EDA ON ZOMATO SALES DATASET

Importing libraries and reading data files

Import libraries

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

To Ignore Warning

```
from warnings import filterwarnings
filterwarnings('ignore')
```

Reading datasets

```
df = pd.read_csv('zomato_dataset.csv')
```

Restaurants dataset preprocessing

Basic EDA

```
# Display top 5 records
df.head()
```

```
  Restaurant Name  Dining Rating  Delivery Rating  Dining Votes  \
0      Doner King           3.9              4.2            39
1      Doner King           3.9              4.2            39
2      Doner King           3.9              4.2            39
3      Doner King           3.9              4.2            39
4      Doner King           3.9              4.2            39

   Delivery Votes   Cuisine  Place Name        City                 Item Name
\
0               0  Fast Food   Malakpet   Hyderabad       Platter Kebab Combo
1               0  Fast Food   Malakpet   Hyderabad   Chicken Rumali Shawarma
2               0  Fast Food   Malakpet   Hyderabad    Chicken Tandoori Salad
3               0  Fast Food   Malakpet   Hyderabad         Chicken BBQ Salad
4               0  Fast Food   Malakpet   Hyderabad  Special Doner Wrap Combo

  Best Seller  Votes  Prices
0  BESTSELLER     84   249.0
1  BESTSELLER     45   129.0
2         NaN     39   189.0
3  BESTSELLER     43   189.0
4    MUST TRY     31   205.0
```

```
# Display last 5 records
df.tail()
```

```
             Restaurant Name  Dining Rating  Delivery Rating  Dining Votes
\
123652  Ariena Boutique Hotel           3.9              4.2            13
123653  Ariena Boutique Hotel           3.9              4.2            13
123654  Ariena Boutique Hotel           3.9              4.2            13
123655  Ariena Boutique Hotel           3.9              4.2            13
123656  Ariena Boutique Hotel           3.9              4.2            13

        Delivery Votes Cuisine  Place Name     City              Item Name  \
123652             523    Pizza     Purena   Raipur     Murgh Reshmi Kebab
123653             523    Pizza     Purena   Raipur      Murgh Large Tikka
123654             523    Pizza     Purena   Raipur  Murgh Chukandri Tikka
123655             523    Pizza     Purena   Raipur     Murgh Golden Kebab
123656             523    Pizza     Purena   Raipur     Gosht Gilawat Chop

       Best Seller  Votes  Prices
123652         NaN      0   525.0
123653         NaN      0   525.0
```

```
123654         NaN      0   525.0
123655         NaN      0   525.0
123656  BESTSELLER      0   595.0
```

```
# Get the total number of records & attributes
print("Shape of Data is:",df.shape)
```

```
Shape of Data is: (123657, 12)
```

```
# Getting statistical info about dataset
df.describe()
```

```
       Dining Rating  Delivery Rating   Dining Votes  Delivery Votes  \
count   91421.000000    122377.000000  123657.000000   123657.000000
mean        3.822264         3.963184     152.729858      115.763725
std         0.408693         0.245900     232.214061      243.970828
min         2.500000         2.500000       0.000000        0.000000
25%         3.600000         3.800000       0.000000        0.000000
50%         3.900000         4.000000      30.000000        0.000000
75%         4.100000         4.100000     217.000000       23.000000
max         4.800000         4.600000     997.000000      983.000000

               Votes         Prices
count  123657.000000  123657.000000
mean       24.666772     241.378399
std       125.236009     192.830713
min         0.000000       0.950000
25%         0.000000     130.000000
50%         0.000000     208.570000
75%        15.000000     299.000000
max      9750.000000   12024.000000
```

```
#correlation value between 1 and -1
df.corr()
```

```
                Dining Rating  Delivery Rating  Dining Votes  Delivery Votes
\
Dining Rating         1.000000         0.311651      0.229514       -0.138843
Delivery Rating       0.311651         1.000000      0.132089       -0.065398
Dining Votes          0.229514         0.132089      1.000000       -0.244525
Delivery Votes       -0.138843        -0.065398     -0.244525        1.000000
Votes                 0.040707         0.049216      0.007271       -0.063766
Prices                0.074197         0.054198      0.016136        0.007060

                   Votes    Prices
Dining Rating    0.040707  0.074197
Delivery Rating  0.049216  0.054198
Dining Votes     0.007271  0.016136
Delivery Votes  -0.063766  0.007060
Votes            1.000000 -0.058036
Prices          -0.058036  1.000000
```

```
#columns names
df.columns
```

```
Index(['Restaurant Name', 'Dining Rating', 'Delivery Rating', 'Dining Votes',
       'Delivery Votes', 'Cuisine ', 'Place Name', 'City', 'Item Name',
       'Best Seller', 'Votes', 'Prices'],
      dtype='object')
```

```
# Getting more info about dataset
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 123657 entries, 0 to 123656
Data columns (total 12 columns):
 #   Column           Non-Null Count   Dtype  
---  ------           --------------   -----  
 0   Restaurant Name  123657 non-null  object 
 1   Dining Rating    91421 non-null   float64
 2   Delivery Rating  122377 non-null  float64
 3   Dining Votes     123657 non-null  int64  
 4   Delivery Votes   123657 non-null  int64  
 5   Cuisine          123657 non-null  object 
 6   Place Name       123657 non-null  object 
 7   City             123657 non-null  object 
 8   Item Name        123657 non-null  object 
 9   Best Seller      27942 non-null   object 
 10  Votes            123657 non-null  int64  
 11  Prices           123657 non-null  float64
dtypes: float64(3), int64(3), object(6)
memory usage: 11.3+ MB
```

```
df['Prices'].unique()
```

```
array([ 249. ,  129. ,  189. , ..., 1789. , 1139. ,   22.5])
```

Categorical Columns

- Restaurant Name
- Cuisine
- Place Name
- City

- Item Name
- Best Seller

Numerical Columns

- Dining Rating
- Delivery Rating
- Dining Votes

- Delivery Votes
- Votes
- Prices

How much missing values there are in these dataset?

Advance EDA

```
# Check for Null Values
df.isnull().sum()
```

```
Restaurant Name        0
Dining Rating      32236
Delivery Rating     1280
Dining Votes           0
Delivery Votes         0
Cuisine                0
Place Name             0
City                   0
Item Name              0
Best Seller        95715
Votes                  0
Prices                 0
dtype: int64
```

```
# Impute Missing Values for Age
df['Votes'].isnull().sum()
```

```
0
```

DataCleaning

```
#  Drop the Prices column
df.drop('Prices', axis=1,inplace=True)

#To rename a column from bag_weight to bag_weight\kg
df1=df.rename(
columns={
'Best Seller': 'Best Seller/food'
})

df.head(2)
```

```
  Restaurant Name  Dining Rating  Delivery Rating  Dining Votes  \
0      Doner King           3.9              4.2            39
1      Doner King           3.9              4.2            39

   Delivery Votes   Cuisine  Place Name        City                Item Name
\
0               0  Fast Food   Malakpet   Hyderabad      Platter Kebab Combo
1               0  Fast Food   Malakpet   Hyderabad  Chicken Rumali Shawarma
```

```
  Best Seller  Votes
0  BESTSELLER     84
1  BESTSELLER     45
```

```
# To check whether the given rows have duplicate values
df.duplicated()
```

```
0         False
1         False
2         False
3         False
4         False
          ...
123652     True
123653     True
123654     True
123655     True
123656     True
Length: 123657, dtype: bool
```

```
df.duplicated().sum()
```

```
22284
```

Observation

- I have performed a detailed analysis on Indian Restaurants Dataset from Zomato.
- Used as a manual to perform basic to intermediate EDA on any dataset.
- Importing.
- Preprocessing.
- Exploring data.
- Removing duplicates.
- Dealing with missing values.
- There are no Duplicate Values.
- The above code shows that there are some null values in the data.
- Shows the total rows, name and number of columns and their datatypes
- There are 123657 rows in all column and there are no missing data in numeric columns.

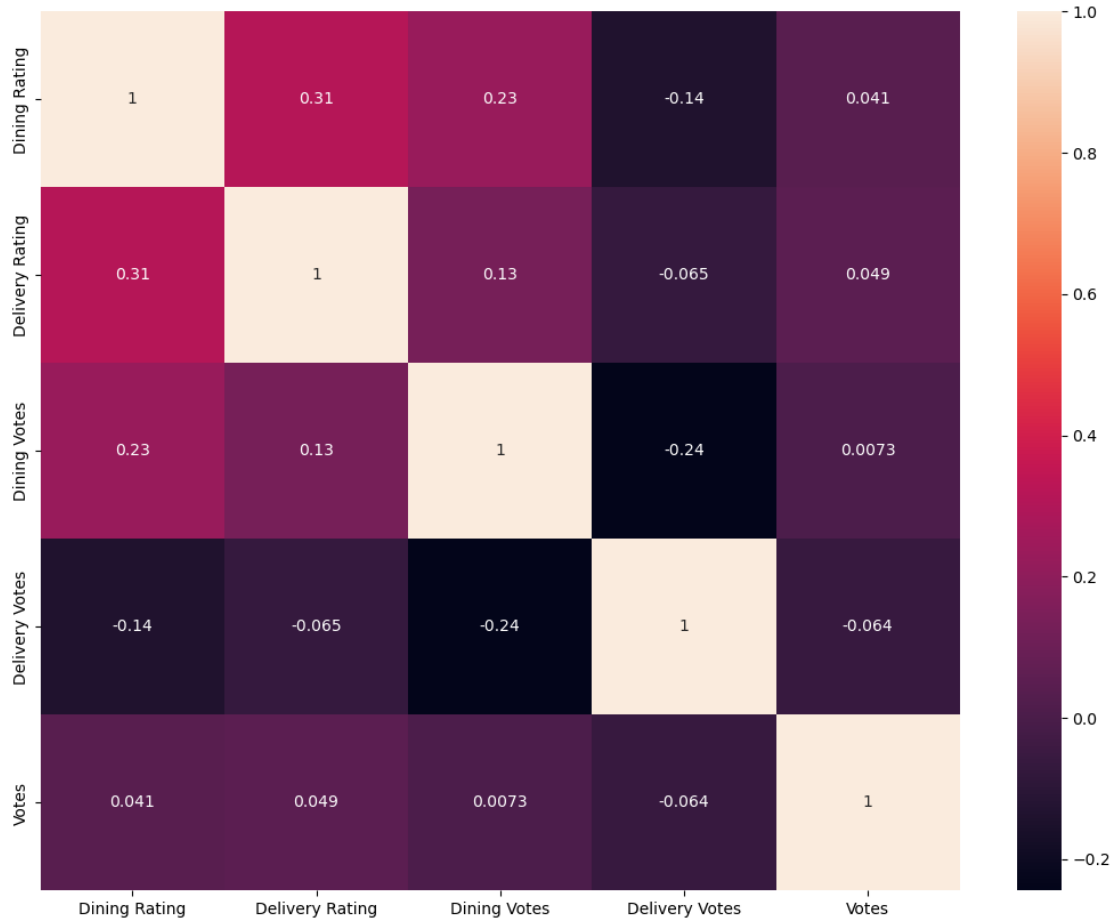
Seaborn library

Visualization of various themes/types of the graphs of datasets

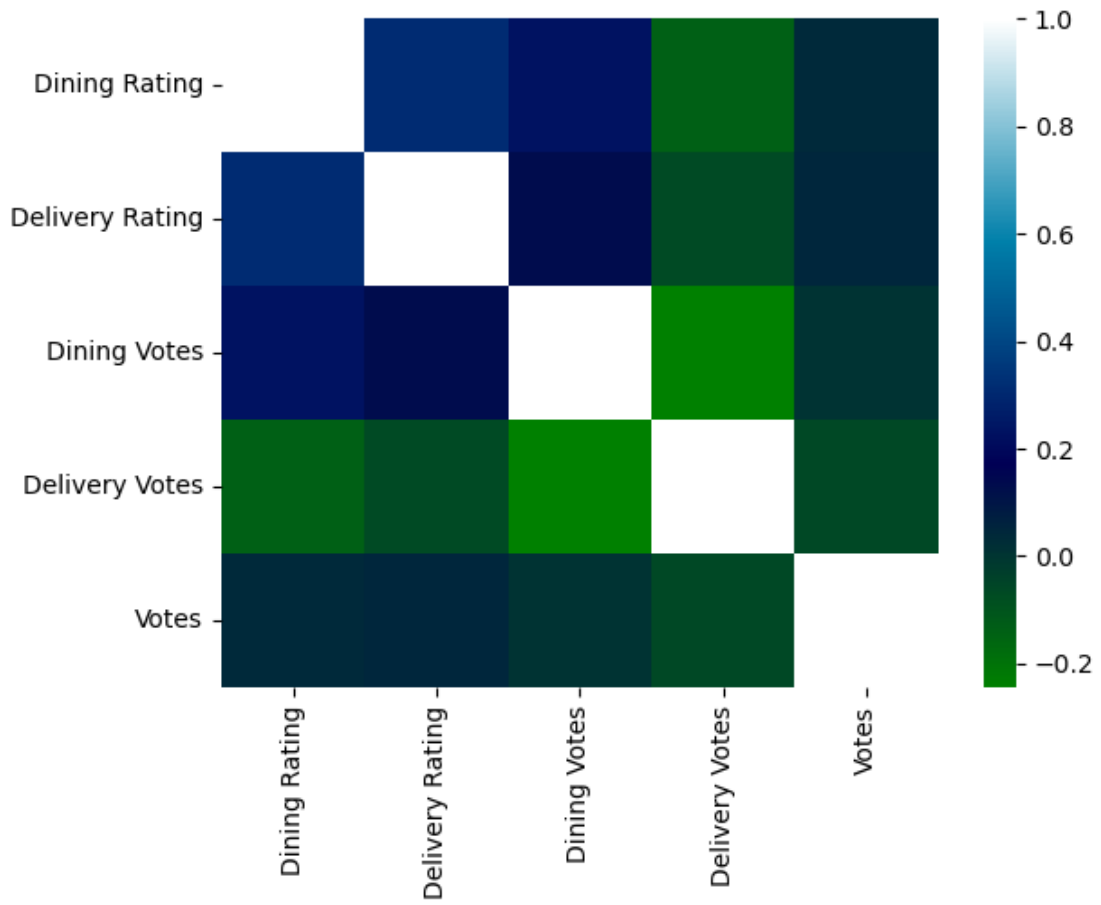
Heatmap for two conditions

- first condition true
- second condition false

```
#heatmap of true condition
plt.figure(figsize=(13,10))
sns.heatmap(df.corr(), annot=True);
```



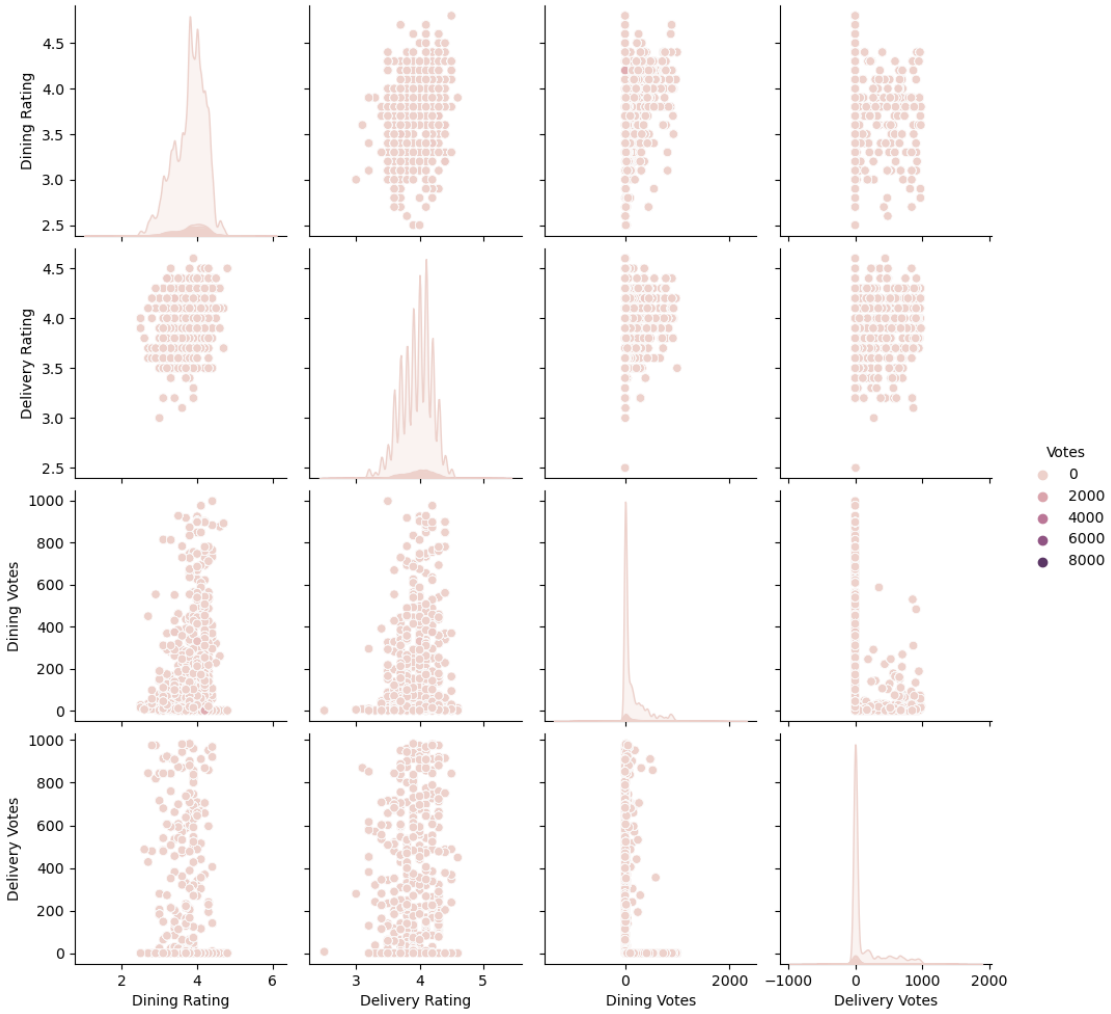

```
#heatmap of false condition
sns.heatmap(df.corr(), annot=False,cmap='ocean');
```



pairplot of datasets

```
#pair plot
sns.pairplot(df[['Dining Rating','Delivery Rating','Dining Votes','Delivery Votes','Votes']],hue='Votes')
```

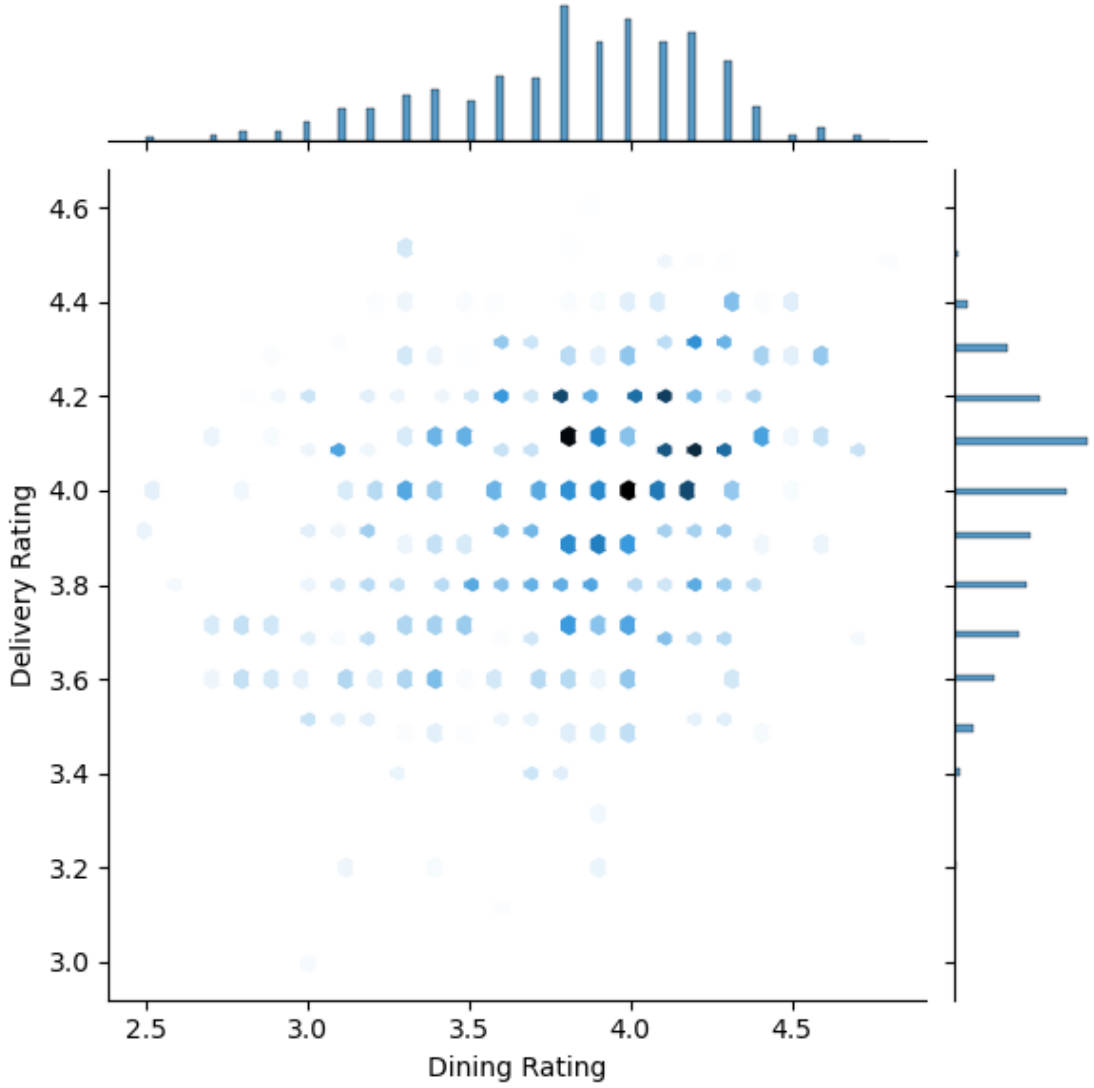
```
<seaborn.axisgrid.PairGrid at 0x21096f09850>
```



jointplot of datasets two columns

```
#Generate Joinplot for Dining Rating vs Delivery Rating.
sns.jointplot(x='Dining Rating',y='Delivery Rating',data=df,kind='hex')
#plot_kinds = ["scatter", "hist", "hex", "kde", "reg", "resid"]
```

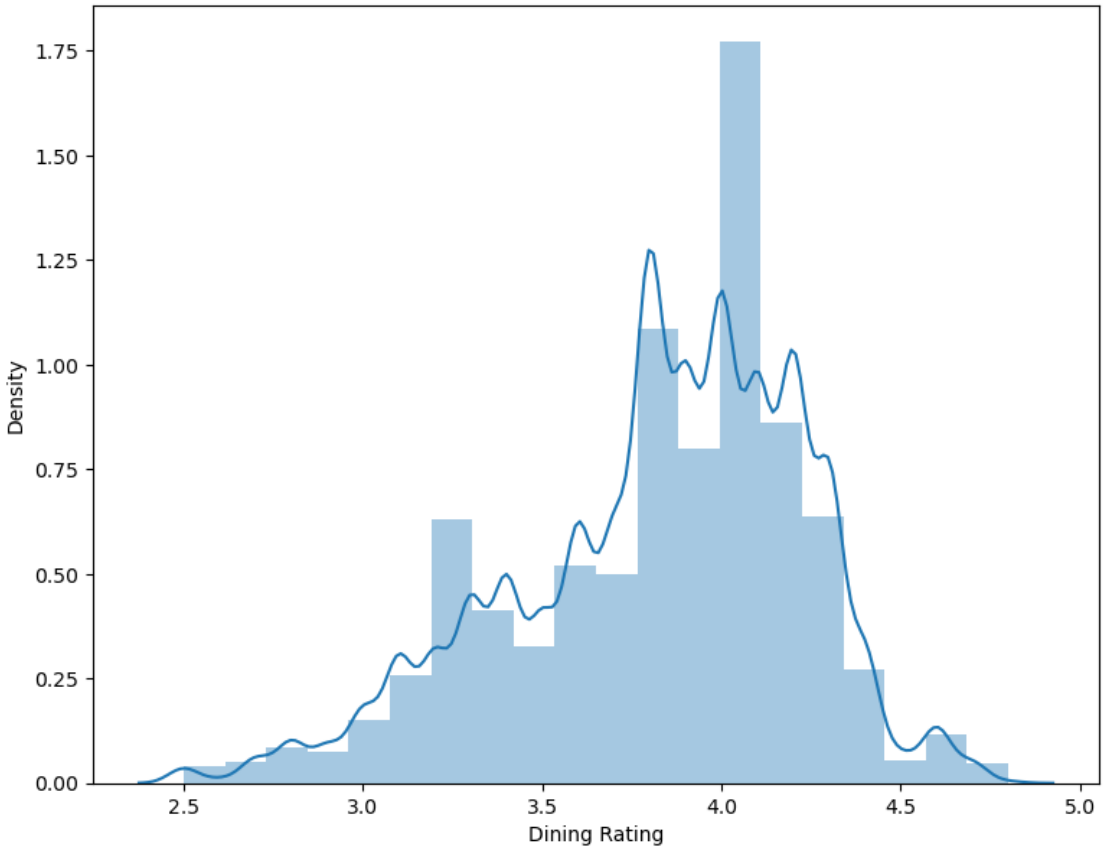
```
<seaborn.axisgrid.JointGrid at 0x21099721210>
```



Rating Distributions

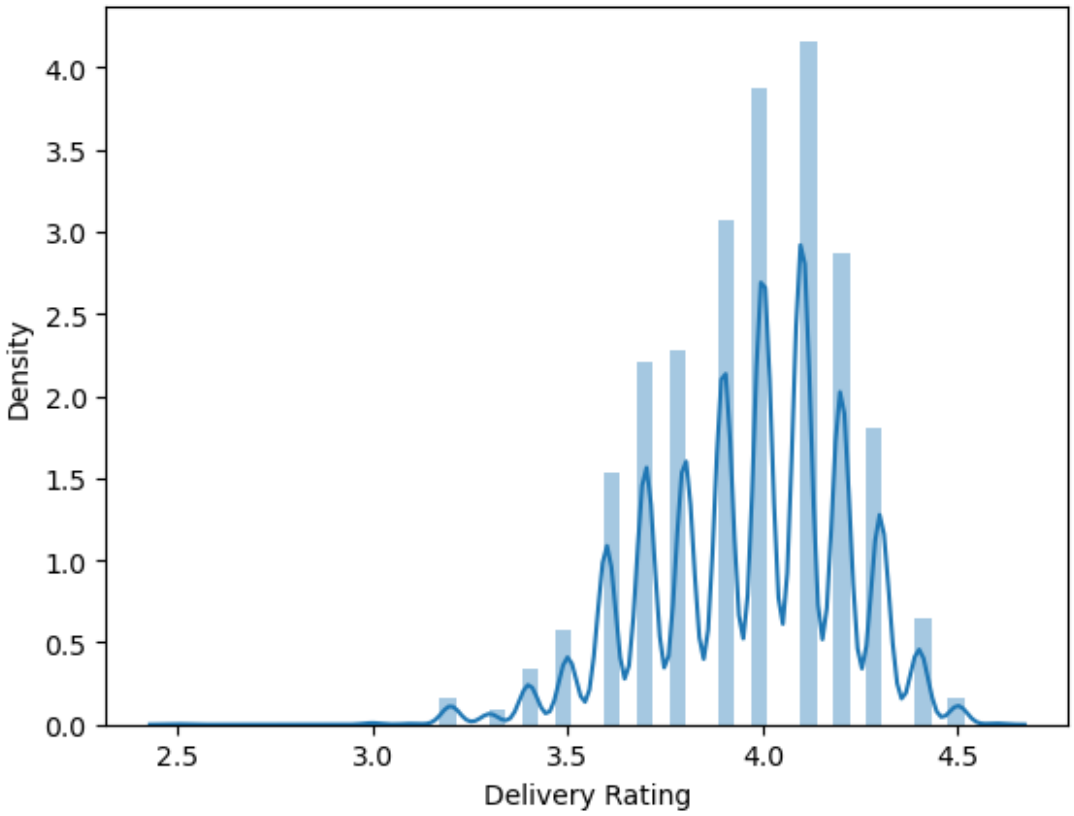
```
#How ratings are distributed
plt.figure(figsize=(9,7))

sns.distplot(df['Dining Rating'],bins=20)
```

```
<Axes: xlabel='Dining Rating', ylabel='Density'>
```




```
#distribution plot
sns.distplot(df['Delivery Rating'])
```

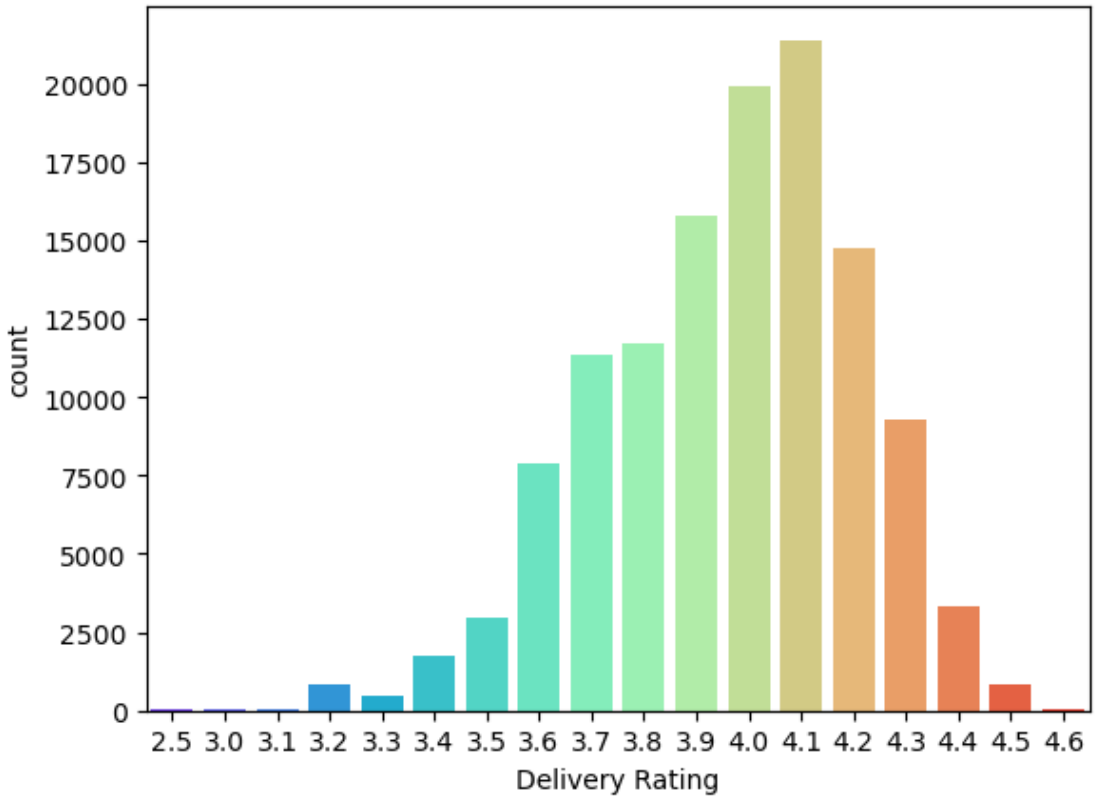
```
<Axes: xlabel='Delivery Rating', ylabel='Density'>
```



Services rating

```
#count plot
sns.countplot(x=df['Delivery Rating'],palette='rainbow')
```

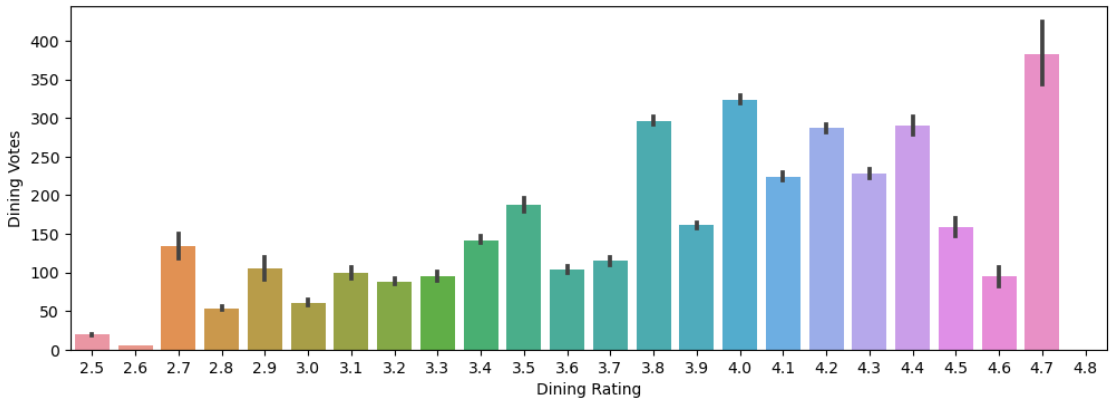
```
<Axes: xlabel='Delivery Rating', ylabel='count'>
```



Multivariate Analysis

```
#bar plot
import matplotlib.pyplot as plt
plt.figure(figsize=(12,4))

sns.barplot(x='Dining Rating',y='Dining Votes',data=df)
```

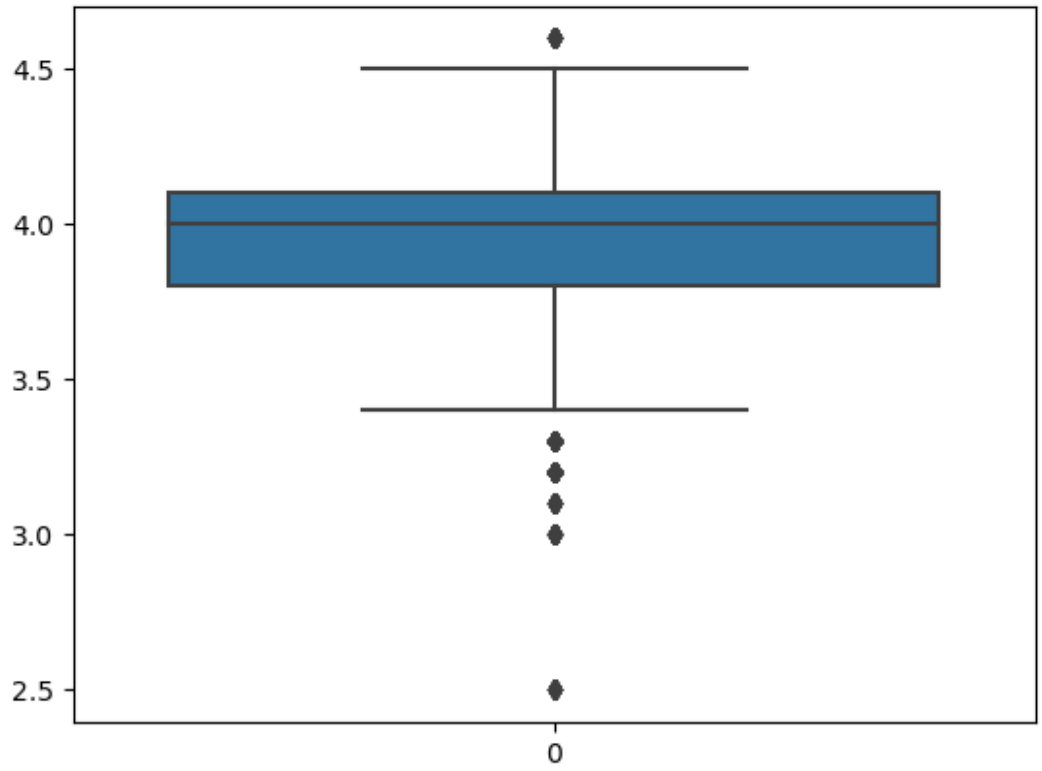
```
<Axes: xlabel='Dining Rating', ylabel='Dining Votes'>
```



Boxplot

```
# Boxplot for Delivery Rating
sns.boxplot(df['Delivery Rating'])
```

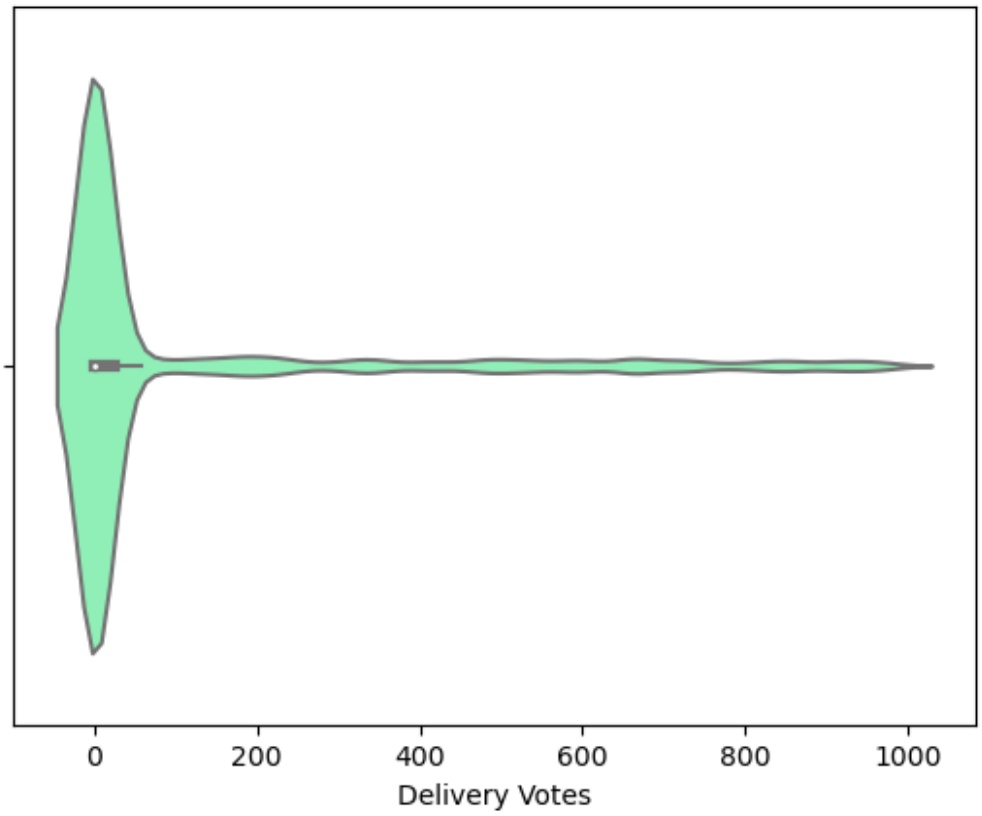
```
<Axes: >
```



Violin plot of Delivery Votes

```
#violin plot
sns.violinplot(x=df['Delivery Votes'],palette='rainbow')
# shift tab
```

```
<Axes: xlabel='Delivery Votes'>
```



Observation

- To use of seaborn libarary we analysis or plot the various themes/type of graph of the datasets.
- we find
- heatmap for two condition (true/false)
- pairplot of datasets
- Rating Distributions of distribution plot
- Services rating of countplot
- Barplot is Multivariate Analysis
- Boxplot of Delivery Votes
- Violin plot of Delivery Votes

Pandas profiling.

```
! pip install ydata-profiling
```

```
Requirement already satisfied: ydata-profiling in
c:\users\acer\anaconda3\lib\site-packages (4.3.1)
Requirement already satisfied: scipy<1.11,>=1.4.1 in
c:\users\acer\anaconda3\lib\site-packages (from ydata-profiling) (1.10.1)
Requirement already satisfied: pandas!=1.4.0,<2.1,>1.1 in
c:\users\acer\anaconda3\lib\site-packages (from ydata-profiling) (1.5.3)
```

```
Requirement already satisfied: matplotlib<4,>=3.2 in
c:\users\acer\anaconda3\lib\site-packages (from ydata-profiling) (3.7.1)
Requirement already satisfied: pydantic<2,>=1.8.1 in
c:\users\acer\anaconda3\lib\site-packages (from ydata-profiling) (1.10.11)
Requirement already satisfied: PyYAML<6.1,>=5.0.0 in
c:\users\acer\anaconda3\lib\site-packages (from ydata-profiling) (6.0)
Requirement already satisfied: jinja2<3.2,>=2.11.1 in
c:\users\acer\anaconda3\lib\site-packages (from ydata-profiling) (3.1.2)
Requirement already satisfied: visions[type_image_path]==0.7.5 in
c:\users\acer\anaconda3\lib\site-packages (from ydata-profiling) (0.7.5)
Requirement already satisfied: numpy<1.24,>=1.16.0 in
c:\users\acer\anaconda3\lib\site-packages (from ydata-profiling) (1.23.5)
Requirement already satisfied: htmlmin==0.1.12 in
c:\users\acer\anaconda3\lib\site-packages (from ydata-profiling) (0.1.12)
Requirement already satisfied: phik<0.13,>=0.11.1 in
c:\users\acer\anaconda3\lib\site-packages (from ydata-profiling) (0.12.3)
Requirement already satisfied: requests<3,>=2.24.0 in
c:\users\acer\anaconda3\lib\site-packages (from ydata-profiling) (2.29.0)
Requirement already satisfied: tqdm<5,>=4.48.2 in
c:\users\acer\anaconda3\lib\site-packages (from ydata-profiling) (4.65.0)
Requirement already satisfied: seaborn<0.13,>=0.10.1 in
c:\users\acer\anaconda3\lib\site-packages (from ydata-profiling) (0.12.2)
Requirement already satisfied: multimethod<2,>=1.4 in
c:\users\acer\anaconda3\lib\site-packages (from ydata-profiling) (1.9.1)
Requirement already satisfied: statsmodels<1,>=0.13.2 in
c:\users\acer\anaconda3\lib\site-packages (from ydata-profiling) (0.13.5)
Requirement already satisfied: typeguard<3,>=2.13.2 in
c:\users\acer\anaconda3\lib\site-packages (from ydata-profiling) (2.13.3)
Requirement already satisfied: imagehash==4.3.1 in
c:\users\acer\anaconda3\lib\site-packages (from ydata-profiling) (4.3.1)
Requirement already satisfied: wordcloud>=1.9.1 in
c:\users\acer\anaconda3\lib\site-packages (from ydata-profiling) (1.9.2)
Requirement already satisfied: dacite>=1.8 in
c:\users\acer\anaconda3\lib\site-packages (from ydata-profiling) (1.8.1)
Requirement already satisfied: PyWavelets in
c:\users\acer\anaconda3\lib\site-packages (from imagehash==4.3.1->ydata-
profiling) (1.4.1)
Requirement already satisfied: pillow in c:\users\acer\anaconda3\lib\site-
packages (from imagehash==4.3.1->ydata-profiling) (9.4.0)
Requirement already satisfied: attrs>=19.3.0 in
c:\users\acer\anaconda3\lib\site-packages (from
visions[type_image_path]==0.7.5->ydata-profiling) (22.1.0)
Requirement already satisfied: networkx>=2.4 in
c:\users\acer\anaconda3\lib\site-packages (from
visions[type_image_path]==0.7.5->ydata-profiling) (2.8.4)
Requirement already satisfied: tangled-up-in-unicode>=0.0.4 in
c:\users\acer\anaconda3\lib\site-packages (from
visions[type_image_path]==0.7.5->ydata-profiling) (0.2.0)
Requirement already satisfied: MarkupSafe>=2.0 in
c:\users\acer\anaconda3\lib\site-packages (from jinja2<3.2,>=2.11.1->ydata-
```

```
profiling) (2.1.1)
Requirement already satisfied: contourpy>=1.0.1 in
c:\users\acer\anaconda3\lib\site-packages (from matplotlib<4,>=3.2->ydata-
profiling) (1.0.5)
Requirement already satisfied: cycler>=0.10 in
c:\users\acer\anaconda3\lib\site-packages (from matplotlib<4,>=3.2->ydata-
profiling) (0.11.0)
Requirement already satisfied: fonttools>=4.22.0 in
c:\users\acer\anaconda3\lib\site-packages (from matplotlib<4,>=3.2->ydata-
profiling) (4.25.0)
Requirement already satisfied: kiwisolver>=1.0.1 in
c:\users\acer\anaconda3\lib\site-packages (from matplotlib<4,>=3.2->ydata-
profiling) (1.4.4)
Requirement already satisfied: packaging>=20.0 in
c:\users\acer\anaconda3\lib\site-packages (from matplotlib<4,>=3.2->ydata-
profiling) (23.0)
Requirement already satisfied: pyparsing>=2.3.1 in
c:\users\acer\anaconda3\lib\site-packages (from matplotlib<4,>=3.2->ydata-
profiling) (3.0.9)
Requirement already satisfied: python-dateutil>=2.7 in
c:\users\acer\anaconda3\lib\site-packages (from matplotlib<4,>=3.2->ydata-
profiling) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in
c:\users\acer\anaconda3\lib\site-packages (from pandas!=1.4.0,<2.1,>1.1-
>ydata-profiling) (2022.7)
Requirement already satisfied: joblib>=0.14.1 in
c:\users\acer\anaconda3\lib\site-packages (from phik<0.13,>=0.11.1->ydata-
profiling) (1.1.1)
Requirement already satisfied: typing-extensions>=4.2.0 in
c:\users\acer\anaconda3\lib\site-packages (from pydantic<2,>=1.8.1->ydata-
profiling) (4.6.3)
Requirement already satisfied: charset-normalizer<4,>=2 in
c:\users\acer\anaconda3\lib\site-packages (from requests<3,>=2.24.0->ydata-
profiling) (2.0.4)
Requirement already satisfied: idna<4,>=2.5 in
c:\users\acer\anaconda3\lib\site-packages (from requests<3,>=2.24.0->ydata-
profiling) (3.4)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in
c:\users\acer\anaconda3\lib\site-packages (from requests<3,>=2.24.0->ydata-
profiling) (1.26.16)
Requirement already satisfied: certifi>=2017.4.17 in
c:\users\acer\anaconda3\lib\site-packages (from requests<3,>=2.24.0->ydata-
profiling) (2023.5.7)
Requirement already satisfied: patsy>=0.5.2 in
c:\users\acer\anaconda3\lib\site-packages (from statsmodels<1,>=0.13.2-
>ydata-profiling) (0.5.3)
Requirement already satisfied: colorama in c:\users\acer\anaconda3\lib\site-
packages (from tqdm<5,>=4.48.2->ydata-profiling) (0.4.6)
Requirement already satisfied: six in c:\users\acer\anaconda3\lib\site-
```

```
packages (from patsy>=0.5.2->statsmodels<1,>=0.13.2->ydata-profiling)
(1.16.0)
```

```
df = pd.read_csv('zomato_dataset.csv')
df.head()
```

```
  Restaurant Name  Dining Rating  Delivery Rating  Dining Votes  \
0      Doner King           3.9              4.2            39
1      Doner King           3.9              4.2            39
2      Doner King           3.9              4.2            39
3      Doner King           3.9              4.2            39
4      Doner King           3.9              4.2            39

   Delivery Votes   Cuisine  Place Name        City                 Item Name
\
0               0  Fast Food   Malakpet   Hyderabad       Platter Kebab Combo
1               0  Fast Food   Malakpet   Hyderabad   Chicken Rumali Shawarma
2               0  Fast Food   Malakpet   Hyderabad    Chicken Tandoori Salad
3               0  Fast Food   Malakpet   Hyderabad         Chicken BBQ Salad
4               0  Fast Food   Malakpet   Hyderabad  Special Doner Wrap Combo

  Best Seller  Votes  Prices
0  BESTSELLER     84   249.0
1  BESTSELLER     45   129.0
2         NaN     39   189.0
3  BESTSELLER     43   189.0
4    MUST TRY     31   205.0
```

```
from pydantic import BaseModel
```

```
from ydata_profiling import ProfileReport
report = ProfileReport(df)
report.to_file("zomato_dataset_eda.html")
```


```
{"model_id":"2ac3e51abeef4a1c96b467088b587cc3","version_major":2,"version_min
or":0}
```

```
{"model_id":"8aed0a00837b4916a1f895918047ae02","version_major":2,"version_min
or":0}
```

```
{"model_id":"5fc57864873c4382a964125598b846c7","version_major":2,"version_min
or":0}
```

```
{"model_id":"9aeedb16c7784feb89f515a2c822e0a6","version_major":2,"version_min
or":0}
```


Observation

we use pandas profiling to create csv (Comma Separated Value) file to html (Hypertext Markup Language) file.

Conclusions

After working on this data, we can conclude the following things:-

- Approx. 35% of restaurants in India are part of some chain
- Quick bites and casual dining type of most number of highest average ratings, votes.
- Banglore has most number of restaurants
- Mumbai and New Delhi dominates for most photo uploads per outlet
- Most restaurants are rated between 3 and 4
- Majority of restaurants are budget friendly with average cost of two between Rs.250 to Rs.800
- There are less number of restaurants at higher price ranges
- As the average cost of two increases, the chance of a restaurant having higher rating increases