Local Business Reviews and Data Analysis

[1]Hemamalini Madhanguru, [2]Mahsa Tayer Farahani, [3]Ruchi Singh, [4]Yashaswi Ananth
Grad Students, Department of Information Systems
California State University, Los Angeles
[1]hmadhan@calstatela.edu, [2]mtayerf@calstatela.edu, [3]rsingh26@calstatela.edu, [4]yananth@calstatela.edu

Abstract

In this project, we have made an attempt to analyze the local business data and reviews to get insights on the popularity of a business and factors responsible for it. To get better insights from the reviews, we also analyzed the sentiments of the customer review. The total size of the dataset is 180 MB. We have used IBM BigInsight and Azure HDInsight platforms for our research project and HiveQL and Pig for query.

Through this project we shall determine the following:

1. What actions could be taken by existing business to do better like their competitors.
2. Which are the most prospering businesses in a particular region.
3. The importance of customer satisfaction.
4. Feeling of the customers while writing the reviews for the services they used.

1. Introduction

All local business in today's time, engage with the consumers online for marketing, feedback, repeat business etc. One of the most popular and widely used business and services review platforms are Yelp and Google Local. Data collected from these platforms is the lifeline of the registered businesses. Analyzing this data is the key to critical marketing decisions, fine-tune product/services strategy or improve customer relationship. The dataset spans a variety of businesses such as restaurants, shopping, nightlife, medical, education entertainment, common services, etc. The user reviews collected from Google Local contains valuable information on consumer and business trends. This paper derives a lot of inferences about the local business and lists the factors responsible for making a business popular on the basis of the reviews and ratings. In depth analysis of the business reviews helps us understand the sentiments of the customers.

2. Technical Specifications

We have worked on Hadoop ecosystem to perform the Big data analytics and have built the project on IBM BigInsight as well as Azure HDInsight. Server details are as follows

Azure Specifications:

- Worker Nodes: 4
- Head Nodes: 4
- Number of Cores: 24
- RAM: 14 GB
- Disk Size: 200 GB
- Operating System: Linux

IBM Bluemix Specifications:

- Management Nodes: 1
  - vCPU: 12
  - RAM (GB): 48
- Data Nodes: 1
  - vCPU: 4
  - RAM (GB): 24
  - Data Disk: 1 TB SATA
- CPU Speed: 2.4 GHz Intel Xeon® ES-2673

HiveQL and Pig are the querying tools built on top of Hadoop that is used to query data within HDFS. Hive and Pig inherit all of Hadoop's fault tolerance features and is scalable for Big Data. The Hive language resembles SQL, which makes it useful for creating reports by Data Analysts whereas Pig is a Procedural Data Flow language used for programming by researchers and programmers.

Visualizations are generated in Tableau and Excel power view that create multi-faceted views of data and help communicate complex ideas simply.

3. Data Specifications

The Local Business dataset collected from Yelp and Google APIs are dated between 2005 and 2016, is rich in information about the local businesses in cities from 14 states in U.S and 4 cities that include: U.K: Edinburgh, Germany: Karlsruhe, Canada: Montreal and Waterloo. The yelp data set is of size 90MB in CSV file format with 334,335 rows and 108 columns. The Google Local reviews dataset is of size 85MB in JASON file format with 117,486 and 10 columns. To be able to use this data in Hadoop, the JSON file was converted to a CSV file using SerDe (Serializer/Deserializer) in Hive. Alternatively, JSON to CSV file format conversion can also be done in PIG using the built-in function JSONLoader().

The Business Data contained a lot of junk data which made it difficult to import into Hive tables for analysis. To clean the data, we removed the duplicate rows, eliminated complete NULL valued columns. We also had to format the date column to yyyy-mm-dd time columns to timestamps.

4. Implications of Data Analysis

This data set has a huge potential for research. Thus it has been studied and analyzed earlier by researchers, students as well as Yelp and Google themselves. Students and researchers have worked on a subset of this data set to analyze business in particular cities, areas or zones. Yelp and Google use large scale dataset(super set this dataset) to find results in real time and do to reflective analysis which is beyond human decision making.

For data analysis it is important to clearly understand for whom and for what purpose is the analysis conducted. This research is performed on the complete dataset for all location and years to analyze business from the users point of view. This research provides insights for new upcoming business as well as the betterment of the existing business. The research paper is a good source of information about local business and can be easily understood by readers from all background. It also opens new doors for researchers to have a holistic approach in studying the business data in US. This paper further supports that Big Data analysis is a useful tool for investigating a large chunks of business data.

5. Analysis of the Data Set

The clean data was imported into HDFS, exploratory data analysis was performed on the dataset using LOAD and DUMP in Pig to gather some initial facts about our data. For example, the Local Business data includes 2,903,217 entries having attributes such as business id, address, name, longitude, latitude, etc. The Reviews data includes 1,174,850 hive entries having attributes that included business id, user id, review id, stars, review date, and the user comments.
Separate tables are created for business and reviews in HDFS using Hive QL. Reviews table creation is as follows:

```
CREATE EXTERNAL TABLE IF NOT EXISTS reviews(funny int, useful int, cool int, userid string, reviewid string, stars int, reviewday date, type string, businessid string, comment string)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE LOCATION '/user/rsingh26/LocalBussinessReviews/' TBLPROPERTIES ('skip.header.line.count'='1');
```

A master table is created by joining the business and reviews table on common business id in each table. There are 452 categories of business, we grouped them broadly into 6 categories listed below. For each category separate tables are created to further analyse the business grouped under them.

a) Education
b) Entertaiment
c) Food
d) Medical
e) Services
f) Shopping

The data has been analyzed in the form of answers to questions related to business data.

1) *What is the review count for each category of business in order to understand which business category gets most number of reviews by the customers?*

The Hive QL to get the sum of the count for each category is as follows and its visualization is in Fig. 1.

```
select sum(review_count)from Education;
select sum(review_count)from Entertaiment;
select sum(review_count)from Food;
select sum(review_count)from Medical;
select sum(review_count)from Services;
select sum(review_count)from Shopping;
```

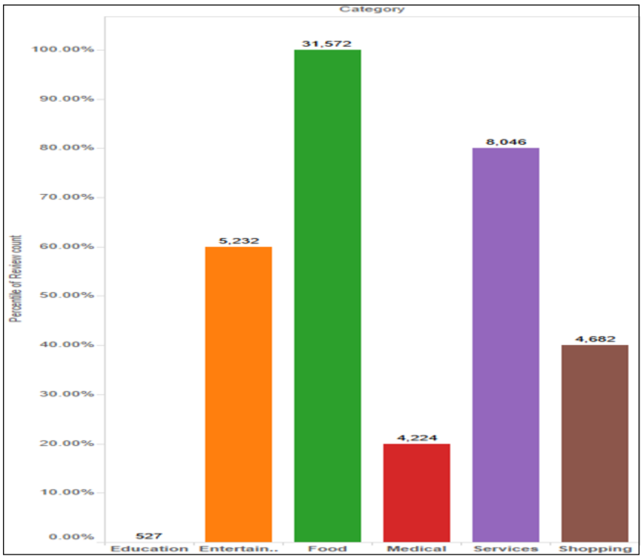


Fig. 1 Review count category-wise

Inference: Users tend to review the food businesses over the other categories like education, medical, shopping and services. Categories like Education and Medical are equally important categories as Food.

2) *Which sub category is most popular in the perticular state in US?*

The business data was collected from 14 states in US as shown in Fig.2 is visualized though powerview.

State	Category	Color
Alabama	Food	■
Arizona	Home Services	■
Arkansas	Home Services	■
California	Entertainment	■
Illinois	Food	■
Minnesota	Local Services	■
Nevada	Food	■
New Mexico	Steak Houses	■
North Carolina	Active Life	■
Oregon	Hotels and Travel	■
Pennsylvania	Food	■
South Carolina	Food	■
Texas	Banks and Credit Union	■
Wisconsin	Food	■

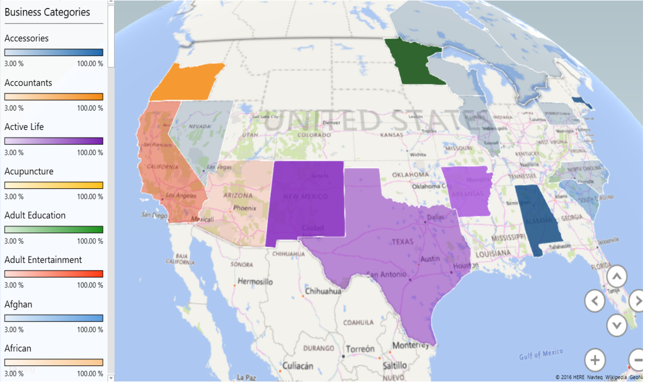


Fig. 2 Most popular sub categories grouped by state in US

3) *Which are the most popular cities for each business category?*

Tree map representation in Tableau is used to visualize the business categories grouped by city. The size of the rectangle represents the review count for a particular city and every business category is represented with a different color [refer fig.1]

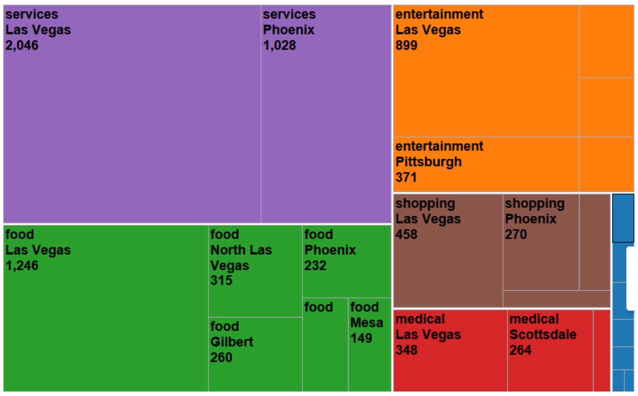


Fig. 3 Businesses categories grouped by city

Inference: In every category of business maximum number of review count come from Las Vegas. Possible reason could be the huge number of tourists that visit the city every year, use the services and leave feedbacks.

4) What is the count of reviews for the sub-categories of Shopping?

Bubble chart in Tableau is used to visualize the shopping sub-categories grouped by city in fig.4. Each sub-category is represented by a distict color and size of the bubble represents the review count.

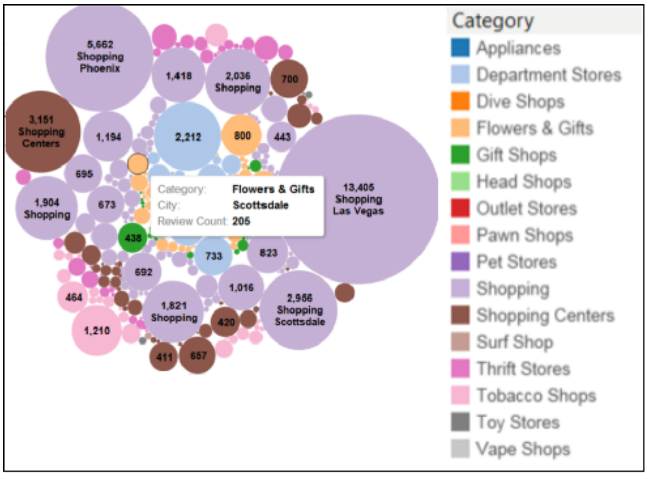


Fig. 4 Count of reviews for the sub-categories of the Shopping category

5) What is the sentiment analysis the customer reviews for all the Local Businesses in a city?

Customer's sentiment is analysed by the sum total of the positive, negetive and nutral words, used in the review, in accordance to the English dictionary. Cumulative sentiment is calculated for all the business reviews for a city and representend in a pie chart for each city. The size of the pie graph is the number of people who voted the review useful.

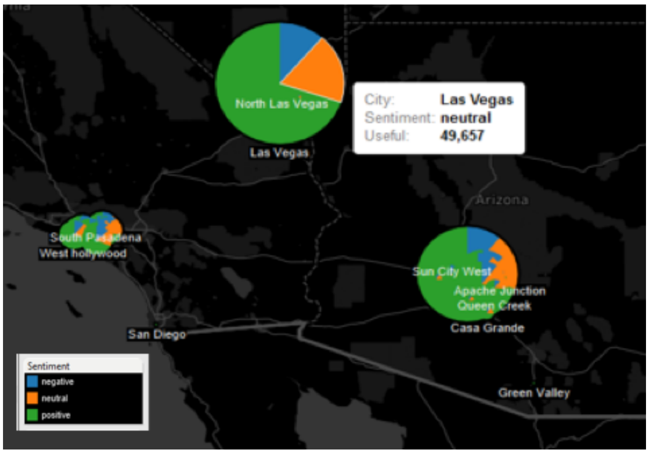


Fig. 5 Sentiment Analysis based on the customer reviews

Inference: More than 60% of the people in a city write positive reviews for the business and services they use. Las Vegas has the highest number of useful review votes.

6) What is the percentage of positive, neutral and negative reviews by the customers for the sub category of Local Businesses?

Sentiment analysis for the sub categories of Services is done in fig. 6. Similar analysis can be done for Education, Entertainment, Food, Medical and Shopping as well.

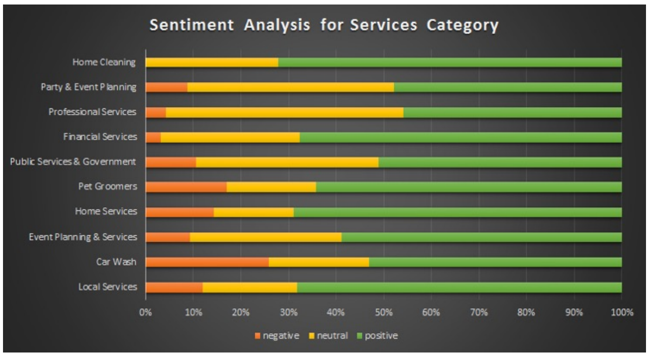


Fig. 6 Percentage of positive, neutral and negative customer reviews for the Services category

Inference: Services businesses mostly received positive reviews from the customers and the Home cleaning department did not receive any negative feedback.

7) What is the maximum number of business reviewed by individual users?

The following Hive QL was used to find the maximum number of businesses reviewed by a single reviewer and fig. 6 visualizes it in Excel charts.

```
select userid, count(businessid) from
review_analysis group by userid limit 10;
```

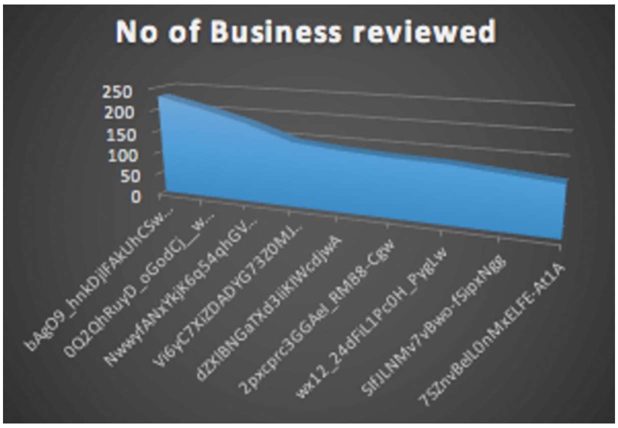
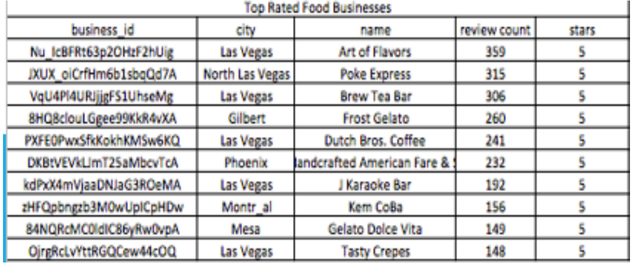


Fig. 6 Maximum count of reviews made by individual users

Inference: One particular user had written reviews for almost 250 different business over a span of 10 years. Further text analysis of the reviews could help us in understanding the authenticity of these reviews.

8) *What are the factors responsible for the popularity of the Local Businesses?*

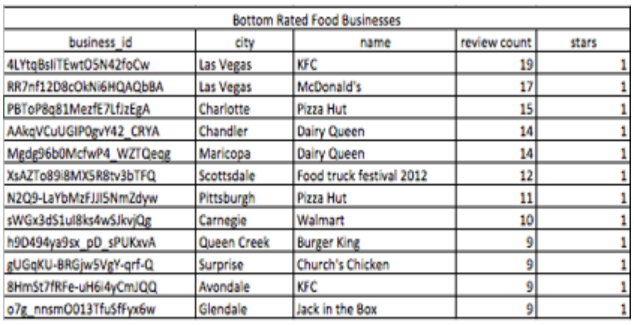
The top and bottom rated Food business are listed below in fig.7 and fig.8 respectively. Various attributes like happy hour, reservation, parking, vegan etc were compared and analysed. It led us to the conclusion that the attributes listed in fig.9 are hugely responsible for the good reviews and ratings of this business.Similar analysis can be done with other categories as well.



Top Rated Food Businesses

business_id	city	name	review count	stars
Nu_IcBFRt63p2OHzF2hUig	Las Vegas	Art of Flavors	359	5
JXUX_oiCrfHm6b1sbqQd7A	North Las Vegas	Poke Express	315	5
VqU4Pl4URjjjgFS1UhseMg	Las Vegas	Brew Tea Bar	306	5
8HQ8clouLGgee99KkR4vXA	Gilbert	Frost Gelato	260	5
PXFE0PwxSfkKokhKMSw6KQ	Las Vegas	Dutch Bros. Coffee	241	5
DKBtVEVkLJmT25aMbcvTcA	Phoenix	landcrafted American Fare &	232	5
kdPxX4mVjaaDNJaG3ROeMA	Las Vegas	J Karaoke Bar	192	5
zHFQpbngzb3M0wUpICpHDw	Montr_al	Kem CoBa	156	5
84NQRcMC0ldIC86yRw0vpA	Mesa	Gelato Dolce Vita	149	5
OjrgRcLvYttRGQCew44cOQ	Las Vegas	Tasty Crepes	148	5

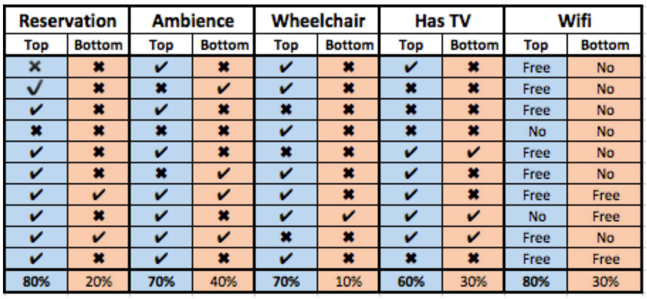
Fig. 7 Top rated food businesses



Bottom Rated Food Businesses

business_id	city	name	review count	stars
4LYtqBsIITEwtO5N42foCw	Las Vegas	KFC	19	1
RR7nf12D8cOkNi6HQAQbBA	Las Vegas	McDonald's	17	1
PBToP8q81MezfE7LfJzEgA	Charlotte	Pizza Hut	15	1
AAkqVCuUGIP0gvY42_CRYA	Chandler	Dairy Queen	14	1
Mgdg96b0McfwP4_WZTQeog	Maricopa	Dairy Queen	14	1
XsAZTo89i8MX5R8tv3bTFQ	Scottsdale	Food truck festival 2012	12	1
N2Q9-LaYbMzFJJI5NmZdyw	Pittsburgh	Pizza Hut	11	1
sWGx3dS1ul8ks4wSJkvjQg	Carnegie	Walmart	10	1
h9D494ya9sx_pD_sPUKxvA	Queen Creek	Burger King	9	1
gUGqKU-BRGjw5VgY-qrf-Q	Surprise	Church's Chicken	9	1
8HmSt7fRFe-uH6i4yCmJQQ	Avondale	KFC	9	1
o7g_nnsmO013TfuSfFyx6w	Glendale	Jack in the Box	9	1

Fig. 8 Bottom rated food businesses



Reservation		Ambience		Wheelchair		Has TV		Wifi	
Top	Bottom	Top	Bottom	Top	Bottom	Top	Bottom	Top	Bottom
×	✖	✓	✖	✓	✖	✓	✖	Free	No
✓	✖	✖	✓	✓	✖	✖	✖	Free	No
✓	✖	✓	✖	✖	✖	✖	✖	Free	No
✖	✖	✖	✖	✓	✖	✖	✖	No	No
✓	✖	✓	✖	✖	✖	✓	✓	Free	No
✓	✖	✖	✓	✓	✖	✓	✖	Free	No
✓	✓	✓	✓	✓	✖	✓	✖	Free	Free
✓	✖	✓	✖	✓	✓	✓	✓	No	Free
✓	✓	✓	✓	✖	✖	✓	✓	Free	No
✓	✖	✓	✖	✓	✖	✖	✖	Free	Free
80%	20%	70%	40%	70%	10%	60%	30%	80%	30%

Fig. 9 Factors influencing the popularity of the food businesses

Inference: More than 70% of the top 10 food businesses have reservation, ambience, wheelchair facility, TV and free wifi whereas less than 30% of bottom 10 do not have the same.

6. Conclusion

Using Hadoop, Hive, PIG, and Tableau greatly enhanced the exploratory possibilities and analytics capability with Big Data. The findings from the reserch are, food is the most popular category of Local Business based on the review count. Las Vegas is the most popular city for local business in every category. Reservation, ambience, wifi are the main factors responsible for the popularity of food businesses. More than 60% of people in a city write positive reviews for local business. And on record the maximun number of businesses reviewed by a reviewer is 250 in the span of 10 years.

7. Future Research

A detailed text analysis can be done to find the authenticity of the reviews. A detailed study of the popularity of a business and its attributes and help in predicting the business growth.

Acknowledgement

The analysis derived from this paper uses the Open Source components. The source code of this Open Source project along with the related information can be found on GitHub. We acknowledge and are grateful to Prof. Jongwook Woo for his guidance and encouragement.

References

1. Data source links: https://s3.amazonaws.com/hipicdatasets/yelp_raw_fall_2016.csv https://docs.google.com/uc?id=0B9kspRX6SWaaMlRvREQ3NmUxOE0&export=download
2. Tableau http://www.tableau.com
3. Hadoop Tutorial https://hortonworks.com/tutorials
4. Apache Hadoop Project, http://hadoop.apache.org/
5. Apache Hive, http://hive.apache.org/