Capstone Project
Applied Data Science

Author: Eduardo Suarez

February 2021

Contents

Viva Las Vegas Strip! Bet on investment 5

1 Introduction 5

2 Business Problem 5

2.1 Stating and refining the Question 6

3 Analytic Approach 6

4 Stages of the project 7

5 Data 7

5.1 Data Requirements 7

5.2 Data Collection 9

5.2.1 Michael Bluejay's easy vegas website 9

5.2.2 Foursquare Platform 10

6 Data Cleansing 10

6.1 Hotel data set 10

6.2 Nearby Business dataset 11

7 Exploratory Data Analysis 11

7.1 Dataset Summary 12

7.2 Hotel dataset 13

7.3 Nearby business dataset 14

8 Data Preparation 17

8.1 Feature Engineering 17

8.1.1 One Hot encoding 17

8.1.2 Binning features 18

8.2 Venue Analysis 18

9 Cluster Hotels (using k-means algorithm, see justification in point 3) 18

9.1 Standardization 18

9.2 Get the best k value. 19

9.3 k-means algorithm (Results) 19

9.4 Information of the groups (Clusters) 20

9.4.1 Graphical information of the clusters 20

9.4.2 Detailed information by cluster 21
10 Results 23
11 Discussion 24
12 Conclusion 24

Viva Las Vegas Strip! Bet on investment

Capstone Project

Author: Eduardo Suarez

February 2021

1 Introduction

The Las Vegas Strip is the world capital of entertainment, gambling, events and conventions. In this line of approximately 6.8 kilometers located in the city of Las Vegas Nevada are the largest hotels and entertainment centers in the city for young people, adults, and children. The Strip is visited annually by millions of tourists, businessmen, and celebrities from all over the world and is an ideal place to invest.

On the other hand, there are also thousands of businesses on the Strip and its surroundings that offer a variety of services and products for the millions of customers annually, so the competition in this area is very high, but with the millions of people that visit, the Strip is an ideal place to invest if data-driven feasibility studies are conducted to find patterns and signals that help an investor make the right decisions on where and in which business to invest.

Despite the limited time to deliver the Capstone Project and limited data sources, this project has a high level of analysis and a holistic approach to the business environment in the Strip, ideal as input for the conceptual phase of an investment project.

2 Business Problem

A group of investors would like to invest in the US in the city of Las Vegas, Nevada, specifically on the Las Vegas Strip. They are looking to do business in this major world entertainment metropolis and would like to know what type of business to invest their money in, and that business has little or no competition to help them recoup their investment in a relatively short time.

2.1 Stating and refining the Question

To successfully develop this case study, it is necessary to delve into the business problem posed, and define some questions that allow obtaining a more precise and exact understanding of what is wanted with respect to the type of business, investment and competition, these questions would be:

- Do you have any type of business in mind that you want to invest in?
- The investment will be towards a small, medium or large enterprise,
- Do you have in mind to invest in an innovative business or will it be a venture that will compete with those established on the Las Vegas Strip?
- What specific location on the Las Vegas Strip do you want to invest in?

These questions will help stakeholders (investors) to specify a more specific question about their business problem that allows the data scientist to make a good analytical approach. After feedback from stakeholders, the business problem was reformulated as follows:

"Invest in businesses that are not hotels, amusement parks or entertainment, that is, small or medium enterprises such as groceries, shops, stores, cafes, retails, among others, that are located near the most important hotels on the Las Vegas Strip. Depending on how it is targeted and localized competition in areas of interest to invest, entrepreneurship could be an innovative business or a business that can compete with existing ones".

Based on this new approach, the question of interest for the case study was established:

What are the types of businesses that are located near the major hotels on the Las Vegas Strip (other than amusement parks, entertainment, other hotels, or large businesses) and how are these businesses distributed or grouped in these areas?

3 Analytic Approach

Once the business problem has been clearly established, now let us define our analytical approach to solve the business problem, according to the question asked was determined that the analysis will be exploratory:

- For the point of identifying the main hotels and nearby businesses on the Las Vegas Strip I used descriptive statistics to describe, characterize and summarize

the data set through tables and graphs that allow us to find patterns or references.

- For the point referred to how these businesses are distributed or grouped in the areas of interest (hotels), I used the unsupervised learning machine learning technique (K-means) that allows us to identify groups or clusters of interest for investors that allows them to identify where direct your investment.

4 Stages of the project

The stages of the project are detailed in the following figure:

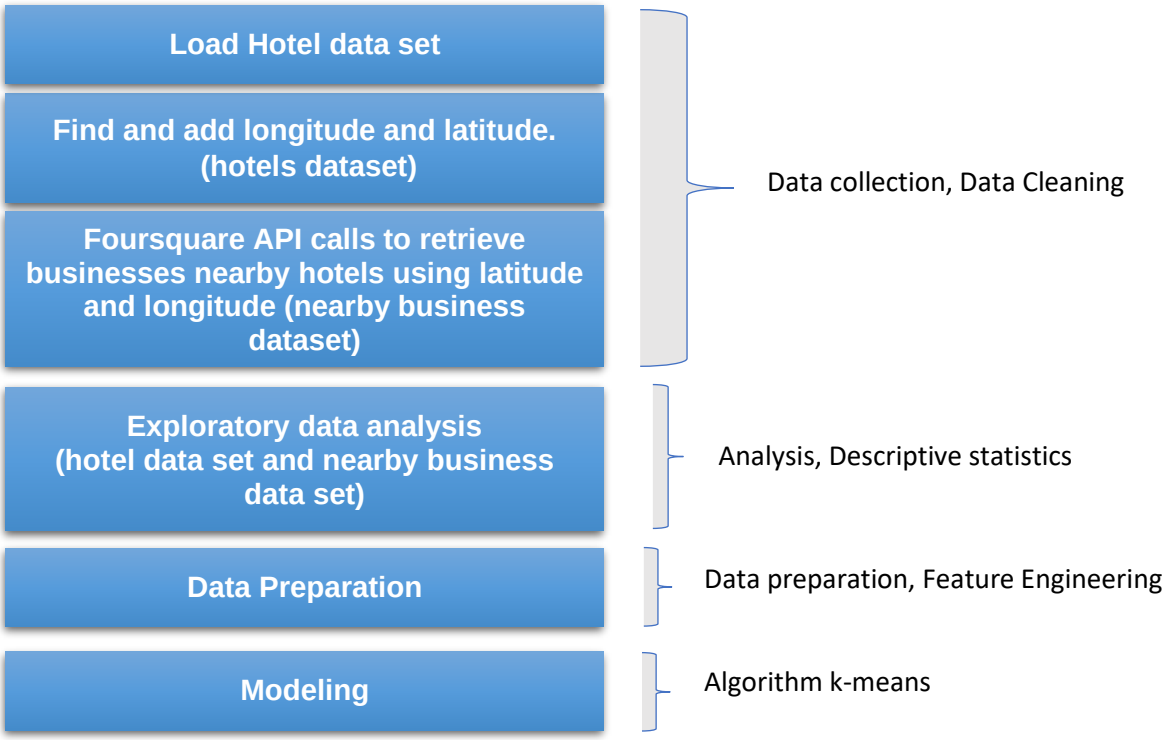


Figure 1. Stages of the project

5 Data

5.1 Data Requirements

Due to the limited time for the study, it was not possible to perform an exhaustive search of several data sources to determine and compare the quantity and quality of the information and data, however, the two data sources that I consulted for this project are an excellent reference for its quality and veracity.

Based on the analytical approach established in the previous point, the following data requirements (numerical and categorical) and data sources were identified:

- Records of the main hotels located on the Las Vegas Strip, can be a web page, database, XLS and CSV files or another file that contains at least the following information: Hotel name, address, location or zip code.

  After an internet search, the website https://easy.vegas/casinos/list-interactive (Figure 2) was identified as a data source that contains information about main hotels y casinos on the Strip.

  This website allowed us to select the minimum required data and other data of interest through the following options (below an example of the data):

  Show: All hotels were selected only in the Strip area
  Columns: Hotel price level, address and zip code were selected
  Sort: It was ordered by Area and Hotel Name

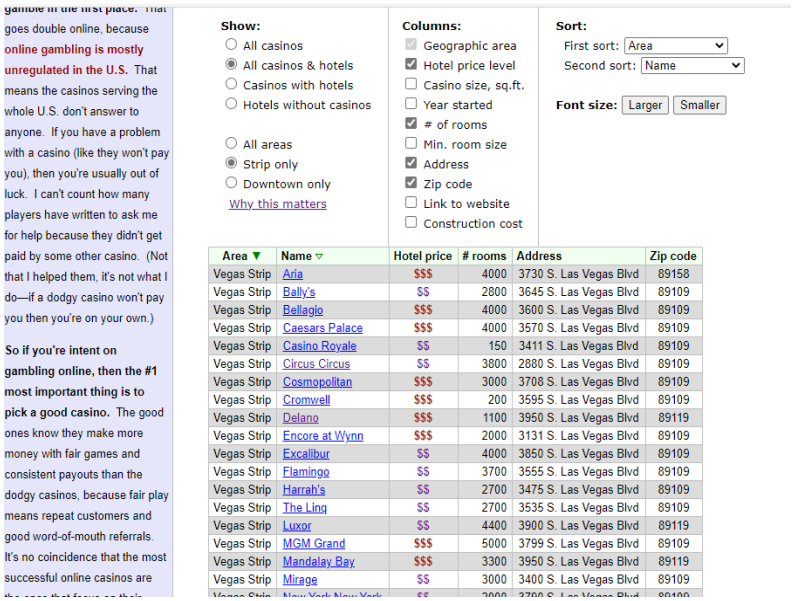


Figure 2. Michael Bluejay's easy vegas website.
Source: https://easy.vegas/casinos/list-interactive

- Records of businesses located on the Las Vegas Strip, can be a web page, database, XLS and CSV files or other file that contains at least the following information: Business name, address, location or zip code, type, or business class.

  To obtain the data of the businesses near the main hotels on the Strip I used the location platform based on social networks Foursquare (Figure 3) through its API to obtain a JSON file (below an example of the data). It contains important characteristics for the study (venues, types, categories among other).

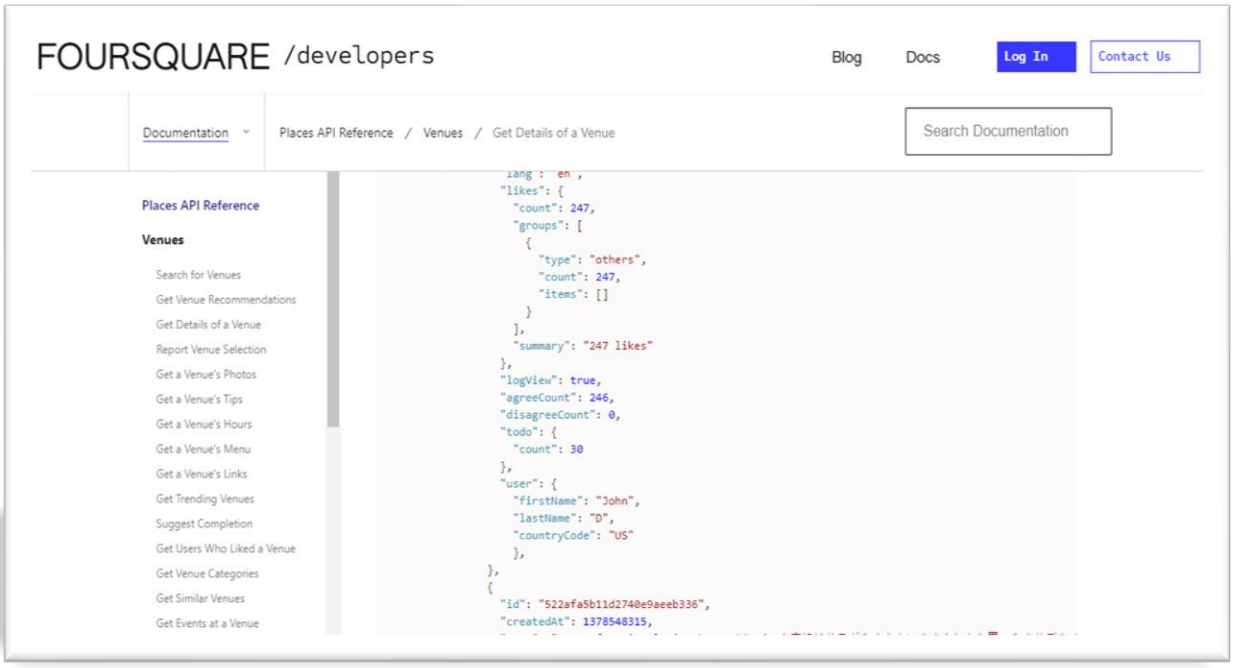


Figure 3. Foursquare platform. Example Venues Detail. API Reference

5.2 Data Collection

5.2.1 Michael Bluejay's easy vegas website

Since the data obtained through the website https://easy.vegas/casinos/list-interactive to identify the main hotels in the Las Vegas Strip is very small (32 records), it will be converted to an Excel table (dataset). A row was also detected that does not correspond to a hotel but to a casino, this row was eliminated.

This data was uploaded to the Project Notebook in IBM Cloud Pak for Data, Watson Service. In total, 32 records were retrieved with the following information from the hotels:

Area: Vegas Strip (object)

Name: Hotel name (object)

Price: Accommodation appreciation price (object)

rooms: Number of rooms in the hotel (integer)

Address: Physical address of the hotel (object)

Zip code: area code (integer)

5.2.2 Foursquare Platform

For the collection of businesses near the hotels on the Las Vegas Strip, the data was obtained from the Foursquare platform through API service using Python getting details of the businesses using hotel coordinates and locating businesses within a 500-meter radius of each hotel. This data was uploaded to the Project Notebook in IBM Cloud Pak for Data, Watson Service. In total, 2603 records were retrieved with the following information:

Venue ID: Unique string identifier venue (object)

Venue: Name of business or venue (object)

Distance: Distance in meters of each business with respect to each hotel in a radius of 0-500 meters

Latitude: Geographic coordinate (float)

Longitude: Geographic coordinate (float)

Venue Category: Classification of business according to its economic activity, such as hotel, bar, restaurant, etc. (object)

6 Data Cleansing

6.1 Hotel data set

The focus of the investment will be on business in hotel areas with large accommodation capacity, so small hotels on the Strip were not taken in this project, this data was eliminated. For this reason, our hotel data set decreased to 29 records.

Latitude and Longitude fields of floating type were created in the dataset.

On the other hand, the geopy library and the Photon geocoder were used to locate the coordinates of the hotels, but the latitude and longitude were wrong due to the structure of the address field. The structure of the address field was modified by removing the word "Blvd", after that, the data was imported successfully.

Later, the Folium library was used to visualize the geospatial location of the hotels according to the coordinates obtained (Figure 4), however, it was possible to visualize 4 hotels with wrong coordinates on the map.

These wrong coordinates were corrected and updated (Figure 5) each one using google map and updated in the data set.

All hotels are now on the Strip which is an area or a strip that goes from north to south in Las Vegas.

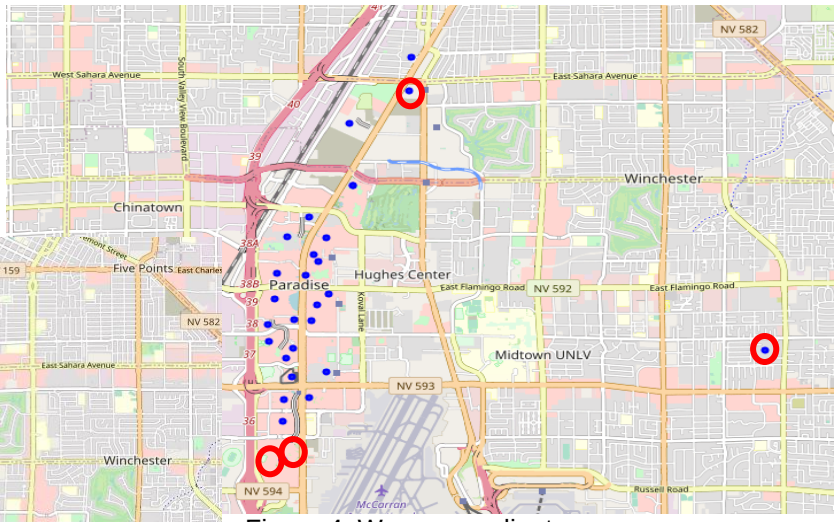


Figure 4. Wrong coordinates

Subsequently, the Name and Hotel Price fields were renamed to “Hotel” and “Price” in the data set.

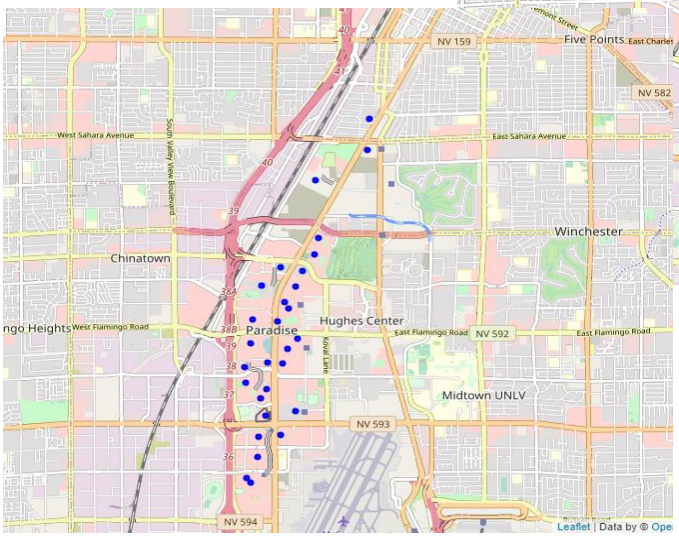


Figure 5. Corrected coordinates

6.2 Nearby Business dataset

The “Name” and “Venue Category” fields in the dataset were renamed to “Hotel” and “Category” respectively.

Businesses or places that do not add value to our case study were filtered (no deleted), for example, hotels, zoos, theme parks, museums, among others.

7 Exploratory Data Analysis

After cleaning the data, an exploratory analysis of the two data sets was carried out. Each variable of interest was analyzed to discover signals and better understand the data.

7.1 Dataset Summary

A quick descriptive view of the two data sets was made to validate type of data, columns and total of records.

```
In [82]: coord_hotels.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 29 entries, 0 to 28
Data columns (total 7 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Hotel      29 non-null     object
 1   Price      29 non-null     object
 2   rooms      29 non-null     int64
 3   Address    29 non-null     object
 4   Zip code   29 non-null     int64
 5   Latitude   29 non-null     float64
 6   Longitude  29 non-null     float64
dtypes: float64(2), int64(2), object(3)
memory usage: 1.7+ KB
```

Figure 6. Summary Hotel dataset

```
In [86]: ID_hotbuss.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2603 entries, 0 to 2602
Data columns (total 7 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Name            2603 non-null   object
 1   Venue ID        2603 non-null   object
 2   Venue           2603 non-null   object
 3   Distance        2603 non-null   int64
 4   Latitude        2603 non-null   float64
 5   Longitude       2603 non-null   float64
 6   Venue Category  2603 non-null   object
dtypes: float64(2), int64(1), object(4)
memory usage: 142.5+ KB
```

Figure 7. Summary nearby business dataset

The data types of each feature were validated ok, there were no null data and the number of records per data set was verified.

On the other hand, a statistical summary of the data set "Nearby business dataset" (Table 1) was also made, where I could see some interesting data:

```
n [87]: ID_hotbuss.describe(include='all')
ut[87]:
```

	Name	Venue ID	Venue	Distance	Latitude	Longitude	Venue Category
count	2603	2603	2603	2603.000000	2603.000000	2603.000000	2603
unique	29	925	864	NaN	NaN	NaN	166
top	Flamingo	53054fd5498e2d9d5a17a5b9	Starbucks	NaN	NaN	NaN	Hotel
freq	100	7	19	NaN	NaN	NaN	186
mean	NaN	NaN	NaN	291.230119	36.113537	-115.172092	NaN
std	NaN	NaN	NaN	128.162502	0.012260	0.004783	NaN
min	NaN	NaN	NaN	4.000000	36.087974	-115.181944	NaN
25%	NaN	NaN	NaN	192.000000	36.105534	-115.175530	NaN
50%	NaN	NaN	NaN	309.000000	36.112758	-115.172997	NaN
75%	NaN	NaN	NaN	396.000000	36.121181	-115.169979	NaN
max	NaN	NaN	NaN	500.000000	36.149996	-115.151481	NaN

Table 1. Statistical summary Nearby business dataset

a. One of the businesses with the highest number of occurrences was Starbucks (19)

b. The number of categories or business classification was 166.

c. The Venue ID variable which is a unique identifier per business has 925 unique values out of a total of 2,603 records (analysis that will be developed later)

d. The shortest distance of a business with respect to a hotel is 4 meters and the greatest distance is 500 meters and a median of 309 meters.

7.2 Hotel dataset

First, taking the variables “Hotel”, “rooms”, and “Price” I grouped the hotels by number of rooms showing appreciation prices of the rooms by levels low-medium-high (Figure 8).

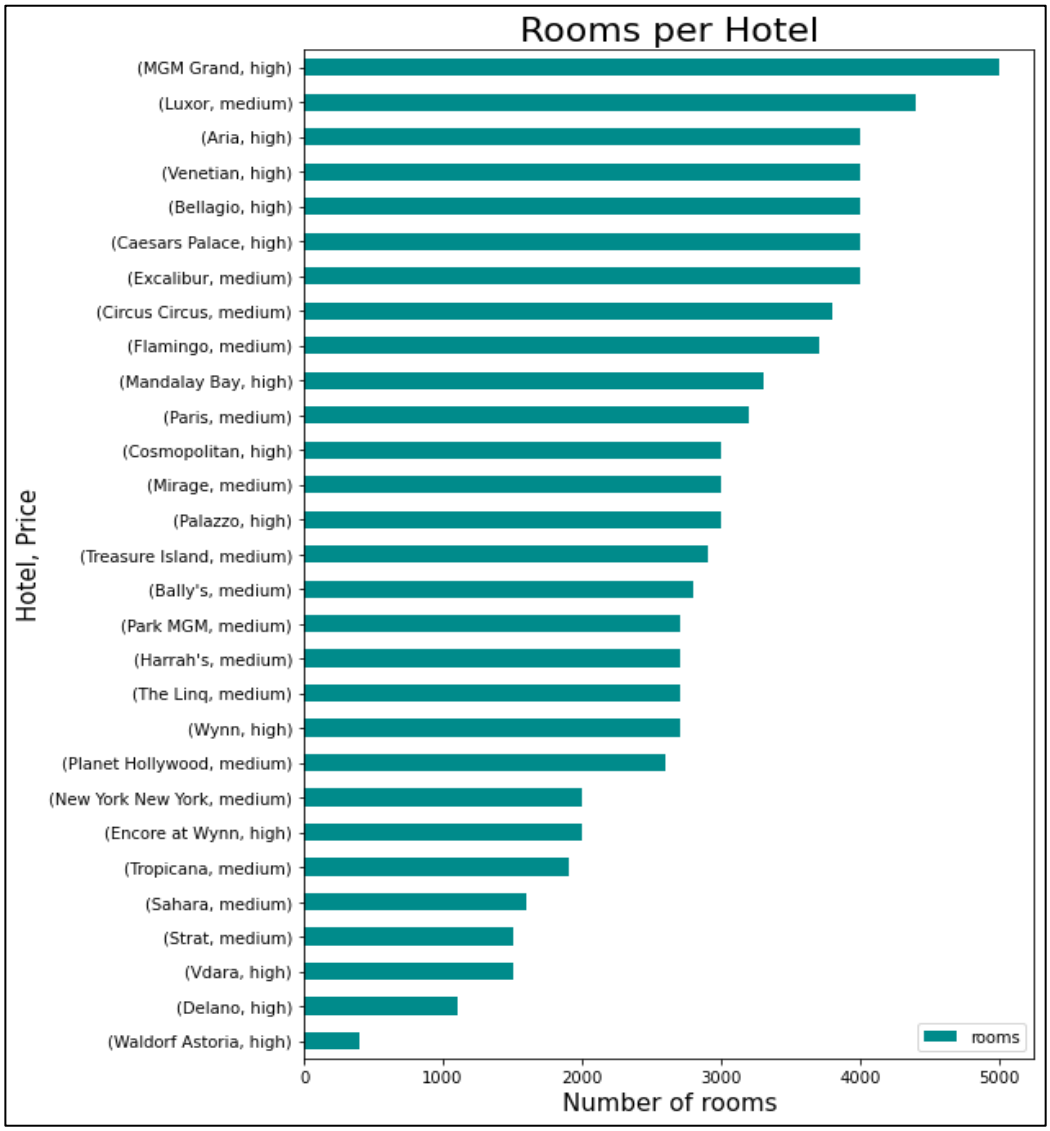


Figure 8. Rooms and appreciation prices per Hotel.

Then, I visualized the hotels according to their accommodation capacity using Folium (Figure 9).

The size of the circles indicates the capacity of each hotel.

The graphs in Figures 8 and 9 at first glance provide us with first-hand information about the distribution of hotels in terms of accommodation, appreciation price, and geographical location.

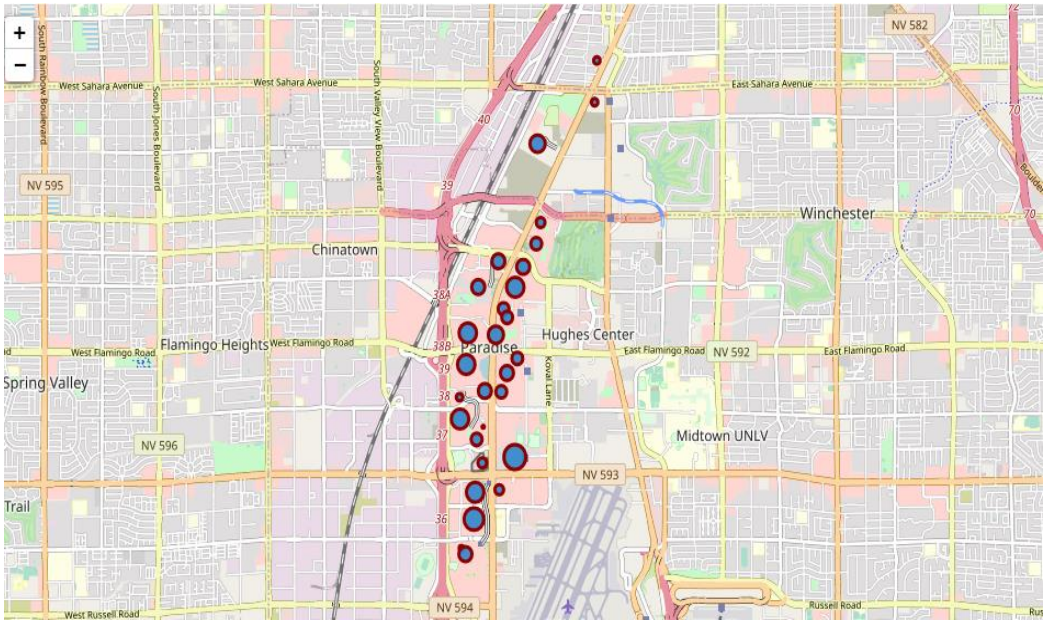


Figure 9. Accommodation capacity by hotels, geospatial view.

7.3 Nearby business dataset

Taking the variables "Hotel" and "Venue" (businesses) I grouped the number of businesses close to each hotel (Figure 10)

According to the data previously retrieved through Foursquare API, most hotels had more than 100 venues, however, Foursquare limits a maximum of 100 venues per coordinate given. For the purposes of this project, I only used the maximum limit allowed by coordinate.

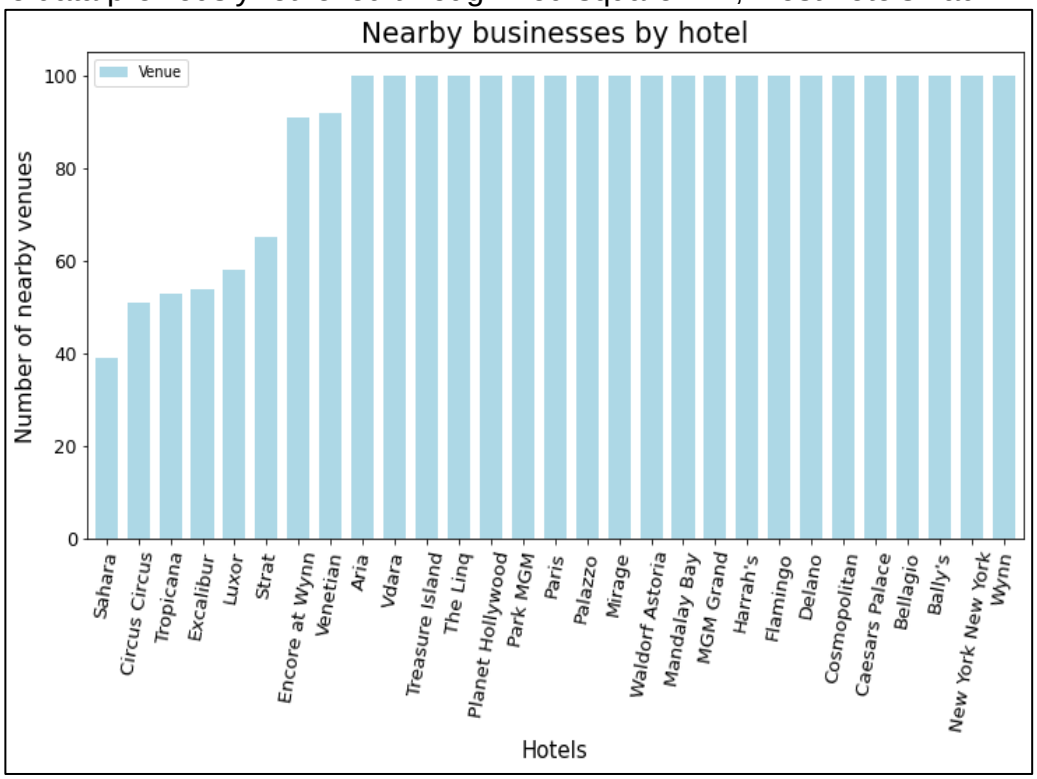


Figure 10. Nearby business by hotel

The Venue ID variable displayed 925 unique records out of 2603 records. I breakdown the data until I obtained a single record per row of the variable "Venue ID", an example of the analysis is shown in the Table 2:

As we can see in table 2, some Venue IDs are repeated for more than one hotel, this because the same business has different distances for each hotel due to the proximity of the hotels to each other. These records will not be eliminated from the data set since the distance feature is important for the cluster model.

Out[90]:

Venue ID	Venue	Category	Hotel	Distance	Latitude	Longitude
41326e00f964a52006141fe3	Coyote Ugly Saloon - Las Vegas	Bar	New York New York	92	1	1
41326e00f964a52038151fe3	Venetian Resort & Casino	Casino	Harrah's	386	1	1
			Mirage	455	1	1
			Palazzo	238	1	1
			The Linq	483	1	1
			Treasure Island	267	1	1
			Venetian	132	1	1
41326e00f964a52057141fe3	House of Blues	Music Venue	Delano	191	1	1
			Luxor	317	1	1
			Mandalay Bay	174	1	1

Table 2. Unique records for Venue ID feature (Breakdown)

Records that do not apply to our case study (parks, museums, zoos, among others) were excluded. In total, I identified 660 business near hotels.

To analyze the variable "Category" I used the Word Cloud technique to visualize the most frequent business or venues by category.

Figure 11. Most frequent businesses or venues by category

We can see above that most businesses are focused on the food and beverage business.

The "Distance" feature represents the distance in meters of each business with respect to each hotel in a radius of 0 to 500 meters, I used a histogram and a boxplot graph to analyze this variable.

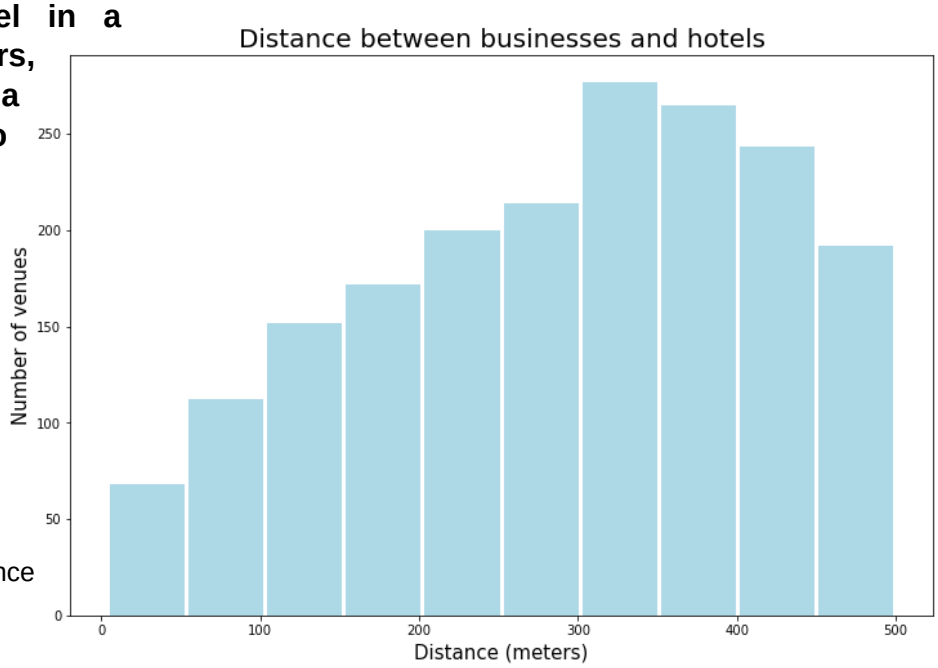
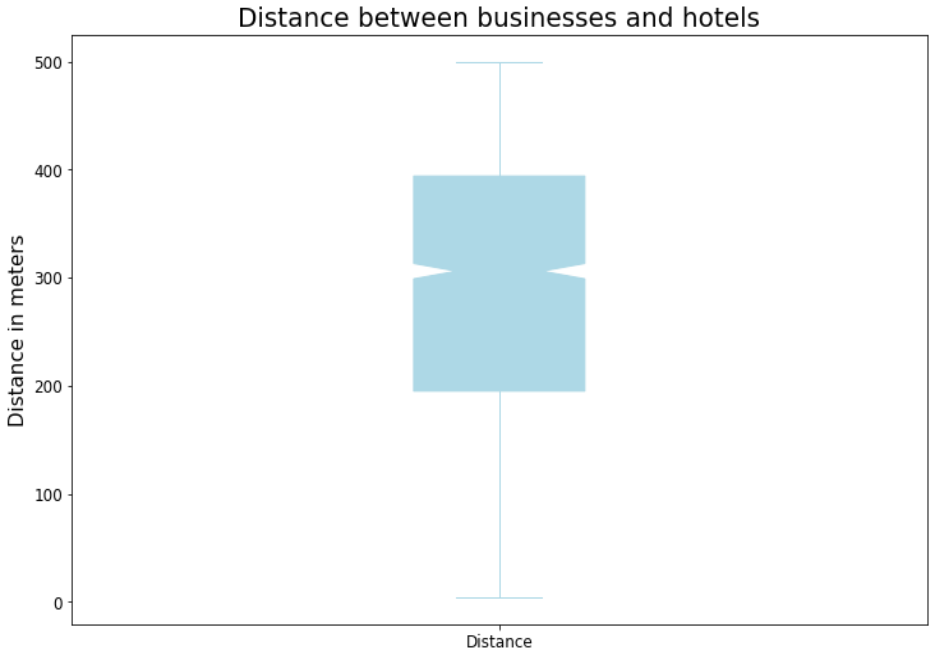


Figure 12. Histogram. Distance between business and hotels

In the graphs we can see that:

a. 25% of the businesses (~470 businesses) are at a radial distance between 0 and 200 meters from the coordinates of the hotels.

b. 25% of the businesses (~470 businesses) are at a radial distance between 400 and 500 meters from the coordinates of the hotels.



c. 50% of the businesses (~940 businesses) are at a radial distance >200 and <400 meters from the coordinates of the hotels.

Figure 13. Boxplot. Distance between business and hotels

8 Data Preparation

Preparing the data for modeling involves selecting the correct variables in the proper format for our k-means machine learning algorithm to work correctly. I used the features, Category", "Distance", and “Price” for our hotel clustering model. Lastly, I united the three data sets (Category, Price and Distance) into one, the algorithm was applied to this last data set.

8.1 Feature Engineering

8.1.1 One Hot encoding

Category Feature:

Our qualitative "Category" variable was converted to a numeric variable using dummy encoding. Subsequently, I grouped the data by hotel, averaging the frequency of occurrence of each category.

Out[97]:

	Hotel	Accessories Store	Advertising Agency	American Restaurant	Arcade	Arts & Crafts Store	Asian Restaurant	BBQ Joint	Bagel Shop	Bakery	...	Tapas Restaurant	Tattoo Parlor	Tea Room	Tennis Court	Thai Restaurant	T Resta
0	Aria	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
1	Aria	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
2	Aria	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
														0	0	0	
														0	0	0	

Out[98]:

	Hotel	Accessories Store	Advertising Agency	American Restaurant	Arcade	Arts & Crafts Store	Asian Restaurant	BBQ Joint	Bagel Shop	Bakery	...	Tapas Restaurant	Tattoo Parlor	Tea Room	Tennis Court	R
0	Aria	0.0	0.0	0.014286	0.000000	0.000000	0.014286	0.000000	0.000000	0.0	...	0.0	0.0	0.014286	0.0	
1	Bally's	0.0	0.0	0.044776	0.000000	0.014925	0.000000	0.014925	0.000000	0.0	...	0.0	0.0	0.000000	0.0	
2	Bellagio	0.0	0.0	0.014286	0.000000	0.000000	0.014286	0.000000	0.000000	0.0	...	0.0	0.0	0.000000	0.0	
3	Caesars Palace	0.0	0.0	0.041096	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	...	0.0	0.0	0.000000	0.0	
4	Circus Circus	0.0	0.0	0.071429	0.071429	0.000000	0.000000	0.000000	0.035714	0.0	...	0.0	0.0	0.000000	0.0	

Table 3. One hot encoding “Category” feature

Price Feature:

In the same way, the qualitative variable "Price" was converted to a numerical variable (dummy encoding) to be adequately interpreted by our k-means model.

Out[99]:

	Hotel	high_price	medium_price
0	Aria	1	0
1	Bally's	0	1
2	Bellagio	1	0
3	Caesars Palace	1	0
4	Circus Circus	0	1

Table 4. One hot encoding “Price” feature

8.1.2 Binning features

Distance Feature:

I divided the "Distance" feature into three new features (using the minimum, average and maximum value of the distances) that I call minimum, medium(average) and maximum. This is because each Hotel has more than one distance (meters) per category associated with all businesses.

Out[103]:

	Hotel	minimum_distance	medium_distance	maximum_distance
0	Aria	43	312.514286	493
1	Bally's	26	305.119403	500
2	Bellagio	19	314.657143	497
3	Caesars Palace	65	304.205479	481
4	Circus Circus	33	203.321429	499

Table 5. Binning “Distance” feature

8.2 Venue Analysis

To simplify the analysis of the large number of business categories associated with each hotel, I determined the 10 most common venues (sorted by their relative frequency) per hotel to later append to our cluster data set.

Out[107]:

	Hotel	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	Aria	Lounge	Spanish Restaurant	Italian Restaurant	Cocktail Bar	Steakhouse	Sushi Restaurant	Buffet	Seafood Restaurant	Spa	Bar
1	Bally's	French Restaurant	Steakhouse	Cocktail Bar	American Restaurant	Gift Shop	Hotel Bar	Italian Restaurant	Burger Joint	Bar	Lounge
2	Bellagio	Lounge	Italian Restaurant	French Restaurant	Spa	Japanese Restaurant	Cocktail Bar	Steakhouse	Buffet	Burger Joint	Beer Garden
3	Caesars Palace	Italian Restaurant	Clothing Store	Lounge	Bar	Women's Store	American Restaurant	French Restaurant	Cosmetics Shop	Boutique	Lingerie Store
4	Circus Circus	Bar	Fast Food Restaurant	American Restaurant	Arcade	Gym	Convenience Store	Coffee Shop	Donut Shop	Sandwich Place	Snack Place

Table 6. Top 10 most common venues or business (see Jupiter notebook for full view)

9 Cluster Hotels (using k-means algorithm, see justification in point 3)

9.1 Standardization

Since minimum_distance, medium_distance, and maximum_distance features have high values compared to the other variables and have a Gaussian behavior, I carried out a standardization of the data to be modeled (errata in the Jupiter notebook where it says Normalization).

9.2 Get the best k value.

To obtain the best k for our model, I performed a routine or loop to calculate k-means (k) and the sum of squared errors (SSE) in a range from 1 to 10, then plot the results (Elbow method).

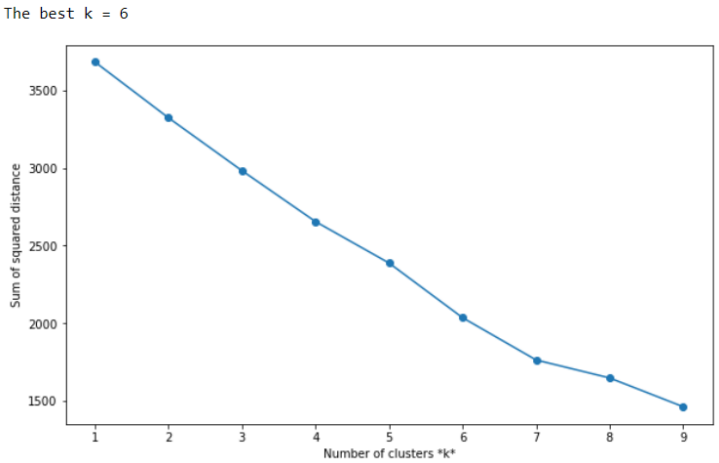


Figure 14. Best k, Elbow method

9.3 k-means algorithm (Results)

Let us see the results of our model graphically using the folium map after running the k-means clustering algorithm.

On the map of the Las Vegas Strip, we can see the 6 groups obtained from the algorithm according to the “Category”, “Distance”, and “Price” features. The largest group of clusters (violet and blue color) are found between Tropicana and Desert Inn avenues, practically the heart of the Strip.

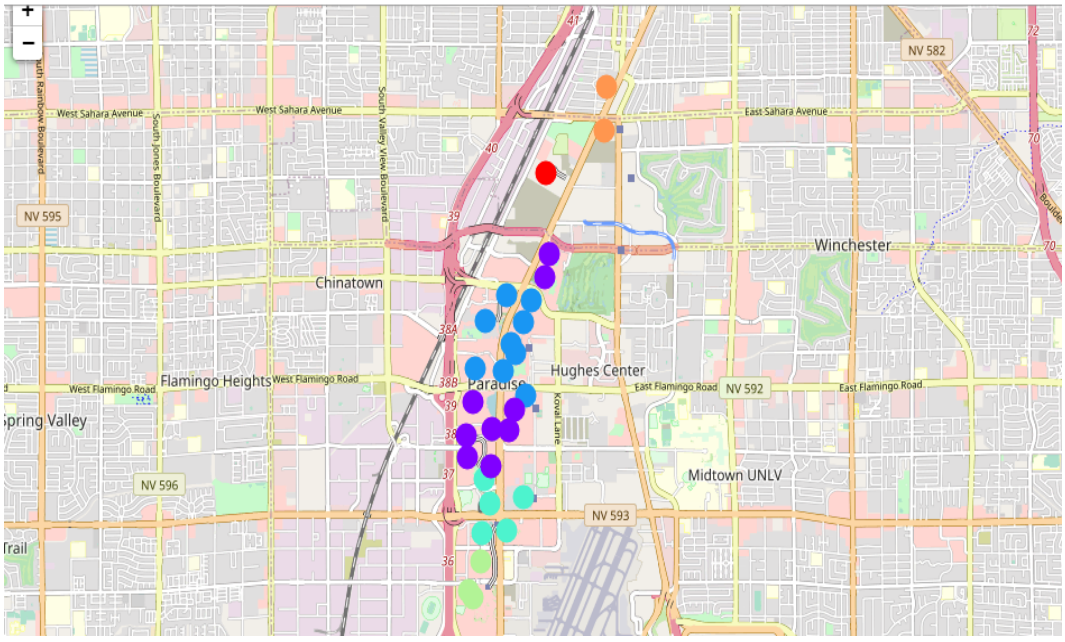


Figure 15. Clustering Hotels

9.4 Information of the groups (Clusters)

9.4.1 Graphical information of the clusters

Figure 16 shows the number of hotels grouped by cluster, in the graph we can see that 18 hotels (62%) are in clusters 1 and 2.

In section 9.4.2 we will see in detail the information of each cluster.

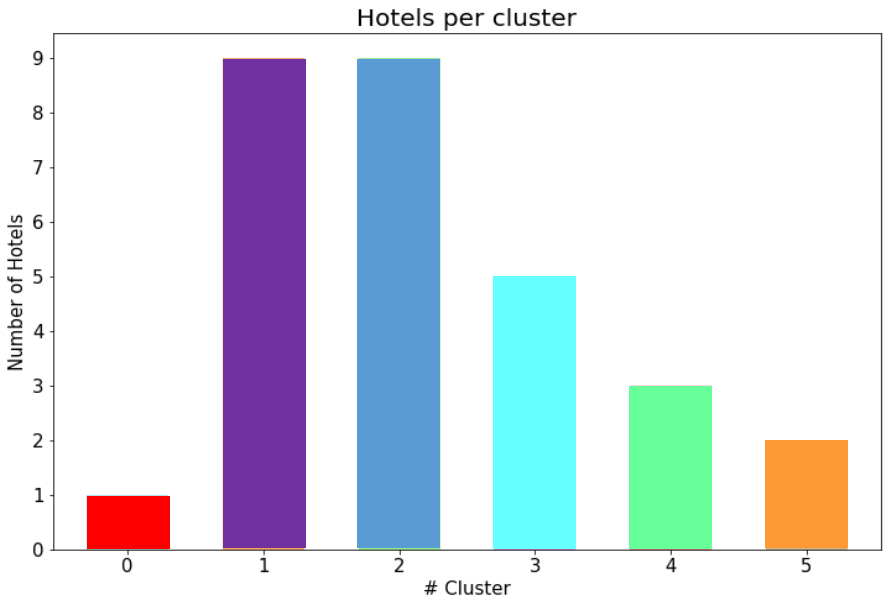


Figure 16. Hotels grouped by cluster

On the other hand, figure 17 shows the first most common businesses by cluster. The description on the "y" axis indicates the cluster and the business category. The food and beverage category are the most common business.

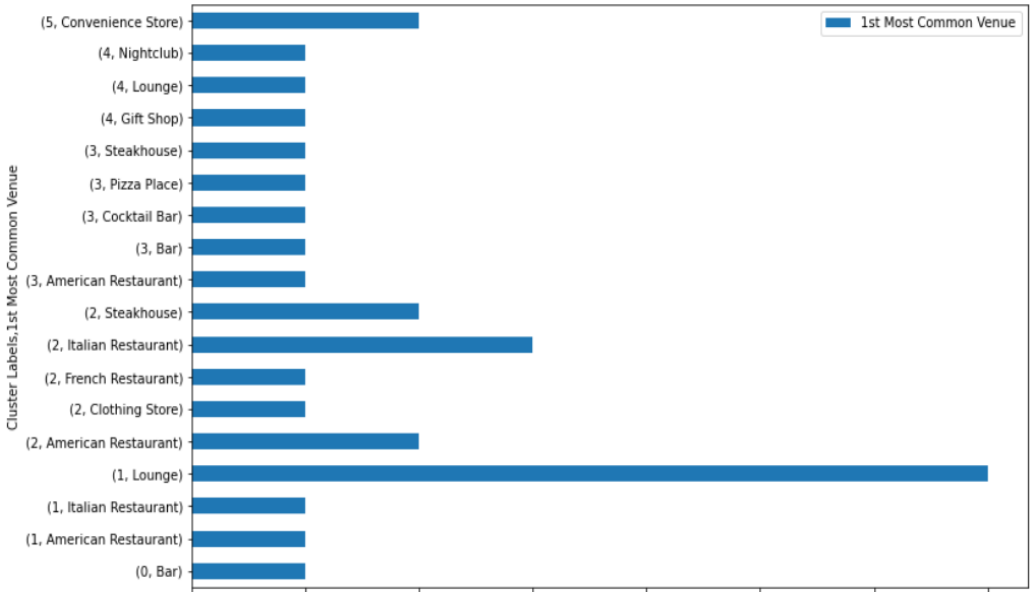


Figure 17. 1st Most common business by cluster

9.4.2 Detailed information by cluster

To finish, the information of the clusters is detailed below (see Jupiter notebook for full view:

a. Cluster 1 groups the hotels with the highest price and the most common business category are "Lounge".

b. Cluster 2 groups the hotels with medium prices and the most common business category are "Restaurant"

c. Cluster 3 groups the hotels with a medium price and the most common business category are "Bar".

d. Clusters 0, 4, 5 group hotels between high and medium prices, the business categories are more diverse, these are located on the north and south ends of the Strip.

Cluster 0

Common: (Restaurants, Foods) Not common: Arcade, Gym, Bar

```
[115]: hotels_merged.loc[hotels_merged['Cluster Labels'] == 0, hotels_merged.columns[[0] + list(range(1, hotels_merged.shape[1]))]]
```

[115]:

	Hotel	Price	rooms	Address	Zip code	Latitude	Longitude	Cluster Labels	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
4	Circus Circus	medium	3800	2880 S. Las Vegas	89109	36.13772	-115.165846	0	Bar	Fast Food Restaurant	American Restaurant	Arcade	Gym	Convenience Store	Coffee Shop

Cluster 1

Common: (Restaurants, Foods, Bar, Lounge, Spa) Not common: Clothing Store, Nightclub, Boutique, Leather goods store, Lingerie Store

```
In [116]: hotels_merged.loc[hotels_merged['Cluster Labels'] == 1, hotels_merged.columns[[0] + list(range(1, hotels_merged.shape[1]))]]
```

Out[116]:

	Hotel	Price	rooms	Address	Zip code	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
0	Aria	high	4000	3730 S. Las Vegas	89158	36.107035	-115.177976	1	Lounge	Spanish Restaurant	Italian Restaurant	Cocktail Bar	Steakhouse	Sushi Restaurant
2	Bellagio	high	4000	3600 S. Las Vegas	89109	36.113115	-115.177006	1	Lounge	Italian Restaurant	French Restaurant	Spa	Japanese Restaurant	Cocktail Bar
5	Cosmopolitan	high	3000	3708 S. Las Vegas	89109	36.110157	-115.174091	1	Lounge	Seafood Restaurant	French Restaurant	Steakhouse	Spa	Spanish Restaurant

Cluster 2

(Restaurants, Foods, Bar, Clothing Store, Nightclub) Not common: Jewelry Store, Cosmetics Shop, Pharmacy, Lingerie Store, Women's Store

In [117]:
```
hotels_merged.loc[hotels_merged['Cluster Labels'] == 2, hotels_merged.columns[[0] + list(range(1, hotels_merged.shape[1]))]]
```

Out[117]:

	Hotel	Price	rooms	Address	Zip code	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
1	Bally's	medium	2800	3645 S. Las Vegas	89109	36.113784	-115.169011	2	French Restaurant	Steakhouse	Cocktail Bar	American Restaurant	Gift Shop	Hotel Bar
3	Caesars Palace	high	4000	3570 S. Las Vegas	89109	36.116628	-115.176757	2	Italian Restaurant	Clothing Store	Lounge	Bar	Women's Store	American Restaurant
9	Flamingo	medium	3700	3555 S. Las Vegas	89109	36.116425	-115.172364	2	Italian Restaurant	American Restaurant	Lounge	Bar	French Restaurant	Steakhouse
10	Harrah's	medium	2700	3475 S. Las Vegas	89109	36.119303	-115.171162	2	American Restaurant	Italian Restaurant	Steakhouse	Nightclub	Bar	Cocktail Bar

Cluster 3 ¶

Common: (Restaurants, Foods, Bar, Gym, Candy Store, Spa) Not Common: Pharmacy, Nightclub, Irish Pub

In [118]:
```
hotels_merged.loc[hotels_merged['Cluster Labels'] == 3, hotels_merged.columns[[0] + list(range(1, hotels_merged.shape[1]))]]
```

Out[118]:

	Hotel	Price	rooms	Address	Zip code	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
8	Excalibur	medium	4000	3850 S. Las Vegas	89109	36.098873	-115.175608	3	Cocktail Bar	Bar	Pizza Place	Steakhouse	Burger Joint	Spa
13	MGM Grand	high	5000	3799 S. Las Vegas	89109	36.102790	-115.169399	3	Pizza Place	American Restaurant	Bar	Cocktail Bar	Steakhouse	Gift Shop
16	New York New York	medium	2000	3790 S. Las Vegas	89109	36.102144	-115.174495	3	Bar	American Restaurant	Pizza Place	Burger Joint	French Restaurant	Clothing Store

Cluster 4

Common: (Restaurants, Foods, Bar,Lounge,Gift Shop) Not common: Gym, Music Venue, Salon/Barbershop, Rental Car

In [120]:
```
hotels_merged.loc[hotels_merged['Cluster Labels'] == 4, hotels_merged.columns[[0] + list(range(1, hotels_merged.shape[1]))]]
```

Out[120]:

	Hotel	Price	rooms	Address	Zip code	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
6	Delano	high	1100	3950 S. Las Vegas	89119	36.092639	-115.177773	4	Lounge	Cocktail Bar	Gift Shop	Mexican Restaurant	Coffee Shop	Italian Restaurant
12	Luxor	medium	4400	3900 S. Las Vegas	89119	36.095872	-115.175807	4	Gift Shop	Cocktail Bar	Lounge	Gym	Pizza Place	Nightclub
14	Mandalay Bay	high	3300	3950 S. Las Vegas	89119	36.091963	-115.177131	4	Nightclub	Cocktail Bar	American Restaurant	Lounge	Italian Restaurant	Pizza Place

Cluster 5

Common: (Restaurants, Foods, Bar, Convenience Store) Not common: Gift Shop, Gym

```
In [121]: hotels_merged.loc[hotels_merged['Cluster Labels'] == 5, hotels_merged.columns[[0] + list(range(1, hotels_merged.shape[1]))]]
```

Out[121]:

	Hotel	Price	rooms	Address	Zip code	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7 C
21	Sahara	medium	1600	2535 S. Las Vegas	89109	36.142342	-115.156912	5	Convenience Store	Restaurant	Steakhouse	Sandwich Place	Gym	French Restaurant	Medit Re
22	Strat	medium	1500	2000 S. Las Vegas	89103	36.146995	-115.156564	5	Convenience Store	American Restaurant	Gift Shop	Cocktail Bar	Mexican Restaurant	Fast Food Restaurant	Re

10 Results

At this point I will present a summary of the results obtained in project which will allow us to have an overview of the findings:

1. 125 business categories (types of business) were detected that covered areas such as restaurants, bars, gyms, clothing stores, pharmacy, Candy Store, among others.

2. 660 businesses or venues near the hotels were identified.

3. The shortest distance between a business and a hotel location was 4 meters and the longest 500 meters. Also 50% of the businesses are in a radius between 4 and 309 meters near the hotels.

4. Some of the most frequent categories or classes of business identified were American Restaurant, Italian Restaurant, Bars, Nightclubs, Steakhouse, Lounge, Pizza Place, Clothing Store, Gift Shop, the food and beverage business are the most frequent.

5. 25% of the businesses are at a radial distance between 0 and 200 meters, another 25% are at a distance between 400 and 500 meters and 50% of the businesses have a radial distance> 200 and <400 meters from the proximity of the hotels.

6. The largest clusters of hotels (according to the analysis of the features "Category", "Distance", and "Price") are physically located between Tropicana and Desert Inn avenues, these are clusters 1 and 2.

7. 62% of the hotels (18) are in clusters 1 and 2. Cluster 1 groups the hotels with high prices and cluster 2 the hotels with medium prices.

8. Clusters 0, 4, and 5 group hotels with mixed prices (high and medium) and are located on the north and south limits of the Las Vegas Strip.

11 Discussion

During the development of the project, some observations (lessons learned) were detected that would be important to mention for future projects like this one or if this project is to be replicated.

Business problem: For me it is the most important stage in any machine learning project, a misinterpretation or lack of clarity in the business problem or not correctly establishing the question of interest to solve the business case, would result in failure of the project and therefore in the loss of time, money, and resources.

Data collection: It is necessary to have more time to locate a greater number of data sources that allow to compare the quantity and quality of the data necessary to strengthen the analysis and the development of the model.

Foursquare is an excellent source of data to consult business or venues, this platform focuses on the interaction between people and the businesses they visit (social media), however, it does not represent a complete database of all existing businesses on a site or location. For this reason, it is necessary to have other additional data sources so that the results are more precise and accuracy in the project and to strengthen the machine learning model.

Data cleaning and preparation: This phase occupies a large amount of time in the project; therefore, time and resources (human and technological) must be available to shorten the time in this stage.

12 Conclusion

The project from its conception focused on identifying investment opportunities in the vicinity of the main hotels located on the Strip of the City of Las Vegas, identifying businesses near these hotels and discovering patterns or interest groups for investors interested in doing business on the Strip.

In summary, during the development of the project, the following results were obtained: number of businesses located in the vicinity of the main hotels, categories or classification of those businesses, the distance or proximity that these businesses have around the hotels, the most frequent businesses and the clusters of hotels (using the k-means machine learning algorithm) related to the variables Category, Distance and Appreciation Price.

However, due to the short time to deliver the capstone project, only two data sources could be consulted, the lack of data from many businesses on the Strip was not possible to obtain, which limited our scope for the development of the project.

Despite the limitation, the results and findings discovered in the project are ideal for investors who focus on the food and beverage business since the most complete data collected focuses on these areas.