**Data Analysis on YELP**

**application reviews**

**Project by:** Sujeeth Nidumolu

**Under the supervision of:** Dr. Someswar Kesh

**Name of the document:** Final report documentation

**Term:** Fall 2018

**Student ID:** 700667613

**Email:** [sxn76130@ucmo.edu](mailto:sxn76130@ucmo.edu)

Table of Contents

[**1.** **Glossary of terms** 3](#_Toc531838141)

[**2.** **Statement of requirements** 4](#_Toc531838142)

[**3.** **Functional requirements** 6](#_Toc531838143)

[**4.** **System diagram** 7](#_Toc531838144)

[**5.** **Use cases** 8](#_Toc531838145)

[**6.** **Class diagram** 10](#_Toc531838146)

[**7.** **Project notes** 11](#_Toc531838147)

[**8.** **Test cases** 18](#_Toc531838148)

[**9.** **Conclusion** 19](#_Toc531838149)

[**10.** **References** 20](#_Toc531838150)

# **Glossary of terms**

**Yelp:** A well-known website and application which records and displays the reviews given by users on nearby restaurants, motels, resorts, coffee shops etc. which helps the customers to look at the reviews before making a visit. Also, it helps the existing business owners to know the feedback from the customers to improvise or expand their business accordingly.

**Business/Businesses:** Local bodies making money through servicessuch as restaurants, motels, coffee shops, bars etc. and are linked with Yelp.

**User:** A person who has registered with Yelp for reading and writing reviews on local business bodies.

**Rating(Stars):** It is measured on a scale of 0 to 5 where a user can rate his/her experience on a business body and the average is calculated and displayed on the yelp application.

**Review(Tip):** Along with the rating, user can share his/her experience in words.

**Entrepreneur/Future business owner:** A person who is willing to start a new business organization.

**Existing business owner:** Person actively running business and tied with Yelp for reviews from users and track business trends.

**Dataset:** Logs of information recorded from an application which is used for data analytics.

# **Statement of requirements**

**Goal:** The goal is to analyze this unorganized data, design a system to generate insights from the Yelp data which could help the entrepreneurs and business owners to study the progress of the business and take wise decisions towards betterment of the business. With the Yelp’s data, the ultimate task is to derive insights from various reviews from customers, data being acquired from various places across USA.

**Problem statement:** There is a tremendous increase in data from year to year which increases the demand of cloud storage such as iCloud, Google drive, OneDrive, Dropbox, Facebook, Whatsapp etc. Every application now a day uses cloud storage for saving the personal data which keeps expanding day to day. Statistics says that on an average, 3 exabytes (which is equivalent to 1e+9 GB) of data is generated every day from the internet.

Although Yelp displays all the reviews and ratings as an average from all the users in their website and application as a list, it can be a hard job for a restaurant owner to go through each review and analyze. Therefore, the challenge is to develop an analytics system which could take the raw data from Yelp datasource and derive insights which could clearly visualize the sectors to the existing business owners or future business owners who are planning to establish their business.

**Proposed solutions:**

1. To provide analytics to existing business owners which could help them in improving their services, look at their business trends and make important decisions of their business such as expansion of their franchise in other cities or countries.
2. To help the future business owners or Entrepreneurs in providing analysis on the success rate a restaurant can yield if they set up the business at a certain region, or country.

**Proposed questions:**

Below are the proposed questions to develop insights on the *Yelp* dataset:

1. Finding the day and time in a week receiving highest number of check-ins.
2. Predicting the top 50 reviewed businesses in USA.
3. To find the cities playing major role in making business.
4. Determining cities with highest reviews and best ratings for their businesses.
5. Determining cities with greater than 50k reviews ranked by average star ratings.
6. To determine frequently given tips(feedback) from customers.
7. Finding the closed businesses.
8. Taking one business with highest rating and one business with lowest rating to compare the tips recorded by the users in both the cases to compare feedback given by the users.

# **Functional requirements**

Following are the tools and functional requirements to perform the analysis:

1. Python3 programming.
2. Jupyter Notebook – IDE for python
3. Revised dataset extracted from *Yelp official* and *Kaggle.*
4. Machine learning algorithms to load, clean and manipulate the data.

**FURPS table:**

|  |  |
| --- | --- |
| **Functionality** | 1. Machine learning algorithms are used to load, process and manipulate the data. 2. Graphs and charts are displayed as output in the Notebook |
| **Usability** | 1. User can see bar charts as outputs by running the entire Notebook at once. |
| **Reliability** | 1. System will discard the rows and cells if they are empty or corrupt to refine the useful data. |
| **Performance** | 1. System will perform analysis on 3.2GB dataset for the prediction |

# **System diagram**

Results displayed to the user in IDE

Python and machine learning programs

Yelp Dataset

Fig. System diagram of Data analysis

**Actors and goals**

|  |  |
| --- | --- |
| **ACTOR** | **GOALS** |
| **Existing business owner** | Who is actively running business and is linked with Yelp to analyze the business to expand and improve the business. |
| **Future business owner** | Who is looking to open a new business by looking at the Yelp analysis on past data. |

# **Use cases**

Detailed description of the use cases:

|  |  |  |
| --- | --- | --- |
| **Use Case** | **Identifier** | **Description** |
| UC-1 | Identify day and time | Finding the day and time in a week receiving highest number of check-ins. |
| UC-2 | Top 50 reviewed businesses | To find the list of top 50 businesses which has received highest number of reviews |
| UC-3 | Cities role in businesses | Determining the cities playing major role in making business |
| UC-4 | Cities with highest reviews for business | Determining cities with highest reviews and best ratings for their businesses. |
| UC-5 | Cities with greater than 50k reviews | Determining cities with greater than 50k reviews ranked by average star ratings. |
| UC-6 | Frequent customer tips | To determine frequently given tips(feedback) from customers. |
| UC-7 | Closed businesses | Finding the closed businesses which are registered but closed |
| UC-8 | Highest and lowest rated businesses | Taking one business with highest rating and one business with lowest rating to compare the tips recorded by the users in both the cases to compare feedback given by the users. |

Table. Use cases description with their identifiers

**Use case diagram:**

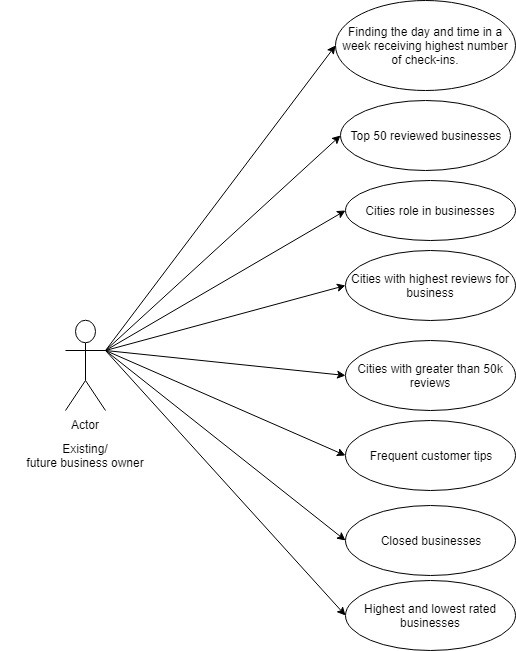


Fig. Use case diagram for the analysis

# **Class diagram**

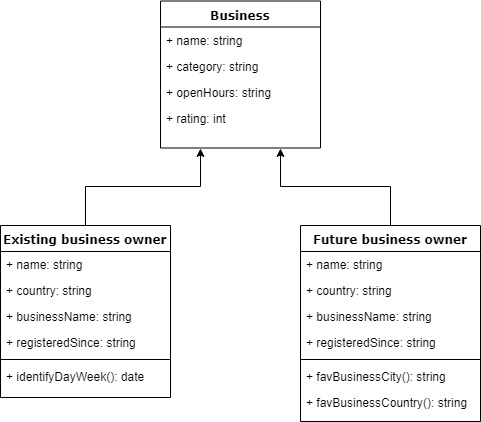


Fig. Class diagram of the analysis

# **Project notes**

The following are the topics I have researched and learnt to work on this project.

**NumPy (Numerical Python)**

This is one of the efficient and widely used libraries in Python with an intention of storing and manipulating numerical arrays (two dimensional) often treated as a list, but in a larger size. As the IDE of Jupyter Notebook itself comes with the packages installed, so the package is imported, and the necessary operations are performed. Unlike the other programming languages, NumPy supports the same type of variables, arrays/lists. Along with this, there are some general concepts in NumPy as defined following. Attributes of arrays: is the process of finding the shape, size, consumption of memory and types of data in the arrays. Arrays Indexing: Setting up the value of individual array elements. Slicing of arrays divides a large section of arrays into smaller for a smoother operation. Joining and splitting of arrays performs the combination and division operations one array.

Apart from the above defined attributes, there are several other factors which are often used in NumPy functions such as finding the minimum and maximum in the list, determining the aggregations in the list such as mean, median, maximum, standard deviation etc. Also, NumPy library supports two dimensional Boolean arrays. To summarize, NumPy efficiently stores the data in the form of arrays and lists to store and manipulate the data efficiently.

**Pandas**

Pandas feeds the dataset we read and load as a Dataframe, similarly like NumPy, Pandas is also imported as a package and the operations are performed. DataFrames can be can be defined as a multidimensional array with columns and rows attached as labels. Pandas consists of Series which is a single-dimensional indexed array which is like the creation of an array which additionally constitutes of indexes included. Pandas is majorly used for mathematical operations such as add(), sub(), mul(), div(), pow(). Pandas is highly efficient when handling the missing data in the dataset such as the rows with missing values as: “Null”, “NaN (Not a Number)”, “NA”. It is easy to find the number of missing values and fill them by substituting the values as average or mean of the columns. For the null values, there are some predefined functions such as isnull(): generates a Boolean mask for the missing values, notnull(): contrasting feature of isnull(), dropna(): which returns a filtered version of the data and fillna(): which returns the copy of the data with missing values filled.

Pandas also supports the usage of methods: len(), isupper(), islower(), ljust(), rjust(), split() which performs the string operations aggressively reducing the load on the system.

**Matplotlib**

This is categorized as one of the necessary and important library for the visualization of our data. Alike the last two libraries, matplotlib is also imported into the IDE for making the plots of the data we have manipulated using NumPy and Pandas. One good advantage of Matplotlib is that is supported by all kinds of operating systems even which has a mediocre graphic system.

To display the plots, plt.show() is the function used. The beauty of Jupyter Notebook is “%matplotlib inline” which displays the visualizations of matplotlib dynamically. There are various attributes which enables the programmers to change the colors of the graphs or increase/decrease the frequencies. Unlike the line plots, there are scatter plots which demonstrates the data as points in a multidimension by using the function plt.plot() and plt.scatter().

Density and contour plots are the visualization techniques which displays the outputs in a three-dimensional perspective. Histograms, Binning and Density are the bar chart type visualization which can interpret the data to a non-technical person with an ease. Alike Matplotlib, Seaborn is one of the visualization library which is used for plotting the outputs. Now, let us take a look at the Machine Learning algorithms and how they are applied through the above-mentioned libraries.

**K-NN classification**

Before getting into this algorithm, it is good to get familiar with noted terms in machine learning. There are two types of machine learning algorithms, namely Supervised and Unsupervised. Supervised machine learning is defined as predicting the future results using the past data. For example, estimating the next year’s annual budget for the organization using the past decade datasets. In supervised machine learning a set of cases are used to train, test and evaluate the model. Each case is comprised of the values of one or more features and a label value. The features are variables used by the model to predict the value of the label. Minimizing the errors between the true value of the label and the prediction supervises the training of this model. Once the model is trained and tested, it can be evaluated based on the accuracy in predicting the label of a new set of cases. The goal is to predict the type or class of the label, which makes the machine learning model a classification model.

The k-nearest-neighbor algorithm is conceptually simple. In fact, there is no formal training step. Given a known set of cases, a new case is classified by majority vote of the K (where k=1,2,3 , etc.) points nearest to the values of the new case; that is, the nearest neighbors of the new case. There are two features, the values of one shown on the horizontal axis and the values of the other shown on the vertical axis. To summarize, each case has a value for the two features, and a class. The goal of the KNN algorithm is to classify cases with unknown labels.

**Preparation of a dataset**

Data preparation is an important step before training any machine learning model. These data require only two preparation steps: Scale the numeric values of the features. It is important that numeric features used to train machine learning models have a similar range of values. Otherwise, features which happen to have large numeric values may dominate model training, even if other features with smaller numeric values are more informative. In frequent cases Z-score normalization is used. This normalization process scales each feature so that the mean is 0 and the variance is 1.0.

Secondly, split the dataset into randomly sampled training and evaluation data sets. The random selection of cases seeks to limit the leakage of information between the training and evaluation cases.

**Exploratory data analysis: Regression technique**

Before creating analytical models, a data scientist must develop an understanding of the properties and relationships in a dataset. There are two goals for data exploration and visualization. First to understand the relationships between the data columns. Second to identify features that may be useful for predicting labels in machine learning projects. Additionally, redundant, collinear features can be identified. Thus, visualization for data exploration is an essential data science skill.

The first goal is to explore a dataset that includes information about the features in the dataset. This type of predictive modeling, in which the attempt to predict a real numeric value, is known as regression. Python supports the matplotlib library; which provides extensive graphical capabilities. Additionally, the Python Pandas library and the Seaborn library add a higher-level graphics capability. Pandas and Seaborn abstract a lot of the low-level details. These features make Python a useful language to create visualizations of your data when exploring relationships between the data features. Further, there is a liberty as we can identify features that may be useful for predicting labels in machine learning projects. As specified earlier, there are various kinds of visualizations to represent the data we have gathered after regression. One of it is Bar charts which is explained below.

**Bar charts**

A bar chart displays frequencies of each category. In most cases, the categories should be ordered by frequency; ascending or descending. Ordering categories by frequency aids in viewer interpretation. The figure and axes are defined using Matplotlib methods. The bar plot is created using the Pandas plot.bar method. Annotations are added to the plot using Matplotlib methods. Since Pandas plotting is built on Matplotlib, it is always possible to add additional plot attributes using methods in this package.

**Histograms**

Histograms are related to bar plots. Whereas, a bar plot shows the counts of unique categories, a histogram shows the number of data values within a bin for a numeric variable. The bins divide the values of the variable into equal segments. The vertical axis of the histogram shows the count of data values within each bin. Some of the variables have distributions that are right-skewed or skewed to the right side. This skewed distribution will affect the statistics of any machine learning model.

**Two dimensional plots**

Two dimensional plots help to develop an understanding of the relationship between two variables. For machine learning, the relationship of greatest interest is between the features and the label. It can also be useful to examine the relationships between features to determine if the features are co-variate or not. Such a procedure can prove more reliable than simply computing correlation when the relationship is nonlinear.

**Classification technique**

Visualization for classification problems shares much in common with visualization for regression problems. Colinear features should be identified so they can be eliminated or otherwise dealt with. However, for classification problems we will look for features that help separate the label categories. Separation is achieved when there are distinctive feature values for each label category. Good separation results in low classification error rate.

Class imbalance: It means that there are unequal numbers of cases for the categories of the label. Class imbalance can seriously bias the training of classifier algorithms. It many cases, the imbalance leads to a higher error rate for the minority class. Most real-world classification problems have class imbalance, sometimes severe class imbalance, so it is important to test for this before training any model. Also, classes can be classified because of categorical and numerical features which is entirely dependent on the perspective we are looking at the data.

**Data preparation**

Data preparation is a vital step in the machine learning pipeline. Just as visualization is necessary to understand the relationships in data, proper preparation or data munging is required to ensure machine learning models work optimally. The process of data preparation is highly interactive and iterative. A typical process includes at least the following steps: Visualization of the dataset to understand the relationships and identify possible problems with the data. Data cleaning and transformation to address the problems identified. It many cases, the initial step is then repeated to verify that the cleaning and transformation had the desired effect. Construction and evaluation of a machine learning models. Visualization of the results will often lead to understanding of further data preparation that is required; going back to first step.

**Working on the missing values**

Missing values are a common problem in data set. Failure to deal with missing values before training a machine learning model will lead to biased training at best, and in many cases actual failure. The Python scikit-learn package will not process arrays with missing values. There are two problems that must be deal with when treating missing values: Firstly, find the missing values. This can be difficult as there are no standard way missing values are coded. Some common possibilities for missing values are: Coded by some character string, or numeric value like -999. A NULL value or numeric missing value such as a NaN. The following explains how to treat the missing values: by removing the features with substantial numbers of missing values. In many cases, such features are likely to have little information value. Second is to remove rows with missing values. If there are only a few rows with missing values, it might be easier and more certain to simply remove them.

Third is to impute values. Imputation can be done with simple algorithms such as replacing the missing values with the mean or median value. There are also complex statistical methods such as the expectation maximization (EM) or SMOTE algorithms. Finally, using nearest neighbor values. Alternatives for nearest neighbor values include, averaging, forward filling or backward filling.

**Removing duplicate rows**

Duplicate cases can seriously bias the training of machine learning models. In simple terms, cases which are duplicates add undue weight to that case when training a machine learning model. Therefore, it is necessary to ensure there are no duplicates in the dataset before training a model. One must be careful when determining if a case is a duplicate or not. It is possible that some cases have identical values, particularly if most or all features are categorical. On the other hand, if there are columns with values guaranteed to be unique these can be used to detect and remove duplicates. Another consideration when removing duplicate cases is determining which case to remove. If the duplicates have different dates of creation, the newest date is often selected. In the absence of such a criterion, the choice is often arbitrary.

**Scaling numeric features**

Now that the dataset is split, the numeric feature column must be re-scaled. Rescaling of numeric features is extremely important. The numeric range of a feature should not determine how much that feature determines the training of the machine learning model. For example, consider a data set with two features, age in years, typically measured in a few tens, and income, typically measured in tens or hundreds of thousands. There is no reason to believe that income is more important than age in some model, simply because its range of values is greater. To prevent this problem numeric features are scaled to the same range. There are many possible scaling methods. One simple method is known as Min-Max normalization.

# **Test cases**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Use case** | **Function to test** | **Initial system state** | **Input** | **Expected output** | **Actual output** |
| Existing business owners | Checking which time and day of a week more check-ins are observed | HighestCheckins page in a notebook is displayed with code | Run the whole sheet using “Run all” option in the notebook | Outputs are displayed in the form of bar charts | Outputs are displayed in the form of bar charts |
| Future business owners | Checking cities with highest reviews and best ratings for their businesses. | BusinessAnalysis page in a notebook is displayed with code | Run the whole sheet using “Run all” option in the notebook | Outputs are displayed in the form of bar charts | Outputs are displayed in the form of bar charts |

# **Conclusion**

After the complete analysis, the following stats were observed meeting our requirements:

1. Saturday at 19:00hrs recorded highest check-ins for businesses.
2. Results shown that “Las Vegas” city has shown doing the best business and receiving a better overall rating compared with other cities of USA. Which says that, future business owners have highly chances of establishing their business in Las Vegas irrespective of the category of the business they chose. Also, it can be deduced that majority of the customers rated “Happy” and “Cool” to the business and a very little portion of customers rated their disappointment as “Bad”.
3. Also, the analysis indicates that majority check-ins were recorded during 16:00hrs to 23:00hrs during the weekends; Friday and Saturdays. This can help the existing business owners to allot additional staff during these hours to serve the customers better.
4. Coffee shop “Starbucks” received best ratings along with a positive feedback from customers.

# **References**

1. VanderPlas, J. T. (2017). Python data science handbook: Essential tools for working with data. Sebastopol, CA: O’Reilly Media.
2. Richert, W. & Coelho, L. P. (2013). Building Machine Learning Systems with Python. Birmingham, UK: Packt Publishing Ltd.
3. Yelp, Inc. (2018, August 02). Yelp Dataset. Retrieved from https://www.kaggle.com/yelp-dataset/yelp-dataset/home.
4. Project Jupyter. (n.d.). Retrieved from <http://jupyter.org>