**CSE 587 DATA INTENSIVE COMPUTING**

**TERM PROJECT**

*ANALYSIS OF YELP DATA IN ARIZONA   
FOR EATING JOINTS*

Members :

**HRISHIKESH SARAF**  
*hsaraf  
50205927*

**YASH NAVIN KUMAR JAIN***yashnavi  
 50206851*

**INTRODUCTION TO DATASET**

We have taken the Yelp Dataset from Yelp Academic Dataset challenge website. It consists of various tables in JSON format. We have chosen the business table as it is relatively of lower size (~115mb) hence can be easily processed.

As the scope of the dataset it large, we have **chosen the city of Arizona** to work with as the data for Arizona was quite rich.   
Fields/dimensions are:   
**Business\_id, name, city, state, latitude, longitude, stars, review\_count, attributes, categories.**  
Highlighted ones are fields which we have used.  
Additionally, we are only concentrating on Eating Joints for the state, hence we select only those records who are of categories: **Bakeries, Bars, Burgers, Cafes, Ice Cream & Frozen Yogurts, Pizza, Sandwich & Steakhouse.**

**Tools used: We have used Tableau for data analysis and creating stories and prediction from our data.***Note: We have filtered our data in each step for state of Arizona.*

**ACTIVITY 1  
STORY TELLING WITH OUR DATA**

Based on the data of Arizona we have created a story about eating joints in the state where we do some analysis and answer some questions.

**Dash Board 1: Category and Map**Input: Business data extract created from dataset. **Question: What are the popular categories of eating joints in Arizona?**This dashboard involves 2 sheets.

1. **Category**
   1. This sheet maps all the categories of eating joints we are considering, and creates **packed bubble chart**.
   2. The size of the bubbles is controlled by the distinct number of Businesses (Eating Joints) in that area.
   3. The bubbles are colored according to the categories.
2. **Map**
   1. This sheet displays all the businesses in Arizona on a Map.
   2. We have applied a filter of Categories, where we can select the category and see the distribution of it in the whole state.
3. **Action Filter**To combine both the sheets, we have created on - select action where we select a category in one sheet and it filters out the result in the Map.

**Conclusion:   
We can see that Bars is the most popular categories of eating joints followed by Sandwiches, Pizza and Burgers.**

**CHART: DISTRIBUTION BY CITY**Input: Business data extract created from dataset. **Question: How eating joints are distributed over the state of Arizona?**  
  
To answer this question we have created a **bar graph**, by selecting:

1. **Rows:** City – Arranged in decreasing order of number of businesses.
2. **Columns:** Distinct count of number of business.
3. **Filters:** 
   1. Selecting cities of Arizona.
   2. Selecting only the top 10 cities with the most number of eating joints.

Categories of eating joints. We can change the input to our preferred category.

1. **Color:** Colored by category.
2. **Labels:** Labelled by distinct count of eating joints of each category.

**Conclusion:   
We can see that the city with most number of eateries will obviously be Phoenix which is the capital of Arizona. The other cities on the list are the other popular cities of Arizona like Scottsdale, Mesa, Tempe, etc.**

**Dash Board 2: IMPORTANCE OF TV IN BARS**

As bars are quite popular in Arizona we dive deep into them and try to answer: **Question: What are the popular categories of eating joints in Arizona?**This dashboard involves 3 sheets.

1. **Distribution of Bars with/without TV**
   1. **Rows:** City – Distinct count of business IDs.
   2. **Columns:** Attributes field filtering (Bars with TV, Bars without TV)
   3. **Filters:** 
      1. Selecting all those eateries which are of Bar category.
      2. Selected attributes – Bars having TV and not having TV.
2. **Map of all Bars**
   1. This sheet displays all the bars in Arizona on a Map.
   2. Displaying business with the selected attributes (Bars having/not having TV).
   3. The color of the bars is depicted by average star ratings received.
3. **Number of Reviews**
   1. **Rows:** City – Attributes (Bars Having TV/Bars not having TV)
   2. **Filters: on Action filter from Distribution which will return the number of bars of that attributes.**
4. **Action Filter**To combine both the sheets, we have created on - select action where we select an attribute on **Distribution of Bars with/without TV** sheet and it filters out the map result on **Map of all Bars** sheet and from **Number of review sheets** it gives out the number of bars **(of that attributes)** which have received attribute.

**Conclusion:**

**We can observe that there are more bars with TV than without TV. That is probably because bars are a good place where people come together and watch football or meet friends.**

**Dash Board 3: Price Range and Location**Input: Business data extract created from dataset. **Question: How are eateries distributed across the state based on price range?**This dashboard involves 2 sheets.

1. **Price Range Chart**
   1. This sheet maps all the eating joints we are considering, and creates **pie chart** on the distribution of joints based on price range.
   2. Price range can be: **Cheap, Moderate, Expensive, Very Expensive.**
   3. The size of the angle depends on distinct number of records for that price range.
   4. The angles are colored according to the price range.
2. **Location based on Price Range**
   1. This sheet displays all the eateries in Arizona on a Map.
   2. We have applied a filter of price range attribute, where we can select the category and see the distribution of it in the whole state.
   3. Color of the points are controlled by the average star rating for that joint.
3. **Action Filter**To combine both the sheets, we have created on - select action where we select a price range on pie chart in price range sheet and it filters out the result on the map.

**Conclusion:**

**We can see from the pie chart and map results, that there are more eateries which are cheap or moderate compared to expensive ones. People on an average have lower household incomes in these areas, which is why they prefer cheap/moderate places.**

**ACTIVITY 2  
PREDICTIONS**

The Story talks about three elements:

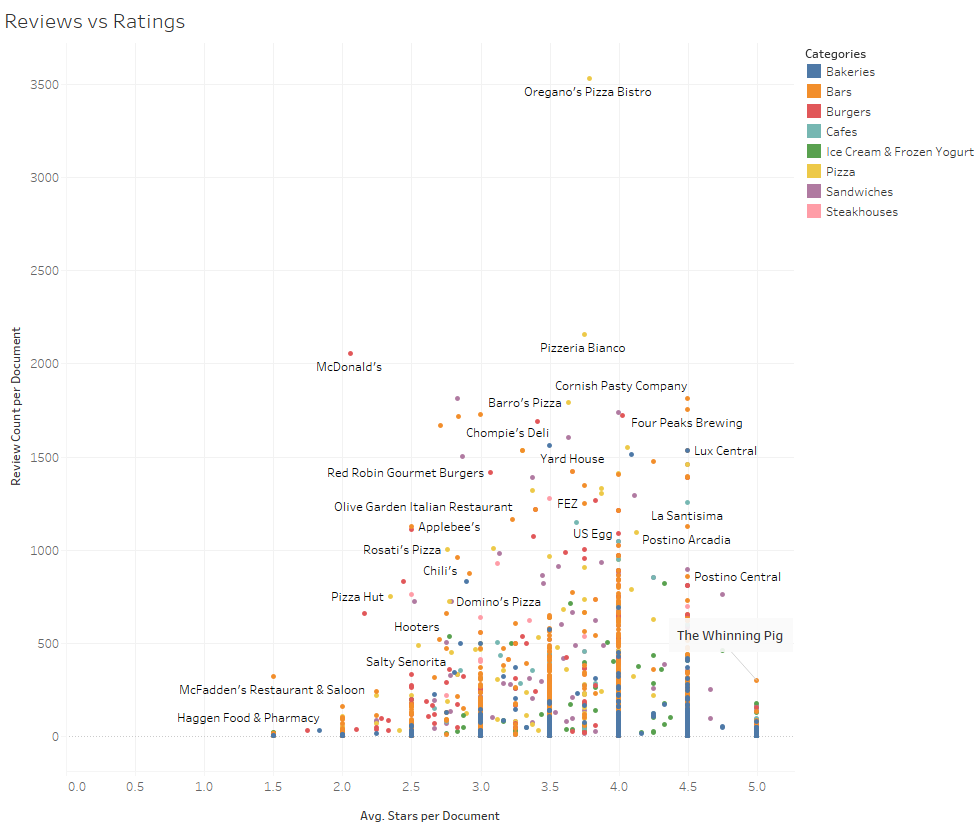
1. **Reviews (Review Count per Document) v/s Ratings (Avg. Stars per Document):**

We plotted a scatter-plot using the measures and made an interesting observation.  
The graph shows what a user might consider as “good” restaurant on Yelp! app.  
Whenever a user searches for a restaurant, he usually considers both factors (Reviews, Ratings).  
The graph below shows what makes a food joint “Good”, “Moderate”, “Bad” or “Can’t say”.  
Let’s first consider the example of McDonald’s in the plot. The average ratings received by MacDonald’s is slightly greater than 2 but people who have reviewed it are more than 2000 which makes it a “Moderate” food joint.

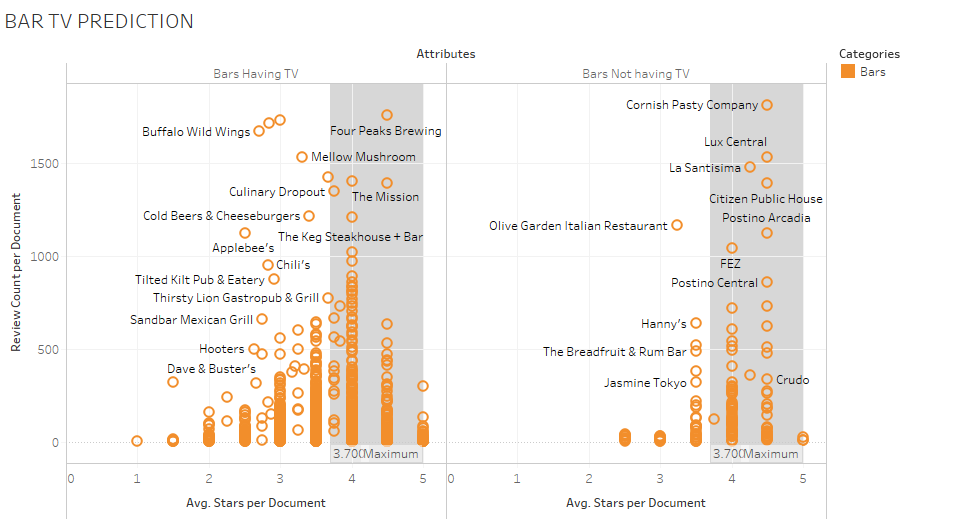
Oregano’s Pizza Bistro is at the top in reviews lies between 3.5 and 4 in reviews which makes a “Good” food joint.

McFadden’s Restaurant and Saloon is joint that has 1.5 in ratings and less than 500 in reviews which makes it a “bad” restaurant.

Whinning Pig is a restaurant which is rated 5 but hasn’t been reviewed by many people. So, we “Can’t say” if the food joint is good or bad.



1. **Prediction based on Bars with TV and Bars without TV:**



The two scatter plots compare the densities of bars having different values of Review Count and Avg. Ratings received. We can say that a bar which is reviewed more is more likely to be more frequented by people than bar which isn't.

**Obervation1:** The density of points in the highly-rated region is more in the case of Bars Having TV compared to Bars Not Having TV.

**Observation2:** There are more bars with reviews above 1500 in the case of Bars Having TV compared to Bars Not Having TV.

**Considering both observations we can predict that a bar is likely to be reviewed more and rated highly and as a result frequented more if it has a TV.**

1. **Prediction based on Success Metric:**

We have defined a “Success Metric” based on a simple assumption that the food joints that are most reviewed are also most visited. Taking this assumption into consideration we define “Success Metric” of a food joint category as:  
**Success Metric = Sum of *Reviews per joint* in the category/ Number of Distinct *Food Joint IDs* in the category.**

For prediction, we have shown how Success Metric in a food category varies in different cities.

Let’s consider the city of Phoenix:

We have 57 Steakhouses with Success Metric as high as 171.4 whereas, we have 399 Sandwich joints with Success Metric as low as 77.1. With this data, we can predict that a new Steakhouse business can do well in Phoenix because there are only 57 there and they’re reviewed/frequented more. Whereas, a Sandwich business may not do good as there are 399 of them and not reviewed/frequented often.

**So, with the help of “Success Metric” we can predict whether a certain *type* of joint may succeed in a city or not.**

