

# Where to Install a New Hostel In Rio de Janeiro?

*A Data Analysis Approach*

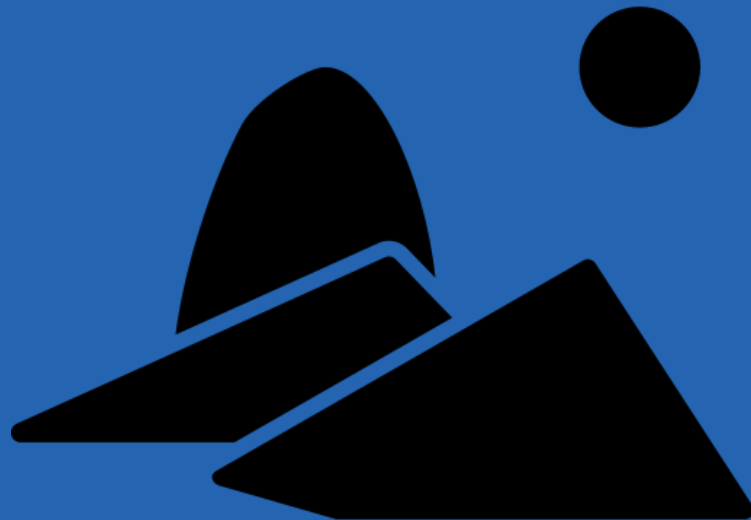
---

Capstone Project

---

IBM Data Scientist Professional Certificate

*July, 2019*



Felipe N. Ruperti



**Data Science**

Powered By  
**coursera**

# Summary

---

<b>1</b>	<b>Executive Summary</b>	<b>4</b>
<b>2</b>	<b>Introduction</b>	<b>5</b>
<b>3</b>	<b>Methodology</b>	<b>6</b>
3.1	Data Preparation . . . . .	6
3.2	Descriptive Analysis . . . . .	8
3.3	K-Means . . . . .	8
3.4	Ranking . . . . .	9
<b>4</b>	<b>Results</b>	<b>11</b>
4.1	A Descriptive Analysis . . . . .	11
4.2	The K-Means Model . . . . .	15
4.3	The Ranking . . . . .	17
<b>5</b>	<b>Discussion</b>	<b>18</b>
<b>6</b>	<b>Conclusion</b>	<b>20</b>
<b>A</b>	<b>Image's Copyright</b>	<b>22</b>

---

# Executive Summary


---

This report is a result of the IBM Data Scientist Professional Certificate Capstone Project. The scope is the tourism sector and the main objective is to analyze where are the best spots in the city of Rio de Janeiro for the installment of a new hostel. The methodology employed is mainly based on exploratory data analysis applied to spatial data from Foursquare. As Rio is a huge city, the scope of the analysis was restricted to the *Zona Sul* (South Zone), *Centro* (Downtown) and *Grande Tijuca* districts.

In the first part, heatmaps, histograms and the summary of certain key statistics were intensively used in order to obtain insight into the characteristics of many of Rio de Janeiro's neighborhoods. The k-means method was applied to group the neighborhoods according to their similarities/dissimilarities. In the second part, a ranking of the best neighborhoods, based on three different metrics, was created from the min-max normalization method.

Results indicate that the top 20 best locations for a new hostel are all located in the South Zone and Downtown districts. As the ranking penalizes competition from other installed hostels from a 350 meter radius, the ranking was able to spot some interesting unexplored opportunities, specially in Downtown Rio, but, also, in some upscale neighborhoods in the touristic South Zone. Regarding the 5 top spots, the report recommends:

 South Zone: parts of Leblon, Leme and Gávea.

 Downtown: Cinelândia and parts midway of Cinelândia to Lapa (*Praça Passeio Público*).

Some important limitations of the data are: lack of information regarding security and rent prices for the neighborhoods/locations and the quality of the venues (as restaurants). A specially useful information would be to have the crime rate and the rent price of each neighborhood. Another important limitation is that the database with the spatial informations (latitude and longitude) of the neighborhoods in Rio was collected manually, being subject to some error. Nevertheless, the results found are consistent to Rio de Janeiro's touristic reality. Lastly, it should be noted that different weights for the metrics and other choices for the radius would modify the positions in the ranking. Nevertheless, the results are consistent with Rio de Janeiro's reality.

# Introduction

---

Despite being considered the Brazilian capital of tourism, Rio de Janeiro is a city with vastly unexplored touristic opportunities. According to the [Price of Travel](#) ranking the number of touristic accommodations for the year of 2011, Rio was placed on the 77th position among 90 internationally famous cities, with only 75 bed & breakfast and hostel options. Taking this into account this report aims to provide a data analysis approach in order to appoint the neighborhoods in the city with higher potential for the installment of a new hostel.

# Methodology

---

The methodology is divided into four parts.

- **Data Preparation:** presents how the dataset was obtained as well as some key aspects of the data.
- **Descriptive Analysis:** gives a brief overview of the statistics and graphs used in order to have a better understanding of the data. The descriptive analysis, *per se*, is shown in the result section.
- **K-means:** explains the parameters and how the optimal number of cluster was obtained.
- **Ranking:** explains its main features and how it was built.

## 3.1. Data Preparation

The dataset can be found on my [GitHub](#) webpage and was created following two steps:

1. The initial data consists of 139 geospatial points/spots, called for simplification as neighborhoods<sup>1</sup>, which had their latitude and longitude gathered manually from the [GPS Coordinates](#) website. Besides, each neighborhood was assigned to its corresponding districts.
2. The initial data latitude/longitude of each point was used in order to obtain two variables: main/primary and specific/secondary venue types from the [Foursquare](#) developer api. The radius from each point was set in 350 meters and the limit of venues retrived per neighborhood was set in 500<sup>2</sup>. The venues were then merged with the neighborhoods/districts data frame of the previous step. As there are many venues per neighborhood, the dataset has a total of 4,737 observations and 8 columns/variables. Table 3.1 shows the first 3 observations of the dataset.

---

<sup>1</sup>Many observations account for more than one neighborhood, given their dimensions.

<sup>2</sup>The limit was not reached, as can be seen in the table.

	Neighborhood	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Main Venue Category	Venue Category
1	Leme 1	-22.962373	-43.168171	Bar do David	-22.960691	-43.168878	nightlife	Bar
2	Leme 1	-22.962373	-43.168171	Ponta do Leme	-22.962557	-43.166152	parks_outdoors	Beach
3	Leme 1	-22.962373	-43.168171	Calçadão do Leme	-22.962679	-43.167250	parks_outdoors	Pedestrian Plaza

Table 3.1: First 3 Observations of the Dataset

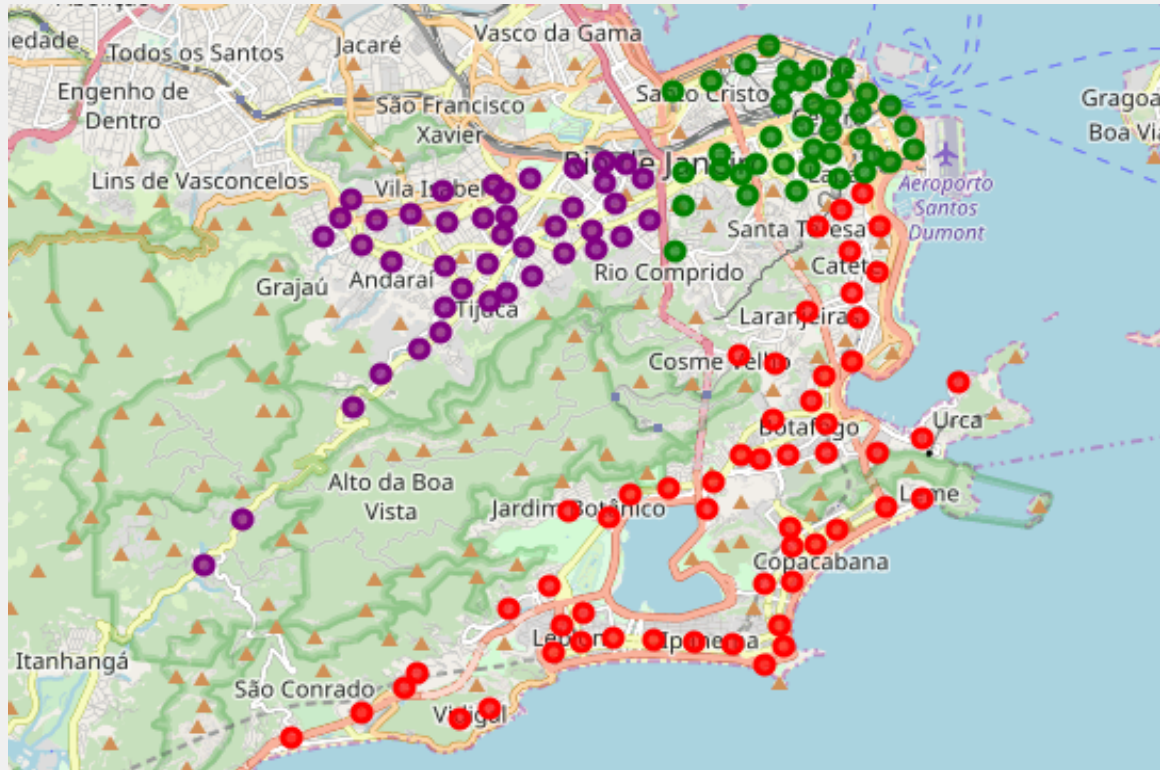


Figure 3.1: Rio's Districts - Neighborhoods

The geospatial graphs were rendered using the folium package from Python. Figure 3.1 shows all the points or neighborhoods based on their latitudes and longitudes and colored according to their districts: South Zone is in red, Downtown is in green and the *Grande Tijuca* is in purple. The table 3.2 on the right summarizes the number of points/neighborhoods per district.

	District	N <sup>o</sup> of Neighborhoods
1	Centro	39
2	Grande Tijuca	42
3	Zona Sul	58

Table 3.2: Number of Neighborhoods per District

## 3.2. Descriptive Analysis

Descriptive analysis was used in order to have a better understanding of the venues and its spatial distribution across neighborhoods. Statistics as mean, median, mode and graphs as heatmaps, bar charts and histograms were employed to enhance insight.

## 3.3. K-Means

The unsupervised data clustering was chosen in this report in order to have further insight into the data, by grouping the neighborhoods according to their secondary venue types. The k-means was the algorithm herein applied for data clustering, because it is simple to understand and deploy<sup>3</sup>. The euclidian distance was selected in order to quantify intra and inter cluster distances. In the data initialization the k-means++ algorithm developed by [Arthur and Vassilvitskii \(2007\)](#) was employed to improve efficiency. The k-means algorithm was run iteratively from 3 until 20 clusters,  $k = \{x \in \mathbb{N} \mid 3 \leq x \leq 20\}$ , and the Silhouette-Elbow method was applied in order to chose the best value of  $k$ , based on the Total Within Sum of Squares.

Before running the k-means, the dataset was filtered, keeping only the neighborhoods containing 5 or more venues. There were 5 observations excluded, as can be seen on table 3.3.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Counts
1	Alto da Boa Vista I	-22.965528	-43.277408	4
2	Catumbi II	-22.916378	-43.196243	4
3	Tijuca - Usina I	-22.943596	-43.255090	4
4	Urca 2	-22.944836	-43.162149	4
5	Sambódromo	-22.913160	-43.197164	1

Table 3.3: Neighborhoods with Less than 5 Observations

The justification for the previous procedure is related to the cluster analysis step (after the k-means is run), where a description for each cluster is given based on the 5 most common venues of each neighborhood assigned to that cluster, as in shown, for exemplification, on table 3.4. It should be noted that the description/labeling of each cluster is the last part of the k-means procedure.

<sup>3</sup>For further details regarding the k-means algorithm, please refer to ([Lantz, 2015](#), pp. 289-294).



	Neighborhood	District	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Av. Franklin Roosevelt	Centro	Brazilian Restaurant	Restaurant	Hotel	Vegetarian / Vegan Restaurant	Juice Bar
1	Av. Marechal Floriano	Centro	Brazilian Restaurant	Italian Restaurant	Middle Eastern Restaurant	Restaurant	Miscellaneous Shop
2	Av. Salvador de Sá	Centro	Brazilian Restaurant	Coffee Shop	Restaurant	Bistro	Soccer Stadium
3	Botafogo 1	Zona Sul	Bar	Brazilian Restaurant	Café	Vegetarian / Vegan Restaurant	Restaurant
4	Botafogo 2	Zona Sul	Brazilian Restaurant	Vegetarian / Vegan Restaurant	Nightclub	Health Food Store	Snack Place

Table 3.4: Sample of 5 Neighborhoods from Cluster 0: Top 5 Most Frequent Venues

In the last step, a label/description is added to each cluster based on the most common top 5 venues. In order to complete this step the first most common venue to the fifth most common venue columns are summed over the neighborhoods for each cluster and, then, the mode is employed in order to retrieve the most frequent venue. In case of different venues with the same frequency, both are included in the new column separated by comma<sup>4</sup>.

### 3.4. Ranking

The ranking is based on three metrics (new variables):

- Number of Venues for Tourism: it is related to the main categories that may have more touristic relevance. This metric was obtained by the summation of the number of venues in the following categories: arts & entertainment, event, food, nightlife and parks & outdoors.
- Number of Venues for Tourism/Per Hostel + Newcomer: this variable penalizes competition. It gives the average number of touristic venues per hostel, considering the installation of a newcomer. This adjustment of the number of hostels + 1 newcomer is important to avoid division by zero.
- Counts of Scenic Lookouts and Beaches: this variable contains the venues that make Rio internationally known.

In order to make the variables comparable the **min-man normalization** was applied and, thus, the range of each new variable transformed to fit between  $[0, 1]$ . Distinct weights were used: 60% was assigned to the Number of Venues for Tourism, most important variable; 20% to Number of Venues for Tourism/Per Hostel, the justification is that penalizing for new hostels should have a limited impact, as neighborhoods that already have hostels may be one of the most attractive for tourists; 10% to Counts of Scenic Lookouts and Beaches, this low weight was given in order to

<sup>4</sup>For a better understanding proceed to table 4.2 in the Results section.

reduce possible bias as the Foursquare data does not account properly some of Rio's beaches, as Copacabana beach.

**Ranking = 60%N of Venues for Tourism+30%N of Venues for Tourism/Per Hostel & Newcomer+  
10%N of Scenic Lookouts & Beaches**

# Results

The results consists of three parts:

- Descriptive Analysis
- K-means Model
- Ranking the Neighborhoods

## 4.1. A Descriptive Analysis

The mean and median number of venues per neighborhood is of 35 and 27, respectively. The distribution of venues per neighborhood has a positive skewed format, i.e., it has more neighborhoods with 0 to 30 venues than otherwise, as seen on figure 4.1. The most common number of venues per neighborhood (mode) is 18.

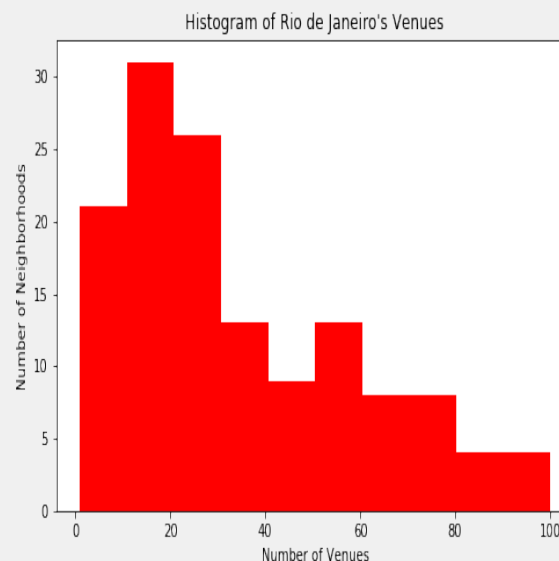


Figure 4.1

The number of distinct specific/secondary venues for Rio totals 296. On the other hand, there are 8 main/primary venue categories found on Foursquare: arts and entertainment, building, education, events, food, nightlife, parks & outdoors and shops.

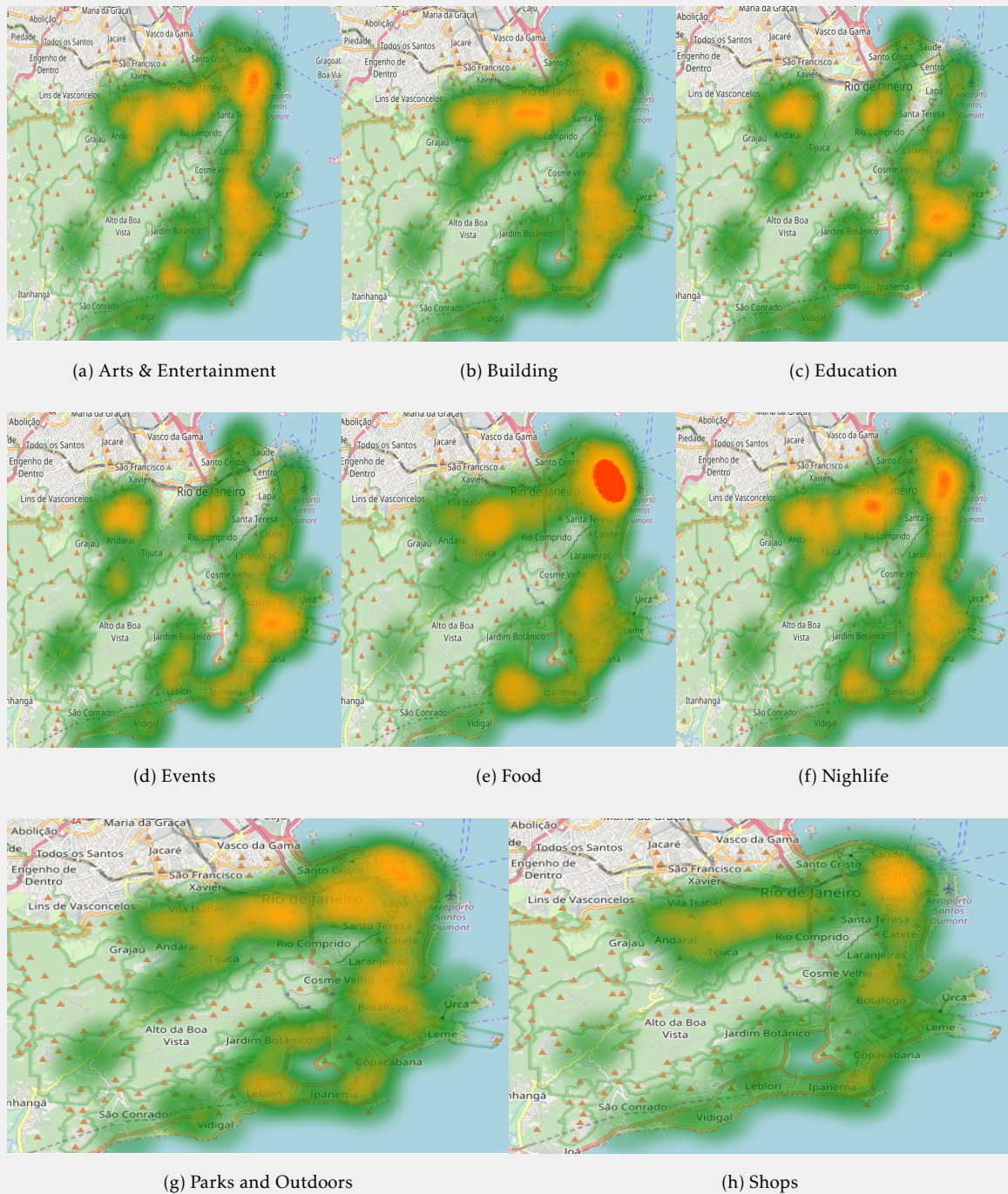
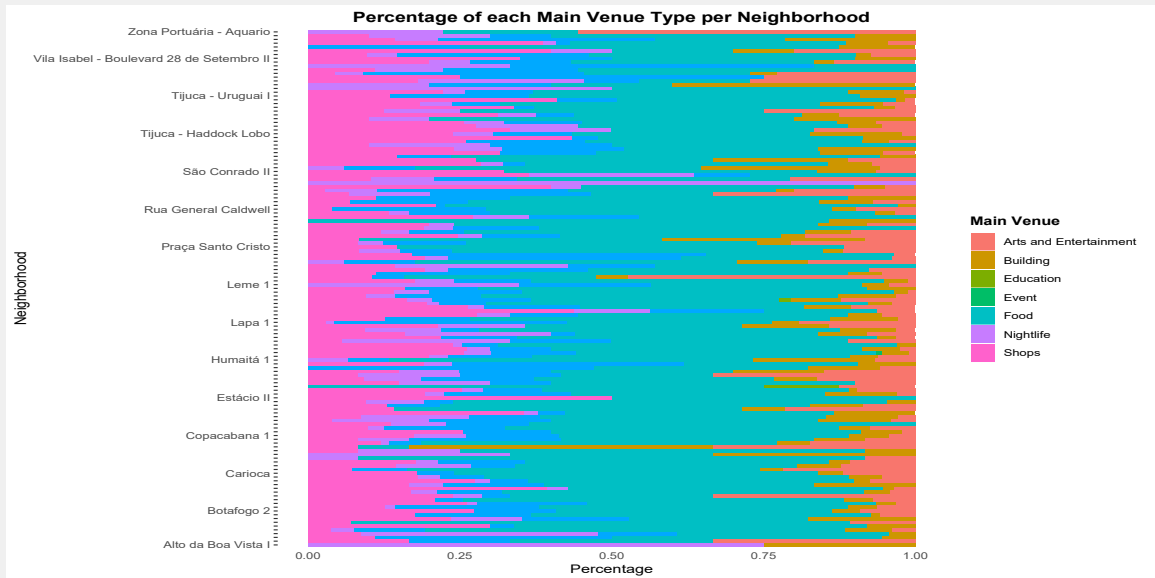


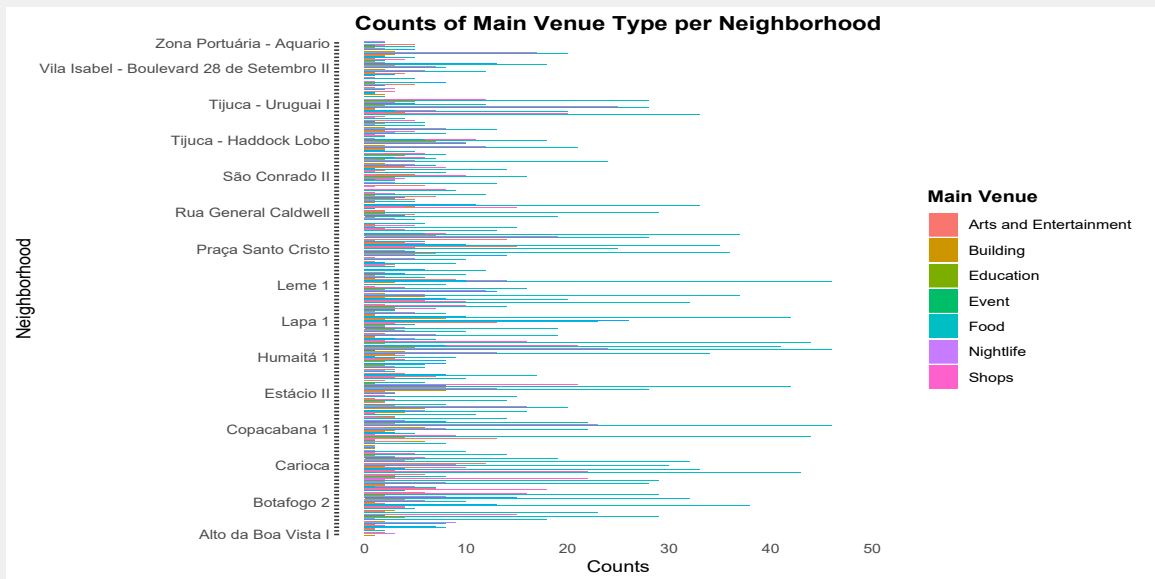
Figure 4.2: Primary Venue Heatmaps

The figure 4.2 presents a heatmap for each type of primary venue category. Downtown concentrates venues on arts & entertainment, nightlife, shops and food; the South Zone events and education; and *Grande Tijuca* ranks well in number of shops. Outside of Downtown there are important concentrations of food, nightlife and arts & entertainment venues in the neighborhoods

of Botafogo, Copacabana and Leblon, in the South Zone, and in Tijuca, Maracanã and Praça da Bandeira, in *Grande Tijuca*. The parks and outdoors venues has a greater density in Downtown regarding numerous plazas and squares, but also in Tijuca and around Laranjeiras, Santa Teresa and Flamengo, in the South Zone.



(a) Percentage Bar Chart



(b) Counts per Category Bar Chart

Figure 4.3: Percentage and Number of Counts Stacked Bar Charts

Both the percentage and count bar charts show that the food venue type has, on average, more counts and a higher percentage in the neighborhoods' composition. Other important main types, but on a lesser degree, are: the shops, building and arts & entertainment.

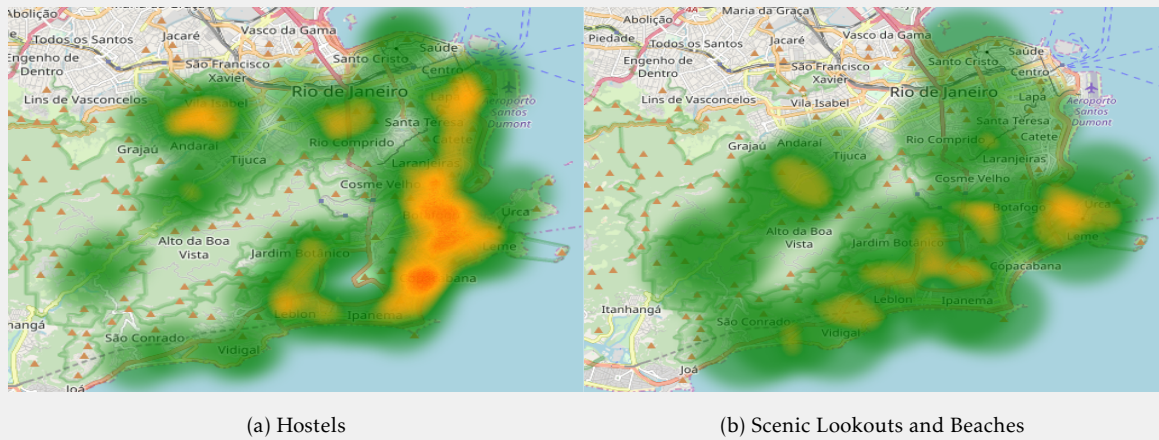


Figure 4.4: Hostels, Scenic Lookouts and Beaches Heatmap

Figure 4.4 represents the heatmap of hostels and scenic lookouts & beaches, which are important for our analysis in terms of competition and touristic value, respectively. There are only 61 hostels in the dataset. It can be seen from the figure the density of hostels in the South Zone and in Lapa, in Downtown. Figure 4.1 shows the top 10 neighborhoods by number of hostels and it can be found that 9 of them are located in the South Zone, with a higher number in Copacabana, Santa Teresa and Botafogo. The exception is Lapa, in Downtown, tied for the second place. The scenic lookouts & beaches venues is mostly found in the South Zone, but the density is lesser than would be expected, as Foursquare didn't have entries for some important venues, as Copacabana beach.

	Neighborhood	Nº Hostels
1	Copacabana 8	7
2	Lapa 1	6
3	Glória - Santa Teresa	6
4	Copacabana - Bairro Peixoto	4
5	Botafogo 3	4
6	Botafogo 4	3
7	Santa Teresa - Largo dos Guimarães	3
8	Leblon 3	3
9	Botafogo 5	3
10	Copacabana 2	2

Table 4.1: Top 10 Number of Hostels per Neighborhood



## 4.2. The K-Means Model

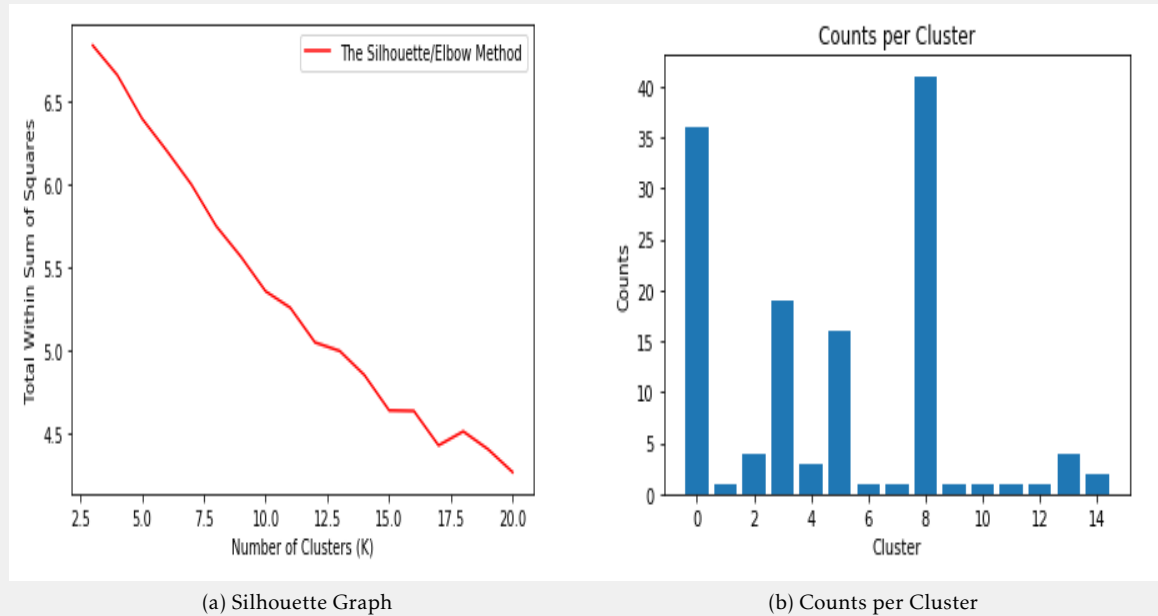


Figure 4.5: Silhouette Graph and Counts per Cluster Label Bar Chart

The elbow method applied to the Total Within Sum of Squares of the Silhouette plot 4.5.(a) indicates that the optimal number of clusters is  $k = 15^1$ . The bar plot 4.5.(b) shows the number of neighborhoods per cluster label. The most numerous are cluster number 8, 0, 3 and 5, respectively, with more than 15 observations each. The remaining clusters have all less than 5 neighborhoods.

Table 4.2 presents the five most common venues per cluster. After an analysis of the most frequent venues from the 1st to the fifth most common venue and between the apparent differences between clusters, a description is assigned to each. As bars, Brazilian restaurants and restaurants are the most common venues, there is some similarities/homogeneity between some of the clusters, as can be noted of clusters 0, 3, 4 and 8. There are also some very distinct clusters, as 1, 10 and 5, being the last two the less attractive for tourism.

In figure 4.6 it can be seen a greater distribution of clusters of label 8 (Bars, Restaurants, Pizza and Coffee), highlighted in light blue, in the South Zone and of clusters of label 0 (Restaurants, Bars and Culture), in red, in Downtown. *Grande Tijuca* on the other hand is more heterogeneous, but with a more frequent type of cluster 3 (Bars, Bakeries and Restaurants), in dark blue.

<sup>1</sup>Note that  $k = 16$  leads to an increase in the Total Within Sum of Squares.

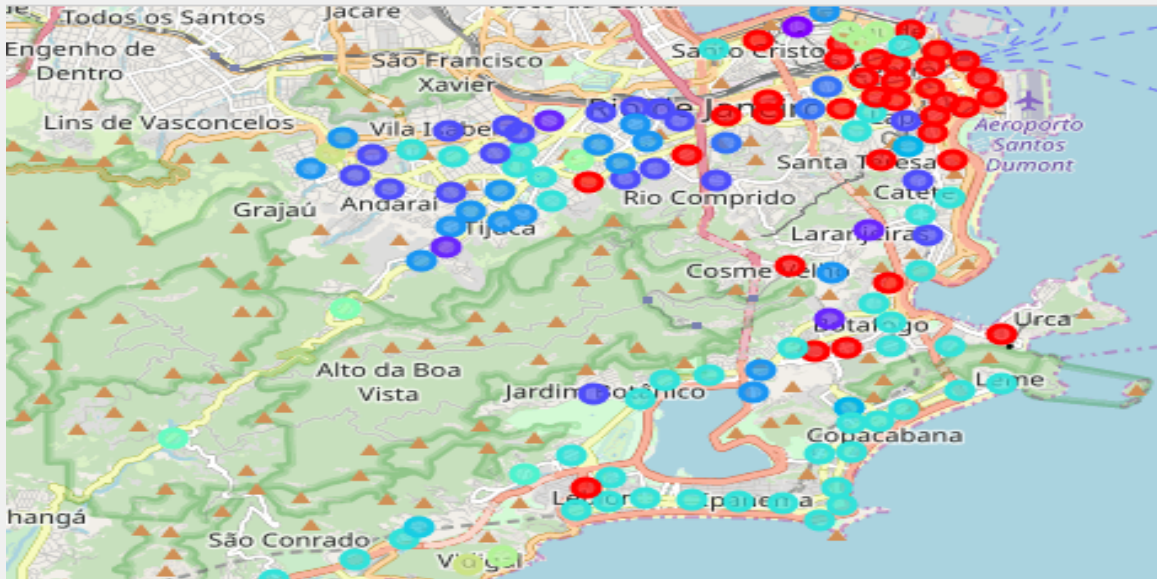


Figure 4.6: Spatial Distribution of Clusters

Cluster	Cluster Description	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	0 Restaurants, Bars and Culture	Brazilian Restaurant	Theater	Bar	Bar	Coffee Shop, Plaza, Restaurant
1	1 Area Under Revitalization	Factory	Performing Arts Venue	Theater	Recreation Center	Samba School
2	10 Soccer Field	Soccer Field	Big Box Store	Bakery	Nightclub	Fast Food Restaurant
3	11 Outdoors & Recreation	Outdoors & Recreation	Bakery	Nightclub	Gym	Social Club
4	12 Café and Tea	Café	Tea Room	Dance Studio	Plaza	Martial Arts Dojo
5	13 Nighclubs, Music and Historic Sites	Nightclub	Music Venue	Bar, Brazilian Restaurant, Empanada Restaurant...	Historic Site	Dive Bar, Historic Site, Hostel, Plaza
6	14 Bars, Hostels and Restaurants	Bar	Hostel, Pizza Place	Gym, Trail	Bakery, Restaurant	Brazilian Restaurant, Scenic Lookout
7	2 Good Variety of Options	Bakery, Convenience Store, Grocery Store, Socc...	Coffee Shop, Food Truck, Snack Place, Stadium	Flea Market, Pizza Place, Plaza, Scenic Lookout	Bistro, Museum, Pet Store, Theme Park	Gym
8	3 Bars, Bakeries and Restaurants	Bar	Bakery, Brazilian Restaurant	Bakery, Brazilian Restaurant, Dive Bar, Restau...	Hotel, Japanese Restaurant	Bakery
9	4 Restaurants and Food Truck	Brazilian Restaurant	Restaurant	Bank, Food Truck, Library	Coffee Shop, Farmers Market, Restaurant	Coffee Shop, Fabric Shop, Public Art
10	5 Fitness and Gyms	Gym	Bakery, Gym	Pizza Place	Gym / Fitness Center	Pizza Place
11	6 Hostels, Bars and Culture	Hostel	Bar	Art Museum, Hotel	Acai House, Bistro	Historic Site, Pie Shop
12	7 Sushi and Burger	Sushi Restaurant	Burger Joint	Pizza Place	Snack Place	Gym / Fitness Center
13	8 Bars, Restaurants, Pizza and Coffe	Brazilian Restaurant	Bar	Pizza Place	Pizza Place	Coffee Shop, Restaurant
14	9 Brazilian Restaurants and Snack Bars	Brazilian Restaurant	Empada House	Amphitheater	Empanada Restaurant	Sushi Restaurant

Table 4.2: Cluster Description



### 4.3. The Ranking

index	Cluster Label	Cluster Description	Neighborhood	Borough	Neighborhood Latitude	Neighborhood Longitude	Ranking
0	1	8 Bars, Restaurants, Pizza and Coffe	Leblon 4	Zona Sul	-22.985782	-43.227286	92.50
1	2	0 Restaurants, Bars and Culture	Praça Passeio Público	Centro	-22.912930	-43.177173	79.48
2	3	0 Restaurants, Bars and Culture	Cinelândia	Centro	-22.910366	-43.175907	78.17
3	4	8 Bars, Restaurants, Pizza and Coffe	Leme 2	Zona Sul	-22.963840	-43.173703	78.17
4	5	8 Bars, Restaurants, Pizza and Coffe	Gávea 1	Zona Sul	-22.975750	-43.228035	76.86
5	6	8 Bars, Restaurants, Pizza and Coffe	Ipanema 1	Zona Sul	-22.984576	-43.198638	70.57
6	7	0 Restaurants, Bars and Culture	Camelôdromo	Centro	-22.903379	-43.182878	70.28
7	8	0 Restaurants, Bars and Culture	Botafogo 1	Zona Sul	-22.956530	-43.193978	68.97
8	9	0 Restaurants, Bars and Culture	Praça Tiradentes	Centro	-22.906478	-43.182956	68.97
9	10	8 Bars, Restaurants, Pizza and Coffe	Lapa 2	Centro	-22.912367	-43.186272	67.65
10	11	0 Restaurants, Bars and Culture	Rua do Ouvidor	Centro	-22.903613	-43.178112	67.65
11	12	0 Restaurants, Bars and Culture	Praça XV	Centro	-22.902859	-43.173220	66.34
12	13	0 Restaurants, Bars and Culture	Praça Mauá	Centro	-22.897001	-43.180842	65.03
13	14	0 Restaurants, Bars and Culture	Castelo	Centro	-22.911183	-43.173301	61.08
14	15	8 Bars, Restaurants, Pizza and Coffe	Humaitá 2	Zona Sul	-22.955930	-43.197163	61.08
15	16	8 Bars, Restaurants, Pizza and Coffe	Botafogo 5	Zona Sul	-22.947694	-43.185878	60.06
16	17	0 Restaurants, Bars and Culture	Rua do Senado	Centro	-22.909510	-43.185593	59.77
17	18	8 Bars, Restaurants, Pizza and Coffe	Copacabana 2	Zona Sul	-22.969604	-43.185210	59.28
18	19	8 Bars, Restaurants, Pizza and Coffe	Ipanema 2	Zona Sul	-22.984335	-43.204910	58.49
19	20	3 Bars, Bakeries and Restaurants	Lapa 1	Centro	-22.913835	-43.181443	56.15

Table 4.3: Ranking - Top 20 Neighborhoods for Hostels

Table 4.3 presents the top 20 highest rated neighborhoods in the ranking. The most common clusters are 0 (Restaurants, Bars and Culture), with 10 observations, and 8 (Bars, Restaurants, Pizza and Coffe), with 9 observations. Cluster 3 (Bars, Bakeries and Restaurants) has only one observation, being placed on the bottom. The distribution of neighborhoods per district is: 11 in Downtown, 9 in the South Zone and 0 in *Grande Tijuca*. The highest rated observation, with a score of 92.50 out of 100 is Leblon 4 and the lesser ranked in the top 20 list, with a score of 56.15, is Lapa 1.

# Discussion

---

The top 20 results are all located in the South Zone and Downtown, the most touristic districts of Rio, what would be expected of where an entrepreneur should consider installing a new hostel. In this section we will focus the analysis on the top 5 neighborhoods of the ranking, with a score above 75. It should be noted that there are some variables that were not included in this study, due to the lack of a readily available database, that should be taken into account.

The first one is crime (pickpockets, robbery, etc), that tends to be more concentrated in Downtown. Places as Cinelândia and Praça Passeio Público are more secure during business days, but generally desert, and thus, less secure, at nights and weekends. Leblon 4, Gávea 1 and Leme 2 are all relatively safe for Rio standards, as they are more heavily policed and although these neighborhoods are considered residential, they also have a good offer of restaurants and bars, which keep them with some nightlife.

The second factor to be considered is rent price. The South Zone is the most expensive district, due to its restricted real state offer, but high demand, as it concentrates the safest neighborhoods, the most famous beaches and good options of services, restaurants and bars.

Below is a description of pros and cons of the top 5 neighborhoods:

- Leblon 4 - Pros: excellent offer of restaurants, bars, leisure options, beach and transportation. Very safe neighborhood. Cons: expensive rent price as well as expensive restaurants and bars.
- Praça Passeio Público - Pros: good location, near Lapa (with a strong nightlife) and Aterro do Flamengo (good leisure/scenic option), very strong offer of cultural and historic venues, has excellent transportation options. Cons: not very safe at night.
- Cinelândia - it is very near Praça Passeio Público, so, the same applies to Cinelândia.
- Leme 2 - Pros: has one of the best beaches in the urban part of the city, good offer of restaurants and bars, near Copacabana beach, usually safe. Cons: not so good transportation offer (despite being near Copacabana) and expensive rent price.

- Gávea 1 - Pros: good options of bars, usually safe and near good scenic lookouts & beaches.  
Cons: expensive rent price / restricted transportation options.

# Conclusion

---

The results of the top 20 best neighborhoods for the installment of a new hostel based on the number of venues, the competition from existing hostels and nearby scenic lookouts and beaches is consistent to what would be expected when considering the characteristics of the neighborhoods and touristic opportunities. There will be a trade-off regarding the neighborhoods in the city Rio de Janeiro between rent price vs safety at night and between beaches vs number of cultural/historic venues, that is reflected in the differences between the South Zone and Downtown district. This creates specific strategies, notably for marketing, that the entrepreneur should take into account, because the neighborhood/district/location will determine not only the price per stay to be charged, but, specially, the hostel's touristic profile.

# Bibliography

---

David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

Foursquare. <https://developer.foursquare.com/>. Accessed on: 2019-07-04.



GPS Coordinates. <https://www.gps-coordinates.net/>. Accessed on: 2019-07-04.

Brett Lantz. *Machine learning with R*. Packt Publishing Ltd, 2015.

Price of Travel. Cities with the Most Hotels (including Hostels). <https://www.priceoftravel.com/1588/cities-with-the-most-hotels-around-the-world/>. Accessed on: 2019-07-04.

# Image's Copyright

---

-  Brazil sculpture of Christ the redeemer free icon from Flaticon.
-  Rio de janeiro Brazilian mountains landscape free icon from Flaticon.