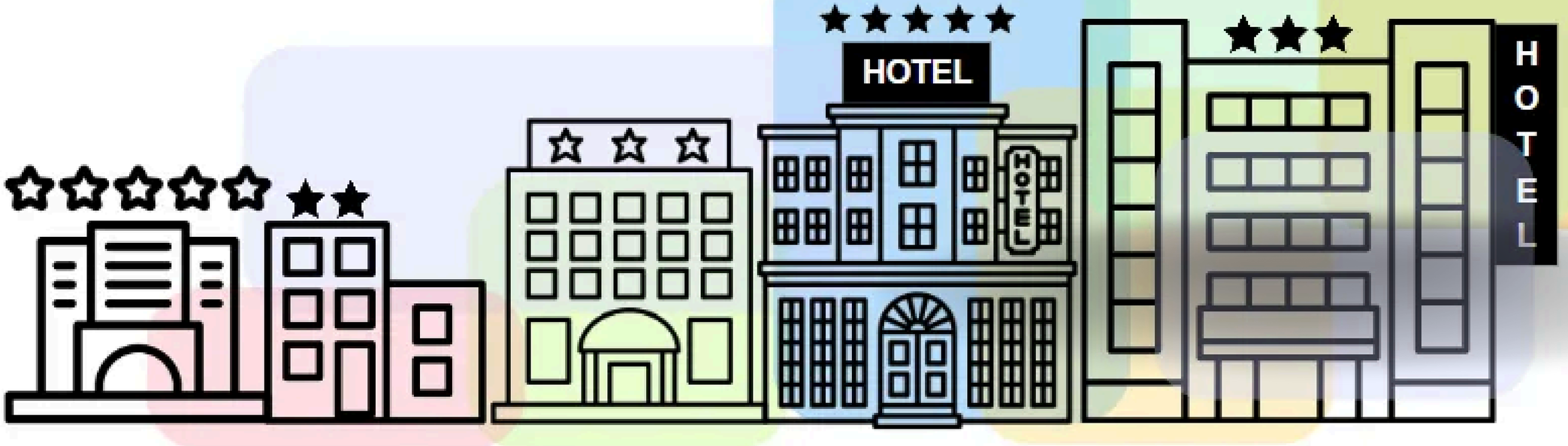
a Live Project on

Hotel Review Model

Under the guidance of Mr. Ambrish Soni
Mentor : Mr. Kartik Tyagi

Team Members

Prakriti (Leader)
Sumanyu Harjai
Harshit Malik
Vansh Imlani
Anmol Aggarwal
Harish Kumar
Kalpana Bhardwaj



Problem Statement:

- Users often rely on online hotel reviews to make decisions. However, manually analyzing and understanding the sentiment of reviews is time-consuming and challenging due to the high volume of data.

Goal:

- To build a machine learning model capable of predicting hotel review ratings based on textual data.

Objective:

- To train a model that predicts ratings (1-5 stars) for reviews using a 20% test dataset to evaluate performance.
- To assist users in quickly understanding the overall sentiment and quality of reviews

About the dataset

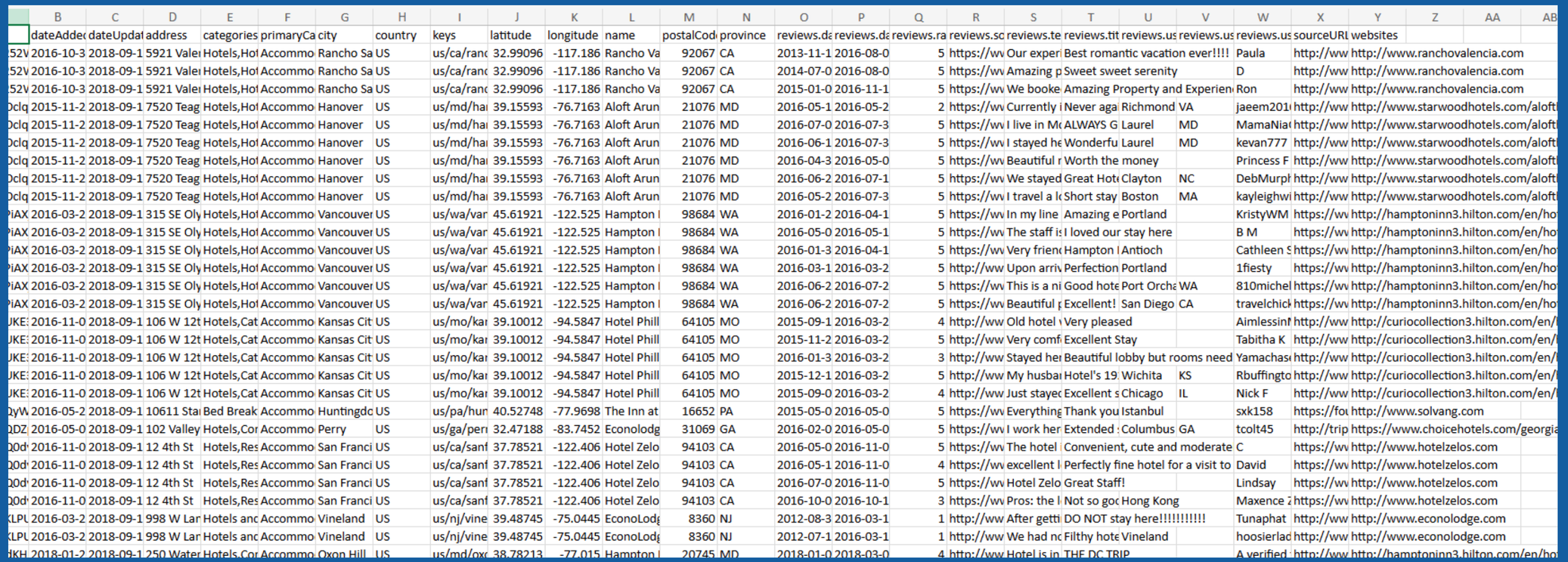
No. of rows- 10000

No of columns- 24

id, dateAdded , dateUpdated, address, categories , primaryCategories, city, country, keys, latitude, longitude, name, postalCode, province, reviews.date, reviews.dateSeen, reviews.rating, reviews.sourceURLs, reviews.text, reviews.title, reviews.userCity, reviews.userProvince, reviews.username, sourceURLs, websites

Null Values

reviews.text	1
reviews.title	1
reviews.userCity	5836
reviews.userProvince	7295



	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
	dateAdde	dateUpdat	address	categories	primaryCa	city	country	keys	latitude	longitude	name	postalCod	province	reviews.da	reviews.da	reviews.ra	reviews.sc	reviews.te	reviews.tit	reviews.us	reviews.us	reviews.us	sourceURL	websites			
	2016-10-3	2018-09-1	5921 Vale	Hotels,Hol	Accommo	Rancho Sa	US	us/ca/ranc	32.99096	-117.186	Rancho Va	92067	CA	2013-11-1	2016-08-0	5	https://ww	Our exper	Best romantic vacation ever!!!!			Paula	http://ww	http://www.ranchovalencia.com			
	2016-10-3	2018-09-1	5921 Vale	Hotels,Hol	Accommo	Rancho Sa	US	us/ca/ranc	32.99096	-117.186	Rancho Va	92067	CA	2014-07-0	2016-08-0	5	https://ww	Amazing p	Sweet sweet serenity			D	http://ww	http://www.ranchovalencia.com			
	2016-10-3	2018-09-1	5921 Vale	Hotels,Hol	Accommo	Rancho Sa	US	us/ca/ranc	32.99096	-117.186	Rancho Va	92067	CA	2015-01-0	2016-11-1	5	https://ww	We booke	Amazing Property and Experien			Ron	http://ww	http://www.ranchovalencia.com			
	2015-11-2	2018-09-1	7520 Teag	Hotels,Hol	Accommo	Hanover	US	us/md/hai	39.15593	-76.7163	Aloft Arun	21076	MD	2016-05-1	2016-05-2	2	https://ww	Currently i	Never aga	Richmond	VA		jaeem201	http://ww	http://www.starwoodhotels.com/aloft			
	2015-11-2	2018-09-1	7520 Teag	Hotels,Hol	Accommo	Hanover	US	us/md/hai	39.15593	-76.7163	Aloft Arun	21076	MD	2016-07-0	2016-07-3	5	https://ww	I live in M	ALWAYS G	Laurel	MD		MamaNia	http://ww	http://www.starwoodhotels.com/aloft			
	2015-11-2	2018-09-1	7520 Teag	Hotels,Hol	Accommo	Hanover	US	us/md/hai	39.15593	-76.7163	Aloft Arun	21076	MD	2016-06-1	2016-07-3	5	https://ww	I stayed he	Wonderfu	Laurel	MD		kevan777	http://ww	http://www.starwoodhotels.com/aloft			
	2015-11-2	2018-09-1	7520 Teag	Hotels,Hol	Accommo	Hanover	US	us/md/hai	39.15593	-76.7163	Aloft Arun	21076	MD	2016-04-3	2016-05-0	5	https://ww	Beautiful r	Worth the money				Princess F	http://ww	http://www.starwoodhotels.com/aloft			
	2015-11-2	2018-09-1	7520 Teag	Hotels,Hol	Accommo	Hanover	US	us/md/hai	39.15593	-76.7163	Aloft Arun	21076	MD	2016-06-2	2016-07-1	5	https://ww	We stayed	Great Hot	Clayton	NC		DebMurph	http://ww	http://www.starwoodhotels.com/aloft			
	2015-11-2	2018-09-1	7520 Teag	Hotels,Hol	Accommo	Hanover	US	us/md/hai	39.15593	-76.7163	Aloft Arun	21076	MD	2016-05-2	2016-07-3	5	https://ww	I travel a l	Short stay	Boston	MA		kayleighwi	http://ww	http://www.starwoodhotels.com/aloft			
	2016-03-2	2018-09-1	315 SE Oly	Hotels,Hol	Accommo	Vancouver	US	us/wa/var	45.61921	-122.525	Hampton I	98684	WA	2016-01-2	2016-04-1	5	https://ww	In my line	Amazing e	Portland			KristyWM	https://ww	http://hamptoninn3.hilton.com/en/ho			
	2016-03-2	2018-09-1	315 SE Oly	Hotels,Hol	Accommo	Vancouver	US	us/wa/var	45.61921	-122.525	Hampton I	98684	WA	2016-05-0	2016-05-1	5	https://ww	The staff is	I loved our stay here				B M	https://ww	http://hamptoninn3.hilton.com/en/ho			
	2016-03-2	2018-09-1	315 SE Oly	Hotels,Hol	Accommo	Vancouver	US	us/wa/var	45.61921	-122.525	Hampton I	98684	WA	2016-01-3	2016-04-1	5	https://ww	Very frien	Hampton I	Antioch			Cathleen S	https://ww	http://hamptoninn3.hilton.com/en/ho			
	2016-03-2	2018-09-1	315 SE Oly	Hotels,Hol	Accommo	Vancouver	US	us/wa/var	45.61921	-122.525	Hampton I	98684	WA	2016-03-1	2016-03-2	5	http://ww	Upon arriv	Perfection	Portland			1fiesty	https://ww	http://hamptoninn3.hilton.com/en/ho			
	2016-03-2	2018-09-1	315 SE Oly	Hotels,Hol	Accommo	Vancouver	US	us/wa/var	45.61921	-122.525	Hampton I	98684	WA	2016-06-2	2016-07-2	5	https://ww	This is a ni	Good hote	Port Orcha	WA		810michel	https://ww	http://hamptoninn3.hilton.com/en/ho			
	2016-03-2	2018-09-1	315 SE Oly	Hotels,Hol	Accommo	Vancouver	US	us/wa/var	45.61921	-122.525	Hampton I	98684	WA	2016-06-2	2016-07-2	5	https://ww	Beautiful p	Excellent!	San Diego	CA		travelchick	https://ww	http://hamptoninn3.hilton.com/en/ho			
	2016-11-0	2018-09-1	106 W 12t	Hotels,Cat	Accommo	Kansas Cit	US	us/mo/kar	39.10012	-94.5847	Hotel Phill	64105	MO	2015-09-1	2016-03-2	4	http://ww	Old hotel v	Very pleased				Aimlessin	http://ww	http://curiocollection3.hilton.com/en/			
	2016-11-0	2018-09-1	106 W 12t	Hotels,Cat	Accommo	Kansas Cit	US	us/mo/kar	39.10012	-94.5847	Hotel Phill	64105	MO	2015-11-2	2016-03-2	5	http://ww	Very comf	Excellent Stay				Tabitha K	http://ww	http://curiocollection3.hilton.com/en/			
	2016-11-0	2018-09-1	106 W 12t	Hotels,Cat	Accommo	Kansas Cit	US	us/mo/kar	39.10012	-94.5847	Hotel Phill	64105	MO	2016-01-3	2016-03-2	3	http://ww	Stayed her	Beautiful lobby but rooms need				Yamachas	http://ww	http://curiocollection3.hilton.com/en/			
	2016-11-0	2018-09-1	106 W 12t	Hotels,Cat	Accommo	Kansas Cit	US	us/mo/kar	39.10012	-94.5847	Hotel Phill	64105	MO	2015-12-1	2016-03-2	5	http://ww	My husbar	Hotel's 19	Wichita	KS		Rbuffingto	http://ww	http://curiocollection3.hilton.com/en/			
	2016-11-0	2018-09-1	106 W 12t	Hotels,Cat	Accommo	Kansas Cit	US	us/mo/kar	39.10012	-94.5847	Hotel Phill	64105	MO	2015-09-0	2016-03-2	4	http://ww	Just stayec	Excellent s	Chicago	IL		Nick F	http://ww	http://curiocollection3.hilton.com/en/			
	2016-05-2	2018-09-1	10611 Sta	Bed Break	Accommo	Huntingdo	US	us/pa/hun	40.52748	-77.9698	The Inn at	16652	PA	2015-05-0	2016-05-0	5	https://ww	Everything	Thank you	Istanbul			sxk158	https://fo	http://www.solvang.com			
	2016-05-0	2018-09-1	102 Valley	Hotels,Cor	Accommo	Perry	US	us/ga/per	32.47188	-83.7452	Econolodg	31069	GA	2016-02-0	2016-05-0	5	https://ww	I work her	Extended :	Columbus	GA		tcolt45	http://trip	https://www.choicehotels.com/georgia			
	2016-11-0	2018-09-1	12 4th St	Hotels,Res	Accommo	San Franci	US	us/ca/sanf	37.78521	-122.406	Hotel Zelo	94103	CA	2016-05-0	2016-11-0	5	https://ww	The hotel	Convenient, cute and moderate				C	https://ww	http://www.hotelzelos.com			
	2016-11-0	2018-09-1	12 4th St	Hotels,Res	Accommo	San Franci	US	us/ca/sanf	37.78521	-122.406	Hotel Zelo	94103	CA	2016-05-1	2016-11-0	4	https://ww	excellent l	Perfectly fine hotel for a visit to				David	https://ww	http://www.hotelzelos.com			
	2016-11-0	2018-09-1	12 4th St	Hotels,Res	Accommo	San Franci	US	us/ca/sanf	37.78521	-122.406	Hotel Zelo	94103	CA	2016-07-0	2016-11-0	5	https://ww	Hotel Zelo	Great Staff!				Lindsay	https://ww	http://www.hotelzelos.com			
	2016-11-0	2018-09-1	12 4th St	Hotels,Res	Accommo	San Franci	US	us/ca/sanf	37.78521	-122.406	Hotel Zelo	94103	CA	2016-10-0	2016-10-1	3	https://ww	Pros: the l	Not so goc	Hong Kong			Maxence Z	https://ww	http://www.hotelzelos.com			
	2016-03-2	2018-09-1	998 W Lar	Hotels and	Accommo	Vineland	US	us/nj/vine	39.48745	-75.0445	EconoLodg	8360	NJ	2012-08-3	2016-03-1	1	http://ww	After getti	DO NOT stay here!!!!!!!!!!				Tunaphat	http://ww	http://www.econolodge.com			
	2016-03-2	2018-09-1	998 W Lar	Hotels and	Accommo	Vineland	US	us/nj/vine	39.48745	-75.0445	EconoLodg	8360	NJ	2012-07-1	2016-03-1	1	http://ww	We had nc	Filthy hote	Vineland			hoosierlad	http://ww	http://www.econolodge.com			
	2018-01-2	2018-09-1	250 Water	Hotels,Cor	Accommo	Oxon Hill	US	us/md/oxo	38.78213	-77.015	Hampton I	20745	MD	2018-01-0	2018-03-0	4	http://ww	Hotel is in	THE DC TRIP				A verified	http://ww	http://hamptoninn3.hilton.com/en/ho			

Libraries Used

- numpy – Arrays
- pandas – DataFrames
- matplotlib.pyplot – Plotting
- seaborn – Visualization
- sklearn.model_selection – Splitting
- sklearn.feature_extraction.text – Vectorization
- sklearn.linear_model – Regression
- sklearn.metrics – Evaluation
- pickle – Serialization
- sklearn.ensemble – Ensemble
- sklearn.svm – SVM
- sklearn.naive_bayes – Naive Bayes
- gensim.models – Word Embeddings
- sklearn.preprocessing – Scaling
- sklearn.utils – Utilities

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, classification_report
import pickle
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC
from sklearn.naive_bayes import MultinomialNB
from sklearn.naive_bayes import GaussianNB
from gensim.models import Word2Vec
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import LabelEncoder
from sklearn.utils import Bunch
from sklearn.model_selection import GridSearchCV
import nltk
from nltk.tokenize import word_tokenize
from imblearn.over_sampling import SMOTE
✓ 9.2s
```

Exploring the Dataset

Data Information

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 25 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   id                    10000 non-null  object
 1   dateAdded             10000 non-null  object
 2   dateUpdated           10000 non-null  object
 3   address               10000 non-null  object
 4   categories            10000 non-null  object
 5   primaryCategories     10000 non-null  object
 6   city                  10000 non-null  object
 7   country               10000 non-null  object
 8   keys                  10000 non-null  object
 9   latitude              10000 non-null  float64
 10  longitude             10000 non-null  float64
 11  name                  10000 non-null  object
 12  postalCode            10000 non-null  object
 13  province              10000 non-null  object
 14  reviews.date          10000 non-null  object
 15  reviews.dateSeen      10000 non-null  object
 16  reviews.rating        10000 non-null  float64
 17  reviews.sourceURLs    10000 non-null  object
 18  reviews.text          9999 non-null   object
 19  reviews.title         9999 non-null   object
 20  reviews.userCity      4164 non-null   object
 21  reviews.userProvince  2705 non-null   object
 22  reviews.username      10000 non-null  object
 23  sourceURLs            10000 non-null  object
 24  websites              10000 non-null  object
dtypes: float64(3), object(22)
memory usage: 1.9+ MB
```

Data Describe

```
data.describe()
```

[4] ✓ 0.0s

	latitude	longitude	reviews.rating
count	10000.000000	10000.000000	10000.000000
mean	37.003630	-92.675934	4.034265
std	5.517273	19.347989	1.162240
min	19.438604	-159.474930	1.000000
25%	33.927588	-111.622343	3.350000
50%	37.785060	-84.452114	4.000000
75%	40.416380	-77.052700	5.000000
max	70.133620	-68.203990	5.000000

Data Shape

```
data.shape
```

[8] ✓ 0.0s

```
(10000, 25)
```

Null Values

```
data.isnull().sum()
```

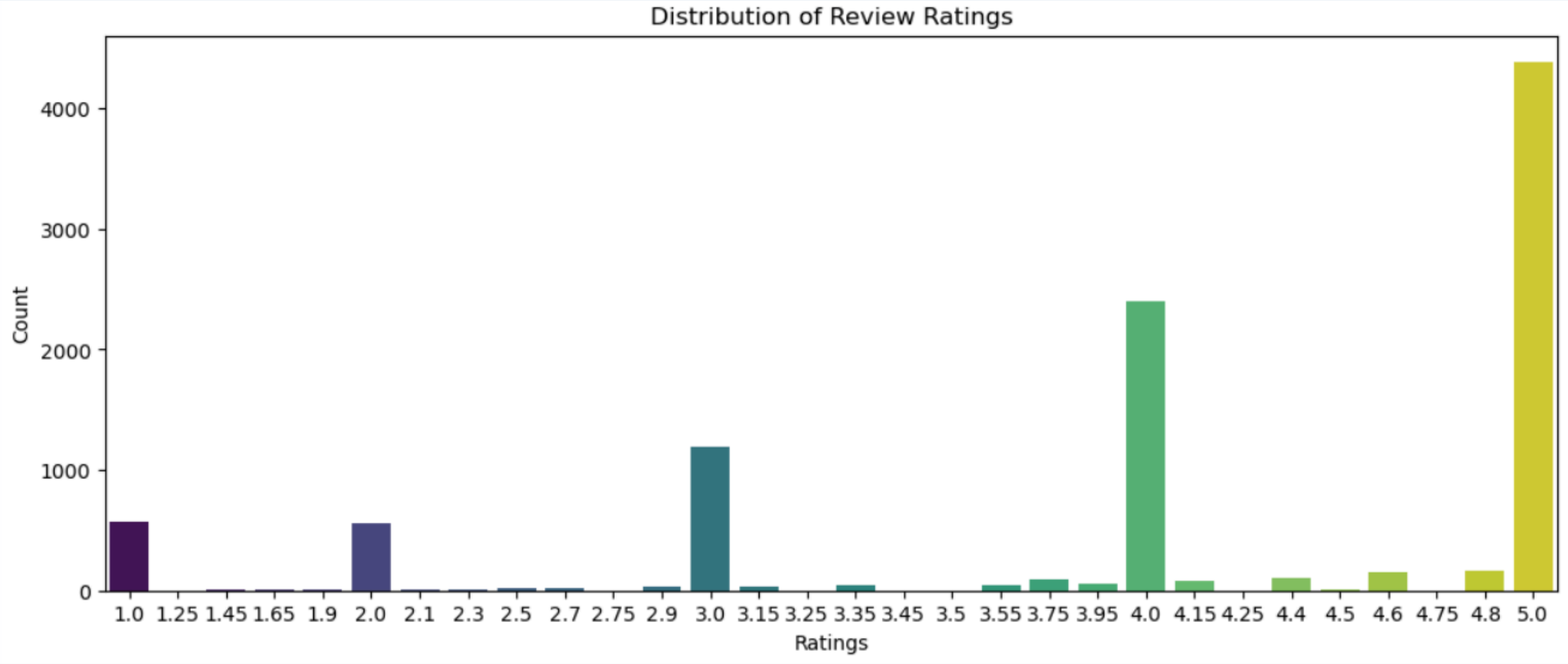
[6] ✓ 0.0s

```
id                        0
dateAdded                 0
dateUpdated               0
address                   0
categories                0
primaryCategories         0
city                      0
country                   0
keys                      0
latitude                  0
longitude                 0
name                      0
postalCode                0
province                  0
reviews.date              0
reviews.dateSeen          0
reviews.rating            0
reviews.sourceURLs        0
reviews.text              1
reviews.title             1
reviews.userCity       5836
reviews.userProvince   7295
reviews.username          0
sourceURLs                0
websites                  0
dtype: int64
```

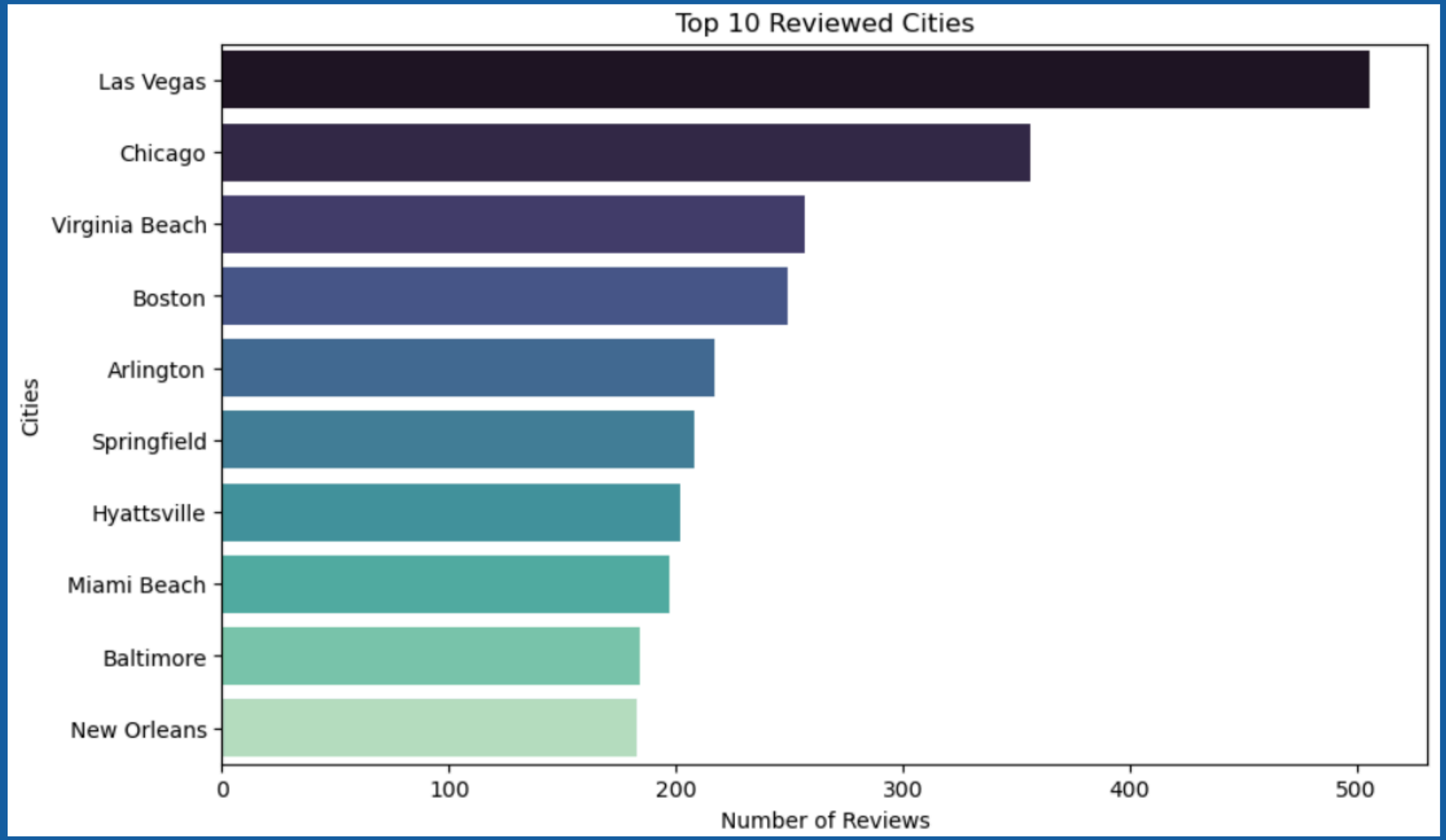
VISUALIZATION

Data visualization is the graphical representation of information and data using charts, graphs, and plots. It helps identify patterns, trends, and insights in the data, making it easier to understand and communicate complex information effectively.

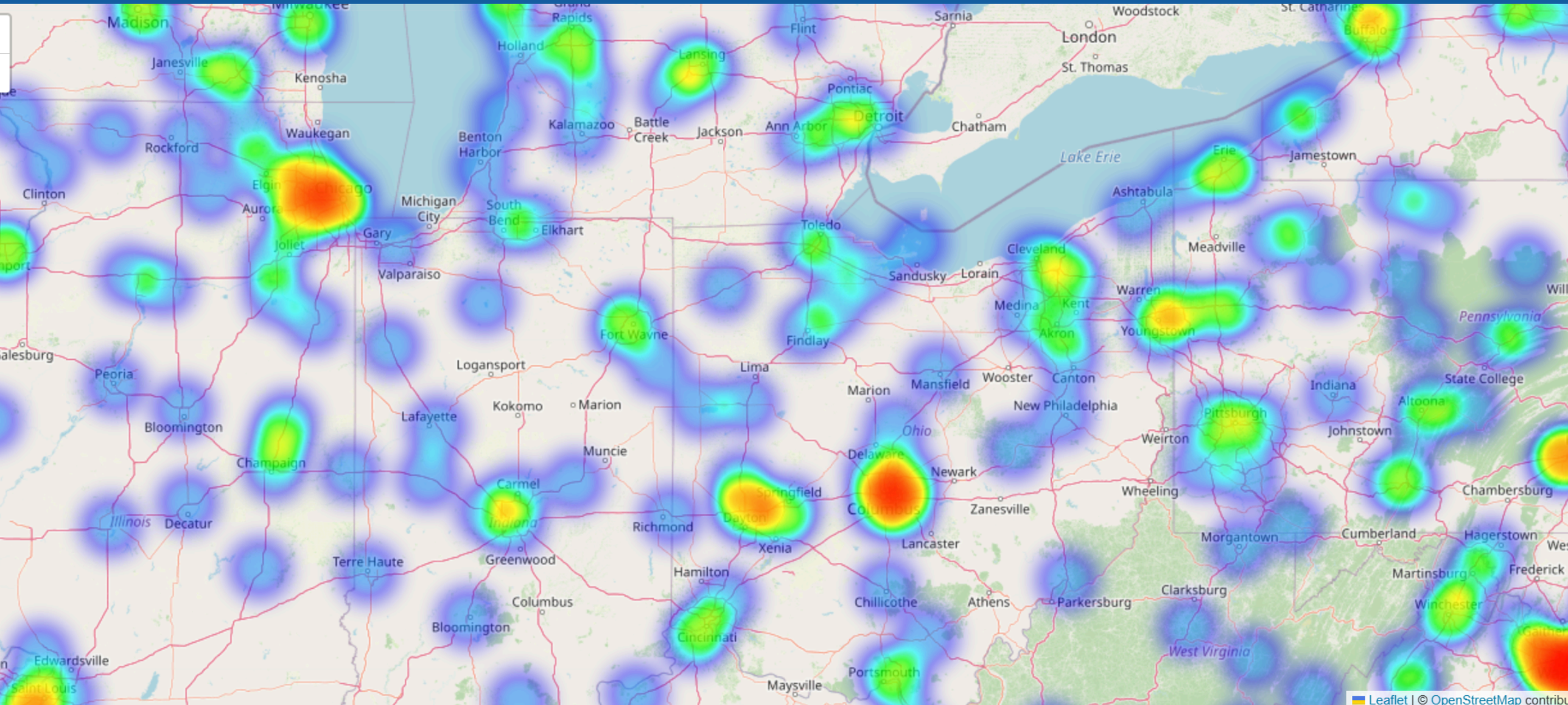
Ratings Distribution



Review Distribution



Geographical Map

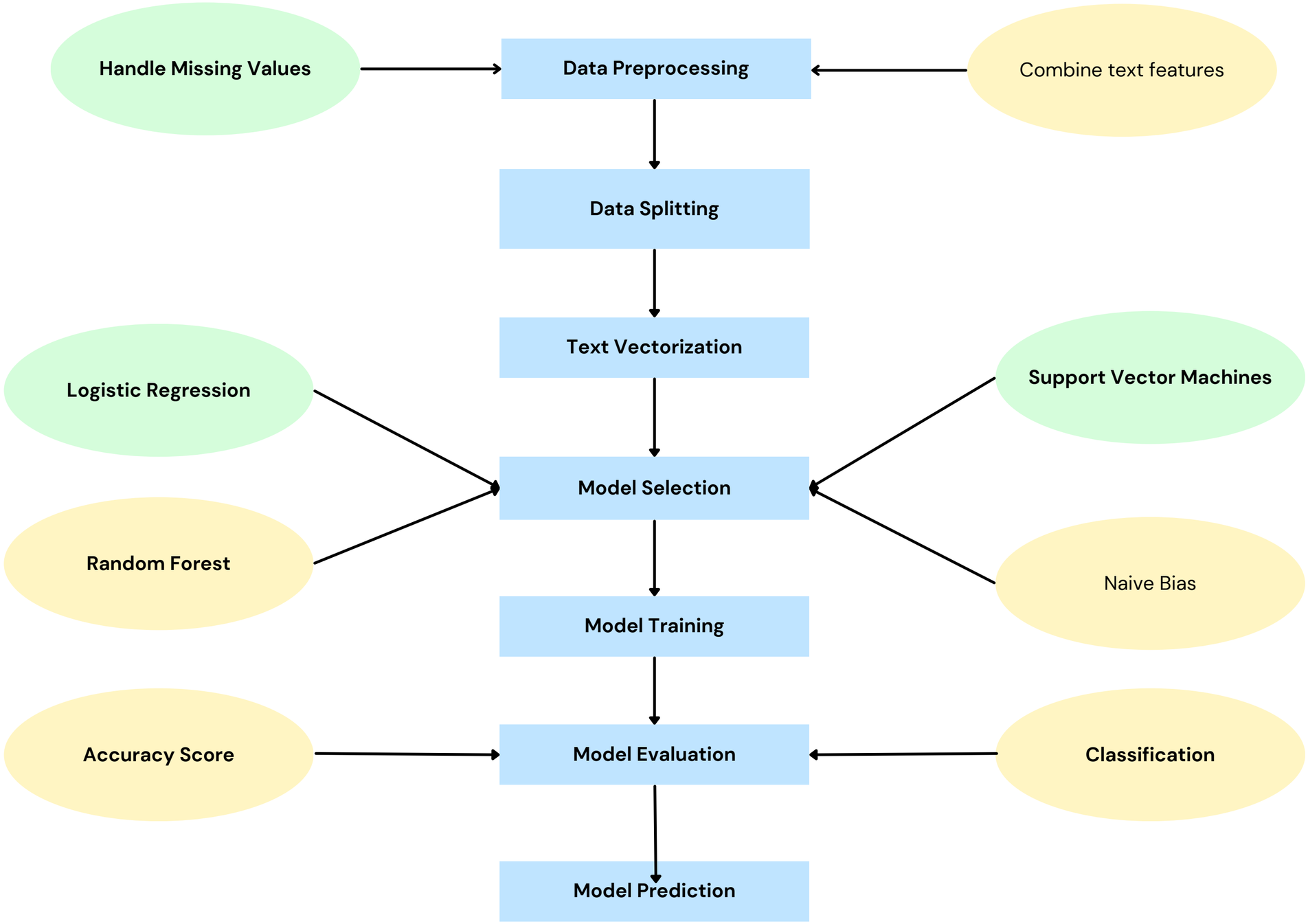


```python
# Handle missing values
data['primaryCategories'] = data['primaryCategories'].fillna('')
data['reviews.text'] = data['reviews.text'].fillna('')
data['reviews.title'] = data['reviews.title'].fillna('')

# Combine relevant text columns into a single feature for prediction
data['review_title.updated'] = data['primaryCategories'] + ' ' + data['reviews.text'] + ' ' + data['reviews.title']

# Drop rows with missing target values
data = data.dropna(subset=['reviews.rating'])

# Convert 'reviews.rating' to integer for modeling
data['reviews.rating'] = data['reviews.rating'].astype(int)

# Features and target
X = data['review_title.updated']
y = data['reviews.rating']

# Split the data into training (80%) and testing (20%) sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

```python
# Function for predicting new input ratings
def predict_rating(primary_category, review_text, review_title, model_type='LogisticRegression'):
    # Load the model and vectorizer
    with open(f'review_rating_{model_type}_model.pkl', 'rb') as model_file:
        model = pickle.load(model_file)
    with open('tfidf_vectorizer.pkl', 'rb') as vectorizer_file:
        vectorizer = pickle.load(vectorizer_file)

    # Combine the input features
    review_combined = primary_category + ' ' + review_text + ' ' + review_title

    # Transform input text using the vectorizer
    review_tfidf = vectorizer.transform([review_combined])

    # Predict the rating
    predicted_rating = model.predict(review_tfidf)[0]
    return predicted_rating

# Example: Predicting for a new input (Generalization)
primary_category = "Accommodation & Food Services"
review_text = "Our experience at Rancho Valencia was absolutely perfect from beginning to end!!!! We felt special and very happy during our stayed. I
review_title = "Best romantic vacation ever!!!!"
predicted_rating = predict_rating(primary_category, review_text, review_title, model_type)
print(f"The predicted rating is: {predicted_rating} star")
```

```python
# Model selection
model_type = 'LogisticRegression'  # Change this to 'RandomForest', 'SVC', or 'NaiveBayes' for different models

if model_type == 'LogisticRegression':
    model = LogisticRegression(max_iter=1000)
elif model_type == 'RandomForest':
    model = RandomForestClassifier(n_estimators=100, random_state=42)
elif model_type == 'SVC':
    model = SVC(kernel='linear', C=1)
elif model_type == 'NaiveBayes':
    model = MultinomialNB()
else:
    raise ValueError("Invalid model_type! Choose from 'LogisticRegression', 'RandomForest', 'SVC', or 'NaiveBayes'.")

# Train the selected model
model.fit(X_train_tfidf, y_train)

# Save the model for future use
with open(f'review_rating_{model_type}_model.pkl', 'wb') as model_file:
    pickle.dump(model, model_file)

# Generalization: Evaluate the model on unseen test data
X_test_tfidf = vectorizer.transform(X_test)  # Transform the test set
y_pred = model.predict(X_test_tfidf)

# Output evaluation metrics
print(f"Accuracy Score with {model_type}: {accuracy_score(y_test, y_pred)}")
print("Classification Report:")
print(classification_report(y_test, y_pred))
```

Let us understand how the code is working here


Handle Missing Values
Data Preprocessing
Combine text features
Data Splitting
Text Vectorization
Logistic Regression
Support Vector Machines
Model Selection
Random Forest
Naive Bias
Model Training
Accuracy Score
Model Evaluation
Classification
Model Prediction

Statistics

Performance Summary

- Model Used: Logistic Regression
- Accuracy: 58.35%
- Best Performance: 5-star reviews (Precision: 67%, Recall: 85%).
- Weak Areas: 2-star reviews (Precision: 44%, Recall: 17%).

58.25
Accuracy

67.00%
Precision

85.00%
Recall

```
Accuracy Score with LogisticRegression: 0.5835
Classification Report:
              precision    recall  f1-score   support

           1       0.59      0.53      0.56       102
           2       0.44      0.17      0.24       130
           3       0.49      0.35      0.40       312
           4       0.45      0.41      0.43       575
           5       0.67      0.85      0.75       881

    accuracy                           0.58      2000
   macro avg       0.53      0.46      0.48      2000
weighted avg       0.56      0.58      0.56      2000

The predicted rating is: 5 star
```

THANK YOU!

"We sincerely thank the jury members and our mentors for their invaluable feedback and encouragement throughout this project."