Career Foundry Achievement 6 Project Brief: AirBnb renting trends in Berlin (data from March 2021)

1 - PROJECT DETAILS

Introduction

Airbnb offer the possibility to renting out a room or an entire flat through their platform service. Since their launch in 2008, the platform popularity increased till becoming one of the primary accommodation providers for travellers across the globe. The project will focus on Berlin, with 19857 listings (as of March 2021).

Objective:

To build an interactive dashboard showing the most popular neighbourhood and average price per overnight stay.

Context & Data requirements:

For this task, data have been open sourced from Airbnb inside:

http://insideairbnb.com/get-the-data.html

making it an official source, as data is directly collected from the platform on a monthly basis

Table of content

Column name	Column description
id	Listing ID assigned to the host
name	Housing´s name on Airbnb
host_id	Host ID assigned from Airbnb when registering to the platform
host_name	Host Name
neighbourhood_group	Location of the housing based on Berlin Areas
neighbourhood	more specific housing location based on Berlin districts
latitude	GPS coordinates
longitude	GPS coordinates
room_type	Housing category (flat/room)
price	Overnight price
minimum_nights	Minimum overnight stay offered by the host
number_of_reviews	Number of reviews collected by the host
last_review	Date of the last review
reviews_per_month	Number of reviews per month
calculated_host_listings_count	How many times the host_id is listed
availability_365	Housing available days per year

The AirBnb dataset has been chosen to analyze housing trends among the different Berlin areas and see if there is a specific pattern for each district. Interesting for the analysis would be to understand if there is a specifiv

The Kaggle dataset has been chosen to analyse trends among the different countries and see if they do reflect the continent trends. Interesting for the analysis would be also to understand if the price per night has an impact in the number of bookings and if there is a trend between private room and entire flat reservations.

Analysis criteria

The following analysis will be conducted:

Exploratory analysis through visualizations (scatterplots, correlation heatmaps, pair

plots and categorical plots)

- Geospatial analysis using a shapefile
- Regression analysis
- Cluster analysis
- Time-series analysis
- Analysis narrative and final results (presented in a dashboard)

2 - DATA PROFILE

Clean your data:

FOR dataset 'df_airbnb'

No duplicates were found.

NaN values have been found in following columns:

```
In [26]: # checking data consistency for df_airbnb
         df_airbnb.isnull().sum()
```

```
Out[26]: id                                0
         name                             32
         host_id                           0
         host_name                       932
         neighbourhood_group               0
         neighbourhood                     0
         latitude                          0
         longitude                         0
         room_type                         0
         price                             0
         minimum_nights                    0
         number_of_reviews                 0
         last_review                    4105
         reviews_per_month              4105
         calculated_host_listings_count    0
         availability_365                  0
         dtype: int64
```

Most of the columns can be dropped before conducting the analysis. In "review_per_month" the missing values can be replaced by '0' as the value is derived by the column "number_of_review".

Some basic descriptive analysis

```
In [25]: df_airbnb.describe()
```

Out[25]:

	id	host_id	latitude	longitude	price	minimum_nights	number_of_reviews
count	1.985800e+04	1.985800e+04	19858.000000	19858.000000	19858.000000	19858.000000	19858.000000
mean	2.428805e+07	8.927153e+07	52.510227	13.404362	70.778930	8.604240	21.918622
std	1.418866e+07	1.028498e+08	0.031944	0.062236	120.383995	30.954859	48.038176
min	1.944000e+03	1.581000e+03	52.340410	13.098390	0.000000	1.000000	0.000000
25%	1.201695e+07	1.126449e+07	52.489850	13.367832	35.000000	2.000000	1.000000
50%	2.319536e+07	4.251197e+07	52.509910	13.413860	50.000000	3.000000	4.000000
75%	3.745479e+07	1.383302e+08	52.533090	13.438897	80.000000	5.000000	18.000000
max	4.861566e+07	3.920622e+08	52.655980	13.757580	8000.000000	1124.000000	618.000000

```
In [25]: df_airbnb.describe()
```

Out[25]:

ongitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
8.000000	19858.000000	19858.000000	19858.000000	15753.000000	19858.000000	19858.000000
3.404362	70.778930	8.604240	21.918622	0.674096	3.148454	94.663964
0.062236	120.383995	30.954859	48.038176	1.131620	7.642956	131.877521
3.098390	0.000000	1.000000	0.000000	0.010000	1.000000	0.000000
3.367832	35.000000	2.000000	1.000000	0.090000	1.000000	0.000000
3.413860	50.000000	3.000000	4.000000	0.260000	1.000000	0.000000
3.438897	80.000000	5.000000	18.000000	0.770000	2.000000	178.000000
3.757580	8000.000000	1124.000000	618.000000	45.000000	73.000000	365.000000

Consider limitations and ethics: Not sure if the by combining the host_name and the longitude / latitude the data privacy of the host is guaranteed.

Since the information is saved by the host / user it could be biased from data or typing error, so the localization might be not 100% accurated.

3 – BUSINESS QUESTIONS

- Average overnight price over the city
  ~~*Average overnight price over the city is 70,77 Euro per night*~~
  There where some outlier influencing the prices. For better accuracy, price per night under 10 Euros and above 1000 Euros have been dropped from the dataset.
- Average overnight price for categories “private room” and “entire apartment”
- Average overnight price for each Berlin area
- Average calendar availability across Berlin
  Average calendar availability is 95 days per year

4 – HYPOTHESES

H1. If the accommodation is in a popular area, guests will be willing to pay a higher price for the accommodation

H2. If the availability per year is higher, more reviews can be counted in