Capstone Project Notebook:

Exploratory Analysis of the Airbnb DataSet at Rio de Janeiro

Juliane Nascimento

Introduction

Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present more unique, personalized way of experiencing the world.

Objective

The objective is exploring and clustering this dataset, in a way that I can identify the best places to stay at Rio de Janeiro based on venues around the accommodate. In this notebook I will explore interactions and hidden insights from the available dataset.

DataSet

Detailed Listings data for Rio de Janeiro:

http://data.insideairbnb.com/brazil/rj/rio-de-janeiro/2020-05-24/data/listings.csv.gz

The Dataset has 106 columns and 35.793 rows

```
# shape of the dataset
airbnb_df_complete.shape
```

```
(35793, 106)
```

	id	listing_url	scrape_id	last_scraped	name	summary	space	description	experiences_offered	neighborh
0	17878	https://www.airbnb.com/rooms/17878	20200524171540	2020-05-25	Very Nice 2Br in Copacabana w. balcony, fast WiFi	Discounts for long term stays. Please contact ...	- Beautiful, sunny 2 bedroom (90 square meters...	Discounts for long term stays. Please contact ...	none	This i bests spo
1	21280	https://www.airbnb.com/rooms/21280	20200524171540	2020-05-25	Renovated Modern Apt. Near Beach	Immaculately renovated top-floor apartment ove...	Immaculately renovated top-floor apartment in ...	Immaculately renovated top-floor apartment ove...	none	neighb

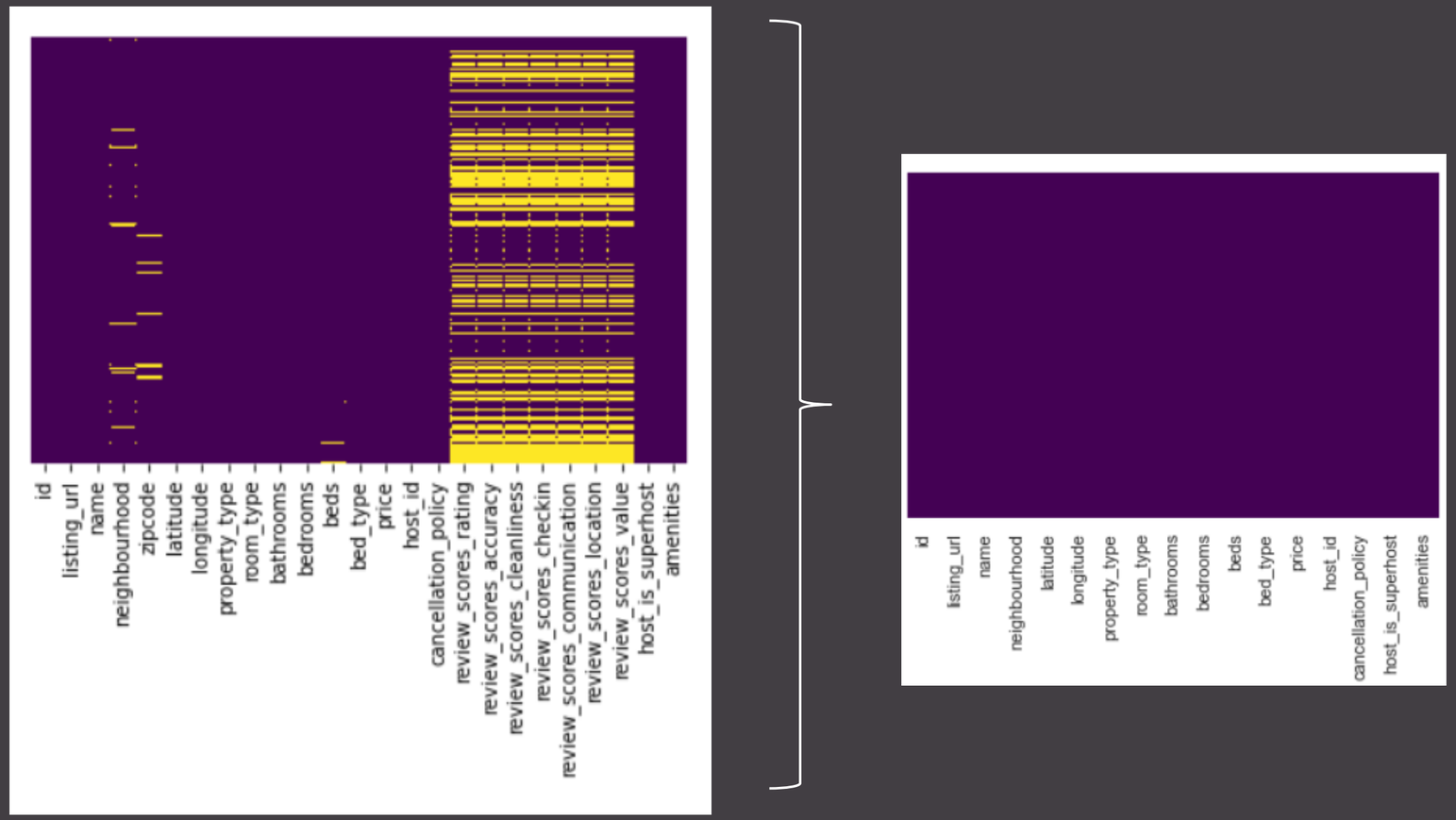
Exploratory Data Analysis and Methodology

Relevant Columns

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 35793 entries, 0 to 35792
Data columns (total 25 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   id                           35793 non-null  int64
 1   listing_url                  35793 non-null  object
 2   name                         35735 non-null  object
 3   neighbourhood                33907 non-null  object
 4   zipcode                      34049 non-null  object
 5   latitude                     35793 non-null  float64
 6   longitude                    35793 non-null  float64
 7   property_type                35793 non-null  object
 8   room_type                    35793 non-null  object
 9   bathrooms                    35726 non-null  float64
 10  bedrooms                     35720 non-null  float64
 11  beds                         35464 non-null  float64
 12  bed_type                     35793 non-null  object
 13  price                        35793 non-null  object
 14  host_id                      35793 non-null  int64
 15  cancellation_policy          35793 non-null  object
 16  review_scores_rating         20002 non-null  float64
 17  review_scores_accuracy       19983 non-null  float64
 18  review_scores_cleanliness    19988 non-null  float64
 19  review_scores_checkin        19982 non-null  float64
 20  review_scores_communication  19990 non-null  float64
 21  review_scores_location       19984 non-null  float64
 22  review_scores_value          19986 non-null  float64
 23  host_is_superhost            35788 non-null  object
 24  amenities                    35793 non-null  object
```

To achieve my goals, I will just use 25 columns that have the essential data about, location, price, score rating, score rating to location, type of room and geographic location.

Missing Values

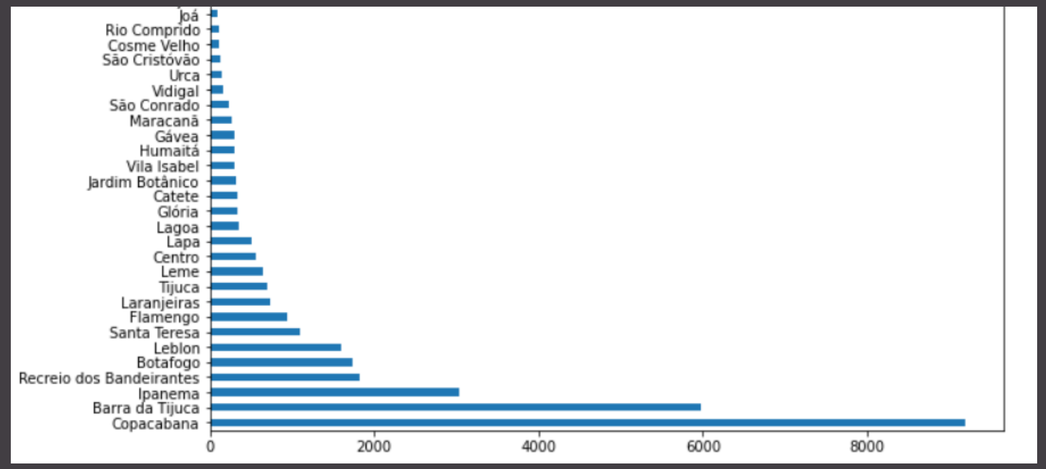


- 'zipcode': Other coluns have this information with more quality
- 'review_scores_rating','review_scores_accuracy','review_scores_cleanliness','review_scores_checkin', 'review_scores_communication','review_scores_location','review_scores_value': I will drop this columns because more than of 50% of the information is missing
- 'name', 'neighbourhood', 'bathrooms','bedrooms','beds', 'host_is_superhost': I will ignore the missing value and drop the rows that does not have information

Exploratory Data Analysis and Methodology

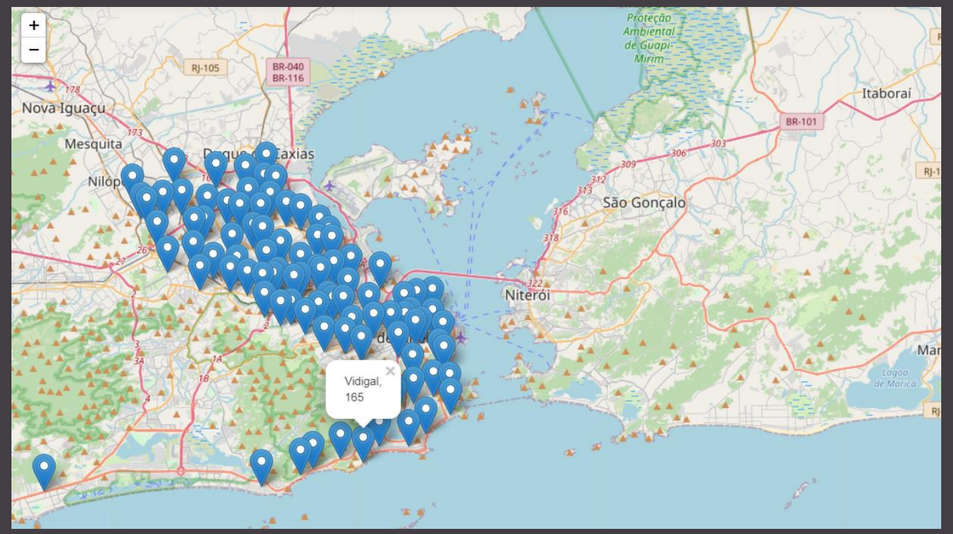
Neighbourhood feature

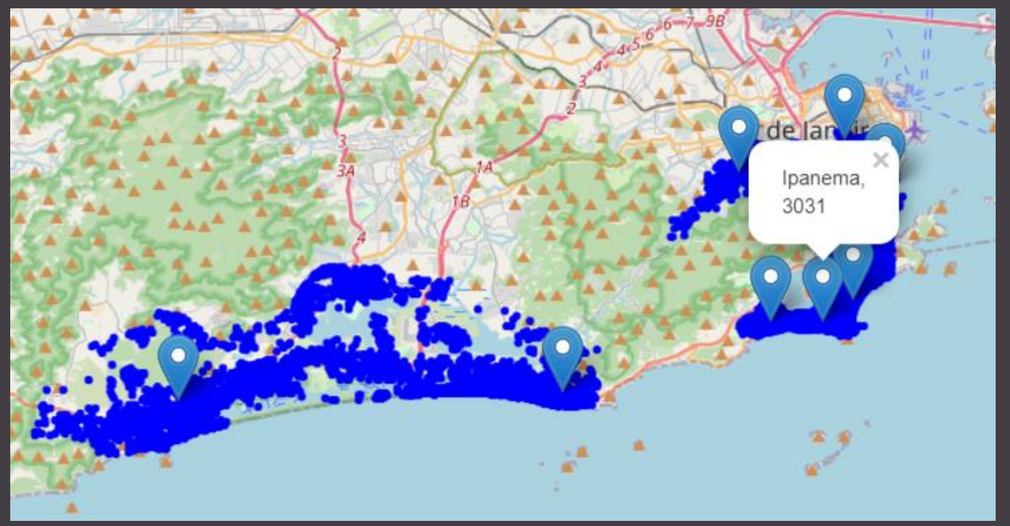
Amount of Airbnbs in the neighborhoods of Rio de Janeiro



Top 10 Neighborhoods that have the largest amount of Airbnb

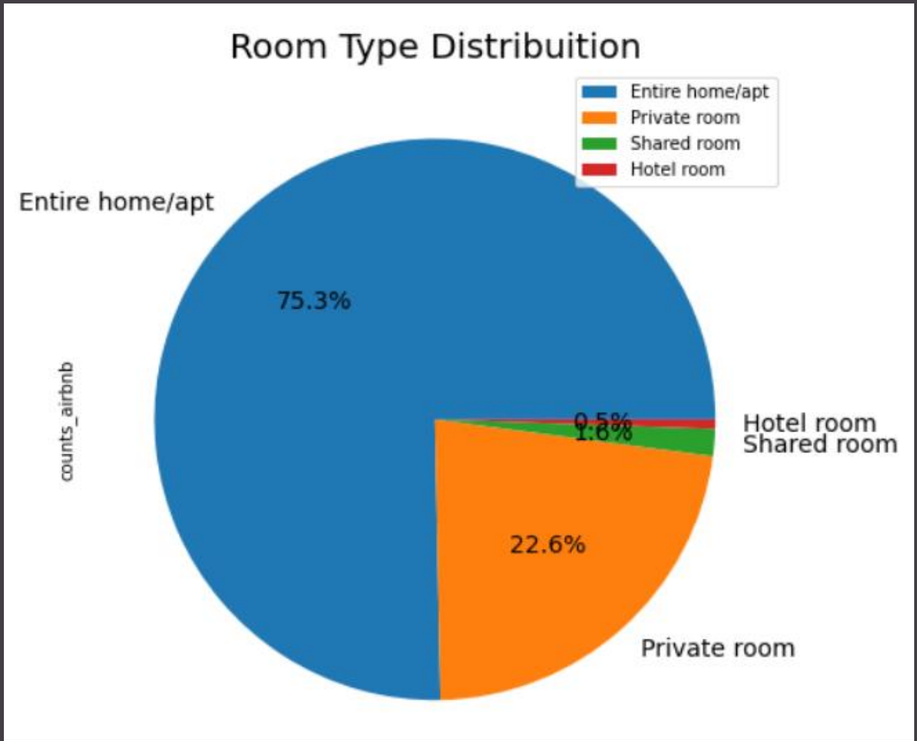
	neighbourhood	counts_airbnb	latitude	longitude
0	Copacabana	9194	-22.97593	-43.18960
1	Barra da Tijuca	5982	-23.01071	-43.30837
2	Ipanema	3031	-22.98441	-43.20180
3	Recreio dos Bandeirantes	1824	-23.01385	-43.46526
4	Botafogo	1729	-22.95138	-43.18413
5	Leblon	1596	-22.98428	-43.22270
6	Santa Teresa	1099	-22.91627	-43.19250
7	Flamengo	936	-22.93409	-43.17604
8	Laranjeiras	727	-21.63333	-41.68333
9	Tijuca	707	-22.92780	-43.23608





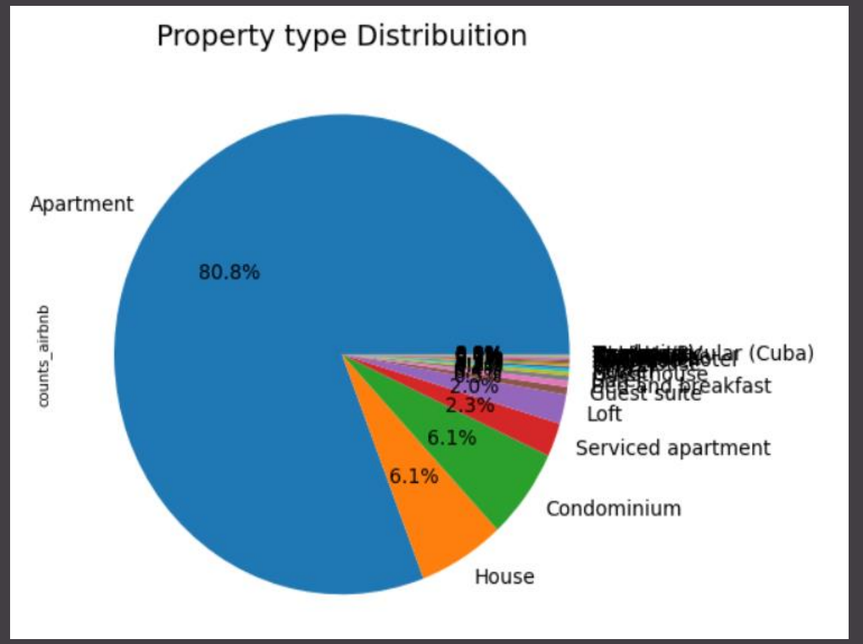
Exploratory Data Analysis and Methodology

Room type feature



The highest concentration of the room type is Entire home/apt with 75%. I believe this happens because guests like to have more privacy

Property type feature

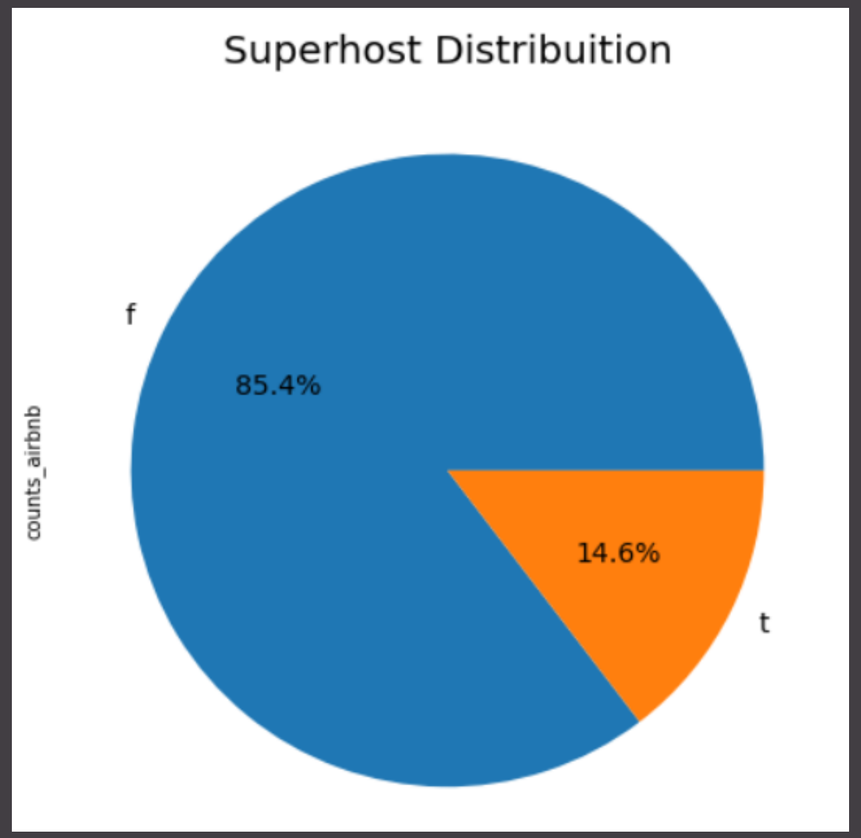


	property_type	counts_airbnb
0	Apartment	21683
1	House	1645
2	Condominium	1635
3	Serviced apartment	611
4	Loft	530

The highest concentration of the property type is Apartment with 80,8%; House with 6,1%; Condominum with 6,1% with 75%. I believe this happens because guests wants more comfort and safety

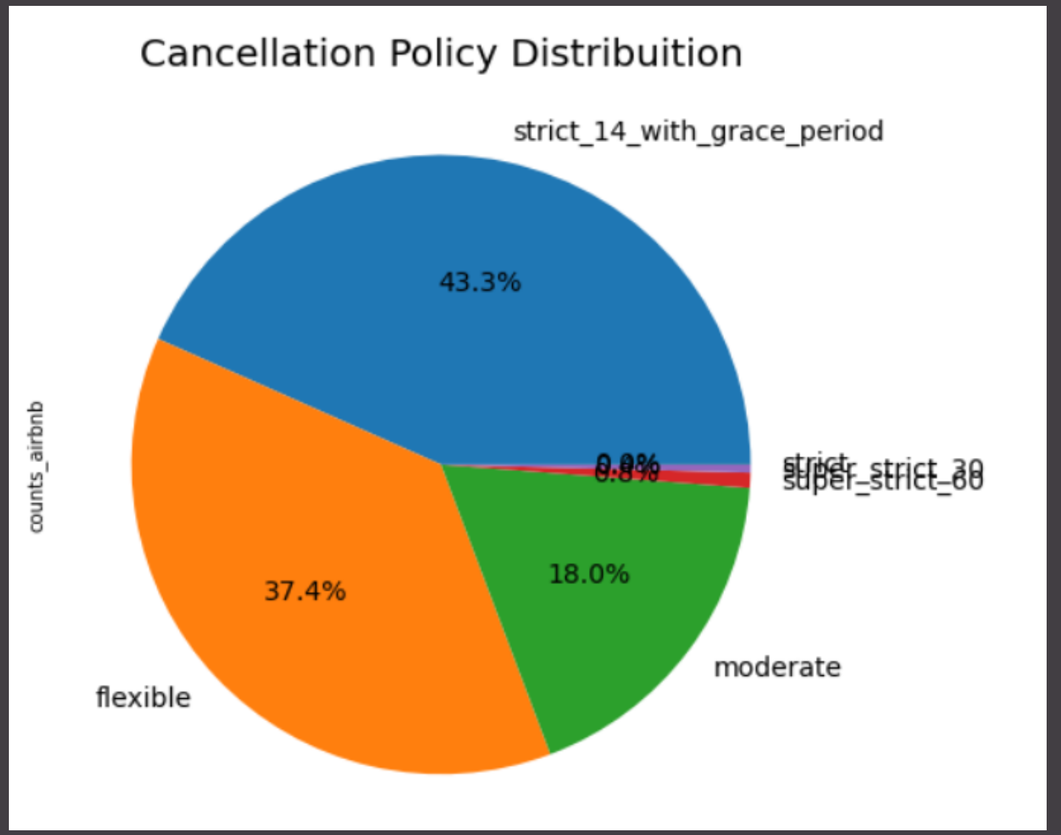
Exploratory Data Analysis and Methodology

Superhost feature



"Superhosts are experienced hosts who are an excellent example for other hosts and receive extraordinary experiences for their guests“
From: https://www.airbnb.com.br/help/article/828/o-que-%C3%A9-um-superhost

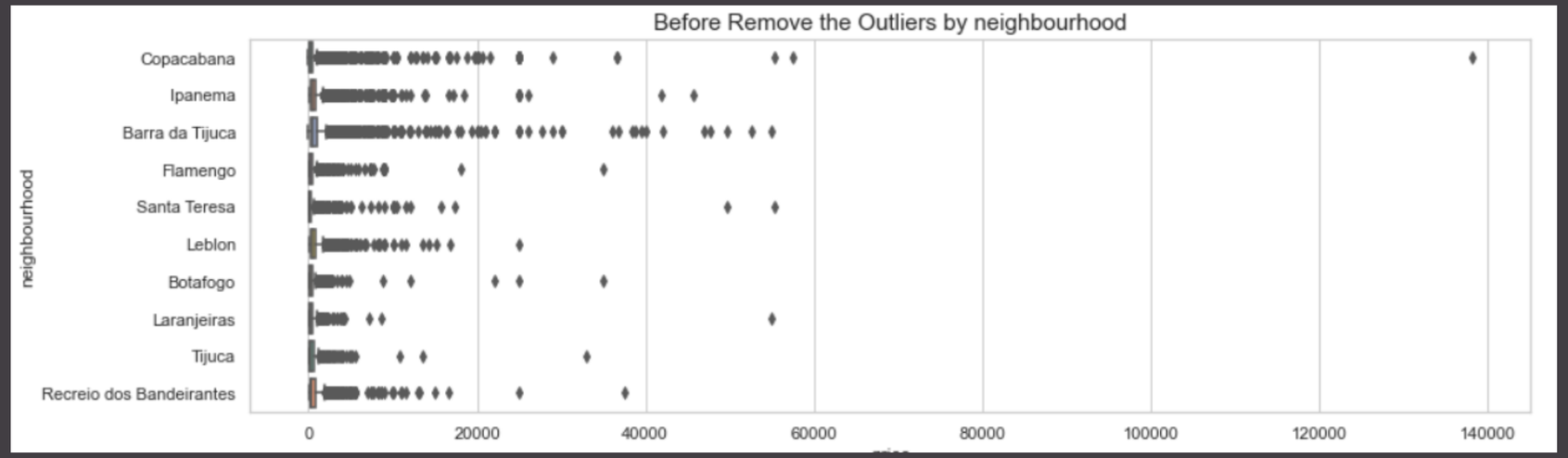
Cancellation Policy feature

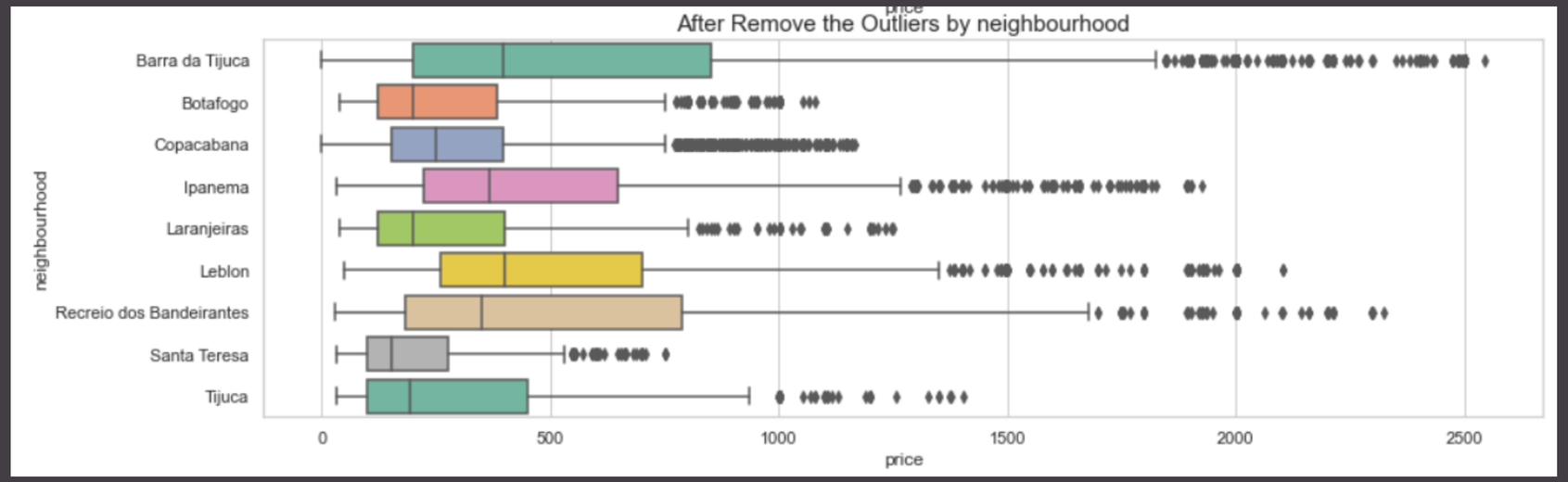


Basically, depending on the cancellation policy if you make the cancellation you can have the full refund or not the reservation. In the top 10 neighborhoods, 37,4% has a flexible policy and 43,3% has a strict 14 with grace period, meaning you can cancel up to 14 days before check-in and get a full refund

Exploratory Data Analysis and Methodology

Price feature





It’s possible to identify that the prices per neighborhood have many outliers. In some cases the daily rate costs more than $ 10,000

As a result of oulier removal 3020 rows of data were removed.

FourSquare Data

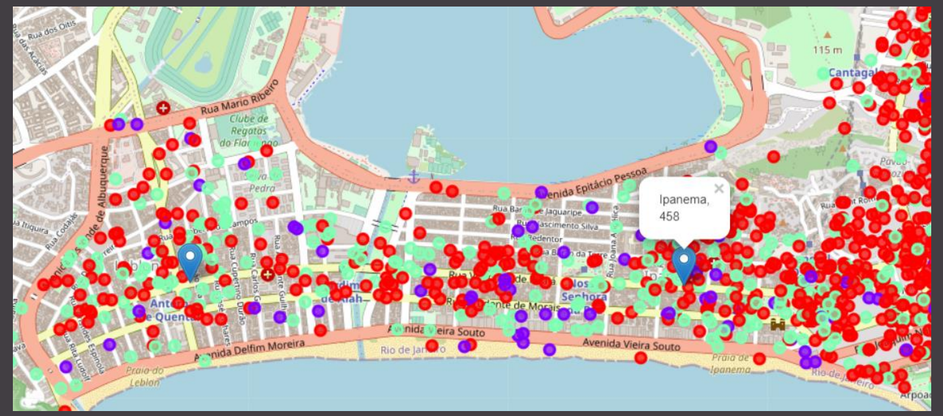
Top10 Categories by Neighborhood

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Barra da Tijuca	Hotel	Beach	Italian Restaurant	Bar	Steakhouse	Beach Bar	Gym / Fitness Center	Bistro	Chocolate Shop	Pizza Place
1	Botafogo	Brazilian Restaurant	Pizza Place	Hostel	Bookstore	Ice Cream Shop	Steakhouse	Gym / Fitness Center	Fruit & Vegetable Store	Bar	Dance Studio
2	Copacabana	Hotel	Bar	Bathing Area	Gym / Fitness Center	Ice Cream Shop	Coffee Shop	Juice Bar	Park	Fruit & Vegetable Store	Spa
3	Ipanema	Hotel	Bathing Area	Gym / Fitness Center	Beach	Bar	Brazilian Restaurant	French Restaurant	Plaza	Café	Italian Restaurant
4	Leblon	Steakhouse	Food Stand	Italian Restaurant	Bar	Ice Cream Shop	Bookstore	Bathing Area	Gym / Fitness Center	Fruit & Vegetable Store	Hotel
5	Recreio dos Bandeirantes	Gym / Fitness Center	Bakery	Brazilian Restaurant	Steakhouse	Snack Place	Pizza Place	Bar	Bathing Area	Japanese Restaurant	Park
6	Santa Teresa	Brazilian Restaurant	Hostel	Bar	Music Venue	Plaza	Historic Site	Middle Eastern Restaurant	Pizza Place	Chocolate Shop	Café
7	Tijuca	Bar	Gym	Pizza Place	Bakery	Chocolate Shop	Gym / Fitness Center	Coffee Shop	Bookstore	Pet Store	Japanese Restaurant

Cluster and Results

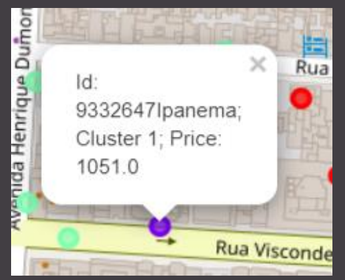
Top10 Categories by Neighborhood

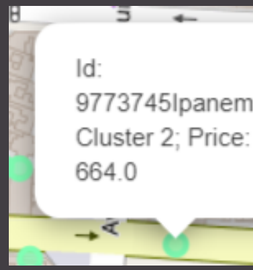
Ipanema Neighborhood



Cluster 0
the average price low

Cluster 1
the high
average price





Cluster 2
the average pr
normal

Conclusion

This notebook was used to explore a dataset of Rio de Janeiro Airbnb's at May 2020. It's a dataset with a variety of columns that allowed me to do deep data exploration. First, I found the Top 10 neighborhoods that have the highest amount of airbnb. After that, I analyzed some features like: property type, room type, cancellation policy, superhost flag and price. I had to work with the pricing resource outliers and removed them from my dataset. Next, I applied some filters that I thought could may be relevant to my analysis.

After that, I used the FourSquare API to find the main places aroun the top 10 neighborhood. Finally, I used Kmeans to find groups. Overall, I found a very good number of interesting relationships between data set resources and locations around the neighborhood.

This project allowed me to learn more about the distribution of venues that are rented through the Airbnb platform at Rio de Janeiro.

Thanks! ;)