Data 228

Analyze and Visualize US Business Restaurant Insights using Yelp

Project Report

Abraca Data (Group 7)

Yelp Insights

Table of Contents

Objective **2**

Data Source **3**

Yelp Dataset 3

US Household Income Dataset 4

Zip Code Dataset 4

Project Architecture **5**

Data Model **6**

Data Wrangling **7**

Other Data Wrangling Methods 9

Statistics of Original and Processed Dataset 11

Visualization **12**

Word Cloud of Cuisines 12

Top Ten Restaurants by Count 12

Average Rating of Top Ten Restaurants 13

Relation between Average Household Income of County and Number of Restaurant (Top 12) 14

Distribution of Restaurant Price Range based on Average Household Income 14

Number of Reviews Vs Restaurant Rating 15

Relation between Average Length of Review and Review Rating 16

Comparison between Ambience vs Estimated Average Income Range 16

Effect of Food Categories on Ratings 19

Review By Years 20

Restaurant Price ($) and stars 21

Restaurant Attributes Compare 22

Restaurant Attributes Compare of restaurants with 2 stars 23

Business review counts with more analysis on stars count 24

Website hosted on AWS S3 website hosting **27**

Conclusion **31**

Future Scope **32**

Reference Links **33**

Team Members **34**

Detailed Project Roles and Responsibilities Error! Bookmark not defined.

Abstract and Objective

Yelp Inc. is an online portal which provides crowd-sourced reviews for businesses and extends other services like restaurant table reservation. During the COVID-19 pandemic situation, ordering food online was the preferred source for people due to shelter in place orders issued by the Government and their safety. Yelp being the most popular review portal provided an edge to make informed decisions for the people during the Pandemic. This can be an opportunity for business owners to expand their sales by keeping their Yelp profile competitive with increase in dependency of people on reviews and ratings.

Our project focuses on analyzing the current trends of correlation between income and restaurants business, impact of business attributes, customer's reviews on business ratings and their price range

In this project we will be utilizing AWS-provided services and third-party software, Python, MySQL and Tableau to model, wrangle and analyze the data. The project findings will help investors to figure out the feasibility of a new venture and guide existing owners how to upscale their business margins

Data Source

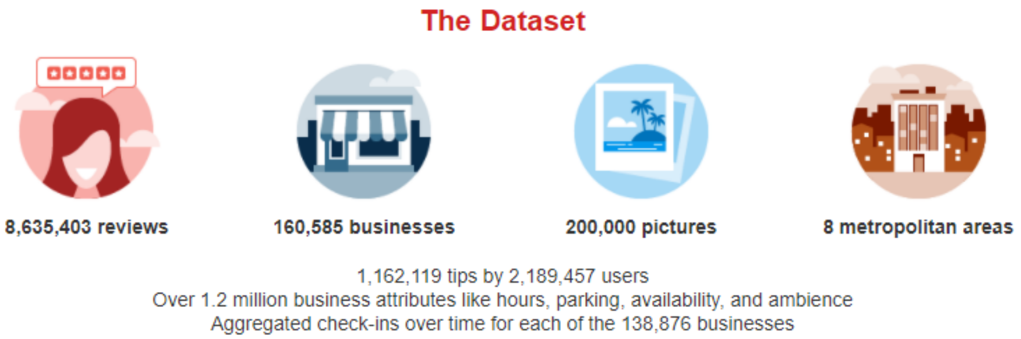
We have used three different datasets in the project to bring insights to Yelp Restaurant Business

Yelp Dataset

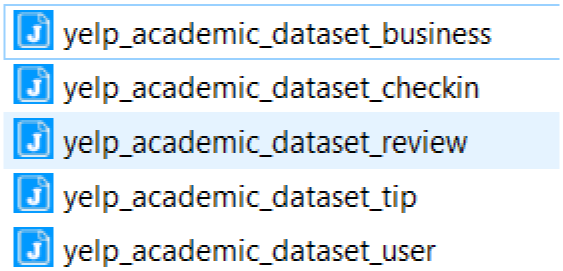
Yelp Data

The Yelp dataset is a subset of their businesses, reviews, and user data for use in personal, educational, and academic purposes. Available as JSON files, it can be used to teach students about databases, to learn NLP, or for sample production data while you learn how to make mobile apps.[1]

Figure Reference: https://www.yelp.com/dataset


The Dataset
8,635,403 reviews
160,585 businesses
200,000 pictures
8 metropolitan areas
1,162,119 tips by 2,189,457 users
Over 1.2 million business attributes like hours, parking, availability, and ambience
Aggregated check-ins over time for each of the 138,876 businesses

The scope of this project is to focus on Restaurants Category of Business and their corresponding Reviews mentioned in the dataset.
Original Dataset consist of 5 JSON file, of which, we have used Business and Review files



yelp_academic_dataset_business
yelp_academic_dataset_checkin
yelp_academic_dataset_review
yelp_academic_dataset_tip
yelp_academic_dataset_user

US Household Income Dataset

US Census Data

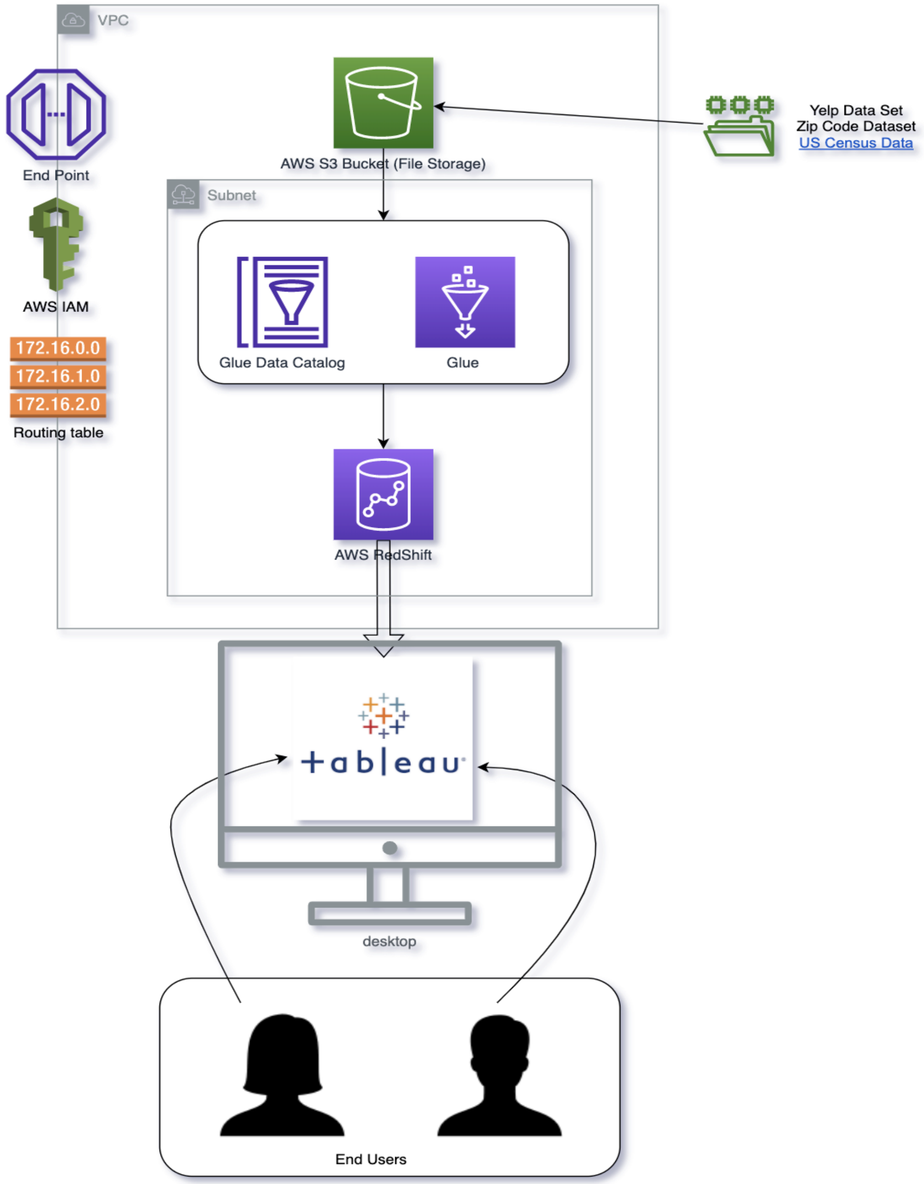
The data has been taken from United States Census Bureau site. The bureau puts together income in the past twelve months every year (inflation adjusted dollars) based on results of survey conducted by the American Community Survey (ACS).
ACS produces population, demographic and housing unit estimates, it is the Census Bureau's Population Estimates Program that produces and disseminates the official estimates of the population for the nation, states, counties, cities, and towns and estimates of housing units for states and counties.[2]
As part of the project, we have taken the latest data available at the site that is income in the past twelve months for year 2019.
The survey was based on an interview conducted with 2,059,945 housing Units across United States and their responses. [3]
The dataset consists information of 840 counties within United States and Puerto Rico.

Zip Code Dataset

unitedstateszipcodes.org
This dataset is used to bring together Yelp Business Dataset and US Household Income Dataset. It is required as Yelp Business Dataset contains information of Postal Codes, City and State while US Household Income Dataset contains information of County and State
The dataset contains information of Zip, County, State and Country.

Project Architecture

Project utilizes Amazon web services. Amazon S3(Simple Storage Services) to store raw data sets. ETL processes like data wrangling and cleansing will be performed using AWS Glue. Further cleansed data will be made accessible through AWS Redshift. Analytical processes will be performed on cleansed data for data visualization using Tableau. A website is created for end users to view the work



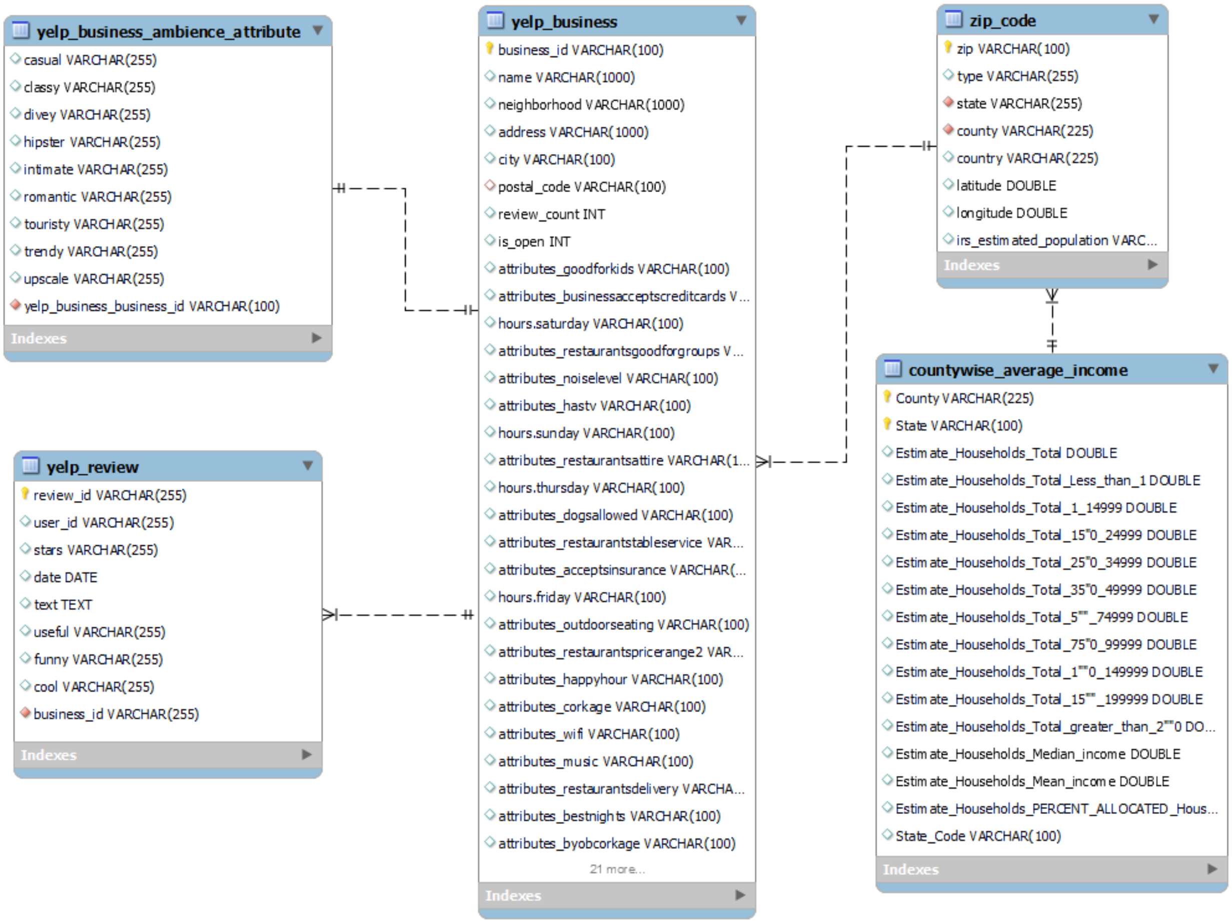
Data Model

Our AWS Redshift Database consist of five table. The relationship between them is defined as below:

Database Name: yelp

Schema Name: yelp

Schema Diagram [designed using MySQL EER Diagram]:



yelp_business_ambience_attribute
casual VARCHAR(255)
classy VARCHAR(255)
divey VARCHAR(255)
hipster VARCHAR(255)
intimate VARCHAR(255)
romantic VARCHAR(255)
touristy VARCHAR(255)
trendy VARCHAR(255)
upscale VARCHAR(255)
yelp_business_business_id VARCHAR(100)
Indexes

yelp_business
business_id VARCHAR(100)
name VARCHAR(1000)
neighborhood VARCHAR(1000)
address VARCHAR(1000)
city VARCHAR(100)
postal_code VARCHAR(100)
review_count INT
is_open INT
attributes_goodforkids VARCHAR(100)
attributes_businessacceptscreditcards V...
hours.saturday VARCHAR(100)
attributes_restaurantsgoodforgroups V...
attributes_noiselevel VARCHAR(100)
attributes_hastv VARCHAR(100)
hours.sunday VARCHAR(100)
attributes_restaurantsattire VARCHAR(1...
hours.thursday VARCHAR(100)
attributes_dogsallowed VARCHAR(100)
attributes_restaurantstableservice VAR...
attributes_acceptsinsurance VARCHAR(...
hours.friday VARCHAR(100)
attributes_outdoorseating VARCHAR(100)
attributes_restaurantspricerange2 VAR...
attributes_happyhour VARCHAR(100)
attributes_corkage VARCHAR(100)
attributes_wifi VARCHAR(100)
attributes_music VARCHAR(100)
attributes_restaurantsdelivery VARCHA...
attributes_bestnights VARCHAR(100)
attributes_byobcorkage VARCHAR(100)
21 more...
Indexes

zip_code
zip VARCHAR(100)
type VARCHAR(255)
state VARCHAR(255)
county VARCHAR(225)
country VARCHAR(225)
latitude DOUBLE
longitude DOUBLE
irs_estimated_population VARC...
Indexes

yelp_review
review_id VARCHAR(255)
user_id VARCHAR(255)
stars VARCHAR(255)
date DATE
text TEXT
useful VARCHAR(255)
funny VARCHAR(255)
cool VARCHAR(255)
business_id VARCHAR(255)
Indexes

countywise_average_income
County VARCHAR(225)
State VARCHAR(100)
Estimate_Households_Total DOUBLE
Estimate_Households_Total_Less_than_1 DOUBLE
Estimate_Households_Total_1_14999 DOUBLE
Estimate_Households_Total_15"0_24999 DOUBLE
Estimate_Households_Total_25"0_34999 DOUBLE
Estimate_Households_Total_35"0_49999 DOUBLE
Estimate_Households_Total_5""_74999 DOUBLE
Estimate_Households_Total_75"0_99999 DOUBLE
Estimate_Households_Total_1""0_149999 DOUBLE
Estimate_Households_Total_15""_199999 DOUBLE
Estimate_Households_Total_greater_than_2""0 DO...
Estimate_Households_Median_income DOUBLE
Estimate_Households_Mean_income DOUBLE
Estimate_Households_PERCENT_ALLOCATED_Hous...
State_Code VARCHAR(100)
Indexes

Data Wrangling

We have used ETL tool AWS GLUE, python notebook and AWS Redshift ‘COPY’ from S3 to cleanse and Excel to filter unwanted columns and create relationship between files as part of the data wrangling process.

There are two main datasets as part of the project (Yelp and US Census Household Income) and one supporting dataset (Zip Code Information).

Yelp file -Business (JSON file) Data was uploaded in S3 bucket and AWS Glue ETL was used to cleanse, filter required rows and process data into AWS Redshift Cluster

Yelp Review (JSON file) was uploaded in S3 bucket and processed into AWS Redshift using features of COPY command.

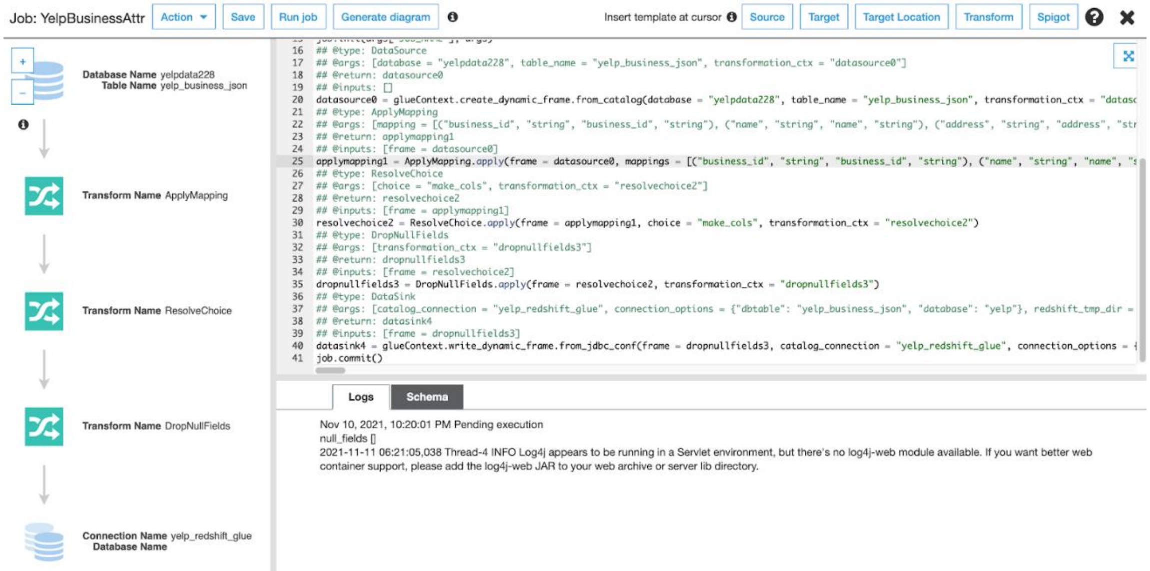
Yelp file -Business Ambience Attributes (nested string format) was fetched from Yelp business file and converted into JSON format using python 3 and after converting to csv using external tool uploaded in S3 in csv format. Further data was copied from S3 to AWS Redshift cluster

US Census Household Income (csv file) was uploaded in S3 bucket and Glue ETL was used to cleanse, filter required rows and process data into AWS Redshift Cluster.

Zip Code (csv file -originally) was converted to text pipe delimited file to avoid truncation of leading zeros in ZIP Code using python and then uploaded in S3 bucket. Further copied to AWS Redshift using features of COPY command.

AWS GLUE Jobs

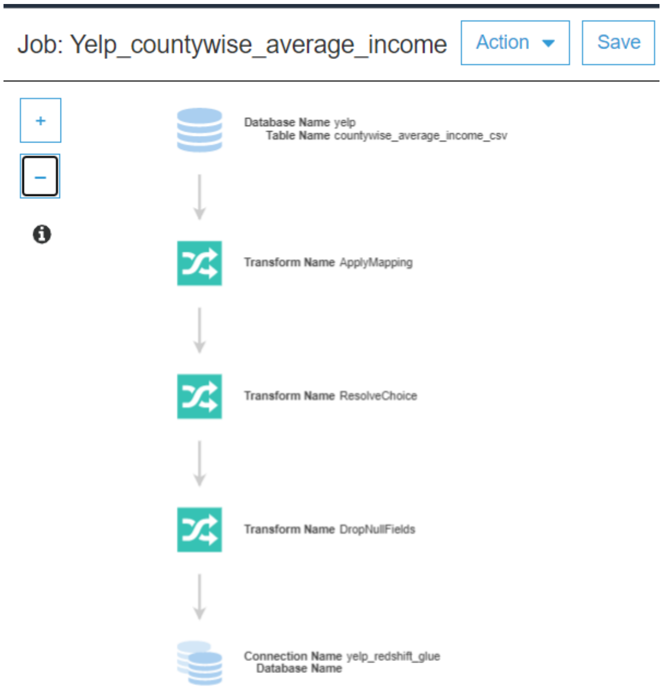
Yelp Business:


Job: YelpBusinessAttr
Action
Save
Run job
Generate diagram
Insert template at cursor
Source
Target
Target Location
Transform
Spigot
+
−
Database Name yelpdata228
Table Name yelp_business_json
Transform Name ApplyMapping
Transform Name ResolveChoice
Transform Name DropNullFields
Connection Name yelp_redshift_glue
Database Name
Logs
Schema

```
## @type: DataSource
## @args: [database = "yelpdata228", table_name = "yelp_business_json", transformation_ctx = "datasource0"]
## @return: datasource0
## @inputs: []
datasource0 = glueContext.create_dynamic_frame.from_catalog(database = "yelpdata228", table_name = "yelp_business_json", transformation_ctx = "datasc
## @type: ApplyMapping
## @args: [mapping = [("business_id", "string", "business_id", "string"), ("name", "string", "name", "string"), ("address", "string", "address", "str
## @return: applymapping1
## @inputs: [frame = datasource0]
applymapping1 = ApplyMapping.apply(frame = datasource0, mappings = [("business_id", "string", "business_id", "string"), ("name", "string", "name", "s
## @type: ResolveChoice
## @args: [choice = "make_cols", transformation_ctx = "resolvechoice2"]
## @return: resolvechoice2
## @inputs: [frame = applymapping1]
resolvechoice2 = ResolveChoice.apply(frame = applymapping1, choice = "make_cols", transformation_ctx = "resolvechoice2")
## @type: DropNullFields
## @args: [transformation_ctx = "dropnullfields3"]
## @return: dropnullfields3
## @inputs: [frame = resolvechoice2]
dropnullfields3 = DropNullFields.apply(frame = resolvechoice2, transformation_ctx = "dropnullfields3")
## @type: DataSink
## @args: [catalog_connection = "yelp_redshift_glue", connection_options = {"dbtable": "yelp_business_json", "database": "yelp"}, redshift_tmp_dir =
## @return: datasink4
## @inputs: [frame = dropnullfields3]
datasink4 = glueContext.write_dynamic_frame.from_jdbc_conf(frame = dropnullfields3, catalog_connection = "yelp_redshift_glue", connection_options = {
job.commit()
```

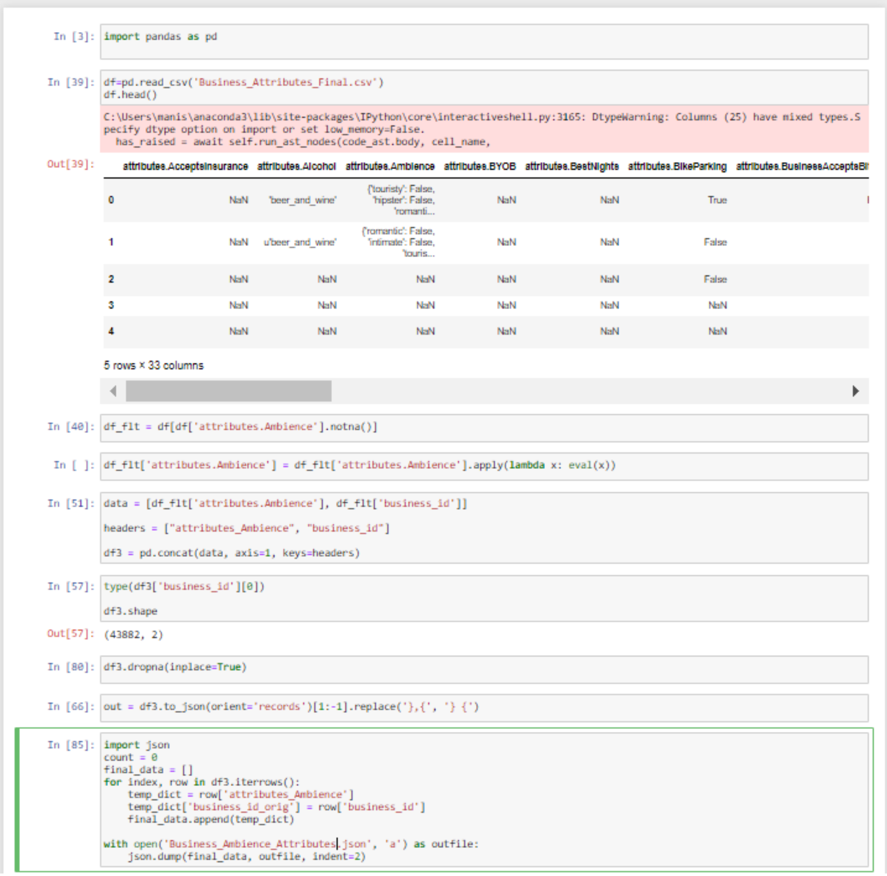
```
Nov 10, 2021, 10:20:01 PM Pending execution
null_fields []
2021-11-11 06:21:05,038 Thread-4 INFO Log4j appears to be running in a Servlet environment, but there's no log4j-web module available. If you want better web
container support, please add the log4j-web JAR to your web archive or server lib directory.
```

US Census Household Income (csv file)


Job: Yelp_countywise_average_income
Action
Save
+
−
Database Name yelp
Table Name countywise_average_income_csv
Transform Name ApplyMapping
Transform Name ResolveChoice
Transform Name DropNullFields
Connection Name yelp_redshift_glue
Database Name

Other Data Wrangling Methods

Yelp file -Business Ambience Attributes (nested string format)



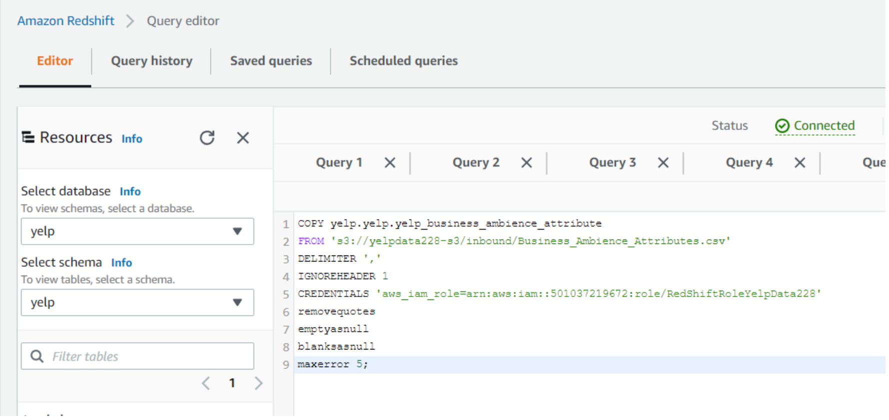
```
In [3]:  import pandas as pd

In [39]: df=pd.read_csv('Business_Attributes_Final.csv')
         df.head()

C:\Users\manis\anaconda3\lib\site-packages\IPython\core\interactiveshell.py:3165: DtypeWarning: Columns (25) have mixed types.S
pecify dtype option on import or set low_memory=False.
  has_raised = await self.run_ast_nodes(code_ast.body, cell_name,
```

Out[39]:

	attributes.AcceptsInsurance	attributes.Alcohol	attributes.Ambience	attributes.BYOB	attributes.BestNights	attributes.BikeParking	attributes.BusinessAcceptsBi
0	NaN	'beer_and_wine'	{'touristy': False, 'hipster': False, 'romanti...	NaN	NaN	True	
1	NaN	u'beer_and_wine'	{'romantic': False, 'intimate': False, 'touris...	NaN	NaN	False	
2	NaN	NaN	NaN	NaN	NaN	False	
3	NaN	NaN	NaN	NaN	NaN	NaN	
4	NaN	NaN	NaN	NaN	NaN	NaN	

5 rows × 33 columns

```
In [40]: df_flt = df[df['attributes.Ambience'].notna()]

In [ ]:  df_flt['attributes.Ambience'] = df_flt['attributes.Ambience'].apply(lambda x: eval(x))

In [51]: data = [df_flt['attributes.Ambience'], df_flt['business_id']]

         headers = ["attributes_Ambience", "business_id"]

         df3 = pd.concat(data, axis=1, keys=headers)

In [57]: type(df3['business_id'][0])

         df3.shape

Out[57]: (43882, 2)

In [80]: df3.dropna(inplace=True)

In [66]: out = df3.to_json(orient='records')[1:-1].replace('},{', '} {')

In [85]: import json
         count = 0
         final_data = []
         for index, row in df3.iterrows():
             temp_dict = row['attributes_Ambience']
             temp_dict['business_id_orig'] = row['business_id']
             final_data.append(temp_dict)

         with open('Business_Ambience_Attributes.json', 'a') as outfile:
             json.dump(final_data, outfile, indent=2)
```

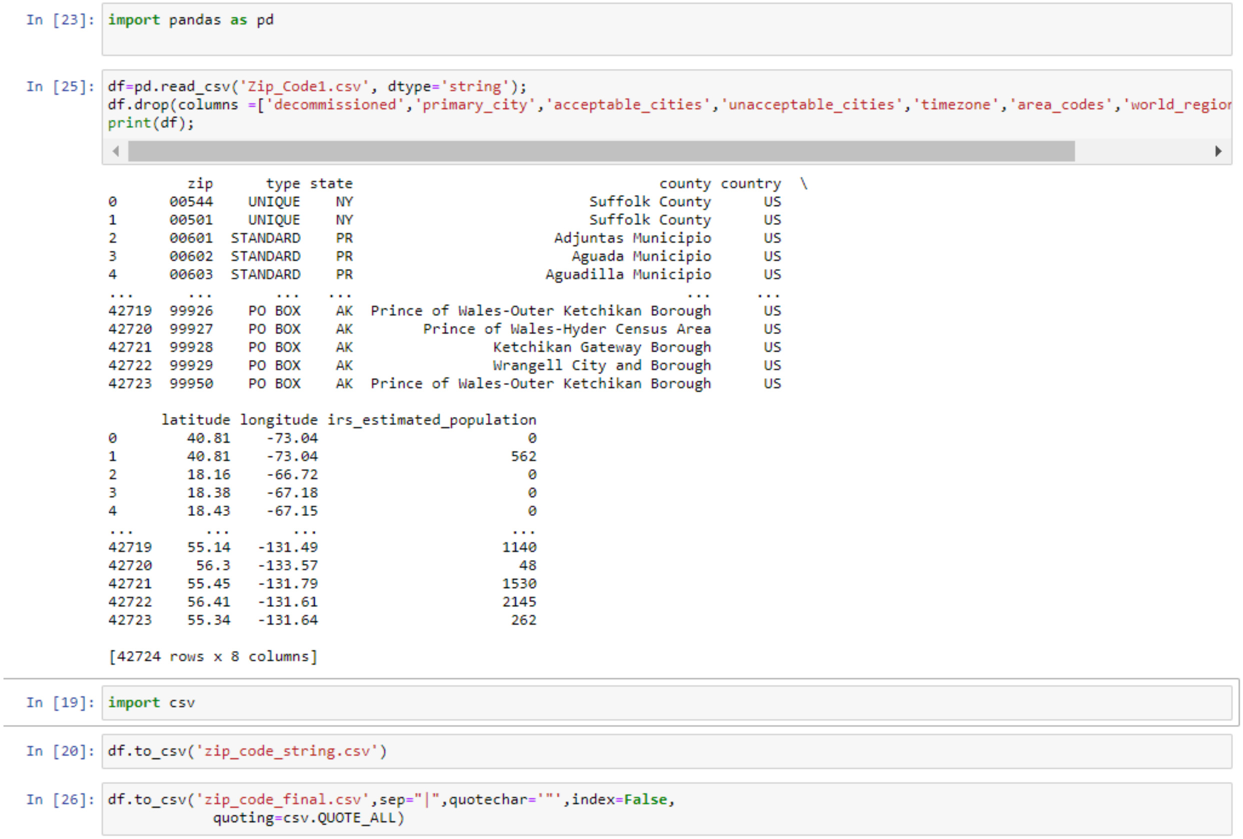


Select database Info
To view schemas, select a database.
yelp

Select schema Info
To view tables, select a schema.
yelp

```
COPY yelp.yelp.yelp_business_ambience_attribute
FROM 's3://yelpdata228-s3/inbound/Business_Ambience_Attributes.csv'
DELIMITER ','
IGNOREHEADER 1
CREDENTIALS 'aws_iam_role=arn:aws:iam::501037219672:role/RedShiftRoleYelpData228'
removequotes
emptyasnull
blanksasnull
maxerror 5;
```

Zip Codes file:



```
In [23]: import pandas as pd
```

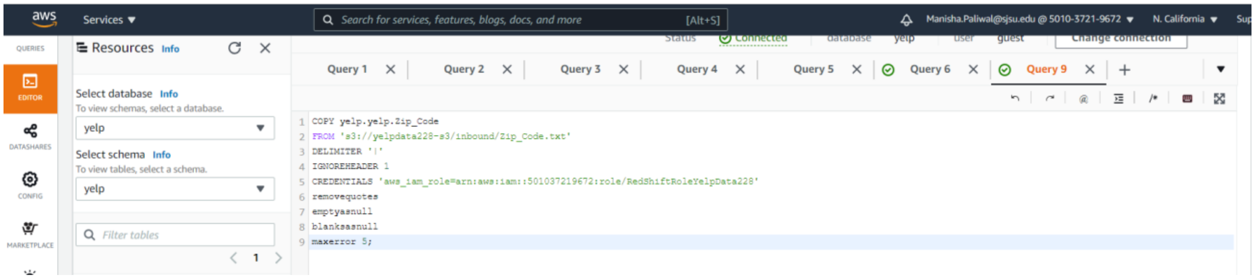
```
In [25]: df=pd.read_csv('Zip_Code1.csv', dtype='string');
         df.drop(columns =['decommissioned','primary_city','acceptable_cities','unacceptable_cities','timezone','area_codes','world_regior
         print(df);
```

```
         zip      type state                                    county country  \
0      00544    UNIQUE    NY                             Suffolk County      US
1      00501    UNIQUE    NY                             Suffolk County      US
2      00601  STANDARD    PR                         Adjuntas Municipio      US
3      00602  STANDARD    PR                           Aguada Municipio      US
4      00603  STANDARD    PR                        Aguadilla Municipio      US
...      ...       ...   ...                                        ...     ...
42719  99926    PO BOX    AK  Prince of Wales-Outer Ketchikan Borough      US
42720  99927    PO BOX    AK        Prince of Wales-Hyder Census Area      US
42721  99928    PO BOX    AK                Ketchikan Gateway Borough      US
42722  99929    PO BOX    AK                 Wrangell City and Borough      US
42723  99950    PO BOX    AK  Prince of Wales-Outer Ketchikan Borough      US

      latitude longitude irs_estimated_population
0        40.81    -73.04                        0
1        40.81    -73.04                      562
2        18.16    -66.72                        0
3        18.38    -67.18                        0
4        18.43    -67.15                        0
...        ...       ...                      ...
42719    55.14   -131.49                     1140
42720     56.3   -133.57                       48
42721    55.45   -131.79                     1530
42722    56.41   -131.61                     2145
42723    55.34   -131.64                      262

[42724 rows x 8 columns]
```

```
In [19]: import csv
```

```
In [20]: df.to_csv('zip_code_string.csv')
```

```
In [26]: df.to_csv('zip_code_final.csv',sep="|",quotechar='"',index=False,
                     quoting=csv.QUOTE_ALL)
```

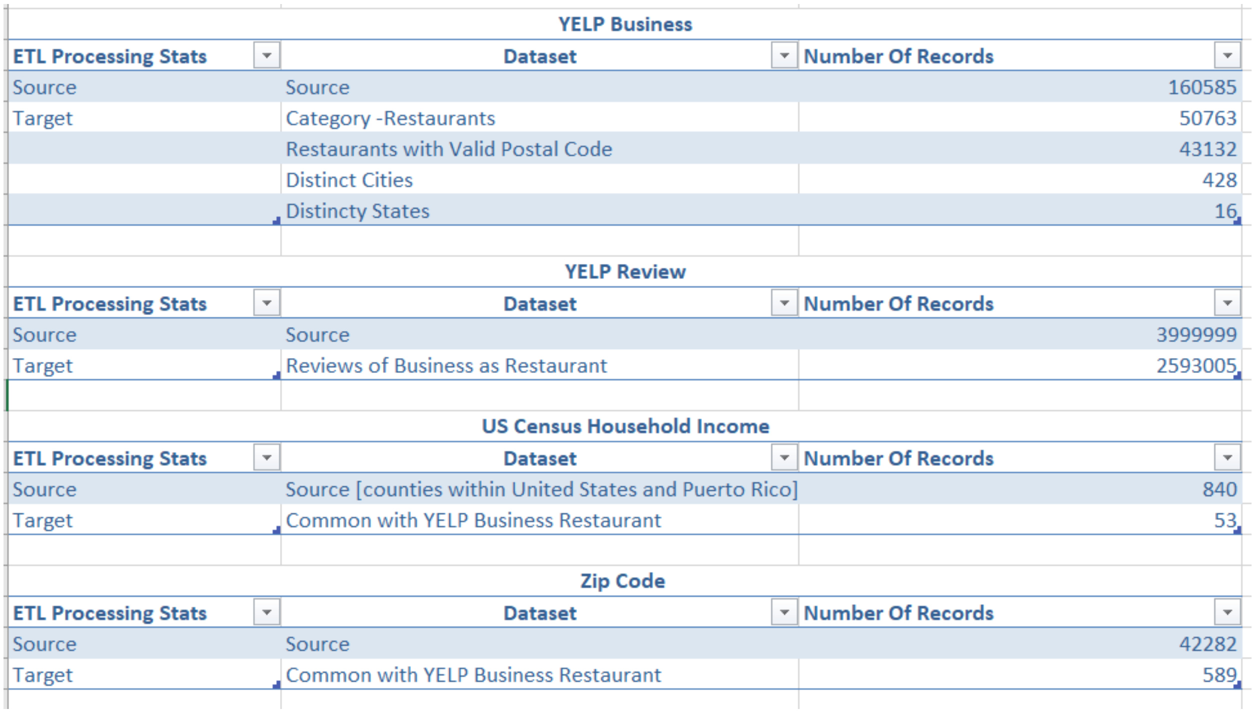


Select database Info
To view schemas, select a database.
yelp

Select schema Info
To view tables, select a schema.
yelp

```
COPY yelp.yelp.Zip_Code
FROM 's3://yelpdata228-s3/inbound/Zip_Code.txt'
DELIMITER '|'
IGNOREHEADER 1
CREDENTIALS 'aws_iam_role=arn:aws:iam::501037219672:role/RedShiftRoleYelpData228'
removequotes
emptyasnull
blanksasnull
maxerror 5;
```

Statistics of Original and Processed Dataset



YELP Business

ETL Processing Stats	Dataset	Number Of Records
Source	Source	160585
Target	Category -Restaurants	50763
	Restaurants with Valid Postal Code	43132
	Distinct Cities	428
	Distincty States	16

YELP Review

ETL Processing Stats	Dataset	Number Of Records
Source	Source	3999999
Target	Reviews of Business as Restaurant	2593005

US Census Household Income

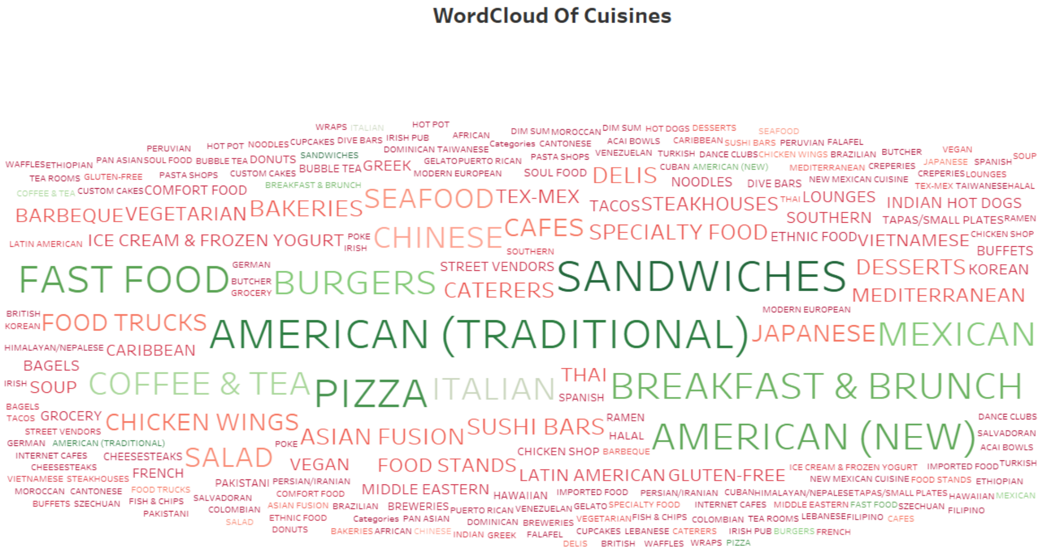
ETL Processing Stats	Dataset	Number Of Records
Source	Source [counties within United States and Puerto Rico]	840
Target	Common with YELP Business Restaurant	53

Zip Code

ETL Processing Stats	Dataset	Number Of Records
Source	Source	42282
Target	Common with YELP Business Restaurant	589

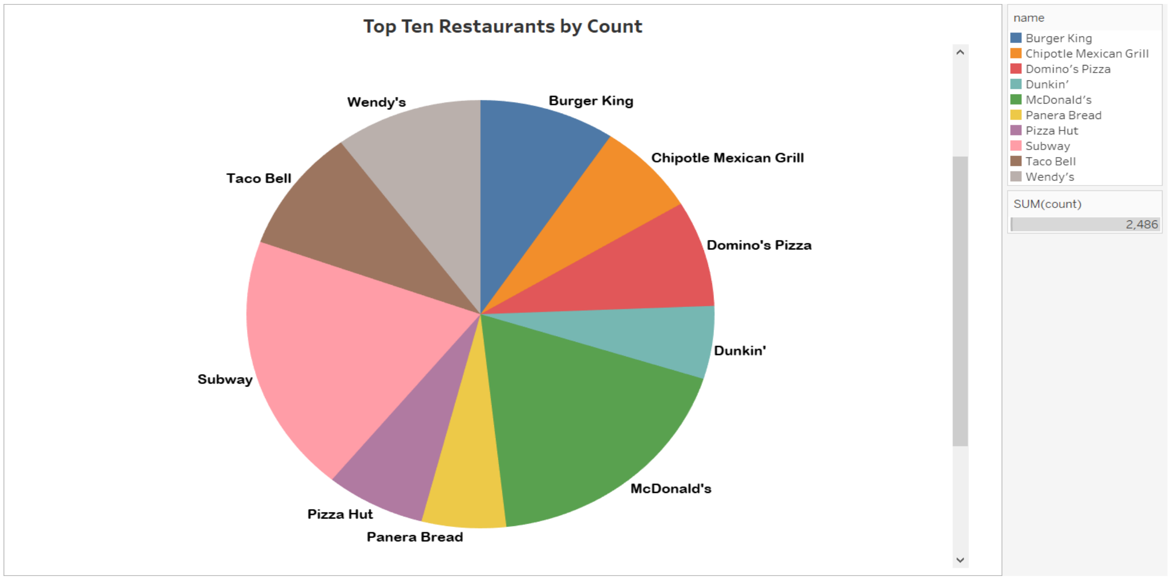
Visualization

Word Cloud of Cuisines



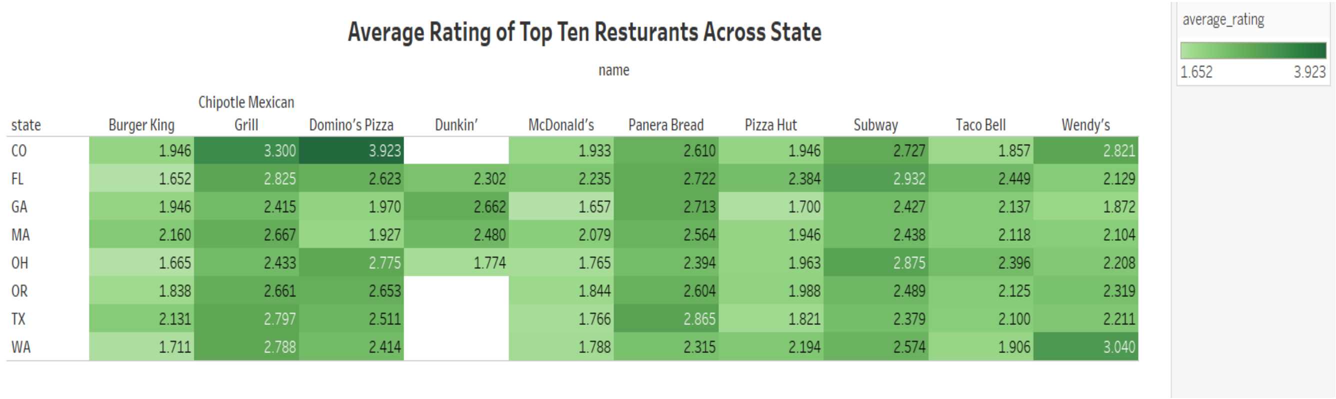
The word cloud of cuisines is created on Yelp Businesses with category Restaurant. It represents the cuisines based on number of restaurants of that type. We can see that American (Traditional), Sandwiches, Pizza, Breakfast & Brunch and Fast Food are the most popular cuisines therefore, there are more restaurants for these type in US

Top Ten Restaurants by Count



Above graph shows the top 10 restaurant chains in US by count based on data available in Yelp. Subway has highest number of restaurants followed by McDonald's.

Average Rating of Top Ten Restaurants



Average Rating of Top Ten Resturants Across State

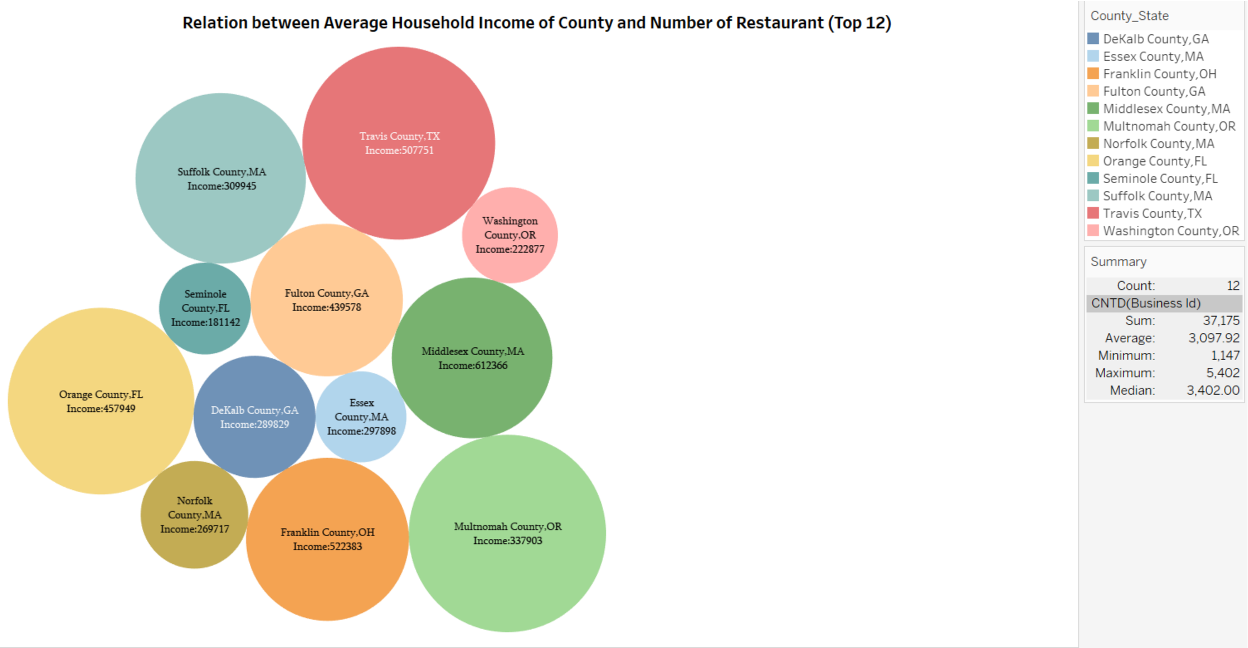
name

state	Burger King	Chipotle Mexican Grill	Domino's Pizza	Dunkin'	McDonald's	Panera Bread	Pizza Hut	Subway	Taco Bell	Wendy's
CO	1.946	3.300	3.923		1.933	2.610	1.946	2.727	1.857	2.821
FL	1.652	2.825	2.623	2.302	2.235	2.722	2.384	2.932	2.449	2.129
GA	1.946	2.415	1.970	2.662	1.657	2.713	1.700	2.427	2.137	1.872
MA	2.160	2.667	1.927	2.480	2.079	2.564	1.946	2.438	2.118	2.104
OH	1.665	2.433	2.775	1.774	1.765	2.394	1.963	2.875	2.396	2.208
OR	1.838	2.661	2.653		1.844	2.604	1.988	2.489	2.125	2.319
TX	2.131	2.797	2.511		1.766	2.865	1.821	2.379	2.100	2.211
WA	1.711	2.788	2.414		1.788	2.315	2.194	2.574	1.906	3.040

Above graph represents the average rating of each restaurant (top ten) in the states they are doing business (as per data available in Yelp) and their average ratings.
None of the most common food chains in the US are doing particularly good when it comes to overall business rating.

McDonald's average rating across all states is only 1.83 closely followed by Pizza Hut with 1.98.
Panera Bread is a consistent performer at an average rating of 2.59 across all states.
Chipotle Mexican Grill has the best average rating across states among the selected food chains at 2.71 closely followed by Domino's at 2.6.

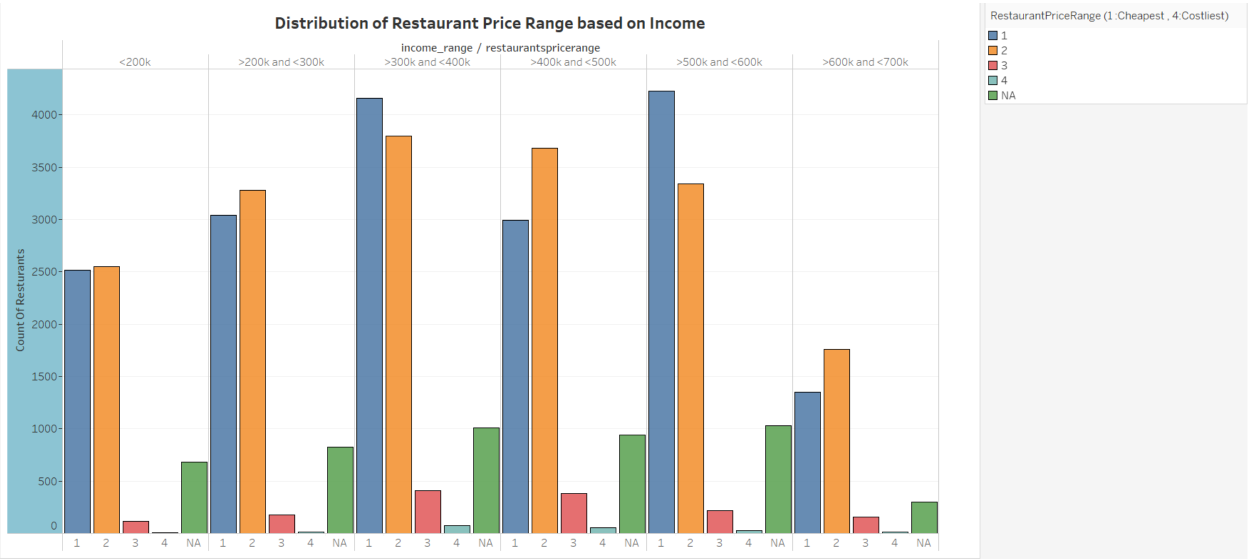
Relation between Average Household Income of County and Number of Restaurant (Top 12)



Here we explore the relationship between the average household income in counties to the total number of restaurants in those counties. We have only taken into consideration the counties with a minimum of 1000 restaurants to keep the data safe from anomalies and outliers.

Prima facie, we see a positive relationship between the average household total income and the number of restaurants, indicating that the number of restaurants does go up as the income goes up. While this is not enough evidence to suggest causality, it certainly shows a correlation between the two.

Distribution of Restaurant Price Range based on Average Household Income



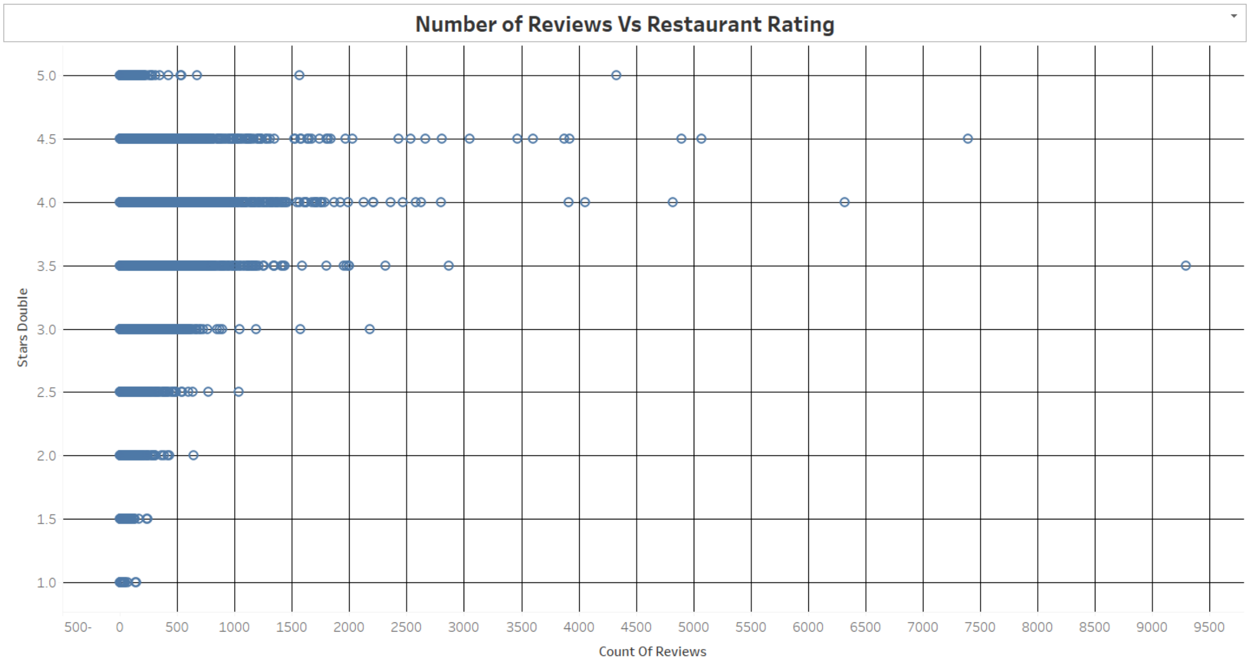
In the above graph, we split the income into range as shown above. Starting from < 200k to 200k – 300k, 300k – 400k and so on. We explore the distribution of restaurant price ranges across counties with different average household income ranges.

The data paints a different picture than one would expect. The neighborhoods with higher income ranges don't have more costlier restaurants. For example, if you look at the range from 500k to 600k, they have the greatest number of cheapest restaurant options.

The observation about the lowest income bracket seems more logical since it suggests that the counties in this income range have more restaurants with cheaper options.

The conclusion that we can draw from the above graph is that the counties with average income in the range of 300k to 500k have the greatest number of restaurants, across all the price ranges.

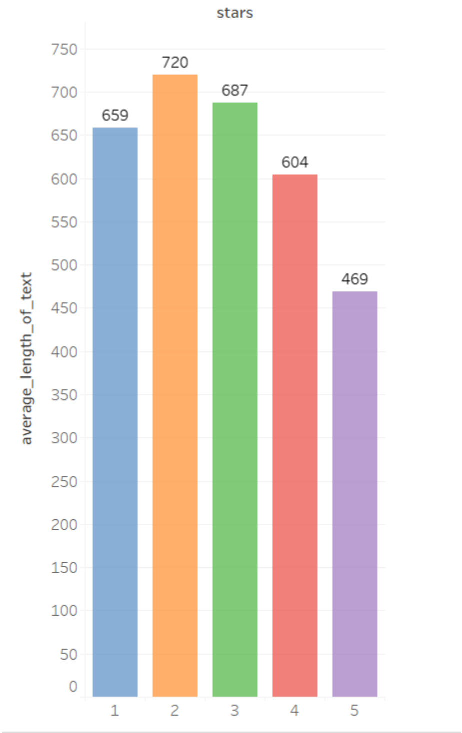
Number of Reviews Vs Restaurant Rating



We have seen restaurant owners asking their customers to review them online. Here we explore whether the number of reviews a restaurant receives changes anything for the restaurant's rating.

As can be seen clearly above, except the elusive 5-star rating, all other ratings have a positive correlation with the number of reviews a restaurant has. The more the number of reviews, the better the rating. 2000 to 2500 seems to be optimum number of ratings a business should have, but of course there are many outliers.

Relation between Average Length of Review and Review Rating



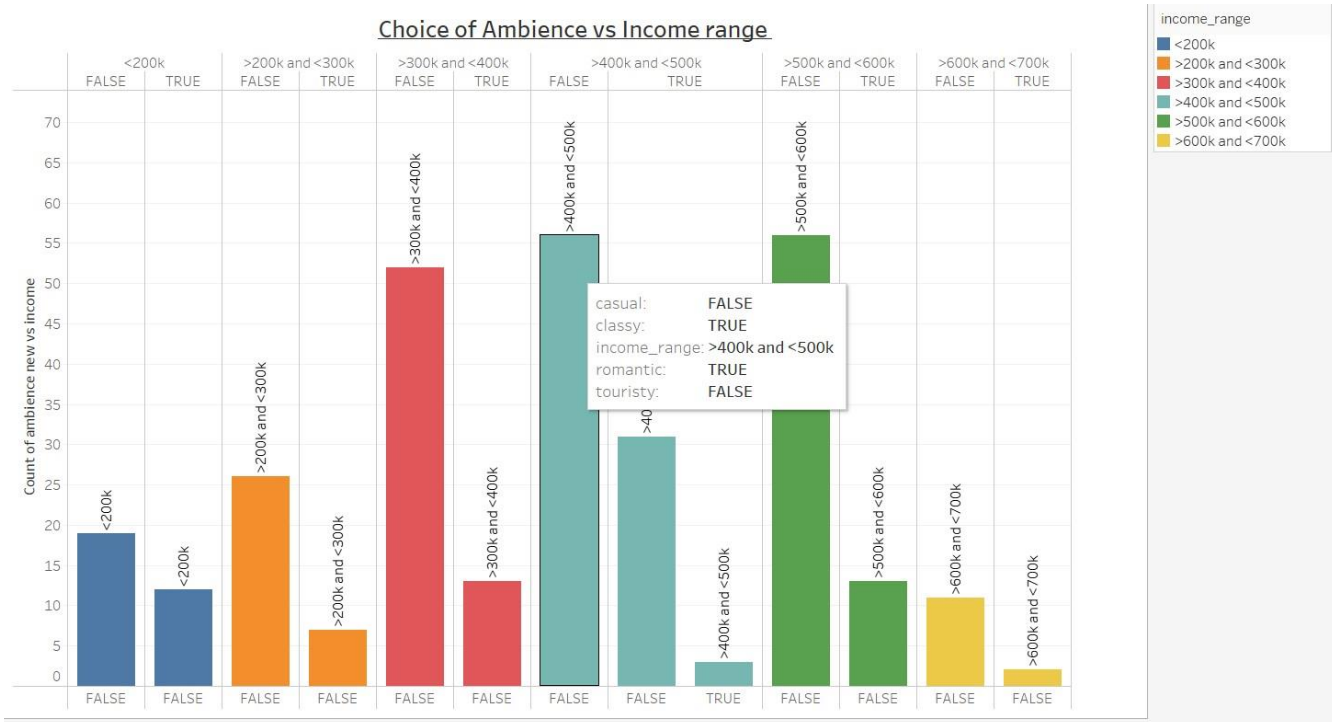
The above graph shows a relationship between the length of a review and the rating the user gave for the same review.

The observation from the above graph suggests that 1-star ratings had generally had around median number of words in the review, which can be interpreted as dissatisfied customers registering their concerns.

The most amount of constructive feedback seems to be happening with the 2- and 3-star rating, where even though the users are dissatisfied, they want to help the restaurant get better and the users providing these ratings are more verbose than others.

And lastly, it can be inferred from above that we humans have fewer words to say when we are praising something.

Comparison between Ambience vs Estimated Average Income Range



casual: FALSE
classy: TRUE
income_range: >400k and <500k
romantic: TRUE
touristy: FALSE

The above Graph provides insights on how the choice of Ambience is affected by the Average Income range.

The Visualization has an income range from 0$ to 700,000$ which is taken from the income Dataset which provides insights on the average annual income. The Ambience attribute from the Yelp dataset gives insights on how the different incomes earned by people affects their choice of ambience.

In this project we have considered 4 main Ambience which are:

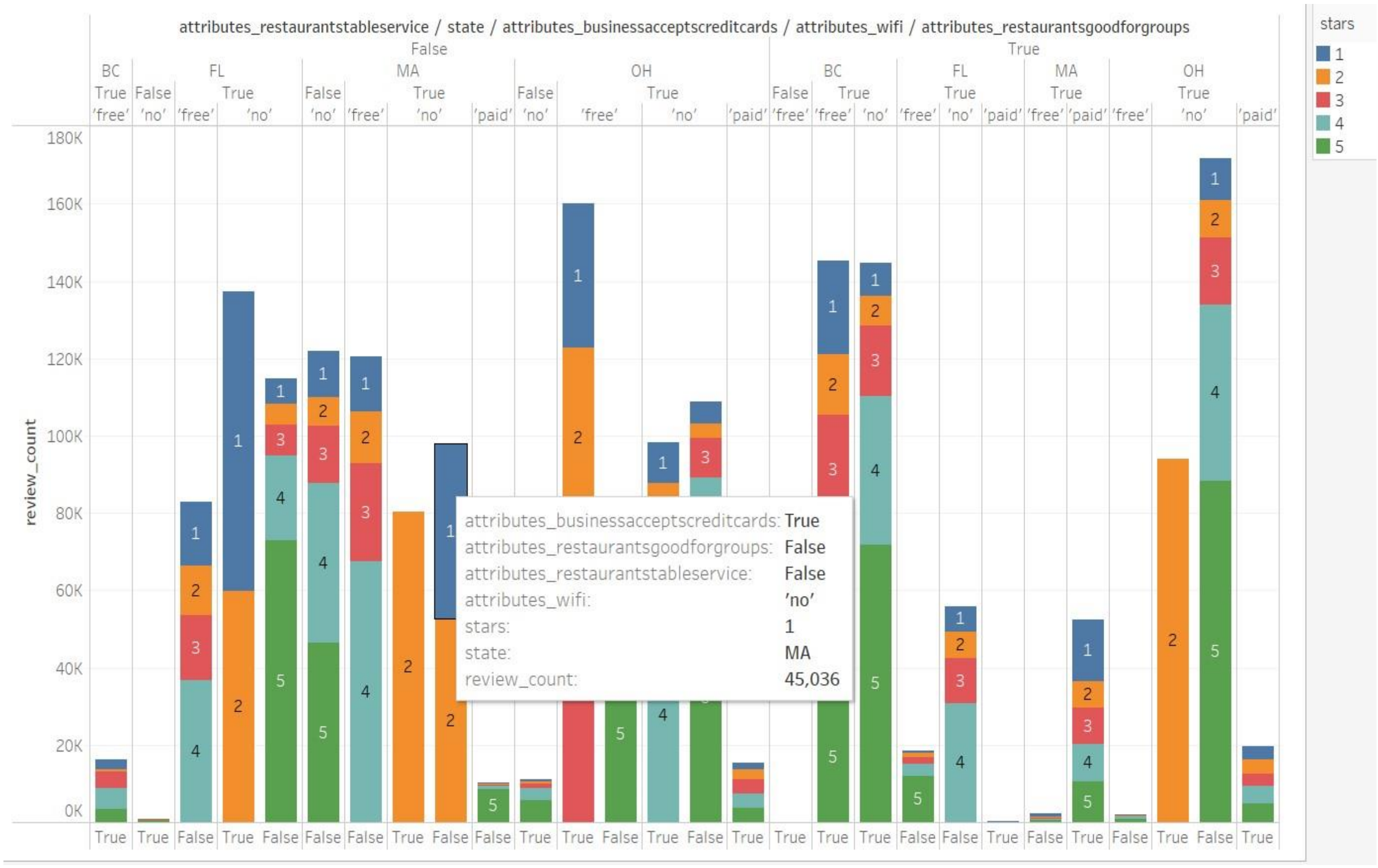
a) Casual
b) Classy
c) Romantic
d) Touristy

From the analysis, we can infer that the people who earn <200,000$ do not usually prefer a casual choice of ambience which is surprising as there is a huge percentage of people who prefer classy and romantic choice of ambience over a casual setting.

However, the people earning < 200,000$ neglect a touristy setting, this is probably because a touristy setting is more expensive when compared to the other attributes.

The people earning in the range between 400,000$ - 500,000$ prefer only a Classy or romantic setting, while the people earning a larger annual income of 500,000$ and above usually tend to prefer Classy and romantic choice of Ambience.

Effect of Business Attributes on Stars and Reviews



attributes_businessacceptscreditcards: True
attributes_restaurantsgoodforgroups: False
attributes_restaurantstableservice: False
attributes_wifi: 'no'
stars: 1
state: MA
review_count: 45,036

The above graph provides insights on how certain Business attributes affects the Stars and Review Counts.

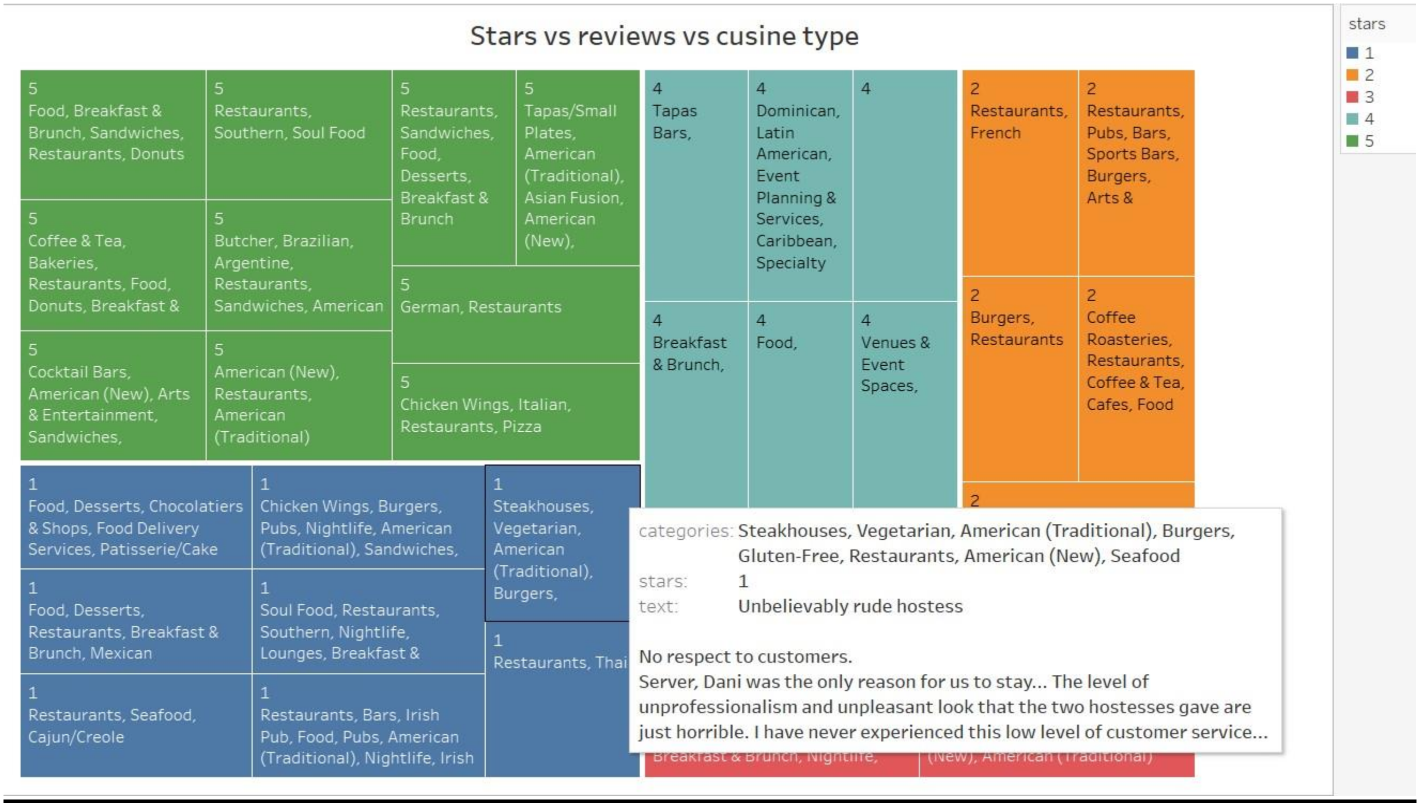
The visualization above provides information on how few Business Attributes can affect the overall ratings for that restaurant in different states.
In this project we have considered 4 attributes which makes a huge impact on the ratings which are:

a) Business accepts credit cards
b) Good for Groups
c) Table service
d) Wi-Fi

From the Graph, we can easily infer that these attributes are one of the main reasons on how the stars are rated. The low ratings (1, 2 and 3) are missing out in of these 4 attributes. The

restaurants which do not have table service generally have a lower rating when compared with the ones that have Table service.
However, the restaurants which have Table service generally have a better rating in both stars as well as review counts.

Effect of Food Categories on Ratings



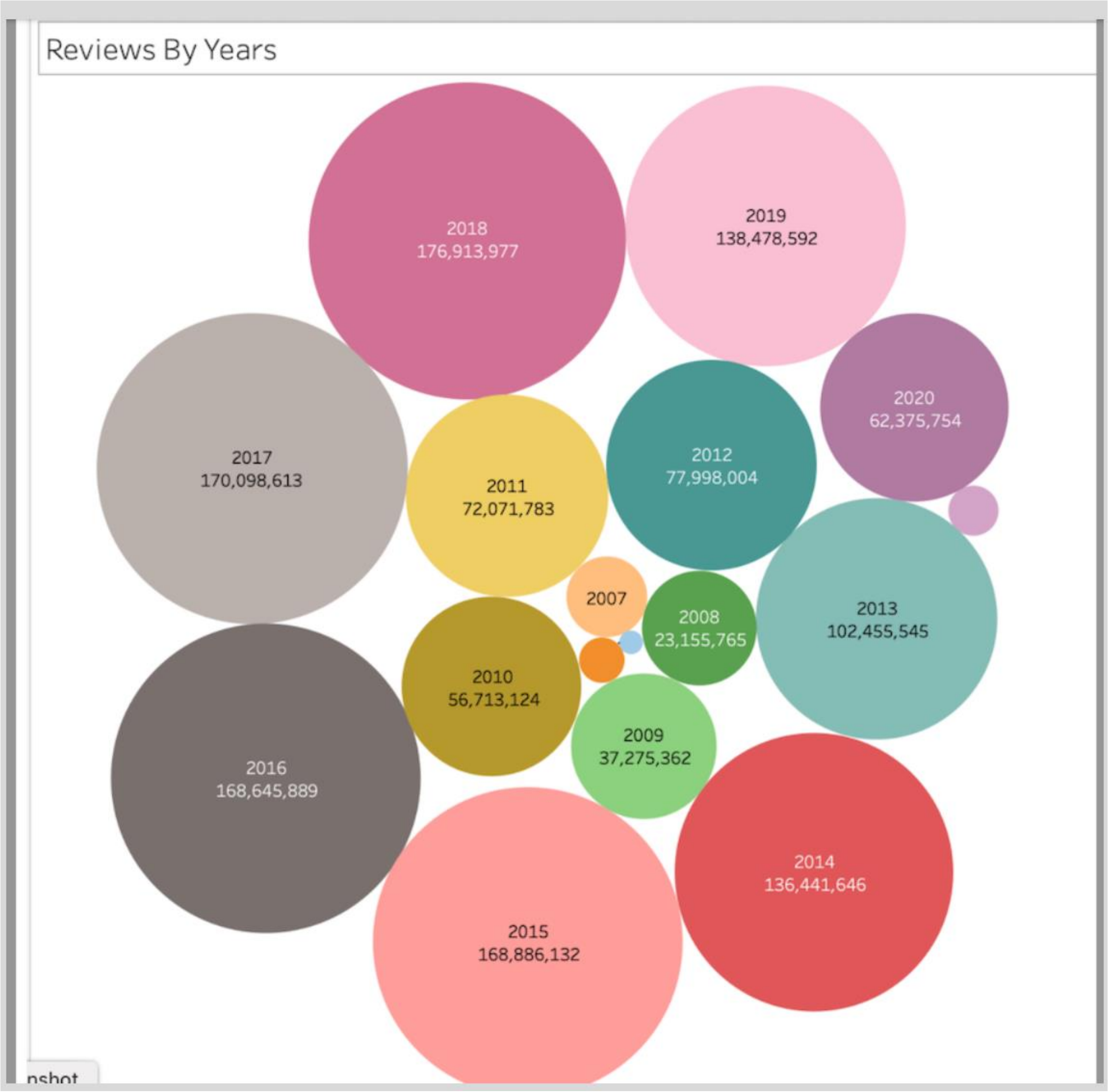
categories: Steakhouses, Vegetarian, American (Traditional), Burgers, Gluten-Free, Restaurants, American (New), Seafood
stars: 1
text: Unbelievably rude hostess

No respect to customers.
Server, Dani was the only reason for us to stay... The level of unprofessionalism and unpleasant look that the two hostesses gave are just horrible. I have never experienced this low level of customer service...

The above Graph provides insights on how Different types of Food Categories in Restaurants can affect their Ratings and reviews.
From the visualization we can infer that certain Categories of Food are more preferred and some are least preferred and how it affects the ratings given by the Customers.
The categories like Fast Food, Coffee, Sandwiches, and sports bars are most preferred over the other Food categories.
Some of the text reviews helps us provide insights on where and how that particular Restaurant went wrong. Some reviews give us information if the dish was not done right or if some other factor such as poor service or if the food didn't meet the customer's expectations were responsible for their poor ratings.

Review By Years

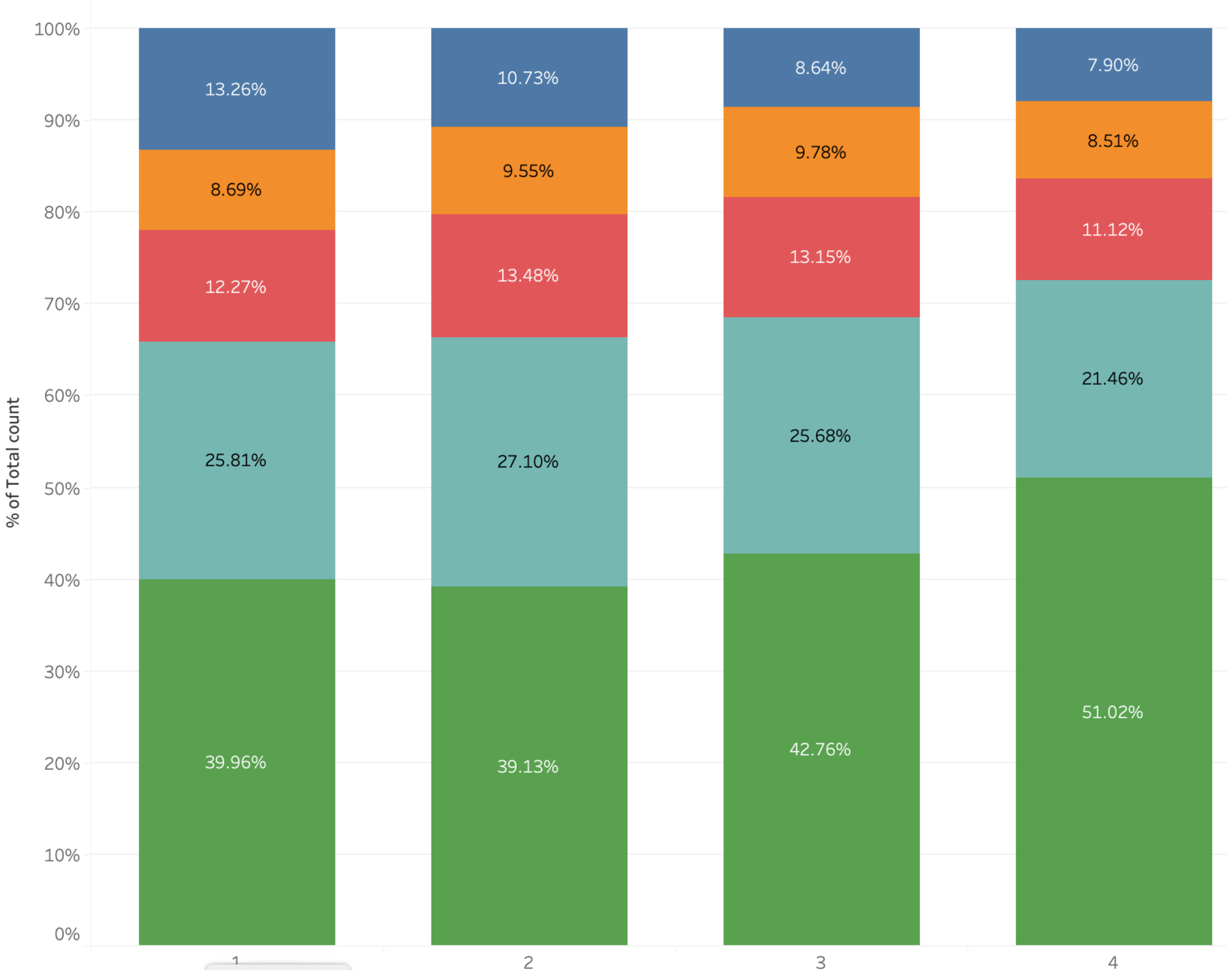
The packed bubble chart shows us the number of reviews given by users for the various restaurants business by the years. The start of the bubble is in the year 2004. The reviews start to decline in 2019 onwards. The 2020 year was during one of our most difficult times with COVID situation. The bubbles started to get smaller.



Restaurant Price ($) and stars

The horizontal axis has the restaurant price ranges. Starting from the left is the least expensive (1/$) to the right being the most expensive restaurant (4/$$$$). On Y-axis we have the number of stars given by the users as a percentage. The stars are rated from least number of stars at 1 to the greatest number of stars at 5. The number of stars rated is converted into % for ease of readability. Across the price range ($) of restaurants versus star ratings there is similarity. Within each price range ($) versus star ratings we do notice disparity

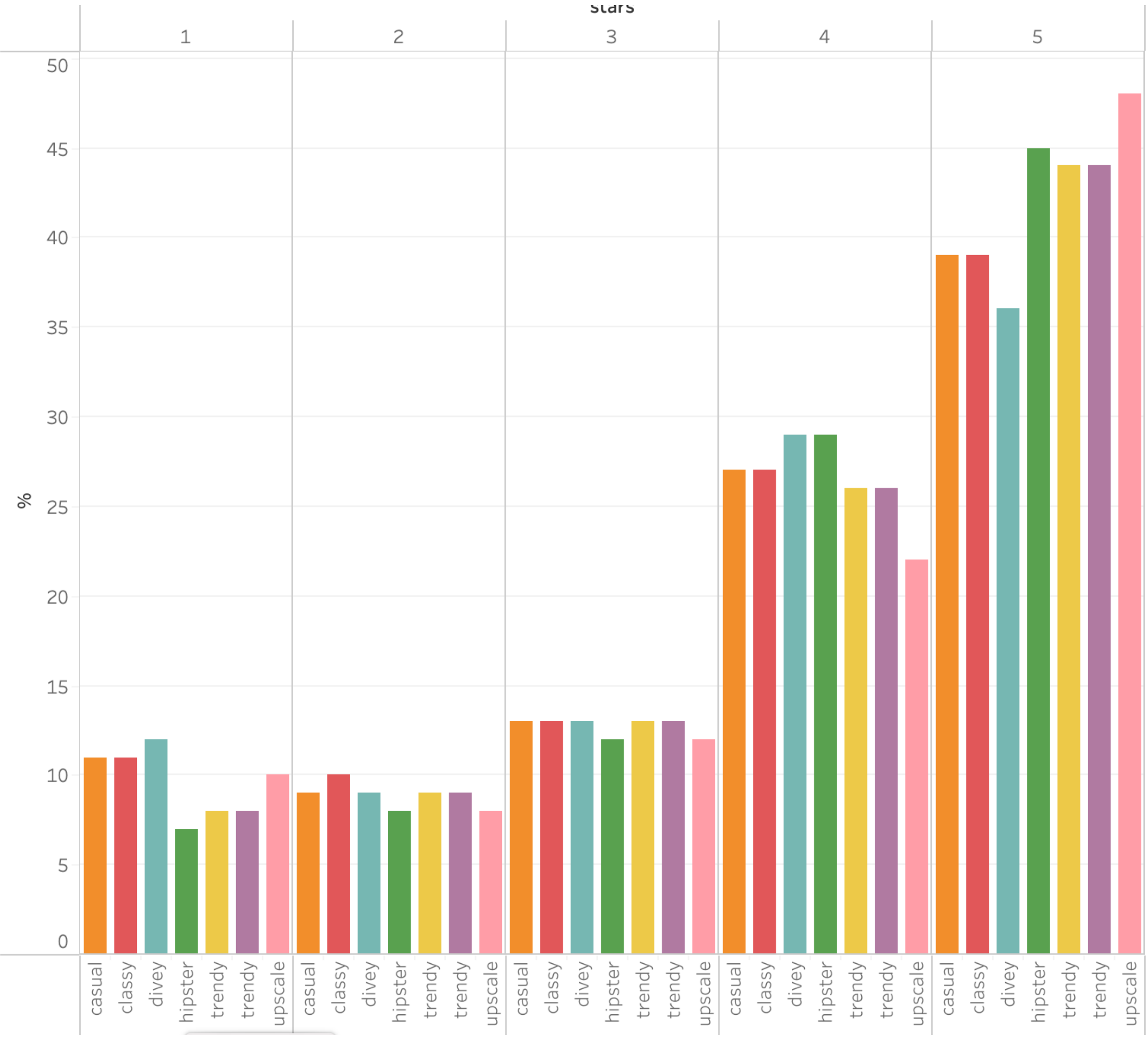
We wanted to run more analysis on the price range ($$) rated restaurants. What can $$ rated restaurants with 2-star ratings do differently so they can grow their business attracting more customers which in turn would imply more ratings.



Restaurant Attributes Compare

Analysis of one of the business attribute ambiences, in the upscale attribute we found that the 2 star/2 review rating results was at 8% while the 2 star/4 reviewer rating was at 22%. The trendy, divey classy attributes were 3 times higher with 4 stars than with 2-star rating.

We decided to go with lower hanging fruits i.e., attributes so the restaurant owners can make smaller changes having a bigger impact on reviewer ratings which in turn boost their business.



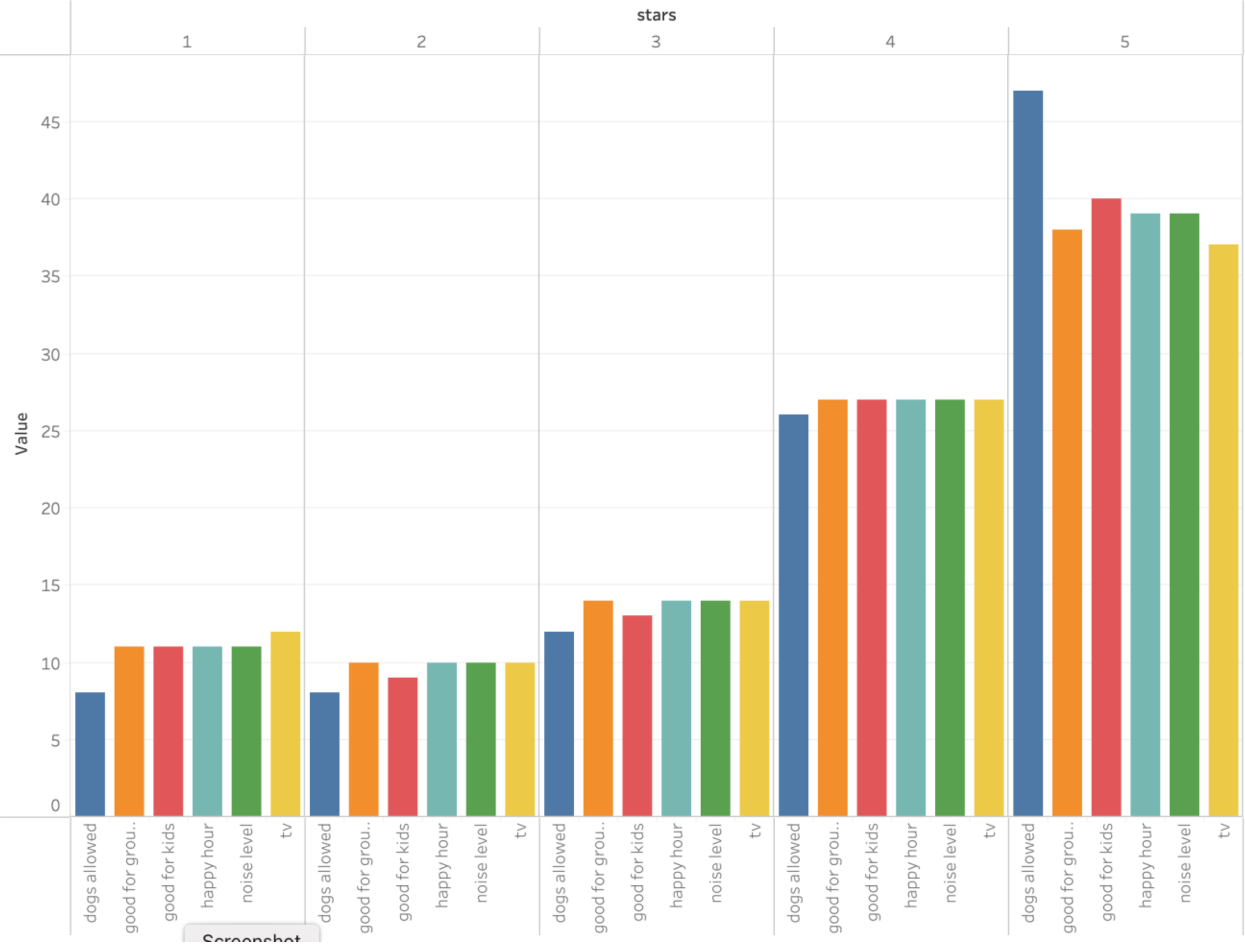
Restaurant Attributes Compare of restaurants with 2 stars

We have considered the attributes “Dogs Allowed”, “Noise Level”, “Good for Groups”, “Happy Hour”, “Has TV”, and “Good for Kids” as a subset of attributes to analyze. We felt these attributes upon small changes could have a big impact on the user rating, which in turn boost business.

The restaurants with 2 stars are at 8% while 4-stars are at 26% that allow dogs.

With COVID the number of dog owners has increased. The restaurant with 2-star rating can make access what it would take to start to allow dogs, maybe have a small outdoor seating area specific for dog owners. Keep small dogs' toys and a bowl with water for dogs.

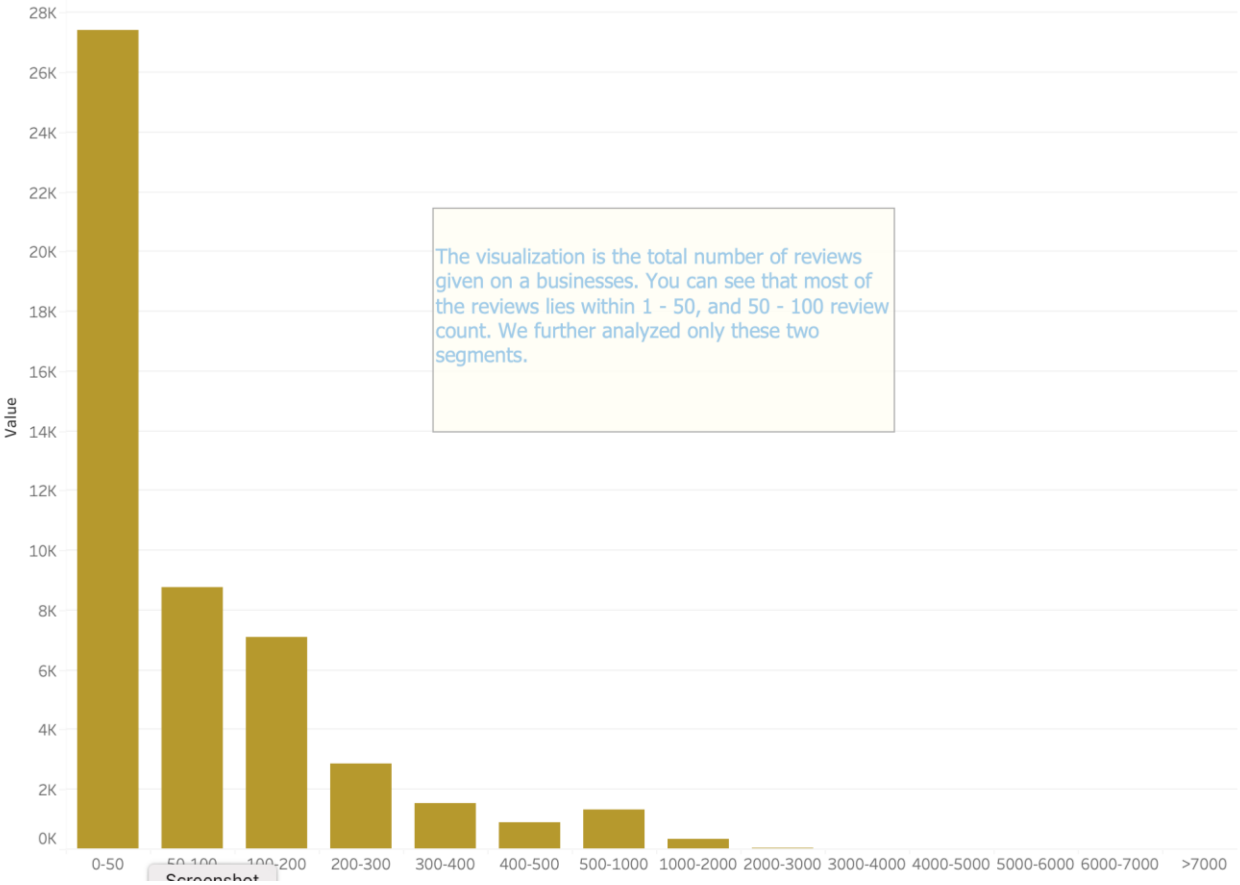
The restaurants with 2-stars are at 9% while 4-stars are at 27% for restaurants that were kid friendly. Look at the menu and add a few kids' friendly menu items. Have some stationary for kids to entertain themselves while waiting for a meal. These small changes do not cost much resulting in a better user experience.



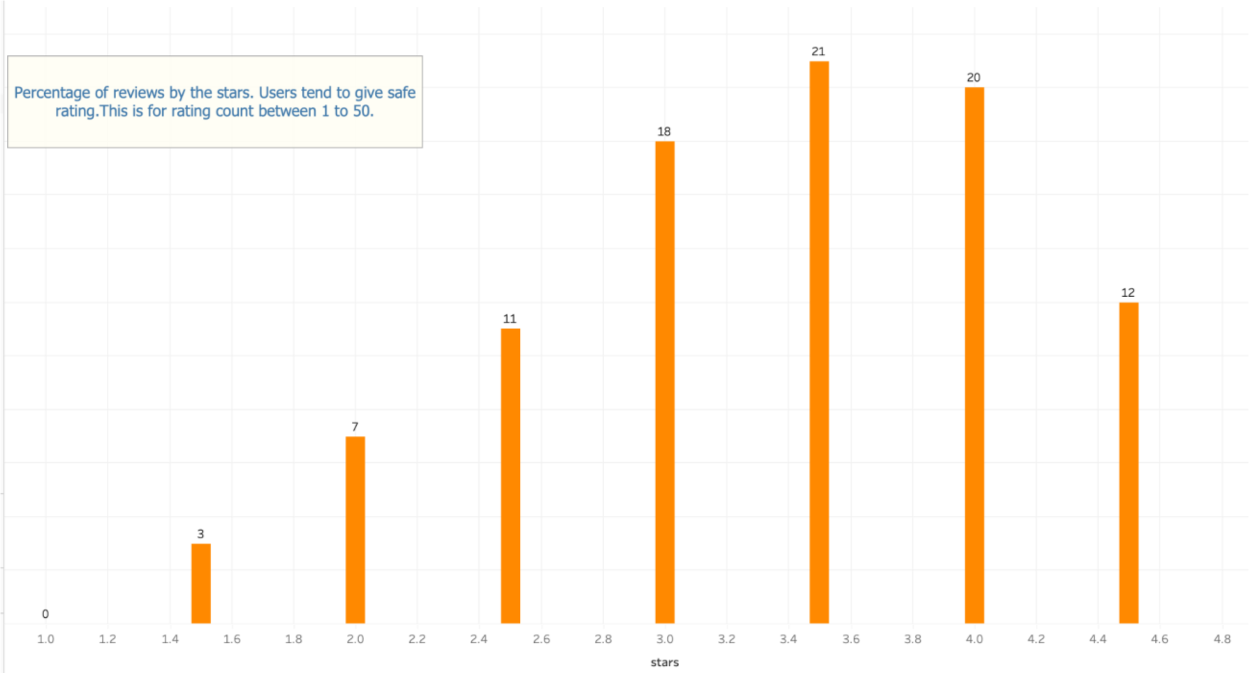
Business review counts with more analysis on stars count

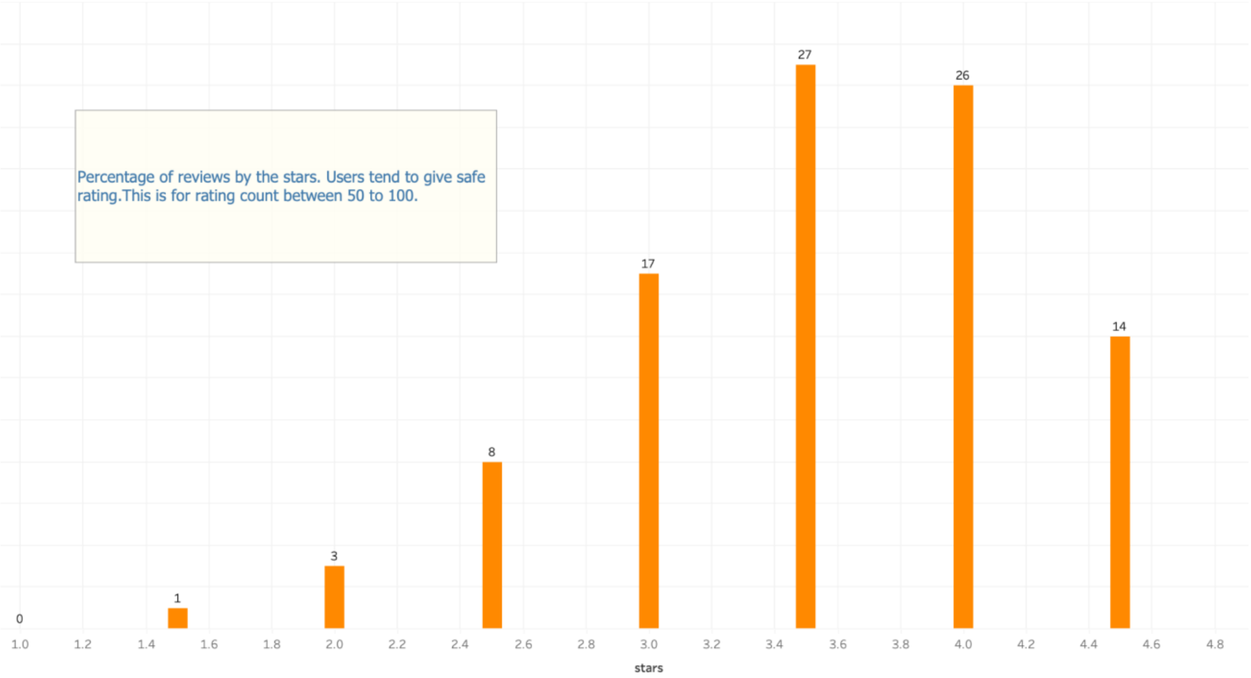
(For review count segments 1 - 50 and 50 - 100)
The visualization on the left is the total number of reviews given to the businesses. You can see that most of the reviews lie with 1 - 50, and 50 - 100 review counts. We further analyzed only these two segments. Their respective visualizations are shown on the right-top and right-bottom. We want to find out what stars were the users giving. On comparing these two segments

wherein most of ratings lie, we can say the bell curve is consistent between the volume of reviews versus stars(rating) on a business implying that users give a median rating.



The visualization is the total number of reviews given on a businesses. You can see that most of the reviews lies within 1 - 50, and 50 - 100 review count. We further analyzed only these two segments.


Percentage of reviews by the stars. Users tend to give safe rating.This is for rating count between 1 to 50.
0
3
7
11
18
21
20
12
1.0
1.2
1.4
1.6
1.8
2.0
2.2
2.4
2.6
2.8
3.0
3.2
3.4
3.6
3.8
4.0
4.2
4.4
4.6
4.8
stars

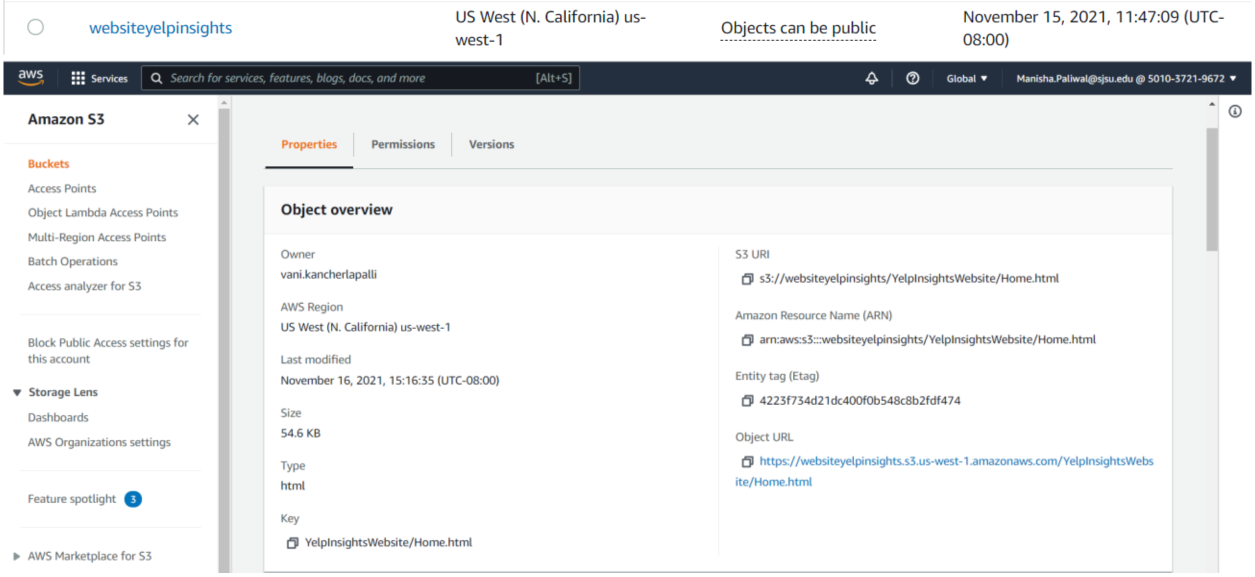

Percentage of reviews by the stars. Users tend to give safe rating.This is for rating count between 50 to 100.
0
1
3
8
17
27
26
14
1.0
1.2
1.4
1.6
1.8
2.0
2.2
2.4
2.6
2.8
3.0
3.2
3.4
3.6
3.8
4.0
4.2
4.4
4.6
4.8
stars

Website hosted on AWS S3 website hosting

[Website Link](#)

A static website is created and uploaded in AWS S3 with public access. Further, the website is hosted using AWS S3 website hosting.

Below are screenshots of S3 folder:



○ websiteyelpinsights | US West (N. California) us-west-1 | Objects can be public | November 15, 2021, 11:47:09 (UTC-08:00)

aws
Services
Search for services, features, blogs, docs, and more [Alt+S]
Global
Manisha.Paliwal@sjsu.edu @ 5010-3721-9672
Amazon S3
Buckets
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
Access analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
AWS Organizations settings
Feature spotlight 3
AWS Marketplace for S3
Properties
Permissions
Versions

Object overview

Owner
vani.kancherlapalli

AWS Region
US West (N. California) us-west-1

Last modified
November 16, 2021, 15:16:35 (UTC-08:00)

Size
54.6 KB

Type
html

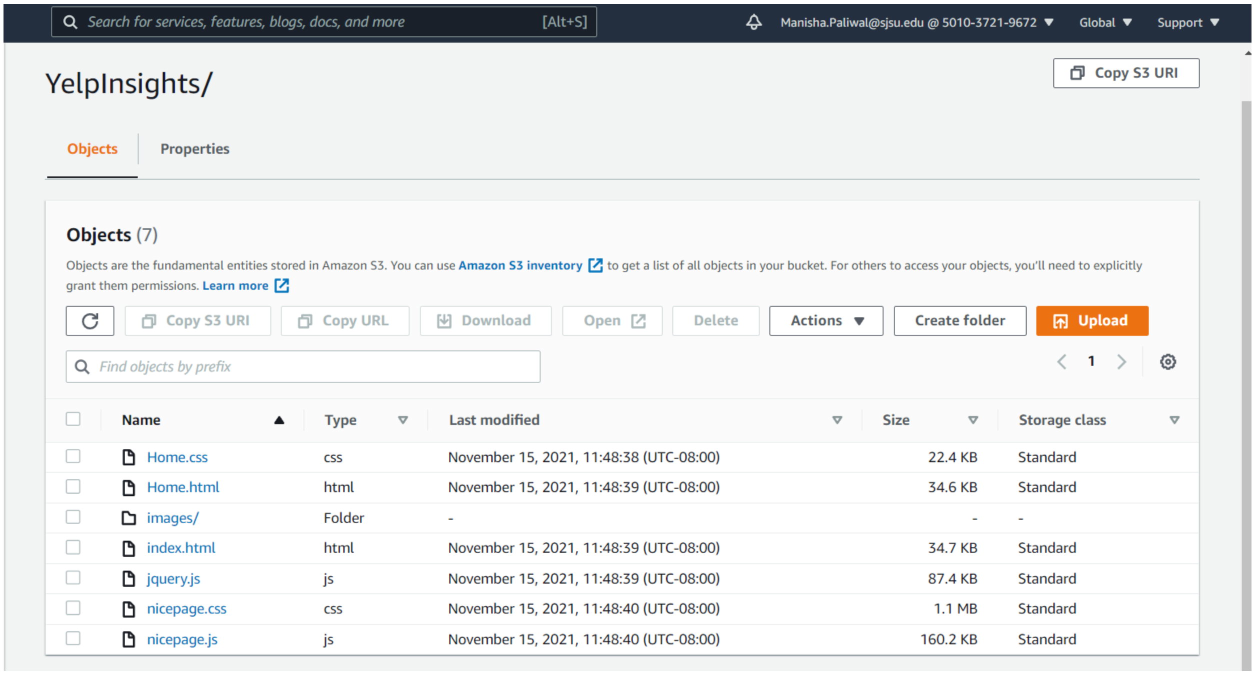
Key
YelpInsightsWebsite/Home.html

S3 URI
s3://websiteyelpinsights/YelpInsightsWebsite/Home.html

Amazon Resource Name (ARN)
arn:aws:s3:::websiteyelpinsights/YelpInsightsWebsite/Home.html

Entity tag (Etag)
4223f734d21dc400f0b548c8b2fdf474

Object URL
https://websiteyelpinsights.s3.us-west-1.amazonaws.com/YelpInsightsWebsite/Home.html

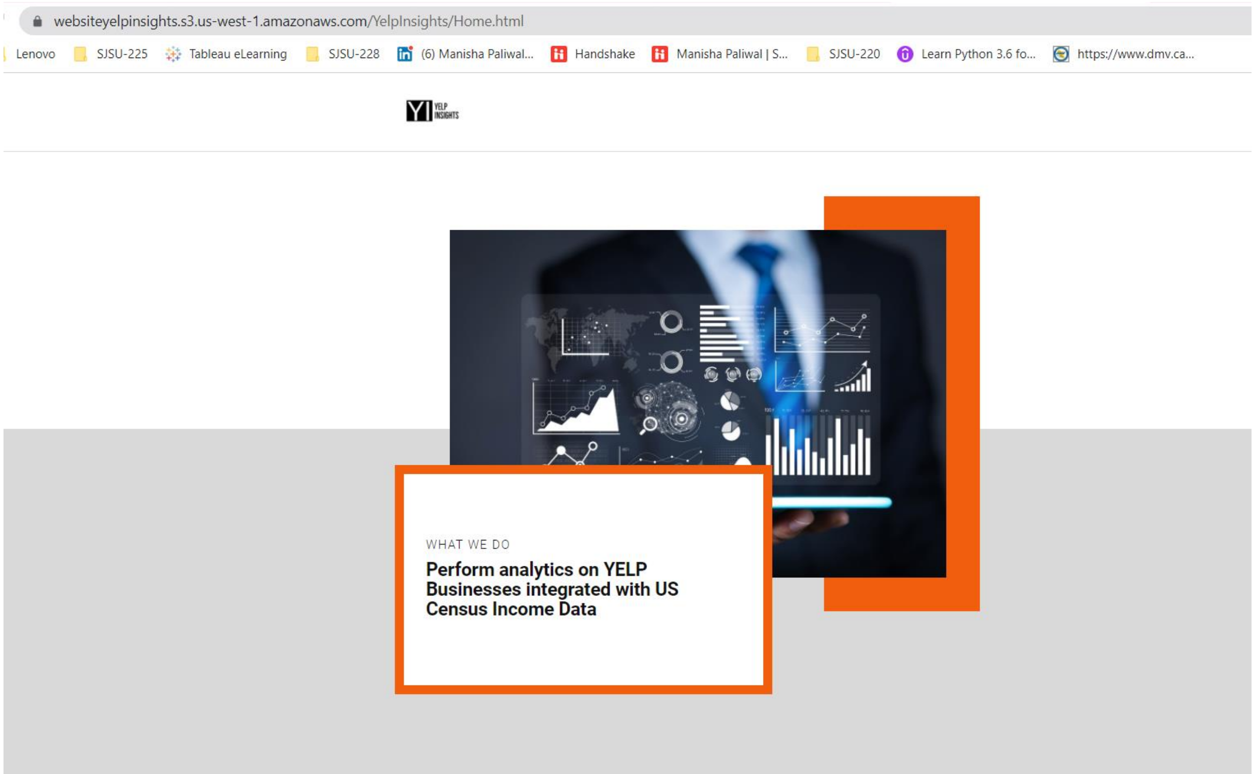


YelpInsights/

Objects (7)

Objects are the fundamental entities stored in Amazon S3. You can use **Amazon S3 inventory** to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. **Learn more**

□	Name	Type	Last modified	Size	Storage class
□	Home.css	css	November 15, 2021, 11:48:38 (UTC-08:00)	22.4 KB	Standard
□	Home.html	html	November 15, 2021, 11:48:39 (UTC-08:00)	34.6 KB	Standard
□	images/	Folder	-	-	-
□	index.html	html	November 15, 2021, 11:48:39 (UTC-08:00)	34.7 KB	Standard
□	jquery.js	js	November 15, 2021, 11:48:39 (UTC-08:00)	87.4 KB	Standard
□	nicepage.css	css	November 15, 2021, 11:48:40 (UTC-08:00)	1.1 MB	Standard
□	nicepage.js	js	November 15, 2021, 11:48:40 (UTC-08:00)	160.2 KB	Standard



WHAT WE DO

Perform analytics on YELP Businesses integrated with US Census Income Data

Testing and Validation

Test No.	001	**Phase:**	1	**Tester:**	Vani/Manisha	Oct 2021
Test Category:		**ETL Testing**				
Software Product:		AWS Glue, Python, AWS S3, Redshift				
Test Title:		Source and destination data count and validation				
Test Purpose:		Source and destination data count and validation				
Test Setup:		Manual Testing was performed by querying source and target				
Prerequisites:		AWS Glue job and source data				
Procedure:		Source Dataset uploaded in AWS S3 was compared with Target Data post ETL using SQL queries				
Expected Results:		Target dataset should be loaded as expected				
Result:		Target dataset was loaded as expected				
Reason for Failure:		No failure				
Remarks:		Test Results Passed				

Test No.	002	**Phase:**	1	**Author:**	Vani, Chidroop, Manisha	Nov 2021
Test Category:		**Visualization Test**				
Software Product:		Tableau				
Test Title:		correct data display on graph				
Test Purpose:		correct data should display on graph				
Test Setup:		rechecks were done to confirm that displayed data is matching with actual data source				
Prerequisites:		data should be loaded in AWS Redshift				
Procedure:		rechecks were done to confirm that displayed data is matching with actual data source				

Checks:	rechecks were done to confirm that displayed data is matching with actual data source
Expected Results:	graph data should match query results from AWS Redshift
Result:	graph data matched query results from AWS Redshift
Reason for Failure:	No failure
Remarks:	Test Results Passed

Test No.	003	**Phase:**	1	**Author:**	Chidroop Sagar	Date: Nov 2021
Test Category:	**Website testing**					
Software Product:	AWS S3 Web hosting and multiple screens					
Test Title:	Website hosted successfully and is compatible with multiple screens					
Test Purpose:	Website hosted successfully and is compatible with multiple screens					
Test Setup:	Browsing website using different browsers					
Prerequisites:	Website source code created and hosted using AWS S3					
Procedure:	Website link tested for accessibility and compatibility over multiple screens					
Expected Results:	Website link should be accessible and compatible					
Result:	Website link was accessible using different browsers					
Reason for Failure:	No failure					
Remarks:	Compatibility on Google chrome is the best. Needs more improvement in internet explorer					

Conclusion

- Our project focuses on analyzing current patterns of Restaurant Business on basis of Yelp Dataset and supporting US Census Household income dataset.
- We brought together the holistic view of below parameters which will enable existing business owners and new investors in Restaurant ventures to make informed decisions
    - ✓ Most popular cuisine
    - ✓ Top restaurants
    - ✓ Citizen's household Income and their influence on existing businesses patterns
    - ✓ Contribution of business attributes towards business ratings and customer experience
- American, Sandwiches, Fast Food and Mexican cuisines are widely available in the US. New venturers can look out for less available food options to cover a broader market
- Restaurants having a greater number of outlets are not necessarily the highest rated businesses
- We see a positive relationship between the household income and restaurants count
- The higher the no. of reviews a business has the more likely it is to have a better rating
- People having a higher income rate does not necessarily mean that they always prefer expensive restaurants
- While the number of reviews has been growing consistently over the years, there has been a decline in past two year due to pandemic
- There is not much disparity between the restaurant price range and the ratings

Future Scope

- Integrate Yelp Dataset with US population for each County to find correlation with Businesses
- Integrate Yelp Dataset with US area for each County to find correlation with Businesses
- Create an interactive Web Application of Data Analytics
- Perform ML Sentiment Analysis on Business Reviews
- Perform deeper analysis of Reviewer comments
- Identify areas of improvements using advanced statistical tools that businesses can incorporate for market value

Reference Links

Content References in this document:

[1] https://www.yelp.com/dataset
[2]https://data.census.gov/cedsci/table?q=Income%20%28Households,%20Families,%20Individuals%29&g=0400000US01%240500000,02%240500000,04%240500000&tid=ACSST1Y2019.S1901&hidePreview=true
[3] https://www.census.gov/acs/www/methodology/sample-size-and-data-quality/sample-size/

Website:

- https://websiteyelpinsights.s3.us-west-1.amazonaws.com/YelpInsights/Home.html

Source:

- https://data.census.gov/cedsci/table?q=Income%20%28Households,%20Families,%20Individuals%29&g=0400000US01%240500000,02%240500000,04%240500000&tid=ACSST1Y2019.S1901&hidePreview=true
- https://www.yelp.com/dataset
- https://www.unitedstateszipcodes.org/

GitHub:

- https://github.com/vanikancherlapalli/Abraca-Data-228

Architecture Diagram:

- https://app.diagrams.net/

Tableau Dashboard:

- https://public.tableau.com/app/profile/manisha.paliwal/viz/YelpInsightsByAbracaData/YelpInsights?publish=yes
- https://public.tableau.com/app/profile/vani.k4703/viz/YelpInsightsByAbracaData_Vani-1/YelpInsights?publish=yes
- https://public.tableau.com/app/profile/chidroop.sagar1502/viz/YelpInsightsbyAbracaData/AbracaDataVisualisationCs?publish=yes

Images:

- https://www.census.gov/
- https://www.yelp.com/dataset
- https://www.dictionary.com/e/zip-code/
- https://www.clipartmax.com/middle/m2H7K9G6Z5K9i8Z5_future-scope-clipart-man-with-binoculars-png/

Website Development Reference:

- Main Image
- Yelp Image
- US Census Image

- Zip Code Image

Team Members

Team Name: Abraca Data (Group 7)

Student Name	**Student ID**	**Role**
Vani Kancherlapalli	014702207	Team Lead
Chidroop Sagar	015926846	Team Member
Manisha Paliwal	015935374	Team Member