Analytics for Big Data

Yelp Review Recommendation System through Sentiment Analysis

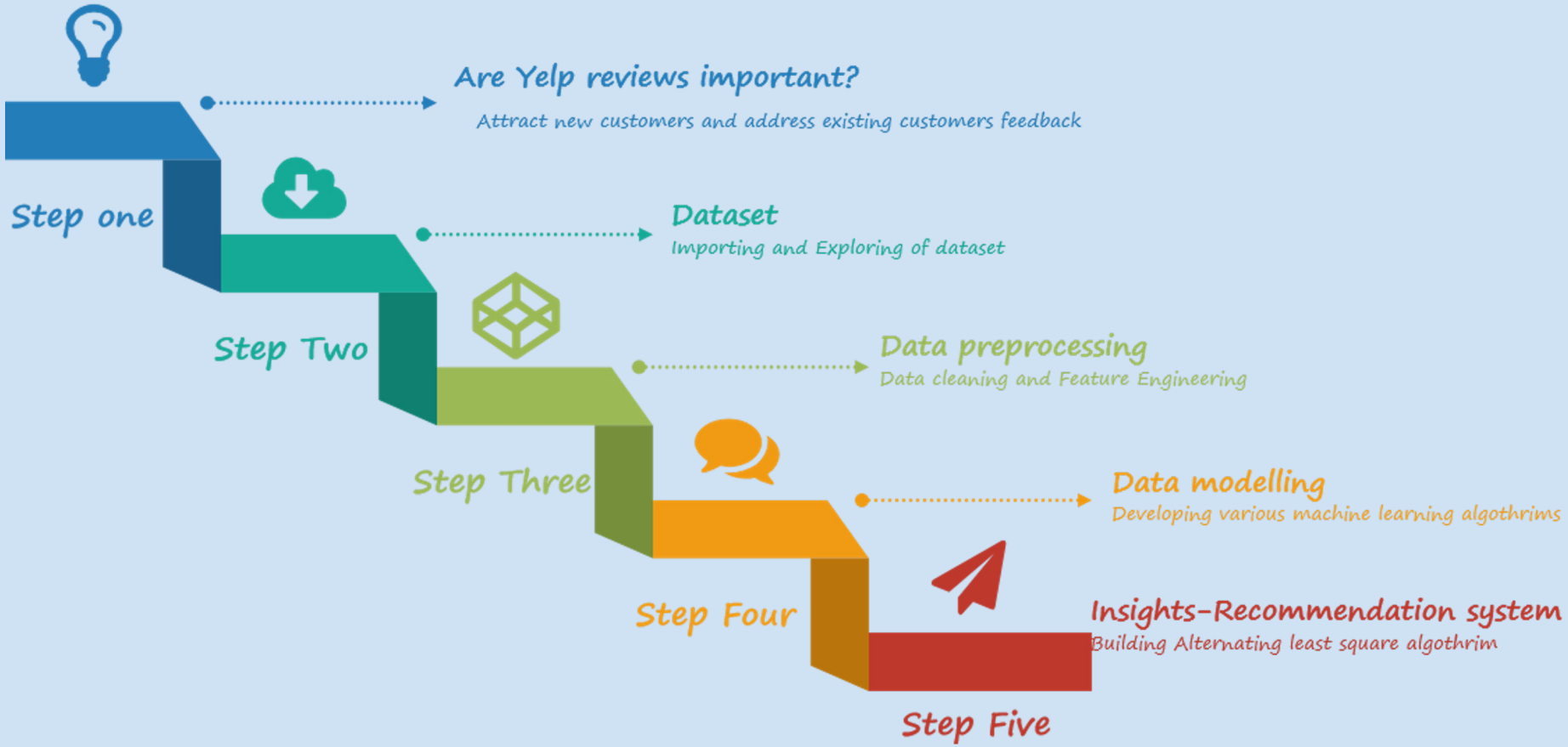
Hoang Nha Nguyen

Nithin S Gowda

Prajwal C P

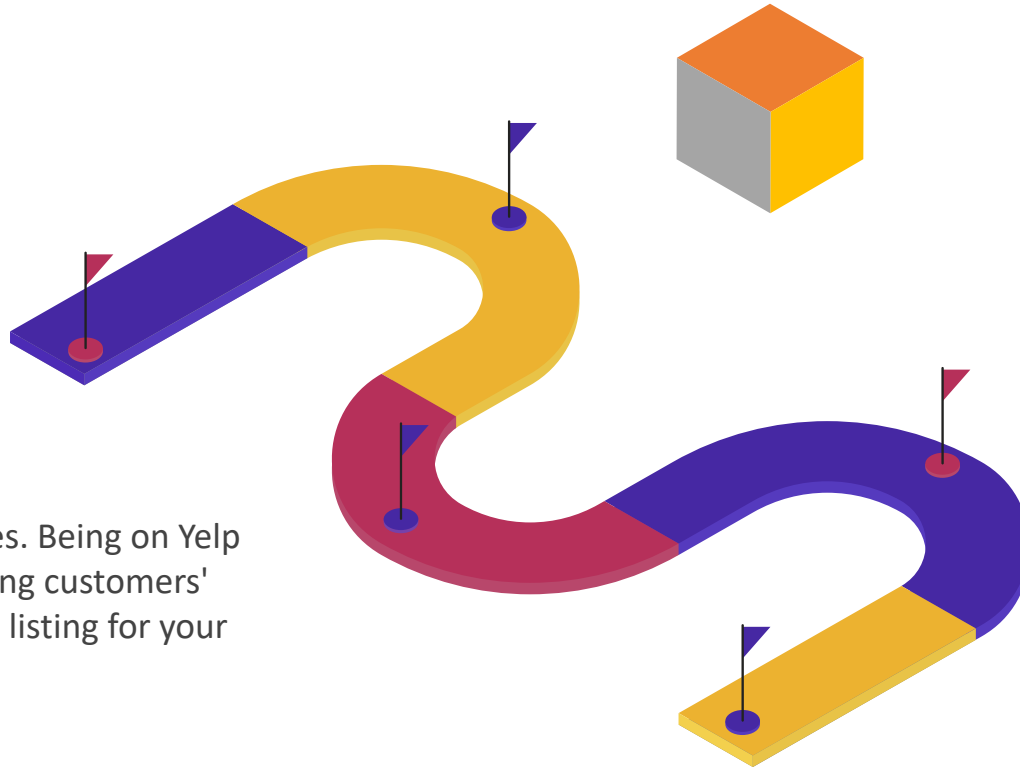
Purva Ladge

Contents


Are Yelp reviews important?
Attract new customers and address existing customers feedback
Step one
Dataset
Importing and Exploring of dataset
Step Two
Data preprocessing
Data cleaning and Feature Engineering
Step Three
Data modelling
Developing various machine learning algothrims
Step Four
Insights-Recommendation system
Building Alternating least square algothrim
Step Five

01

How Yelp reviews mean to Business?



Yelp is a top review site for marketing local businesses. Being on Yelp helps them attract new customers and address existing customers' feedback to improve their company. You can set up a listing for your business on Yelp and respond to negative reviews.

Problem Setting

• Yelp is a popular website and app where crowd-sourced reviews are published about businesses

• Yelp also runs an online reservation service, hosts social events, provides data about businesses

• Our goal is to utilize the vast amounts of data generated by Yelp and build a recommender system for users.

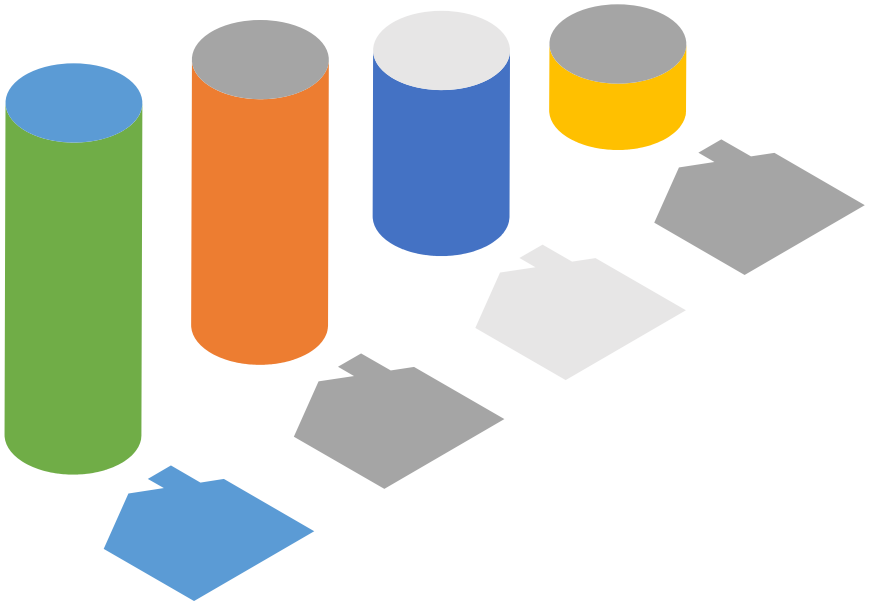
• This will help businesses identify their target customers and increase traffic to the app

Review Analysis Techniques

Sentiment Analysis: Sentiment analysis is the process of determining the emotional tone of a piece of text

Keyword Analysis: This involves identifying the most commonly used words and phrases in reviews. By doing this, restaurants can gain insight into what customers like and dislike about their establishment

Competitor Analysis: By analyzing the reviews of similar restaurants in the area, they can identify areas where they may be falling short and make improvements accordingly



02

Our Dataset

A quick overview of the **Users Story !**

Yelp Dataset Challenge

- Employed the Yelp Open Kaggle Dataset, a subset of Yelp's businesses, reviews, and user data. It was originally from “Yelp Dataset Challenge”.
- We have about 7M reviews on 150k businesses by almost 2M users in three **JSON** files:
  - **business.json** - business data including location data, attributes, and categories.
  - **review.json** - full review text data including the user_id that wrote the review and the business_id the review is written for.
  - **user.json** - user's friend mapping and all the metadata associated with the user.

```
print("Number of reviews: {}".format(yelp_review.count()))
print("Number of business: {}".format(yelp_business.count()))
print("Number of users: {}".format(yelp_user.count()))
```

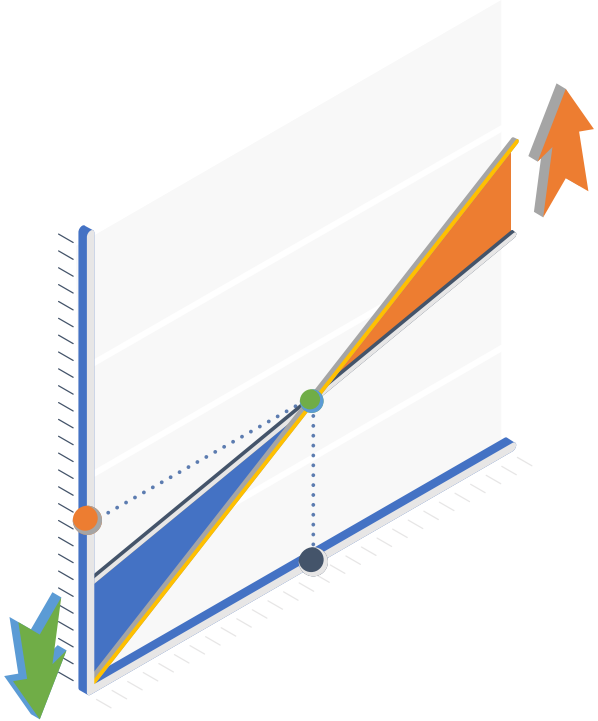
```
Number of reviews: 6990280
Number of business: 150346
Number of users: 1987897
```

Target Variable: ‘stars’

- 1 - star ratings >=3
- 0 - star ratings <3

With regression models, our objective is to help businesses predict Yelp star ratings based on past reviews given by the users.

Objective: we aim to build the grading recommendations system, the **location** of the restaurants is an important factor to consider.



Exploratory Data Analysis

Data Types

```
yelp_business.dtypes
yelp_review.dtypes
yelp_user.dtypes
```

```
[('average_stars', 'double'),
 ('compliment_cool', 'bigint'),
 ('compliment_cute', 'bigint'),
 ('compliment_funny', 'bigint'),
 ('compliment_hot', 'bigint'),
 ('compliment_list', 'bigint'),
 ('compliment_more', 'bigint'),
 ('compliment_note', 'bigint'),
 ('compliment_photos', 'bigint'),
 ('compliment_plain', 'bigint'),
 ('compliment_profile', 'bigint'),
 ('compliment_writer', 'bigint'),
 ('cool', 'bigint'),
 ('elite', 'string'),
 ('fans', 'bigint'),
 ('friends', 'string'),
 ('funny', 'bigint'),
 ('name', 'string'),
 ('review_count', 'bigint'),
 ('useful', 'bigint'),
 ('user_id', 'string'),
 ('yelping_since', 'string')]
```

- We are examining the data types of the three datasets: “business”, “Review”, and “Users”

```
yelp_business.show(5)
yelp_review.show(5)
yelp_user.show(5)
```

address	attributes	business_id	categories	city	hours	is_open	latitude	longitude	
1616 Chapala St, ...	{null, null, null...	Pns2l4eNsfO8kk83d...	Doctors, Traditio...	Santa Barbara	null	0	34.4266787	-119.7111968	Abby
87 Grasso Plaza S...	{null, null, null...	mpf3x-BjTdTEA3yCZ...	Shipping Centers,...	Affton	{8:0-18:30, 0:0-0...	1	38.551126	-90.335695	
5255 E Broadway Blvd	{null, null, null...	tUFrWirKiKi_TAnsV...	Department Stores...	Tucson	{8:0-23:0, 8:0-22...	0	32.223236	-110.880452	
935 Race St	{null, null, u'no...	MTSW4McQd7CbVtyjq...	Restaurants, Food...	Philadelphia	{7:0-21:0, 7:0-20...	1	39.9555052	-75.1555641	St
101 Walnut St	{null, null, null...	mWMc6_wTdE0EUBKIG...	Brewpubs, Breweri...	Green Lane	{12:0-22:0, null,...	1	40.3381827	-75.4716585	Perk

only showing top 5 rows

business_id	cool	date	funny	review_id	stars	text	useful	user_id
XQfwVwDr-v0ZS3_Cb...	0	2018-07-07 22:09:11	0	KU_O5udG6zpxOg-Vc...	3.0	If you decide to ...	0	mh_-eMZ6K5RLWhZyI...
7ATYjTIgM3jUlt4UM...	1	2012-01-03 15:28:18	0	BiTunyQ73aT9WBnpR...	5.0	I've taken a lot ...	1	OyoGAe7OKpv6SyGZT...
YjUWPpI6HXG530lwP...	0	2014-02-05 20:30:30	0	saUsX_uimxRlCVr67...	3.0	Family diner. Had...	0	8g_iMtfSiwikVnbP2...
kxX2SOes4o-D3ZQBk...	1	2015-01-04 00:01:03	0	AqPFMleE6RsU23_au...	5.0	Wow! Yummy, diff...	1	_7bHUi9Uuf5__HHc_...
e4Vwtrqf-wpJfwesg...	1	2017-01-14 20:54:15	0	Sx8TMOWLNuJBWer-0...	4.0	Cute interior and...	1	bcjbaE6dDog4jkNY9...

only showing top 5 rows

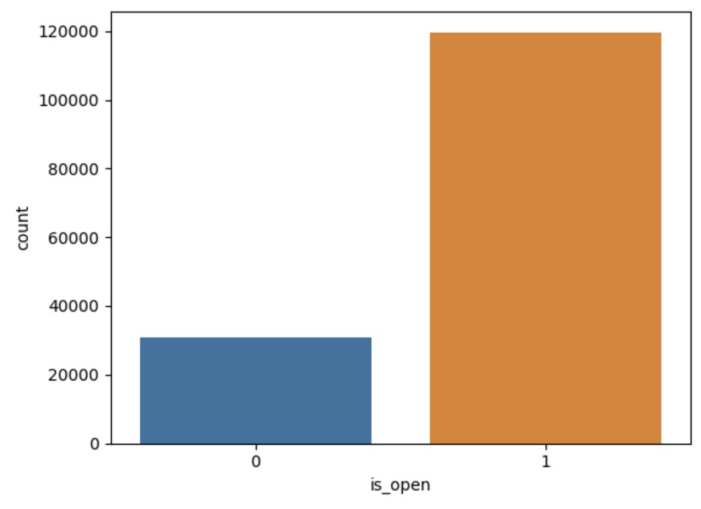
average_stars	compliment_cool	compliment_cute	compliment_funny	compliment_hot	compliment_list	compliment_more	compliment_note	compliment_photos	compliment_
3.91	467	56	467	250	18	65	232	180	
3.74	3131	157	3131	1145	251	264	1847	1946	
3.32	119	17	119	89	3	13	66	18	
4.27	26	6	26	24	2	4	12	9	
3.54	0	0	0	1	0	1	1	0	

only showing top 5 rows

EDA -Univariate Analysis

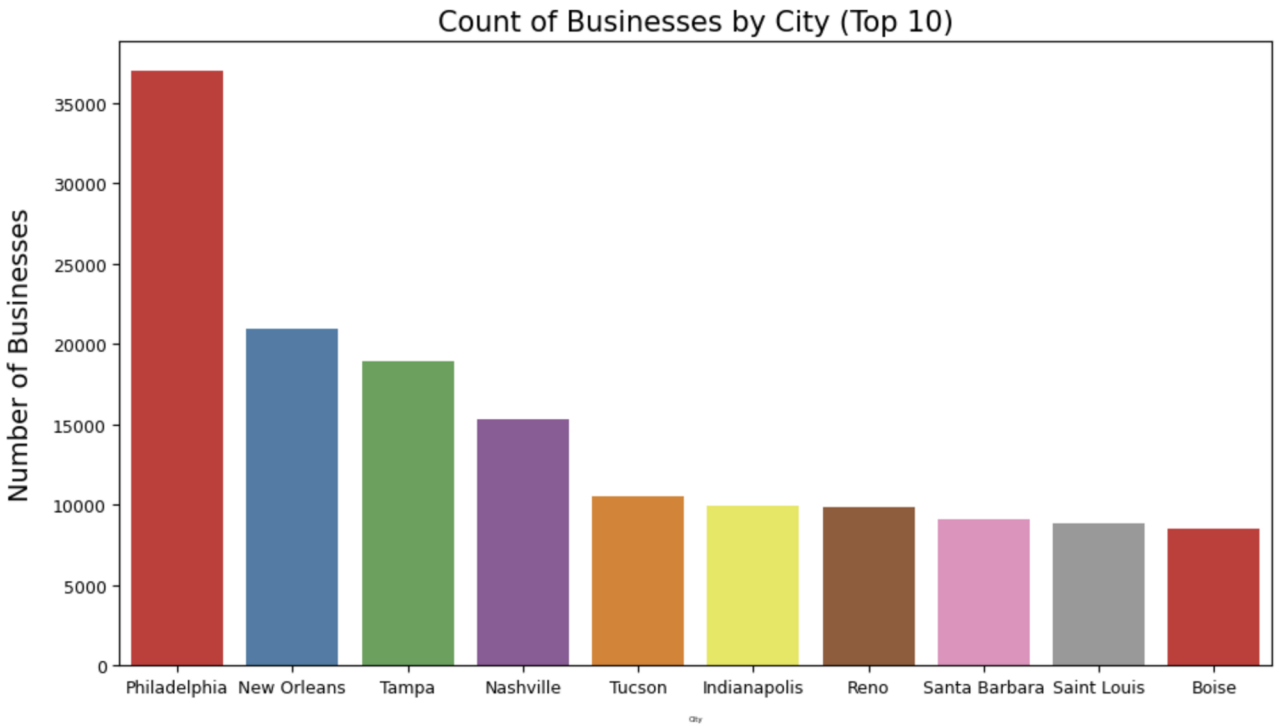
- Top most reviewed businesses where star rating is greater than thee.

business_id	name	categories	count
sk2lZI4zmuGAccd3D...	Boyd Hill Nature ...	Active Life, Park...	33
83IeQHroXEctmMpK1...	The Richel D'Ambr...	Hair Salons, Day ...	39
6e85By5Jy7MMnW2cE...	Wanderwell	Tax Services, Boo...	5
WKMJwqnfZKsAae75R...	Roast Coffeehouse...	Coffee & Tea, Foo...	31
jIBjZcqVOU4pOVT-s...	IRB Sushi	Sushi Bars, Resta...	34
NQhyMw8SOU1HB-V9X...	Champion Chevrole...	Automotive, Auto ...	20
q6661I3CGW0UB740E...	India House	Home & Garden, Ho...	32
lpbt16sSm4BTcfeq4...	Super Wok	Restaurants, Chinese	38
AwmeLVLEfdFoCa0La...	The Beer Store	Food, Beer, Wine ...	16
RZ-FNTXvqHKngyLGD...	Gaetano's of West...	Italian, Food, Re...	15



EDA - Univariate Analysis

- Top 10 cities with most businesses

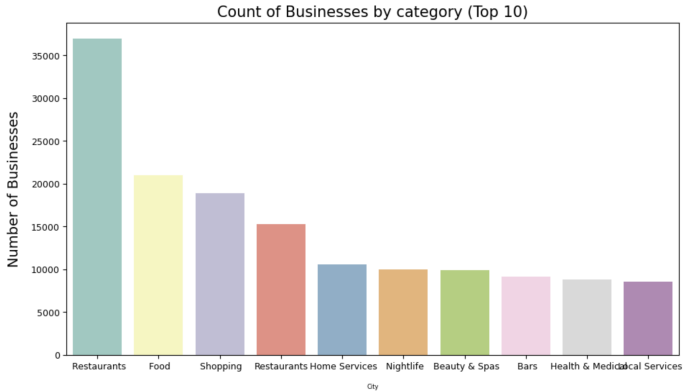


city	sum(count)
Philadelphia	967552
New Orleans	635364
Tampa	454889
Nashville	451571
Tucson	404880
Indianapolis	361489
Reno	351573
Santa Barbara	269630
Saint Louis	253437
Boise	105366

Univariate Analysis

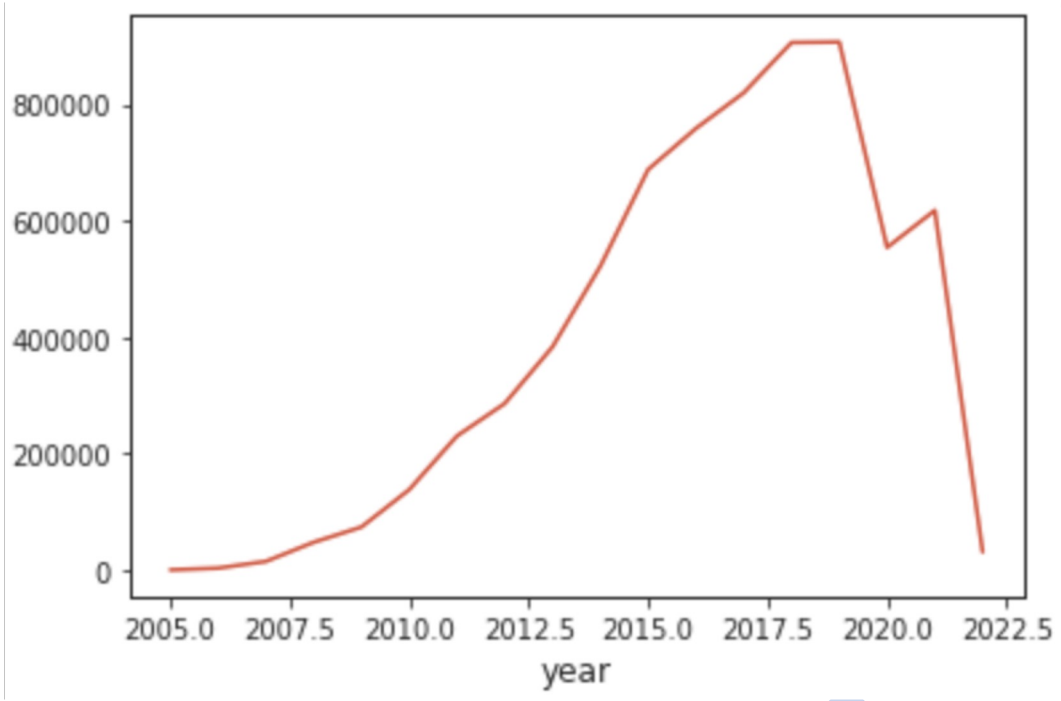
- Total number of Businesses of different categories

category	count
Restaurants	36978
Food	20998
Shopping	18915
Restaurants	15290
Home Services	10563
Nightlife	9990
Beauty & Spas	9907
Bars	9130
Health & Medical	8832
Local Services	8556



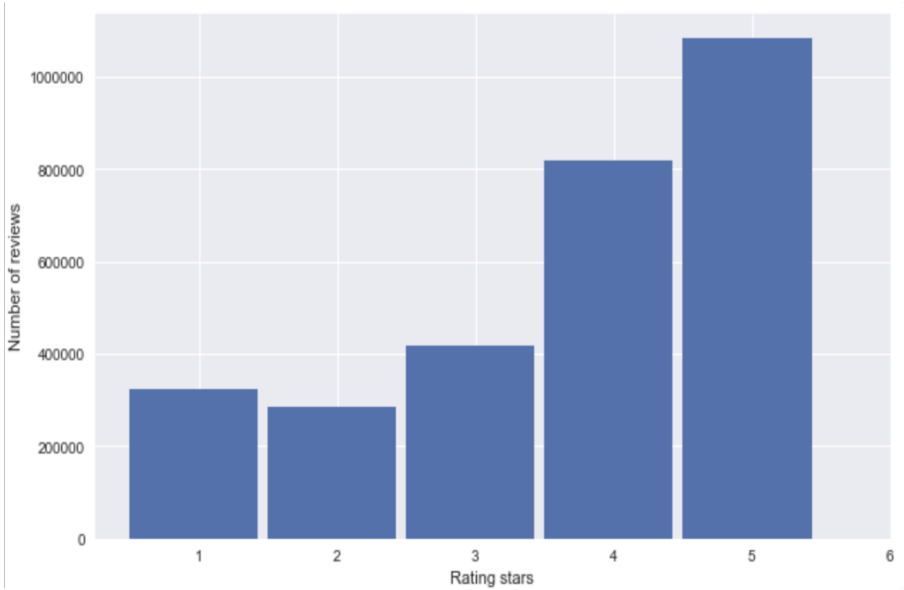
Univariate Analysis

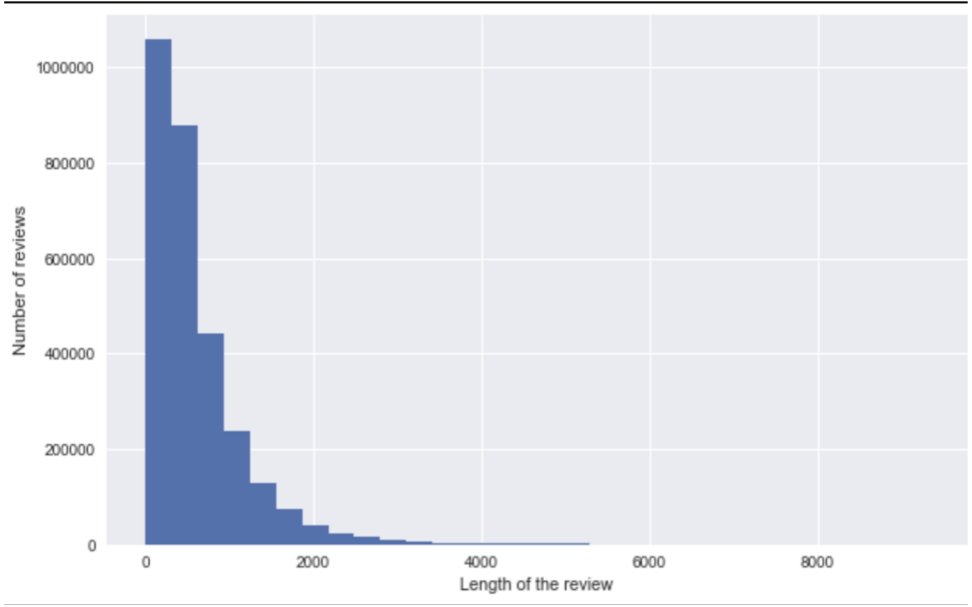
- Distribution of Reviews over years



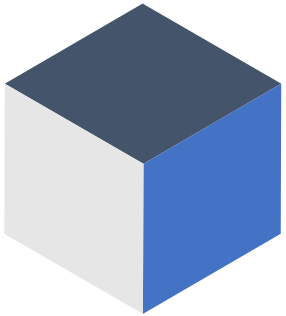
Bivariate Analysis

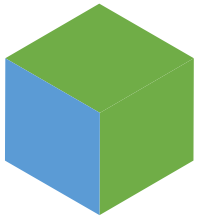
- Distribution of **Number of reviews** and **Star ratings**





- Distribution of **number of reviews** and the **length of reviews**




03

Text Pre-processing Sentiment Analysis

Imported Libraries

- Libraries for Machine Learning features and clustering

```
from pyspark.ml.feature import StandardScaler
from pyspark.ml.clustering import KMeans
from pyspark.ml.evaluation import ClusteringEvaluator
```

- Libraries for sentiment analysis

```
from wordcloud import WordCloud, STOPWORDS
```

- Libraries for Spark

```
from pyspark import *
from pyspark.python.pyspark.shell import spark
from pyspark.sql.functions import *
from datetime import datetime
from pyspark.sql.functions import udf, to_date, to_utc_timestamp, lit, col
from pyspark.sql.types import StringType, DateType
from pyspark import SparkContext
from pyspark.sql import SQLContext
```

```
from pyspark.sql.functions import split,explode
from pyspark.sql.functions import *
from pyspark.sql.functions import udf
from pyspark.sql.types import IntegerType

from pyspark.mllib.classification import SVMModel, SVMWithSGD
from pyspark.mllib.regression import LabeledPoint
from pyspark.mllib.linalg import Vectors as MLLibVectors
from pyspark.ml import Pipeline
from pyspark.ml.evaluation import BinaryClassificationEvaluator, MulticlassClassificationEvaluator
from pyspark.ml.feature import *
from pyspark.ml.feature import IDF
from pyspark.ml.tuning import CrossValidator
from pyspark.ml.tuning import ParamGridBuilder


from nltk.stem.porter import *
from nltk.tokenize import word_tokenize, sent_tokenize
from nltk.corpus import stopwords

from pyspark.mllib.classification import SVMModel, SVMWithSGD
from pyspark.mllib.regression import LabeledPoint
from pyspark.mllib.linalg import Vectors as MLLibVectors
```

TEXT PRE-PROCESSING WORKFLOW

review_id	text	label
KU_O5udG6zpxOg-Vc...	If you decide to ...	1
BiTunyQ73aT9WBnpR...	I ve taken a lot ...	1
saUsX_uimxRlCVr67...	Family diner Had...	1
AqPFMleE6RsU23_au...	Wow Yummy diff...	1
Sx8TMOWLNuJBWer-0...	Cute interior and...	1

only showing top 5 rows

review_id	text	label	words	words_new
KU_O5udG6zpxOg-Vc...	If you decide to ...	1	[if, you, decide,...	[decide, eat, , a...
BiTunyQ73aT9WBnpR...	I ve taken a lot ...	1	[i, ve, taken, a,...	[ve, taken, lot, ...
saUsX_uimxRlCVr67...	Family diner Had...	1	[family, diner, ,...	[family, diner, ,...
AqPFMleE6RsU23_au...	Wow Yummy diff...	1	[wow, , , yummy, ...	[wow, , , yummy, ...
Sx8TMOWLNuJBWer-0...	Cute interior and...	1	[cute, interior, ...	[cute, interior, ...

only showing top 5 rows

Step 1:
Text data after removal of punctuations & label defined using the threshold value = 3

Step 2:
Text data after tokenization and removal of stop words

review_id	text	label	words	words_new
KU_O5udG6zpxOg-Vc...	If you decide to ...	1	[if, you, decide,...	[decide, eat, , a...
BiTunyQ73aT9WBnpR...	I ve taken a lot ...	1	[i, ve, taken, a,...	[ve, taken, lot, ...
saUsX_uimxRlCVr67...	Family diner Had...	1	[family, diner, ,...	[family, diner, ,...
AqPFMleE6RsU23_au...	Wow Yummy diff...	1	[wow, , , yummy, ...	[wow, , , yummy, ...
Sx8TMOWLNuJBWer-0...	Cute interior and...	1	[cute, interior, ...	[cute, interior, ...

only showing top 5 rows

Step 3:
Text data after feature extraction:
1. Count Vectorization
2. Term Frequency and Inverse Document Frequency (tf-idf)

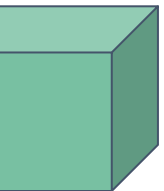
04

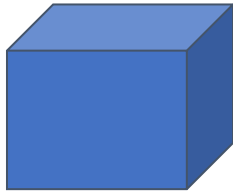
Data Modeling

- Logistic Regression (Baseline Model)
- Random Forest
- Naive Bayes
- SVM

Supervised Models - Results

	Naive Bayes	Logistic Regression	Random Forest	SVM
F1 Score	0.846	0.8816	0.766	0.9122
Parameters used	Smoothing = 1.0 Model Type = 'Multinomial'	Lambda = 0.02 Alpha = 0.3	Max Depth = 2	Regular Parameter = 0.3 Number of iterations = 50





05 How to implement?

Yelp Review Recommendation System

• We have implemented a recommender system that could suggest a user his/her preferred business around his or her location.

• As we found the top businesses are 'Restaurants' and the most reviewed city is 'Philadelphia'. We have chosen both these categories to build the system.

• We have built the recommender system using Alternating Least Square algorithm

• The parameters used are:
1. maxIter=5,
2. regParam=0.09,
3. rank=25

• Model is evaluated on RMSE: 1.44

Insights

• Support Vector Machine gives the highest accuracy of 92% in predicting the sentiment of a business.

• The Recommender system is able to give the user = 952 to following restaurant recommendations in the city "Philadelphia".

business_id	user_id	stars	categories
4_W5pstoN1166TGjj...	8zBwGPQIzuvnjbrRc...	4.0	American (Traditi...
kxKai8GE5oDMPevV7...	8zBwGPQIzuvnjbrRc...	3.0	Restaurants, Bars...
RkSs_qLitbI320DP2...	8zBwGPQIzuvnjbrRc...	5.0	Nightlife, Restau...
fdNvkw1Z9L6TkLnfX...	8zBwGPQIzuvnjbrRc...	2.0	Restaurants, Food...
I4Szupt_YHzR9dczc...	8zBwGPQIzuvnjbrRc...	4.0	Restaurants, Dive...
HcddEbhaQ3wgyEFoE...	8zBwGPQIzuvnjbrRc...	4.0	Thai, Restaurants
HcddEbhaQ3wgyEFoE...	8zBwGPQIzuvnjbrRc...	4.0	Thai, Restaurants
e1nfM_rumu9l6dLCc...	8zBwGPQIzuvnjbrRc...	2.0	Food, Ethnic Food...
y1Z9tymuBGVDZnYZo...	8zBwGPQIzuvnjbrRc...	3.0	Burgers, American...
1Ck2drCXGh1h5bhc9...	8zBwGPQIzuvnjbrRc...	5.0	Restaurants, Amer...
6JFMbFYVbl8ufz74N...	8zBwGPQIzuvnjbrRc...	5.0	American (New), R...
aJV-u_8zf5vVIaHy7...	8zBwGPQIzuvnjbrRc...	3.0	Nightlife, Restau...
x39G7-aTCVh-972fg...	8zBwGPQIzuvnjbrRc...	4.0	Food, Coffee & Te...