BIG DATA FRAMEWORK PROJECT

Zomato Recommendation System using PySpark

Team Member:

1. Abhineet Raj 20MIA1146
2. Ayush Madurwar 20MIA1009
3. Tanmay Tiwari 20MIA1097

In [1]:
```
#Importing Libraries
import numpy as np
import pandas as pd
import seaborn as sb
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import LogisticRegression
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import r2_score
```

In [2]:
```
#reading the dataset
zomato_real=pd.read_csv("zomato.csv")
zomato_real.head()
```

Out[2]:

	url	address	name	online_order	book_table	r
0	https://www.zomato.com/bangalore/jalsa-banasha...	942, 21st Main Road, 2nd Stage, Banashankari, ...	Jalsa	Yes	Yes	4
1	https://www.zomato.com/bangalore/spice-elephan...	2nd Floor, 80 Feet Road, Near Big Bazaar, 6th ...	Spice Elephant	Yes	No	4
2	https://www.zomato.com/SanchurroBangalore?cont...	1112, Next to KIMS Medical College, 17th Cross...	San Churro Cafe	Yes	No	3
3	https://www.zomato.com/bangalore/addhuri-udupi...	1st Floor, Annakuteera, 3rd Stage, Banashankar...	Addhuri Udupi Bhojana	No	No	3
4	https://www.zomato.com/bangalore/grand-village...	10, 3rd Floor, Lakshmi Associates, Gandhi Baza...	Grand Village	No	No	3

In [3]:
```
zomato_real.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 51717 entries, 0 to 51716
Data columns (total 17 columns):
url                            51717 non-null object
address                        51717 non-null object
name                           51717 non-null object
online_order                   51717 non-null object
book_table                     51717 non-null object
rate                           43942 non-null object
votes                          51717 non-null int64
phone                          50509 non-null object
location                       51696 non-null object
rest_type                      51490 non-null object
dish_liked                     23639 non-null object
cuisines                       51672 non-null object
approx_cost(for two people)    51371 non-null object
reviews_list                   51717 non-null object
menu_item                      51717 non-null object
listed_in(type)                51717 non-null object
listed_in(city)                51717 non-null object
dtypes: int64(1), object(16)
memory usage: 6.7+ MB
```

In [4]:
```
zomato=zomato_real.drop(['url','dish_liked','phone'],axis=1) #Dropping the column
```

In [5]:
```
#Removing the Duplicates
zomato.duplicated().sum()
zomato.drop_duplicates(inplace=True)
```

In [6]:
```
#Remove the NaN values from the dataset
zomato.isnull().sum()
zomato.dropna(how='any',inplace=True)
zomato.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 43499 entries, 0 to 51716
Data columns (total 14 columns):
address                      43499 non-null object
name                         43499 non-null object
online_order                 43499 non-null object
book_table                   43499 non-null object
rate                         43499 non-null object
votes                        43499 non-null int64
location                     43499 non-null object
rest_type                    43499 non-null object
cuisines                     43499 non-null object
approx_cost(for two people)  43499 non-null object
reviews_list                 43499 non-null object
menu_item                    43499 non-null object
listed_in(type)              43499 non-null object
listed_in(city)              43499 non-null object
dtypes: int64(1), object(13)
memory usage: 5.0+ MB
```

In [7]:
```
zomato.columns
```

Out[7]:
```
Index(['address', 'name', 'online_order', 'book_table', 'rate', 'votes',
       'location', 'rest_type', 'cuisines', 'approx_cost(for two people)',
       'reviews_list', 'menu_item', 'listed_in(type)', 'listed_in(city)'],
      dtype='object')
```

In [8]:
```
zomato = zomato.rename(columns={'approx_cost(for two people)':'cost','listed_in(ty
                                'listed_in(city)':'city'})
zomato.columns
```

Out[8]:
```
Index(['address', 'name', 'online_order', 'book_table', 'rate', 'votes',
       'location', 'rest_type', 'cuisines', 'cost', 'reviews_list',
       'menu_item', 'type', 'city'],
      dtype='object')
```

In [9]:
```
zomato['cost'] = zomato['cost'].astype(str) #Changing the cost to string
zomato['cost'] = zomato['cost'].apply(lambda x: x.replace(',','.')) #Using lambda
zomato['cost'] = zomato['cost'].astype(float) # Changing the cost to Float
zomato.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 43499 entries, 0 to 51716
Data columns (total 14 columns):
address         43499 non-null object
name            43499 non-null object
online_order    43499 non-null object
book_table      43499 non-null object
rate            43499 non-null object
votes           43499 non-null int64
location        43499 non-null object
rest_type       43499 non-null object
cuisines        43499 non-null object
cost            43499 non-null float64
reviews_list    43499 non-null object
menu_item       43499 non-null object
type            43499 non-null object
city            43499 non-null object
dtypes: float64(1), int64(1), object(12)
memory usage: 5.0+ MB
```

In [10]:
```
zomato['rate'].unique()
```

Out[10]:
```
array(['4.1/5', '3.8/5', '3.7/5', '3.6/5', '4.6/5', '4.0/5', '4.2/5',
       '3.9/5', '3.1/5', '3.0/5', '3.2/5', '3.3/5', '2.8/5', '4.4/5',
       '4.3/5', 'NEW', '2.9/5', '3.5/5', '2.6/5', '3.8 /5', '3.4/5',
       '4.5/5', '2.5/5', '2.7/5', '4.7/5', '2.4/5', '2.2/5', '2.3/5',
       '3.4 /5', '-', '3.6 /5', '4.8/5', '3.9 /5', '4.2 /5', '4.0 /5',
       '4.1 /5', '3.7 /5', '3.1 /5', '2.9 /5', '3.3 /5', '2.8 /5',
       '3.5 /5', '2.7 /5', '2.5 /5', '3.2 /5', '2.6 /5', '4.5 /5',
       '4.3 /5', '4.4 /5', '4.9/5', '2.1/5', '2.0/5', '1.8/5', '4.6 /5',
       '4.9 /5', '3.0 /5', '4.8 /5', '2.3 /5', '4.7 /5', '2.4 /5',
       '2.1 /5', '2.2 /5', '2.0 /5', '1.8 /5'], dtype=object)
```

In [11]:
```
zomato = zomato.loc[zomato.rate !='NEW']
zomato = zomato.loc[zomato.rate !='-'].reset_index(drop=True)
remove_slash = lambda x: x.replace('/5', '') if type(x) == np.str else x
zomato.rate = zomato.rate.apply(remove_slash).str.strip().astype('float')
zomato['rate'].head()
```

Out[11]:
```
0    4.1
1    4.1
2    3.8
3    3.7
4    3.8
Name: rate, dtype: float64
```

In [12]:
```
zomato.name = zomato.name.apply(lambda x:x.title())
zomato.online_order.replace(('Yes','No'),(True, False),inplace=True)
zomato.book_table.replace(('Yes','No'),(True, False),inplace=True)
zomato.cost.unique()
```

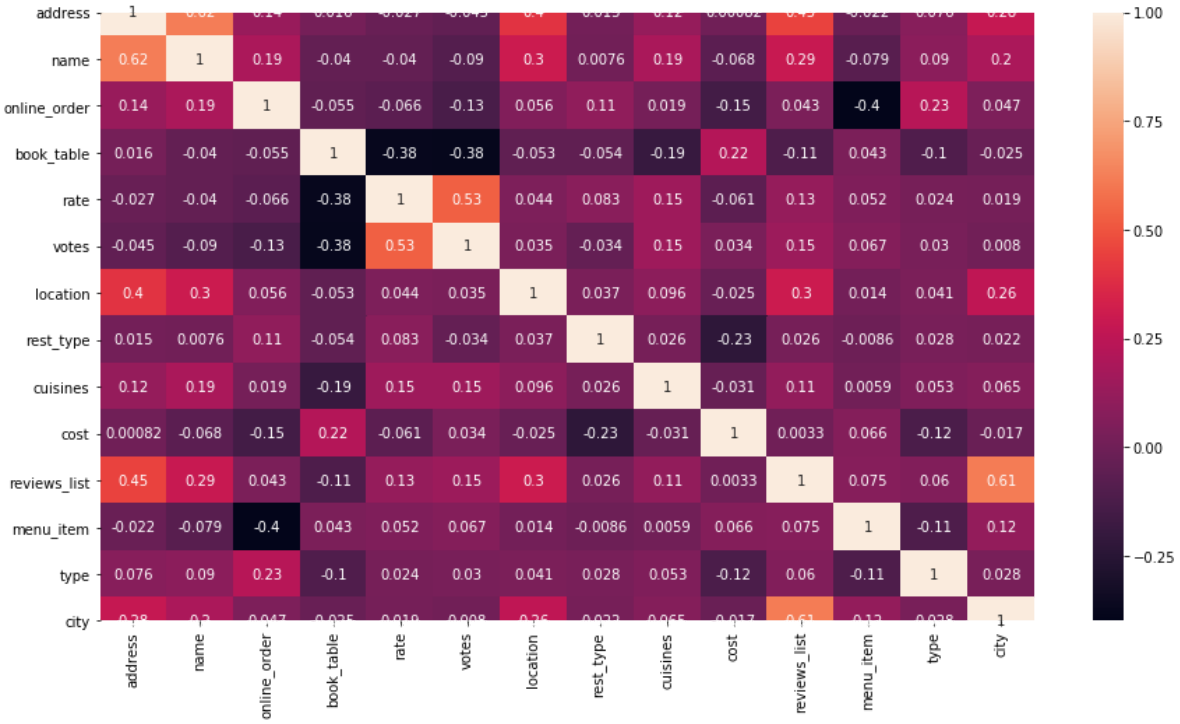
Out[12]:
```
array([800.  , 300.  , 600.  , 700.  , 550.  , 500.  , 450.  , 650.  ,
       400.  , 900.  , 200.  , 750.  , 150.  , 850.  , 100.  ,   1.2 ,
       350.  , 250.  , 950.  ,   1.  ,   1.5 ,   1.3 , 199.  ,   1.1 ,
         1.6 , 230.  , 130.  ,   1.7 ,   1.35,   2.2 ,   1.4 ,   2.  ,
         1.8 ,   1.9 , 180.  , 330.  ,   2.5 ,   2.1 ,   3.  ,   2.8 ,
         3.4 ,  50.  ,  40.  ,   1.25,   3.5 ,   4.  ,   2.4 ,   2.6 ,
         1.45,  70.  ,   3.2 , 240.  ,   6.  ,   1.05,   2.3 ,   4.1 ,
       120.  ,   5.  ,   3.7 ,   1.65,   2.7 ,   4.5 ,  80.  ])
```

In [13]:
```
def Encode(zomato):
    for column in zomato.columns[~zomato.columns.isin(['rate', 'cost', 'votes'])]:
        zomato[column] = zomato[column].factorize()[0]
    return zomato
```

```
zomato_en = Encode(zomato.copy())
```

In [15]:
```
corr = zomato_en.corr(method='kendall')
plt.figure(figsize=(15,8))
sns.heatmap(corr, annot=True)
plt.savefig("correlation.png")
zomato_en.columns
```

Out[15]:
```
Index(['address', 'name', 'online_order', 'book_table', 'rate', 'votes',
       'location', 'rest_type', 'cuisines', 'cost', 'reviews_list',
       'menu_item', 'type', 'city'],
      dtype='object')
```

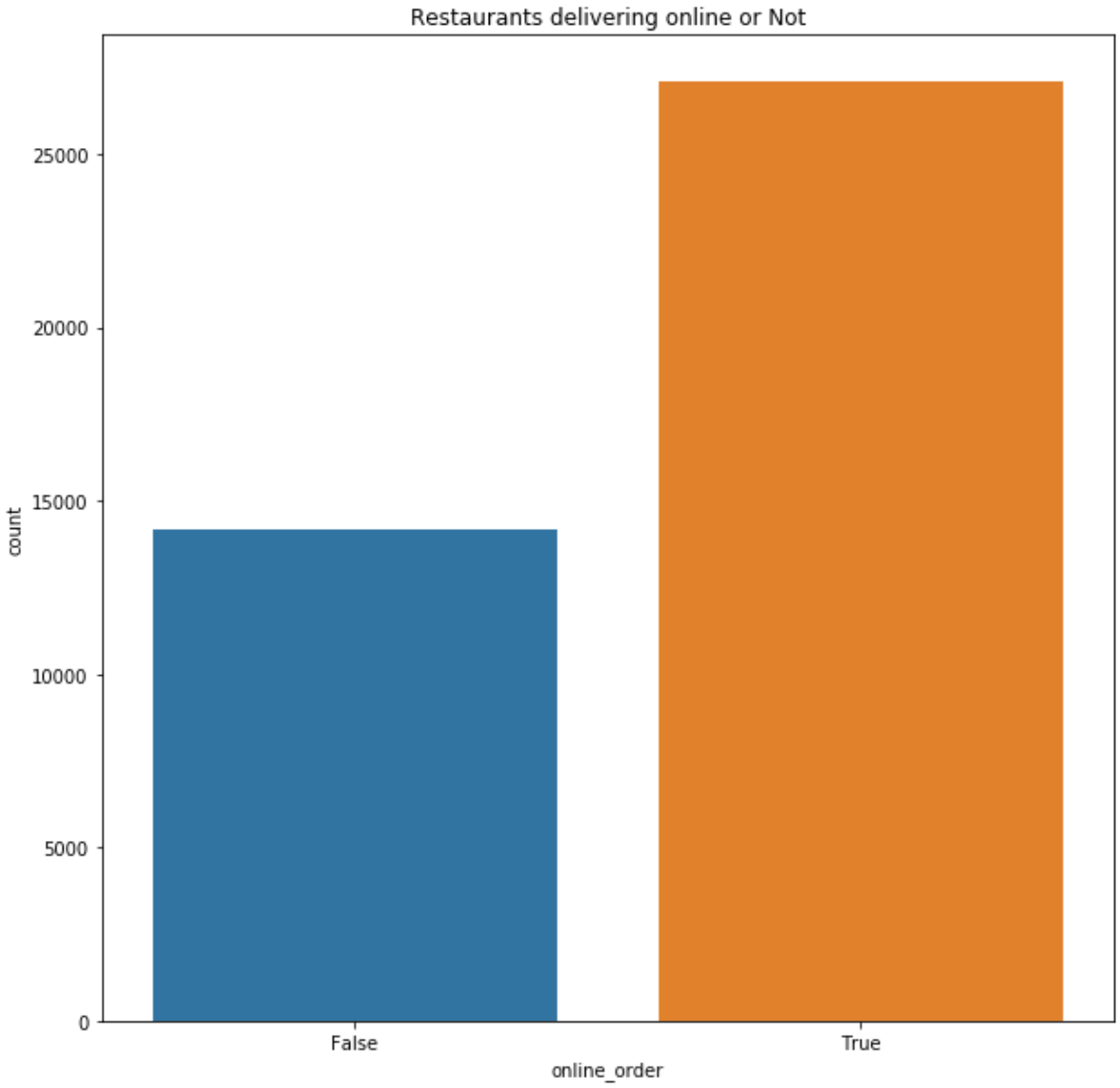


The highest correlation is between name and address which is 0.62 which is not of very much concern

Data Visualization

Restaurants delivering Online or not

In [21]:
```
sns.countplot(zomato['online_order'])
fig = plt.gcf()
fig.set_size_inches(10,10)
plt.title('Restaurants delivering online or Not')
plt.savefig("online.png")
```



Restaurants allowing table booking or not

In [22]:
```
sns.countplot(zomato['book_table'])
fig = plt.gcf()
fig.set_size_inches(10,10)
plt.savefig("Book_Table.png")
plt.title('Restaurants allowing table booking or not')
```

Out[22]:
```
Text(0.5, 1, 'Restaurants allowing table booking or not')
```

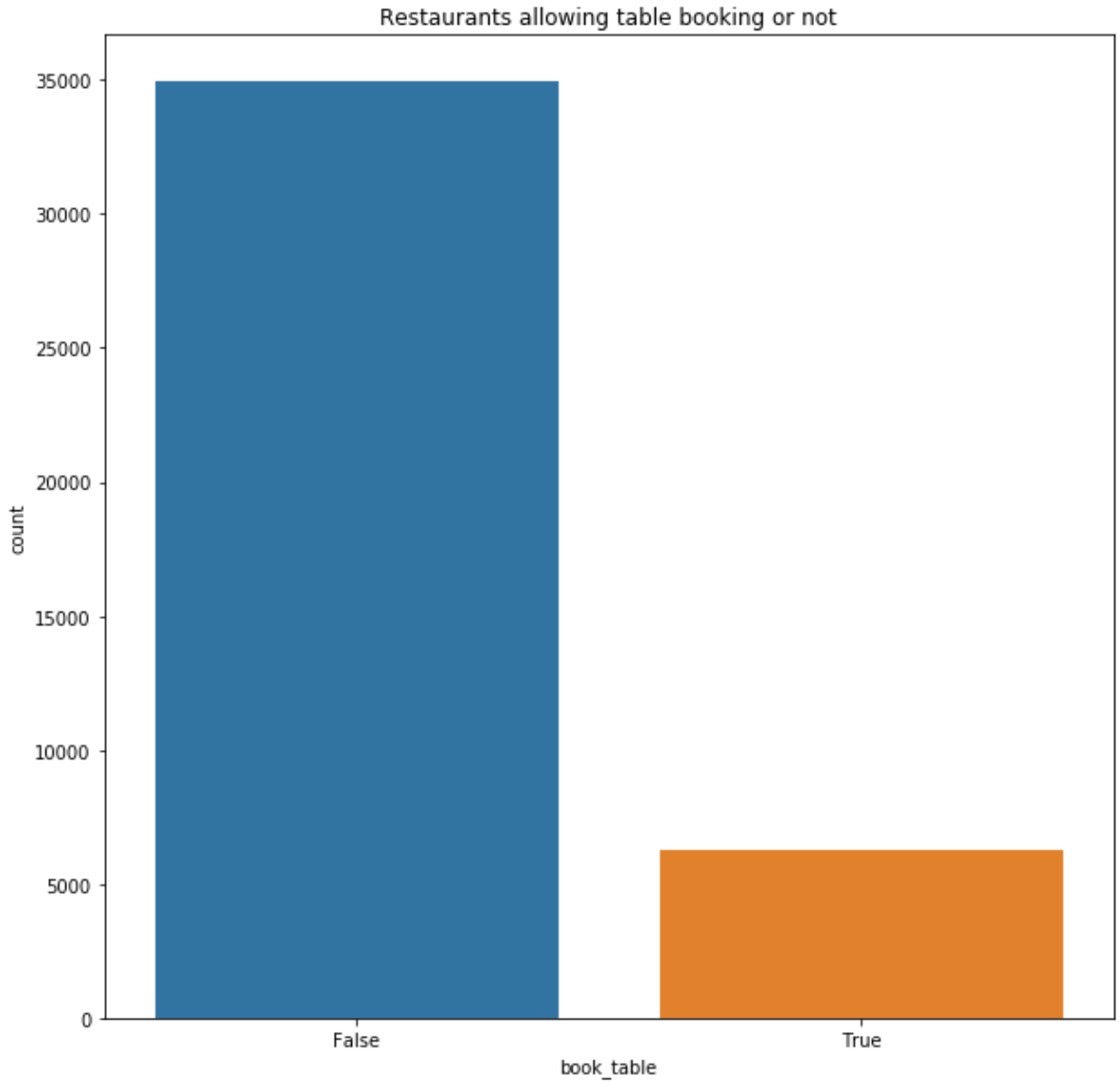
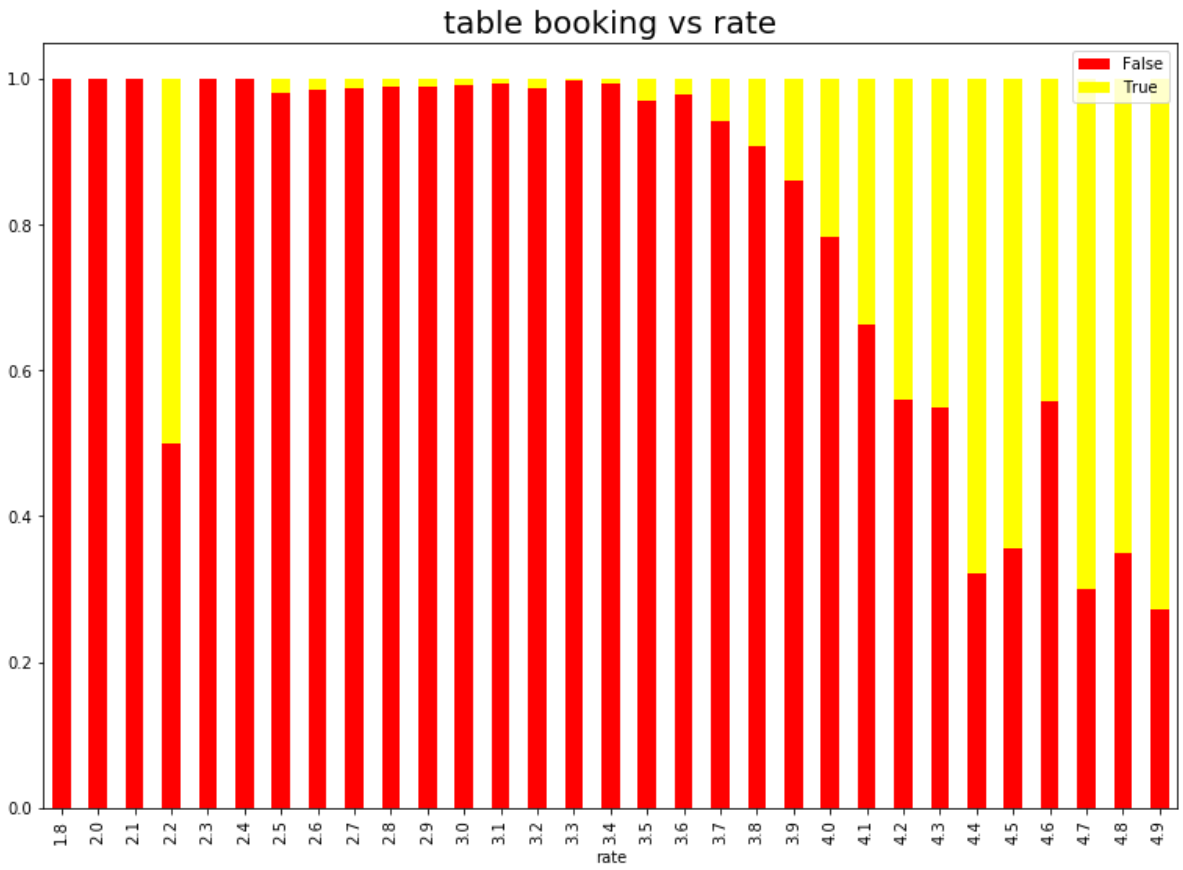


Table booking Rate vs Rate

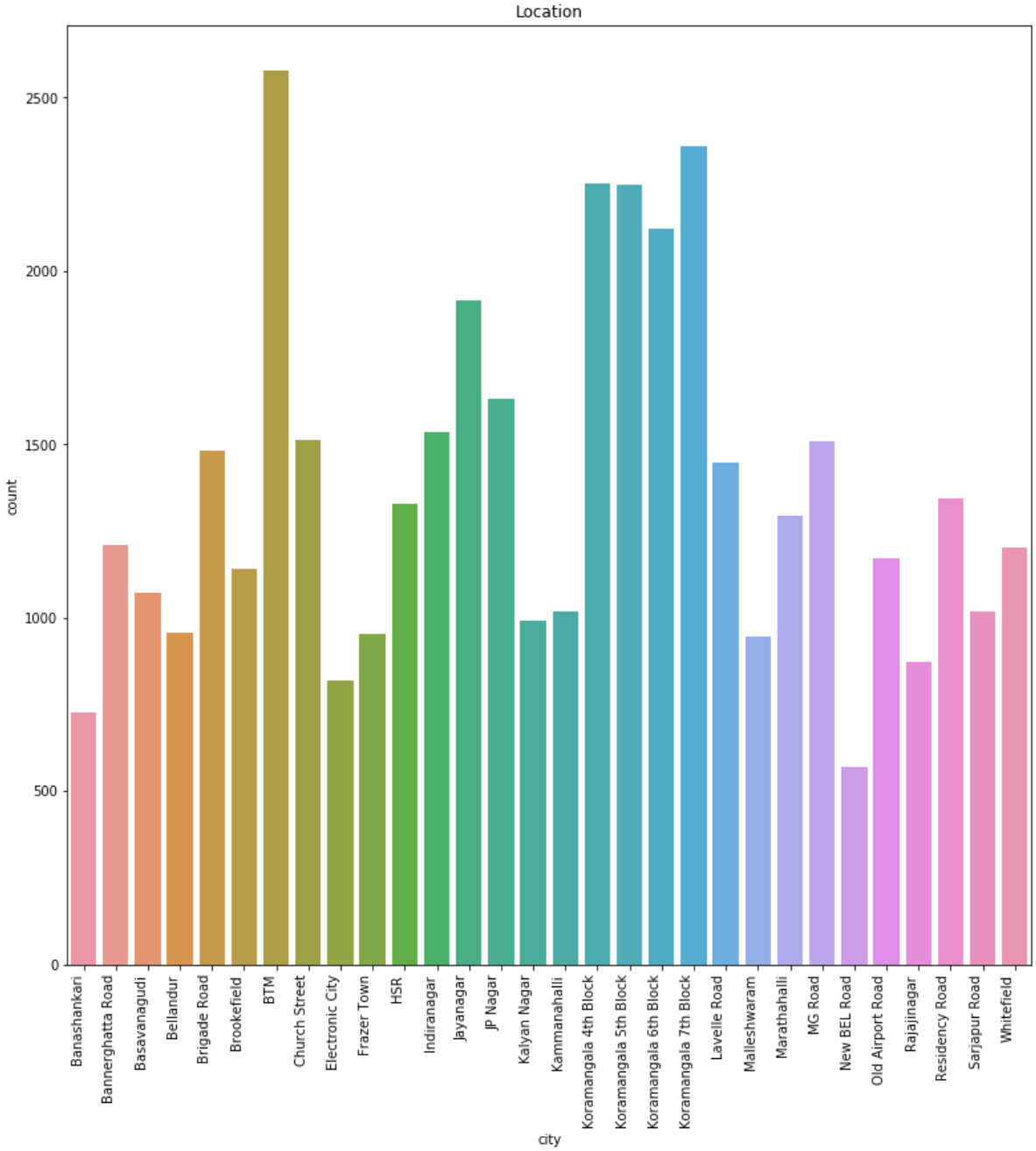
In [23]:
```
plt.rcParams['figure.figsize'] = (13, 9)
Y = pd.crosstab(zomato['rate'], zomato['book_table'])
Y.div(Y.sum(1).astype(float), axis = 0).plot(kind = 'bar', stacked = True,color=['
plt.title('table booking vs rate', fontweight = 30, fontsize = 20)
plt.legend(loc="upper right")
plt.savefig("Table_Booking_Rate.png")
plt.show()
```



Location

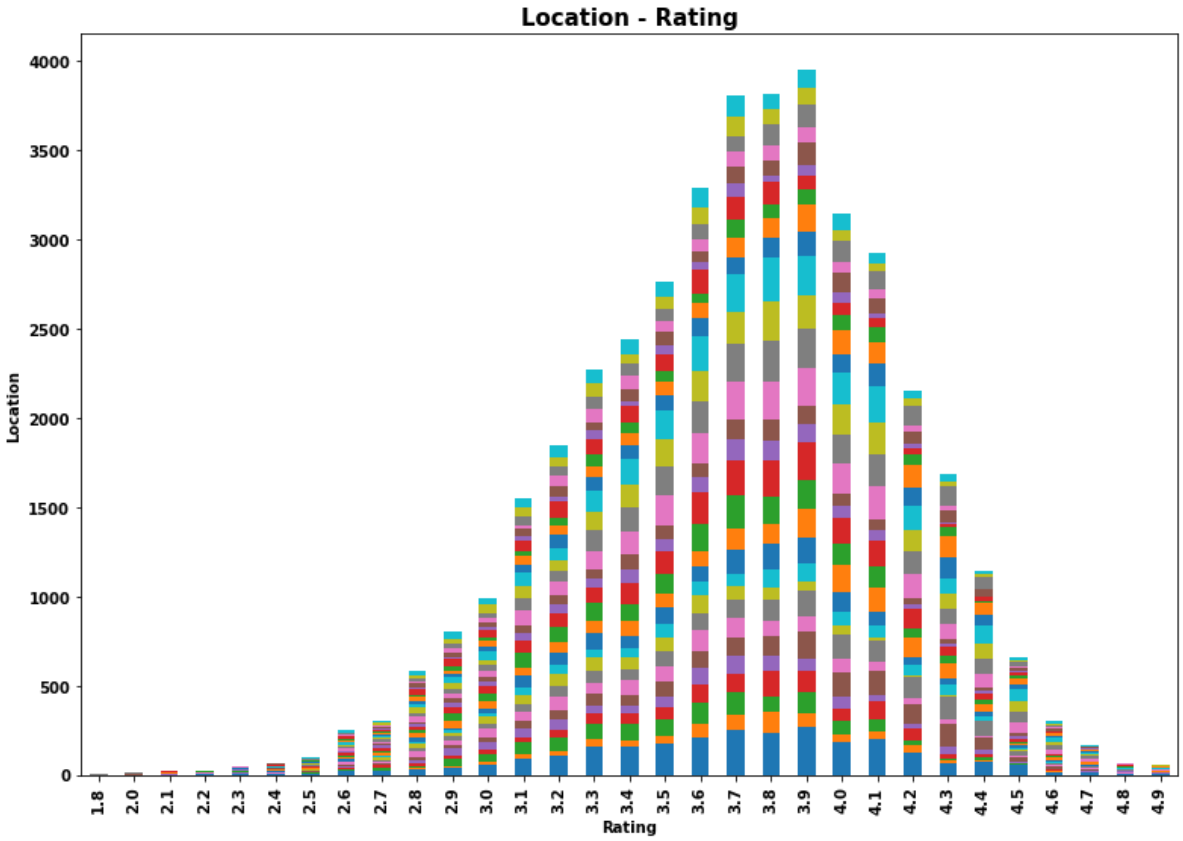
In [24]:
```
sns.countplot(zomato['city'])
sns.countplot(zomato['city']).set_xticklabels(sns.countplot(zomato['city']).get_xt
fig = plt.gcf()
fig.set_size_inches(13,13)
plt.savefig("Location.png")
plt.title('Location')
```

Out[24]:
```
Text(0.5, 1, 'Location')
```



Location and Rating

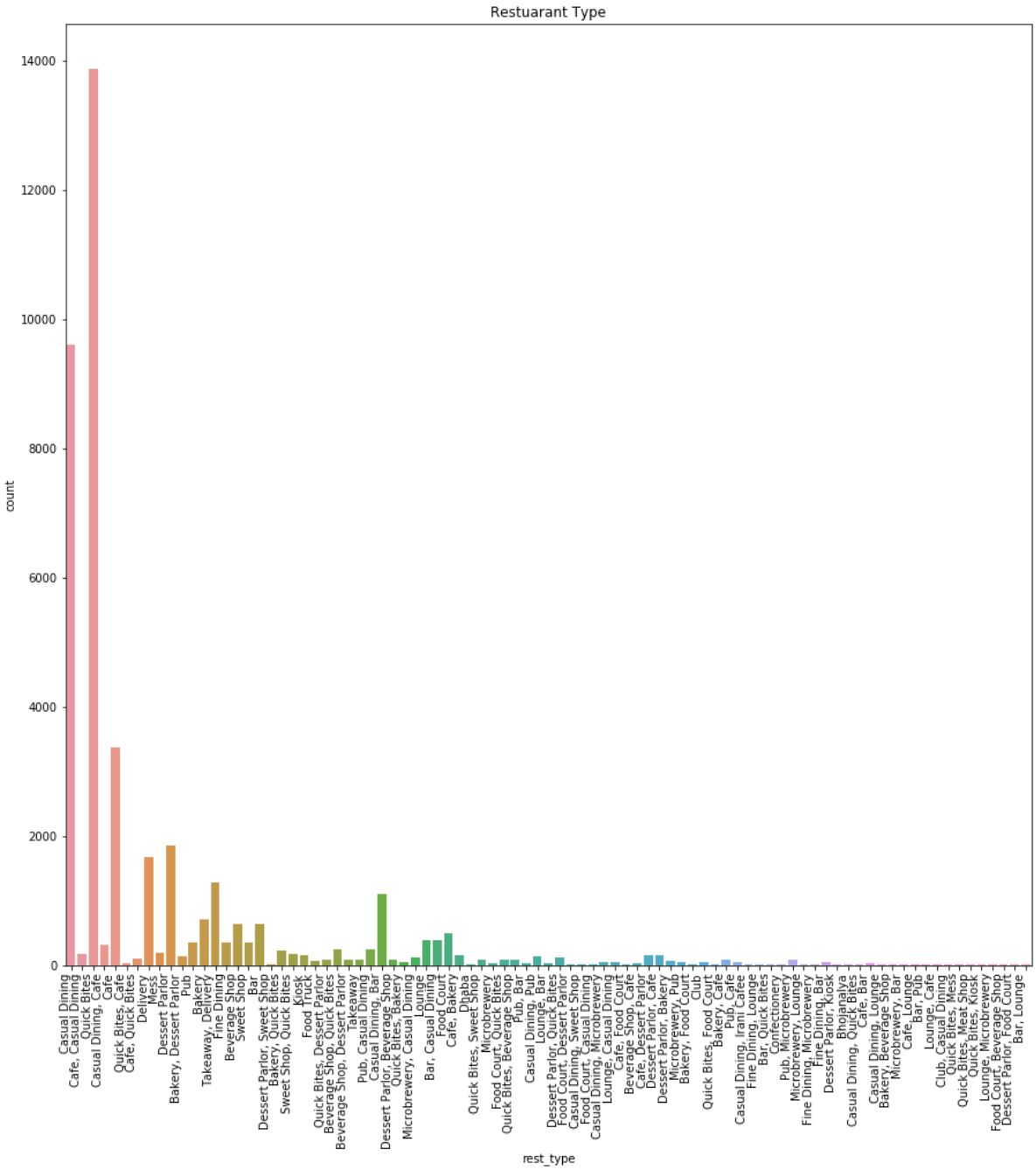
In [25]:
```
loc_plt=pd.crosstab(zomato['rate'],zomato['city'])
loc_plt.plot(kind='bar',stacked=True);
plt.title('Location - Rating',fontsize=15,fontweight='bold')
plt.ylabel('Location',fontsize=10,fontweight='bold')
plt.xlabel('Rating',fontsize=10,fontweight='bold')
plt.xticks(fontsize=10,fontweight='bold')
plt.yticks(fontsize=10,fontweight='bold');
plt.legend().remove();
plt.savefig("Location Rating.png")
```



Restaurant Type

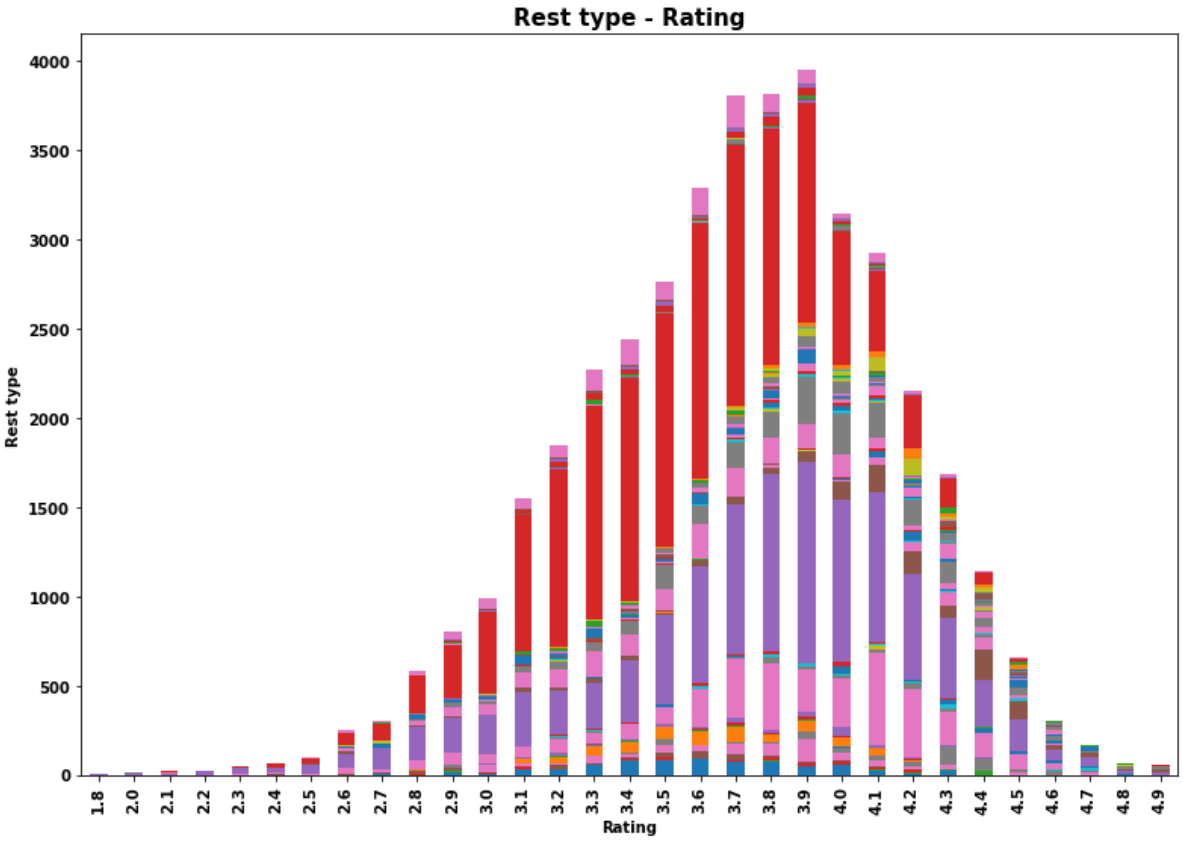
In [26]:
```
sns.countplot(zomato['rest_type'])
sns.countplot(zomato['rest_type']).set_xticklabels(sns.countplot(zomato['rest_type
fig = plt.gcf()
fig.set_size_inches(15,15)
plt.savefig("Restuarant Type")
plt.title('Restuarant Type')
```

Out[26]:
```
Text(0.5, 1, 'Restuarant Type')
```



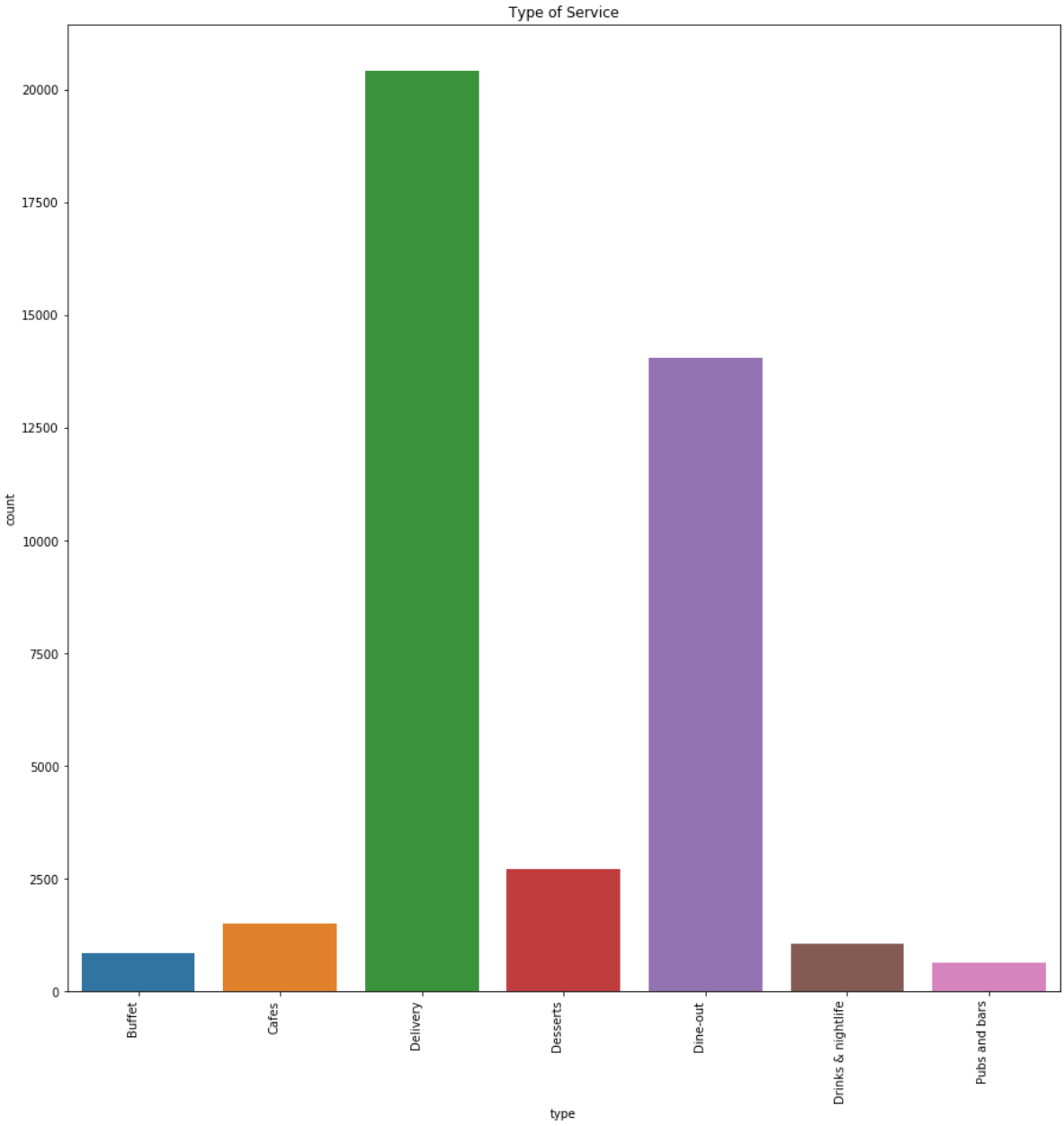
Gaussian Rest type and Rating

In [27]:
```
loc_plt=pd.crosstab(zomato['rate'],zomato['rest_type'])
loc_plt.plot(kind='bar',stacked=True);
plt.title('Rest type - Rating',fontsize=15,fontweight='bold')
plt.ylabel('Rest type',fontsize=10,fontweight='bold')
plt.xlabel('Rating',fontsize=10,fontweight='bold')
plt.xticks(fontsize=10,fontweight='bold')
plt.yticks(fontsize=10,fontweight='bold');
plt.legend().remove();
plt.savefig('Rest Type-Rating')
```



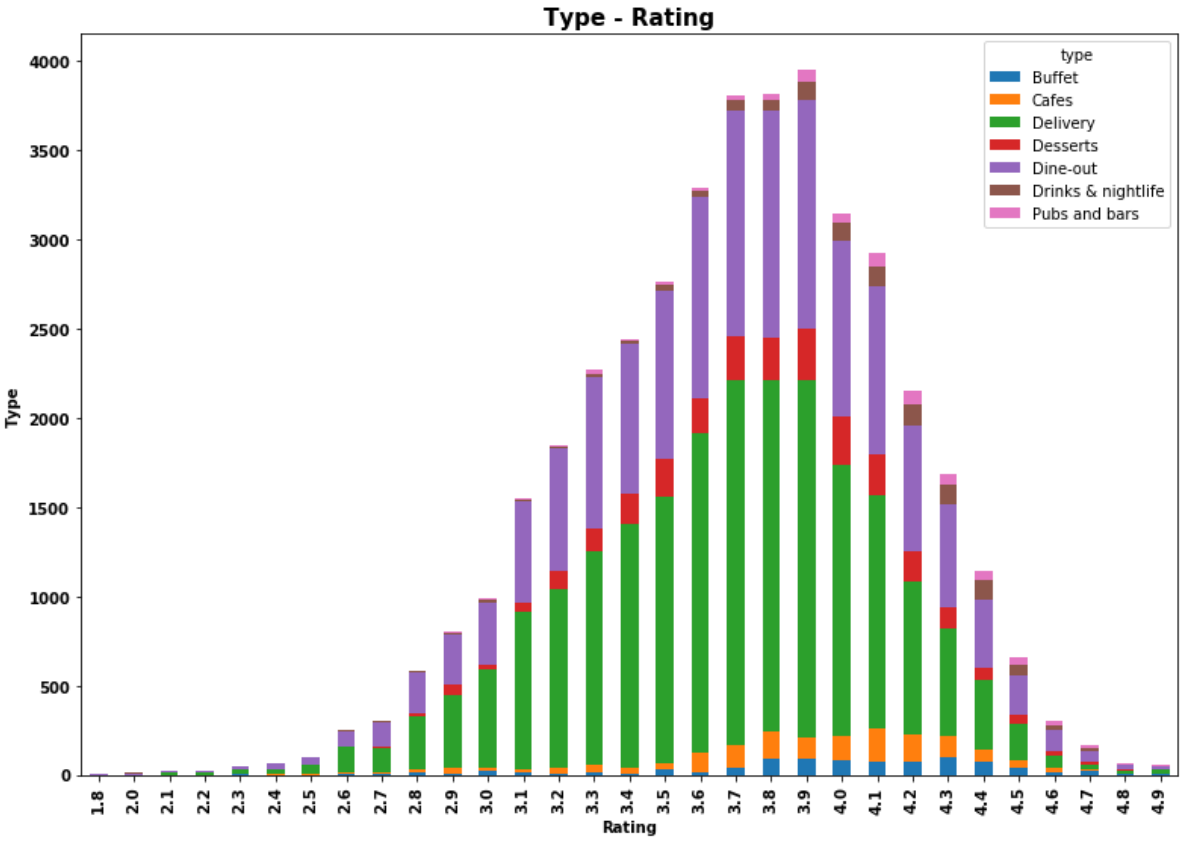
Types of Services

In [28]:
```
sns.countplot(zomato['type'])
sns.countplot(zomato['type']).set_xticklabels(sns.countplot(zomato['type']).get_xt
fig = plt.gcf()
fig.set_size_inches(15,15)
plt.title('Type of Service')
plt.savefig('Types of Service')
```



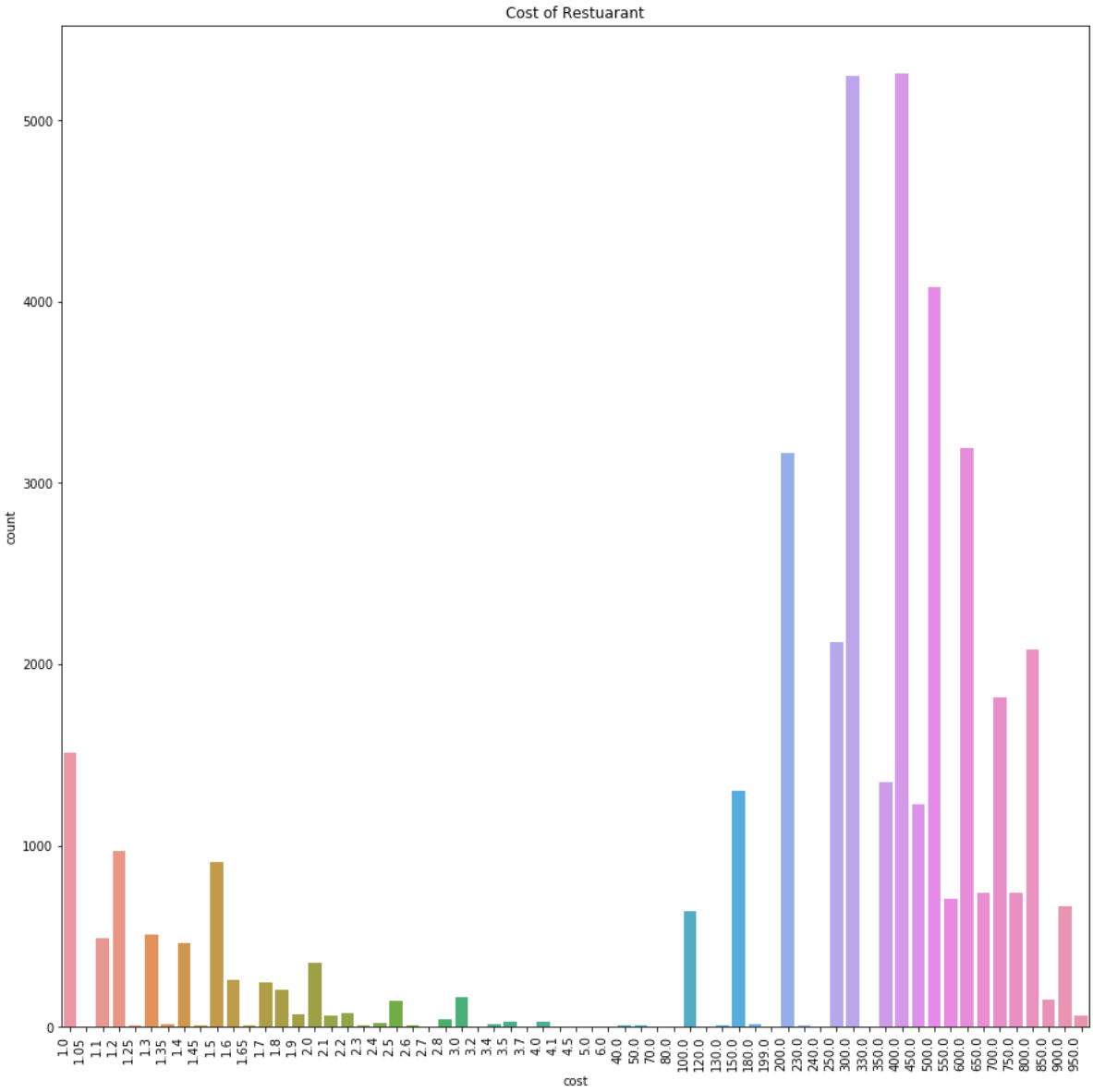
Type and Rating

In [29]:
```
type_plt=pd.crosstab(zomato['rate'],zomato['type'])
type_plt.plot(kind='bar',stacked=True);
plt.title('Type - Rating',fontsize=15,fontweight='bold')
plt.ylabel('Type',fontsize=10,fontweight='bold')
plt.xlabel('Rating',fontsize=10,fontweight='bold')
plt.xticks(fontsize=10,fontweight='bold')
plt.yticks(fontsize=10,fontweight='bold');
plt.savefig('Type and Rating')
```



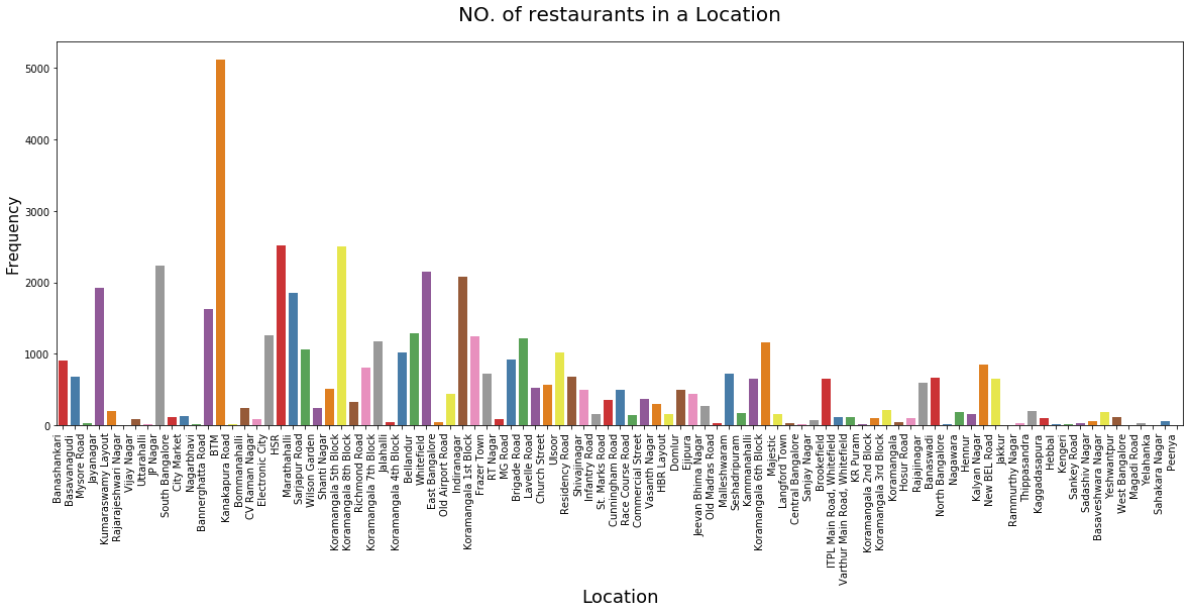
Cost of Restuarant

In [30]:
```
sns.countplot(zomato['cost'])
sns.countplot(zomato['cost']).set_xticklabels(sns.countplot(zomato['cost']).get_xt
fig = plt.gcf()
fig.set_size_inches(15,15)
plt.title('Cost of Restuarant')
plt.savefig('Cost of Restaurant')
```



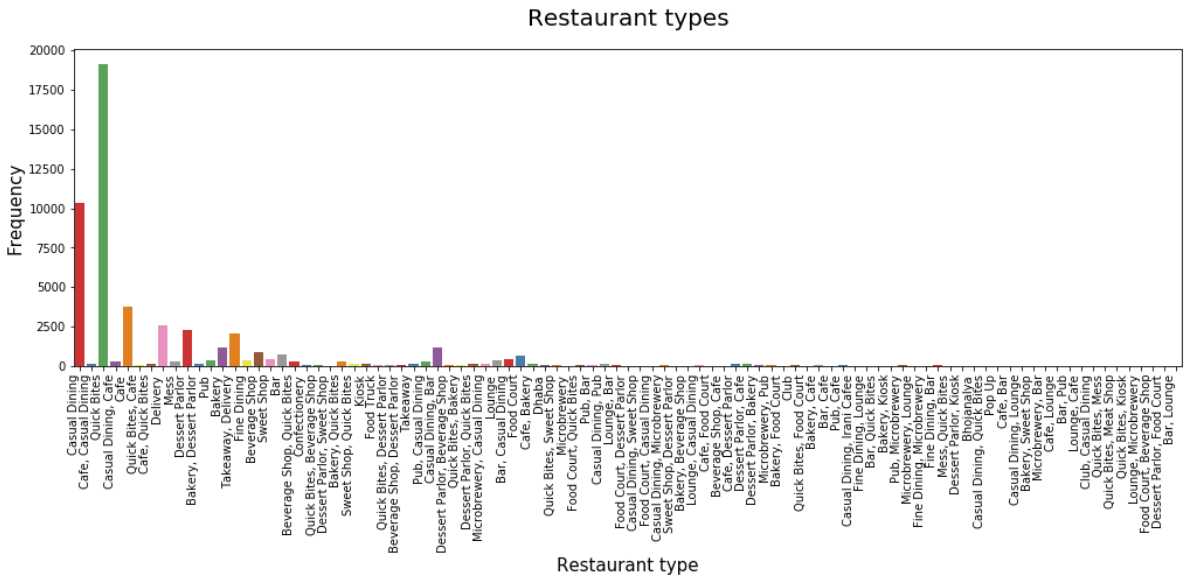
No. of Restaurants in a Location

In [31]:
```
fig = plt.figure(figsize=(20,7))
loc = sns.countplot(x="location",data=zomato_real, palette = "Set1")
loc.set_xticklabels(loc.get_xticklabels(), rotation=90, ha="right")
plt.ylabel("Frequency",size=15)
plt.xlabel("Location",size=18)
loc
plt.title('NO. of restaurants in a Location',size = 20,pad=20)
plt.savefig("Restaurants in Location")
```



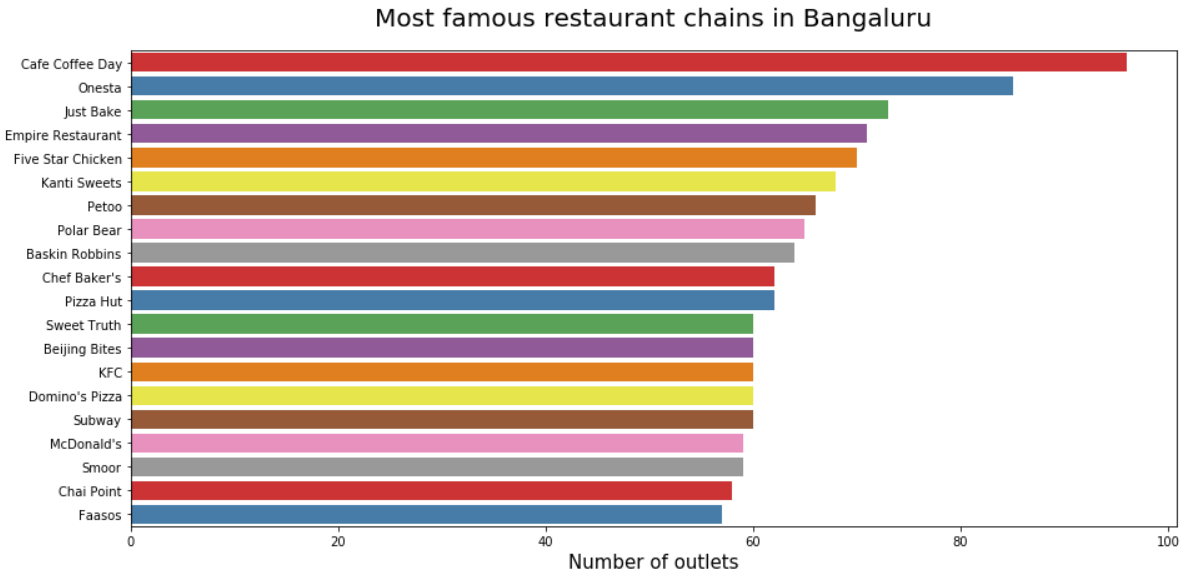
Restaurant type

In [32]:
```
fig = plt.figure(figsize=(17,5))
rest = sns.countplot(x="rest_type",data=zomato_real, palette = "Set1")
rest.set_xticklabels(rest.get_xticklabels(), rotation=90, ha="right")
plt.ylabel("Frequency",size=15)
plt.xlabel("Restaurant type",size=15)
rest
plt.title('Restaurant types',fontsize = 20 ,pad=20)
plt.savefig('Restaurant types')
```



Most famous Restaurant chains in Bengaluru

In [33]:
```
plt.figure(figsize=(15,7))
chains=zomato_real['name'].value_counts()[:20]
sns.barplot(x=chains,y=chains.index,palette='Set1')
plt.title("Most famous restaurant chains in Bangaluru",size=20,pad=20)
plt.xlabel("Number of outlets",size=15)
plt.savefig('Most famous restaurant chains')
```


Most famous restaurant chains in Bangaluru
Cafe Coffee Day
Onesta
Just Bake
Empire Restaurant
Five Star Chicken
Kanti Sweets
Petoo
Polar Bear
Baskin Robbins
Chef Baker's
Pizza Hut
Sweet Truth
Beijing Bites
KFC
Domino's Pizza
Subway
McDonald's
Smoor
Chai Point
Faasos
0
20
40
60
80
100
Number of outlets

Setup Spark

In [1]:
```
!apt-get install openjdk-8-jdk-headless -qq > /dev/null
!wget -q https://dlcdn.apache.org/spark/spark-3.3.2/spark-3.3.2-bin-hadoop3.tgz
!tar -zxvf spark-3.3.2-bin-hadoop3.tgz
```

```
spark-3.3.2-bin-hadoop3/
spark-3.3.2-bin-hadoop3/LICENSE
spark-3.3.2-bin-hadoop3/NOTICE
spark-3.3.2-bin-hadoop3/R/
spark-3.3.2-bin-hadoop3/R/lib/
spark-3.3.2-bin-hadoop3/R/lib/SparkR/
spark-3.3.2-bin-hadoop3/R/lib/SparkR/DESCRIPTION
spark-3.3.2-bin-hadoop3/R/lib/SparkR/INDEX
spark-3.3.2-bin-hadoop3/R/lib/SparkR/Meta/
spark-3.3.2-bin-hadoop3/R/lib/SparkR/Meta/Rd.rds
spark-3.3.2-bin-hadoop3/R/lib/SparkR/Meta/features.rds
spark-3.3.2-bin-hadoop3/R/lib/SparkR/Meta/hsearch.rds
spark-3.3.2-bin-hadoop3/R/lib/SparkR/Meta/links.rds
spark-3.3.2-bin-hadoop3/R/lib/SparkR/Meta/nsInfo.rds
spark-3.3.2-bin-hadoop3/R/lib/SparkR/Meta/package.rds
spark-3.3.2-bin-hadoop3/R/lib/SparkR/Meta/vignette.rds
spark-3.3.2-bin-hadoop3/R/lib/SparkR/NAMESPACE
spark-3.3.2-bin-hadoop3/R/lib/SparkR/R/
spark-3.3.2-bin-hadoop3/R/lib/SparkR/R/SparkR
spark-3.3.2-bin-hadoop3/R/lib/SparkR/R/SparkR.rdb
spark-3.3.2-bin-hadoop3/R/lib/SparkR/R/SparkR.rdx
spark-3.3.2-bin-hadoop3/R/lib/SparkR/doc/
spark-3.3.2-bin-hadoop3/R/lib/SparkR/doc/index.html
spark-3.3.2-bin-hadoop3/R/lib/SparkR/doc/sparkr-vignettes.R
spark-3.3.2-bin-hadoop3/R/lib/SparkR/doc/sparkr-vignettes.Rmd
spark-3.3.2-bin-hadoop3/R/lib/SparkR/doc/sparkr-vignettes.html
spark-3.3.2-bin-hadoop3/R/lib/SparkR/help/
spark-3.3.2-bin-hadoop3/R/lib/SparkR/help/AnIndex
spark-3.3.2-bin-hadoop3/R/lib/SparkR/help/SparkR.rdb
spark-3.3.2-bin-hadoop3/R/lib/SparkR/help/SparkR.rdx
spark-3.3.2-bin-hadoop3/R/lib/SparkR/help/aliases.rds
spark-3.3.2-bin-hadoop3/R/lib/SparkR/help/paths.rds
spark-3.3.2-bin-hadoop3/R/lib/SparkR/html/
spark-3.3.2-bin-hadoop3/R/lib/SparkR/html/00Index.html
spark-3.3.2-bin-hadoop3/R/lib/SparkR/html/R.css
spark-3.3.2-bin-hadoop3/R/lib/SparkR/profile/
spark-3.3.2-bin-hadoop3/R/lib/SparkR/profile/general.R
spark-3.3.2-bin-hadoop3/R/lib/SparkR/profile/shell.R
spark-3.3.2-bin-hadoop3/R/lib/SparkR/tests/
spark-3.3.2-bin-hadoop3/R/lib/SparkR/tests/testthat/
spark-3.3.2-bin-hadoop3/R/lib/SparkR/tests/testthat/test_basic.R
spark-3.3.2-bin-hadoop3/R/lib/SparkR/worker/
spark-3.3.2-bin-hadoop3/R/lib/SparkR/worker/daemon.R
spark-3.3.2-bin-hadoop3/R/lib/SparkR/worker/worker.R
spark-3.3.2-bin-hadoop3/R/lib/sparkr.zip
spark-3.3.2-bin-hadoop3/README.md
spark-3.3.2-bin-hadoop3/RELEASE
spark-3.3.2-bin-hadoop3/bin/
spark-3.3.2-bin-hadoop3/bin/beeline
spark-3.3.2-bin-hadoop3/bin/beeline.cmd
spark-3.3.2-bin-hadoop3/bin/docker-image-tool.sh
spark-3.3.2-bin-hadoop3/bin/find-spark-home
spark-3.3.2-bin-hadoop3/bin/find-spark-home.cmd
spark-3.3.2-bin-hadoop3/bin/load-spark-env.cmd
spark-3.3.2-bin-hadoop3/bin/load-spark-env.sh
spark-3.3.2-bin-hadoop3/bin/pyspark
spark-3.3.2-bin-hadoop3/bin/pyspark.cmd
spark-3.3.2-bin-hadoop3/bin/pyspark2.cmd
spark-3.3.2-bin-hadoop3/bin/run-example
spark-3.3.2-bin-hadoop3/bin/run-example.cmd
spark-3.3.2-bin-hadoop3/bin/spark-class
spark-3.3.2-bin-hadoop3/bin/spark-class.cmd
spark-3.3.2-bin-hadoop3/bin/spark-class2.cmd
spark-3.3.2-bin-hadoop3/bin/spark-shell
```

```
spark-3.3.2-bin-hadoop3/bin/spark-shell.cmd
spark-3.3.2-bin-hadoop3/bin/spark-shell2.cmd
spark-3.3.2-bin-hadoop3/bin/spark-sql
spark-3.3.2-bin-hadoop3/bin/spark-sql.cmd
spark-3.3.2-bin-hadoop3/bin/spark-sql2.cmd
spark-3.3.2-bin-hadoop3/bin/spark-submit
spark-3.3.2-bin-hadoop3/bin/spark-submit.cmd
spark-3.3.2-bin-hadoop3/bin/spark-submit2.cmd
spark-3.3.2-bin-hadoop3/bin/sparkR
spark-3.3.2-bin-hadoop3/bin/sparkR.cmd
spark-3.3.2-bin-hadoop3/bin/sparkR2.cmd
spark-3.3.2-bin-hadoop3/conf/
spark-3.3.2-bin-hadoop3/conf/fairscheduler.xml.template
spark-3.3.2-bin-hadoop3/conf/log4j2.properties.template
spark-3.3.2-bin-hadoop3/conf/metrics.properties.template
spark-3.3.2-bin-hadoop3/conf/spark-defaults.conf.template
spark-3.3.2-bin-hadoop3/conf/spark-env.sh.template
spark-3.3.2-bin-hadoop3/conf/workers.template
spark-3.3.2-bin-hadoop3/data/
spark-3.3.2-bin-hadoop3/data/graphx/
spark-3.3.2-bin-hadoop3/data/graphx/followers.txt
spark-3.3.2-bin-hadoop3/data/graphx/users.txt
spark-3.3.2-bin-hadoop3/data/mllib/
spark-3.3.2-bin-hadoop3/data/mllib/als/
spark-3.3.2-bin-hadoop3/data/mllib/als/sample_movielens_ratings.txt
spark-3.3.2-bin-hadoop3/data/mllib/als/test.data
spark-3.3.2-bin-hadoop3/data/mllib/gmm_data.txt
spark-3.3.2-bin-hadoop3/data/mllib/images/
spark-3.3.2-bin-hadoop3/data/mllib/images/license.txt
spark-3.3.2-bin-hadoop3/data/mllib/images/origin/
spark-3.3.2-bin-hadoop3/data/mllib/images/origin/kittens/
spark-3.3.2-bin-hadoop3/data/mllib/images/origin/kittens/29.5.a_b_EGDP022204.jpg
spark-3.3.2-bin-hadoop3/data/mllib/images/origin/kittens/54893.jpg
spark-3.3.2-bin-hadoop3/data/mllib/images/origin/kittens/DP153539.jpg
spark-3.3.2-bin-hadoop3/data/mllib/images/origin/kittens/DP802813.jpg
spark-3.3.2-bin-hadoop3/data/mllib/images/origin/kittens/not-image.txt
spark-3.3.2-bin-hadoop3/data/mllib/images/origin/license.txt
spark-3.3.2-bin-hadoop3/data/mllib/images/origin/multi-channel/
spark-3.3.2-bin-hadoop3/data/mllib/images/origin/multi-channel/BGRA.png
spark-3.3.2-bin-hadoop3/data/mllib/images/origin/multi-channel/BGRA_alpha_60.png
spark-3.3.2-bin-hadoop3/data/mllib/images/origin/multi-channel/chr30.4.184.jpg
spark-3.3.2-bin-hadoop3/data/mllib/images/origin/multi-channel/grayscale.jpg
spark-3.3.2-bin-hadoop3/data/mllib/kmeans_data.txt
spark-3.3.2-bin-hadoop3/data/mllib/pagerank_data.txt
spark-3.3.2-bin-hadoop3/data/mllib/pic_data.txt
spark-3.3.2-bin-hadoop3/data/mllib/ridge-data/
spark-3.3.2-bin-hadoop3/data/mllib/ridge-data/lpsa.data
spark-3.3.2-bin-hadoop3/data/mllib/sample_binary_classification_data.txt
spark-3.3.2-bin-hadoop3/data/mllib/sample_fpgrowth.txt
spark-3.3.2-bin-hadoop3/data/mllib/sample_isotonic_regression_libsvm_data.txt
spark-3.3.2-bin-hadoop3/data/mllib/sample_kmeans_data.txt
spark-3.3.2-bin-hadoop3/data/mllib/sample_lda_data.txt
spark-3.3.2-bin-hadoop3/data/mllib/sample_lda_libsvm_data.txt
spark-3.3.2-bin-hadoop3/data/mllib/sample_libsvm_data.txt
spark-3.3.2-bin-hadoop3/data/mllib/sample_linear_regression_data.txt
spark-3.3.2-bin-hadoop3/data/mllib/sample_movielens_data.txt
spark-3.3.2-bin-hadoop3/data/mllib/sample_multiclass_classification_data.txt
spark-3.3.2-bin-hadoop3/data/mllib/sample_svm_data.txt
spark-3.3.2-bin-hadoop3/data/mllib/streaming_kmeans_data_test.txt
spark-3.3.2-bin-hadoop3/data/streaming/
spark-3.3.2-bin-hadoop3/data/streaming/AFINN-111.txt
spark-3.3.2-bin-hadoop3/examples/
spark-3.3.2-bin-hadoop3/examples/jars/
spark-3.3.2-bin-hadoop3/examples/jars/scopt_2.12-3.7.1.jar
```

```
spark-3.3.2-bin-hadoop3/python/test_support/sql/orc_partitioned/b=0/c=0/.part-r-00
000-829af031-b970-49d6-ad39-30460a0be2c8.orc.crc
spark-3.3.2-bin-hadoop3/python/test_support/sql/orc_partitioned/b=0/c=0/part-r-000
00-829af031-b970-49d6-ad39-30460a0be2c8.orc
spark-3.3.2-bin-hadoop3/python/test_support/sql/orc_partitioned/b=1/
spark-3.3.2-bin-hadoop3/python/test_support/sql/orc_partitioned/b=1/c=1/
spark-3.3.2-bin-hadoop3/python/test_support/sql/orc_partitioned/b=1/c=1/.part-r-00
000-829af031-b970-49d6-ad39-30460a0be2c8.orc.crc
spark-3.3.2-bin-hadoop3/python/test_support/sql/orc_partitioned/b=1/c=1/part-r-000
00-829af031-b970-49d6-ad39-30460a0be2c8.orc
spark-3.3.2-bin-hadoop3/python/test_support/sql/parquet_partitioned/
spark-3.3.2-bin-hadoop3/python/test_support/sql/parquet_partitioned/_SUCCESS
spark-3.3.2-bin-hadoop3/python/test_support/sql/parquet_partitioned/_common_metada
ta
spark-3.3.2-bin-hadoop3/python/test_support/sql/parquet_partitioned/_metadata
spark-3.3.2-bin-hadoop3/python/test_support/sql/parquet_partitioned/year=2014/
spark-3.3.2-bin-hadoop3/python/test_support/sql/parquet_partitioned/year=2014/mont
h=9/
spark-3.3.2-bin-hadoop3/python/test_support/sql/parquet_partitioned/year=2014/mont
h=9/day=1/
spark-3.3.2-bin-hadoop3/python/test_support/sql/parquet_partitioned/year=2014/mont
h=9/day=1/.part-r-00008.gz.parquet.crc
spark-3.3.2-bin-hadoop3/python/test_support/sql/parquet_partitioned/year=2014/mont
h=9/day=1/part-r-00008.gz.parquet
spark-3.3.2-bin-hadoop3/python/test_support/sql/parquet_partitioned/year=2015/
spark-3.3.2-bin-hadoop3/python/test_support/sql/parquet_partitioned/year=2015/mont
h=10/
spark-3.3.2-bin-hadoop3/python/test_support/sql/parquet_partitioned/year=2015/mont
h=10/day=25/
spark-3.3.2-bin-hadoop3/python/test_support/sql/parquet_partitioned/year=2015/mont
h=10/day=25/.part-r-00002.gz.parquet.crc
spark-3.3.2-bin-hadoop3/python/test_support/sql/parquet_partitioned/year=2015/mont
h=10/day=25/.part-r-00004.gz.parquet.crc
spark-3.3.2-bin-hadoop3/python/test_support/sql/parquet_partitioned/year=2015/mont
h=10/day=25/part-r-00002.gz.parquet
spark-3.3.2-bin-hadoop3/python/test_support/sql/parquet_partitioned/year=2015/mont
h=10/day=25/part-r-00004.gz.parquet
spark-3.3.2-bin-hadoop3/python/test_support/sql/parquet_partitioned/year=2015/mont
h=10/day=26/
spark-3.3.2-bin-hadoop3/python/test_support/sql/parquet_partitioned/year=2015/mont
h=10/day=26/.part-r-00005.gz.parquet.crc
spark-3.3.2-bin-hadoop3/python/test_support/sql/parquet_partitioned/year=2015/mont
h=10/day=26/part-r-00005.gz.parquet
spark-3.3.2-bin-hadoop3/python/test_support/sql/parquet_partitioned/year=2015/mont
h=9/
spark-3.3.2-bin-hadoop3/python/test_support/sql/parquet_partitioned/year=2015/mont
h=9/day=1/
spark-3.3.2-bin-hadoop3/python/test_support/sql/parquet_partitioned/year=2015/mont
h=9/day=1/.part-r-00007.gz.parquet.crc
spark-3.3.2-bin-hadoop3/python/test_support/sql/parquet_partitioned/year=2015/mont
h=9/day=1/part-r-00007.gz.parquet
spark-3.3.2-bin-hadoop3/python/test_support/sql/people.json
spark-3.3.2-bin-hadoop3/python/test_support/sql/people1.json
spark-3.3.2-bin-hadoop3/python/test_support/sql/people_array.json
spark-3.3.2-bin-hadoop3/python/test_support/sql/people_array_utf16le.json
spark-3.3.2-bin-hadoop3/python/test_support/sql/streaming/
spark-3.3.2-bin-hadoop3/python/test_support/sql/streaming/text-test.txt
spark-3.3.2-bin-hadoop3/python/test_support/sql/text-test.txt
spark-3.3.2-bin-hadoop3/python/test_support/userlib-0.1.zip
spark-3.3.2-bin-hadoop3/python/test_support/userlibrary.py
spark-3.3.2-bin-hadoop3/sbin/
spark-3.3.2-bin-hadoop3/sbin/decommission-slave.sh
spark-3.3.2-bin-hadoop3/sbin/decommission-worker.sh
spark-3.3.2-bin-hadoop3/sbin/slaves.sh
```

```
spark-3.3.2-bin-hadoop3/sbin/spark-config.sh
spark-3.3.2-bin-hadoop3/sbin/spark-daemon.sh
spark-3.3.2-bin-hadoop3/sbin/spark-daemons.sh
spark-3.3.2-bin-hadoop3/sbin/start-all.sh
spark-3.3.2-bin-hadoop3/sbin/start-history-server.sh
spark-3.3.2-bin-hadoop3/sbin/start-master.sh
spark-3.3.2-bin-hadoop3/sbin/start-mesos-dispatcher.sh
spark-3.3.2-bin-hadoop3/sbin/start-mesos-shuffle-service.sh
spark-3.3.2-bin-hadoop3/sbin/start-slave.sh
spark-3.3.2-bin-hadoop3/sbin/start-slaves.sh
spark-3.3.2-bin-hadoop3/sbin/start-thriftserver.sh
spark-3.3.2-bin-hadoop3/sbin/start-worker.sh
spark-3.3.2-bin-hadoop3/sbin/start-workers.sh
spark-3.3.2-bin-hadoop3/sbin/stop-all.sh
spark-3.3.2-bin-hadoop3/sbin/stop-history-server.sh
spark-3.3.2-bin-hadoop3/sbin/stop-master.sh
spark-3.3.2-bin-hadoop3/sbin/stop-mesos-dispatcher.sh
spark-3.3.2-bin-hadoop3/sbin/stop-mesos-shuffle-service.sh
spark-3.3.2-bin-hadoop3/sbin/stop-slave.sh
spark-3.3.2-bin-hadoop3/sbin/stop-slaves.sh
spark-3.3.2-bin-hadoop3/sbin/stop-thriftserver.sh
spark-3.3.2-bin-hadoop3/sbin/stop-worker.sh
spark-3.3.2-bin-hadoop3/sbin/stop-workers.sh
spark-3.3.2-bin-hadoop3/sbin/workers.sh
spark-3.3.2-bin-hadoop3/yarn/
spark-3.3.2-bin-hadoop3/yarn/spark-3.3.2-yarn-shuffle.jar
```

In [2]:
```
!apt-get update
```

```
Get:1 https://cloud.r-project.org/bin/linux/ubuntu focal-cran40/ InRelease [3,622
B]
Get:2 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu2004/x86_64
InRelease [1,581 B]
Get:3 http://security.ubuntu.com/ubuntu focal-security InRelease [114 kB]
Hit:4 http://ppa.launchpad.net/c2d4u.team/c2d4u4.0+/ubuntu focal InRelease
Hit:5 http://archive.ubuntu.com/ubuntu focal InRelease
Get:6 http://archive.ubuntu.com/ubuntu focal-updates InRelease [114 kB]
Get:7 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu2004/x86_64
Packages [972 kB]
Hit:8 http://ppa.launchpad.net/cran/libgit2/ubuntu focal InRelease
Get:9 http://archive.ubuntu.com/ubuntu focal-backports InRelease [108 kB]
Hit:10 http://ppa.launchpad.net/deadsnakes/ppa/ubuntu focal InRelease
Get:11 http://security.ubuntu.com/ubuntu focal-security/restricted amd64 Packages
[2,060 kB]
Get:12 http://archive.ubuntu.com/ubuntu focal-updates/universe amd64 Packages [1,3
24 kB]
Hit:13 http://ppa.launchpad.net/graphics-drivers/ppa/ubuntu focal InRelease
Get:14 http://security.ubuntu.com/ubuntu focal-security/multiverse amd64 Packages
[28.5 kB]
Hit:15 http://ppa.launchpad.net/ubuntugis/ppa/ubuntu focal InRelease
Get:16 http://archive.ubuntu.com/ubuntu focal-updates/restricted amd64 Packages
[2,199 kB]
Get:17 http://archive.ubuntu.com/ubuntu focal-updates/multiverse amd64 Packages [3
1.3 kB]
Get:18 http://archive.ubuntu.com/ubuntu focal-updates/main amd64 Packages [3,069 k
B]
Fetched 10.0 MB in 2s (4,096 kB/s)
Reading package lists... Done
```

In [3]:
```
!pip install pyspark
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheel
s/public/simple/
Collecting pyspark
  Downloading pyspark-3.3.2.tar.gz (281.4 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 281.4/281.4 MB 5.4 MB/s eta 0:00:00
  Preparing metadata (setup.py) ... done
Collecting py4j==0.10.9.5
  Downloading py4j-0.10.9.5-py2.py3-none-any.whl (199 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 199.7/199.7 KB 22.8 MB/s eta 0:00:00
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) ... done
  Created wheel for pyspark: filename=pyspark-3.3.2-py2.py3-none-any.whl size=2818
24028 sha256=76652a3f491488c7404051da517552e34fb26885fbed0ca9a8640007f1098255
  Stored in directory: /root/.cache/pip/wheels/6c/e3/9b/0525ce8a69478916513509d436
93511463c6468db0de237c86
Successfully built pyspark
Installing collected packages: py4j, pyspark
  Attempting uninstall: py4j
    Found existing installation: py4j 0.10.9.7
    Uninstalling py4j-0.10.9.7:
      Successfully uninstalled py4j-0.10.9.7
Successfully installed py4j-0.10.9.5 pyspark-3.3.2
```

In [4]:
```
import os
os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-8-openjdk-amd64"
os.environ["SPARK_HOME"] = "/content/spark-3.3.2-bin-hadoop3"
```

In [5]:
```
import pyspark
spark = pyspark.sql.SparkSession.builder.appName("zomato-recommendation").getOrCrea
sc = spark.sparkContext
```

In [6]:
```
#Importing Libraries
import numpy as np
import pandas as pd
import seaborn as sb
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import LogisticRegression
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import r2_score
import warnings
warnings.filterwarnings('always')
warnings.filterwarnings('ignore')
import re
from nltk.corpus import stopwords
from sklearn.metrics.pairwise import linear_kernel
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
```

In [7]:
```
from google.colab import drive
drive.mount('/gdrive')
```

```
Mounted at /gdrive
```

In [67]:
```
data= spark.read.csv('/gdrive/My Drive/Big Data Framework/zomato.csv', inferSchema
```

In [70]:
```
data.show()
```

```
+--------------------+--------------------+--------------------+------------------
--+--------------------+-----+--------------------+--------------------+----------
----------+--------------------+--------------------+--------------------+--------
------------------+--------------------+--------------------+--------------------
+--------------------+
|                 url|             address|                name|        online_ord
er|          book_table| rate|               votes|               phone|
location|           rest_type|          dish_liked|            cuisines|approx_cos
t(for two people)|        reviews_list|           menu_item|     listed_in(type)|
listed_in(city)|
+--------------------+--------------------+--------------------+------------------
--+--------------------+-----+--------------------+--------------------+----------
----------+--------------------+--------------------+--------------------+--------
------------------+--------------------+--------------------+--------------------
+--------------------+
|https://www.zomat...|942, 21st Main Ro...|              Jalsa|                Y
es|                 Yes|4.1/5|                 775|        080 42297555|
null|                null|                null|                null|
null|                null|                null|                null|
null|
|     +91 9743772233"|        Banashankari|       Casual Dining|Pasta, Lunch Buf
f...|North Indian, Mug...|  800|"[('Rated 4.0', '...|        ('Rated 4.0'| 'RATED
\n  You ca...|        ('Rated 5.0'| 'RATED\n  Overde...|        ('Rated 4.0'|
'RATED\n  The pl...|        ('Rated 4.0'| 'RATED\n  The pl...|        ('Rated 4.
0'| 'RATED\n  The pl...|
|https://www.zomat...|2nd Floor, 80 Fee...|      Spice Elephant|                Y
es|                  No|4.1/5|                 787|        080 41714161|        Ba
nashankari|       Casual Dining|Momos, Lunch Buff...|Chinese, North In...|
800|"[('Rated 4.0', '...|rice was well coo...|        ('Rated 5.0'| 'RATED\n  This
p...|
|https://www.zomat...|1112, Next to KIM...|     San Churro Cafe|                Y
es|                  No|3.8/5|                 918|      +91 9663487993|        Ba
nashankari| Cafe, Casual Dining|Churros, Cannello...|Cafe, Mexican, It...|
800|"[('Rated 3.0', "...|        ('Rated 3.0'| ""RATED\n \nWent...| pasta churros
an...|
|https://www.zomat...|1st Floor, Annaku...|Addhuri Udupi Bho...|
No|                  No|3.7/5|                  88|      +91 9620009302|        Ba
nashankari|         Quick Bites|         Masala Dosa|South Indian, Nor...|
300|"[('Rated 4.0', "...|        ('Rated 2.0'| 'RATED\n  Reache...|        ('Rated
4.0'|
|https://www.zomat...|10, 3rd Floor, La...|       Grand Village|
No|                  No|3.8/5|                 166|      +91 8026612447|
null|                null|                null|                null|
null|                null|                null|                null|
null|
|     +91 9901210005"|        Basavanagudi|       Casual Dining| Panipuri, Gol Gap
pe|North Indian, Raj...|  600|[('Rated 4.0', 'R...|                  []|
Buffet|        Banashankari|                null|                null|
null|                null|                null|                null|
null|
|https://www.zomat...|37, 5-1, 4th Floo...|     Timepass Dinner|                Y
es|                  No|3.8/5|                 286|      +91 9980040002|
null|                null|                null|                null|
null|                null|                null|                null|
null|
|     +91 9980063005"|        Basavanagudi|       Casual Dining|Onion Rings, Pas
t...|        North Indian|  600|[('Rated 3.0', 'R...|                  []|
Buffet|        Banashankari|                null|                null|
null|                null|                null|                null|
null|
|https://www.zomat...|19/1, New Timbery...|Rosewood Internat...|
No|                  No|3.6/5|                   8|      +91 9731716688|
null|                null|                null|                null|
null|                null|                null|                null|
```

```
null|
|       080 26740366"|         Mysore Road|       Casual Dining|                nu
ll|North Indian, Sou...|  800|[('Rated 5.0', 'R...|                  []|
Buffet|        Banashankari|                null|                null|
null|                null|               null|                null|
null|
|https://www.zomat...|2469, 3rd Floor, ...|              Onesta|                 Y
es|                 Yes|4.6/5|                2556|        080 48653961|
null|                null|               null|                null|
null|                null|               null|                null|
null|
|       080 48655715"|        Banashankari| Casual Dining, Cafe|Farmhouse Pizza,
...|Pizza, Cafe, Italian|  600|"[('Rated 5.0', '...| and my 5th diffe...|
JP Nagar|        Basavanagudi| Koramangala baka...| it's the unlimit...|
but| but there is a c...| unlimited desser...| one dessert and ...| I have been th
ro...|
|https://www.zomat...|1, 30th Main Road...|      Penthouse Cafe|                 Y
es|                  No|4.0/5|                 324|      +91 8884135549|
null|                null|               null|                null|
null|                null|               null|                null|
null|
|     +91 9449449316"|        Banashankari|                Cafe|Pizza, Mocktail
s,...|Cafe, Italian, Co...|  700|"[('Rated 3.0', "...| it's a very smal...|
('Rated 4.0'|    ""RATED\n  Small|                cosy| covered rooftop ...|
pasta| pizza and sizzle...|        ('Rated 4.0'| ""RATED\n  Small...|        ('Rat
ed 4.0'|
|https://www.zomat...|2470, 21 Main Roa...|           Smacznego|                 Y
es|                  No|4.2/5|                 504|      +91 9945230807|
null|                null|               null|                null|
null|                null|               null|                null|
null|
|     +91 9743804471"|        Banashankari|                Cafe|Waffles, Pasta,
C...|Cafe, Mexican, It...|  550|"[('Rated 4.0', "...|        ('Rated 4.0'| 'RATED
\n  A tuck...| slightly off the...| they served in f...|        ('Rated 5.0'|
'RATED\n  Been h...| although takes a...| do taste quite y...|        ('Rated 5.
0'| 'RATED\n  Being ...|
|https://www.zomat...|12,29 Near PES Un...|CafÃÂÃÂÃÂÃÂÃ...|                 Y
es|                  No|4.1/5|                 402|        080 26724489|
null|                null|               null|                null|
null|                null|               null|                null|
null|
|     +91 7406048982"|        Banashankari|                Cafe|Waffles, Pasta,
C...|                Cafe|  500|"[('Rated 4.0', '...| I wish they gave...| pasta a
rrabiata ...| waffles and nugg...| a good hangout p...| I wouldn't sugge...|
('Rated 3.0'| 'RATED\n  A good...| and this place c...|                  []|
Cafes|
|https://www.zomat...|941, 3rd FLOOR, 2...|        Cafe Shuffle|                 Y
es|                 Yes|4.2/5|                 150|      +91 9742166777|        Ba
nashankari|                Cafe|Mocktails, Peri F...|Cafe, Italian, Co...|
600|"[('Rated 1.0', "...| you get it liter...|        ('Rated 4.0'| ""RATED\n  Whi
le...|
+--------------------+--------------------+--------------------+------------------
--+--------------------+-----+--------------------+--------------------+----------
----------+--------------------+--------------------+--------------------+--------
--------------------+--------------------+--------------------+--------------------
+--------------------+
only showing top 20 rows
```

In [71]:
```
data.columns
```

Out[71]:
```
['url',
 'address',
 'name',
 'online_order',
 'book_table',
 'rate',
 'votes',
 'phone',
 'location',
 'rest_type',
 'dish_liked',
 'cuisines',
 'approx_cost(for two people)',
 'reviews_list',
 'menu_item',
 'listed_in(type)',
 'listed_in(city)']
```

In [72]:
```
data.printSchema()
```

```
root
 |-- url: string (nullable = true)
 |-- address: string (nullable = true)
 |-- name: string (nullable = true)
 |-- online_order: string (nullable = true)
 |-- book_table: string (nullable = true)
 |-- rate: string (nullable = true)
 |-- votes: string (nullable = true)
 |-- phone: string (nullable = true)
 |-- location: string (nullable = true)
 |-- rest_type: string (nullable = true)
 |-- dish_liked: string (nullable = true)
 |-- cuisines: string (nullable = true)
 |-- approx_cost(for two people): string (nullable = true)
 |-- reviews_list: string (nullable = true)
 |-- menu_item: string (nullable = true)
 |-- listed_in(type): string (nullable = true)
 |-- listed_in(city): string (nullable = true)
```

In [73]:
```
#Deleting Unnnecessary Columns
from pyspark.sql.functions import col
zomato=data.drop("url", "dish_liked", "phone")
```

In [74]:
```
from pyspark.sql.functions import count, when
count_all = zomato.count()
count_distinct = zomato.distinct().count()
count_duplicates = count_all - count_distinct
print(count_duplicates)
zomato = zomato.dropDuplicates()
```

```
25720
```

In [75]:
```
from pyspark.sql.functions import col, count
null_counts = zomato.agg(*[count(col(c)).alias(c) for c in zomato.columns])
null_counts.show()
zomato = zomato.dropna(how='any')
zomato.printSchema()
```

```
+-------+-----+------------+----------+-----+-----+--------+---------+--------+---
------------------------+------------+---------+---------------+---------------+
|address| name|online_order|book_table| rate|votes|location|rest_type|cuisines|app
rox_cost(for two people)|reviews_list|menu_item|listed_in(type)|listed_in(city)|
+-------+-----+------------+----------+-----+-----+--------+---------+--------+---
------------------------+------------+---------+---------------+---------------+
|  46010|45942|       39587|     46008|39666|46010|   37248|    37152|   30146|
29552|       29390|    28967|          28595|          28241|
+-------+-----+------------+----------+-----+-----+--------+---------+--------+---
------------------------+------------+---------+---------------+---------------+

root
 |-- address: string (nullable = true)
 |-- name: string (nullable = true)
 |-- online_order: string (nullable = true)
 |-- book_table: string (nullable = true)
 |-- rate: string (nullable = true)
 |-- votes: string (nullable = true)
 |-- location: string (nullable = true)
 |-- rest_type: string (nullable = true)
 |-- cuisines: string (nullable = true)
 |-- approx_cost(for two people): string (nullable = true)
 |-- reviews_list: string (nullable = true)
 |-- menu_item: string (nullable = true)
 |-- listed_in(type): string (nullable = true)
 |-- listed_in(city): string (nullable = true)
```

In [76]:
```
zomato.columns
```

Out[76]:
```
['address',
 'name',
 'online_order',
 'book_table',
 'rate',
 'votes',
 'location',
 'rest_type',
 'cuisines',
 'approx_cost(for two people)',
 'reviews_list',
 'menu_item',
 'listed_in(type)',
 'listed_in(city)']
```

In [77]:
```
zomato = zomato.withColumnRenamed("approx_cost(for two people)", "cost") \
               .withColumnRenamed("listed_in(type)", "type") \
               .withColumnRenamed("listed_in(city)", "city")

print(zomato.columns)
```

```
['address', 'name', 'online_order', 'book_table', 'rate', 'votes', 'location', 're
st_type', 'cuisines', 'cost', 'reviews_list', 'menu_item', 'type', 'city']
```

In [78]:
```
from pyspark.sql.functions import  regexp_replace
zomato = zomato.withColumn("cost", col("cost").cast("string"))
zomato = zomato.withColumn("cost", regexp_replace(col("cost"), ",", "."))
zomato = zomato.withColumn("cost", col("cost").cast("float"))
zomato.printSchema()
```

```
root
 |-- address: string (nullable = true)
 |-- name: string (nullable = true)
 |-- online_order: string (nullable = true)
 |-- book_table: string (nullable = true)
 |-- rate: string (nullable = true)
 |-- votes: string (nullable = true)
 |-- location: string (nullable = true)
 |-- rest_type: string (nullable = true)
 |-- cuisines: string (nullable = true)
 |-- cost: float (nullable = true)
 |-- reviews_list: string (nullable = true)
 |-- menu_item: string (nullable = true)
 |-- type: string (nullable = true)
 |-- city: string (nullable = true)
```

In [79]:
```
zomato.select("rate").distinct().show()
```

```
+------+
|  rate|
+------+
|   800|
| 3.8/5|
|   700|
|   200|
| 2,000|
| 1,400|
| 2.2/5|
|   250|
| 2,100|
|4.0 /5|
| 4.9/5|
| 1,700|
| 1,800|
|4.9 /5|
| 1,600|
| 1,100|
| 3.3/5|
| 2.4/5|
| 4.2/5|
|3.5 /5|
+------+
only showing top 20 rows
```

In [80]:
```
zomato = zomato.filter(col("rate") != "NEW")
zomato = zomato.filter(col("rate") != "-").na.drop()
zomato = zomato.withColumn("rate", regexp_replace(col("rate"), "/5", ""))
zomato = zomato.withColumn("rate", col("rate").cast("float"))
zomato.select("rate").show()
```

```
+----+
|rate|
+----+
| 3.4|
| 3.4|
| 3.6|
| 3.9|
| 4.2|
| 3.6|
| 2.7|
| 3.3|
| 3.7|
| 4.1|
| 4.0|
| 3.9|
| 3.8|
| 4.0|
| 3.0|
| 3.9|
| 4.1|
| 3.3|
| 4.4|
| 3.7|
+----+
only showing top 20 rows
```

In [81]:
```
from pyspark.sql.functions import col, initcap

zomato = zomato.withColumn("name", initcap(col("name")))
zomato = zomato.withColumn("online_order", when(col("online_order") == "Yes", True
zomato = zomato.withColumn("book_table", when(col("book_table") == "Yes", True).otl
zomato.select("cost").distinct().show()
```

```
+-----+
| cost|
+-----+
|550.0|
|500.0|
|180.0|
|  2.5|
|350.0|
|  2.2|
|  3.4|
|100.0|
|  2.0|
|  1.8|
|  3.2|
| 1.45|
|  3.0|
| 70.0|
|  1.5|
| 40.0|
|250.0|
|  1.1|
|  2.8|
|400.0|
+-----+
only showing top 20 rows
```

In [55]:
```
zomato.show(20)
```

```
+--------------------+--------------------+------------+----------+----+-----+----
----------------+--------------------+--------------------+-----+----------------
---+--------------------+--------------------+--------------------+
|             address|                name|online_order|book_table|rate|votes|
location|           rest_type|             cuisines| cost|        reviews_list|
menu_item|                type|                city|
+--------------------+--------------------+------------+----------+----+-----+----
----------------+--------------------+--------------------+-----+----------------
---+--------------------+--------------------+--------------------+
|715, Hari Complex...|Karavali Fish Center|       false|     false| 3.4|   10|
Banashankari|         Quick Bites|Mangalorean, Chinese|300.0|"[('Rated 4.0', '...|
This place serve...|        ('Rated 3.0'| 'RATED\n  Ambien...|
|Shop 984, Btm 2nd...|          Lassi Berg|        true|     false| 3.4|   13|
BTM|            Takeaway|              Juices|100.0|[('Rated 5.0', 'R...|
[]|           Delivery|   Bannerghatta Road|
|25/26,1st Cross, ...|            Kuttanad|        true|     false| 3.6|  140|   B
annerghatta Road|         Quick Bites|Kerala, South Indian|200.0|"[('Rated 3.0',
'...|                  []|            Dine-out|   Bannerghatta Road|
|Hi Street Mall, 1...|           Keventers|        true|     false| 3.9|  210|
Jayanagar|Beverage Shop, De...|Desserts, Beverag...|400.0|"[('Rated 4.0', "...|
('Rated 1.0'| ""RATED\n  Could...|        ('Rated 3.0'|
|2477, 24th Cross ...|       Frozen Bottle|        true|     false| 4.2|  146|
Banashankari|       Beverage Shop|Beverages, Desser...|400.0|"[('Rated 4.0', '...|
[]|           Delivery|        Basavanagudi|
|28 East End Main ...|Jayanagara Donne ...|        true|     false| 3.6|   60|
Jayanagar|         Quick Bites|Biryani, South In...|400.0|"[('Rated 4.0', "...|
('Rated 1.0'| 'RATED\n  (outda...| never order anyt...|
|C.K.B Layout, Nea...|     Mahek Of Punjab|        true|     false| 2.7|   75|
Marathahalli|         Quick Bites|        North Indian|400.0|"[('Rated 2.0', "...|
and ended up ord...| and quite fast. ...| with too much on...|
|95, Vydehi Hospit...|        Estaa Sweets|        true|     false| 3.3|    5|
Whitefield|          Sweet Shop|              Mithai|300.0|[('Rated 1.0', 'R...|
['Premium Dry Fru...|           Delivery|         Brookefield|
|58, Ground Floor,...|            Tata Cha|        true|     false| 3.7|   31|Kora
mangala 7th B...|               Cafe|Cafe, Tea, North ...|500.0|"[('Rated 4.0',
"...| and cranks it up...|        ('Rated 4.0'| ""RATED\n  The a...|
|70/1, 4th Cross, ...| Tapri By The Corner|        true|     false| 4.1|  205|Kora
mangala 5th B...|         Quick Bites|           Fast Food|250.0|"[('Rated 3.0',
"...| however been loc...| which was modera...| however not pick...|
|1016, 1st Floor, ...|         Koshe Kosha|        true|     false| 4.0|  571|Kora
mangala 1st B...|       Casual Dining|    Bengali, Seafood|600.0|"[('Rated 4.0',
'...| honestly there a...| considering thei...|        ('Rated 4.0'|
|100 Feet Ring Roa...|   Empire Restaurant|        true|     false| 3.9|   95|
BTM|  Takeaway, Delivery|Kerala, Seafood, ...|400.0|[('Rated 3.0', 'R...|
[]|           Delivery|                 BTM|
|353/354, 1st A cr...|Begin With Us.......|        true|     false| 3.8|   36|Kora
mangala 7th B...|       Beverage Shop|      Beverages, Tea|300.0|"[('Rated 4.0',
'...| Smoor and Krispy...| we could see tha...| can visit this p...|
|20/A, 3rd Floor, ...|      Little Lucknow|        true|      true| 4.0|  235|Kora
mangala 5th B...|       Casual Dining|Mughlai, North In...|800.0|"[('Rated 3.0',
'...| food is very nic...| in a nice thin w...| it would've been...|
|1/84, Ground Floo...|Invitation Bar & ...|       false|     false| 3.0|   15|
BTM|                 Bar|        North Indian|  1.0|[('Rated 2.0', 'R...|
[]|           Delivery|                 BTM|
|Sector 6, HSR Lay...|          #1-81 Cafe|        true|     false| 3.9|   48|
HSR|         Quick Bites|Fast Food, Beverages|400.0|"[('Rated 4.0', '...| just cam
e here b...| especially comfo...|        ('Rated 5.0'|
|39, 15th Cross, B...|Two Friends Cauldron|        true|      true| 4.1|  454|
JP Nagar|       Casual Dining|Continental, Ital...|700.0|"[('Rated 4.0', '...|
Flying Faith|           Flit Flat|        Mumble Munch|
|3B, 1st Main, Kor...|Banashankari Donn...|       false|     false| 3.3|    4|Kora
mangala 7th B...|         Quick Bites|South Indian, Bir...|300.0|
[]|                  []|            Dine-out|                 BTM|
|JW Marriott, 24/1...|Jw Kitchen - Jw M...|       false|      true| 4.4| 2119|
```

```
Lavelle Road|         Fine Dining|North Indian, Con...|  2.2|"[('Rated 4.0', '...|
('Rated 4.0'| 'RATED\n  5 star...|        ('Rated 5.0'|
|Keonic, 5th Main ...|            Varenyam|       false|     false| 3.7|   56|
Electronic City|       Casual Dining|South Indian, Nor...|500.0|"[('Rated 1.0',
'...| it's your normal...|      kadhai chicken| butter kulcha an...|
+--------------------+--------------------+------------+----------+----+-----+----
---------------+--------------------+--------------------+-----+----------------
---+--------------------+--------------------+--------------------+
only showing top 20 rows
```

In [56]:
```
zomato.select("location").distinct().show()
```

```
+--------------------+
|            location|
+--------------------+
|           Bellandur|
|      East Bangalore|
|         Indiranagar|
|                 BTM|
|        Banashankari|
|Koramangala 7th B...|
|             JP Nagar|
|        Lavelle Road|
|        Kammanahalli|
|Koramangala 3rd B...|
|Koramangala 2nd B...|
|      St. Marks Road|
|            Majestic|
|ITPL Main Road, W...|
|           Jayanagar|
|        Brigade Road|
|     Electronic City|
|         Frazer Town|
|       Church Street|
|                 HSR|
+--------------------+
only showing top 20 rows
```

In [83]:
```
from pyspark.sql.functions import mean
from pyspark.sql.window import Window
from pyspark.sql.functions import col

window = Window.partitionBy("name")
zomato = zomato.withColumn("Mean Rating", mean(col("rate")).over(window))
```

In [84]:
```
zomato.show(10)
```

```
+-------------------+-------------------+-----------+----------+----+-----+----
----------------+------------+-------------------+-----+----------------------+---
-----------------+-------------------+-------------------+------------------+
|            address|               name|online_order|book_table|rate|votes|
location|     rest_type|           cuisines| cost|          reviews_list|
menu_item|                 type|                 city|       Mean Rating|
+-------------------+-------------------+-----------+----------+----+-----+----
----------------+------------+-------------------+-----+----------------------+---
-----------------+-------------------+-------------------+------------------+
|Sector 6, HSR Lay...|          #l-81 Cafe|        true|     false| 3.9|   48|
HSR|  Quick Bites|Fast Food, Beverages|400.0|"[('Rated 4.0', '...| just came here
b...| especially comfo...|        ('Rated 5.0'|3.9000000953674316|
|Sector 6, HSR Lay...|          #l-81 Cafe|        true|     false| 3.9|   48|
HSR|  Quick Bites|Fast Food, Beverages|400.0|"[('Rated 4.0', '...| just came here
b...| especially comfo...|        ('Rated 5.0'|3.9000000953674316|
|Sector 6, HSR Lay...|          #l-81 Cafe|        true|     false| 3.9|   48|
HSR|  Quick Bites|Fast Food, Beverages|400.0|"[('Rated 4.0', '...| just came here
b...| especially comfo...|        ('Rated 5.0'|3.9000000953674316|
|16, 17th A Main, ...|            1000 B.c|        true|     false| 3.2|   49|Kora
mangala 5th B...|  Quick Bites|Arabian, Sandwich...|300.0|"[('Rated 1.0', "...| pl
ease don't put...|        ('Rated 4.0'| ""RATED\n  The t...| 3.200000047683716|
|16, 17th A Main, ...|            1000 B.c|        true|     false| 3.2|   49|Kora
mangala 5th B...|  Quick Bites|Arabian, Sandwich...|300.0|"[('Rated 1.0', "...| pl
ease don't put...|        ('Rated 4.0'| ""RATED\n  The t...| 3.200000047683716|
|688, Thanish Corn...|100ãâãâãâãâãâãâãâ...|       false|     false| 3.7|   41|
BTM|Casual Dining|Biryani, North In...|450.0|"[('Rated 1.0', '...| we specially en
j...| We'll definitely...|        ('Rated 4.5'| 3.700000047683716|
|688, Thanish Corn...|100ãâãâãâãâãâãâãâ...|       false|     false| 3.7|   41|
BTM|Casual Dining|Biryani, North In...|450.0|"[('Rated 1.0', '...| we specially en
j...| We'll definitely...|        ('Rated 4.5'| 3.700000047683716|
|688, Thanish Corn...|100ãâãâãâãâãâãâãâ...|       false|     false| 3.7|   41|
BTM|Casual Dining|Biryani, North In...|450.0|"[('Rated 1.0', '...| we specially en
j...| We'll definitely...|        ('Rated 4.5'| 3.700000047683716|
|Near Velankani Bu...|11 To 11 Express ...|        true|     false| 3.5|   22|
Electronic City|  Quick Bites|      Biryani, Kebab|300.0|[('Rated 1.0', 'R...|
[]|             Delivery|     Electronic City|               3.5|
|Near Velankani Bu...|11 To 11 Express ...|        true|     false| 3.5|   22|
Electronic City|  Quick Bites|      Biryani, Kebab|300.0|[('Rated 1.0', 'R...|
[]|             Dine-out|     Electronic City|               3.5|
+-------------------+-------------------+-----------+----------+----+-----+----
----------------+------------+-------------------+-----+----------------------+---
-----------------+-------------------+-------------------+------------------+
only showing top 10 rows
```

In [86]:

```python
from pyspark.ml.linalg import Vectors
from pyspark.sql.functions import udf
from pyspark.sql.types import DoubleType
from pyspark.ml.linalg import VectorUDT
to_vector = udf(lambda x: Vectors.dense(x), VectorUDT())

zomato = zomato.withColumn("Mean Rating", to_vector(col("Mean Rating")))
```

In [91]:

```python
from pyspark.ml.feature import MinMaxScaler
from pyspark.ml.linalg import Vectors
from pyspark.sql.functions import col, udf
from pyspark.sql.types import DoubleType

vector_to_double = udf(lambda x: float(x[0]), DoubleType())
scaler = MinMaxScaler(inputCol="Mean Rating", outputCol="Scaled Mean Rating", min=
scalerModel = scaler.fit(zomato)
zomato = scalerModel.transform(zomato)
zomato = zomato.withColumn("Scaled Mean Rating", vector_to_double(col("Scaled Mean
```

```
zomato = zomato.withColumn("Scaled Mean Rating", F.round(col("Scaled Mean Rating")
zomato.sample(0.1).show(3)
```

```
+--------------------+--------------+-----------+----------+----+-----+---------
-----------+-------------+--------------------+-----+--------------------+--------
------------+--------------------+--------------------+--------------------+-------
-----------+
|             address|          name|online_order|book_table|rate|votes|
location|    rest_type|             cuisines| cost|        reviews_list|
menu_item|                type|                city|        Mean Rating|Scaled Mea
n Rating|
+--------------------+--------------+-----------+----------+----+-----+---------
-----------+-------------+--------------------+-----+--------------------+--------
------------+--------------------+--------------------+--------------------+-------
-----------+
|3, 80 Feet Road, ...| 1522 - The Pub|       false|      true| 4.2| 1743|Koramanga
la 4th B...|          Pub|Chinese, Continen...|  1.4|"[('Rated 5.0', '...| peanuts
and chic...|        ('Rated 4.0'| 'RATED\n  A good...|[4.199999809265137]|
4.0|
|11, Smondoville N...|       4foodiez|        true|     false| 3.6|   82|     Elec
tronic City|  Quick Bites|South Indian, Nor...|300.0|[('Rated 4.0', 'R...|
[]|            Dine-out|     Electronic City|[3.6499999761581421]|              3.2
1|
|107/P4, Ground Fl...|A Southern Fair|        true|     false| 3.0|   19|
Whitefield|Casual Dining|Biryani, Seafood,...|750.0|"[('Rated 3.0', '...| the seer
fish fr...| because it shows...|                  []|[3.1000000238418579]|
2.43|
+--------------------+--------------+-----------+----------+----+-----+---------
-----------+-------------+--------------------+-----+--------------------+--------
------------+--------------------+--------------------+--------------------+-------
-----------+
only showing top 3 rows
```

In [92]:
```
zomato.select('reviews_list', 'cuisines').sample(False, 0.05)
```

Out[92]:
```
DataFrame[reviews_list: string, cuisines: string]
```

In [93]:
```
from pyspark.sql.functions import lower

zomato = zomato.withColumn("reviews_list", lower(zomato["reviews_list"]))
zomato.select('reviews_list', 'cuisines').sample(False, 0.05)
```

Out[93]:
```
DataFrame[reviews_list: string, cuisines: string]
```

In [94]:
```
from pyspark.sql.functions import udf
from pyspark.sql.types import StringType
import string

PUNCT_TO_REMOVE = string.punctuation

def remove_punctuation(text):
    return text.translate(str.maketrans('', '', PUNCT_TO_REMOVE))

remove_punctuation_udf = udf(remove_punctuation, StringType())

zomato = zomato.withColumn("reviews_list", remove_punctuation_udf(zomato["reviews_
zomato.select('reviews_list', 'cuisines').sample(False, 0.05)
```

Out[94]:
```
DataFrame[reviews_list: string, cuisines: string]
```

In [96]:
```
import nltk
nltk.download('stopwords')
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
```

Out[96]:
```
True
```

In [99]:
```
from pyspark.ml.feature import StopWordsRemover
from pyspark.sql.functions import udf
from pyspark.sql.types import StringType, ArrayType
from nltk.corpus import stopwords

STOPWORDS = set(stopwords.words('english'))

def remove_stopwords(text):
    return " ".join([word for word in str(text).split() if word not in STOPWORDS])

remove_stopwords_udf = udf(remove_stopwords, StringType())

split_udf = udf(lambda x: x.split(), ArrayType(StringType()))

zomato = zomato.withColumn('reviews_list_split', split_udf('reviews_list'))

remover = StopWordsRemover(inputCol="reviews_list_split", outputCol="reviews_list_
zomato = remover.transform(zomato)

zomato = zomato.withColumn("reviews_list", remove_stopwords_udf(zomato["reviews_li
            .drop("reviews_list_filtered", "reviews_list_split")
```

In [100…
```
import re
from pyspark.sql.functions import udf
from pyspark.sql.types import StringType

def remove_urls(text):
    url_pattern = re.compile(r'https?://\S+|www\.\S+')
    return url_pattern.sub(r'', text)

remove_urls_udf = udf(remove_urls, StringType())

zomato = zomato.withColumn("reviews_list", remove_urls_udf(zomato["reviews_list"]))
```

In [101…
```
zomato.select('reviews_list', 'cuisines').sample(False, 0.05)
```

Out[101]:
```
DataFrame[reviews_list: string, cuisines: string]
```

In [104…
```
from pyspark.ml.feature import CountVectorizer
from pyspark.sql.functions import udf,lit,array
from pyspark.sql.types import ArrayType, StructType, StructField, StringType, Integ

def get_top_words(column, top_nu_of_words, nu_of_word):
    vec = CountVectorizer(ngram_range=nu_of_word, stopWords='english')
    bag_of_words = vec.fit(column).transform(column)
    sum_words = bag_of_words.sum(axis=0)
    words_freq = [(word, sum_words[0, idx]) for word, idx in vec.vocabulary_.items
    words_freq = sorted(words_freq, key=lambda x: x[1], reverse=True)
    return words_freq[:top_nu_of_words]

get_top_words_udf = udf(get_top_words, ArrayType(StructType([
    StructField("word", StringType(), True),
    StructField("frequency", IntegerType(), True)
])))

zomato = zomato.withColumn("top_words", get_top_words_udf(zomato["reviews_list"],
```

In [113…
```
from pyspark.sql.functions import col
zomato = zomato.drop(*['address', 'rest_type', 'type', 'menu_item', 'votes'])
```

In [115…
```
print("Shape of dataframe: ", (zomato.count(), len(zomato.columns)))
print("Columns in dataframe: ", zomato.columns)
```

```
Shape of dataframe:  (16186, 12)
Columns in dataframe:  ['name', 'online_order', 'book_table', 'rate', 'location',
'cuisines', 'cost', 'reviews_list', 'city', 'Mean Rating', 'Scaled Mean Rating',
'top_words']
```

In [116…
```
from pyspark.sql.functions import rand

df_percent = zomato.orderBy(rand()).limit(int(zomato.count() * 0.5))
```

In [117…
```
print("Shape of dataframe: ", (df_percent.count(), len(df_percent.columns)))
```

```
Shape of dataframe:  (8093, 12)
```

In [123…
```
df_percent = df_percent.withColumn("name", F.col("name").cast("string"))
df_percent = df_percent.select("*").orderBy("name")
df_percent.createOrReplaceTempView("df_percent")
df_percent = spark.sql("SELECT * FROM df_percent ORDER BY name")
df_percent = df_percent.withColumnRenamed("name", "index").withColumnRenamed("perc
```

In [124…
```
from pyspark.sql.functions import col

indices = df_percent.select(col("index")).rdd.flatMap(lambda x: x).collect()
```

In [125…
```
from pyspark.ml.feature import HashingTF, IDF, Tokenizer

tokenizer = Tokenizer(inputCol="reviews_list", outputCol="words")
wordsData = tokenizer.transform(df_percent)

hashingTF = HashingTF(inputCol="words", outputCol="rawFeatures", numFeatures=20)
featurizedData = hashingTF.transform(wordsData)

idf = IDF(inputCol="rawFeatures", outputCol="features")
idfModel = idf.fit(featurizedData)
tfidf_matrix = idfModel.transform(featurizedData).select("features")
```

In [126…
```
from pyspark.ml.feature import Normalizer

normalizer = Normalizer(inputCol="features", outputCol="norm")
data = normalizer.transform(tfidf_matrix)

cosine_similarities = data.rdd.cartesian(data.rdd).map(lambda x: (x[0][0], x[1][0]
```

In [136…
```
```

In [161…
```
from pyspark.sql.functions import col

def recommend(name, cosine_similarities=cosine_similarities):
    recommend_restaurant = []
    idx = indices.index(name)
    score_series = pd.Series(cosine_similarities[idx]).sort_values(ascending=False

    top30_indexes = list(score_series.iloc[0:31].index)

    for each in top30_indexes:
        recommend_restaurant.append(list(df_percent.index)[each])
```

```
    df_new = spark.createDataFrame([], ['cuisines', 'Mean Rating', 'cost'])

    for each in recommend_restaurant:
        df_new = df_new.union(df_percent.select(['cuisines','Mean Rating', 'cost']

    df_new = df_new.dropDuplicates(['cuisines','Mean Rating', 'cost'])
    df_new = df_new.orderBy(col('Mean Rating').desc()).limit(10)

    print('TOP %s RESTAURANTS LIKE %s WITH SIMILAR REVIEWS: ' % (str(df_new.count(

    return df_new
```

In [ ]:
```
recommend('Pai Vihar')
```