Yelp Data Analysis

# Introduction

# Data Set Review

# Exploratory Analysis

### Business Data Analysis

### Reviews Analysis

# Filtering Data and Further Analysis

# Polarity Scores Calculations

# Observations and Results

# Introduction

Yelp is an online business which provide the platform for customers to write reviews and provide a star-rating. One of the research indicates that a one-star increase led to 45% increase in revenue of independent Businesses.

The main objective of our analysis is to conduct deep analysis on 3 different service based Businesses. These businesses are Home Services, Local Services and Home & Garden. We will figure out what makes a good service Business and what concerns customers, and then we shall make recommendations of the future improvement and profit growth.

We will be analyzing users reviews and figuring out the reasons why customers love or dislike the Services. As a example, there may be great reviews primarily due to the friendly service, or negative reviews about high price. We shall also compare among those 3 different cuisine types and figure out differences from reviews and gain valuable insights to make customized recommendations to different types of Services.

# Data Set Review

Yelp has provided the dataset publically to all their users. The Yelp dataset is a subset of the businesses, reviews, and user data for use in personal, educational, and academic purposes.

The dataset is available in JSON format.

We take Json files and convert them into csv for our analysis. We use two files :

Business : yelp\_academic\_dataset\_business.csv

Reviews : yelp\_academic\_dataset\_review.csv

Types of Attributes in Yelp Business Datatypes

* business\_id: ID of the business
* name: name of the business
* neighborhood
* address: address of the business
* city: city of the business
* state: state of the business
* postal\_code: postal code of the business
* latitude: latitude of the business
* longitude: longitude of the business
* stars: average rating of the business
* review\_count: number of reviews received
* is\_open: 1 if the business is open, 0 therwise
* categories: multiple categories of the business

Types of attribute in Yelp Data Reviews

* review\_id: ID of the review
* user\_id: ID of the user
* business\_id: ID of the business
* stars: ratings of the business
* date: review date
* text: review from the user
* useful: number of users who vote a review as usefull
* funny: number of users who vote a review as funny
* cool: number of users who vote a review as cool

Both the data set are connected using business\_id.

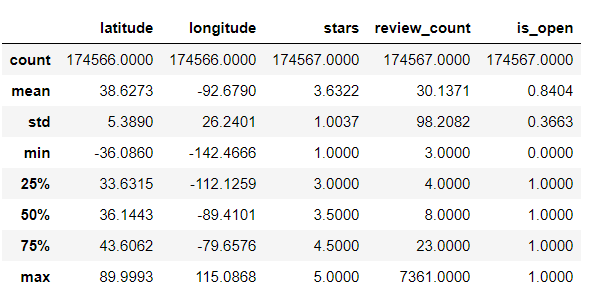
We will use both the dataset and explore them and later on merge them on business\_id to do a collective analysis.

# Exploratory Analysis

### Exploring Business DataSet

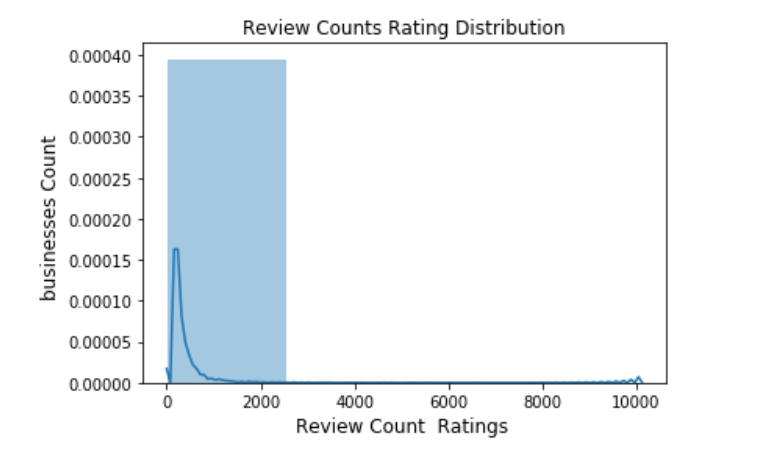
In this section we will analyse all sorts of attributes present in Business as well as reviews dataset.

#### Central Tendencies of Attributes in Business Data Set.



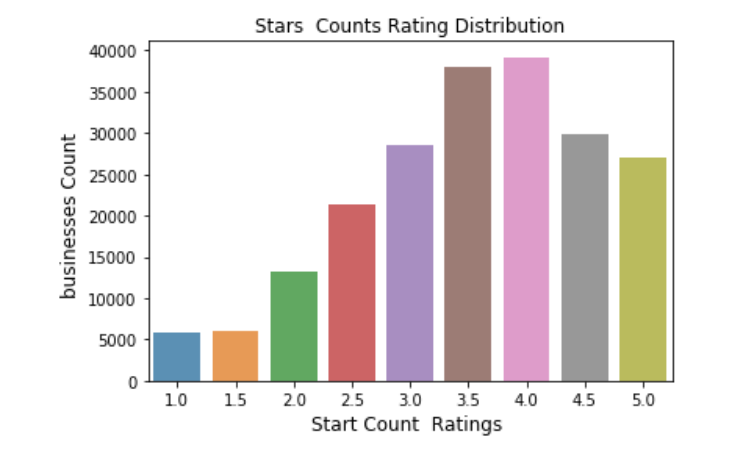
This table represent the stats for all attributes. The field of our interest is mainly stars and review\_count we see that average rating is around 3.63 and average reviewcount any business get is around 30. Given that we have around 209393 businesses.

#### Review Counts Distribution



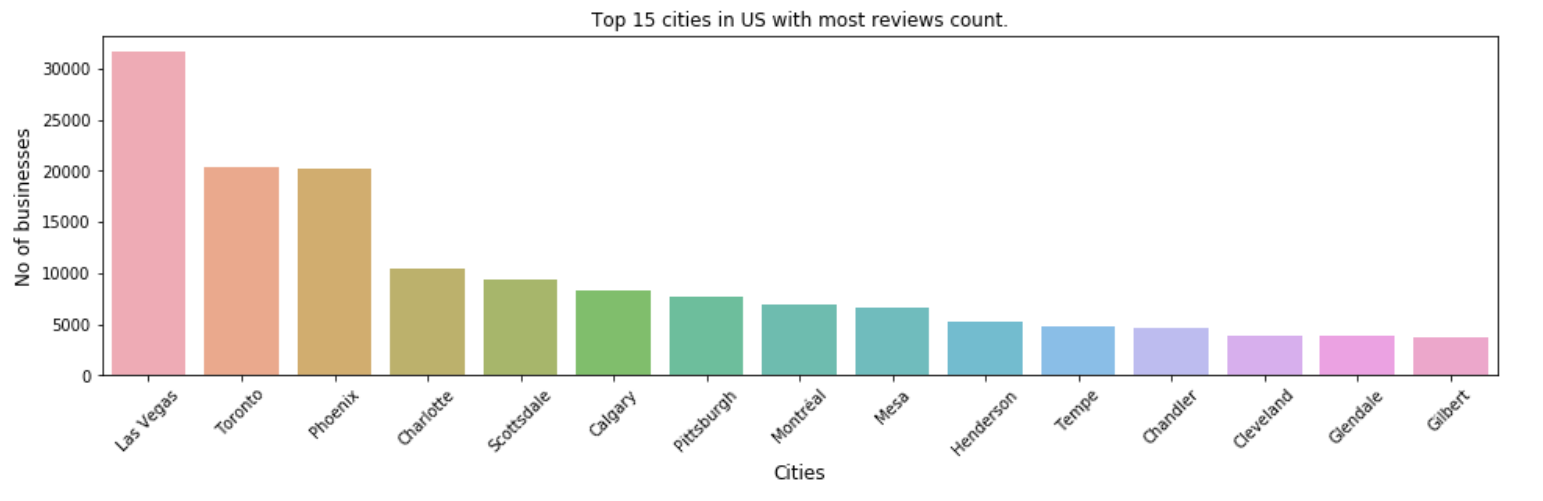
We see that review counts is not very uniform mostly concentrated nearabout 2000 range.

#### Rating Counts Distribution



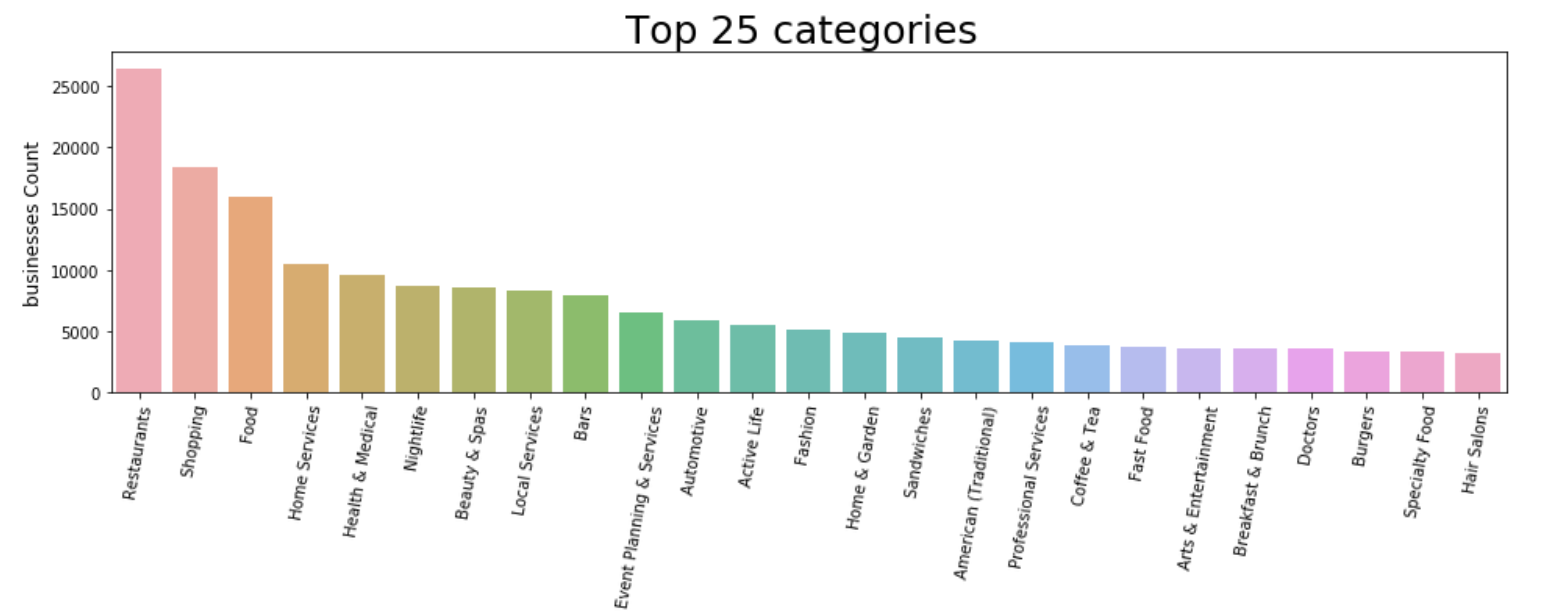
We see that Star counts tends to form somewhat similar to a skewed gaussian where we see the peak is around 3.5-4 and then it decreases.

#### review counts per city (Top 15 cities)



Las Vegas is sure the most active cities(Happening) when it comes to reviews followed by Pheonix and Toronto.

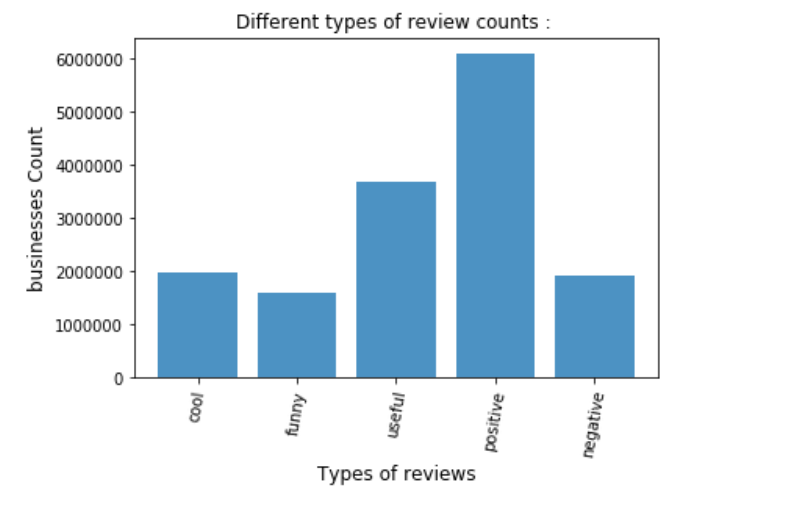
#### Types of categrories of businesses



The top categories are Restaurants , Shopping ,Food,Home Services etc. Business can be categorizes in Product based and Service based businesses we see that majority of business fall inside product based businesses. Our area of interest here would be Service based Businesses.

### Exploring Reviews DataSet

#### Types of Reviews Counts of businesses



We see we have high no of reviews as positive as per our initial analysis . People find a more than half of reviews as useful.

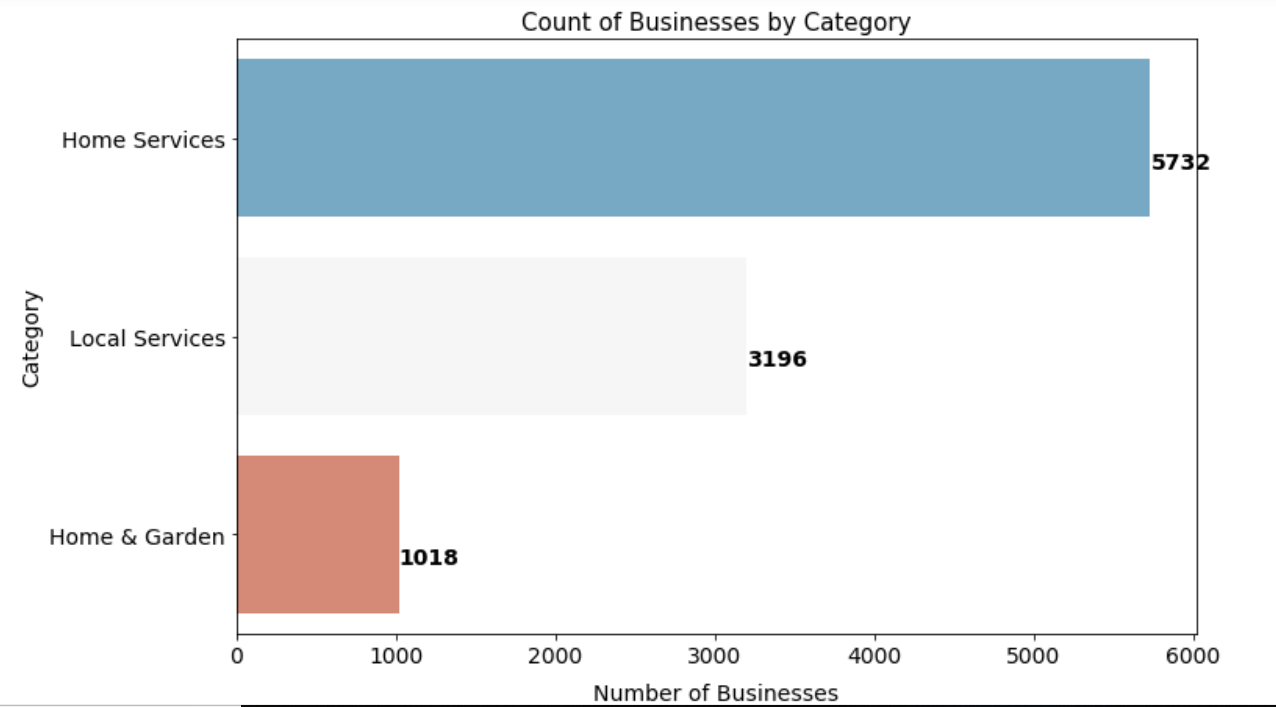
# Filtering Data and Further Analysis

In this section we will filter the data and based on our chosen category {Home Service , Public Service, Home & Garden}.

Filtering Criteria:

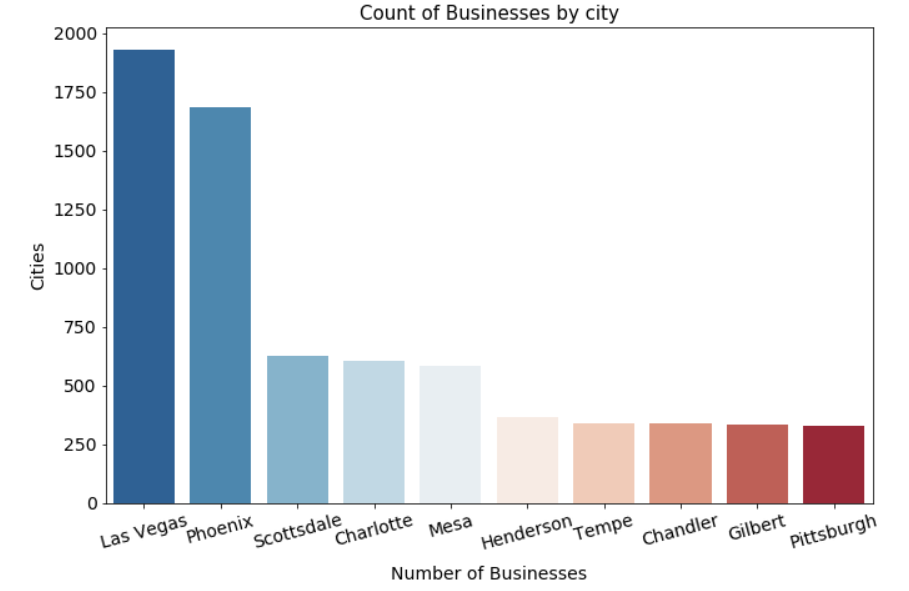
1. We also filter all 50 states from US .
2. We filter 3 category from all the categories.

#### Count Business by categories:



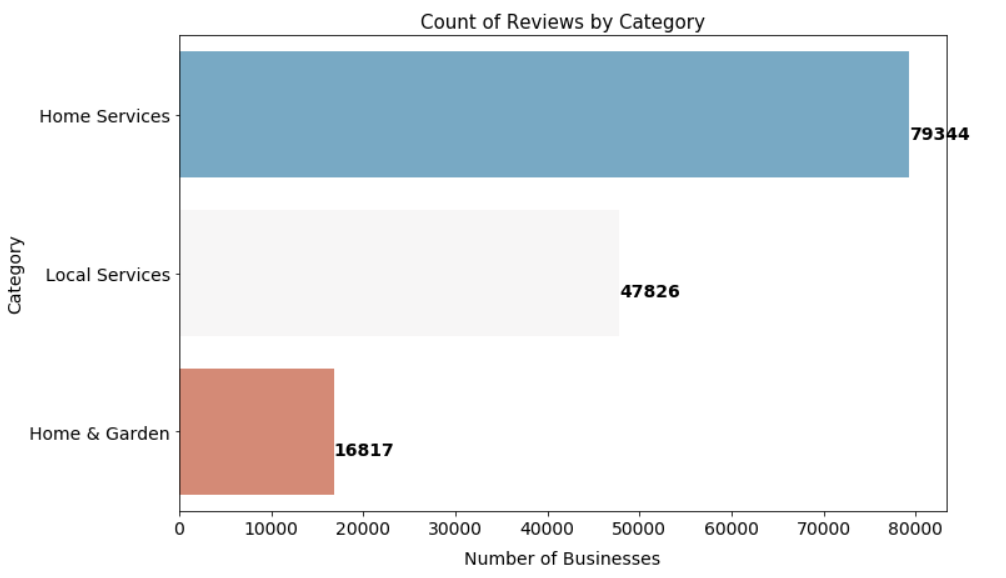
We have about 5732 home services and 3196 local services and about 1018 home and garden to perform analysis.

#### Count Business By city:



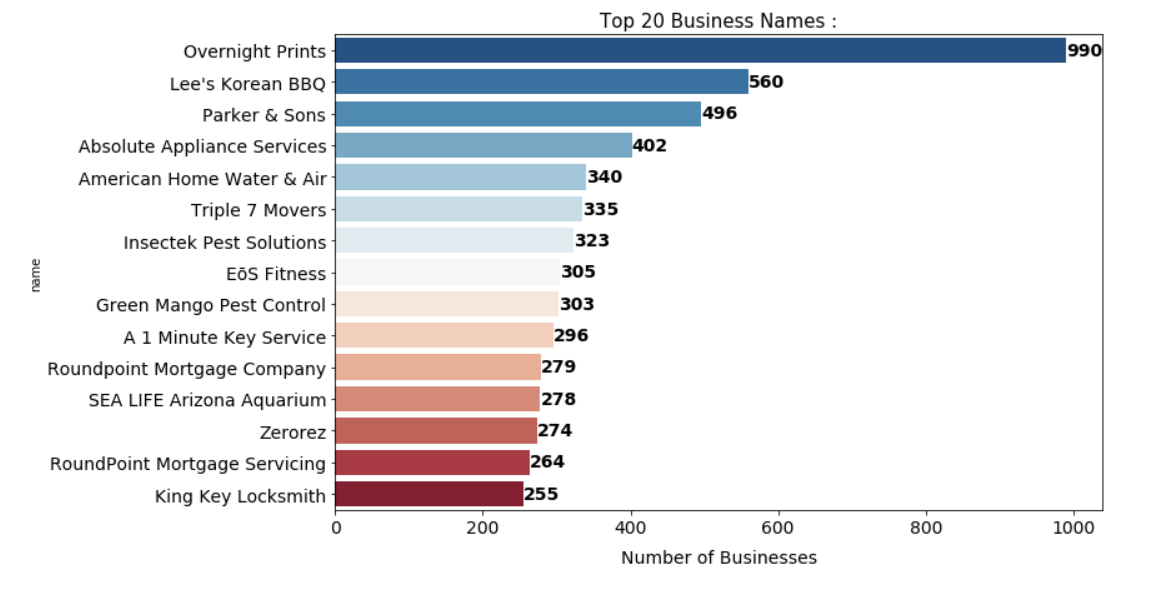
Las Vegas is again here the top again followed by phoenix and then Scottsdale.

#### Count of reviews by category :



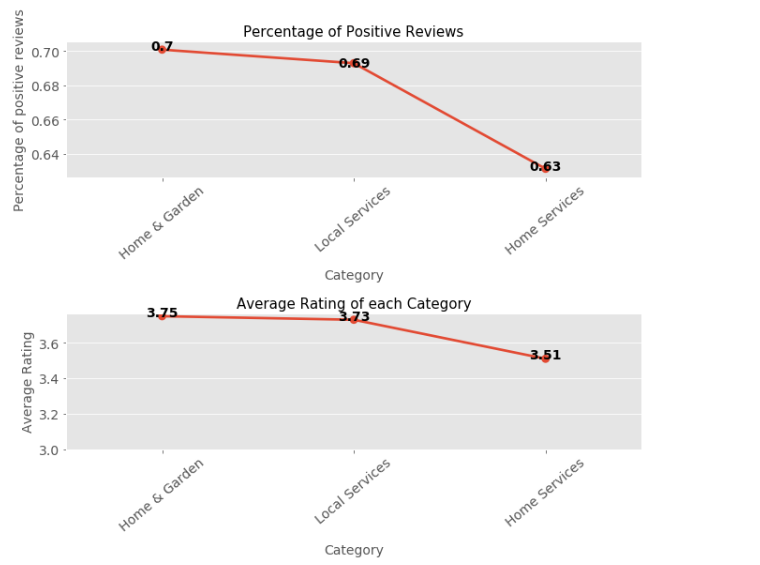
Home service has received the largest sum of reviews and the local service has about the half of it .

#### Top Business names in these category :



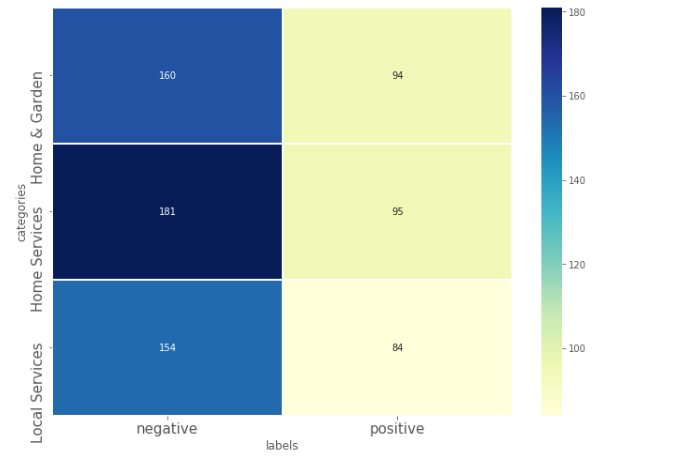
Top services providers are Overnight Prints,Parker and Sons , Triple 7 Movers. etc

#### Distribution of Positive and Negative Reviews in DataSet.



We merge both the dataset and calculate the percentage of positive and negative reviews. If the rating is >4 it is positive and if <3 then negative. We see that Most positive reviews are recieved by Local services whereas Home and garder has recieved less positive recievews upon further analysis we will get to know what we can be done to benefit these business. An interesting graph for average rating for each category average rating for all businesses is above 3.5 that means Businesses are doing good.

#### Average length of words used for reviews



As per the heat map we can see people express their negative reviews with more length.

# Polarity Scores Calculations

Now we clean the text of reviews and perform SVC to calculate word scores:

Linear SVC Model :

The objective of a Linear SVC (Support Vector Classifier) is to fit to the data you provide, returning a "best fit" hyperplane that divides or categorizes the data.

After getting the hyperplane, you can then feed some features to your classifier to see what the "predicted" class is. This algorithm is best for our use case because we feed in class (Positive , Negative ) to categories our data.

Here feature set is our text and Class is positive and negative.

Steps :

1. Removing all punctuation marks and unwanted characters from reviews.
2. Loading Most used Negative and Positive words from textfiles
3. Filter only those words from reviews which are of use (i.e only positive and negative words).
4. Creating test and train dataset. With ratio of 0.9
5. Calculating Polarity Scores
   1. Building features using nltk countVectorizer .
   2. Building LinearSVC model using the features and class.

𝑝𝑜𝑙𝑎𝑟𝑖𝑡𝑦\_𝑠𝑐𝑜𝑟𝑒 𝑡, 𝑐 = 𝑠𝑐𝑜𝑟𝑒(𝑡)×𝑡𝑜𝑡𝑎𝑙\_𝑓𝑟𝑒𝑞𝑢𝑒𝑛𝑐𝑦(𝑡, 𝑐)/𝑛𝑢𝑚𝑏𝑒𝑟\_𝑜𝑓\_ 𝑟𝑒𝑣𝑖𝑒𝑤𝑠(𝑐)

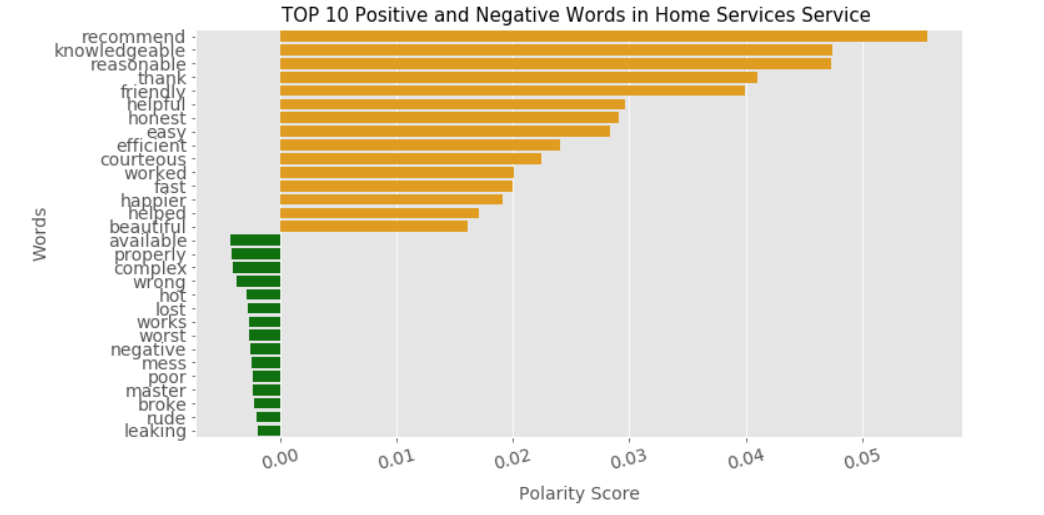
Polarity Score -> s the index for measuring how essential word 𝑡 is among services of type 𝑐.

Total Frequency-> is the total frequency of word 𝑡 in all reviews of type 𝑐 services.

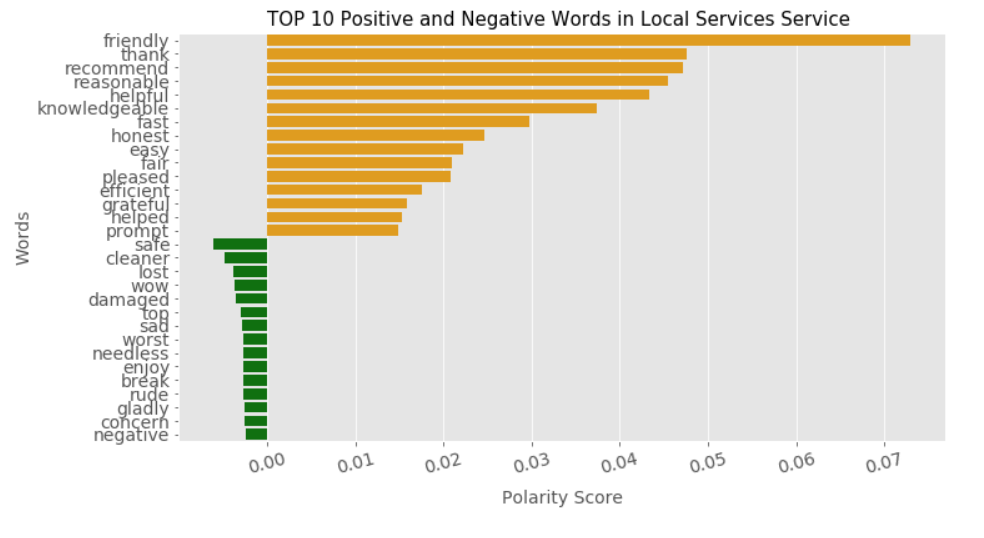
𝑛𝑢𝑚𝑏𝑒𝑟 𝑜𝑓 𝑟𝑒𝑣𝑖𝑒(𝑐) is the total number of reviews of type 𝑐 services.

SVM model actually calculate a total score for each review and this score indicates how satisfied or discontented the customer .Polarity score show how much a word contributes to score of all services. The polarity value of certain words tell us how these business get impacted.

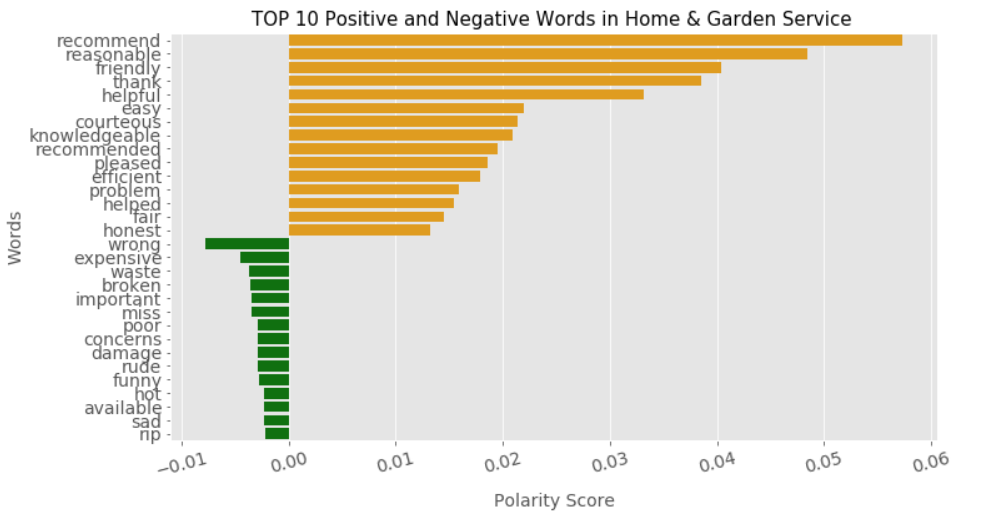
#### Home And Services Category:



#### Local Services Category:



#### Home and Garden Category:



For home and services we notice words like friendly , knowledgable,helpful,reasonable,honesty ,courteous which adds to the positive polarity scores. This simply indicates that users like home services who are polite and are honest . The most impactful review is being friendly users appreciate when service providers are friendly in nature the polarity will increase by 0.0505. Also services whose agents have good knowledge users also appreciate and like that and doing work in time and before that is always appreciated.

Let’s talk about negatives reviews now.

People have used words like wrong (probably wrong installations) ,wrong work , Broken , Poor (maybe quality ,quantity).Support ( customer support etc),rude.,expensive.

#### How can Services Improve?

Services need to focus on these key points to improve the overall experience of user. Today in this digital world every business runs online. People judge the products or services with the help of reviews by other users. A good review increases the sales by many fold be it any business. Businesses should focus on getting more good reviews by improving customer experiences.

* Services should invest time and money to train their employers who provides these services to end users. A bad experience can impact the sales largely. The official should be well behaved kind and courteous.
* Services should hire employee which have good knowledge of the field . A good assessment is must otherwise the lack of knowledge can irritate the customers when these employee provide these services to end customers.
* Most business fail due to a bad customer support . Service based companies must pay close attension to developing a good support system.
* What we see when a company enters a market they launch with less prices and then as they began making revenue the prices go up that’s why a few loyal customers can look elsewhere .
* Service companies should make sure if they are delivering any products by a third party as well make sure they provide a good tracking system and the products are delivered safe. If someone receives a broken product then the customer would never come back.