Project Report

# GitHub URL

<https://github.com/UCDPADanielCarey/Airbnb-Data-Project>

# Abstract

(Short overview of the entire project and features)

My aim of this project is to explore the Airbnb market in Lisbon through the use of python. Which areas are in most demand, and what factors can have an impact on price. I want to also see how the prices in Lisbon compared with a similar city; Madrid.

Airbnb is an online marketplace that allows people who wish to rent out their properties to people who are looking for accommodation in specific locales.

The company has come on in leaps and bounds since 2007, when its co-founders first came up with the idea to invite paying guests to sleep on an air mattress in their front room. Going by latest figures Airbnb has in excess of five and a half million listings in well over 200 countries.

In the last few years, Lisbon has seen a marked increase in Airbnbs available, and it has become particularly popular with “digital nomads” to base themselves there.

# Introduction

*(Explain why you chose this project use case)*

I chose this use case as I have an interest in Airbnb as a business model and I have visited the city of Lisbon before.

# Dataset

*(Provide a description of your dataset and source. Also justify why you chose this source)*

The data sets I worked with were retrieved from *Insider Airbnb* (link here: [http://insideairbnb.com/get-the-data/](file:///C:\Users\danie\Downloads\here)). This website provides a plethora of datasets for most major cities and tourist destinations across the world.

# Implementation Process

*(Describe your entire process in detail)*

I first imported my libraries into Spyder. I then created my data sets from the CSV files that I down loaded from *Insider Airbnb*. I merged the two datasets and removing that rows with NAN values in them.

*#Here I am creating the data frames*

*lisbon\_airbnb = pd.read\_csv(r"C:\Users\danie\OneDrive\Documents\Airbnb\lisbon\_listings.csv")*

*pd.DataFrame(lisbon\_airbnb)*

*print(lisbon\_airbnb)*

*lisbon\_airbnb.dropna() #here I am dropping my NAN values.*

*lisbon\_airbnb.head()*

*madrid\_airbnb = pd.read\_csv(r"C:\Users\danie\OneDrive\Documents\Airbnb\madrid\_listings.csv")*

*pd.DataFrame(madrid\_airbnb)*

*print(madrid\_airbnb)*

*madrid\_airbnb.dropna()*

*#Merging data*

*Lis\_V\_Mad = pd.merge(lisbon\_airbnb, madrid\_airbnb, how='inner', on='price')*

*Lis\_V\_Mad.groupby(['price']).count()['price'].sort\_index(ascending=False)*

*Lis\_V\_Mad.dropna()*

When I ran the below line of code, it allowed me to see what column headings I could drop.

*lisbon\_airbnb.describe()*

I decided to drop a few headings, such as ‘host\_id’. See code below:

*lisbon\_airbnb = lisbon\_airbnb.drop(['id','host\_id','reviews\_per\_month','host\_name','license','last\_review'], axis=1)*

*lisbon\_airbnb.head()*

I then ran the isnull is see if all NANs were removed.

# Results

*(Include the charts and describe them)*

I first used a group by function to see how many room types there were on the dataset:

Graphical user interface

Description automatically generated with medium confidence

I then make a graph in Spyder to visualise the above:

Chart

Description automatically generated

From this we that entire homes are the most popular form of accommodation for Airbnbs in Lisbon’s metropolitan area. I now want to check in which areas of Lisbon had the most listings. I imagine it would be the city centre:

Chart, bar chart

Description automatically generated

There were many neighbourhood groups available so to be able to read the graph better I limited this to the top ten areas. As we can see from the above, Lisbon’s centre has the most listings. Second is the upmarket costal area of Cascais and in third place is Sintra. Sintra is home to several UNESCO word heritage sights.

I then wanted to see which areas of Lisbon have (on average) had the highest prices.

Chart, bar chart, funnel chart

Description automatically generated

Going by the above it is clear that areas further outside of Lisbon’s core can command a higher price. When I delved further into the highest areas, I could see that there were a few outliers pricewise. One property in Alenquer was listed as €5,729! This obviously brought the average up for the whole area.

I then wanted to see on the map where prices varied:

Graphical user interface

Description automatically generated

I then wanted to see geographically where one could stay for under €100 a night. As we can see from the below, the further you go away from Lisbon’s city centre and avoid the coast there are cheaper options available.

Map, scatter chart

Description automatically generated

It was at this stage that I wanted to compare the prices of Lisbon with another major European city. I at first though of comparing it to London, but for the purposes of the below histogram I found Madrid a better fit. When I first compared prices from Lisbon to London, the London prices dwarfed those of Lisbon and made the overlapping histogram visualisation hard to read. I was only interested in properties of under €100 for this graph.

Chart, histogram

Description automatically generated

We can see from the above that Lisbon has a lot more properties in the €70-€100 demo graph.This is interesting as Lisbon is had a population of just under two million inhabitants , whilst Madrid has over six million. Could this be down to Lisbon’s status as a hub for remote working in a last few years? Quit possibly.

I then wanted to see what aspects of the dataset would have a correlation on the price:

Text

Description automatically generated

We can see from the above that reviews and availability throughout the year have a higher correlation on price. I now want to explore this in a heat map:

A picture containing chart

Description automatically generated

I am now trying to see if the skewness and kurtosis for the prices in the dataset were normal and that the price was normally distributed.

Chart, histogram

Description automatically generated

As displayed above, they seem to be slightly off.

I now want to see through the use of modelling what aspects had an impact on price. Firstly I chose to see if there is a correlation between the number of total reviews and price:

Chart, scatter chart

Description automatically generated

We can see from the above there is a correlation between the total number of over all reviews a property has and the price. Though this doesn’t have too much of a sway on the higher priced properties. I then wanted to see if there is an impact on how many reviews per month impacted on the price of the different room types. From the insights gained from the below lmplot, entire homes and hotel rooms reviews appear to have impact on the price. There appears to be only a small amount of shared rooms available in the Lisbon area anyway, so the below findings on the graph aren’t at all surprising. We can conclude that if a person is booking a whole house or apartment that they’re more influenced by reviews. With regards to private rooms there appears to be a dip on rooms costing under €40, but this then steadily increases. This may be that the lower end of the market may not be too fussed about reviews and happy just to have somewhere to lay their head for the night.

Chart, line chart

Description automatically generated

I now want to test if there was a connection between a place being available the whole year round and the price. Obviously from the below graph, one can see that the hotel rooms come up trumps here.

Graphical user interface, chart, line chart

Description automatically generated

# Insights

*(Point out at least 5 insights in bullet points)*

From my exploration of this dataset of the Lisbon Airbnb market through the use of python, I have concluded the following:

1 – That Lisbon seems to have a more diverse Airbnb rental market compared with Madrid, and in some instances overtakes Madrid pricewise.

2- Reviews do have an impact on price, especially in the middle price market. This isn’t as much prevalent in the upper and lower ends of the spectrum as I’ve displayed in several graphs.

3- Availability throughout the year can command a higher price.

4- Certain areas can command a higher price than others.

5- Prices are somewhat skewed in the distribution plot for prices in the middle market. We saw this in the distribution plot graph.

# References

*(Include any references if required)*

<https://www.kaggle.com/code/kotaminegishi/nyc-airbnb-price-modeling/notebook>

<https://github.com/mohamedirfansh/Airbnb-Data-Science-Project/blob/master/Machine%20Learning%20Models.ipynb>

<https://towardsdatascience.com/airbnb-price-prediction-using-linear-regression-scikit-learn-and-statsmodels-6e1fc2bd51a6>

<https://www.kaggle.com/code/benroshan/belong-anywhere-ny-airbnb-price-prediction/notebook>

<https://github.com/Shafiq-Kyazze/AirBnb/blob/master/AirBnb%20%20price-prediction.ipynb>

How Airbnb Works (investopedia.com)

A Lisbon story: short-term rental platforms and the housing market | LSE Business Review