



Business Intelligence  
Part 2 - Data Mining

# AIRBNB LISBON

Belong Anywhere

Presented by: Ricardo Santiago  
Ricardo Silva



# Data Preparation

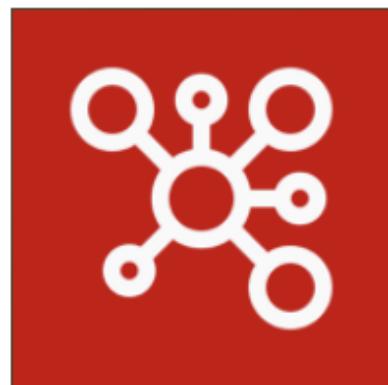


## Dataset used

Lisbon, Lisbon, Portugal		
17 December, 2023 ( <a href="#">Explore</a> )		
Country/City	File Name	Description
Lisbon	<a href="#">listings.csv.gz</a>	Detailed Listings data
Lisbon	<a href="#">calendar.csv.gz</a>	Detailed Calendar Data
Lisbon	<a href="#">reviews.csv.gz</a>	Detailed Review Data

<http://insideairbnb.com/get-the-data>

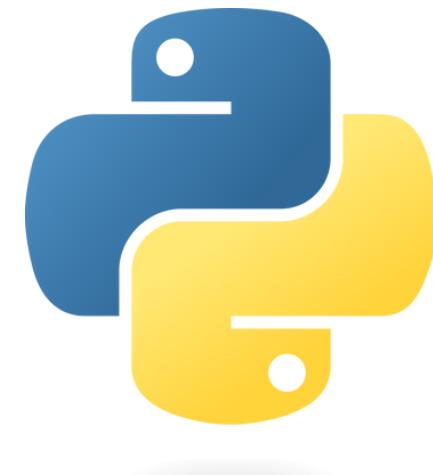
## Software used



**Pentaho**  
ETL Process



**PostgreSQL (pgAdmin 4)**  
Data Storage

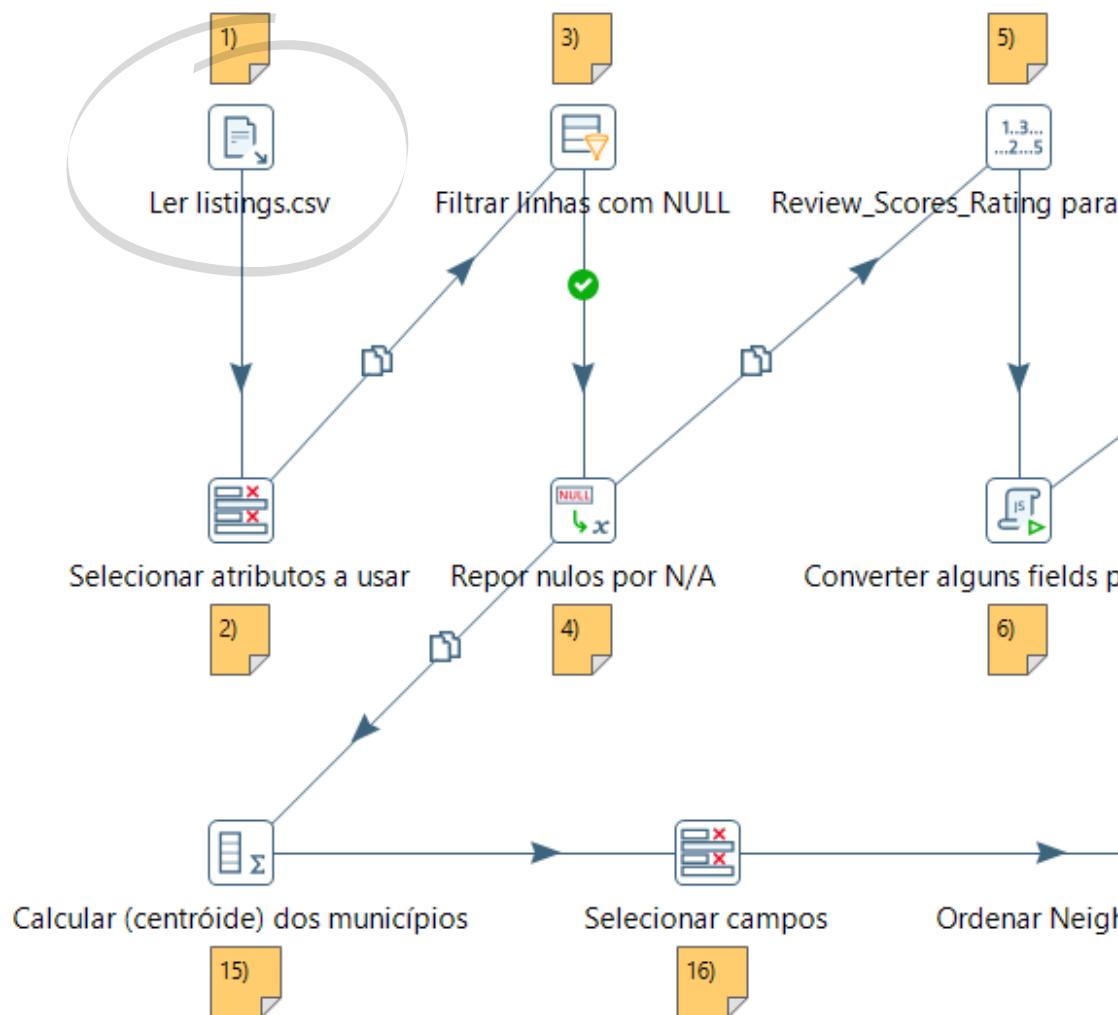


**Python**  
Data Preparation  
Data Mining

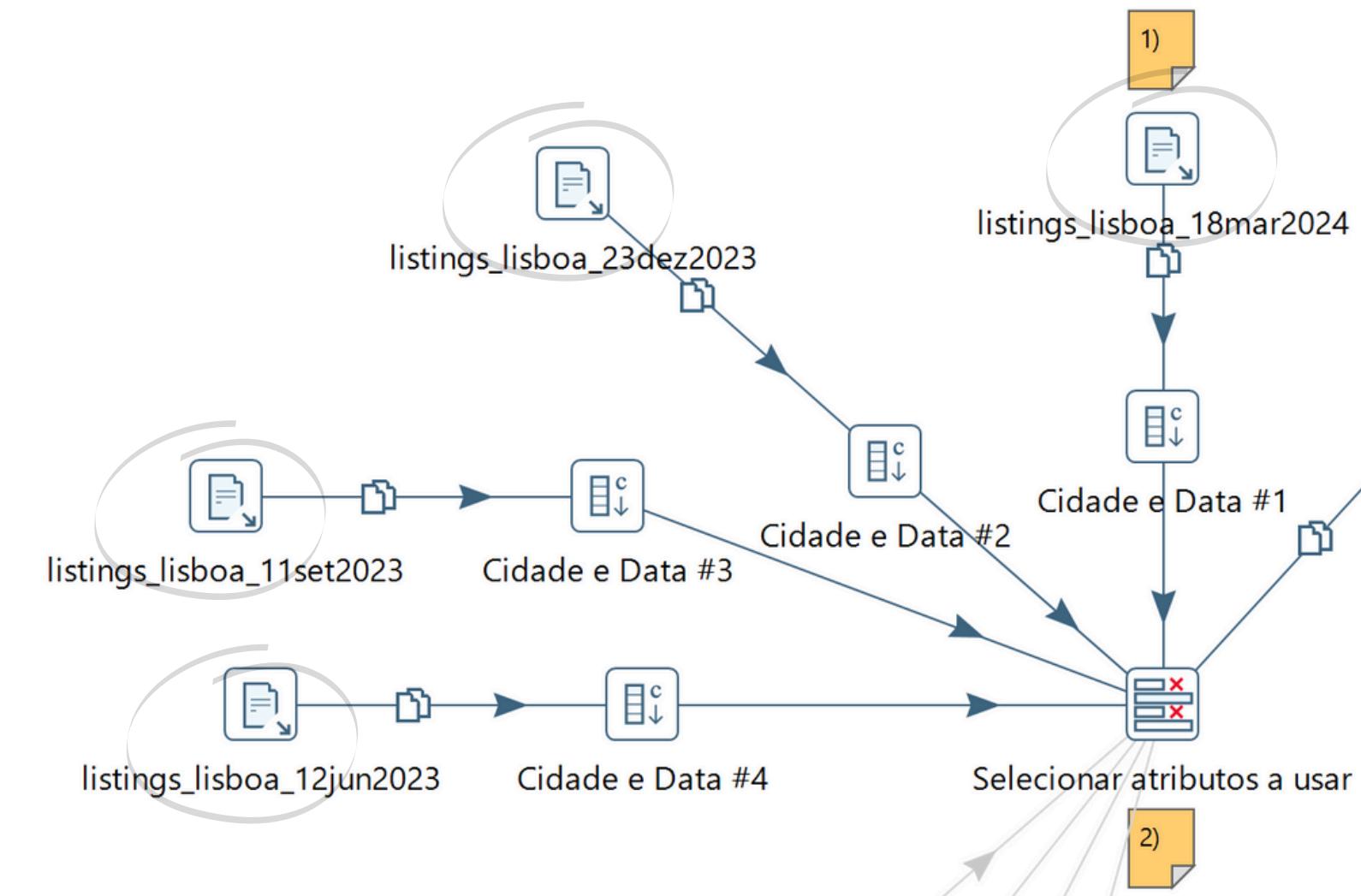
# Change in ETL Process Pentaho



➤ The file about the **most recent** update



# After



# Before

Files about the most recent **four** data updates

# Price Recommendation (Rating Study)

## Goal

Give a price recommendation for accommodation taking into account all its initial characteristics

## Features Used

**latitude**  
**longitude**  
**minimum\_nights**  
**maximum\_nights**  
**has\_availability\_bool**  
**instant\_bookable\_bool**  
**accommodates**  
**bathrooms**  
**bedrooms**  
**beds**

## Algorithms Testing

Tipo de Algoritmo	Algoritmo	Train Accuracy	Test Accuracy	Precision	Recall	F1-Score
Simple Classificator	Decision Trees	0.983	0.845	0.843	0.842	0.842
	KNN	0.980	0.756	0.755	0.754	0.755
Ensembles	RandomForests	0.983	0.840	0.838	0.837	0.837
	AdaBoost	0.493	0.497	0.504	0.486	0.492

# Testing Process

## 1- Hyperparameter search/choice

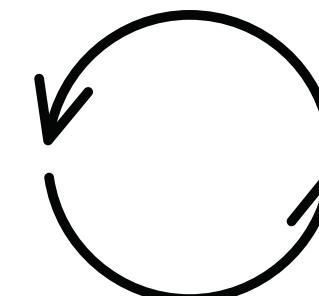
Done with grid\_search.

## 2- Test with X price categories (baseline is X=4)

Testing metrics: Train Accuracy, Test Accuracy, Precision, Recall and F1-Score.  
(Also looking at confusion\_matrix and classification\_report!)

## 3- Check if metrics threshold is reached (never bellow 0.80)

## 4- If not increase categories and repeat 1-3



## Results:

### Lisbon with 10 price categories

Tipo de Algoritmo	Algoritmo	Train Accuracy	Test Accuracy	Precision	Recall	F1-Score
Simple Classifier	Decision Trees	0.969	0.800	0.800	0.799	0.800

### Porto with 6 price categories

Tipo de Algoritmo	Algoritmo	Train Accuracy	Test Accuracy	Precision	Recall	F1-Score
Simple Classifier	Decision Trees	0.930	0.701	0.699	0.698	0.698

# Live Demo

## Airbnb listing price recommendation

The screenshot shows a Jupyter Notebook environment with two code cells and their outputs.

**Code Cell 1:** This cell contains the code for the prediction interface. It includes imports, a function definition, and a call to the function with specific parameters.

```
from IPython.display import clear_output
import numpy as np
import pandas as pd
import requests
import streamlit as st
from sklearn import linear_model
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, MinMaxScaler
```

```
def predict_price():
    # Load data
    df = pd.read_csv('airbnb.csv')
    X = df[['latitude', 'longitude', 'minimum_nights', 'maximum_nights', 'has_availability_bool', 'instant_bookable_bool', 'accommodates', 'bathrooms', 'bedrooms', 'beds']]
    y = df['price']

    # Split data into training and testing sets
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

    # Standardize features
    sc = StandardScaler()
    X_train = sc.fit_transform(X_train)
    X_test = sc.transform(X_test)

    # Train a linear regression model
    lr = linear_model.LinearRegression()
    lr.fit(X_train, y_train)

    # Predict price for a new listing
    new_listing = {'latitude': 38.698, 'longitude': -9.198, 'minimum_nights': 3, 'maximum_nights': 365, 'has_availability_bool': 1, 'instant_bookable_bool': 1, 'accommodates': 3, 'bathrooms': 1, 'bedrooms': 1, 'beds': 3}
    prediction = lr.predict([new_listing])

    return prediction[0]
```

**Code Cell 2:** This cell displays the prediction result in a Streamlit-style modal window.

```
predict_price()
```

**Output:**

**Prediction**

**i** Predicted Price Category: 59.0 - 70.0 \$

OK

# VALUE PROPOSITION



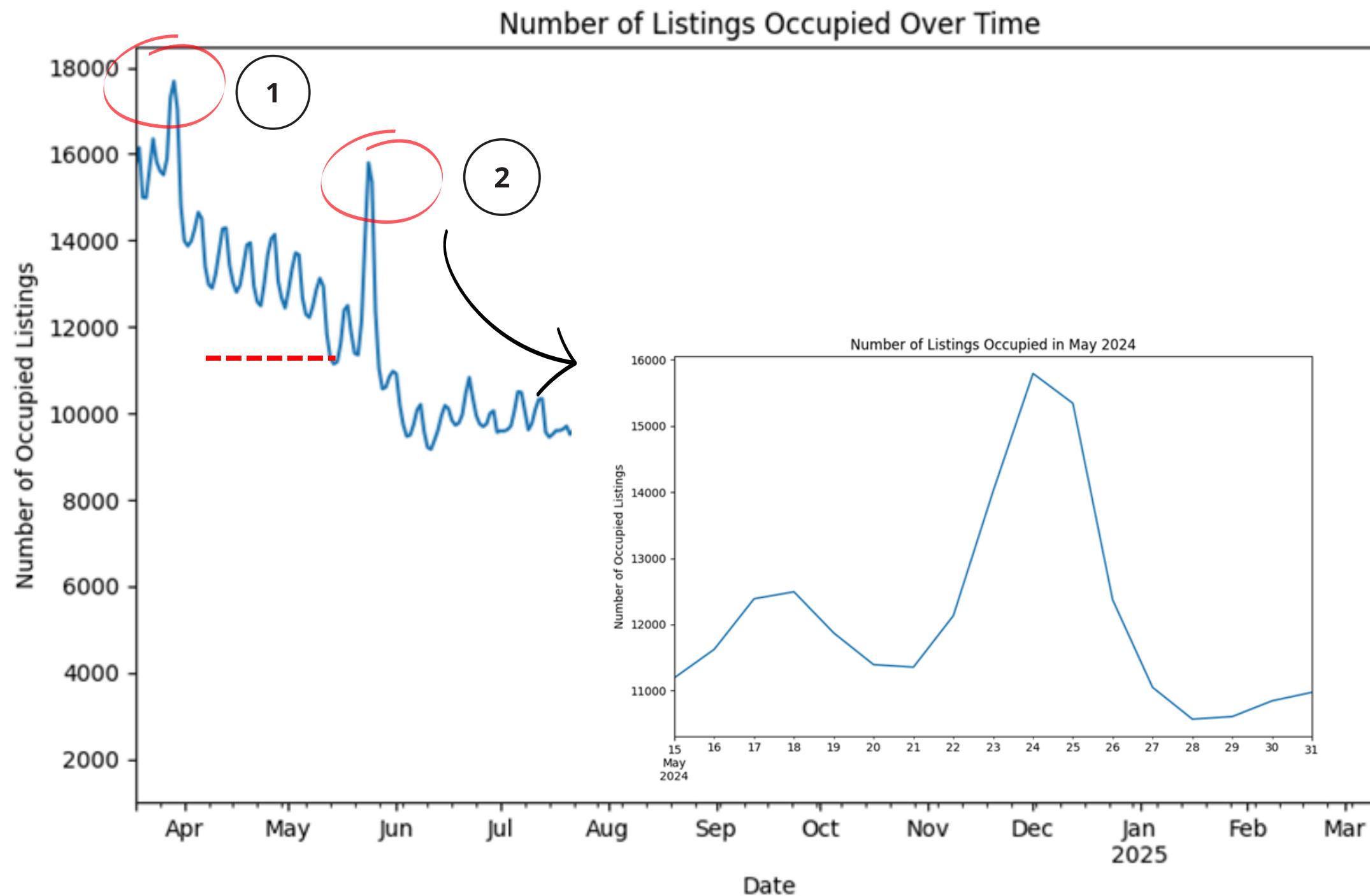
## For the Hosts:

Understand how **different characteristics** can affect the price of their accommodation;

Estimate **potential earnings**, as they have an idea of the price charged per night for accommodation with those characteristics;



# Time series analysis



1

**Holiday (Sexta-Feira Santa) followed by the weekend related to Easter**

(Possibility of **three consecutive days off work**)

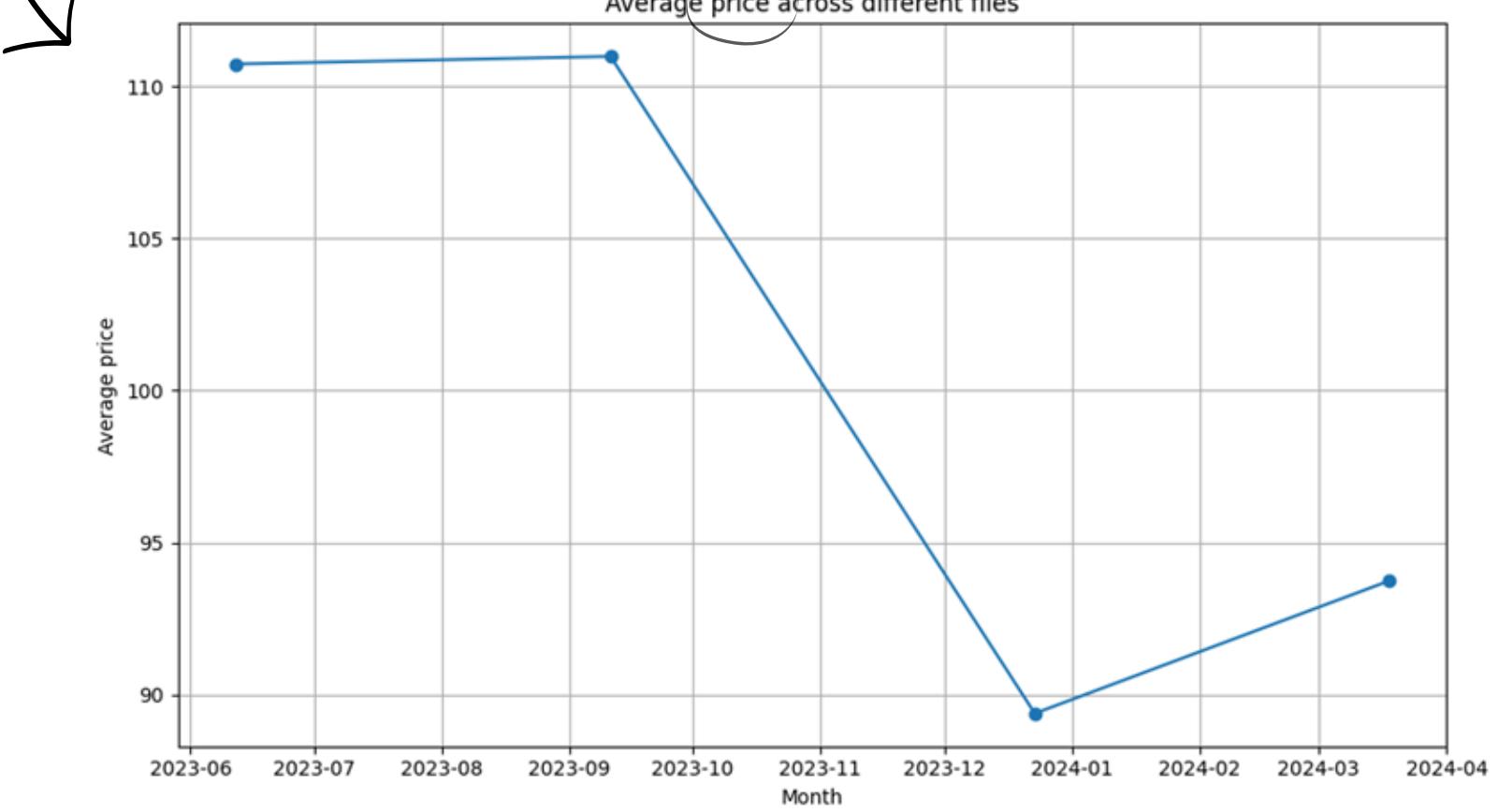
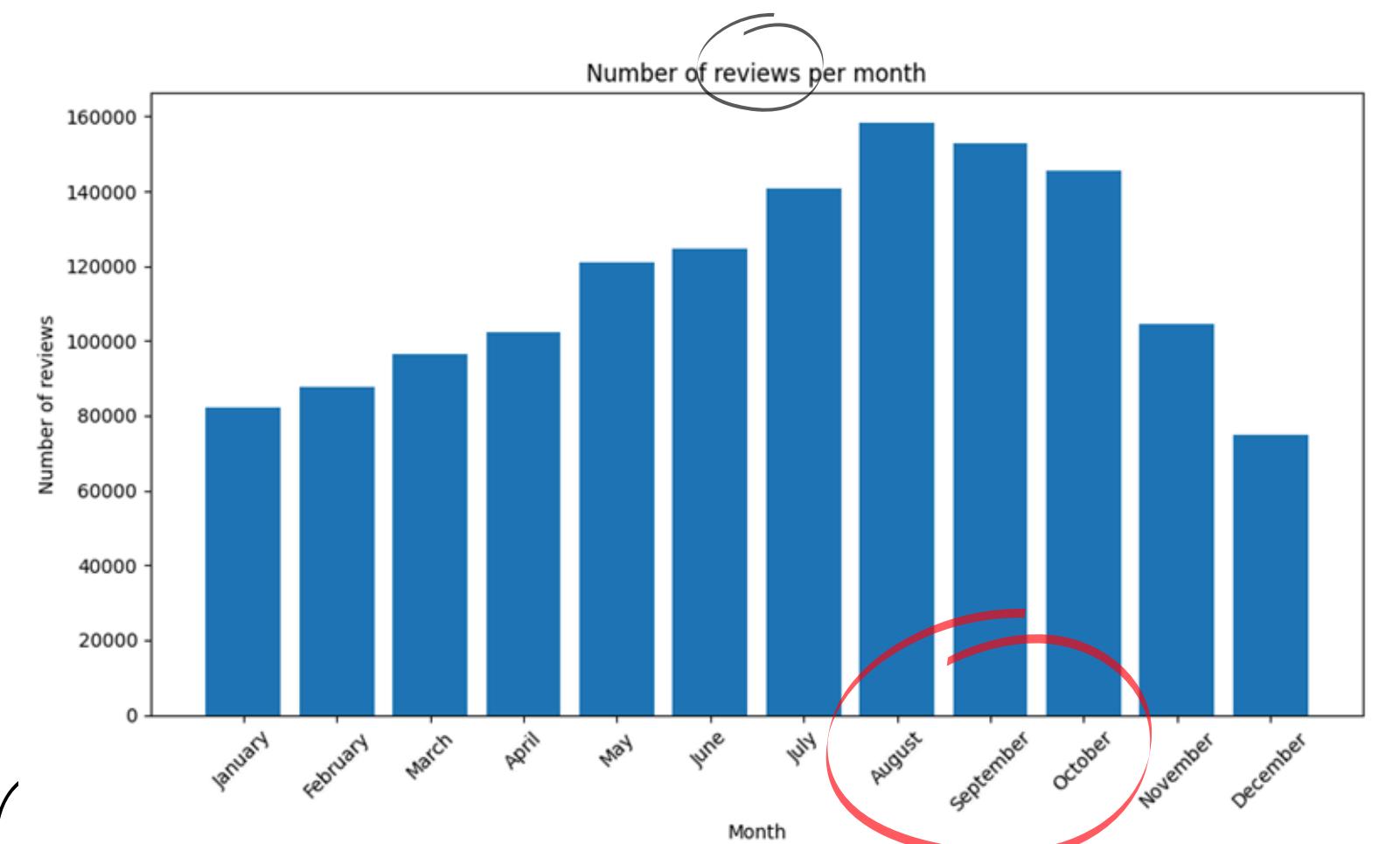
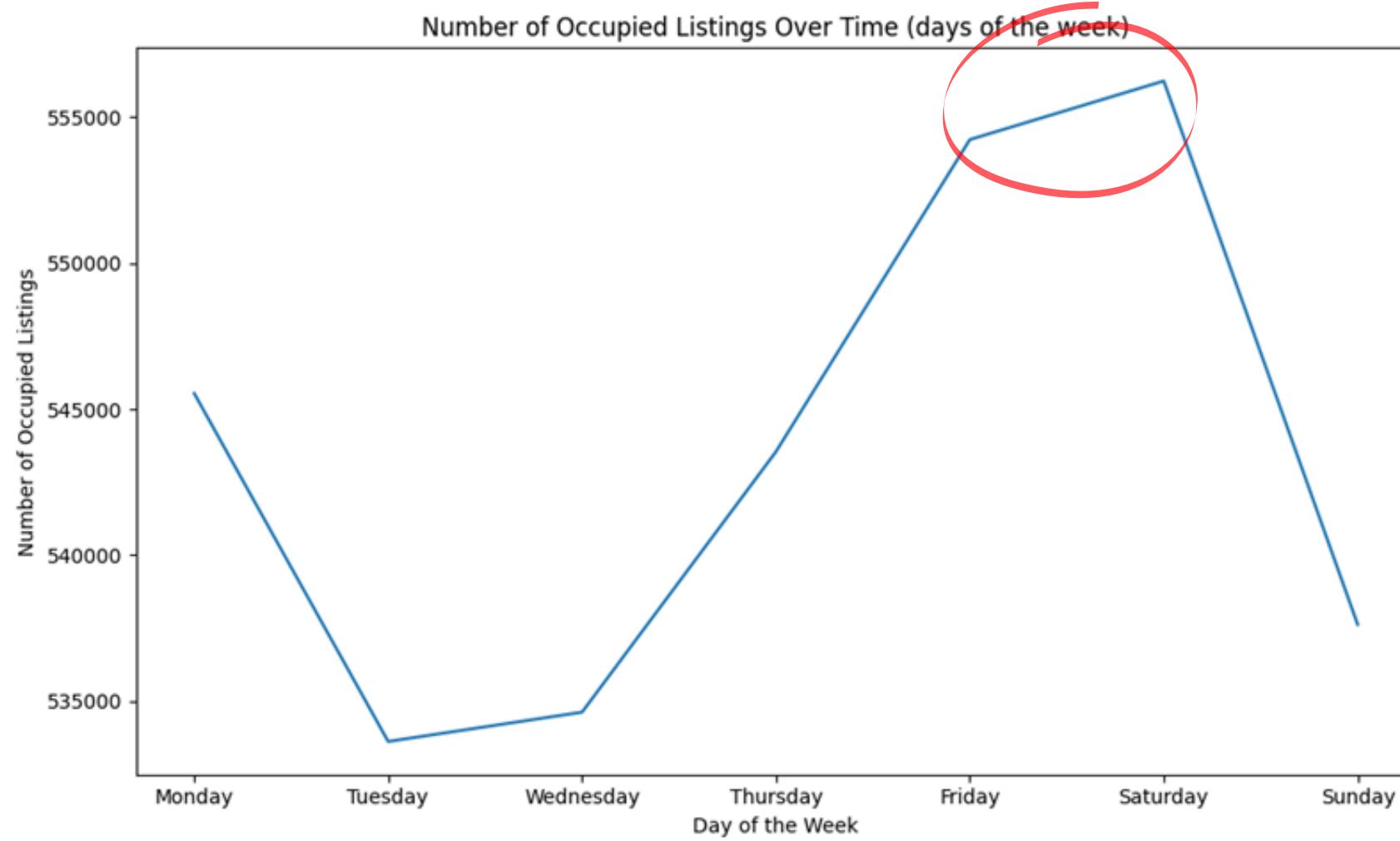
2



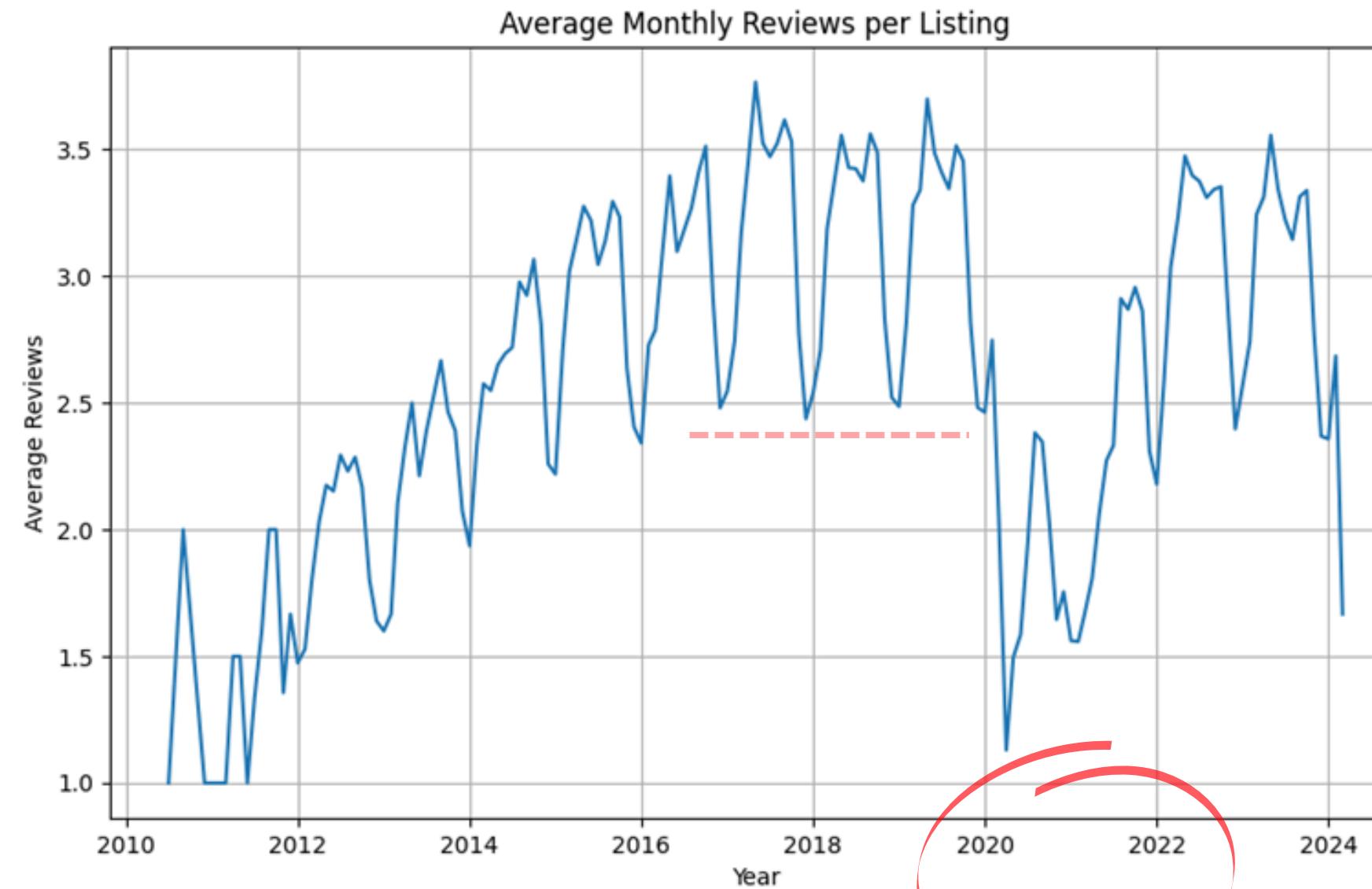
2<sup>nd</sup> SHOW ADDED!  
**TAYLOR SWIFT**  
**THE ERAS TOUR**

24 e 25 Maio 2024  
Estádio da Luz, Lisboa

# Time series analysis



# Time series analysis



P ECONOMY MARKETS COMPANIES BANK WORK AND EMPLOYMENT SITUATION PUBLIC FINANCES INTERNATIONAL

COVID-19

## Pandemic took 16.5 billion from Portuguese tourism in 2020

The drop in tourist demand, which includes consumption by foreigners and Portuguese, shows the “particularly harmful impacts of the pandemic” on this economic activity, highlights the INE.

Luis Villalobos  
May 14, 2021, 11:57 (updated on May 14, 2021, 12:32)

# Occupancy prediction (Time series analysis)

## Goal

Getting an idea of what to expect from the demand for accommodation throughout the different times of the year

## Features Used

(from reviews.csv)

id

listing\_id

date

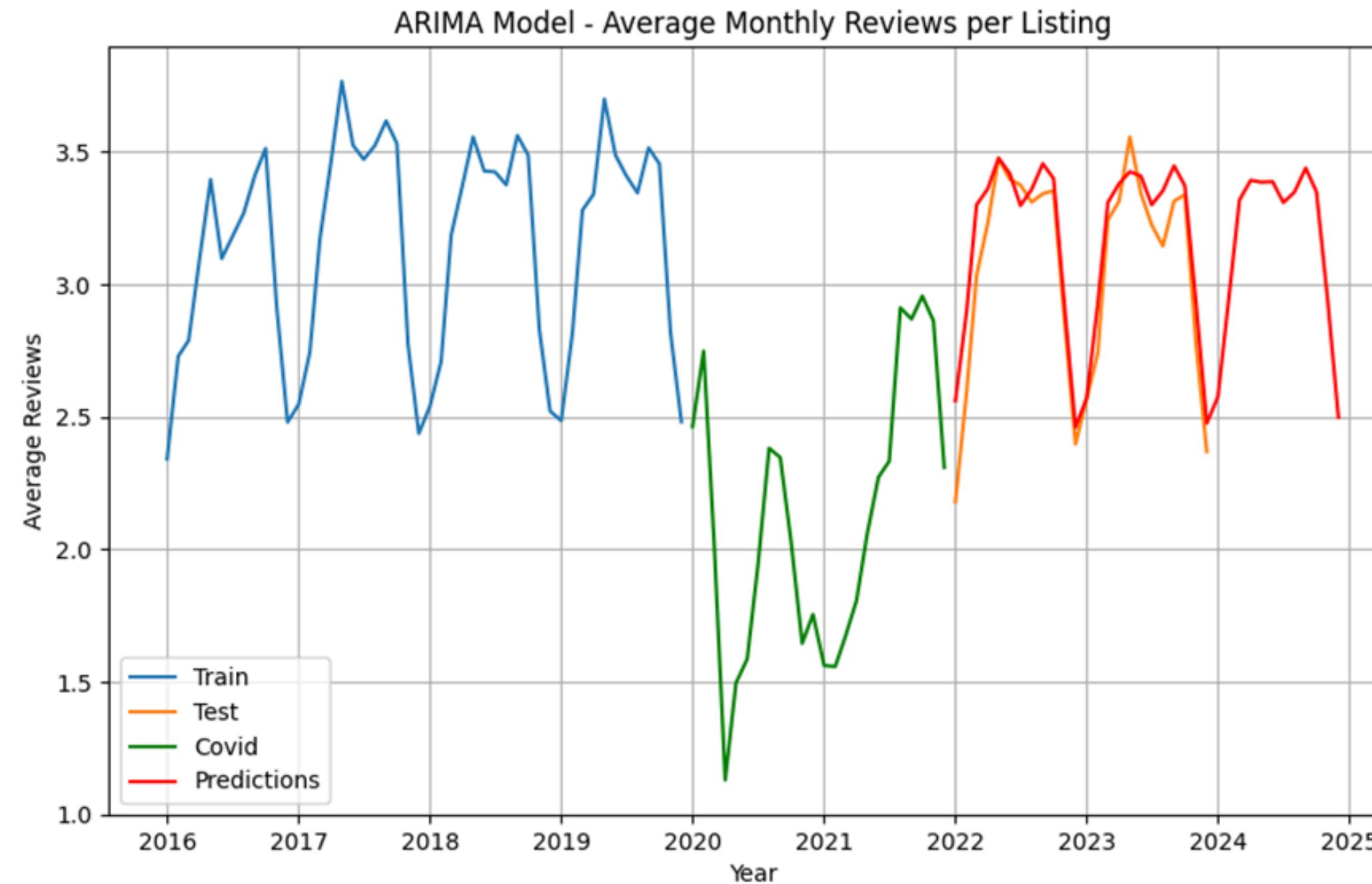
## Algorithm - ARIMA

Statistical analysis model that uses time series data to better understand the dataset and to predict future trends.

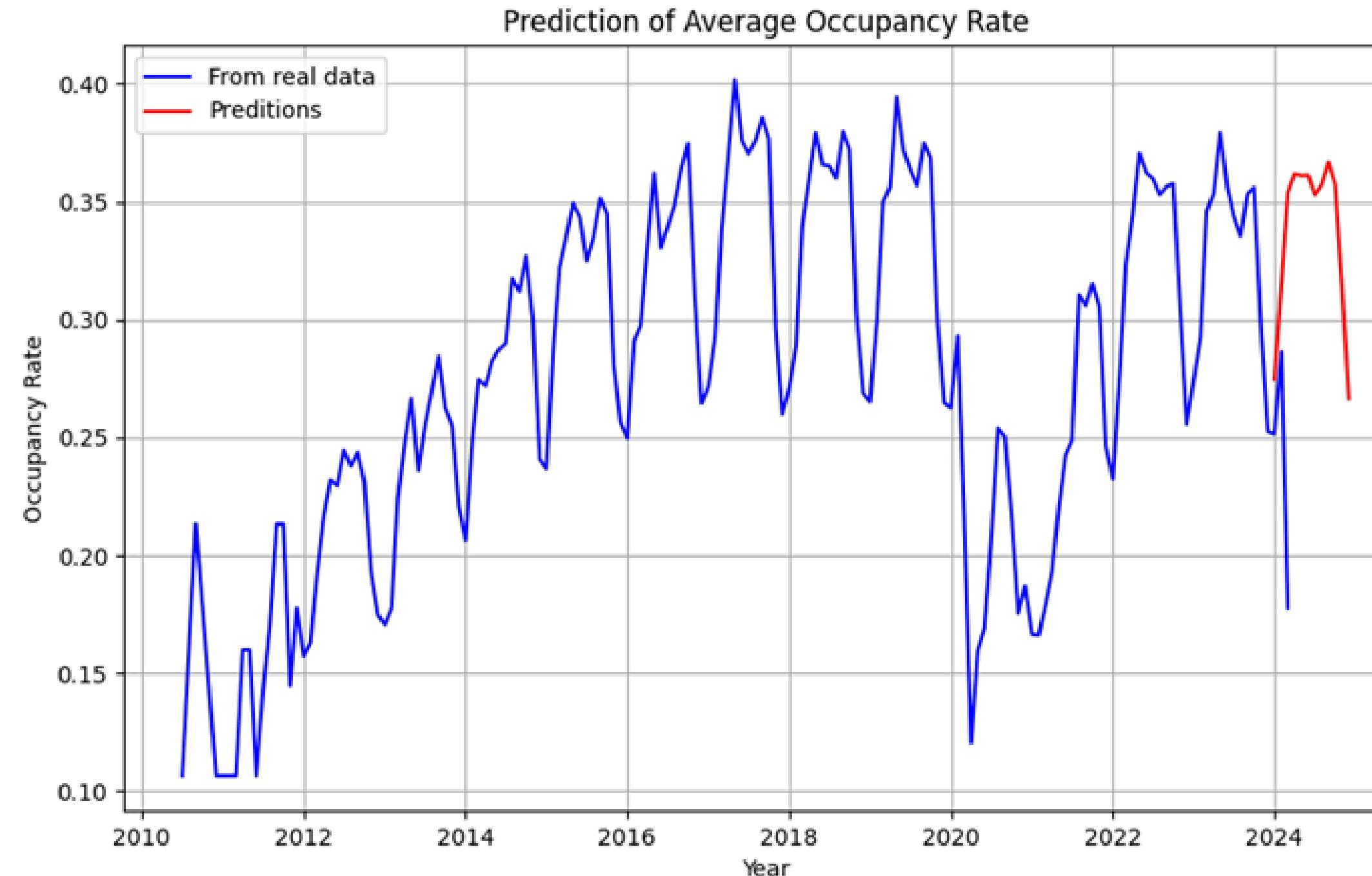
Uses a standard notation with p, d, and q for the parameters to indicate the type of ARIMA model used.

- **p:** the number of lag observations in the model;
- **d:** the number of times the raw observations are differenced;
- **q:** the size of the moving average window.

# Occupancy prediction (Time series analysis)



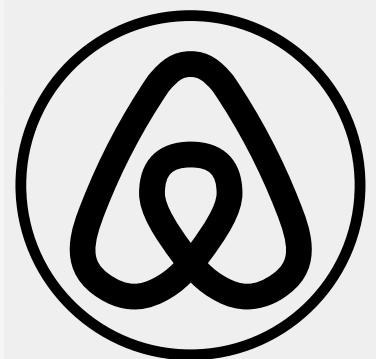
# Occupancy prediction (Time series analysis)



## Assumptions

- Each review corresponded directly to a stay at the accommodation.
- The average length of stay in the city of Lisbon is around 3.2 nights.

# VALUE PROPOSITION



## For the Airbnb:

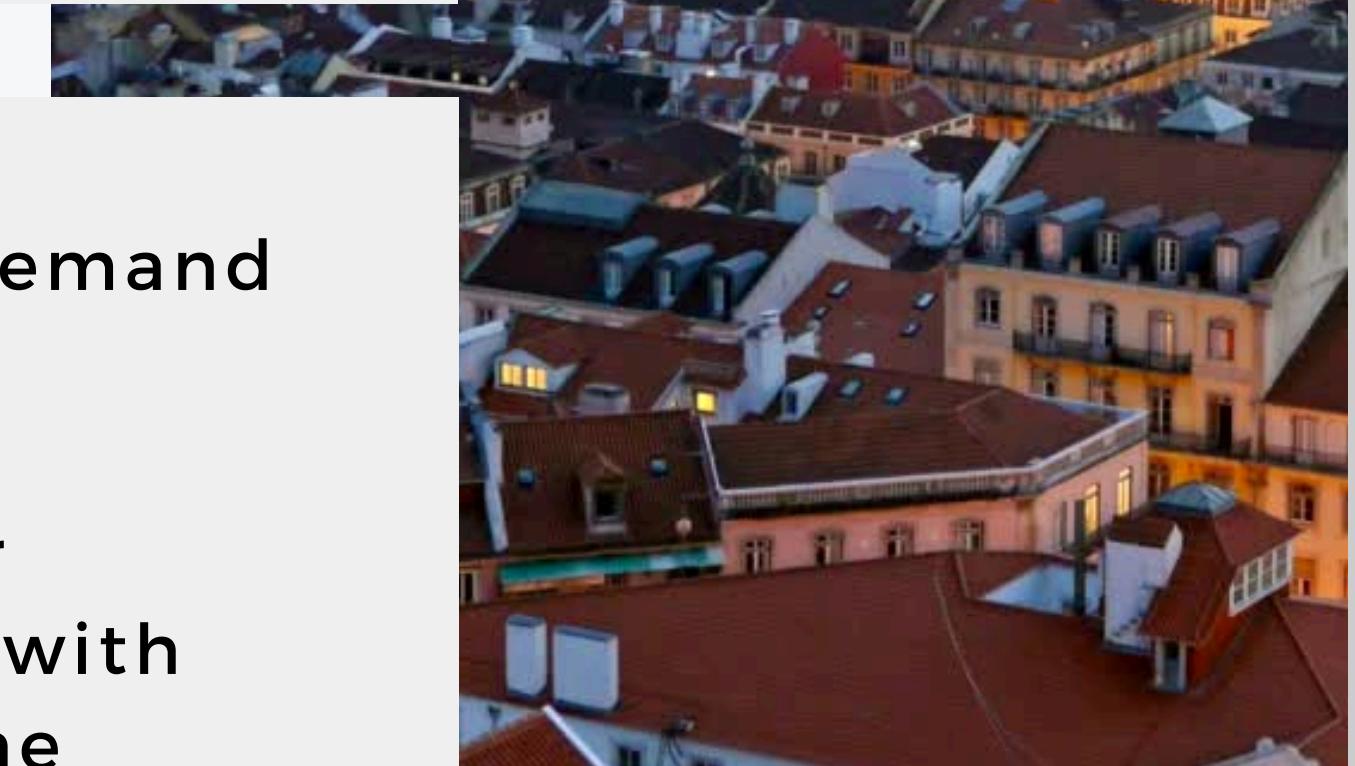
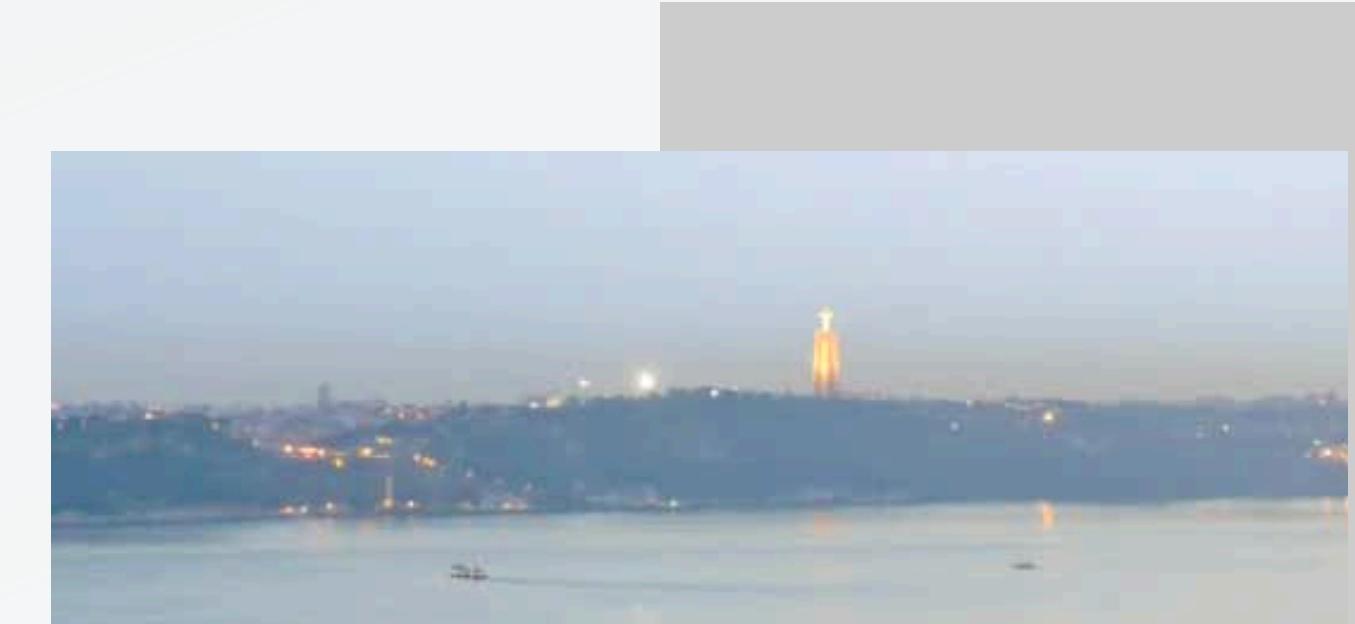
Obtaining statistical data and previewing brand promoting events.



## For the Hosts:

Adaptation of services to seasonal demand variation trends;

Get an idea of the state of your their business by comparing their results with the average statistics obtained by the other hosts.



# AIRBNB LISBON

Belong Anywhere

• • •  
• • •  
• • •

## Q&A

