Pizza! It’s one of America’s most beloved dishes, eaten anywhere, anytime, and by everyone. The objective of this project is to find excellent pizza based on the location, review and price, and finally build a pizza-map focused on New York, which is known as the “Pizza Capital of the U.S.”

I am using the data sources from Jared, Barstool, and Datafiniti. Jared’s data is from top NY pizza restaurants, with a 6-point Likert scale survey on ratings. The Barstool sports dataset has critic, public, and the Barstool Staff’s rating as well as pricing, location, and geo-location. Datafiniti includes 10000 pizza places, their price ranges and geo-locations.

The purpose of the data is to find the best pizza restaurant, focusing New York City. To achieve that Tyler Ricards recorded the web traffic coming through the OneBite application.

I have analyzed and addressed the following business statements-

* Rating wise visualization of Restaurants in New York
* What are the pizza ratings across various price categories?
* What is the pricing of pizza across categories?
* What are the consumer preferences across categories of pizza in United States?
* Do higher priced restaurants have better ratings?
* Cluster pizza restaurants
* visualizing restaurants according to the clusters formed
* Predicting consumer ratings

Approach –

* Data cleaning
* EDA to detect distribution ranges, percentage of missing values and outliers
* Join three datasets to create one that has ratings and price ranges to compliment all of the data analysis
* Visualize ratings of pizza restaurants through interactive maps
* Using multiple graphical and statistical representations analyze trends in pricing, consumer ratings and geographical locations
* Conduct Hypothesis testing to compute if there is any significant difference between pizza ratings across price categories of restaurants
* Cluster restaurants on the basis of variables like price category and various ratings
* Fit a machine learning model to predict the consumer ratings
* K-means clustering technique will be used to form clusters of restaurants
* Use interactive maps and bar graphs to understand the consumer preferences and pricing across various restaurants in New York

Libraries used –

**library**(readr)

**library**(dplyr)

**library**(tidyverse)

**library**(ggplot2)

**library**(ggmap)

**library**(car)

**library**(plotly)

**library**(bootstrap)

**library**(cluster)

**library**(factoextra)

**library**(gridExtra)

**library**(leaflet)

**library**(DataExplorer)

Importing Data -

barstool <- readr::read\_csv("https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2019/2019-10-01/pizza\_barstool.csv")

datafiniti <- readr::read\_csv("https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2019/2019-10-01/pizza\_datafiniti.csv")

jared <- readr::read\_csv("https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2019/2019-10-01/pizza\_jared.csv")

Data cleaning –

Generally we employ data cleaning steps to derive relevant insights from the data and to get rid of garbage values. There are only two missing values in the latitude and longitude column in the Barstool data. Also, there is a value out of range in Dave’s rating .No imputation is being performed as the count is less than 5% of the data in both cases. Critic ratings have 401 zeroes, Community ratings have 41 zeroes, Minimum price range contains 1852 zeroes, Votes have 104 zeroes. Apart from this the data is clean.

**Insights –**

**Graphical user interface

Description automatically generated with low confidence**

Through the Plot Bar chart of Datafiniti, we found that New York is the dominant state (province) in the dataset.

**Correlation between various pizza ratings**

## provider\_rating community\_score critic\_score dave\_score

## provider\_rating 1.00000000 0.31921901 -0.07935913 0.22096952

## community\_score 0.31921901 1.00000000 -0.05570681 0.60522594

## critic\_score -0.07935913 -0.05570681 1.00000000 -0.04922468

## dave\_score 0.22096952 0.60522594 -0.04922468 1.00000000

Correlation between dave score and community score is 0.6, between provider\_rating and community score is 0.32 and correlation between provider\_rating and dave score is 0.22.

**Comparing pizza rating across states:**

**Chart, bar chart, histogram

Description automatically generated**

**Chart, bar chart, histogram

Description automatically generated**

Analyzing ratings across pizza categories

**Chart, box and whisker chart

Description automatically generated**

**Chart, box and whisker chart

Description automatically generated**

Normal Pizza Restaurants have slightly higher All average score as compared to those restaurants which serve Italian pizza, however provider\_rating has very similar distribution across both the categories

**Comparing price range across pizza categories**

## category AVERAGE\_MAX\_PRICE AVERAGE\_MIN\_PRICE

## \* <chr> <dbl> <dbl>

## 1 ALCOHOL SERVING 32.1 26.4

## 2 CATERERS 27.9 23.2

## 3 ITALIAN 30.9 25.4

## 4 NORMAL PIZZA RESTAURANT 27.0 24.1

Alcohol serving pizza restaurants have the highest average min and max price range followed by Italian pizza restaurants. Caterers and Normal Pizza restaurants have similar min and max price range.

**Regression for predicting community ratings - OneBite user ratings**

**Chart

Description automatically generated with low confidence**

Adjusted R square of the model is 0.46. Since we are predicting consumer ratings which has very high variation, we can accept this R square value. Also, p value associated with F test and most of the individual t tests in not significant and we can reject the NULL hypothesis at 95% confidence level.

**Clustering**

Clustered pizza restaurants on the basis of price level and community ratings.

**Line chart, polygon

Description automatically generated**

Elbow Curve to decide the optimum number of clusters looking at the bend

**Chart, line chart

Description automatically generated**

Comparing the clusters

## Cluster review\_stats\_community\_average\_score price\_level

## \* <int> <dbl> <dbl>

## 1 1 5.40 0.922

## 2 2 7.92 2.03

## 3 3 5.78 2.02

## 4 4 7.51 0.932

So we have 4 cluster which signify

* Cluster\_1 - Low Rating and Low Price restaurants
* Cluster\_2 - High Rating and High Price restaurants
* Cluster\_3 - Low Rating and High Price restaurants
* Cluster\_4 - High Rating and Low Price restaurants

Key takeaways from this analysis – the implications to the consumer/business owners

* Upon understanding the consumer preferences, the restaurant owners can better target them.
* Predicting the consumer ratings in advance on the basis of provider ratings then that can immensely help the owners. They can know on a scale how much the consumer is going to like them based on past behavior for similar places and the available provider ratings.
* Through cluster analysis consumers will get to know which restaurant the best is taking into consideration consumer pizza ratings and price category of the restaurant.

We find that there is low correlation between commmunity, provider, critic and jared ratings. Community and Dave ratings have moderately high correlation of ~0.6 New York have lower provider and average pizza ratings on average as compared to rest of the US States with high ratings- IA,OK,FL,OH States with low ratings - WV,MI,NV,SC Restaurants serving Italian Pizza have lower ratings on average as compared to Non Italian Pizza restaurants. Alcohol serving and Italian pizza restaurants have higher priced pizza as compared to those that do not fall in this category High priced restaurants have better pizza ratings as compared to low priced restaurants