
# MS5103 - Business Analytics Project

## YELP DATASET ANALYSIS

**Project Supervisor:**

**Dr. Michael Lang**

**Team Members:**



| Team Members | University ID |
|---|---|
| **Apoorva Phadnis** | **19230364** |
| **Rajan Nayak** | **19230344** |
| **Nikhil Yenni Anjaneya** | **19234025** |

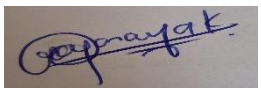
# Declaration of Originality

Module Code: **MS5103** Assignment Title: **M.Sc. Business Analytics - Major Project**

**Group Members: (please use BLOCK CAPITALS)**

| Student ID | Student Name | Contact Details (Email, Telephone) |
|---|---|---|
| 19230364 | Apoorva Phadnis | a.phadnis1@nuigalway.ie |
| 19230344 | Rajan Nayak | r.nayak3@nuigalway.ie |
| 19234025 | Nikhil Yenni Anjaneya | n.yennianjaneya1@nuigalway.ie |

I/We hereby declare that this project is my/our own original work. I/We have read the University *Code of Practice for Dealing with Plagiarism*[*] and am/are aware that the possible penalties for plagiarism include expulsion from the University.

| Signature | Date |
|---|---|
|  | July 2020 |
|  | July 2020 |
|  | July 2020 |

# EXECUTIVE SUMMARY

This report is based on the analysis of YELP dataset for restaurants operating in Las Vegas in Nevada, United States of America.

The report is oriented on several heading and sub-headings. It starts with an introduction which defines our goals and objectives of interest used to perform analyses. The introduction section gives the reader of this document a head start to the Yelp application and its uses across the world. The focus here is completely on the restaurants in Las Vegas. Introduction is followed by a brief background speaking about the history of Yelp and how it has influenced multiple ventures across the world over the years.

The next part of our report focuses on the overview of data providing a description of the fields present in the datasets that are used as a part of the analysis. The type of data, fields and the description will allow the reader of this report to get an essence of the data used.

Following the overview of data are business questions that will be answered through this report in a simple and easy to understand manner for the reader. The first 4 questions are generic in nature and their analysis can be used by restaurant owners for the betterment of the establishment. Next in line, we answer two questions centric to Grand Lux Café on how to improve the business over time. Certain tools that helped us perform the analysis have been jotted down under the tools used section also indicating why these tools used were a better choice over the rest. Initial data cleansing was performed by converting the JSON files into CSV and cleaning the CSV files using python.

The analysis kicks off by performing descriptive statistics to determine the operational status of restaurants and rating analysis based on the count of stars. Geographical analysis is performed to determine the location of the restaurants based on operational status. Following this is the analysis of top influencers who influence the users of yelp by reviewing different restaurants. Factors that affect the closure, word cloud of the restaurants and day-wise visits analysis allows the owners of restaurants make an informed decision with respect to their businesses. Furthermore, restaurant centric analysis is performed on Grand Lux Café that will allow the establishment to focus and improve on their pain points.

The report continues to the limitations section, which gives a brief idea of the shortcomings we as a team faced during the project and is followed by the recommendations section wherein, we offer certain recommendations to external stakeholders who can make use of our research to seek improvements in their business.

The document finishes off with a conclusion which wraps-up our findings and analysis made throughout the document. The final section of the report encompasses the appendix and the bibliography section. The appendix section includes the snippets of coding techniques used while the latter includes the references used throughout the project.

# TABLE OF CONTENTS

# TABLE OF FIGURES

# 1. INTRODUCTION

Recent advancements in technology have led to enhancements in all fields of development. According to the google search, there are 4.54 billion people who have access to the internet which encompasses 59% of the global population. This has led to a wide variety of options available for people to analyse and check the feasibility of their demand. One such platform is Yelp which provides the list of different businesses which allow the users to compare and find the best deals.

In today's fast paced world, people have a bare minimum time for cooking and are always in search of a healthy alternative. Also, people love exploring different varieties and cuisines of food as they call themselves foodies. This has opened multiple opportunities for different ventures.

Yelp is a business directory service and crowd-sourced review forum for different types of ventures. It encompasses range of services such as pet shops, hospitals, salons, spas, restaurants, etc. Las Vegas in Nevada, United States is called the party city and one among the top visited cities in the world. As a result of which, this city comprises of a huge number of businesses that generate high revenue. In our project we focus on providing a detailed analysis of the food ventures listed on Yelp in Las Vegas. Furthermore, restaurant centric analysis is performed on Grand Lux Café that will allow them to focus and improve on their pain points. Also, this analysis opens up opportunities for our primary stakeholders like restaurants and secondary stakeholders such as customers and food critics.

Restaurants currently face wide scale competition all around the world. Many factors come into picture when a restaurant is in business on how to improve the growth and profits of the venture. Categories such as location, opening and closing hours, ratings, user reviews and competition in the neighbourhood play an important role in deciding the future of the business. Restaurants fail to consider these sensitive factors while running the venture which might lead to a downfall in the near future.

On completion of this project, we aim to provide an in-depth analysis to the restaurants in Las Vegas, Nevada based on data gathered from the dataset provided by Yelp. This analysis would in turn help the restaurants to work on their pain points and grow their venture eventually.

## 2. BACKGROUND

Advancements in technology over the past decade have been immense and have bought huge amounts of change in the everyday lives of people. Diligent and genuine efforts are being made constantly trying to improve the existing applications and the way we communicate with the world. [1]

Yelp being one of the most influential websites for consumers has received approximately 178 million unique visitors every month. Founded in 2004, Yelp allows customers to read and write reviews about any type of service and has an approximate of 171 million reviews of almost every type of local business as of Q3 2018. Visitors of Yelp try to find out what is best for the service they are looking for. They search for a business by specific key words or by a specific type they are looking for and can read the reviews posted by others. Writers of reviews on Yelp discuss on what they like about a place, how was the service of the venture and many other factors. People visiting this website and searching for a particular keyword can then read all these reviews that are posted online.

Yelp allows the users to express their thoughts openly and seek information from a variety of sources. This has made the lives of users much easier and access to information available at fingertips. The strength of reviews is such that it can make a business profitable or go into an easy loss. Businesses can take an advantage of such a platform to analyse the user reviews, determine where they went wrong and this opens a window for improvement. Reviews are very influential in deciding the future of a business venture. In the next section we will go through the data in detail.

## 3. OVERVIEW OF DATA

The dataset used is from the Yelp website uploaded on Yelp dataset challenge. The primary analysis on this dataset will be performed keeping in mind the scale of the dataset. The size of the dataset was reduced to allow smooth running and analysis of the current data.

### Vegas Business

| DATATYPE | FIELD | DESCRIPTION |
|---|---|---|
| String | business_id | Business unique id |
| String | name | Name of the restaurant |
| String | address | Address of the restaurant |
| String | city | City where the restaurant is located |
| String | state | State where the restaurant is located |
| Integer | Postal_code | Postal code of restaurant |
| Float | latitude | Geo coordinate - latitude |
| Float | longitude | Geo coordinate - longitude |
| Float | business_stars | Star rating of the business |
| Integer | review_count | Count of the reviews for a particular restaurant |
| Boolean | is_open | Determines if the restaurant is in service or not |
| String | categories | Categories specific to restaurant, e.g. |
| Boolean | restaurants_reservations | Specifies if the restaurant allows reservation or not |
| String | business_parking | Determines the type of parking provided by the restaurant |
| Boolean | caters | Catering services provided or not |
| String | noise_level | Noise levels in the restaurant |
| Boolean | table_service | Does the restaurant provide table service |
| Boolean | take_out | Does the restaurant provide take out option |
| String | price_range | Price range of the restaurant |
| Boolean | outdoor_seating | Does the restaurant provide outdoor seating |
| Boolean | bike_parking | Does the restaurant provide bike parking |
| String | ambience | Ambience type of the restaurant |

| | | |
|---|---|---|
| Boolean | wifi | Does the restaurant provide wifi |
| String | alcohol | Bar type |
| Boolean | restaurants_delivery | Does the restaurant provide delivery |
| Time | monday_hours | Opening hours on monday |
| Time | tuesday_hours | Opening hours on tuesday |
| Time | wednesday_hours | Opening hours on wednesday |
| Time | thursday_hours | Opening hours on thursday |
| Time | friday_hours | Opening hours on friday |
| Time | saturday_hours | Opening hours on saturday |
| Time | sunday_hours | Opening hours on sunday |
| Boolean | accepts_creditcards | Does the restaurant accept credit cards |
| String | music | Music type played at the restaurant |
| Boolean | happy_hour | Does the restaurant provide happy hours |
| Boolean | wheelchair_accessible | Is the restaurant wheelchair accessible |
| Boolean | drive_thru | Does the restaurant have drive-thru service |
| Boolean | smoking | Does the restaurant allow smoking |

## Vegas Review

| DATATYPE | FIELD | DESCRIPTION |
|---|---|---|
| String | review_id | Unique identifier for each review |
| String | user_id | Unique identifier for each user |
| Float | review_stars | Star rating received by each review |
| Integer | useful | Specifies whether the review is useful |
| Integer | funny | Specifies whether the review is funny |
| Integer | cool | Specifies whether the review is cool |
| String | review | Text of the review |
| Date | date | Date of review |

**Vegas Check-ins (Day-Wise Visits)**

| DATATYPE | FIELD | DESCRIPTION |
|---|---|---|
| String | business_id | Unique identifier for each business |
| String | weekday | Specifies the day of the week |
| Time | hour | Indicates the time of visit by the customer |
| Integer | visits | Specifies the number of visits for a restaurant at a particular hour of the day |

## 4. BUSINESS QUESTIONS

From the perspective of a business owner,

1) What factors affected the most and led to the closure of restaurants permanently?
2) What location would pose as an ideal location to start a new venture or a franchise?
3) How would the reviews and ratings provided by Top Influencers prove to be beneficial for an establishment?
4) Which day of the week attracts the most traffic across Las Vegas?

Additionally, we figured out Grand Lux Café to have highest review count with an average rating of star 4. It preceded few restaurants with less review count therefore, we choose GLC to answer few business problems the venture could take steps to resolve using our analysis.

5) What period during the year Grand Lux Café requires better resource management?
6) What are the possible factors that caused the downfall for Grand Lux Café between 2012-13?

## 5. TOOLS USED

In this project we as a team have tried to learn and implement few tools. This will be crucial in order to justify the outcomes of the project and to improve our skills.

**Python:** When we received the data, it was in JSON format. It was essential for us to convert this data into CSV for easy access and field (column) selection. Python was the primary tool

used for text processing, cleaning, making a corpus, running through models, to perform sentimental analysis and further to develop a word cloud. Python supports many machine learning libraries like pandas, NLTK, genism that would prove instrumental in our analysis to solve the business problems.

**TABLEAU:** As a team while using python there were few visualizations that we found difficult to plot. Tableau being an exceptional tool for visualization, provides different templates for better understanding of the data. We specifically, used tableau to confirm if attributes such as Delivery, Drive Thru, etc had impact on restaurants ratings. Its interface also provided us to make optimal use of the latitude and longitude of restaurants to identify if they are closed and open.

**Microsoft Excel:** Microsoft Excel is one of the well-known tools for cleaning, slicing, dicing, pivots, tables and charts in the field of data and analysis. The tool helped us extensively in cleaning and filtering the data specific to Las Vegas.

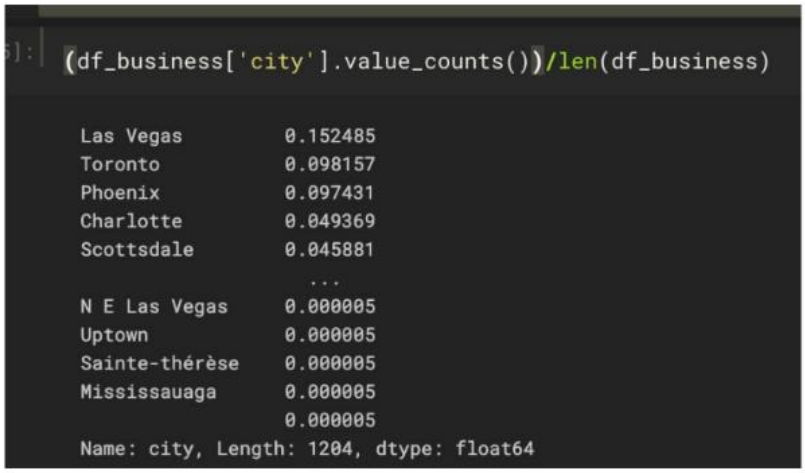
**Microsoft Teams:** All of our collaborations, file sharing, communication and brainstorming was accessed via the collaboration software, Teams. We were able to share our studies, analysis files and most importantly video conference calls seamlessly via the application.

Apart from the above four tools, we did invest our time to figure out possible ways to perform Sentimental Analysis using R. In the course of implementation, we encountered various difficulties and also realized python, as a programming language was more comfortable for our analyses than R.

# 6. DATA CLEANING

## 6.1 Conversion of JSON to CSV

The data provided by Yelp was available on the yelp website in the JSON format. Once the data was converted into csv files from JSON, we considered 3 data files based on our objectives namely vegas_business, vegas_reviews and vegas_users since we will be performing top restaurants analysis and review analysis. Day-wise visits file was downloaded from Kaggle and then filtered out. It was observed that 15% of the business (highest of all) was in Las Vegas and hence our analysis is focused towards only Las Vegas.



```
(df_business['city'].value_counts())/len(df_business)

Las Vegas         0.152485
Toronto           0.098157
Phoenix           0.097431
Charlotte         0.049369
Scottsdale        0.045881
                    ...
N E Las Vegas     0.000005
Uptown            0.000005
Sainte-thérèse    0.000005
Mississauaga      0.000005
                  0.000005
Name: city, Length: 1204, dtype: float64
```

Figure 1. Highest percentage of business in Las Vegas

## 6.2 Data Cleaning the CSV files in Python

Data Cleaning the CSV files in Python:

Once the files have been converted to CSV from JSON we perform some of the basic data cleaning to get the file. For the initial cleaning we use the concepts of pandas creating data frames to load the data.

Step1: Import pandas into python

Step 2: Load the file vegas_business into a data frame df_vegas_business

Step 3: Categorise the df_vegas_business data frame to only include restaurants from the categories

Step 4: Select random 100 restaurants from the data frame and store it into the variable hundred_restaurants and view them to verify

Step 5: Load the file vegas_review and vegas_user into different data frames df_vegas_reviews and df_vegas_users respectively

Step 6: Select only the reviews provided for the hundred_restaurants and omit the rest

Step 7: Merge the df_vegas_reviews and df_vegas_users data frames using a left join

Step 8: From the merged data frame, select only the data applicable to the hundred restaurants

Step 9: Finally merge the df_vegas_business and df_review_users data frames to have the reviews and users of only those hundred restaurants

```
import pandas as pd
import numpy as np
from matplotlib import pyplot as plt
import seaborn as sns
import datetime
```

```
df_business=pd.read_csv("C:/Desktop/NUIG assignments sem2/Projects/FINAL PROJECT/vegas_business.csv")
```

```
df_business.city.value_counts()
```

```
Las Vegas    29370
Name: city, dtype: int64
```

```
df_business=df_business[df_business['categories'].apply(lambda x: True if 'Restaurants' in str(x)else False)].reset_index(dr
df_reviews=pd.read_csv("C:/Desktop/NUIG assignments sem2/Projects/FINAL PROJECT/vegas_review.csv")
```

```
C:\Users\apoor\Anaconda3\lib\site-packages\IPython\core\interactiveshell.py:3057: DtypeWarning: Columns (0) have mixed type
s. Specify dtype option on import or set low_memory=False.
  interactivity=interactivity, compiler=compiler, result=result)
```

```
hundred_business=list(set(df_business['business_id']))[0:100]
print(df_business.head())
df_users=pd.read_csv("C:/Desktop/NUIG assignments sem2/Projects/FINAL PROJECT/vegas_user.csv")
```

```
df_reviews = df_reviews[df_reviews['business_id'].isin(hundred_business)].reset_index(drop=True)
df_review_users = pd.merge(df_users, df_reviews, on='user_id', how='left')
df_review_users = df_review_users[df_review_users['business_id'].isin(hundred_business)].reset_index(drop=True)
```

```
df_final = pd.merge(df_business, df_review_users, on='business_id', how='right').reset_index(drop=True)
```

```
df_final.head()
```

Figure 2. Data Cleaning

# 7. FINDINGS & ANALYSIS

## 7.1 Closed and Open Restaurants

As Las Vegas is a business and tourist hub, there is high number of restaurants operating in this city. Analysis of closed and open restaurants is done using python. In this case duplicate business ids have been dropped. The seaborn count plot has been used to plot a bar graph describing the count of permanently closed v/s open restaurants in Las Vegas.

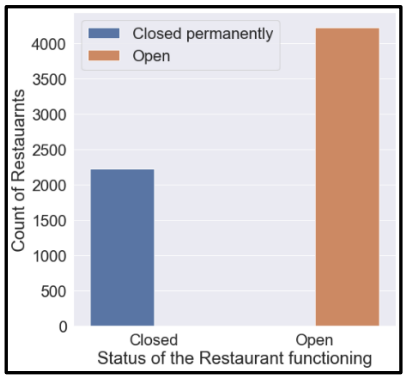



Figure 3. Number of Open and Closed Restaurants of 6450 restaurants

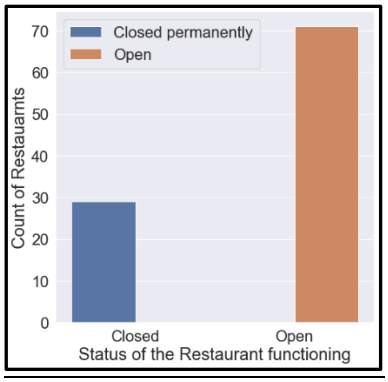



Figure 4. Number of Open and Closed Restaurants of 100 restaurants

**Interpretation and solution**: The 1st bar graph show that out of total 6450 restaurants in Las Vegas, around 2,450 have shut down permanently due to many factors. After analysing the sample data of 100 restaurants, it is evident that out of randomly chosen 100 restaurants that approximately have total of 13548 reviews, 30 restaurants have shut down their operations permanently. From the above bar plots, it is evident that approximately 30% of the businesses have shut down over the years. It is crucial for the consumers to know the restaurants that are open or closed down permanently. This information would help consumers to locate their

desired restaurant without much hassle. The closure of the restaurants may be due to many factors and proper analysis can be performed to find out factors affecting the closure of restaurants permanently using predictive modelling. Due to the limitations of Python, we could run the Logistic Regression algorithm only for 100 restaurants. A memory error is thrown if the model is run for the entire dataset. We will be having a look at that ahead.

### 7.2 Count of Restaurants based on ratings

**Business Problem:** Which is the best location to open up a new restaurant or a franchise of a restaurant running really well?

In today's world, ratings have become an integral part for businesses. The customers rely heavily on business ratings before visiting restaurants. Therefore, it is important for businesses to either maintain their high ratings or improve themselves if they have low ratings. The following bar graph is plotted using the seaborn count plot which counts the number of restaurants having ratings ranging from 1- low to 5 - high. The duplicates have been excluded by dropping duplicate business ids as they appear repeatedly due to various factors. The column stars_x (average rating of the restaurant) has been used to plot the graph.

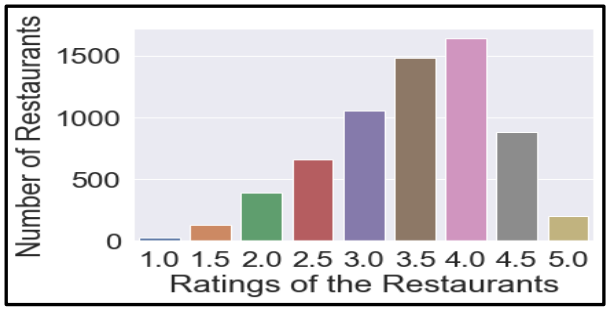



Figure 5. Count of all the Restaurants having ratings ranging from 1- low to 5- high rating


Figure 6. Count of the 100 Restaurants having ratings ranging from 1- low to 5- high rating

**Interpretation and solution**: The bar-graph plotted using python shows that most of the restaurants out of 6450 restaurants in Las Vegas received 4-star ratings, whereas out of the 100 restaurants many have received an average rating of 3.5. The businesses having low ratings can use this statistic for improving their ratings by analysing their shortfalls. On the other hand, businesses with high ratings can either aim for a rating of 5 by improving their services or maintain their current rating. In both cases, only a fraction of restaurants has received an average rating of 5.

### 7.3 Map of 1000 restaurants based on operational status

**Business Problem:** Which is the best location to open up a new restaurant or a franchise of a restaurant running really well?

A map of random 1000 restaurants is plotted using Tableau. The red colour depicts that restaurants are closed permanently, and green colour denotes that the restaurant is open in a particular area.

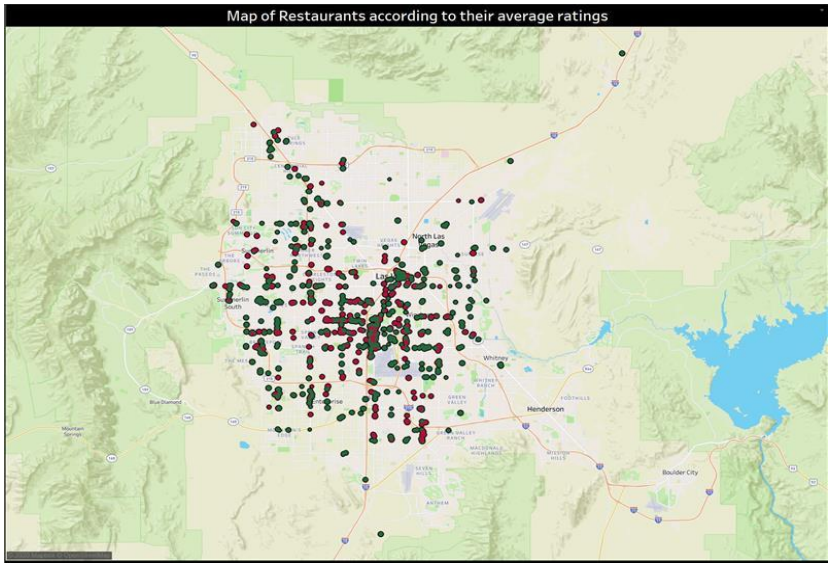



Figure 7. Operational Status of Restaurants (Red = Closed, Green = Open)

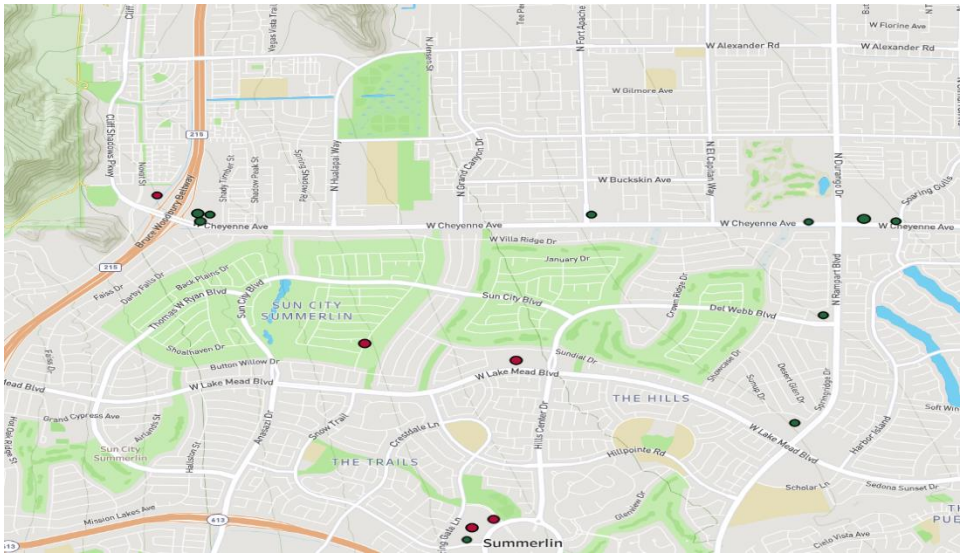



Figure 8. Zoomed in picture of Operational Status of Restaurants

**Interpretation and solution:**

It is clearly evident from the map that W Cheyenne Ave road and the neighbourhood of that area is covered with low rated restaurants and which are still functioning (denoted by green colour). There are only two restaurants which are closed permanently in that area (denoted by red colour) although those restaurants were highly rated. The other restaurants with good ratings (4.5-5 stars) which are doing really good and are thinking of opening up a franchise, they could choose this location for the same. This location would be beneficial to them in terms of profit as there would be less competition because it is surrounded by low rated restaurants.

### 7.4 Top Influencers

**Business Problem:** How would the reviews and ratings provided by Top Influencers prove to be beneficial for an establishment?

The top 3 influencers have been found out based on the highest number of reviews given and how many up votes the review has received. In other words, more the review has number of useful likes, more the review can be trusted. This shows that these users can be considered as influencers since their reviews have received a greater number of up-votes, regardless of the number of reviews they have written. At first, unique users are grouped by the variable “userid” using the code as shown in the snippet below in python. The essence of the review is supported

by the type of reaction the review has received namely useful (how useful the review was to other users), funny, cool and the average rating the review has received. The top influencers along with the other details such as name of the restaurants, average rating given to them and location of the restaurants are displayed using python.

```
user_agg=df4.groupby('user_id').agg({'review_id':['count'],'date':['min','max'],
                                     'useful_y':['sum'],'funny_y':['sum'],'cool_y':['sum'],
                                    'stars_y':['mean']})
```

```
user_agg=user_agg.sort_values([('review_id','count')],ascending=False)
print("          Top 10 Users in Yelp")
user_agg.head(30)
```

| | review_id | date | | useful_y | funny_y | cool_y | stars_y |
|---|---|---|---|---|---|---|---|
| | count | min | max | sum | sum | sum | mean |
| user_id | | | | | | | |
| bLbSNkLggFnqwNNzzq-Ijw | 17 | 2014-02-02 00:44:50 | 2018-10-03 17:07:19 | 238 | 125 | 175 | 3.529412 |
| U4INQZOPSUaj8hMjLlZ3KA | 12 | 2013-10-22 23:01:06 | 2018-10-18 13:20:50 | 72 | 31 | 43 | 3.750000 |
| 3nDUQBjKyVor5wV0reJChg | 12 | 2010-06-12 00:13:27 | 2018-05-21 14:26:29 | 83 | 36 | 74 | 4.416667 |
| L5JFnETi16y2gNiESXBYeA | 11 | 2013-11-20 22:45:43 | 2017-10-29 20:08:10 | 9 | 2 | 3 | 3.090909 |
| uEvusDwoSymbJJ0auR3muQ | 10 | 2012-01-09 18:35:18 | 2013-06-02 00:10:57 | 69 | 33 | 50 | 3.800000 |
| 8DEyKVyplnOcSKx39vatbg | 9 | 2007-11-02 00:46:22 | 2016-07-28 15:04:28 | 172 | 95 | 111 | 3.666667 |
| PKEzKWv_FktMm2mGPjwd0Q | 9 | 2010-04-20 04:27:48 | 2018-09-03 01:52:03 | 65 | 26 | 50 | 3.666667 |
| n86B7lkbU20AkxlFX_5aew | 9 | 2010-02-24 03:20:34 | 2017-08-11 02:12:18 | 33 | 13 | 27 | 3.777778 |
| s2o_JsABvrZVm_T03qrBUw | 8 | 2012-01-10 18:32:03 | 2017-02-02 08:30:38 | 55 | 15 | 39 | 4.375000 |
| HJj82f-csBI7jjgenwqhvw | 8 | 2012-07-10 05:11:53 | 2018-01-21 06:35:57 | 43 | 28 | 31 | 3.750000 |
| Fv0e9RIV9jw5TX3ctA1WbA | 8 | 2012-06-18 03:37:56 | 2017-11-29 06:05:39 | 198 | 116 | 171 | 3.875000 |
| 1O638BDK_fWuxgTVJwff-A | 8 | 2008-05-06 02:34:37 | 2016-05-30 08:37:49 | 16 | 9 | 13 | 3.750000 |
| eZZyuJDoulg4p-GYB3PV_A | 8 | 2011-06-14 10:19:37 | 2018-06-21 03:50:55 | 20 | 5 | 13 | 3.750000 |
| YwaKGmRNnSa3R3N4Hf9jLw | 8 | 2009-04-28 19:25:42 | 2011-04-24 17:36:22 | 13 | 4 | 6 | 3.750000 |
| L15JCA18lb_vMUVJlLxiSw | 7 | 2010-04-16 17:26:18 | 2015-05-15 18:30:13 | 15 | 11 | 6 | 3.428571 |
| L8P5OWO1Jh4B2HLa1Fnbng | 7 | 2009-12-26 23:58:10 | 2015-12-28 10:28:06 | 13 | 4 | 4 | 3.428571 |
| _VMGbmleK71rQGwOBWt_Kg | 7 | 2012-01-05 04:17:09 | 2018-10-01 00:15:24 | 43 | 18 | 33 | 4.571429 |
| I-4KVZ9IqHhk8469X9FvhA | 7 | 2018-04-10 04:10:13 | 2018-09-25 14:59:32 | 125 | 68 | 117 | 4.857143 |
| QJI9OSEn6ujRCtrX06vs1w | 7 | 2010-05-23 01:07:55 | 2013-05-10 12:53:06 | 47 | 21 | 39 | 3.428571 |
| qQecSd0lynfB4g-LPa9JCw | 7 | 2007-09-26 04:24:28 | 2017-12-21 07:00:35 | 13 | 5 | 3 | 3.428571 |

```
[52]: influencers=['bLbSNkLggFnqwNNzzq-Ijw','8DEyKVyplnOcSKx39vatbg','Fv0e9RIV9jw5TX3ctA1WbA']
      top_user_reviews=df_reviews[df_reviews['user_id'].isin(influencers)]
      # Get locations of the places he/she has reviewed
      top_user_locs=pd.merge(top_user_reviews,df_business,on='business_id')
      # UNique list of dates from the user's reviews
      #date_list=list(top_user_locs['date'].unique())
      date_list=list(top_user_locs['date'])
      #rearranging data to suit the format needed for folium
      data=[]
      data1=[]

      df8 = pd.DataFrame()


      for date in date_list:
          subset=top_user_locs[top_user_locs['date']==date]
          df=subset[['latitude','longitude','date','name','stars_x']]
          data12 = pd.DataFrame({"lat":df['latitude'],"lon": df['longitude'],"date": df['date'],"Name": df['name'],"Ratings":df['st
          df8=df8.append(data12)
```

```
df8
```

]:

| | lat | lon | date | Name | Ratings |
|---|---|---|---|---|---|
| 0 | 36.016493 | -115.117069 | 2017-08-16 14:55:24 | Boteco | 3.0 |
| 1 | 36.016493 | -115.117069 | 2017-07-05 08:55:51 | Boteco | 3.0 |
| 2 | 36.171308 | -115.140268 | 2015-02-02 06:36:27 | Gold Spike | 3.0 |
| 3 | 36.169741 | -115.205520 | 2015-01-18 17:54:31 | Taqueria El Buen Pastor | 2.0 |
| 4 | 36.169741 | -115.205520 | 2015-01-09 08:10:18 | Taqueria El Buen Pastor | 2.0 |
| 5 | 36.176102 | -115.260282 | 2014-01-17 03:30:17 | The Bagel Cafe | 3.0 |
| 6 | 36.176102 | -115.260282 | 2014-07-16 16:34:16 | The Bagel Cafe | 1.0 |
| 7 | 36.176102 | -115.260282 | 2012-06-17 22:07:58 | The Bagel Cafe | 4.0 |
| 8 | 36.113693 | -115.307572 | 2014-10-11 22:24:33 | Aranya Thai Bistro | 2.0 |
| 9 | 36.068778 | -115.176840 | 2014-07-28 15:31:37 | Sugar Factory | 3.0 |
| 10 | 36.093063 | -115.176310 | 2016-08-13 07:09:36 | RM Seafood | 4.0 |
| 11 | 36.122328 | -115.170112 | 2015-06-21 09:10:51 | Yardbird Southern Table & Bar | 3.0 |
| 12 | 36.122328 | -115.170112 | 2016-11-03 15:07:46 | Yardbird Southern Table & Bar | 4.0 |
| 13 | 36.148922 | -115.209227 | 2008-05-02 20:55:00 | Amena Bakery & Deli | 4.0 |
| 14 | 36.148922 | -115.209227 | 2014-04-07 18:03:33 | Amena Bakery & Deli | 3.0 |
| 15 | 36.195870 | -115.206444 | 2015-07-28 18:50:05 | Sinaloa Tacos | 3.0 |
| 16 | 36.195870 | -115.206444 | 2013-08-29 04:23:55 | Sinaloa Tacos | 5.0 |
| 17 | 36.268401 | -115.295612 | 2017-09-21 05:15:21 | The Corndog Company LV | 5.0 |
| 18 | 36.268401 | -115.295612 | 2018-01-12 15:18:54 | The Corndog Company LV | 4.0 |
| 19 | 36.145405 | -115.179800 | 2015-07-28 18:50:12 | Goldilocks Salon & Day Spa | 4.0 |
| 20 | 36.167975 | -115.140442 | 2015-08-07 13:27:41 | Bocho Sushi | 3.0 |
| 21 | 36.098347 | -115.299309 | 2017-06-18 07:42:23 | Pop Drinks | 4.0 |
| 22 | 36.193567 | -115.306178 | 2012-11-06 16:10:23 | Bath & Body Works | 4.0 |

Figure 9. Code snippets for Top User Analysis

As shown in the above snippets, the top three influencers have been selected and corresponding latitude and longitudes of the restaurants they have given reviews and ratings to, have been displayed. Along with the names of those restaurants, date and average rating given to the restaurant given by this user are also displayed. Out of the many influencers, only top 3 influencers have been considered to avoid the cluttered look of the map plotted using tableau.

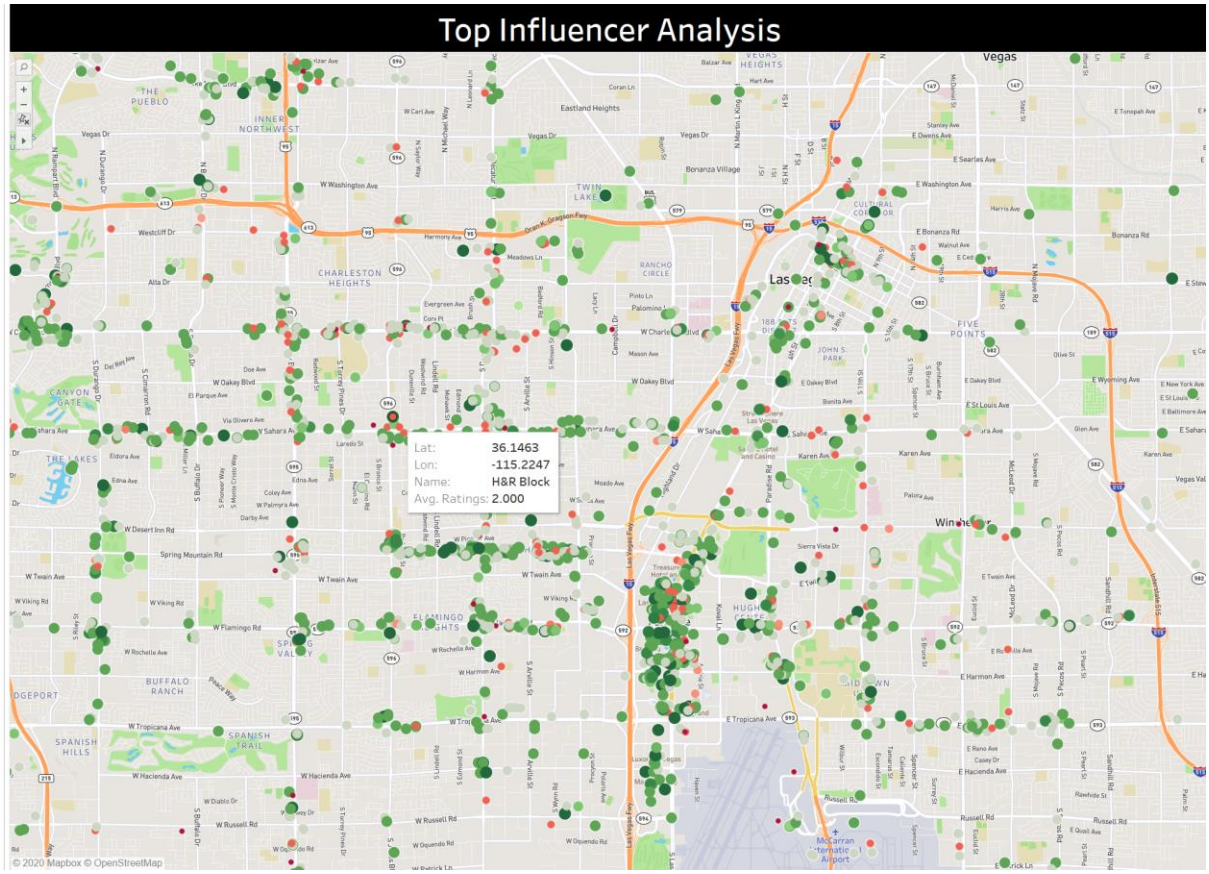



Figure 10. Determining the location and rating of the restaurants based top influencer ratings for Restaurants (Range 1-4: Red = Low, Blue = High)

**Interpretation and conclusion:** The red colour denotes restaurants to which top influencers have given a low rating and blue denotes high rated restaurants by them. For instance, average of the ratings given by the top 3 influencers to a restaurant named H & R Block is only 2. This restaurant could focus on its reviews and improve those certain areas. Also, they could again invite influencers and make their experience better than before. This might provide an opportunity to the restaurants to receive a good review from the influencers, thereby increasing their popularity.

### 7.5 Factors affecting the closure of the restaurants permanently

**Business Problem:** What factors affected the most and led to the closure of restaurants permanently?

We know from section 7.1, that there are about 30 restaurants out of 100 restaurants which are closed permanently, and 70 restaurants are operating successfully. Logistic Regression

algorithm is used to predict the closure of the restaurants, 1 being open and 0 being closed. The data is divided into training set and testing set. Logistic Regression model is fitted to the data and prediction accuracy of 60% is achieved. Only 100 restaurants have been chosen for this analysis due to the limitations of the tool such memory error thrown after running the model and time taken by the model to run for the entire dataset. Also, the impact of other attributes such as alcohol, parking etc. have been analysed using tableau.

```
import xgboost as xgb
from sklearn.model_selection import StratifiedKFold
from sklearn.model_selection import train_test_split
from imblearn.over_sampling import SMOTE
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix,accuracy_score
```

```
type(data1)
```

```
list
```

```
cols1 = ['latitude',
 'longitude',
 'stars_x',
 'review_count_x']
```

```
full_100=pd.read_excel("C:/Desktop/df_final.xlsx")
```

```
unique_df=full_100.drop_duplicates(subset=['business_id'],keep= 'last')
```

```
X = unique_df[cols1]
y = unique_df['is_open']
```

```
X.fillna(0.0,inplace=True)
```

```
C:\Users\apoor\Anaconda3\lib\site-packages\pandas\core\frame.py:4034: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html#indexing-view-versus-copy
  downcast=downcast, **kwargs)
```

```
train_X, test_X, train_y, test_y = train_test_split(X,y,test_size = 0.3, random_state = 42)
```



```
X_res = pd.DataFrame(X_res)
y_res = pd.DataFrame(y_res)
test_X = pd.DataFrame(test_X)
test_y = pd.DataFrame(test_y)
```

```
LR = LogisticRegression()
```

```
LR.fit(X_res,y_res)
pred_y = LR.predict(test_X)
```

```
C:\Users\apoor\Anaconda3\lib\site-packages\sklearn\utils\validation.py:73: DataConversionWarning: A column-vector y was pas
sed when a 1d array was expected. Please change the shape of y to (n_samples, ), for example using ravel().
  return f(**kwargs)
```

```
accuracy = 100*accuracy_score(test_y,pred_y)
print(accuracy)
```

```
60.0
```

```
pred_y
```

```
array([1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1,
       0, 1, 1, 1, 1, 1, 0, 1], dtype=int64)
```

Figure 11. Code Snippet of Predictive Model to predict the closure of Restaurants


Figure 12. Attributes that do not affect the functional status of a Restaurant

**Interpretation and solution:** The factors that play a major role in determining the closure of restaurants are latitude, longitude, average rating of the restaurants and review count. Other attributes such as alcohol, ambience etc. do not affect the functioning status of the restaurants as shown using tableau.

### 7.6 Word Cloud of Restaurants

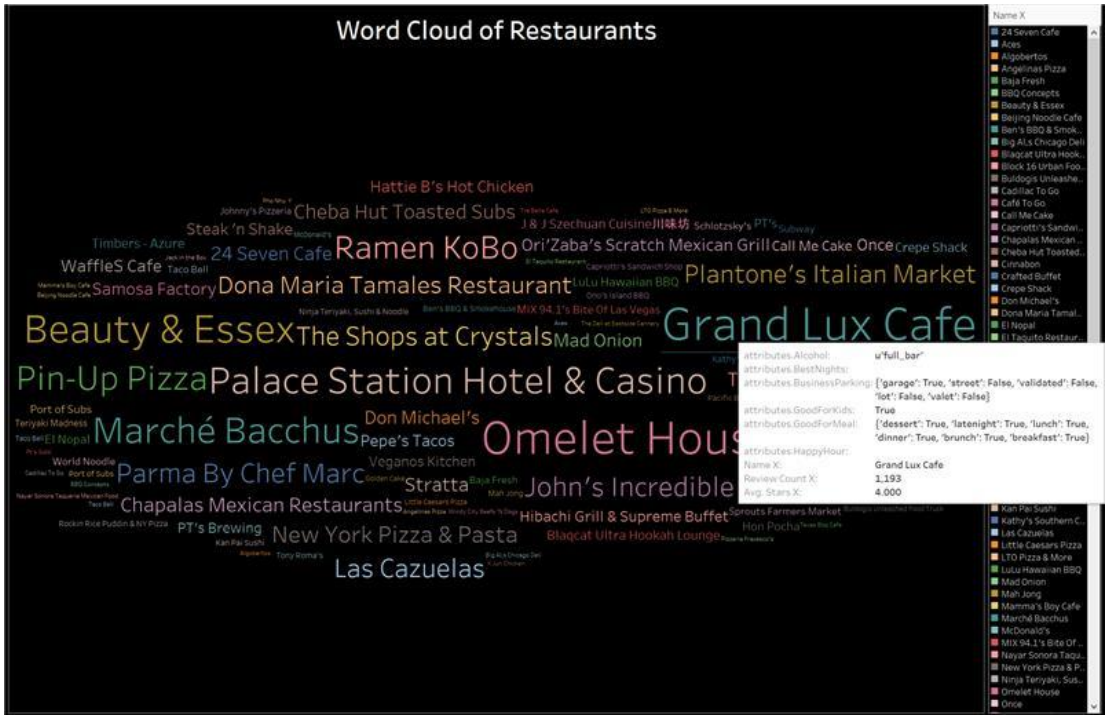



Figure 13. Word Cloud of Restaurants based on the Review Count

The word cloud specifies that Grand Lux Café, Omelet House, Palace Station Hotel & Casino, Pin-Up Pizza, Marche Bacchus and Beauty & Essex are seen to be prominently dominating over all the other restaurants in Las Vegas based on the review count indicating heavier traffic. From the list of above-mentioned restaurants, we particularly select Grand Lux Café taking into consideration the review count and average star rating. The average star rating of the restaurant stands at 4.0 even though the count of reviews received are at 1193 which is much lesser compared to Omelet House that has an average star rating of 4.5 with 1181 reviews. This led to the further analysis of Grand Lux Café as mentioned in Section 8 of this document.

### 7.7 Day-Wise Visits

**Business Problem:** What day of the week attracts the most traffic across Las Vegas?

Day-wise visit data is vital in determining when the business would be at a maximum. So is true in the case of restaurants, where the business is determined by the number of people coming into the restaurant. In common terms, more the customers, higher the business will be. The visits data provided in the yelp dataset consists of visits by customers for all businesses including restaurants. The data preparation is done on the basis of 100 randomly chosen restaurants and hence we filter the visits data for the same restaurants. The analysis here is to find out how many visits are done on a particular day of the week and at what time of the day. The output of analysis for visits based on day of the week is as below:

| hour | Mon | Tue | Wed | Thu | Fri | Sat | Sun |
|---|---|---|---|---|---|---|---|
| 0 | 685 | 490 | 477 | 434 | 490 | 660 | 899 |
| 1 | 864 | 710 | 669 | 669 | 754 | 955 | 1140 |
| 2 | 977 | 878 | 793 | 872 | 821 | 1158 | 1360 |
| 3 | 887 | 754 | 712 | 727 | 810 | 1179 | 1225 |
| 4 | 650 | 532 | 528 | 529 | 677 | 917 | 928 |
| 5 | 444 | 355 | 340 | 344 | 442 | 628 | 619 |
| 6 | 308 | 259 | 257 | 224 | 300 | 414 | 446 |
| 7 | 268 | 199 | 194 | 180 | 193 | 368 | 309 |
| 8 | 186 | 123 | 92 | 102 | 168 | 239 | 223 |
| 9 | 79 | 55 | 56 | 47 | 81 | 119 | 134 |
| 10 | 30 | 14 | 17 | 27 | 38 | 86 | 83 |
| 11 | 19 | 16 | 15 | 14 | 17 | 51 | 55 |
| 12 | 18 | 12 | 7 | 8 | 18 | 38 | 39 |
| 13 | 26 | 16 | 23 | 39 | 31 | 21 | 22 |
| 14 | 78 | 66 | 60 | 68 | 77 | 82 | 90 |
| 15 | 153 | 150 | 115 | 143 | 179 | 251 | 224 |
| 16 | 348 | 305 | 252 | 338 | 387 | 451 | 496 |
| 17 | 421 | 344 | 331 | 375 | 491 | 726 | 745 |
| 18 | 668 | 510 | 495 | 489 | 610 | 1009 | 1067 |
| 19 | 862 | 658 | 669 | 673 | 930 | 1359 | 1457 |
| 20 | 758 | 586 | 647 | 593 | 928 | 1409 | 1322 |
| 21 | 580 | 434 | 447 | 457 | 672 | 1060 | 930 |
| 22 | 421 | 365 | 341 | 405 | 572 | 900 | 713 |
| 23 | 409 | 351 | 342 | 339 | 488 | 728 | 580 |

Figure 14. Day Wise Visits

Also, determining at what hour of the day our maximum visits are taking place allows us to concentrate more on the business at that particular hour and also stock up the necessities for the fast-moving dishes.

Maximum of the visits on Monday: 977

Maximum of the visits on Tuesday: 878

Maximum of the visits on Wednesday: 793

Maximum of the visits on Thursday: 872

Maximum of the visits on Friday: 930

Maximum of the visits on Saturday: 1409

Maximum of the visits on Sunday: 1457

The sum of visits determines the busiest day of the week and hence the businesses can concentrate more on that particular day in terms of stock needed, providing better ambience and so on. Also, they can gather insights on how to improve the customer engagement on the other days where the traffic in and out of the restaurant is less.

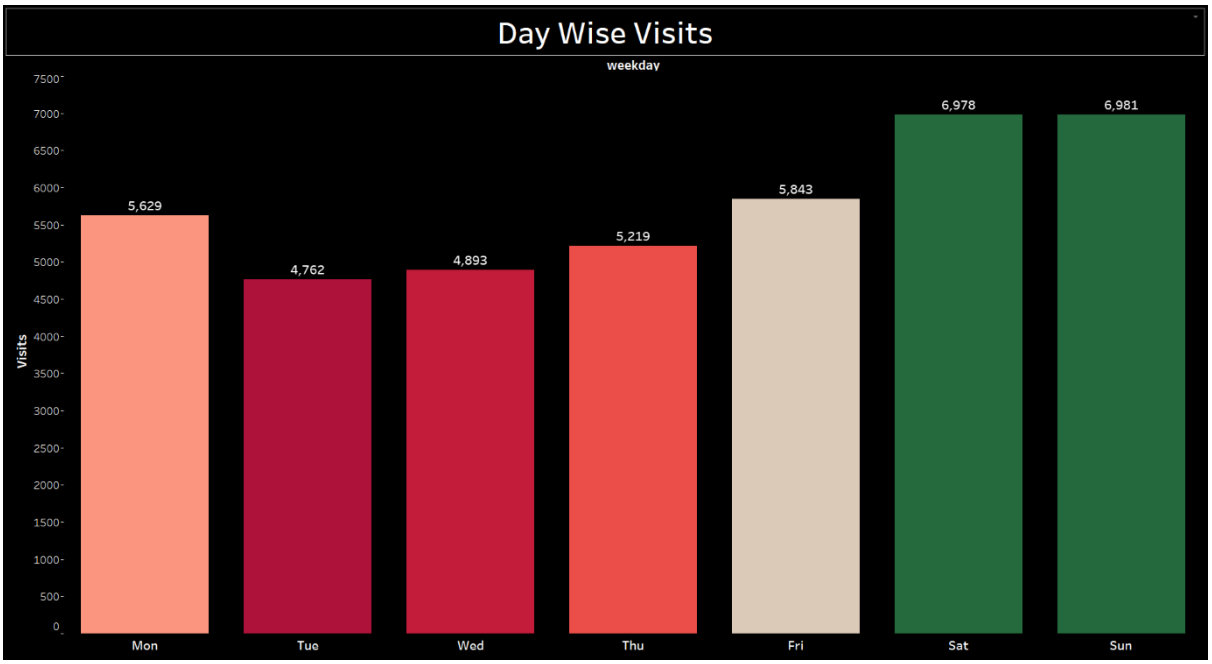



Figure 15. Day-wise customer visits based on the day of the week

**Interpretation and solution:** The graph plotted above using tableau indicates that the customer engagements and visits are at a high on weekends, i.e. 6981 on Sundays and 6978 on Saturdays. It is observed that Monday's and Friday's experience almost same number of visits in the range of 5500 to 6000. It is evident from the bar-graph that less traffic is observed on Tuesdays, Wednesdays and Thursdays, i.e., the mid-days of the week. Promotional activities or marketing campaigns can be implemented to increase sales and profits on these days adding value to the businesses.

# 8. GRAND LUX CAFÉ ANALYSIS

## 8.1 Resource Management

As a restaurant in Las Vegas, Grand Lux Café has seen the highest traffic among all the other restaurants. In order to manage traffic, it is essential for Grand Lux Café to recruit necessary staff for better management. In what months does GLC needs to increase its resources?

**Interpretation and solution:** To solve this problem, we separated Grand Lux Café (GLC) data from the main dataset and transformed the '*date*' column into '*year'*, '*month'* and '*year-month'* columns.

```
# Separate date column into year, month and time
GC_df['date'] = pd.to_datetime(GC_df['date'])
GC_df['year'] = GC_df['date'].dt.to_period('Y')
GC_df['month'] = GC_df['date'].dt.month
GC_df['year_month'] = GC_df['date'].dt.to_period('M')
GC_df.head()

# Save it as csv
GC_df.to_csv('Grand_Lux_Cafe.csv')
```

Figure 16. Transformation of date column into year, year-month and time

This transformation would enable us to analyse the GLC traffic granularly. However, there isn't a column specifying the traffic in GLC. But we do have reviews data given by all unique '*user_id*'. If we distribute the number of reviews over time from July 2008 to November 2018, we would be able to analyse the traffic throughout that timeline.

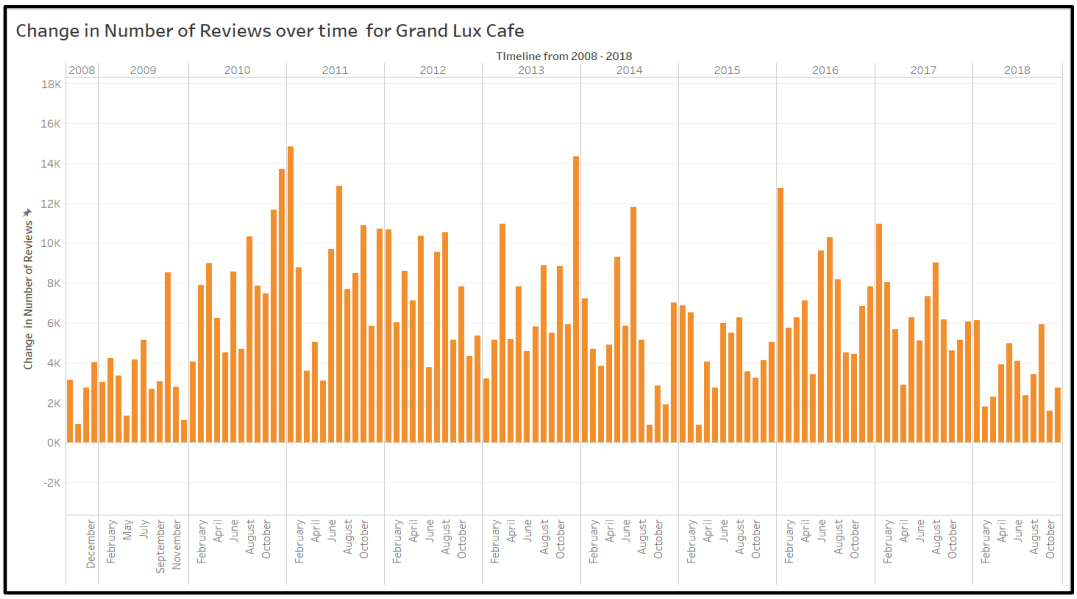



Figure 17. Change in Number of Reviews overtime from July 2008 – November 2018

Figure 17 shows that since its inception, GLC struggled to get traffic for two years. It can be seen that in the summertime (April - June) of 2010, the traffic gradually started rising. The traffic, however, peaked during the year end in the month of Dec 2010 and Jan 2011, eventually to fall during the summertime of 2011. The pattern between summertime and year end continues until 2012 where it seems GLC experienced less traffic throughout. This, however, changed after 2013 and similar summertime and year end pattern were observed throughout 2018.

As to answer the business question, it is clear that from October to January GLC experiences gradual to high traffic. Therefore, from management perspective they should recruit more workforce and manage its food supply from October to January.

### 8.2 Ratings between 2012-13 and Sentimental Analysis

From the above analysis, we could see a possible drop in traffic predominantly between 2012 – 13. It would be interesting to find, if there was a drop, how did that affect GLC overall ratings and what are those factors that might have led to this downfall in traffic?

**Interpretation and solution:** The solution of this business problem was solved in two parts. Firstly, we tried to make sense of the drop-in traffic for GLC since its inception in 2008 till 2018. However, the data for the years 2008 and 2018 was not adequate and therefore, we skipped these 2 years. This was achieved in Tableau by plotting an area graph of percentage difference of ratings from 2009 – 2017.

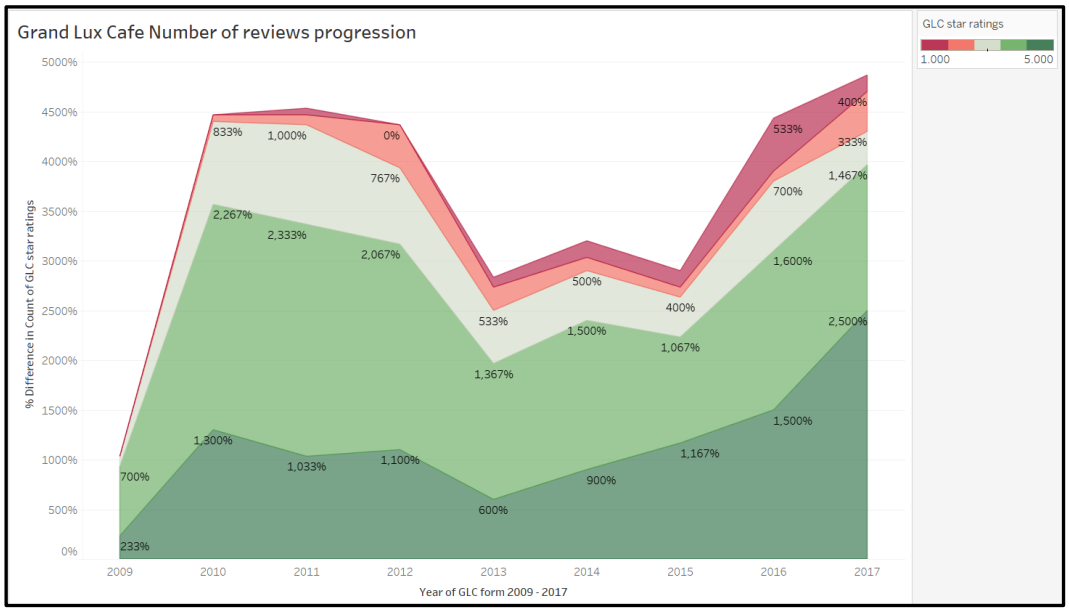



Figure 18. Percentage difference between Star ratings of Grand Lux Café

As we can observe from Figure 18, rating 4 (light green) overwhelmingly covers majority of GLC ratings. This justifies the average rating of GLC to be 4. In the year 2009, the number of reviews having 4 and 5 – star ratings increased by 700% and 233% respectively. Whereas in the year 2010, the business shot up, mostly from Sept-Dec (Figure 17) and the ratings increased for 1 and 2- star ratings - negligible, 3 - 833%, 4 - 2267% and 5 - 1300%.

The possible drop in Figure 17 gets confirmed in Figure 18. This slope from 2012 to 2013 affects the ratings throughout, numerically, for 3 – star rating = 234% drop, 4 – star rating = 700% drop and 5 – star rating = 500% drop. It could be due to this period of operation, the overall ratings of GLC despite receiving the highest traffic is 4.

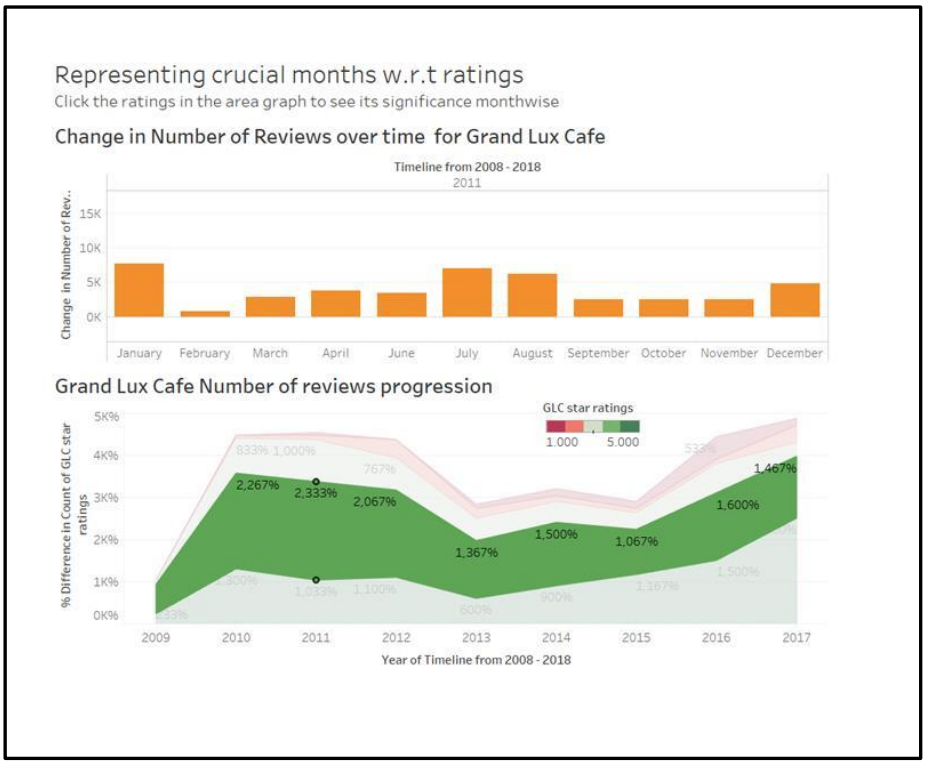



Figure 19. GLC ratings w.r.t months

Additionally, we created a dashboard in tableau using both the analysis representing crucial months w.r.t ratings. As seen in Figure 19, January, July, August and December months generate majority of 4-star ratings. This insight allows the restaurant to understand what months of the year produced high and low ratings.

This leads to the second part of the question as to what could be the factors that might have hampered the ratings. To answer this question, we decided to perform Sentimental Analysis of the '*text*' column in GLC data frame having reviews of users given to GLC between 2012 – 13

to find the factors that might have affected the decline in ratings. To achieve this, we performed Natural Language Programming (NLP) using Natural Language Toolkit (NLTK) package that is primarily used to work with human language in python. In order to use this text, it first needs to be pre-processed. For text processing using NLTK, we used tokenizing in order to split text into words and remove punctuations, numbers, stopwords.

```
# TEXT PRE_PROCESSING - Clean the review text data
# Returning if the word is adj or noun or verb or adverb
def get_wordnet(pos_tag):
    if pos_tag.startswith('J'):
        return wordnet.ADJ
    elif pos_tag.startswith('V'):
        return wordnet.VERB
    elif pos_tag.startswith('N'):
        return wordnet.NOUN
    elif pos_tag.startswith('R'):
        return wordnet.ADV
    else:
        return wordnet.NOUN

def clean_text(text):
    # lower case
    text = str(text).lower()
    # tokenize text and remove punctuations
    text = [word.strip(string.punctuation) for word in text.split(" ")]
    # remove words that contain numbers
    text = [word for word in text if not any(c.isdigit() for c in word)]
     # remove stop words
    stop = stopwords.words('english')
    text = [x for x in text if x not in stop]
    # remove empty tokens
    text = [t for t in text if len(t) > 0]
    # pos tag text
    pos_tags = pos_tag(text)
    print (pos_tags)
    # lemmatize text
    # t[0] is the first word in the pos_tags
    # get_wordnet(t[1]) is the first pos_tag i.e. is it adj, nn, v, rb
    text = [WordNetLemmatizer().lemmatize(t[0], get_wordnet(t[1])) for t in pos_tags]
    # remove words with only one letter
    text = [t for t in text if len(t) > 1]
    # join all
    text = " ".join(text)
    return(text)
```

Figure 20. Review Text Pre-processing and cleaning using NLTK package

Further, these thousands of words, must be differentiated as either verb, noun, adverb or adjective. This is performed by '*POS*' tag (Parts-of-speech), an *'NLTK'* library that assigns every word a tag to determine if it is either verb, noun, adverb or adjective. After which, we performed '*lemmatization*' that groups these words together to be analysed as a single term. It is essential as it brings context to the words. In an attempt to understand the meaning of these lemmatized words, we subjected them through '*wordnet*', a database lexicon, where these words were segregated into nouns, adjectives, adverbs and verbs. Performing these steps give context to the text with respect to noun, say, '*quick'*, '*staff'* where quick and staff are two different words but describes the sentiment of the user collectively, such as quick service. All the pre-processed text is stored into another column called *'drop_text_clean'*.

After pre-processing, these texts were given sentiments using Vader lexicon's '*SentimentIntensityAnalyzer*' that evaluates the intensity of the text as '*pos'*, '*neg'* and '*neu'*.

```
# add sentiment anaylsis columns
# Using vader lexicon of nltk package we assign sentiment to words with the help of SentimentIntensityAnalyzer
sid = SentimentIntensityAnalyzer()
GC_drop["GC_drop_sentiments"] = GC_drop["text"].apply(lambda x: sid.polarity_scores(x))
GC_drop = pd.concat([GC_drop.drop(['GC_drop_sentiments'], axis=1), GC_drop['GC_drop_sentiments'].apply(pd.Series)], axis=1)

GC_drop.shape
```

```
(246, 6)
```

```
# In order to segregate the pos and neg reviews assigning positive and negative polarity.
GC_drop['GC_drop_sentiments_score'] = GC_drop['text'].map(lambda text: TextBlob(text).sentiment.polarity)

#classify sentiment into positive and negative ones
GC_drop['GC_drop_sentiment'] = ''
GC_drop['GC_drop_sentiment'][GC_drop['GC_drop_sentiments_score'] > 0] = 'positive'
GC_drop['GC_drop_sentiment'][GC_drop['GC_drop_sentiments_score'] <= 0] = 'negative'
```

Figure 21. Adding Sentiments to the review text

In addition to that we also used '*Textblob'* that assigns sentiment score to the text, which is then assigned a text to be either positive or negative.

| | text | drop_text_clean | neg | neu | pos | compound | GC_drop_sentiments_score | GC_drop_sentiment |
|---|---|---|---|---|---|---|---|---|
| 0 | 5 stars for the staff.\n\nPacked throughout th... | star staff.\n\npacked throughout day warm mont... | 0.000 | 0.904 | 0.096 | 0.8990 | 0.151786 | positive |
| 1 | Ok so we got here for lunch the other day got ... | review purely late night din order take one ni... | 0.084 | 0.851 | 0.065 | -0.3777 | 0.165472 | positive |
| 2 | Here's an interesting fact: Grand Lux Cafe is... | another dessert adventure middle night feel hu... | 0.014 | 0.891 | 0.095 | 0.7862 | 0.272701 | positive |
| 3 | Ok, maybe its just me but I feel like Grand Lu... | go grand lux cafe palazzo own cheesecake facto... | 0.095 | 0.766 | 0.139 | 0.8313 | -0.004740 | negative |
| 4 | Love the Carrot Cake!! Have had good meals and... | ok get lunch day get seat right away.\n\nbut t... | 0.146 | 0.555 | 0.299 | 0.7835 | 0.240179 | positive |
| 5 | NOM NOM NOM NOM NOM with a side of NOM. So goo... | here's interest fact grand lux cafe own cheese... | 0.000 | 0.658 | 0.342 | 0.9382 | 0.812500 | positive |
| 6 | one of the best places in Vegas, open 24 hours... | use think fan grand lux cafe guiltily love che... | 0.000 | 0.573 | 0.427 | 0.9341 | 0.515000 | positive |
| 7 | I work at the Cheesecake Factory in Chicago so... | love come affordable tasty comfort food vega r... | 0.000 | 0.727 | 0.273 | 0.9915 | 0.340521 | positive |
| 8 | First the hostess with the red bob sucks at he... | ok maybe feel like grand lux look like cheesec... | 0.083 | 0.863 | 0.055 | -0.4215 | 0.040476 | positive |
| 9 | I always enjoy coming here\n\nBig portions and... | love carrot cake good meal bad good service te... | 0.000 | 0.821 | 0.179 | 0.9196 | 0.404004 | positive |

Figure 22. Output after adding sentiments using Vader and Textblob

To build the word clouds the machine must consider all the text as a single document. This was achieved by using two models, namely '*Doc2Vec'* and '*Tf-Id*' in '*genism'* library. '*Doc2Vec'* model converts a document into vector form for the machine to understand it better, whereas '*TF-ID*' model takes this vector as an input to further refine the text such that it could understand the meaning of the sentence.

```
# create doc2vec vector columns

documents = [TaggedDocument(doc, [i]) for i, doc in enumerate(GC_drop["drop_text_clean"].apply(lambda x: x.split(" ")))]

# train a Doc2Vec model with our text data
model_1 = Doc2Vec(documents, vector_size=5, window=2, min_count=1, workers=4)

# transform each document into a vector data
doc2vec_df = GC_drop["drop_text_clean"].apply(lambda x: model_1.infer_vector(x.split(" "))).apply(pd.Series)
doc2vec_df.columns = ["doc2vec_vector_" + str(x) for x in doc2vec_df.columns]
GC_drop = pd.concat([GC_drop, doc2vec_df], axis=1)
GC_drop
```

```
# add tf-idfs columns

tfidf = TfidfVectorizer(min_df = 10)
tfidf_result = tfidf.fit_transform(GC_drop["drop_text_clean"]).toarray()
tfidf_df = pd.DataFrame(tfidf_result, columns = tfidf.get_feature_names())
tfidf_df.columns = ["word_" + str(x) for x in tfidf_df.columns]
tfidf_df.index = GC_drop.index
GC_drop = pd.concat([GC_drop, tfidf_df], axis=1)
tfidf_df
```

Figure 23. Sentiment Analysis using genism models

Further, the reviews were separated and stored into negative (‘*GC_drop _neg’*) and positive (‘*GC_drop _pos’*) data frames. These data frames are then used for words clouds.

After the Sentimental Analysis of the text, we plot the positive and negative words understanding their context and importance to the reviews or text as a whole. However, we still encountered few stopwords and had to remove these words manually from the word cloud.

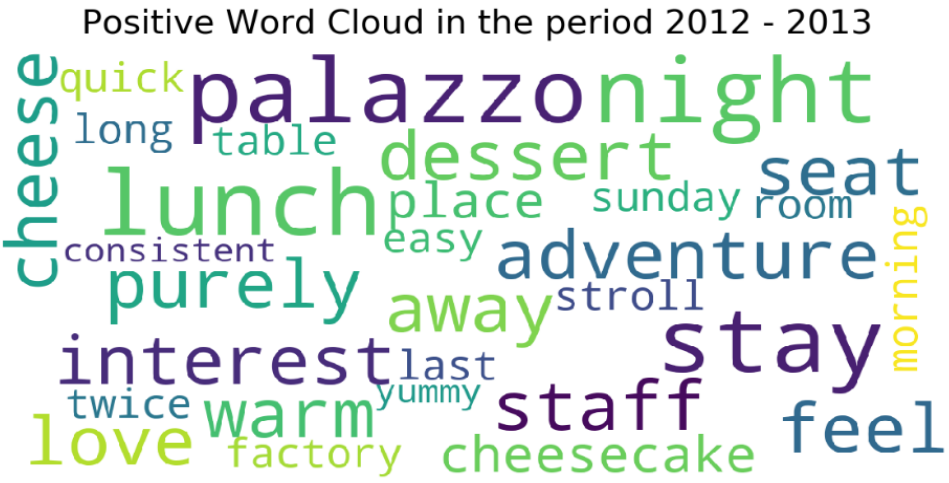



Figure 24. Grand Lux Café positive review text word cloud between 2012 -13

As seen in Figure 24, words like *‘staff’, ‘warm’, ‘interest’, ‘quick’* might signify the positive attitude of staff towards the customers. Moreover, in food the *‘lunch’, ‘dessert’, ‘cheesecake’* reciprocated positive feedback among customers.

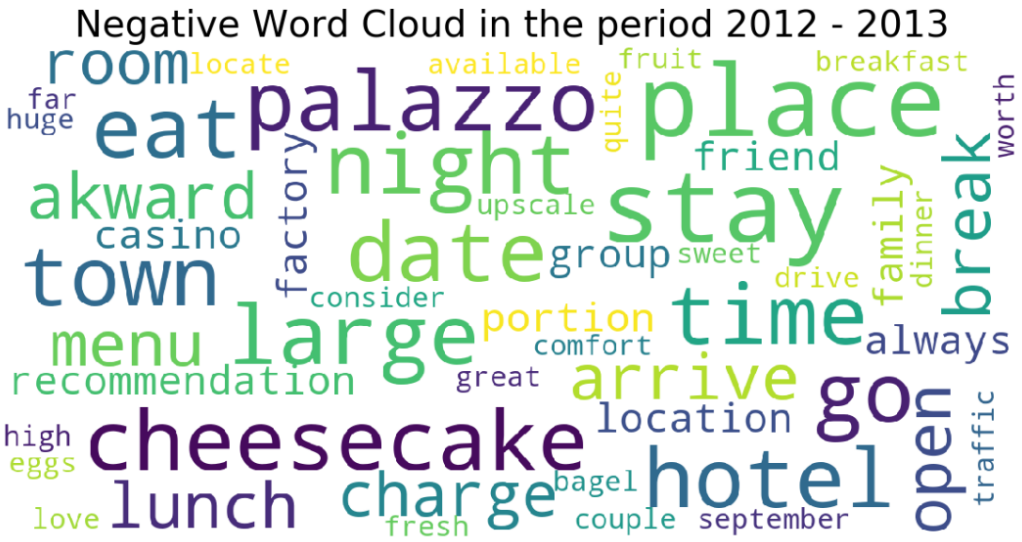



Figure 25. Grand Lux Café negative review text word cloud between 2012 -13

But in order to answer our business question, lets focus on the negative word cloud, Figure 25 where we can clearly identify few major factors for instance, *'comfort', 'akward', 'charge', 'traffic', 'location'*, *'place'*. In addition to that, there are also specific words that relates to food i.e. *'cheesecake', 'breakfast', 'lunch', 'bagel', 'family', 'dinner'*. Therefore, from the above analysis we could infer that, customers might have felt uncomfortable with regards to the location, traffic and charge of restaurant which might have led to the downfall in their rating from 2012 to 2013.

## 9. LIMITATIONS

- Due to performance issue of tools, we restricted the analysis to 100 restaurants as multiple reviews for each restaurant increases the number of rows in turn increasing the data size.
- As we tried to run the predictive model for the entire dataset, a memory error was thrown and hence we decided to run it only for 100 restaurants.
- Huge size of files (almost 10GB), tools such as python and tableau were not able to take the load, which led to merging issues as well as big data problems.
- Lack of certain fields such as profits, revenues, turnovers of the restaurants limited our capability of performing in depth analysis.
- Due to lack of user information such as Age, we were restricted to categorise the users for a particular restaurant into groups, in our case Grand Lux Café. Having this information would have enabled us to identify top users under different age groups and further providing promotional offers for establishing marketing strategy.

# 10. RECOMMENDATIONS

Las Vegas being a tourist hub attracts huge crowd throughout the year that prominently contributes to the food venture business and also proves beneficial for the country's economy. Recommendations from our analysis provides an ideal scenario for an individual or an establishment to make decisions on opening a new venture or franchise based on the geographical analysis of open and closed restaurants. Analysis of the users performed leads us in determining the top influencers whose reviews were found to be beneficial by a majority of users using Yelp. Restaurants can make major decisions by focusing on the reviews provided by top influencers, either negative or positive, in order to work towards attracting customers by providing better experience. Analysis further unfolds that recommends the owners of restaurants to be prepared very well in advance based on the traffic inflow to the restaurant determined by the day-wise visits. Analysis is also concentrated on individual restaurants to generate monthly traffic analysis and monthly restaurant rating analysis using count of reviews and ratings. Additionally, a detailed sentimental analysis is performed to generate positive and negative feedback using review text that determine the Do's and Don'ts for a restaurant. The above recommendations can be used by a stakeholder to make informed business decisions.

# 11. CONCLUSION

From the above analysis performed, we conclude that location, star rating of restaurants and review count play a vital role in determining the functional status of the restaurant. On the basis of this analysis, any restaurant can leverage on these factors to choose a particular location for a new venture or franchise. Furthermore, analysis of top influencers based on the location, rating and reviews will prove beneficial to the establishment to improve on the services that they provide. Day-wise visit analysis provides the organisation with information related to traffic coming into the restaurant on a daily basis and take measures such as stock up groceries, provide exciting offers and discounts. Diving deep into further analysis, we can see that Grand Lux Café has experienced a downfall due to certain factors such as location, traffic and charge of the restaurant which can be seen from the sentimental analysis using word clouds. Grand Lux Café can further aim for the betterment of their business by taking into consideration the factors mentioned above. The entire process of this analysis can be replicated to any restaurant of interest in order to find ways to improve business of the establishment.

# 12. APPENDIX

Analysis Code Snippets

```
In [67]: df_business=pd.read_csv("C:/Desktop/NUIG assignments sem2/Projects/FINAL PROJECT/vegas_business.csv")

In [3]: df_business.columns

Out[3]: Index(['Unnamed: 0', 'business_id', 'name', 'address', 'city', 'state',
       'postal_code', 'latitude', 'longitude', 'stars', 'review_count',
       'is_open', 'categories', 'hours', 'attributes.GoodForKids',
       'attributes.RestaurantsReservations', 'attributes.GoodForMeal',
       'attributes.BusinessParking', 'attributes.Caters',
       'attributes.NoiseLevel', 'attributes.RestaurantsTableService',
       'attributes.RestaurantsTakeOut', 'attributes.RestaurantsPriceRange2',
       'attributes.OutdoorSeating', 'attributes.BikeParking',
       'attributes.Ambience', 'attributes.HasTV', 'attributes.WiFi',
       'attributes.Alcohol', 'attributes.RestaurantsAttire',
       'attributes.RestaurantsGoodForGroups', 'attributes.RestaurantsDelivery',
       'hours.Monday', 'hours.Tuesday', 'hours.Wednesday', 'hours.Thursday',
       'hours.Friday', 'hours.Saturday', 'hours.Sunday',
       'attributes.BusinessAcceptsCreditCards', 'attributes',
       'attributes.BusinessAcceptsBitcoin', 'attributes.ByAppointmentOnly',
       'attributes.AcceptsInsurance', 'attributes.Music',
       'attributes.GoodForDancing', 'attributes.CoatCheck',
       'attributes.HappyHour', 'attributes.BestNights',
       'attributes.WheelchairAccessible', 'attributes.DogsAllowed',
       'attributes.BYOBCorkage', 'attributes.DriveThru', 'attributes.Smoking',
       'attributes.AgesAllowed', 'attributes.HairSpecializesIn',
       'attributes.Corkage', 'attributes.BYOB',
       'attributes.DietaryRestrictions', 'attributes.Open24Hours',
       'attributes.RestaurantsCounterService'],
      dtype='object')

In [68]: df_business=df_business[df_business['categories'].apply(lambda x: True if 'Restaurants' in str(x)else False)].reset_index(dr
         df_reviews=pd.read_csv("C:/Desktop/NUIG assignments sem2/Projects/FINAL PROJECT/vegas_review.csv")
```

```
In [69]: df10=df_business.drop_duplicates(subset=['business_id'],keep= 'last')

In [70]: df10['business_id'].count()

Out[70]: 6450

In [9]:  1 df10=df10.replace(to_replace=[0,1], value = ['Closed','Open'])

In [23]: plt.figure(figsize=(8,8))

         sns.set(font_scale=2)
         ax=sns.countplot(x='is_open',data=df10,hue='is_open')
         ax.set(xlabel='Status of the Restaurant functioning',ylabel='Count of Restauarnts')
         plt.legend(['Closed permanently','Open'])
```

```
In [26]: df_businessone= df_business.drop_duplicates(subset=['business_id'],keep= 'last')

In [29]: ay=sns.countplot(x='stars',data=df_businessone)
         ay.set(xlabel='Ratings of the Restaurants',ylabel='Number of Restaurants')

Out[29]: [Text(0, 0.5, 'Number of Restaurants'),
          Text(0.5, 0, 'Ratings of the Restaurants')]
```

```
In [236]: df_final['business_id'].count()

Out[236]: 17191

In [32]: df4=pd.read_excel("C:/Desktop/df_final.xlsx")

In [237]: df6=df4.drop_duplicates(subset=['business_id'],keep= 'last')

In [238]: df6

Out[238]:
```

| | Unnamed: 0 | Unnamed: 0.1 | business_id | name_x | address | city | state | postal_code | latitude | longitude | ... | compliment_ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 2 | 563 | THOZYeKWKyW_-6U30E1b5w | Angelinas Pizza | O' Aces Bar, 3003 N Rainbow | Las Vegas | NV | 89108 | 36.213848 | -115.242603 | ... | |
| 12 | 12 | 590 | 707ST4xhNavXNvZtEL-RgQ | Windy City Beefs 'N Dogs | 7006 W Charleston Blvd | Las Vegas | NV | 89145 | 36.159616 | -115.251523 | ... | |
| 167 | 167 | 918 | r-wz3rmeNGLBMOqtfXnxog | Don Michael's | 4864 W Lone Mountain Rd | Las Vegas | NV | 89130 | 36.247280 | -115.208920 | ... | |
| 366 | 366 | 997 | 0G1y7uV7w1D2uziS4Lt_Dw | Chapalas Mexican Restaurants | 3331 E Tropicana Ave | Las Vegas | NV | 89121 | 36.099613 | -115.103913 | ... | |
| | | | | | 6475 W | | | | | | | |

```
In [239]: df6=df6.replace(to_replace=[0,1], value = ['Closed','Open'])

In [240]: df6

Out[240]:
```

| | Unnamed: 0 | Unnamed: 0.1 | business_id | name_x | address | city | state | postal_code | latitude | longitude | ... | compliment_ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 2 | 563 | THOZYeKWKyW_-6U30E1b5w | Angelinas Pizza | O' Aces Bar, 3003 N Rainbow | Las Vegas | NV | 89108 | 36.213848 | -115.242603 | ... | |
| 12 | 12 | 590 | 707ST4xhNavXNvZtEL-RgQ | Windy City Beefs 'N Dogs | 7006 W Charleston Blvd | Las Vegas | NV | 89145 | 36.159616 | -115.251523 | ... | |
| 167 | 167 | 918 | r-wz3rmeNGLBMOqtfXnxog | Don Michael's | 4864 W Lone Mountain Rd | Las Vegas | NV | 89130 | 36.247280 | -115.208920 | ... | |
| 366 | 366 | 997 | 0G1y7uV7w1D2uziS4Lt_Dw | Chapalas Mexican Restaurants | 3331 E Tropicana Ave | Las Vegas | NV | 89121 | 36.099613 | -115.103913 | ... | |

```
In [241]: plt.figure(figsize=(8,8))

          sns.set(font_scale=2)
          ax=sns.countplot(x='is_open',data=df6,hue='is_open')
          ax.set(xlabel='Status of the Restaurant functioning',ylabel='Count of Restauarnts')
          plt.legend(['Closed permanently','Open'])

Out[241]: <matplotlib.legend.Legend at 0x1691ef1c128>

In [242]: ay=sns.countplot(x='stars_x',data=df6)
          ay.set(xlabel='Ratings of the Restaurants',ylabel='Number of Restaurants')
```

Code 1: Code for bar plots of open and closed restaurants & count of restaurants

## 13. BIBLIOGRAPHY

[1] Hicks, A., Comp, S., Horovitz, J., M., Miki, M., & Bevan, J. L. (2012). *Why people use Yelp.com: An exploration of uses and gratifications - ScienceDirect*. https://www.sciencedirect.com/science/article/pii/S0747563212001951

[2] Yelp.com. (2020). *Yelp Dataset*. [online] Available at: https://www.yelp.com/dataset/challenge [Accessed 1 Mar. 2020].

[3] Gnanaskandan, K., Shadbad, F., Goud, P., & Mereddy, S. *Yelp Data Set Challenge (What drives restaurant ratings?)*. https://www.slideshare.net/prashanth1957/yelp-data-set-challenge-what-drives-restaurant-ratings?from_action=save