

Problem / Question

To provide insightful analytics from large datasets obtained from the internet about the existing Local Business.

Hypothesis

- All the data in the dataset are authentic
- All the local business in a city are either registered in Yelp or Google Local
- Insights drawn from data collected from 2005 and 2016 is generalizes for the business

Project Overview

- We have analyzed a huge data set from Yelp and Google Local that spanned a variety of businesses such as restaurants, shopping, nightlife, medical, education, entertainment, common services, etc. in various cities across the world using Big Data.
- We have tried to understand various aspects of the Local Businesses; factors driving their popularity, customer review patterns and regions that favor certain businesses the most.
- Analyzing user reviews has helped us in deriving insights to achieve our objective through this project.

Variables / Research

- Controlled variables
 - These are kept the same throughout your experiments

Independent variable

- The **one** variable you purposely change and test

Dependent variable


- The measure of change observed because of independent variable
- Decide how you will measure the change

Data Specifications

IBM Bluemix Components	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
Dataset Components	Local Business	
	File Type	CSV
	File Size	90 MB
	Rows	334,35
	Columns	108
	Customer Reviews	
	File Type	JSON
	File Size	85 MB
	Rows	117,486
	Columns	10

Procedure

Acquire Data




Identification and authenticated access to Yelp and Google Local data. Transportation of 200 MB data from sources to distributed files systems.

Data Engineering



- Exploring and understanding data.
- Removing junk data, duplicate rows, eliminating NULL values and formatting data to yyyy-mm-dd.

Analyze Data



HiveQL and Pig are the querying tools built on top of Hadoop that is used to query data within HDFS. Hive makes it useful for creating reports whereas Pig is a Procedural Data Flow language used for programming by researchers and programmers.

Interpret Data



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

Visualizations

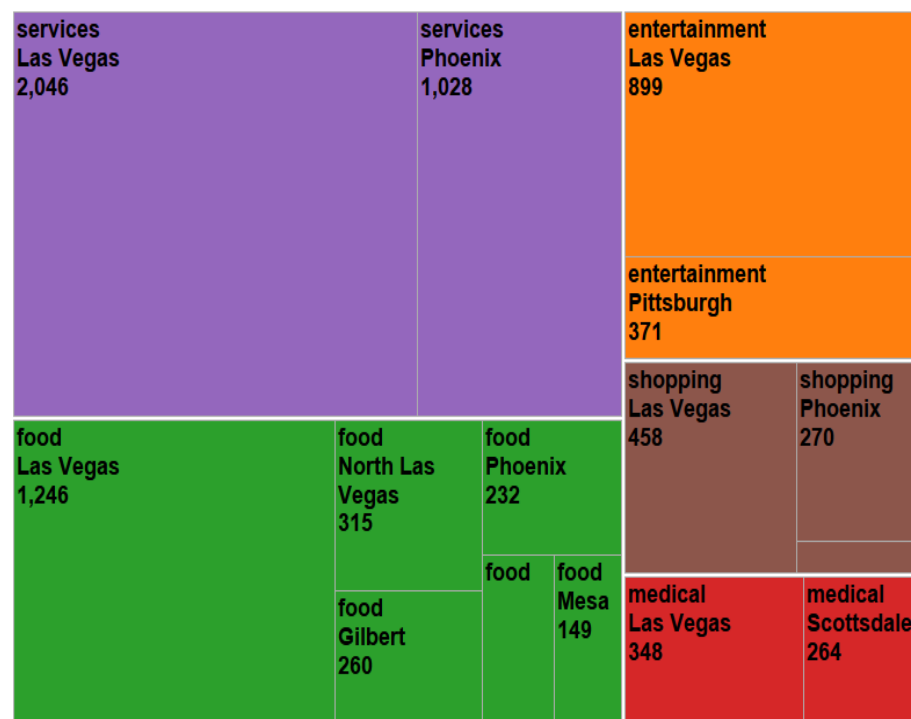


Figure. 1 Businesses categories grouped by city

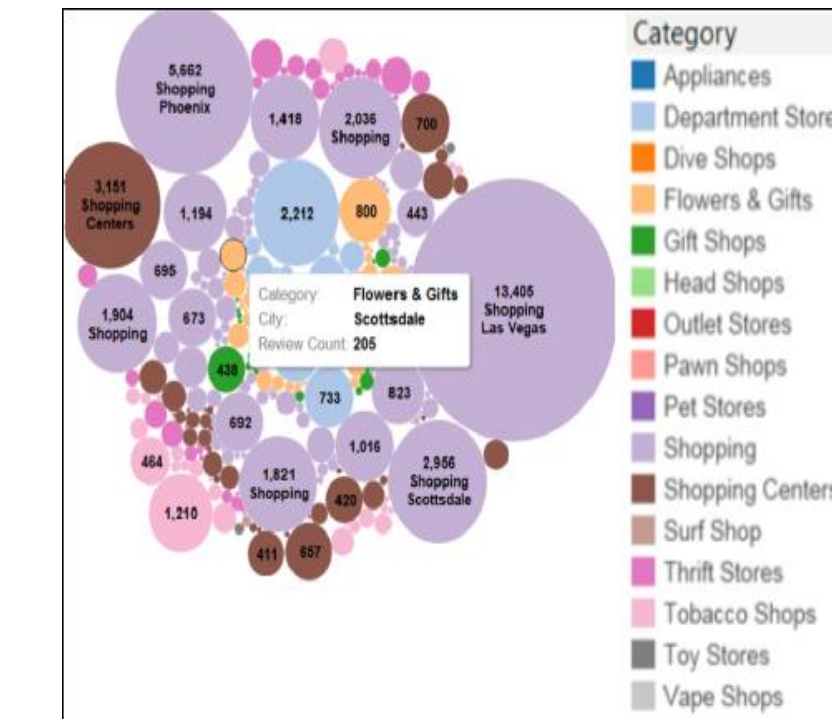


Figure. 2 Count of reviews for the sub-categories of the Shopping category



Figure. 3 Maximum count of reviews made by individual users

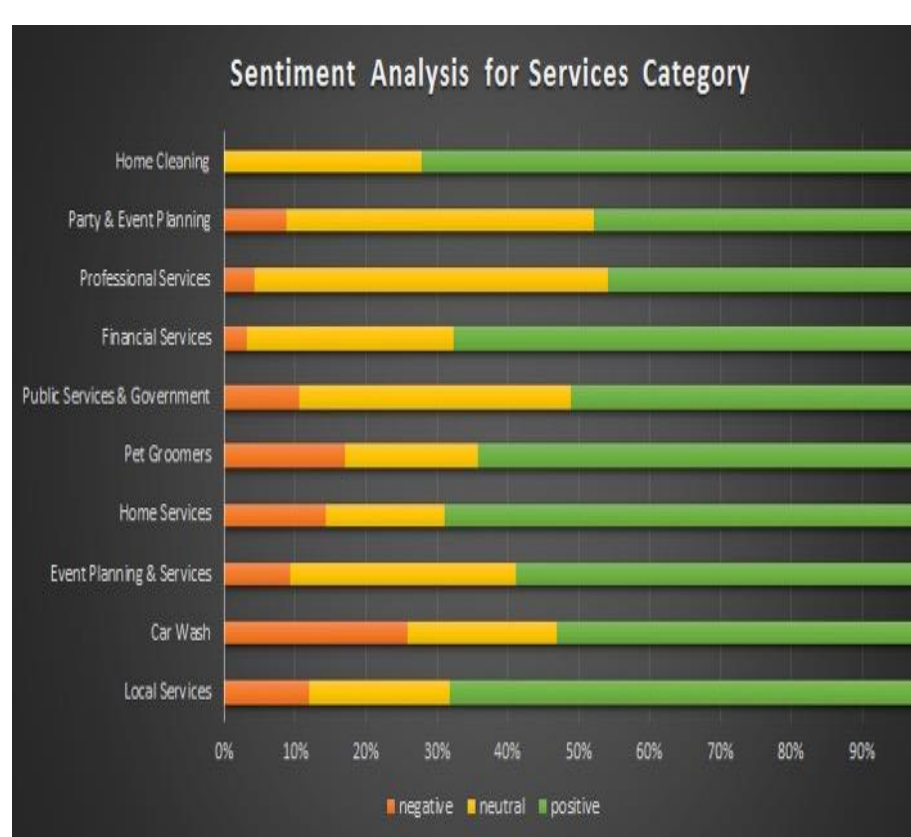


Figure. 4 Percentage of positive, neutral and negative customer review for the Services category

Top-Rated Food Businesses				
City	Name	Reviews	Stars	
Las Vegas	Art of Fish	359	5	
Las Vegas	Bake House	315	5	
Las Vegas	Pretzels Tea Bar	306	5	
Las Vegas	Fresh Fish	260	5	
Las Vegas	Dutch Bros Coffee	243	5	
Las Vegas	Handcrafted Sandwiches & Fare	232	5	
Las Vegas	Konoko Bar	162	5	
Monroe	Kern Cuts	150	5	
Mass	Grainito Cuban Vibe	148	5	
Las Vegas	Treat Crepes	148	5	

Table 1. Top rated food businesses				
City	Name	Reviews	Stars	
Las Vegas	HFC	17	1	
Las Vegas	McDonald's	19	1	
Charlotte	Pizza Hut	15	1	
Charlotte	Dairy Queen	14	1	
Maricopa	McDonald's	14	1	
Scottsdale	Food Truck Festival 2012	12	1	
Las Vegas	Pizza Hut	11	1	
Las Vegas	McDonald's	10	1	
Queen Creek	Burger King	9	1	
Las Vegas	Church's Chicken	8	1	

Table. 2 Bottom rated food businesses

Reservation		Amenities		Wheelchair		TV		WiFi	
Top	Low	Top	Low	Top	Low	Top	Low	Top	Low
No	No	Yes	No	Yes	No	Yes	No	Yes	No
Yes	No	Yes	No	Yes	No	Yes	No	Yes	No
Yes	No	Yes	No	Yes	No	No	No	Yes	No
Yes	No	No	Yes	No	No	No	No	Yes	No
No	No	No	No	Yes	No	No	No	No	No
Yes	No	Yes	No	No	No	Yes	Yes	Yes	No
Yes	No	No	Yes	Yes	No	Yes	No	Yes	No
Yes	Yes	Yes	Yes	Yes	No	Yes	No	Yes	Yes
Yes	No	Yes	Yes	No	Yes	Yes	Yes	No	Yes
Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No
80%	20%	70%	40%	70%	10%	60%	30%	80%	30%

Table 3 Factors influencing the popularity of the food businesses

Conclusion

- Using Hadoop, Hive, Pig, and Tableau enhanced the exploratory possibilities and analytics capability to store and process Big Data in parallel
- Factors driving their popularity, customer review patterns, regions that favor certain businesses the most can be determined by analyzing data
- Analyzing the customer sentiments based on their reviews has helped us in realizing the importance of customer satisfaction

Works Cited

- Apache Hadoop Project, <http://hadoop.apache.org/>
- Apache Hive, <http://hive.apache.org/>
- Yelp Data of HiPIC,
https://s3.amazonaws.com/hipicdatasets/yelp_raw_fall_2016.csv
- Yelp Review Data Set,
<https://docs.google.com/uc?id=0B9kspRX6SWaaMIRvREQ3NmUxOE0&export=download>
- GitHub Link:
https://github.com/shamaahsaa/Local_Business_DataAnalysis