host, reviews, Booking calendar, and location.

For **Reviews and Locations**, we define a hierarchy, which means that a tree is embedded in each dimension, with these functional dependencies.

Neighbourhood → Longitude

Neighbourhood → Latitude

Neighborhood → Neighbourhood Group

Number of Reviews → Reviews per month

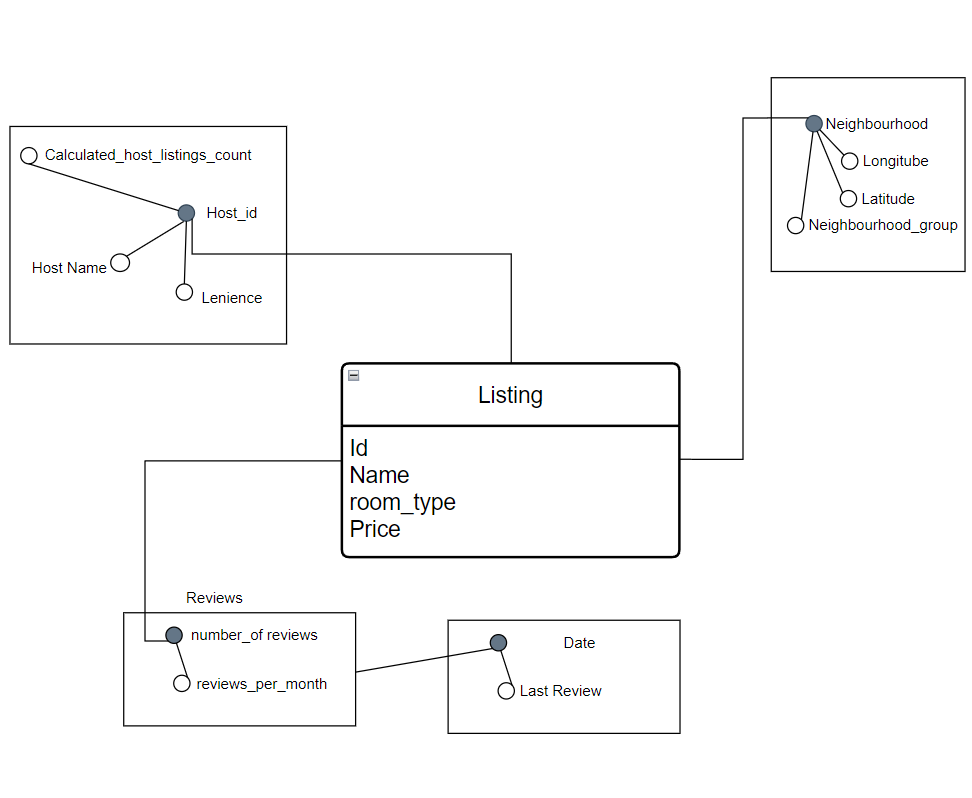
Date → Last Review

Host ID → Calculated\_host\_listing\_count

The **Host** dimensional attribute has the following descriptive attributes (i.e. not used for aggregation in the hierarchy):

License

Host Name



# **ETL Process**

Before proceeding with the analysis, there are 3 main steps to prepare the data: Extraction, Transformation and Loading, plus another step called Cleansing (or cleaning) that comes after Extraction (in fact 4 steps)

**Extraction**: Aimed at collecting data from sources (also called Integration when several heterogeneous sources exist). In this report, the data has been extracted from the Kaggle website.

**Cleansing (or cleaning):** Aimed at fixing incorrect, incomplete, duplicate, or otherwise erroneous data in a data set. It involves identifying data errors and then changing, updating, or removing data to correct them.

**Transformation**: Aimed at adjusting the format of data from source schemata to reconcile schemata, useful for Data Warehouse (DWH) analysis.

**Loading**: Aimed at entering data into the DW/Data Marts and updating the already existing data.

In this report, Tableau Prep Builder has been used for cleaning, transformation and loading. For our project, on Tableau Prep Builder, 4 input tables were inserted and 3 sheets were output in different excel sheets: Listing, Reviews and Calendar. The four one is neighborhood contains two columns: neighborhood and neighborhood \_group. However, the neighborhood \_group column consists entirely of null values. Considering this, it has been decided to exclude this file from the ELT process as it does not contribute any meaningful information to the project. By excluding the file, the focus remains on the relevant data, ensuring a streamlined ELT workflow that maximizes efficiency and effectiveness. The report describes how each of the three is obtained.

**Listing**

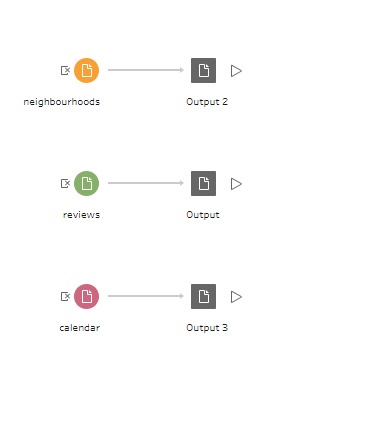
The ELT workflow in Tableau Prep for the 'listing.csv' file involves cleaning and transforming the data to optimize it for analysis. This includes removing unnecessary fields such as autogenerated row numbers, the neighborhood group column, and the last review column. Empty values in the BNB's name and host's name are replaced with 'Unspecified,' while null values in the count of reviews per month are replaced with 0. The License column, containing irrelevant data, is also removed. Additionally, the reviews per month values are rounded off for better presentation and analysis. These steps ensure a streamlined dataset ready for further exploration and visualization in Tableau.

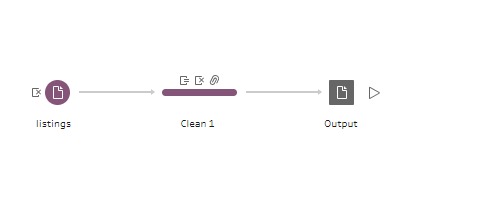
**Calendar**

The workflow for the calendar.csv table involves data cleaning and sampling techniques to optimize processing for a large dataset. Firstly, the auto-generated field for source row numbers is removed to ensure a cleaner dataset. Additionally, due to the large data size, a sampling technique is applied to collect a subset of 100,000 rows, allowing for quicker processing and analysis. This approach strikes a balance between representative data and efficient handling of the dataset, facilitating smoother exploration and visualization of the calendar data.

**Reviews**

The workflow for the reviews.csv table focuses on data cleaning and optimization. As a first step, the auto-generated field for source row numbers is removed, ensuring a cleaner dataset.





# **Dashboard**

In order to answer the business questions, an explanatory dashboard was created for each of the three main topics, and they were organized within a story. All of them comply with the emphasis guidelines.

## **Main Dashboard**

The Main Dashboard provides a comprehensive summary of data related to listings, hosts, reviews, and cities in a convenient and easy-to-understand format. Here are the key insights it offers:

**Total number of listings:** This indicates the overall count of available listings in the dataset, representing the number of properties that can be rented.

**Total number of hosts:** This represents the number of individuals or entities who are offering their properties for rent.

**Total number of reviews:** This indicates the cumulative number of reviews received by all the listings. Reviews are important for potential guests to assess the quality and reliability of a property.

**Price range:** The dashboard presents the minimum and maximum prices among all the listings, giving an overview of the price distribution.

**Top five cities with highest average price:** This insight identifies the five cities where the average price per night is the highest, considering different room types (e.g., entire homes, private rooms). It helps users understand which cities tend to have higher-priced accommodations.

**Top five cities with highest number of reviews:** This highlights the five cities that have received the most reviews across different room types. This information can indicate popular destinations or cities that attract a high volume of visitors.

**Top five cities with greatest number of listings:** This insight displays the five cities that have the highest number of available listings.

**Demanded forecast:** The dashboard provides insights into the upcoming demand for rentals. For example, it may indicate that the upcoming month of August is expected to have a very high demand.

A screenshot of a computer

Description automatically generated with low confidence

## **Price – Host – City Dashboard**

The price-Host-City analysis dashboard provides a detailed comparison of various aspects related to the correlation between Price, Host, and Cities. Here are the different columns and their respective analyses:

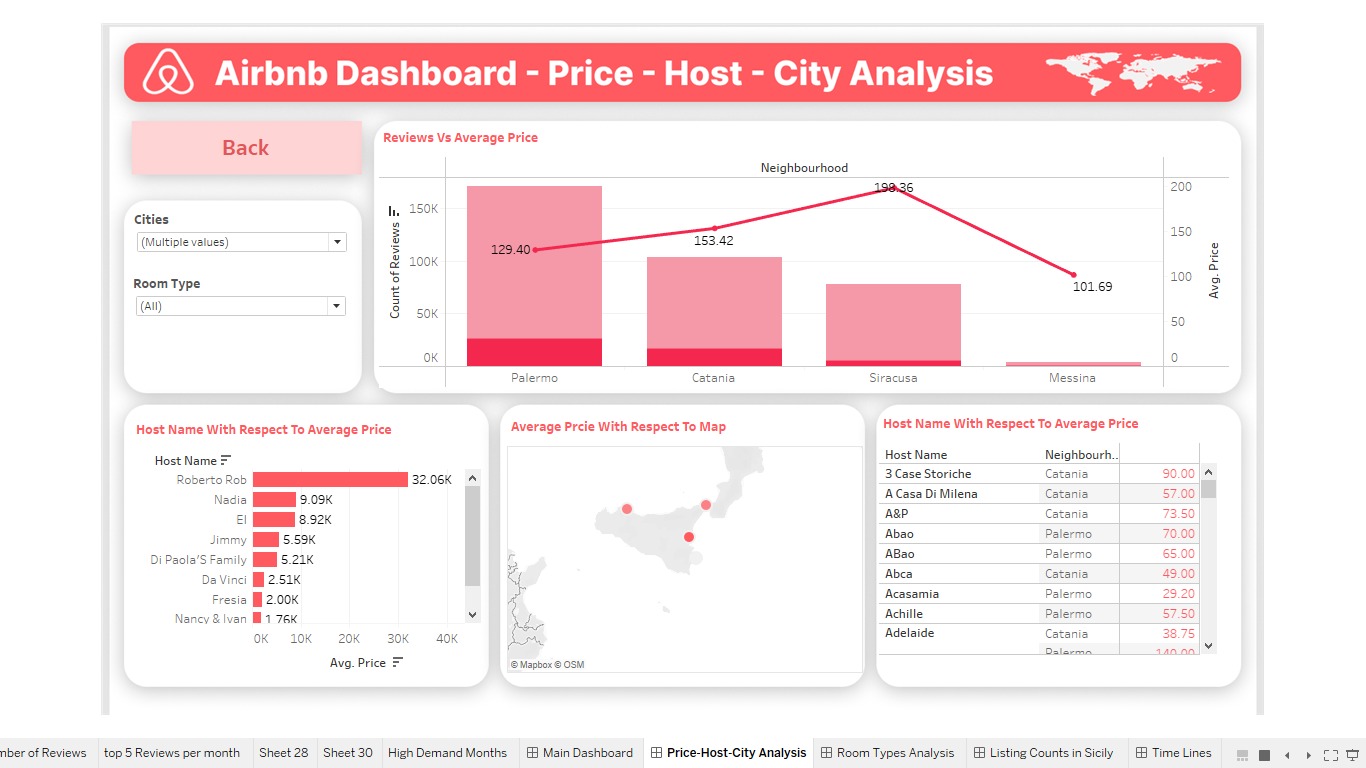
**Average price and total number of reviews by city:** This analysis explores the relationship between the average price of listings and the total number of reviews received, considering different cities. It helps identify if there is any correlation between price and popularity (as indicated by the number of reviews).

**The average price on the map:** This feature visually represents the average price of listings on a map. It allows users to quickly grasp the price distribution across different geographic locations.

**Hosts' names and their average price:** This chart presents the names of hosts alongside their corresponding average price. It enables users to identify specific hosts and understand their pricing patterns.

**Table displaying city, Hostname, and average price:** This table presents a comprehensive view of cities, hostnames, and their respective average prices. It allows users to easily compare prices across cities and identify hosts offering listings at different price points.

These analyses and visualizations in the price-Host-City dashboard provide valuable insights into the relationship between price, hosts, and cities. Users can explore correlations between price and popularity, visually assess average prices on a map, and compare prices across different hosts and cities using the provided table.



## **Room Type Analysis.**

The room type analysis dashboard offers a comprehensive overview of different room types and their important attributes. Here is what you can find in this analysis:

**Count of listings by room type in cities:** This section provides a count of listings categorized by different room types (e.g., entire home, private room, shared room, hotel room) within each city. It helps users understand the distribution and availability of various room types in different locations.

**Average reviews of room types:** This analysis highlights the average number of reviews received by each room type. It allows users to assess the level of feedback and guest satisfaction for different types of accommodations.

**Comparison of review count and average price:** This comparison displays the relationship between the count of reviews and the average price for each room type. It helps users identify if there is any correlation between the number of reviews and the pricing strategy for different room types.

Based on these insights, you have observed that even though shared rooms and hotel rooms have relatively low numbers of listings, they tend to have a high number of reviews. This suggests that hosts or establishments offering these room types actively encourage guests to provide feedback, indicating a focus on improving guest experiences. It implies that hosts offering shared rooms and hotel rooms value guest feedback and strive to enhance their services based on guest input.

A screenshot of a graph

Description automatically generated with low confidence

## **Timelines.**

The timeline dashboard in the advanced booking revenue section provides insights into revenue generated from advanced bookings, along with the split between different months and availability. Here is what you can find in this dashboard:

**Advanced booking revenue:** This section displays the revenue generated from advanced bookings. It helps users understand the financial impact of bookings made in advance.

**Split between different months:** The dashboard displays the revenue distribution across different months. This allows users to identify patterns or trends in booking behavior throughout the year. For example, it can highlight peak months with high booking volumes and potential off-peak periods with lower booking activity.

By examining the timeline dashboard, you can gain a comprehensive understanding of the revenue generated from advanced bookings, identify booking patterns by month, and assess the availability of listings during different periods.

A screenshot of a graph

Description automatically generated with low confidence

## **Listings in Sicily.**

The Sicily Cities dashboard provides valuable information about the distribution and percentage of listings in different cities within Sicily. Here is what you can find in this dashboard:

Cities and their percent of the number in Sicily: This section presents a list of cities in Sicily along with the percentage of listings they contribute to the total number of listings in the region. It helps users understand the concentration of listings in different cities and highlights the cities with the highest number of listings.

**Example observation:** The dashboard mentions specific cities like Palermo, Siracusa, and Catania, indicating that they have the highest number of listings in Sicily, accounting for more than 1/4th (25%) of the total number of listings. This observation gives users an understanding of the dominant cities in terms of available accommodations.

**Map view of cities and a number of listings**: This visualization on the map provides a spatial representation of cities in Sicily, with each city being represented by a color. The darker the color, the higher the number of listings in that particular city. It helps users visualize the distribution of listings across different regions of Sicily.

By examining this dashboard, users can identify the cities with the highest concentration of listings in Sicily, understand the percentage contribution of each city, and gain a visual understanding of the number of listings in different regions

A screenshot of a computer

Description automatically generated with medium confidence

# **Conclusion**

Let us answer the business questions based on the information provided in the dashboard:

**(1) Average price of room types in Sicily and comparing with any specific city:**

Entire Home and Apartment: 172.8 Euros

Hotel Room: 163.8 Euros

Private Room: 108.2 Euros

Shared Room: 46.4 Euros

Users can compare the prices of different room types in different cities by selecting different locations in the filter option. As we compare Catania average price with average price of in rest of Sicily.

|  |  |  |
| --- | --- | --- |
| **Room Type** | **Average Price in CT** | **Average Price in Sicily** |
| Entire Home / Apartment | 184.10 Euro | 172.8 Euro |
| Hotel Room | 91.45 Euro | 163.8 Euro |
| Private Room | 76.17 Euro | 108.2 Euro |
| Shared Room | 40.60 Euro | 46.4 Euro |

**(2) The highest number of listings with respect to room types:**

The cities with the highest number of listings are Palermo, Syracuse, and Catania. Among these cities, Palermo has the largest number of listing in all room types available.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Room Type** | **Listing count** | **City** | **Average Price** | **Average Review** |
| Entire Home / Apartment | 4140 | Palermo | 125.2 Euro | 35.00 |
| Hotel Room | 53 | Palermo | 102.9 Euro | 18.19 |
| Private Room | 1459 | Palermo | 143.2 Euro | 17.38 |
| Shared Room | 15 | Palermo | 40.6 Euro | 9.20 |

**(3) Highest number of advanced bookings by month:**

Based on the data extracted from March, the month of April has the highest number (39.85K) of advanced bookings and it is follow by August (32.78K) and July (30.52k) in terms of advanced bookings.

**(4) Highest average price by month:**

The month of August has the highest average price (196.47 euro) for accommodations in Sicily.

This can be an attributed to the high demand during that month. July and June has second and third highest average price 176.10 euro & 167.85 euro respectively.

**(5) Discrepancy between visitor numbers and bookings:**

According to Google, Taormina has the highest number of visitors in Sicily. However, the analysis from the dashboard indicates that cities like Palermo, Syracuse, and Catania have the highest number of bookings. This suggests that tourists prefer staying in the larger cities rather than the most visited destination.

|  |  |
| --- | --- |
| **City** | **Listing Percentage in Sicily** |
| Palermo | 10.97% |
| Syracuse | 6.91% |
| Catania | 6.49% |

**(6) Impact of the pandemic on reviews:**

In 2020, there was a drop in the total number of reviews, which can be an attributed to the pandemic's effect on travel.



However, the growth rates in reviews continue to show the increasing popularity of Sicily as a tourist destination, indicating a recovery and growing interest in the region.

These insights provide valuable information for businesses in the hospitality and tourism industry, enabling them to make data-driven decisions and tailor their strategies to meet customer demands and market trends in Sicily.

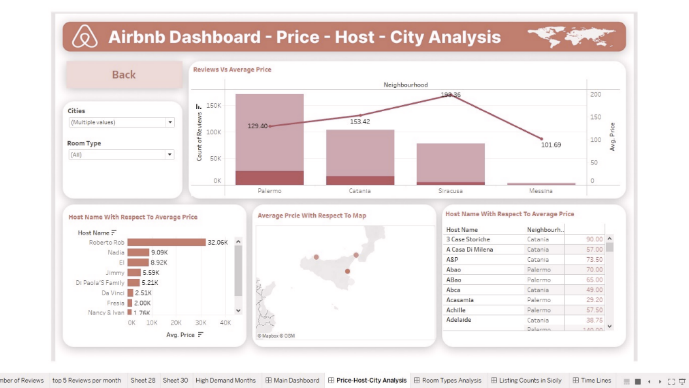
# **Color blindness test**

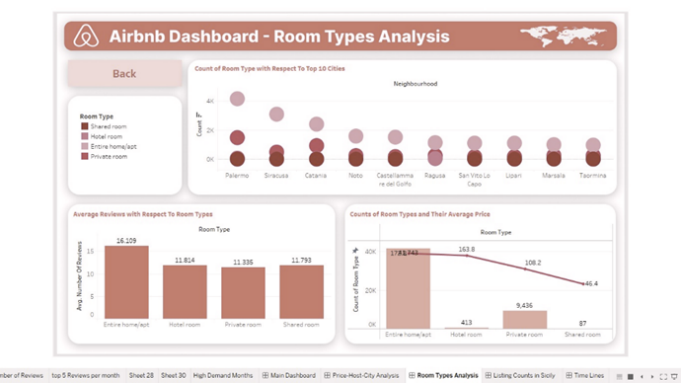
For each dashboard, the color blindness test is carried out, to evaluate whether people who have anomalies in the perception of colors can easily analyze the dashboards achieved. The

A Pilestone simulator was used (https://pilestone.com/pages/color-blindness-simulator-1# )

## **Red Weak test**

A picture containing text, screenshot, font, diagram

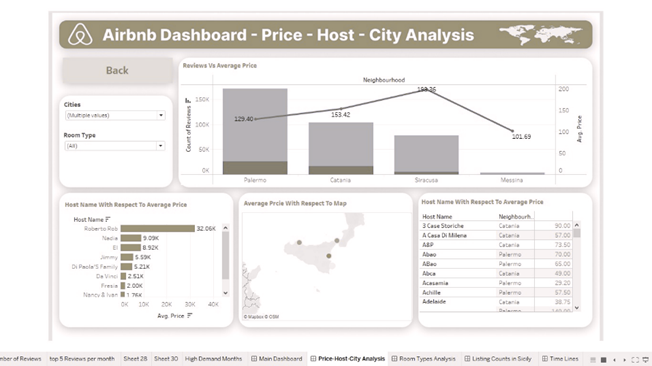
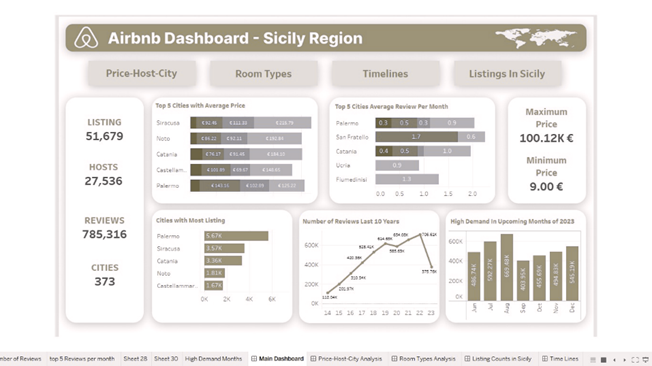
Description automatically generated

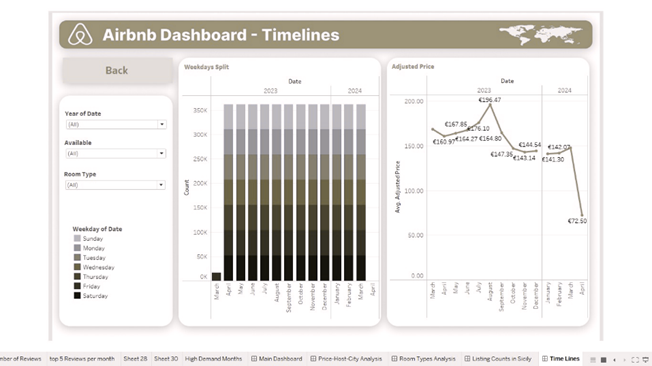
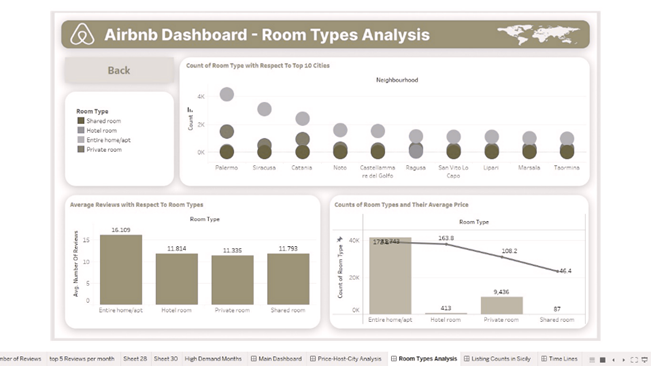


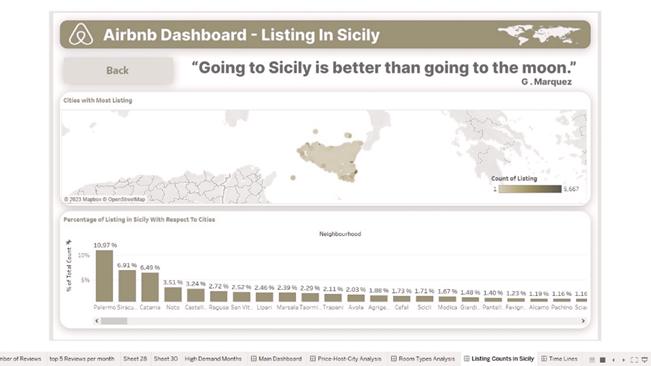
A screenshot of a computer

Description automatically generated with medium confidence

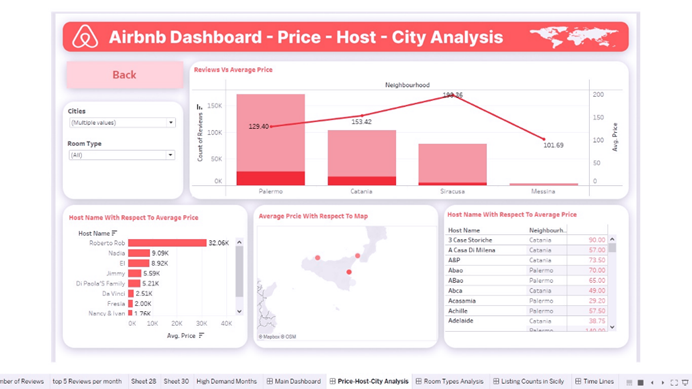
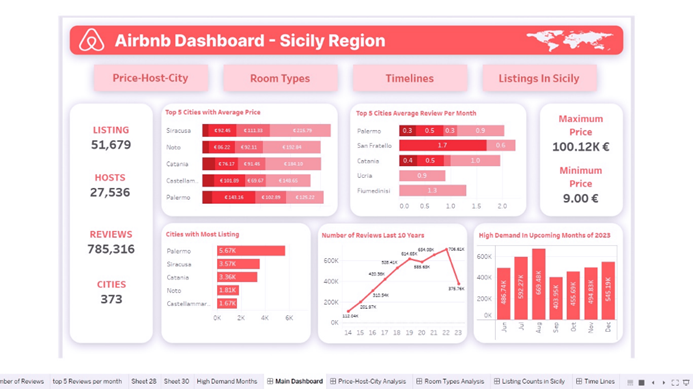
## **Red Blind test**

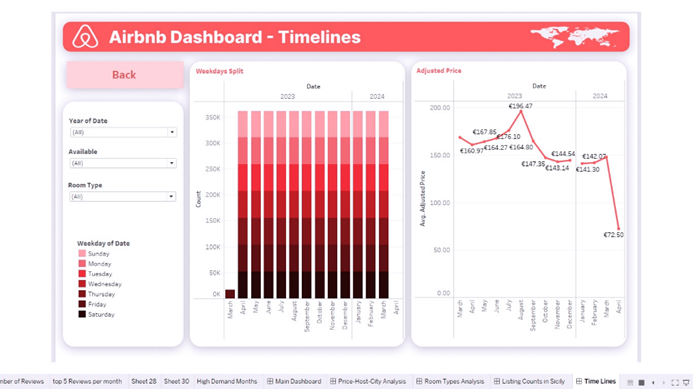
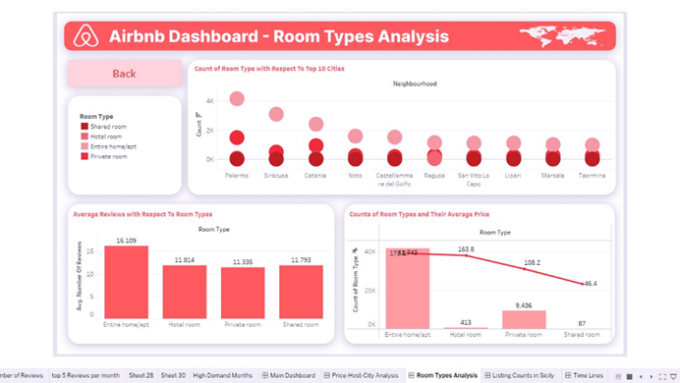


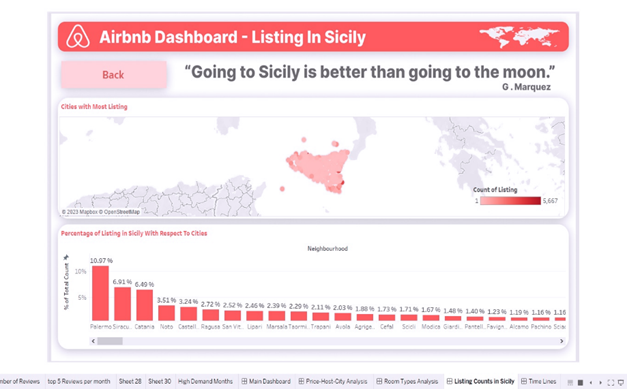




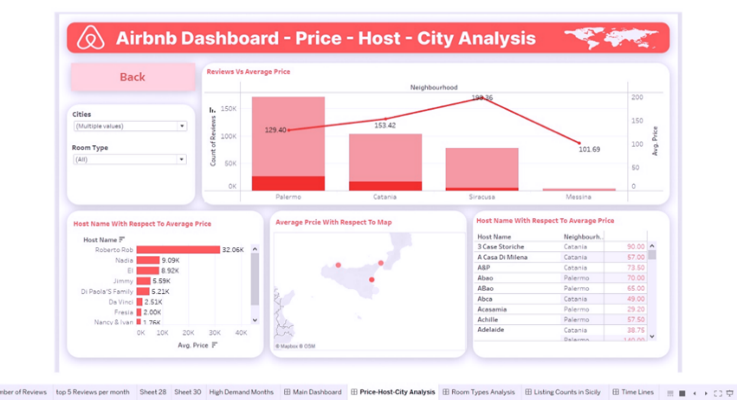
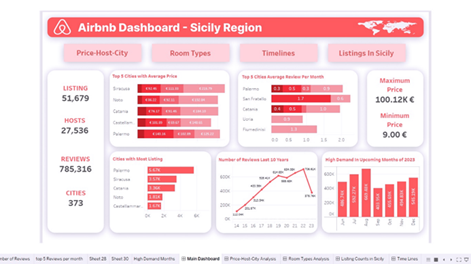
## **Blue Weak test**

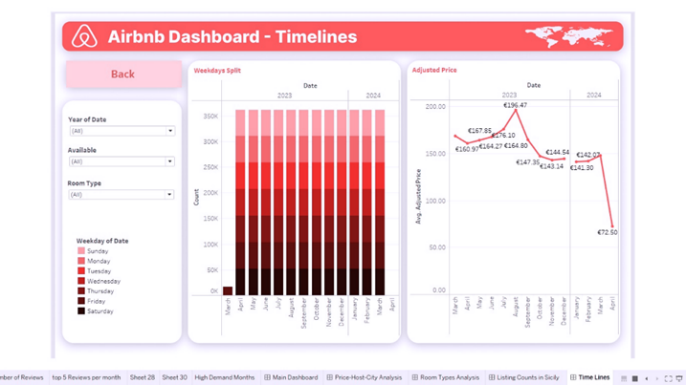
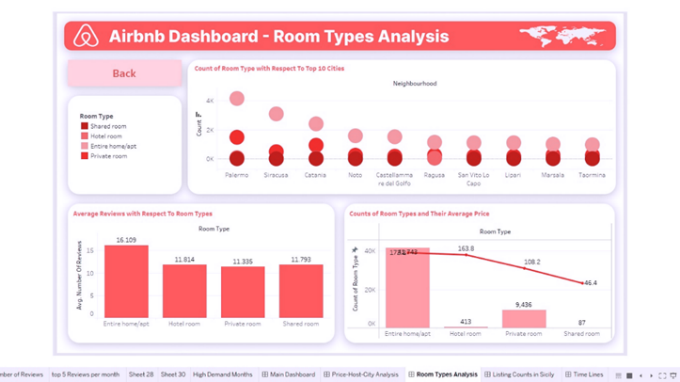


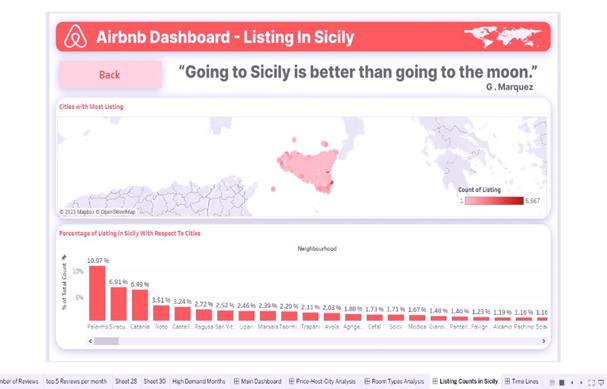




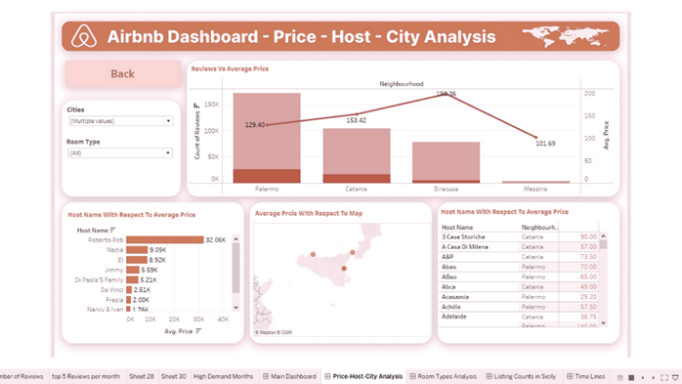
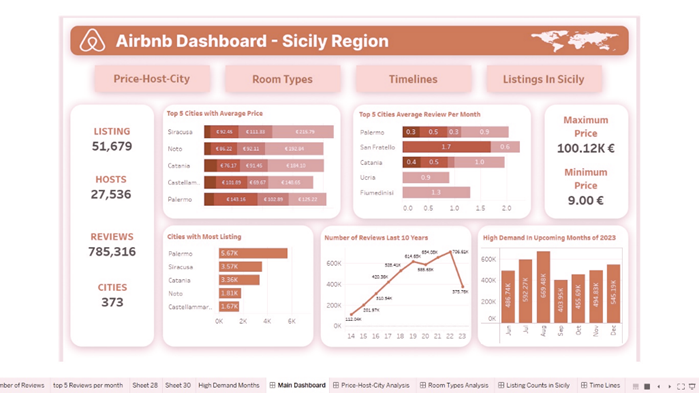
## **Blue Blind test**

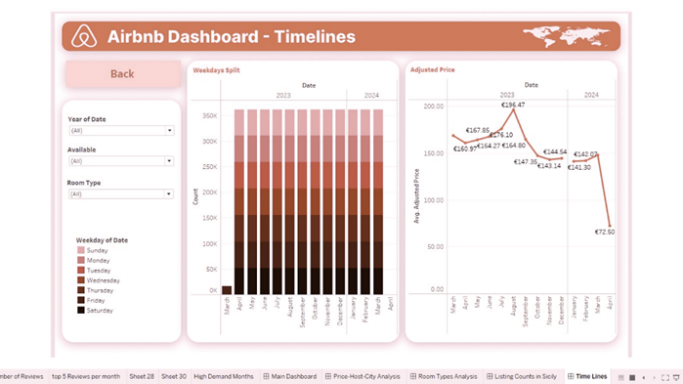
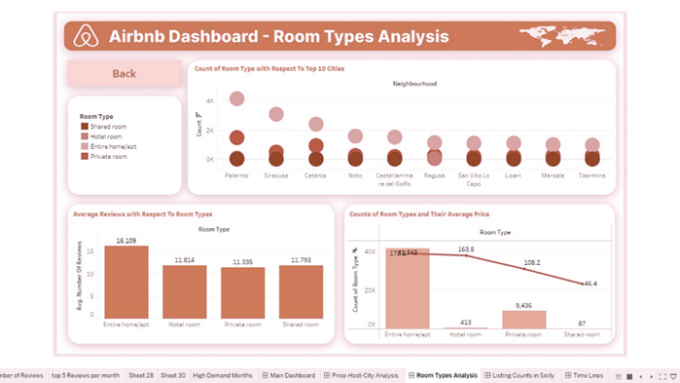


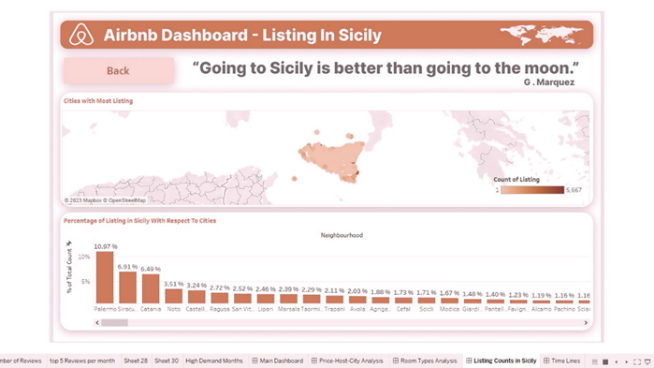




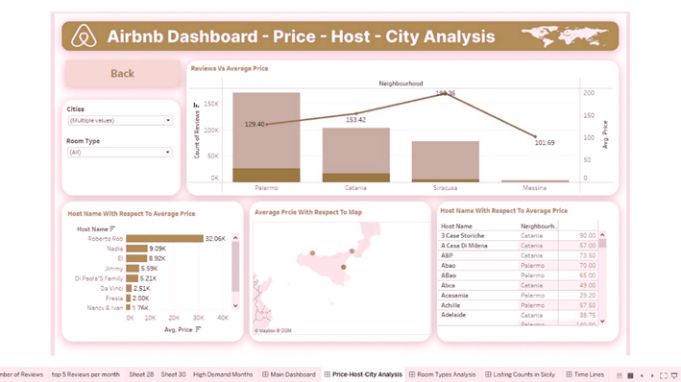
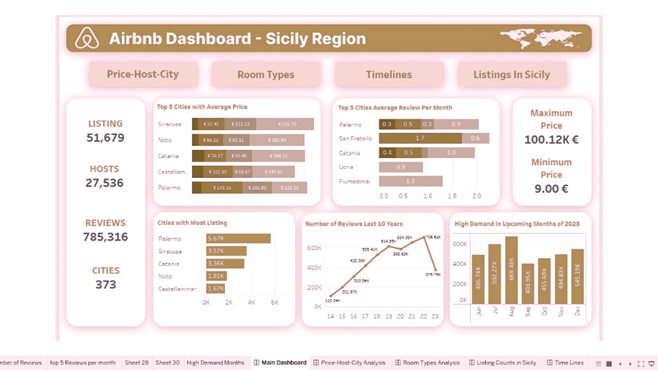
## **Green Weak test**

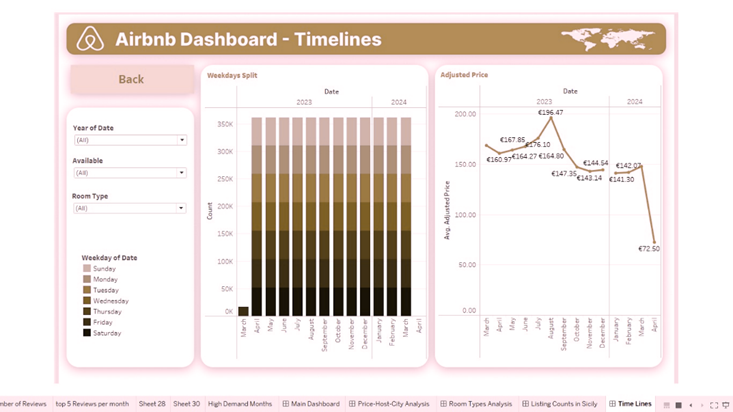
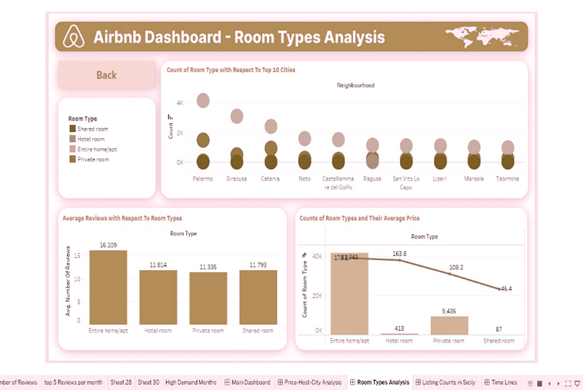


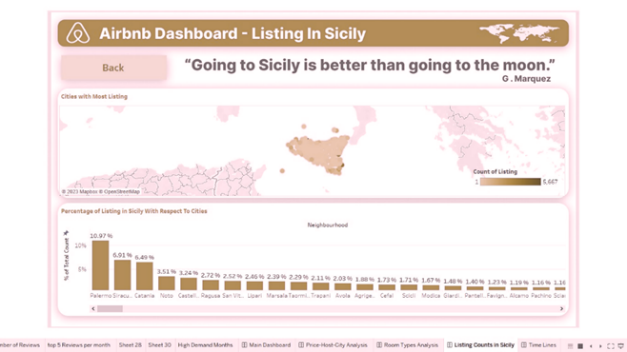




## **Green Blind test**







## **Monochromic Test**

