TEACHNOOK

Major Project Name: Take any Dataset of your choice, perform EDA (Exploratory Data Analysis) and apply a suitable Classifier, Regressor or Clusterer and calculate the accuracy of the model.

Duration of the project: Two Weeks

Developed by Mr. Saguturu Kishan Sai

Computer Science Engineering,

Specialization: Artificial Intelligence and Machine Learning.

Team supported by:

- Mr. Abhi Deshmukh
- Mr. Ajith Reddy
- Mr. Arunram Ilayarasan
- Miss. Chandrika Agarwal
- Mr. Jashmeshsign Kocsher
- Mr. Joannsanders
- Mr. Mayankbhuse
- Mr. Madhav Panchar
- Mr. Rahul Biswas
- Mr. Sarashasware
- Mr. Susruteepogade

Project Content:

- Data Set
- Exploratory Data Analysis
- Classifier
- Regressor
- Clusterer
- Correlation Between Features
- Accuracy Of the Model

Abstract:

A dataset is a set or collection of data. This set is normally presented in a tabular pattern. Every column describes a particular variable. And each row corresponds to a given member of the data set, as per the given question. This is a part of data management.

What is Exploratory Data Analysis?

Exploratory Data Analysis or EDA is used to take insights from the data. Data Scientists and Analysts try to find different patterns, relations, and anomalies in the data using some statistical graphs and other visualization techniques. Following things are part of EDA:

1. Missing Values
2. Explore About the Numerical Variables
3. Explore About Categorical variables
4. Finding relationship Between Features

What is a **Classifier**? In data science, a classifier is a type of machine learning algorithm used to assign a class label to a data input. An example is an image recognition classifier to label an image (e.g., “car,” “truck,” or “person”).

Regression analysis is primarily used for two conceptually distinct purposes.

First, regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning.

Second, in some situation’s regression analysis can be used to infer causal relationships between the independent and dependent variables. Importantly, regressions by themselves only reveal relationships between a dependent variable and a collection of independent variables in a fixed dataset. To use regressions for prediction or to infer causal relationships, respectively, a researcher must carefully justify why existing relationships have predictive power for a new context or why a relationship between two variables has a causal interpretation. The latter is especially important when researchers hope to estimate causal relationships using observational data.

In machine learning too, we often group examples as a first step to understand a subject (data set) in a machine learning system. Grouping unlabelled is called **clustering**.

As the examples are unlabelled, clustering relies on unsupervised machine learning. If the examples are labelled, then clustering becomes classification. For a more detailed discussion of supervised and unsupervised methods see Introduction to Machine Learning Problem Framing.

Correlation Between Features:

- First thing before modelling is always to remove correlated factors
- That will make your model more accurate and
- fits well Convert your dataset into pandas dataframe and use Corr ()

Accuracy Of the Model:

$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN}$$

If we consider a binary classification problem,
C_{00} represents the count of true negative
C_{01} represents the count of false positive
C_{10} represents the count of false negative and
C_{11} represents the count of true positive.

Input and Output of Codes:

Data Set Link:

https://docs.google.com/spreadsheets/d/1624JXPuE8DTAosrXinm4e_OFAzTZLT5FXIWyMrNPAGQ/edit?usp=sharing

EDA (Exploratory Data Analysis):

Missing Values 1:

```
In [8]: df.isnull().sum()

Out[8]: Restaurant ID         0
        Restaurant Name       0
        Country Code          0
        City                  0
        Address               0
        Locality              0
        Locality Verbose      0
        Longitude             0
        Latitude              0
        Cuisines              9
        Average Cost for two  0
```

Missing Values 2:

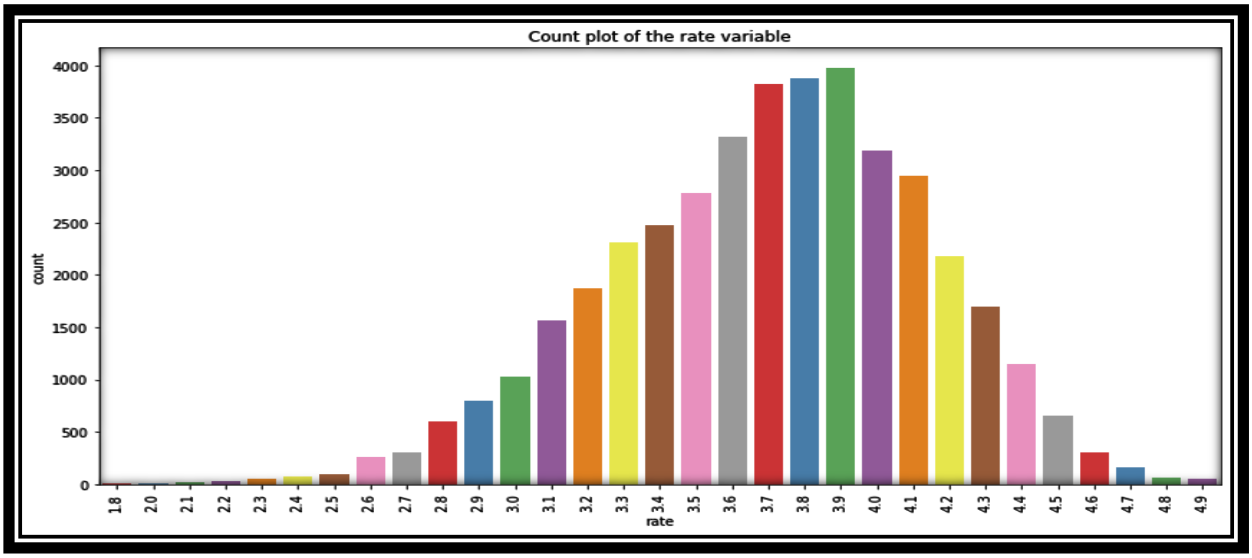
```
Currency              0
Has Table booking     0
Has Online delivery   0
Is delivering now     0
Switch to order menu  0
Price range           0
Aggregate rating      0
Rating color          0
Rating text           0
Votes                 0
dtype: int64
```

2. Explore About the Numerical Variables:

```
In [11]: df.shape
Out[11]: (9551, 21)
```

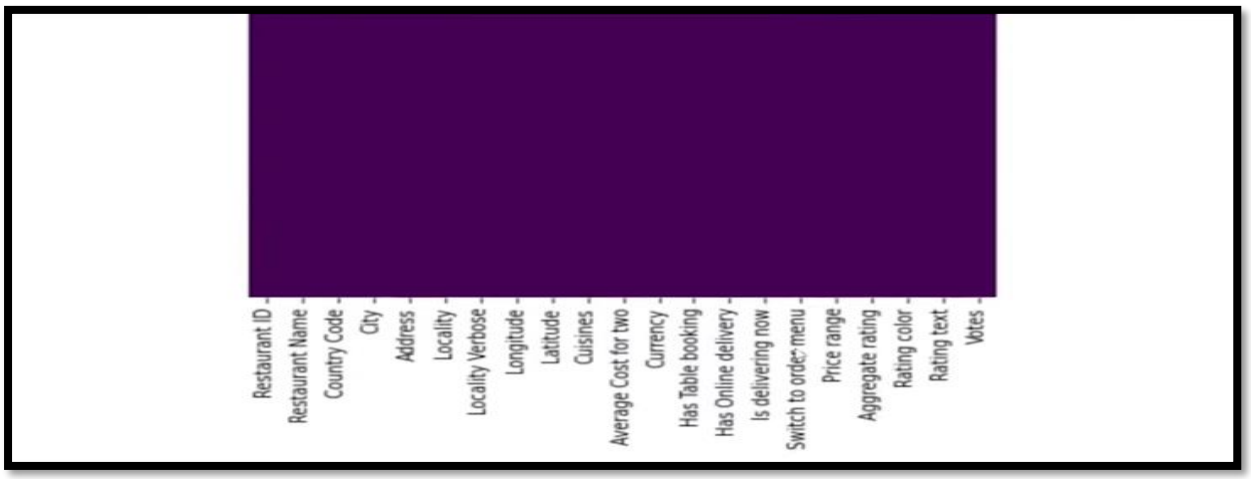
3. Explore About Categorical variables

```
plt.rcParams['figure.figsize'] = 14,7
sns.countplot(df['rate'], palette='Set1')
plt.title("Count plot of the rate variable")
plt.xticks(rotation = 90)
plt.show()
```



Heat Map Analysis:

```
In [10]: sns.heatmap(df.isnull(),yticklabels=False,cbar=False,cmap='viridis')
Out[10]: <AxesSubplot:>
```



4.Finding relationship Between Features:

```
In [9]: [features for features in df.columns if df[features].isnull().sum()>0]
Out[9]: ['Cuisines']
```

Classifier:

```
#!/usr/bin/env python
# coding: utf-8
# In[87]:
import pandas as pd
import numpy as np
import seaborn as sns
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings("ignore")
# In[88]:
df = pd.read_csv(r".....\zomato.csv")
df
# In[89]:
df=df.drop(['Restaurant Name'],axis=1)
df=df.drop(['City'],axis=1)
df=df.drop(['Address'],axis=1)
df=df.drop(['Locality'],axis=1)
df=df.drop(['Locality Verbose'],axis=1)
df=df.drop(['Cuisines'],axis=1)
df=df.drop(['Currency'],axis=1)
df=df.drop(['Has Table booking'],axis=1)
df=df.drop(['Has Online delivery'],axis=1)
df=df.drop(['Is delivering now'],axis=1)
df=df.drop(['Switch to order menu'],axis=1)
df=df.drop(['Rating color'],axis=1)
df=df.drop(['Rating text'],axis=1)
# In[90]:
df.head()
# In[91]:
df.tail()
# In[92]:
df.describe()
# In[93]:
df.info
# In[94]:
df.isna().sum()
# In[95]:
df.hist()
```

```
# In[96]:
df.shape
# # Regression
# In[97]:
X=df.iloc[:,:-1]
y=df.iloc[:,-1]
# In[98]:
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,
random_state=50)
# In[99]:
from sklearn.metrics import mean_absolute_error,r2_score,mean_squared_error
from sklearn.linear_model import LinearRegression
model = LinearRegression()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
print("R^2 : ", r2_score(y_test, y_pred))
print("MAE :", mean_absolute_error(y_test,y_pred))
print("RMSE:",np.sqrt(mean_squared_error(y_test, y_pred)))
# # Classification
# In[100]:
X=df.iloc[:,:-1]
y=df.iloc[:,-1]
# In[101]:
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,
random_state=50)
# In[102]:
from sklearn.tree import DecisionTreeClassifier
from sklearn import tree
# In[103]:
dtree=DecisionTreeClassifier(max_depth=2)
# In[104]:
dtree.fit(X_train,y_train)
# In[117]:
y_pred = dtree.predict(X_test)
from sklearn.metrics import accuracy_score
accuracy_score(y_pred,y_test)
# In[ ]:
```

Regression:

Applying Linear Regression Algorithm

[59]:
```
from sklearn.linear_model import LinearRegression

lr = LinearRegression()

lr.fit(x_train, y_train)

lr_pred = lr.predict(x_test)
```

[60]:
```
r2 = r2_score(y_test,lr_pred)
print('R-Square Score: ',r2*100)
```

```
R-Square Score:  32.866970444722085
```

[61]:
```
# Calculate the absolute errors
lr_errors = abs(lr_pred - y_test)
# Print out the mean absolute error (mae)
print('Mean Absolute Error:', round(np.mean(lr_pred), 2), 'degrees.')
```

```
Mean Absolute Error: 3.7 degrees.
```

[62]:
```
# Calculate mean absolute percentage error (MAPE)
mape = 100 * (lr_errors / y_test)
# Calculate and display accuracy
lr_accuracy = 100 - np.mean(mape)
print('Accuracy for Logistic Regression is :', round(lr_accuracy, 2), '%.')
```

```
Accuracy for Logistic Regression is : 92.95 %.
```

Clusterer:

Results – Analysis of Cluster

```
locality_merged.loc[locality_merged['Cluster Labels'] == 0]
```

	venue	latitude	longitude	locality	price_for_two	price_range	rating	Cluster Labels
1	Shiro	12.971758	77.595922	UB City	3000	4	4.4	0
11	Cafe Noir	12.972126	77.596441	UB City	1500	3	4.2	0
12	Skyye	12.971632	77.596371	UB City	2500	4	4.3	0
14	McDonald's	12.976243	77.598372	MG Road	500	2	3.8	0
15	Rim Naam - The Oberoi	12.972776	77.618641	MG Road	3000	4	4.6	0
17	Toast & Tonic	12.966665	77.608927	Richmond	2000	4	4.6	0

Cluster 0

```
locality_merged.loc[locality_merged['Cluster Labels'] == 1]
```

	venue	latitude	longitude	locality	price_for_two	price_range	rating	Cluster Labels
0	ROYCE' Chocolate	12.972469	77.595103	Lavelle Road	1000	3	3.5	1
2	Mathsya Darshini	12.975296	77.588858	Lavelle Road	350	1	3.4	1
3	Truffles	12.971769	77.601137	St. Marks Road	900	2	4.4	1
4	Smoke House Deli	12.971659	77.598318	Lavelle Road	1600	3	4.7	1
5	Hard Rock Cafe	12.976034	77.601567	St. Marks Road	2500	4	4.5	1
6	Corner House Ice Cream	12.973186	77.599967	Lavelle Road	350	1	4.4	1
8	Harima	12.967536	77.599901	Residency Road	2000	4	4.3	1

Cluster 1

```
locality_merged.loc[locality_merged['Cluster Labels'] == 2]
```

	venue	latitude	longitude	locality	price_for_two	price_range	rating	Cluster Labels
18	Brahmin's Coffee Bar	12.954032	77.568948	Basavanagudi	100	1	4.8	2
26	Hari Super Sandwich	12.932848	77.582555	Jayanagar	200	1	4.4	2
30	CTR	12.998270	77.569455	Malleshwaram	150	1	4.7	2
38	Upahara Darshini	12.939350	77.571491	Basavanagudi	150	1	4.2	2
39	Fishland	12.975600	77.578557	Majestic	500	2	4.1	2
42	S R & Sons Bakery	12.983502	77.606561	Commercial Street	100	1	3.5	2
45	Mavalli Tiffin Room (MTR)	12.955176	77.585622	Basavanagudi	250	1	4.5	2
47	Cookie Man	13.011356	77.555020	Malleshwaram	150	1	3.6	2

Cluster 2

```
locality_merged.loc[locality_merged['Cluster Labels'] == 3]
```

	venue	latitude	longitude	locality	price_for_two	price_range	rating	Cluster Labels
7	Masala Klub - The Taj West End	12.984113	77.583968	Race Course Road	4000	4	4.4	3
9	Edo Restaurant & Bar - ITC Gardenia	12.967392	77.596392	ITC Gardenia	4000	4	4.3	3
34	Dum Pukht Jolly Nabobs - ITC Windsor	12.994669	77.585355	ITC Windsor	5000	4	4.3	3
56	Cubbon Pavilion - ITC Gardenia	12.967401	77.596393	ITC Gardenia	2500	4	4.3	3
60	Blue Bar - The Taj West End	12.984111	77.583966	Race Course Road	2500	4	4.0	3
126	The Raj Pavilion - ITC Windsor	12.994645	77.585229	ITC Windsor	2000	4	4.2	3

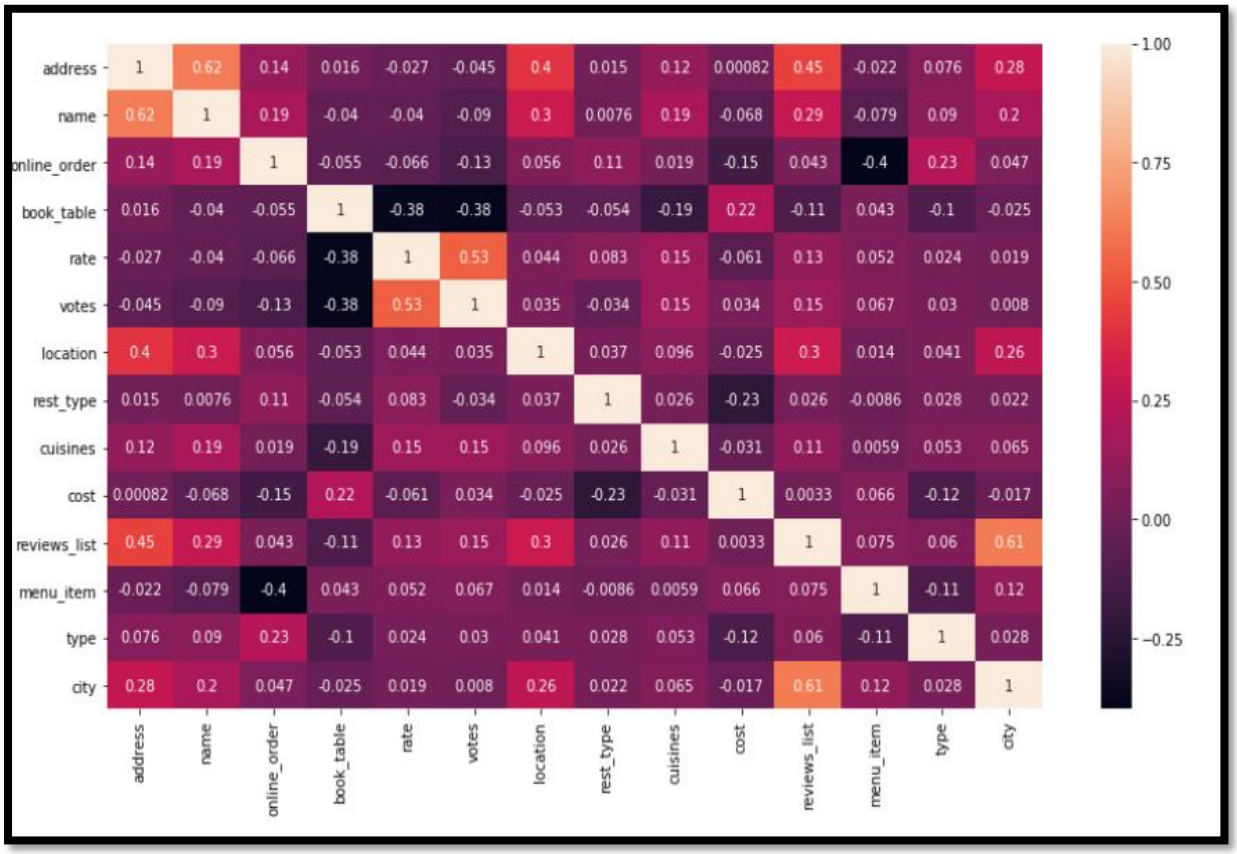
Cluster 3

Correlation Between Feature: Data frame and using use corr()

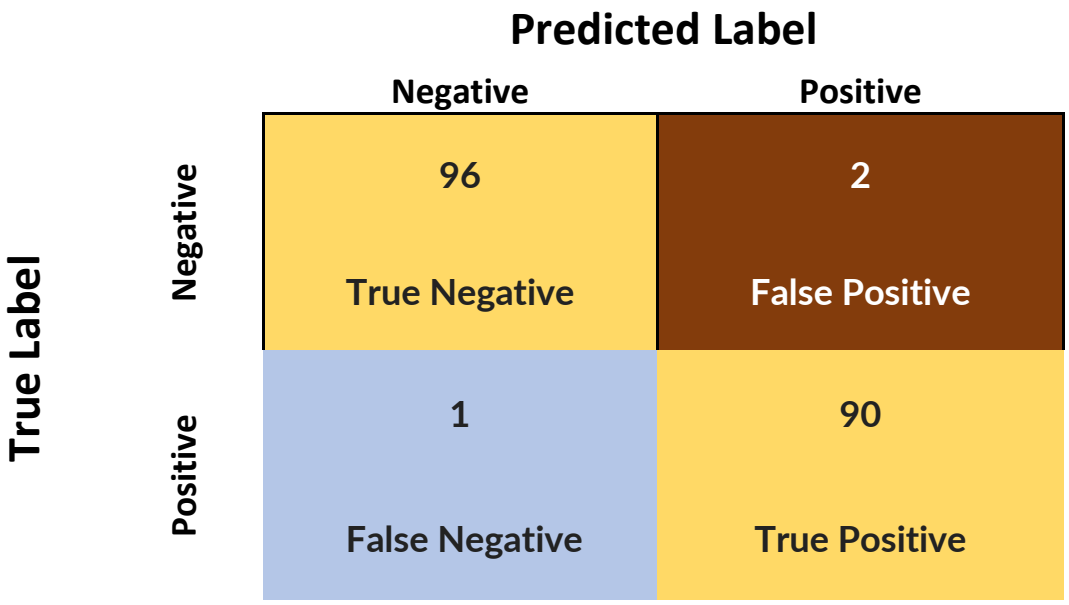
Get Correlation between different variables

In [11]:
```
#Get Correlation between different variables
corr = zomato_en.corr(method='kendall')
plt.figure(figsize=(15,8))
sns.heatmap(corr, annot=True)
zomato_en.columns
```

Out[11]:
```
Index(['address', 'name', 'online_order', 'book_table', 'r
ate', 'votes',
       'location', 'rest_type', 'cuisines', 'cost', 'revie
ws_list',
       'menu_item', 'type', 'city'],
      dtype='object')
```



Accuracy Of the Model:



Accuracy = 96+90 / 96+2+90+1

= 0.9841 x 100

= 98.41%

Conclusion:

With this project in python, I have successfully developed given "dataset" tasks are executed with given problem statement.