Jaypee Institute of Information Technology, Noida

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING AND INFORMATION TECHNOLOGY

Project Title: Yelp Review Analysis

Enrol. No.	Name of Student
9921103030	Ayushi Tripathi
9921103032	Vishal Kumar

Course Name: Introduction to Big Data and Data Analytics

Course Code: 20B12CS333

Program: B.Tech. CS&E/B.Tech. IT

3rd Year 5th Sem

2023-24

Table of Contents

1. ABSTRACT:

In the digital age, where information is at our fingertips, online reviews have become a powerful influence on consumer decisions. Review analysis, the systematic examination of user-generated feedback, has emerged as a crucial aspect of understanding customer sentiments, preferences, and overall satisfaction. This practice involves employing various techniques, often rooted in natural language processing and machine learning, to extract valuable insights from the vast pool of reviews available on the internet.

Review analysis provides a panoramic view of how products, services, or experiences are perceived by the public. In e-commerce, for instance, potential buyers heavily rely on reviews to gauge the quality and reliability of a product. Similarly, in the realm of hospitality, restaurants, hotels, and other establishments thrive or falter based on the feedback they receive. Review analysis enables businesses to comprehend the dynamics of customer satisfaction, identify areas of improvement, and capitalize on strengths.

Review analysis empowers businesses to adopt a customer-centric approach. By delving into the language and sentiments expressed in reviews, companies can gain a nuanced understanding of their customers' needs and expectations. This information is invaluable for refining existing products, introducing new features, or tailoring services to align with customer preferences. In essence, it facilitates a continuous feedback loop that fosters adaptive and customer-focused decision-making.

In a competitive market, staying ahead requires more than just a quality product; it demands a comprehensive understanding of the market landscape. Review analysis equips businesses with the tools to benchmark their performance against competitors. By studying the strengths and weaknesses highlighted in reviews, companies can strategically position themselves, identify unique selling propositions, and differentiate their offerings in the market.

Review analysis is not merely retrospective; it extends to predictive analytics. By identifying patterns in historical reviews, businesses can anticipate trends and forecast customer preferences. This foresight is invaluable for staying agile in a rapidly evolving market, enabling companies to proactively adapt to changing consumer expectations and emerging industry trends.

In the decision-making hierarchy of modern businesses, data-driven insights hold a paramount position. Review analysis contributes to evidence-based decision-making by providing a rich source of qualitative and quantitative data. Whether it's fine-tuning marketing strategies, optimizing customer service protocols, or refining product development, reviews offer a treasure trove of actionable intelligence.

2. INTRODUCTION:

2.1.PROBLEM STATEMENT

In the realm of culinary entrepreneurship, the decision to open a restaurant is multifaceted, with location playing a pivotal role in determining success. This project aims to leverage advanced data analysis techniques on a cleaned and processed dataset containing restaurant reviews to provide actionable insights for prospective restaurant owners. The focus is on advising whether a given location within the city is conducive to establishing a new restaurant.

2.2.MOTIVATION

The motivation behind undertaking this project stems from the recognition of the challenges faced by aspiring restaurant owners and investors in the highly competitive and dynamic hospitality industry. Establishing a restaurant involves a myriad of decisions, with one of the most critical being the choice of location. This project seeks to empower individuals in the culinary domain with data-driven insights, transforming the traditionally intuition-based decision-making process into a more informed and strategic endeavor.

- Enhancing Decision-Making: The project recognizes the limitations of subjective decision-making in the context of restaurant establishment. By applying advanced analytical techniques to the dataset, it seeks to enhance decision-making processes, offering stakeholders a more comprehensive and objective evaluation of potential locations. In doing so, the project aligns with the broader trend of embracing data-driven strategies in various industries.

- Addressing Industry Challenges: The restaurant industry is constantly evolving, influenced by factors such as changing consumer preferences, economic fluctuations, and global events. This project aims to address the dynamic nature of the industry by providing actionable insights derived from a thorough analysis of restaurant reviews. By understanding current trends and sentiments, entrepreneurs can position themselves strategically in the market.

- Facilitating Entrepreneurship: Encouraging entrepreneurship in the culinary sector is a key motivation. By offering valuable information on popular cuisines, successful restaurant types, and economically viable locations, the project serves as a catalyst for individuals aspiring to enter the restaurant business. This aligns with the broader goal of fostering innovation and diversity in the entrepreneurial landscape.

2.3.OBJECTIVE

- Sentiment Analysis: Conduct sentiment analysis on the reviews to gauge the overall sentiment towards restaurants in different locations. Understand the prevailing sentiments, both positive and negative, to assess the general perception of dining establishments.

- Identifying Popular Cuisines and Dishes: Analyze the dataset to identify the most popular cuisines and dishes among customers. This information can guide menu planning for a new restaurant, ensuring alignment with local preferences.

- Comparative Analysis of Ratings: Compare and contrast the ratings and reviews of existing restaurants in various locations. Determine if there are particular areas with a higher concentration of well-rated establishments, indicating a potentially favorable environment for a new restaurant.

- Correlation with Economic Factors: Explore any correlations between the economic factors of a location and the success of restaurants. Factors such as the average cost for two people, the type of restaurant, and customer reviews could provide insights into the economic viability of opening a new establishment.

- Spatial Analysis: Utilize spatial analysis to map out restaurant distribution across the city. Identify clusters or gaps in restaurant availability, helping to pinpoint areas with untapped market potential.

2.4.CONTRIBUTION

Ayushi Tripathi – Preprocessing and SVM
Vishal Kumar – Visualization

3. DETAILED DESCRIPTION OF THE PROJECT:

3.1.PROPOSED WORK WITH TOOLS AND DATASETS USED

In this project, Jupyter Notebook served as the primary software for performing data analysis using Python. Employing essential libraries such as scikit-learn (sklearn), matplotlib, seaborn, pandas, and numpy, we executed robust data cleaning procedures. The dataset underwent thorough preprocessing to ensure its readiness for analysis. Leveraging the SVM (Support Vector Machine) machine learning algorithm, we performed insightful analysis on the cleaned data. Visualizations, crafted with matplotlib and seaborn, provided a clear and intuitive representation of key patterns and trends, enhancing the interpretability of the dataset and contributing to the overall efficacy of the project.

Software Requirement:

- ❖ Anaconda environment
- ❖ Jupyter Notebook

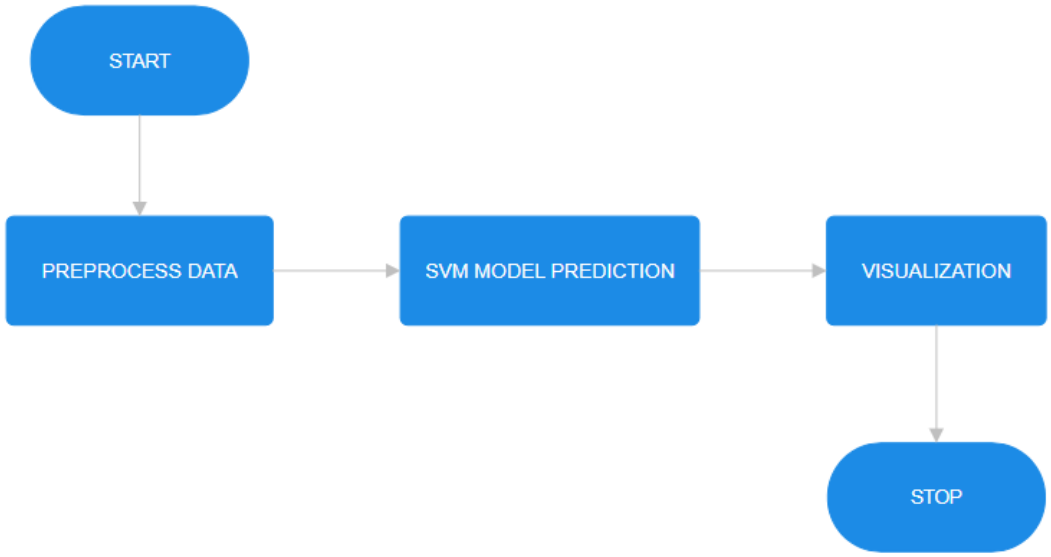
Language Used:

- ❖ Python

Dataset Used:

https://www.kaggle.com/datasets/himanshupoddar/zomato-bangalore-restaurants

3.2.WORKFLOW DIAGRAM



4. IMPLEMENTATION:

4.1.PROGRAM CODE

```
#!/usr/bin/env python
# coding: utf-8
# In[270]:


import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.model_selection import train_test_split
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score,roc_curve
plt.style.use('dark_background')
```

```
# In[271]:


df=pd.read_csv('zomato.csv')
df.head()


# In[272]:


df.shape


# In[273]:


df.columns


# In[274]:


df=df.drop(['url','address','phone','menu_item','dish_liked','reviews_list'],a
xis=1)
df.head()


# In[275]:


df.info()


# In[276]:


df.drop_duplicates(inplace=True)
df.shape


# In[277]:


df['rate'].unique()


# In[278]:


def handlerate(value):
    if(value=='NEW' or value=='-'):
        return np.nan
    else:
        value=str(value).split('/')
        value=value[0]
```

```
        return float(value)

df['rate']=df['rate'].apply(handlerate)
df['rate'].head()


# In[279]:


df.rate.isnull().sum()


# In[280]:


df['rate'].fillna(df['rate'].mean(),inplace=True)
df['rate'].isnull().sum()


# In[281]:


df.info()


# In[282]:


df.dropna(inplace=True)
df.head()


# In[283]:


df.rename(columns={'approx_cost(for two
people)':'cost2plates','listed_in(type)':'type'},inplace=True)
df.head()


# In[284]:


df['location'].unique()


# In[285]:


df['listed_in(city)'].unique()


# In[286]:


df = df.drop(['listed_in(city)'], axis=1)
df.columns
```

```
# In[287]:


df['cost2plates'].unique()


# In[288]:


def handlecomma(value):
    value=str(value)
    if ',' in value:
        value=value.replace(',','')
        return float(value)
    else:
        return float(value)

df['cost2plates']=df['cost2plates'].apply(handlecomma)
df['cost2plates'].unique()


# In[289]:


df.head()


# In[290]:


df['rest_type'].value_counts()


# In[291]:


rest_types=df['rest_type'].value_counts()
rest_types


# In[292]:


rest_types_lessthan1000=rest_types[rest_types<1000]
rest_types_lessthan1000


# In[293]:


def handle_rest_type(value):
    if(value in rest_types_lessthan1000):
        return 'others'
    else:
        return value
```

```
df['rest_type']=df['rest_type'].apply(handle_rest_type)
df['rest_type'].value_counts()


# In[294]:


df.head()


# In[295]:


df['location'].value_counts()


# In[296]:


location=df['location'].value_counts()
location_lessthan300=location[location<300]

def handle_location(value):
    if(value in location_lessthan300):
        return 'others'
    else:
        return value

df['location']=df['location'].apply(handle_location)
df['location'].value_counts()


# In[297]:


df.head()


# In[298]:


df['cuisines'].value_counts()


# In[299]:


cuisines=df['cuisines'].value_counts()
cuisines_lessthan100=cuisines[cuisines<100]

def handle_cuisines(value):
    if(value in cuisines_lessthan100):
        return 'others'
    else:
        return value
```

```
df['cuisines']=df['cuisines'].apply(handle_cuisines)
df['cuisines'].value_counts()


# In[300]:


df.head()


# In[301]:


vectorizer = CountVectorizer()
X = vectorizer.fit_transform(df['cuisines'])

y = df['rate'] > 4

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

svm = SVC()
svm.fit(X_train, y_train)

y_pred = svm.predict(X_test)

accuracy_score(y_test, y_pred)


# In[302]:


df['type'].value_counts()


# In[303]:


location_counts = df['location'].value_counts()

plt.figure(figsize=(16,10))
sns.barplot(x=location_counts.index, y=location_counts.values)
plt.xticks(rotation=90)
plt.show()


# In[304]:


df['online_order']


# In[305]:


online_orders = df['online_order'].value_counts()
online_orders.plot(kind='bar',color=['blue', 'magenta'])
```

```
plt.title("Online Order Distribution")
plt.xlabel("Ordered Online?")
plt.ylabel("Count")
plt.show()


# In[306]:


book = df['book_table'].value_counts()

plt.figure(figsize=(6,6))
sns.barplot(x=book.index, y=book.values)
plt.xticks(rotation=90)
plt.show()


# In[307]:


plt.figure(figsize=(6,6))
sns.boxplot(x='online_order',y='rate',data=df)


# In[308]:


plt.figure(figsize=(6,6))
sns.boxplot(x='book_table',y='rate',data=df)


# In[309]:


df1=df.groupby(['location','online_order'])['name'].count()
df1.to_csv('location_online.csv')
df1=pd.read_csv('location_online.csv')
df1=pd.pivot_table(df1,values=None,index=['location'],columns=['online_order']
,fill_value=0,aggfunc=np.sum)
df1


# In[310]:


df1.plot(kind='bar',figsize=(15,8))


# # Book Table vs Location

# In[311]:


df2=df.groupby(['location','book_table'])['name'].count()
df2.to_csv('location_booktable.csv')
df2=pd.read_csv('location_booktable.csv')
```

```
df2=pd.pivot_table(df2,values=None,index=['location'],columns=['book_table'],f
ill_value=0,aggfunc=np.sum)
df2


# In[312]:


df2.plot(kind='bar',figsize=(15,8))


# # Type of Restaurant vs Rate

# In[313]:


plt.figure(figsize=(14,8))
sns.boxplot(x='type',y='rate',data=df,palette='inferno')


# # Type of Restaurant vs Location

# In[314]:


df3=df.groupby(['location','type'])['name'].count()
df3.to_csv('location_type.csv')
df3=pd.read_csv('location_type.csv')
df3=pd.pivot_table(df3,values=None,index=['location'],columns=['type'],fill_va
lue=0,aggfunc=np.sum)
df3


# In[315]:


df3.plot(kind='bar',figsize=(36,8))


# # No. of votes vs Location

# In[316]:


df4=df[['location','votes']]
df4.drop_duplicates()
df5=df4.groupby(['location'])['votes'].sum()
df5=df5.to_frame()
df5=df5.sort_values('votes',ascending=False)
df5.head()


# In[317]:


plt.figure(figsize=(15,8))
sns.barplot(x=df5.index,y=df5['votes'])
```

```
plt.xticks(rotation=90)


# In[318]:


df6=df[['cuisines','votes']]
df6.drop_duplicates()
df7=df6.groupby(['cuisines'])['votes'].sum()
df7=df7.to_frame()
df7=df7.sort_values('votes',ascending=False)
df7.head()


# In[319]:


df7=df7.iloc[1:,:]
df7.head()


# In[320]:


plt.figure(figsize=(15,8))
sns.barplot(x=df7.index,y=df7['votes'])
plt.xticks(rotation=90)
```

4.2.RESULT ANALYSIS

From figure 1, we can see that BTM location in the city has most number of restaurants whereas St. Marks Road has the least. One should open a restaurant in an area with less number of restaurants like Old Airport Road.

From figure 4, we can see that the restaurants that offer online ordering have higher ratings as compared to the ones that don't.

From figure 5, we can see that the restaurants that offer table booking have higher ratings as compared to the ones that don't.

From figure 6, we can conclude that opening a restaurant in areas where there are less restaurants offering online order will prove beneficial.

From figure 7, we can conclude that opening a restaurant in areas where there are less restaurants offering table booking will prove beneficial.

From figure 8, we can see that 'Drinks & nightlife' and 'Pubs and bars' have the highest average ratings.

From figure 9, we can see that, for example, opening 'Drinks & nightlife' type restaurant in 'Commercial Street' would prove beneficial as there as not such restaurants in that location.

From figure 11, we can see that North Indian, Chinese and South Indian cuisine restaurants have the most votes. Restaurants serving these cuisines can benefit a lot.

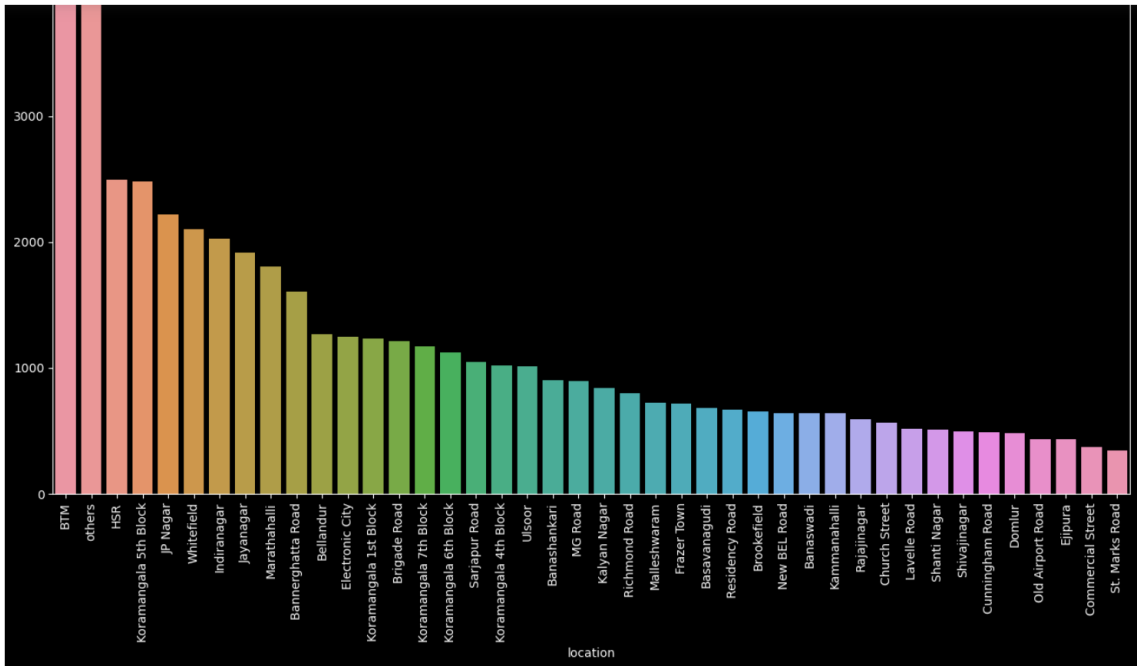


Fig 1. Number of restaurants in a location

Fig 2. Number of restaurants offering online order

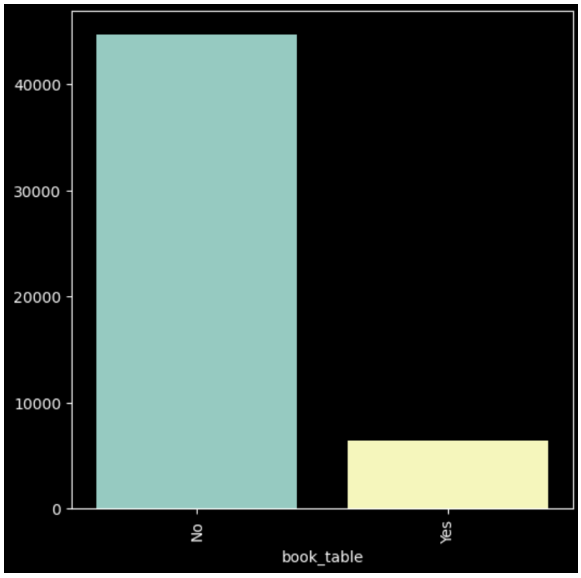


Fig 3. Number of restaurants offering table booking

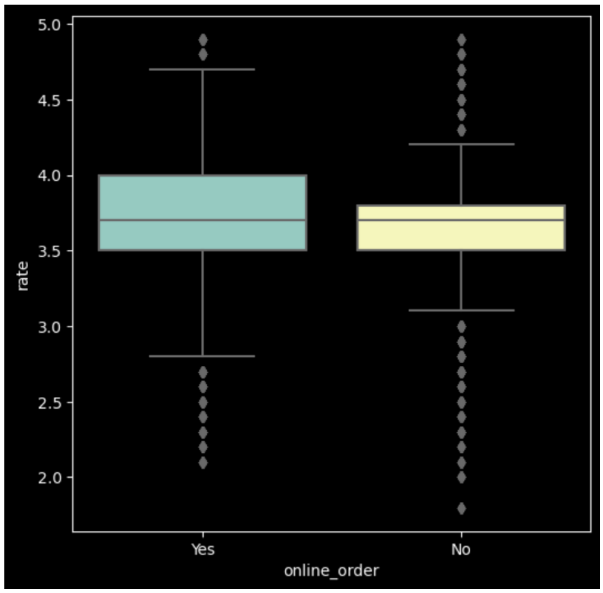


Fig 4. Rating vs Online order

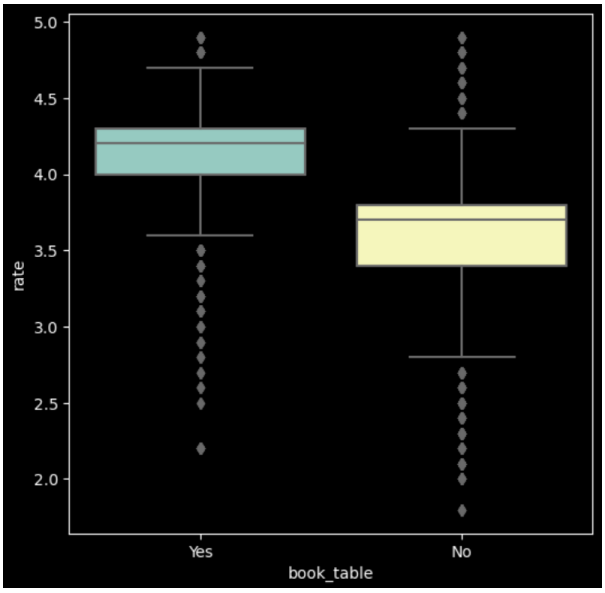


Fig 5. Rating vs Book table

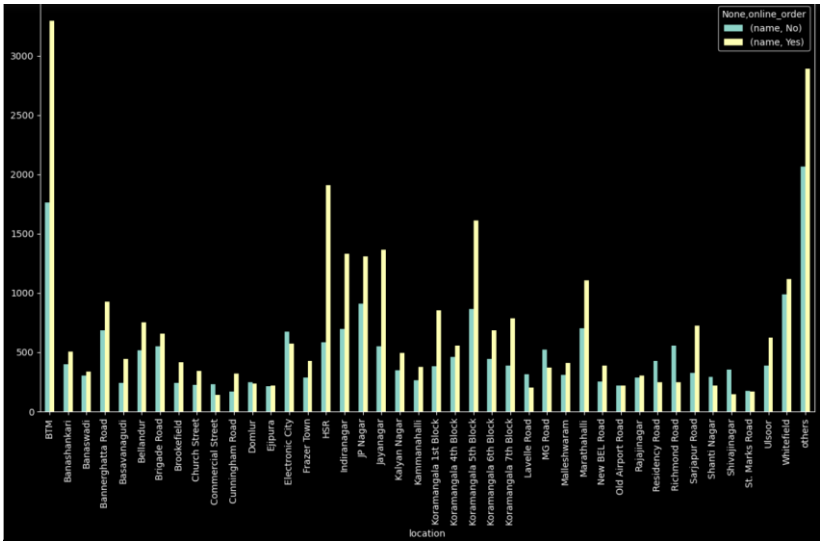


Fig 6. Location vs Online order of restaurants in a location

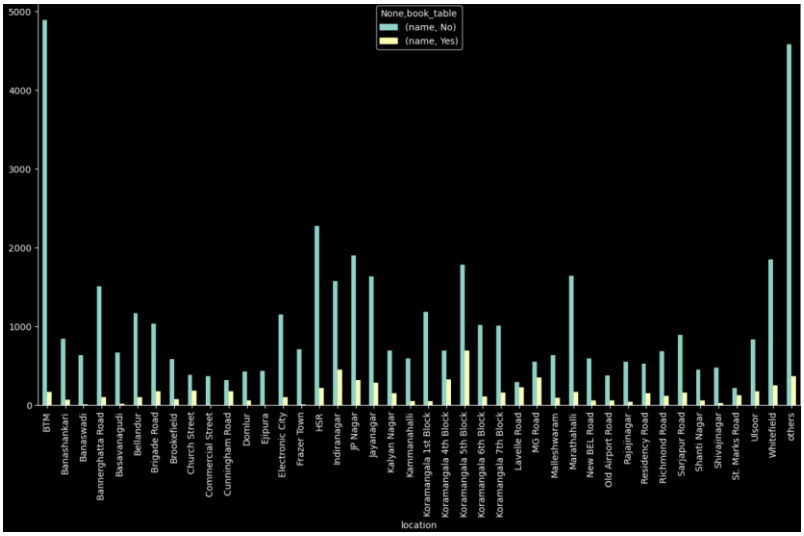


Fig 7. Location vs Book table

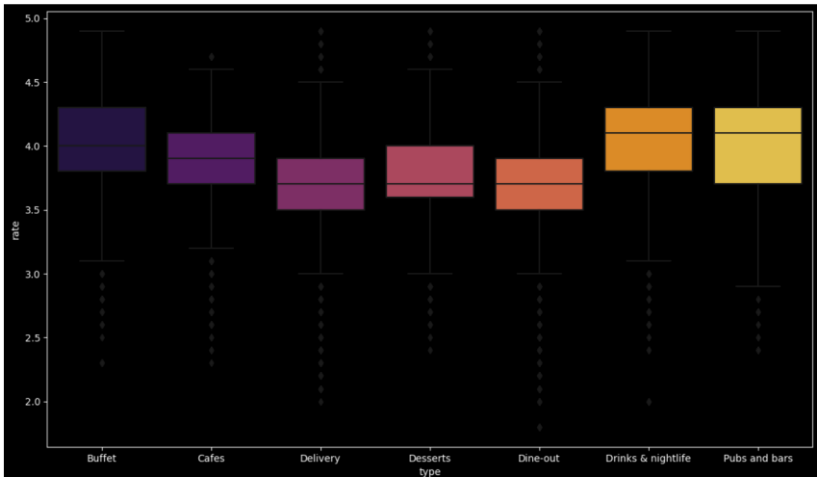


Fig 8. Type of restaurant vs Rate

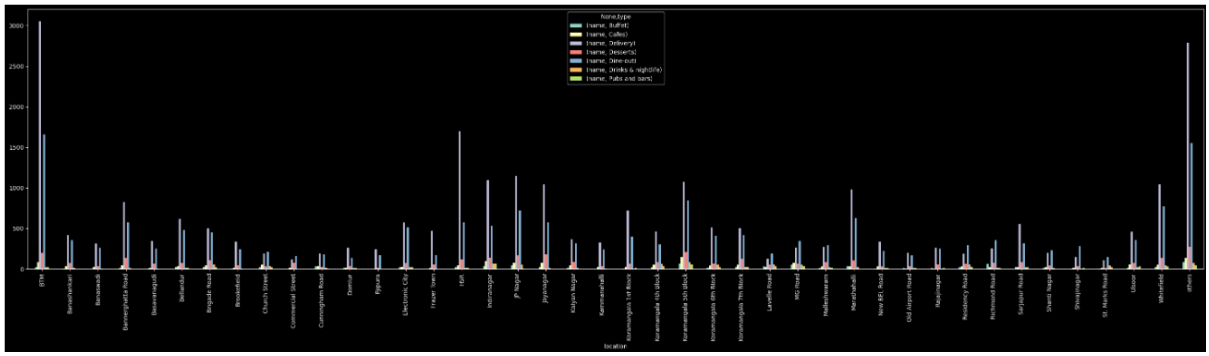


Fig 9. Type of restaurant vs Location

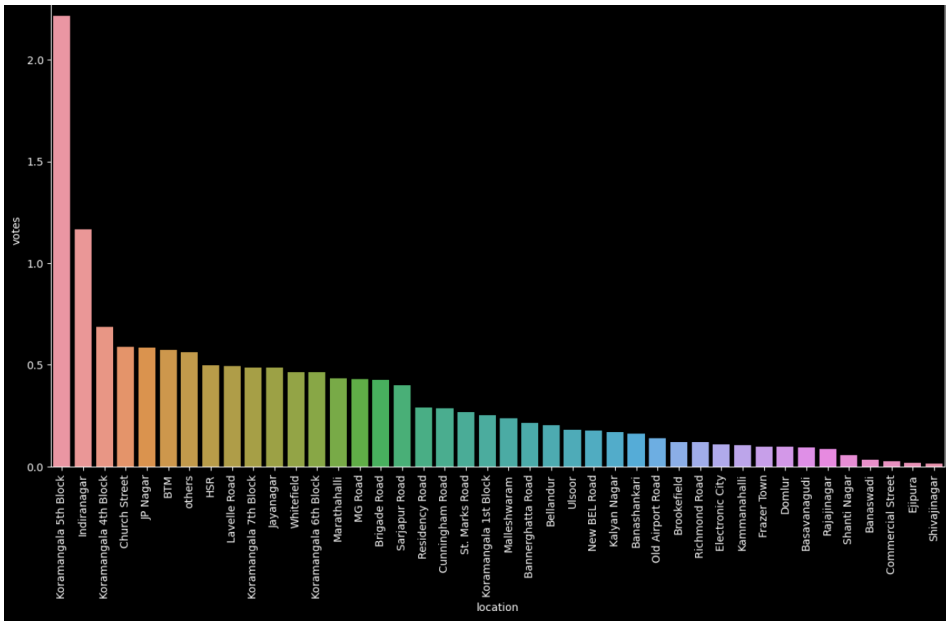


Fig 10. Number of votes vs Location

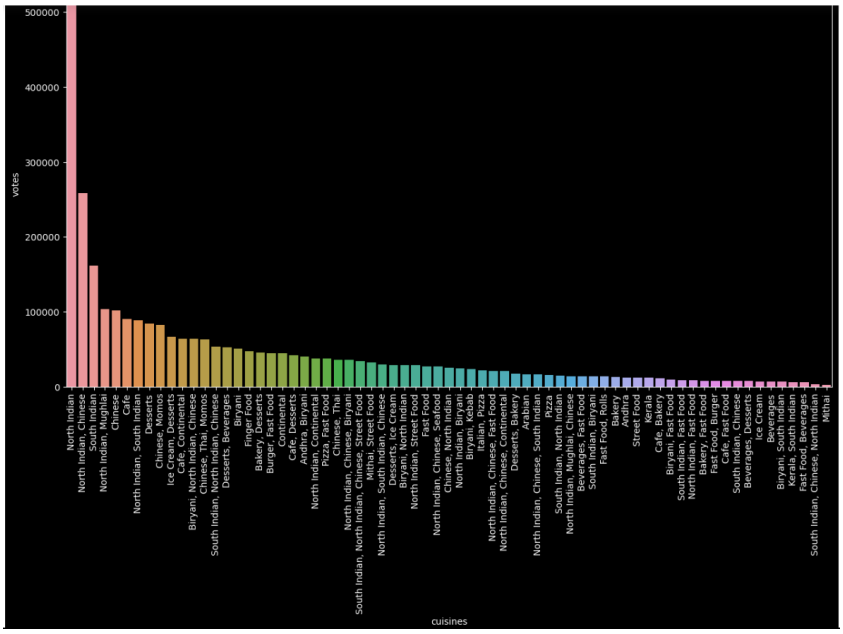
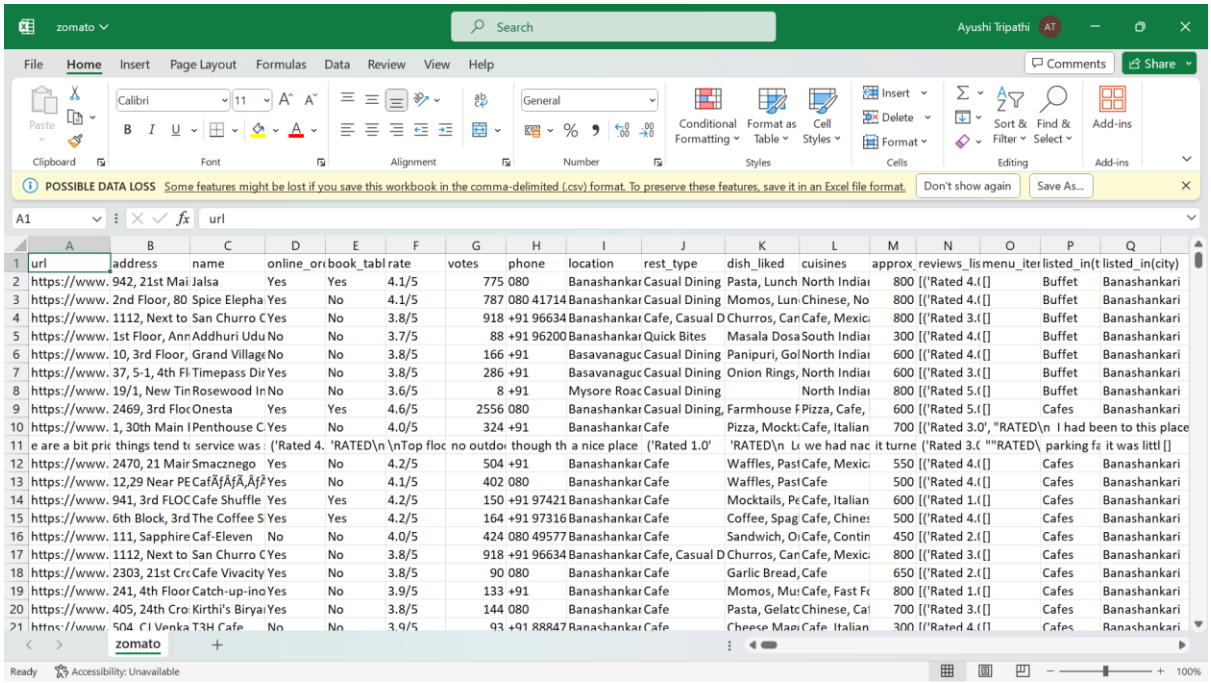


Fig 11. Cuisines vs Votes

Table 1. First few records of the dataset



5. CONCLUSION:

In wrapping up this project, we've journeyed through the world of restaurant reviews and data analysis. By cleaning and processing the data using tools like Jupyter Notebook and powerful Python libraries, we unveiled meaningful insights. From understanding customer sentiments to identifying popular cuisines and using machine learning with SVM, we've equipped aspiring restaurant owners with valuable information. Visualizations added a touch of clarity to our findings. This project isn't just about numbers—it's about helping entrepreneurs make informed decisions when deciding where to open their restaurants. It's a step towards turning dreams of successful culinary ventures into reality.

6. REFERENCES:

1. SmartDraw. (2023). SmartDraw - Flowchart Maker. Retrieved December 1, 2023, from https://www.smartdraw.com/flowchart/flowchart-maker.htm

2. GeeksforGeeks. (2023). Support Vector Machine Algorithm. Retrieved December 1, 2023, from https://www.geeksforgeeks.org/support-vector-machine-algorithm/

3. GeeksforGeeks. (2023). Python Seaborn Tutorial. Retrieved December 1, 2023, from https://www.geeksforgeeks.org/python-seaborn-tutorial/

4. GeeksforGeeks. (2023). Matplotlib Tutorial. Retrieved December 1, 2023, from https://www.geeksforgeeks.org/matplotlib-tutorial/