BIG DATA MANAGEMENT

ANALYSIS OF BUSINESS TRENDS USING YELP

Group G - Abhilash Sunkam, Lokesh Koppaka, Vishal Lella, Anuja Patil

Abstract:
Yelp is a local business directory service and review site with social networking features. It allows users to give ratings and review businesses. It publishes crowd-sourced reviews about local businesses regarding Restaurants, Department Stores, Bars, Home-Local Services, Cafes, Automotives etc. There is an increasing number of users and visitors, who use it as a local search engine, giving feedback and starting or advertising their own businesses. Yelp has 148 million reviews so it is a big repository for making interesting insights which will be helpful in making better business decisions. This huge data can help us to dig out vital information like providing the best restaurants in a particular city, finding out the factors behind the prices of some businesses going high etc. We can identify certain patterns in the trends, perform some predictions with existing data, which can give entrepreneurs and business persons an idea about investing in particular areas and understand users' interests. All facilities are there for businesses; it is just a matter of unlocking them, giving it a go and playing with the tools. The objective of our project is to perform data processing on Yelp dataset and derive interesting insights to help existing and future business owners. Some such interesting insight like business distribution per rating as heat map, business scatter plot resembling the exact city map , popular business, top cities are presented.

Introduction:
Yelp, a local search crowdsource review forum, generates and holds a huge amount of local businesses datasets like restaurant ratings, recommendations, customer reviews etc. In the near past, Yelp announced a challenge to conduct research or analysis on their data and share their discoveries with them. This resulted in the huge involvement of various students and researchers resulting in a lot of publications on Natural Language Processing, Sentiment Analysis, Graph Mining, photo classification, region food choices analysis etc. This motivated us to look into various such literature approaches and derive one of our own to come with analytical insights which make the goal of our project.

Related work:
To give a kickstart to our project, we went through few research papers which gave us an idea of how a dataset can be proved to be useful for different category people like businessmen, entrepreneurs, food lovers, thrifty citizens, etc. We went through some papers which described the methods used by authors to calculate useful information for entrepreneurs to initiate their business, forecasting trends based on current ones to improve their work. Some of the methods are mentioned here :

To examine how a user would rate his/her experience with a business, [2] identified whether user rating is influenced by extrinsic factors surrounding a restaurant and intrinsic factors that lie

within a restaurant. For instance, the level of satisfaction or dissatisfaction reflected by a user in their rating for a business could be driven by various factors such as service, food quality, quantity with respect to the price of the meal which are the intrinsic factors and hygiene at the locality, the popularity of the neighborhood in which restaurant is situated comes under extrinsic factors.

To identify different features for restaurants of different cuisines an innovative method was proposed in [2]. The method was based on a high-accuracy SVM model, calculating word scores and measuring the polarity. The essential features they discovered might not only help customers to choose their favorite cuisine, but also provide restaurants with their advantages and shortages. On the other hand, similar procedures can be reproduced for reviews and comments in other areas like movie reviews and social media posts. That would be the anticipations in the future: gathering opinions from people, extracting information from opinions and generating suggestions from information.

In [7], the authors have guessed the rating of the businesses while reducing the reviewer bias by using naive Bayes. They divided different reviews into word clouds and word frequencies. While using the document term matrix and looking at the data, the star ratings were predicted based on Yelp reviews.

To demonstrate ratings vs quality trends of the restaurants using Yelp dataset, the authors in [4] have designed models using Linear second order differential equation like the one given below.

$$m\frac{d^2}{dt^2}x + c\frac{d}{dt}x + kx = q\cos(\omega t)$$

x = change in rating; q = quality

The solution to the differential equation tends to have periodic behavior which demonstrates the rising and falling as momentum for businesses. This helps in deriving insights such as identifying food preferences, preference for ethnic restaurants, top ethnicity food preferences vs regions.

The comparison between the franchise and non-franchise quick-serve restaurants in selected states in the US was made by the authors in [5]. Both numeric and text reviews were analyzed. The authors tested differences in the mean values between the franchise and non-franchise locations using a two-sample t-test. The authors have found correlations between the ratings and the number of guests. These numeric ratings were analyzed using relatively traditional statistical techniques like t-tests and correlations. Word clouds were generated from the Yelp text reviews for the quick-serve franchises and to compare reviews based on the rating of the restaurant.

In [1], the evidence was presented that, Yelp data can complement government surveys by measuring economic activity in real time at any geographical scale. Changes in the number of businesses and restaurants reviewed on Yelp can predict changes in the number of overall establishments and restaurants in County Business Patterns. To assess the overall predictive power of Yelp, the authors have used a random forest algorithm to predict the growth in CBP establishments. They repeated this exercise using Yelp and CBP data at the restaurant level.

They then found that Yelp is more predictive in some industries than others using a regression framework.

Considering the approaches mentioned above, in our implementation we have used business types, longitude, latitude, stars and review count attributes of the dataset to derive insights from the data like the regional popular restaurants, and rating trends of the restaurants. These insights are derived using simple aggregation operations over the attributes. We have found out the number of business categories and shown the top 20 categories trending currently. We have also filtered on top 2 regions for businesses and aggregated over the rating to generate popular restaurants.

Motivation:
Yelp dataset has a rich variety of the ratings, comments, and metadata of businesses. Every day some amount of data is created using the internet. Storing huge amount of data and retrieving knowledge/insights out of it is challenging task these days. After researching for several topics, we came across **Yelp Data Challenge** - *The challenge is a chance for students to conduct research or analysis on our data and share their discoveries with us*. We decided to take up this challenge and worked to gain interesting insights that would help business make better decisions. We will describe how we are storing and process data using big data tool and derive interesting analysis.

Dataset:
The dataset we worked on is a subset of Yelp's businesses, reviews, and user data. We have 7 data tables - business, business_attributes, business_hours, checkin, review, user and tips. We have mostly worked on business tables. In total, there are

- 5,200,000 user reviews
- Information on 174,000 businesses
- The data that spans 11 metropolitan areas

The dataset covers only the cities in the US. We restricted this dataset because of three reasons:

- Reducing the data size improves computation time
- Avoid relationships being masked due to differing dynamics of different cities
- Satisfy curiosity about a city of personal interest

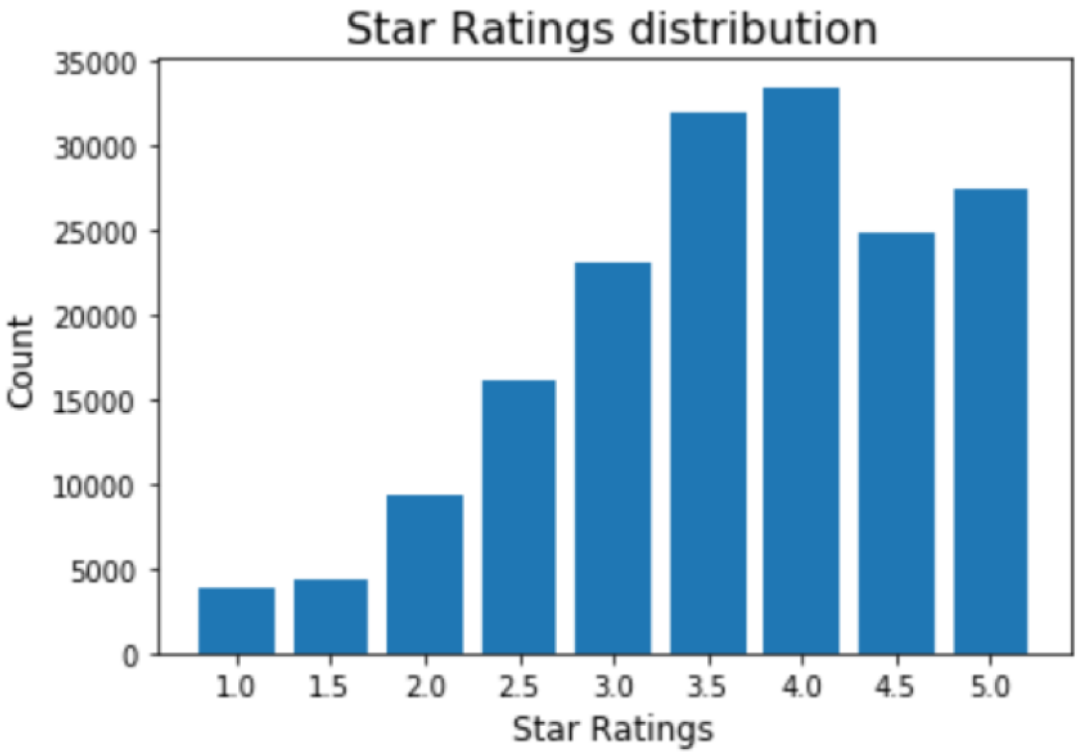
Problem Definition:
Based on the dataset, we defined our problem statement to be - finding businesses trends among different cities across US and analyzing different business distribution in top cities. We have also visualized them on maps and charts to get a clear picture of trends which might be useful for future entrepreneurs.

Approach:
The dataset is obtained from the Yelp's official website for our study. The original data is in JSON format. We chose to work on version 6 of the dataset, which is in CSV, as we did not want the nature of our data to be unstructured. The business dataset has many columns but we considered few columns which are necessary for our analysis (dimensionality reduction).

The first thing in our approach after obtaining the data was to perform some preprocessing on it. The CSV files contained extra characters which we needed to remove to ensure proper structure. We used Apache Spark - PySpark - for data processing to explore the implementation in Spark and also to get experience with this framework. We used MatPlotLib library of Python and Jupyter notebook to perform visualization.

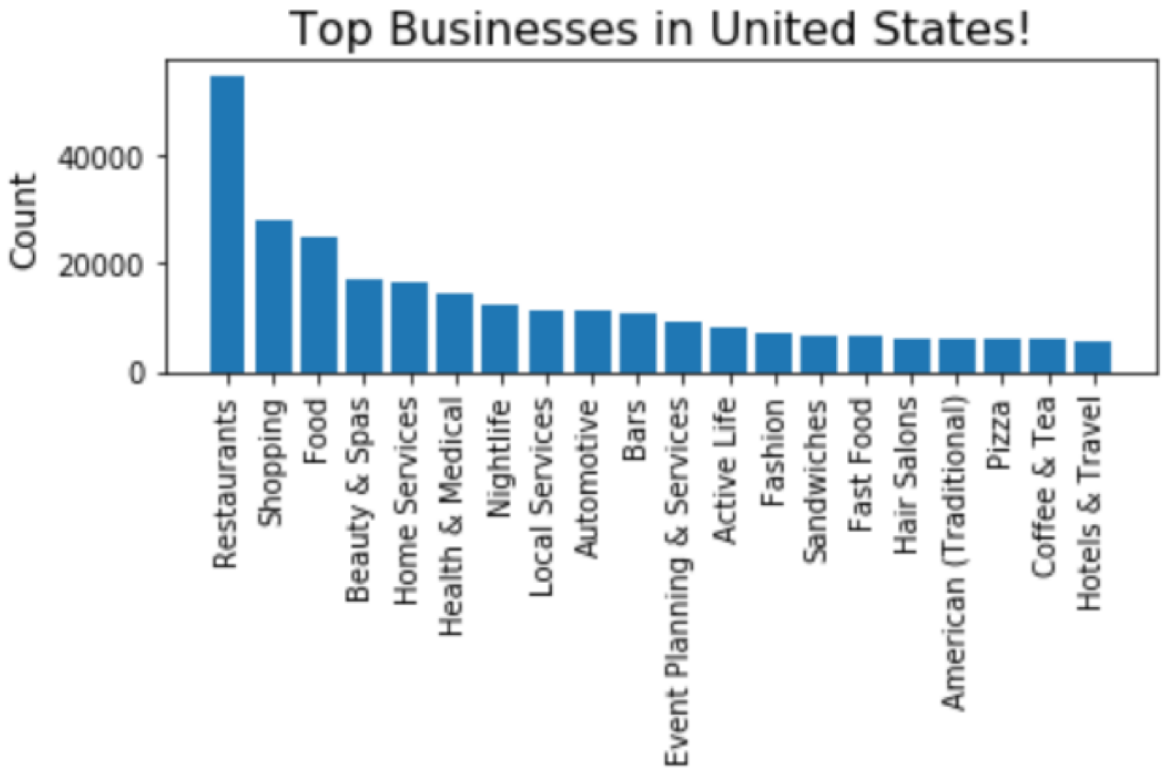
Experiments and Results:
Data Processing and Insight Generation:

- We gathered information about the average star ratings Americans give to businesses and how the ratings are distributed which indicates how good the businesses are



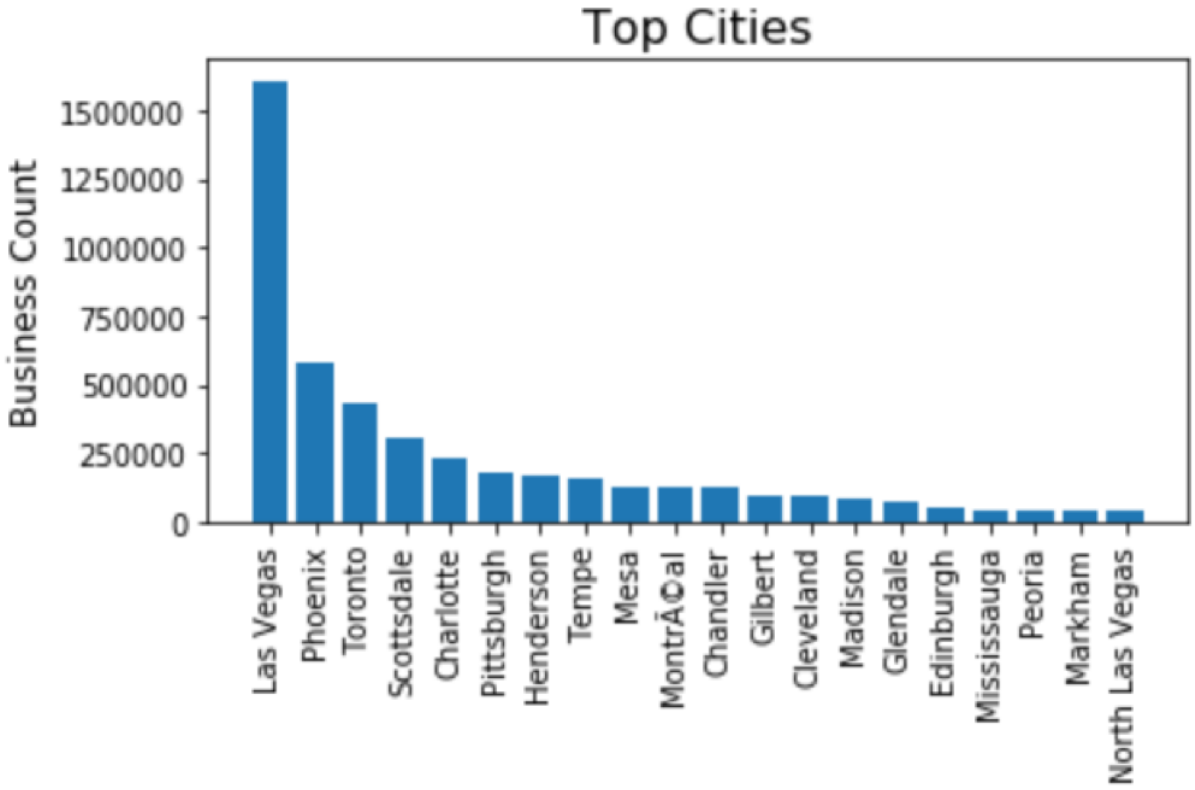
From this graph, we get an idea that Americans usually give 4 star rating to businesses which indicates that they are enthusiastic about trying different things.

- We plotted the popular business flourishing in the US



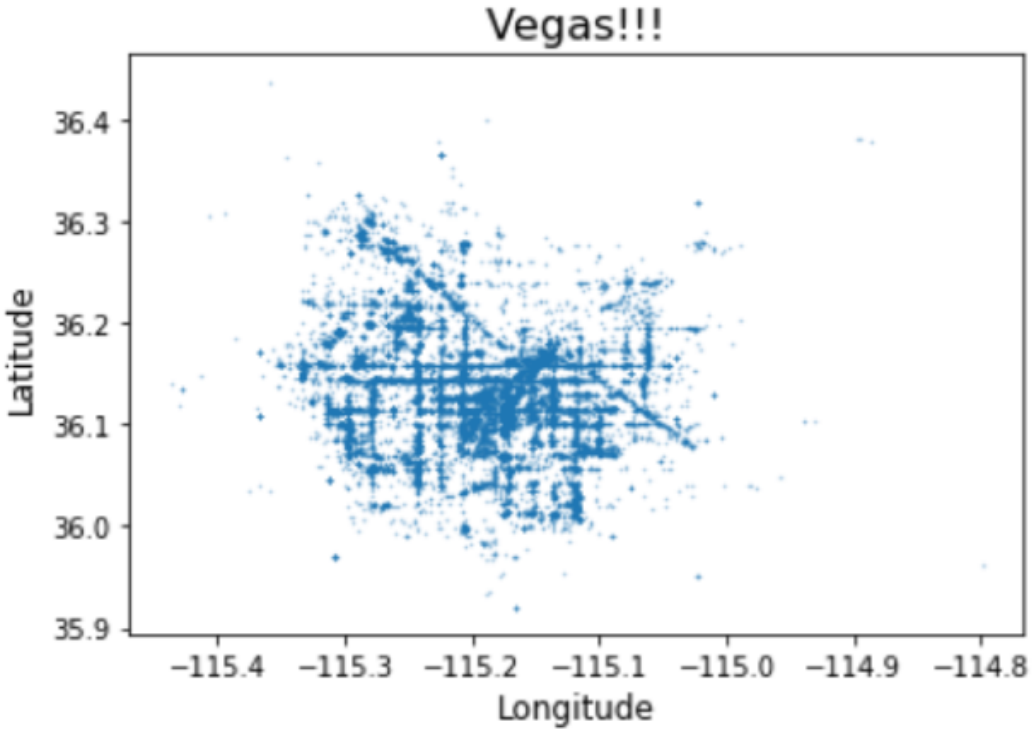
We have just plotted top 20 businesses. There are about 1294 business categories.

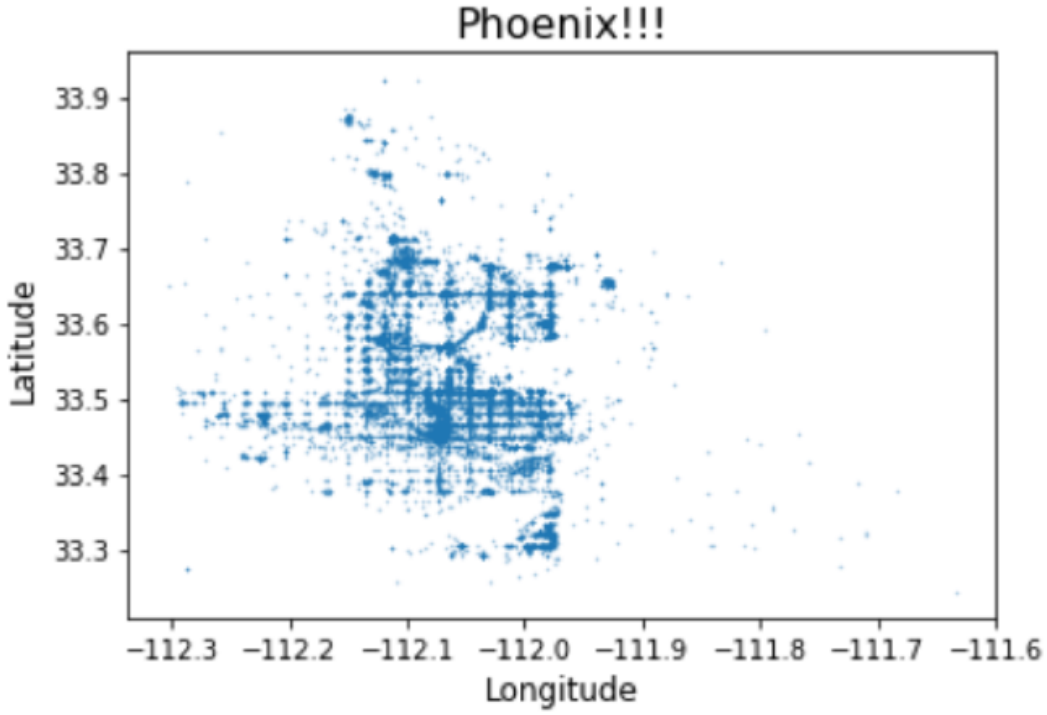
- We then analyzed the top cities having comparatively large number of business reviews



We can see that the top 2 cities having more number of reviews are Las Vegas and Phoenix. So we dived into the details of these 2 cities.

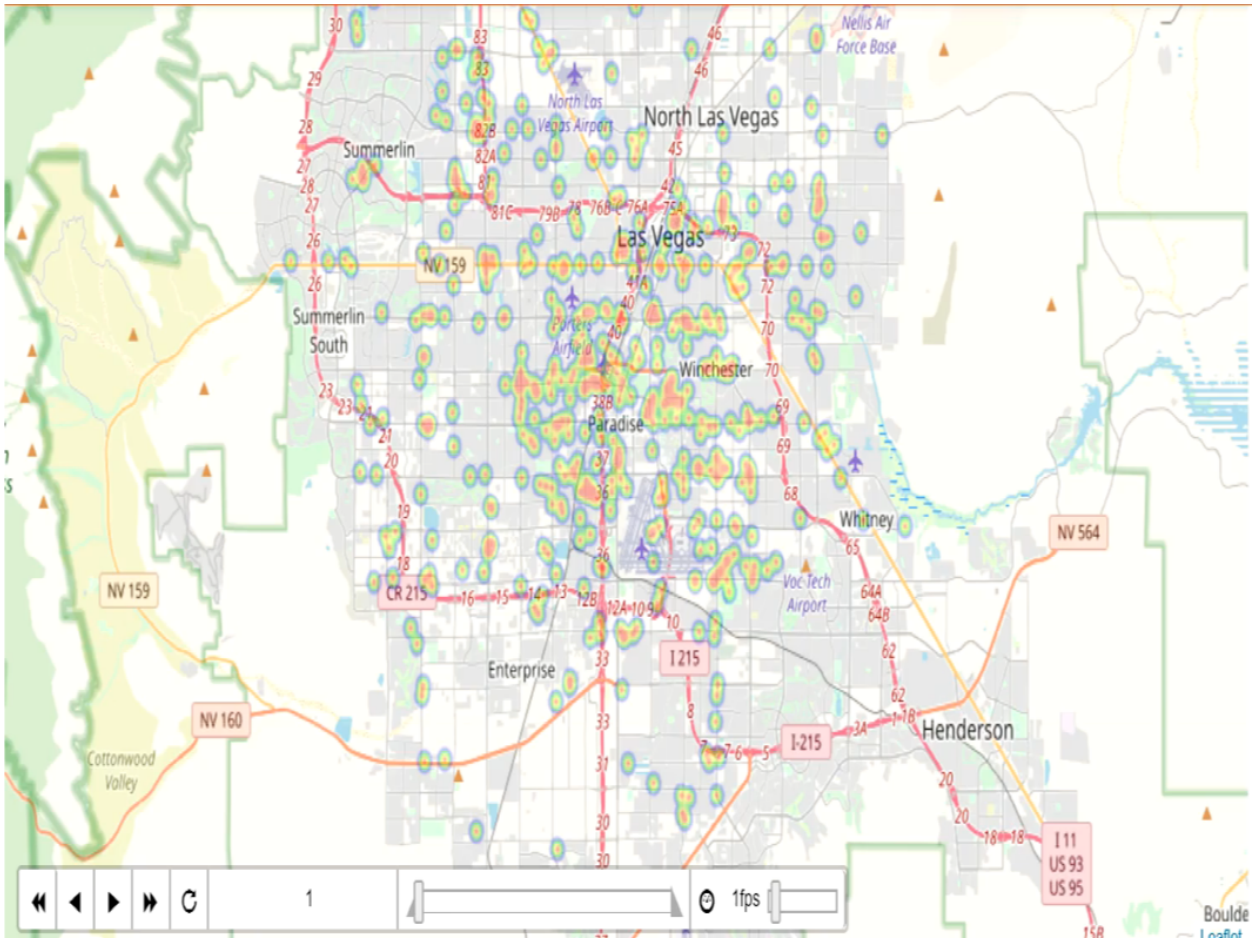
- We visualized the top 2 cities on scatter plots





Both the maps look quite similar to the actual maps of respective cities. This indicates that the businesses are distributed everywhere in those cities.

- Now, we analyze how the businesses are distributed across Las Vegas. We visualize this using a heat map of the city



Each spot is a different rating business. So, the darkest spots indicate top star rated business and lightest spots indicate lowest star rated business. This can give an idea where businesses are in demand. For each category, we can show a heat map to differentiate among businesses.

Conclusion:

The project demonstrates the importance of big data systems along with big data repositories to derive interesting analysis. We have presented useful insights like Rating Distributions, Popular business, Top Cities vs business review, Top cities to invest and a heat map around one popular city. These insights can help any entrepreneur in having the right investment in the right place.

Apart from providing insights, the project has given us an opportunity to explore various big data frameworks like Hadoop, Spark, hands-on experience on these frameworks and querying big data set.

Future Work:

We will consider building a web interface for users for allowing them to browse results. Our project can be extended for performing sentiment analysis based on reviews. Sentiment analysis can be useful for predicting trends according to users' interests. This is a challenging task and we might need to spend a little more time researching and implementing. This section of enhancements includes natural language processing, sentiment analysis, model generation, and a few other challenging machine learning techniques.

We can also build a success model which evaluates whether a new business will succeed or fail. This again requires knowledge of machine or deep learning to predict the scenarios.

Our project can also be extended to find the flourishing businesses based on users' reviews.

Acknowledgement:

We take this opportunity to acknowledge Prof. Ahmed Eldawy for teaching the big data concepts in an interactive way which helped us to think about approaching a problem. We would also like to acknowledge our TA Saheli Ghosh for being available whenever we had issues in our project.

References:

1. *Nowcasting the Local Economy: Using Yelp Data to Measure Economic Activity*: Edward L. Glaeser Hyunjin Kim Michael Luca
2. *Identifying Restaurant Features via Sentiment Analysis on Yelp Reviews*: Boya Yu, Jiaxu Zhou, Yi Zhang, Yunong Cao
3. *Analysis of Mexican Restaurant Reviews with Yelp Data Challenge Data set*: Coursera
4. *Analysis of Yelp Reviews*: Peter Hajas, Louis Gutierrez, Mukkai S. Krishnamoorthy
5. *Does Yelp Matter? Analyzing (And Guide to Using) Ratings for a Quick Serve Restaurant Chain*: Bogdan Gadidov, Jennifer Lewis Priestley
6. *Sentiment analysis for Yelp review classification*: Vivian Rajkumar
7. *Using Naive Bayes to predict ratings based on Yelp reviews*: DJMcClellan
8. *Prediction of Useful Reviews on Yelp Dataset*: Yanrong Li, Yuhao Liu, Richard Chiou, Pradeep Kalipatnapu
9. *Sentiment Analysis of Big Data*: Methods, Applications, and Open Challenges: Shahid Shayaa, Noor Ismawati Jaafar, Shamshul Bahri, Ainin Sulaiman, Phoong Seuk Wai, Yeong Wai Chung, Arsalan Zahid Piprani, Mohammed Ali Al-Garadi
10. *Big Data: Technologies, Trends and Applications*: Sudhakar Singh, Pankaj Singh, Rakhi Garg, P K Mishra
11. *Setting Up a Big Data Project: Challenges, Opportunities, Technologies and Optimization*: Roberto V. Zicari, Marten Rosselli, Todor Ivanov, Nikolaos Korfiatis, Karsten Tolle, Raik Niemann and Christoph Reichenbach
12. *Managing a Big Data/Analytics Project: A Systematic Literature Review*: G.O. Cordeiro, F. Deschamps, E. Pinheiro de Lima