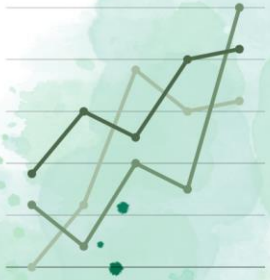
Airbnb Mallorca

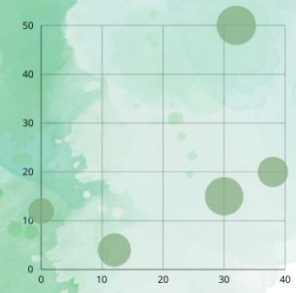


Project Goals

The goals for this project were sourcing my own dataset, explore relationships within the data and create a hypothesis to test. After this I used the results that I created in Python to build a business case in Tableau to answer the key questions:

- What is the relationship between property location and guest satisfaction?
- Does exceptional value of a listing such as close to the beach or mountain view have influence on the guest satisfaction regarding the location?

I chose the dataset of Airbnb Mallorca because I live there partly and because the data is of high quality.



Data that I used

Python: last modified csv: 11964 rows,
43 columns
Regions: Mallorca (Spain)
Other data: GEOJSON file of the neighbourhoods, Zillow housing data for time series
Created: coding, clusters and boolean columns from listing characteristics. Visuals in both Python and Tableau
Data Sources: Inside Airbnb, Nadaq

Skills that I applied

Sourcing data
Consistency checks
Exploring relationships with matrix heatmaps and plots
Spatial analysis
Supervised Machine Learning: regression
Unsupervised Machine Learning: clustering
Deriving new columns with boolean data
Sourcing and analyzing time series data
Creating a business case in Tableau

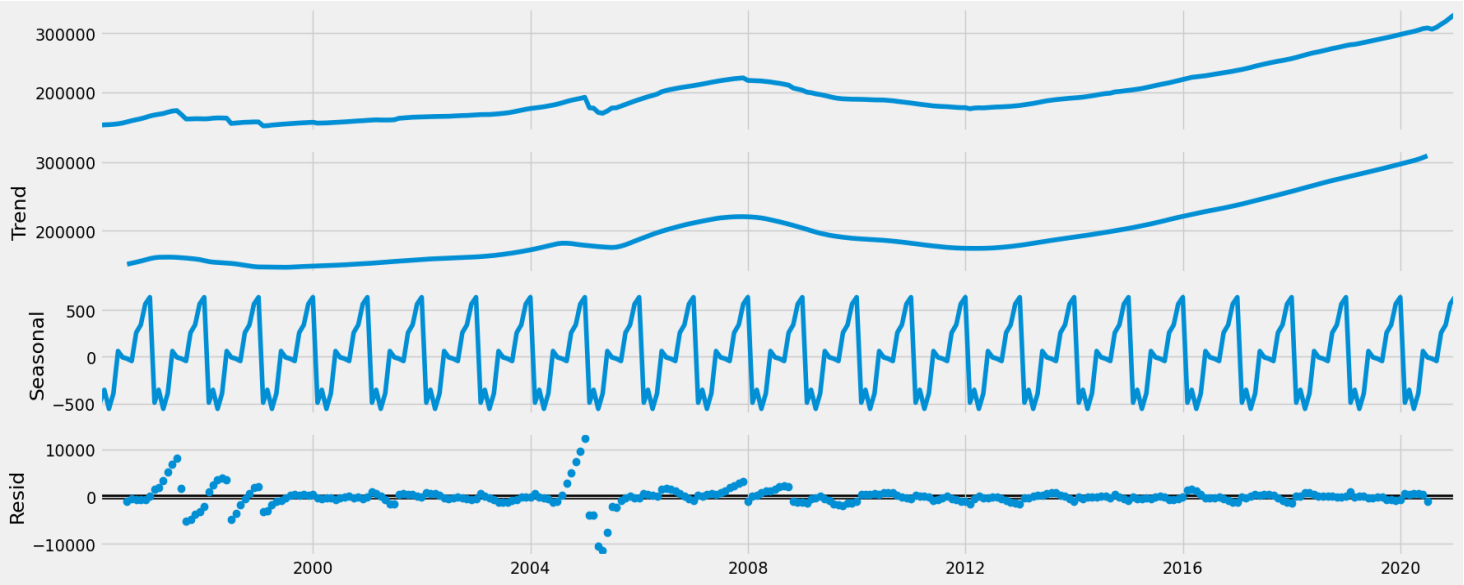
airbnb

The process

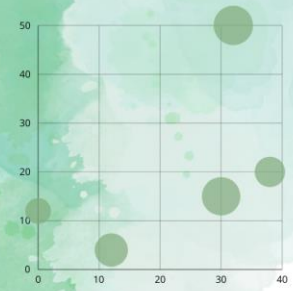
The first task was the foundation of this project as I had to source my own data. There are quite some some sources online and organizations that make their data public. Choosing the data depended on different criteria points. To mention a few: it needed to include a geographical component with at least 2 different values, 2 continuous variables, 2 categorical variables and have at least 1500 rows.

Once I found the dataset of Airbnb Mallorca, I started analysing in Python. After cleaning the data, I explored the dataset by conducting several types of analysis explained in the following sheets. From my key findings I created a business case in the shape of an interactive storyboard in Tableau.

An important part of data analytics can be time series analysis, to discover trends and to be able to forecast predictions if the data is stationary. To stay as close to the subject of accommodation, I found a dataset on the U.S. housing market between 1996 and 2020 from Zillow.

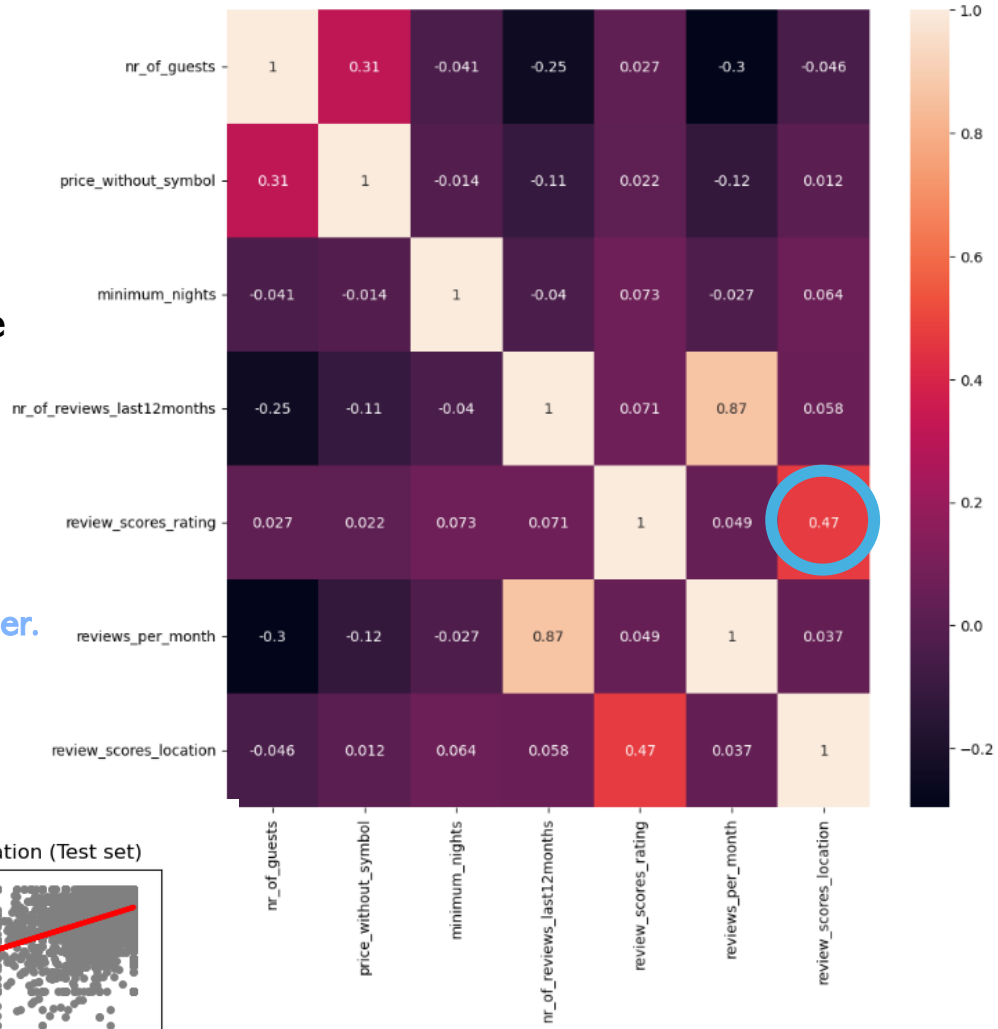


I had to set the date as the index of the dataset to conduct the analysis and use code to decompose the data in order to see the different components that can affect the time series: trend, seasonality, cycle and noise.



Analysis

After exploring the relationships between numeric values of the Airbnb Mallorca listings, I decided to further analyze the correlation between review scores rating (total score) and review scores location.



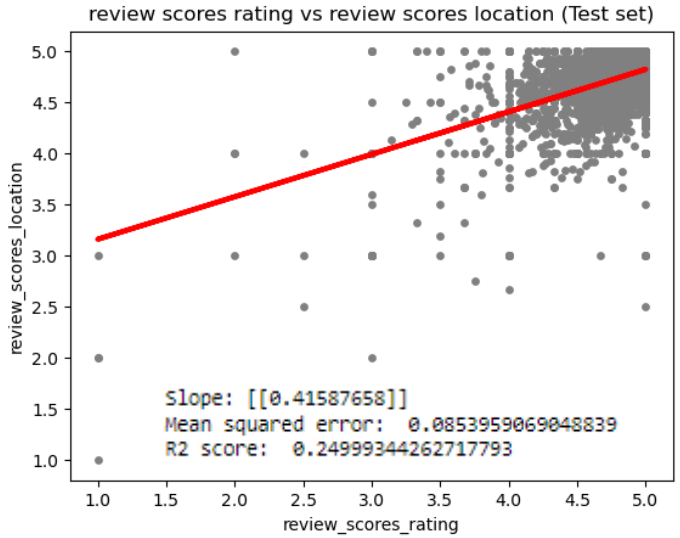
This ultimately resulted in the hypothesis:
If the review scores rating is higher, then the review scores of the location are also higher.

Conducting supervised machine learning resulted in the following linear regression.
The summary statistics model tell the following:

The slope indicates that when the review score ratings get higher, the review scores of the location get also slightly higher.

The MSE is very small, which means that the regression line passes fairly close to the observations, making it a good fit.

The R2 hower is 0.25 which can be seen as 25% which indicates it is not a good fit.

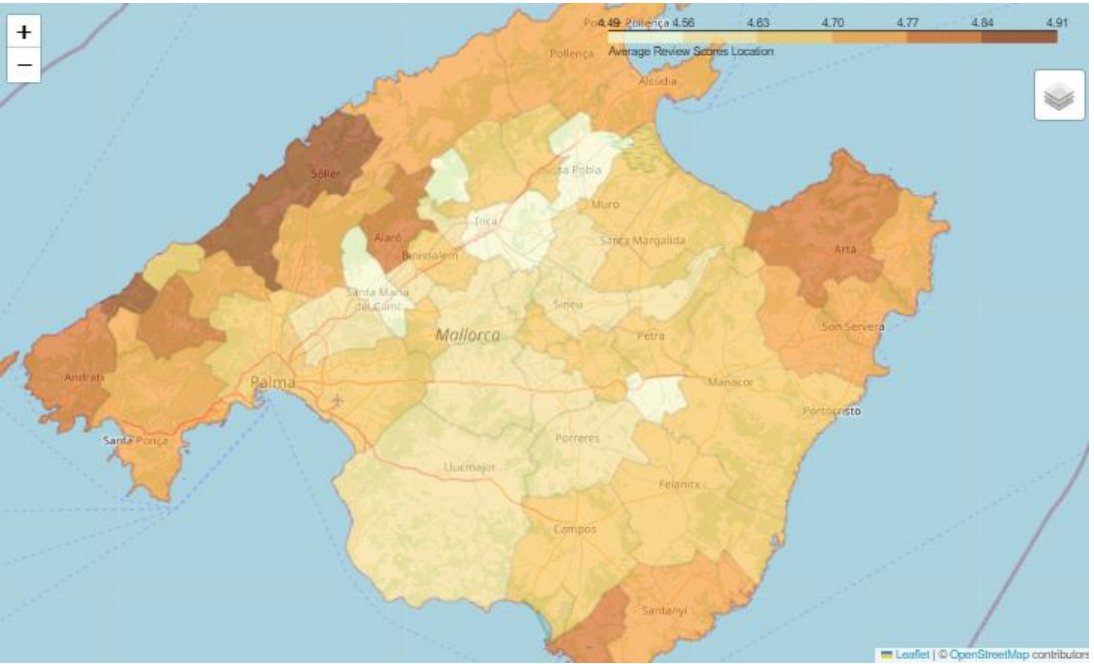


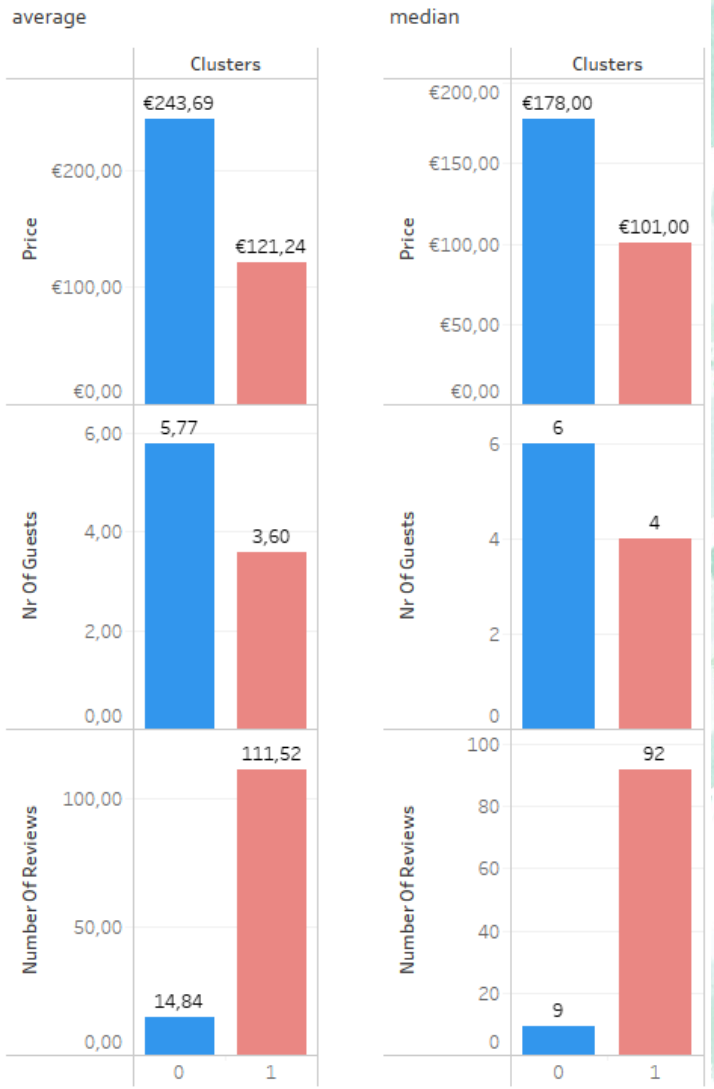
airbnb

Analysis

To fully explain the data I standardized it and used unsupervised machine learning to create clusters. It did not give more information on any of the review scores, but it did show a clear distribution for other variables about the Airbnb listings. There are two clusters and cluster 0 is much bigger than 1.

The choropleth map created in Python does give more information on the relationship between property location and guest satisfaction. Even though the average rating scores for location are close to each other, between 4.49 and 4.91, it does clearly show that the west coast of the island has the highest ratings.



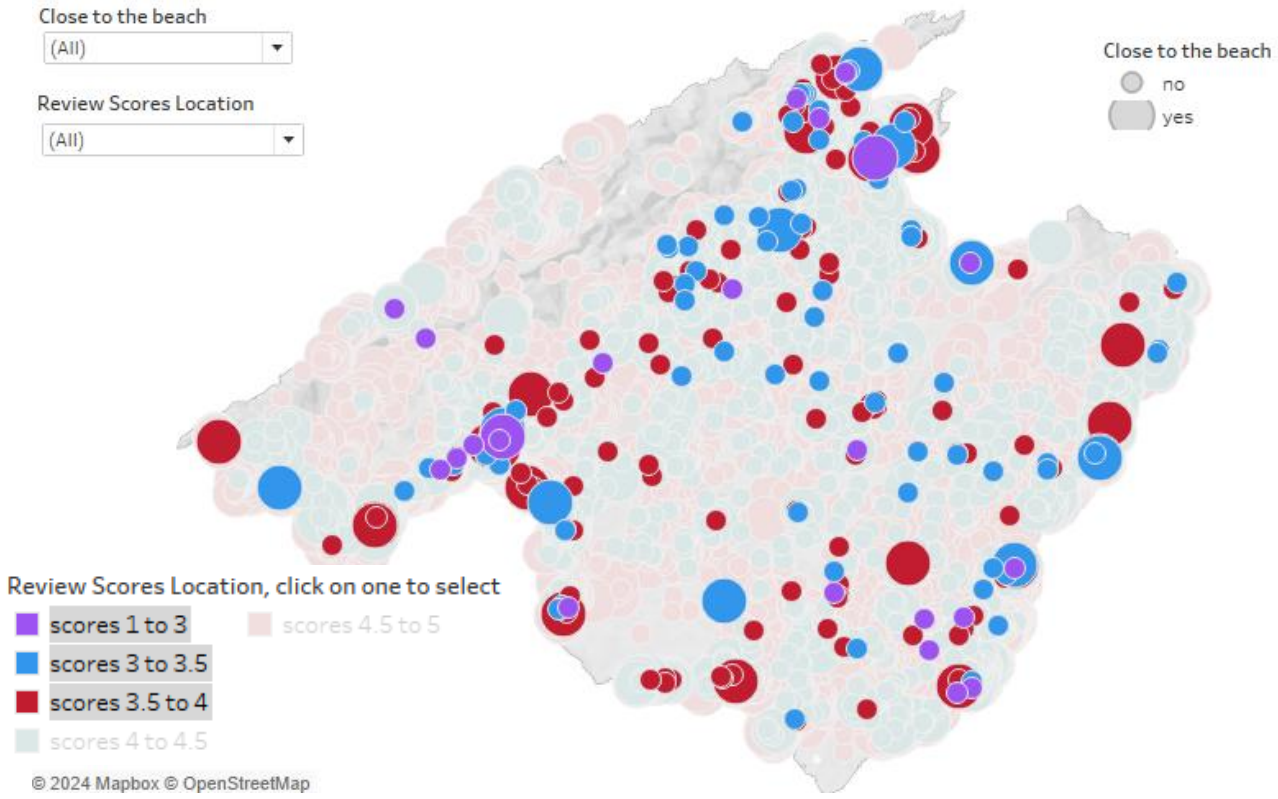


airbnb

Business Case

In Tableau I created a business case for a client who is looking for a property in Mallorca, to use as an Airbnb listing and for personal use. I was responsible finding out the best possible locations for her and I based my analysis around the following two key questions:

- What is the relationship between property location and guest satisfaction?
- Does exceptional value of a listing such as close to the beach or mountain view have influence on the guest satisfaction regarding the location?



To answer the second question, I created Boolean columns for the amenities ‘close to the beach’ and ‘mountain view’. Then I added the values into separate maps, combined with location review scores segmented in groups.

There is no clear majority in high location scores for listings that are close to the beach or have mountain view, compared to low scores.

However, for both variables there are only listings with the highest scores between 4 and 5 on the west coast. The same area as discussed on previous slide

Recommendation

After conducting exploratory, linear regression and cluster analysis it became clear that the correlation between review scores rating and review scores location is moderate but not strong.

Therefore the hypothesis '**If the review scores rating is higher, then the review scores of the location are also higher**' is not entirely true.

The spatial analysis resulted in the choropleth map of the neighbourhoods of Mallorca, with each its average review score location. This map does show that guests are more satisfied about the location on the west side of the island, where there are both mountains and beaches on the coastline, giving a fascinating landscape. When analyzing the location related amenities 'close to the beach' and 'mountain view' it became clear that there are only high scores between 4 and 5 in this region. Therefore I would recommend my client to start and Airbnb in one of the following 6 neighbourhoods:

Estellencs
Valldemossa
Deyá
Fornalutx
Sóller
Escorca

Next Steps

Next possible step is analyzing more of the data available for the specific interesting neighbourhoods. Mallorca has a lot of airbnb listings, focusing on a smaller area will help analyzing data more in detail.

Limitations

The data from Airbnb includes information about reviews, but not on bookings. Reviews are optional and not given by every guest. Data on bookings would give more certainty whether a listing is popular or not.

Analyzing the amenities 'close to the beach' and 'mountain view' makes me realize that hosts add these themselves as an extra value of the listing, but it shows that this is not always trustworthy.

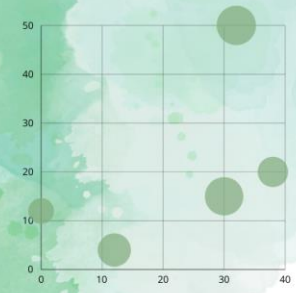
Personal evaluation

Successes

As the subject is about accommodations on Mallorca, I used a lot of spatial analysis. Especially in Tableau I was able to try and test creating different maps and filters, which I learned a lot from.

I believe I have chosen an interesting hypothesis based on the dataset and the relationships in it. My mentor was impressed how I turned it into the business case.

Challenges and lessons learned



The dataset of Airbnb Mallorca was not the first one I chose and tried working with. First I found a dataset of B Corp, a social enterprise that inspires me. Unfortunately, previous datasets were either incomplete or incorrect and I learned that it would have saved me time if I would already have started searching during my previous project. I sent this as feedback to CareerFoundry.

When I conducted spatial analysis in Python, using the Airbnb dataset and their geojson file with the neighbourhoods, I kept getting black areas on the map, while this did not show when my mentor opened the same Jupyter notebook. After trying different things, I found out the names of the neighbourhoods were not the same in both files. Spain uses a lot of accents on letters, so I managed to correct it by opening the geojson file in VS Code and edit them.

airbnb