

# Identifying Insightful Analytics for Restaurant Owners via Implementing Sentiment Analysis on Yelp Data

ITEC 2600 - C

Syed Shah	216005126
Mir Umar Naeem	216762478
Megh Ladumor	216763922

## Abstract:

Our objective is to create a model which will use existing yelp data to provide insightful analytics. Using this data model we will help existing and future restaurant owners make important decisions such as how they can improve their services and for those who are opening a new restaurant we will suggest the best modeled dimensions which will enable them to achieve not only a good rating but also a valuable business by attracting more customers and giving them what they like the most.

## Introduction:

What makes a good restaurant? What are the major concerns of customers? Common knowledge may give general answers like delicious food, great services or pleasant environments, but they might not be true for different types of restaurants. In this analysis, we are going to unveil those essential features behind all kinds of restaurants via sentiment analysis on Yelp data.

We will be using a variety of dimensions found in the data set to execute sentiment analysis. Focusing on the attributes that most business owners over see such as hours of operation, ambience, payment methods, bar, parking etc. To complete the extraction or classification of sentiment from sentiment reviews we will be using Text Analysis. Getting an insight of what the customers like and dislike. All of this will be done using Matlab.

## Data and Methods

### A. Data Description

The dataset (Data\_Source) originated from the online Yelp Dataset. Consisting of basic business information such as hours, address, ambience etc. For this analysis, we focused on customer reviews for restaurants along with the business attributes data.

```
%%% Importing the data source in matlab  
  
Data_Source = readtable('yelp_business.csv');
```

Ratings in our case is the discriminator for good or bad sentiment. Business category has been used to group data into smaller chunks. Type of restaurants and their quantity. The selected categories are as follows:

1. American
2. Chinese
3. Japanese
4. Mexican
5. Pakistani/Indian
6. Italian

From the data set we will be using the following columns to carry out certain methods which will help us move towards our goal.

1. Categories
2. Star Rating
3. Review Count
4. Noise Level
5. Alcohol

## B. Data Cleaning

As the imported data set consisted of a vast variety of businesses we first filtered out only restaurants.

### 1. Restaurant\_Table

All rows related to restaurants were extracted and saved in a table called **Restaurant\_Table**.

```
%% selecting index of rows that are related to Restaurants

pattern = "Restaurants";
TF = contains(Data_Source.categories, pattern);

%% Looping through the Data_Source to extract all rows that consist of %%Restaurants

j = 1;
for i = 1: 42153
    if(var(i,1)==1)
        Restaurant_Table(j,:) = Data_Source(i,:);
        j = j+ 1;
    end
end
```

### 2. Classifying Types of Restaurants

We then stemmed, lemmatized and filtered out types of restaurants. Storing them in the respective tables so further method execution can be simplified and fast.

```
%% Tables to separate different types of restaurants
American_Rest_Tbl = table;
Chinese_Rest_Tbl = table;
Mexican_Rest_Tbl = table;
Italian_Rest_Tbl = table;
Japanese_Rest_Tbl = table;
Pak_Ind_Rest_Tbl = table;

%% Separating Tables based on rest type

j = 1;
text = "Mexican";

for i = 1: 14303
    if(var(i,1)==1)
        Mexican_Rest_Tbl(j,:) = Restaurant_Table(i,:);
        j = j+ 1;
    end
end
```

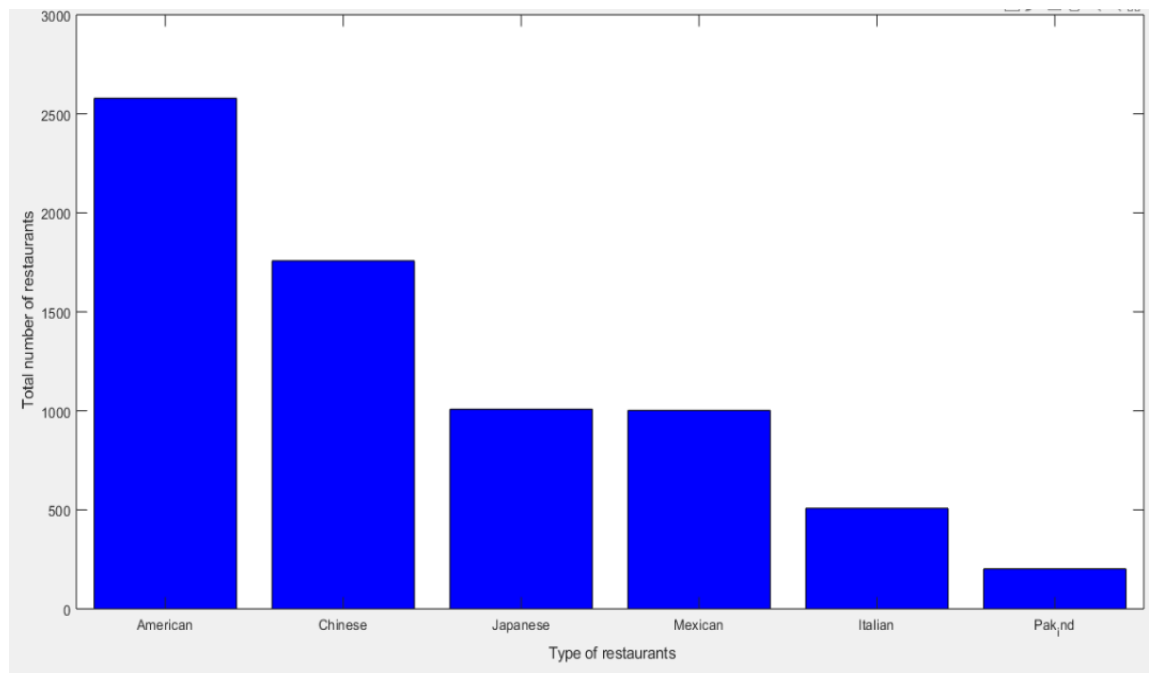
The same was done for all types of restaurants.

### C. Methods

We developed a number of methods to compute various dimensions which will help us move towards a conclusion.

#### 1. Total number of restaurant types

We identified the different types of restaurants and their quantity. This will give us the population of restaurant we can work with.



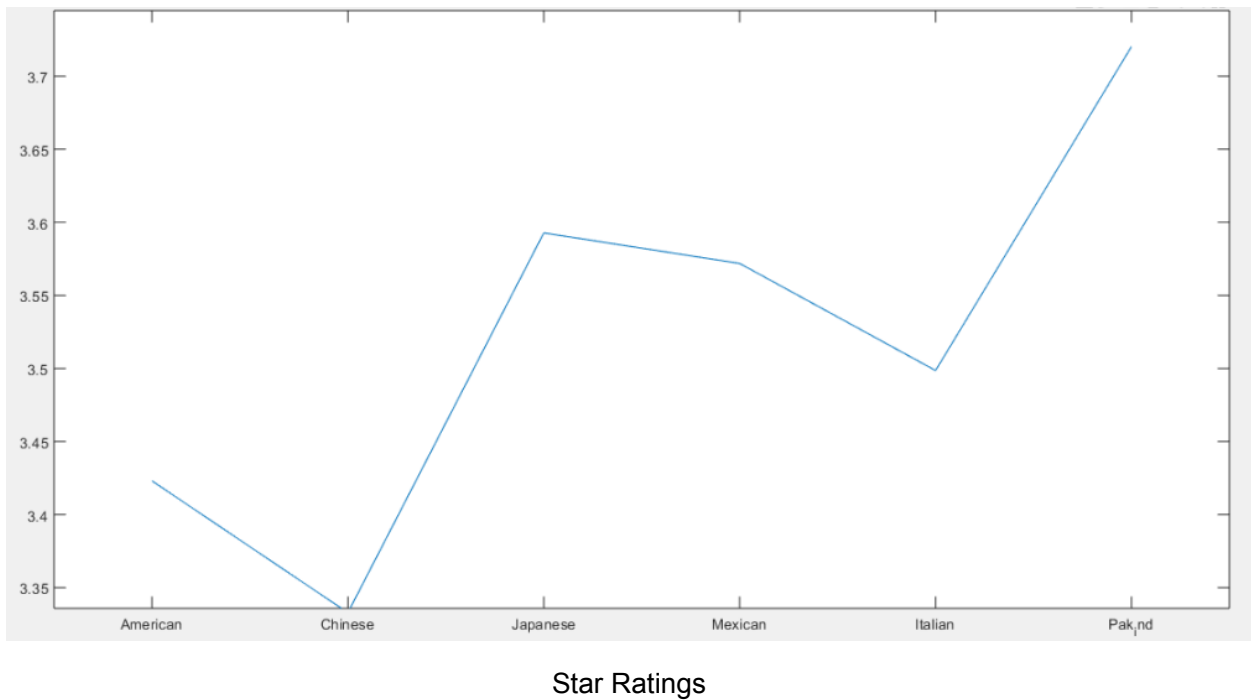
#### CODE :

```
X = Num;  
Y = categorical(Names);  
Y = reordercats(Y,Names);  
bar(y, x, 'b')  
xlabel('Type of restaurants')  
ylabel(' Total number of restaurants')
```

## 2. Distribution of Stars

Using the star rating column from the data set, we took the average of all ratings for each type of restaurant. Plotting a graph gives us a better view of their current standing.

Using this we can select the best restaurant which will tell us what attribute customers like the most, hence the good star rating



CODE:

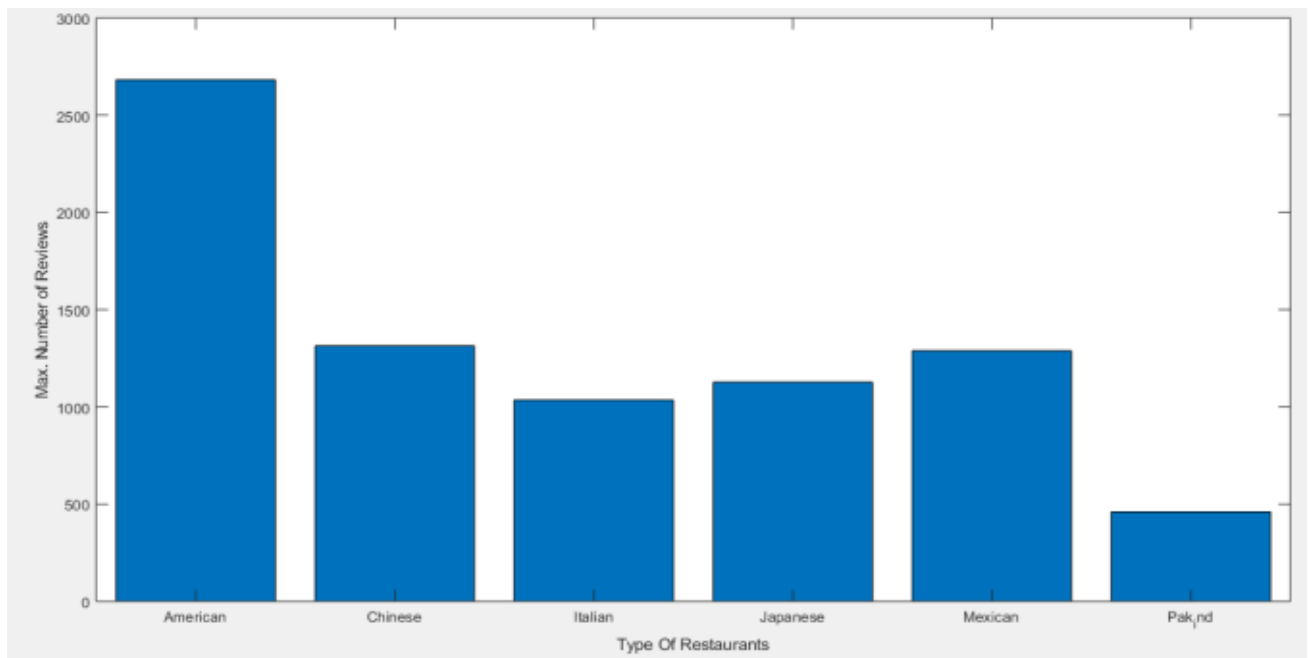
```
RestaurantMean=[mean(American_Rest_Tbl.stars),  
mean(Chinese_Rest_Tbl.stars),  
mean(Italian_Rest_Tbl.stars),  
mean(Japanese_Rest_Tbl.stars),  
mean(Mexican_Rest_Tbl.stars),  
mean(Pak_Ind_Rest_Tbl.stars)]  
y=categorical(Names);  
y=reordercats(y,Names);  
plot(y,RestaurantMean)
```

### 3. Highest Reviewed Restaurants

Determining the highest reviewed restaurants gives us accuracy. The more number of reviews lead to a more accurate rating. High volume of reviews show that a larger population actually experienced that restaurant giving credibility to the rating it holds.

If a restaurant has high rating with a large number of reviews, this is proof of the restaurant being authentic as despite a large number of reviews the mean star rating is still high.

Below we plotted out the number of reviews against the type of restaurants.



#### 4. Top 50 Based on stars

As seen in our previous graph, there is a large number of restaurants that have a good rating ( above 4 / 5).

By selecting the top 50 restaurants based on their star rating we can develop a subset of our already filtered dataset that will help us deduce that the attribute they have are most liked by customers.

Doing so will enable us to dive in deeper and have a look at what differences they hold among themselves.

Executing the following code gives us the new dataset

```
-----  
code for top 50 using star ratings  
-----  
j =1;  
>> for i = 1:2585  
if(American_Rest_Tbl.stars(i) >4)  
American_Top_Star(j,:) = American_Rest_Tbl(i,:);  
j=j+1;  
end  
end  
-----  
j =1;  
for i = 1:508  
if(Japanese_Rest_Tbl.stars(i) >4)  
Japanese_Top_Star(j,:) = Japanese_Rest_Tbl(i,:);  
j=j+1;  
end  
end  
-----  
j =1;  
for i = 1:1002  
if(Chinese_Rest_Tbl.stars(i) >4)  
Chinese_Top_Star(j,:) = Chinese_Rest_Tbl(i,:);  
j=j+1;  
end  
end  
-----  
>> j =1;  
for i = 1:1008  
if(Italian_Rest_Tbl.stars(i) >4)  
Italian_Top_Star(j,:) = Italian_Rest_Tbl(i,:);  
j=j+1;  
end  
end  
end
```

```
>> j =1;
for i = 1:1749
if(Mexican_Rest_Tbl.stars(i) >4)
Mexican_Top_Star(j,:) = Mexican_Rest_Tbl(i,:);
j=j+1;
end
End
-----
```

```
>> j =1;
for i = 1:202
if(Pak_Ind_Rest_Tbl.stars(i) >4)
Pak_Ind_Top_Star(j,:) = Pak_Ind_Rest_Tbl(i,:);
j=j+1;
end
end
end
```



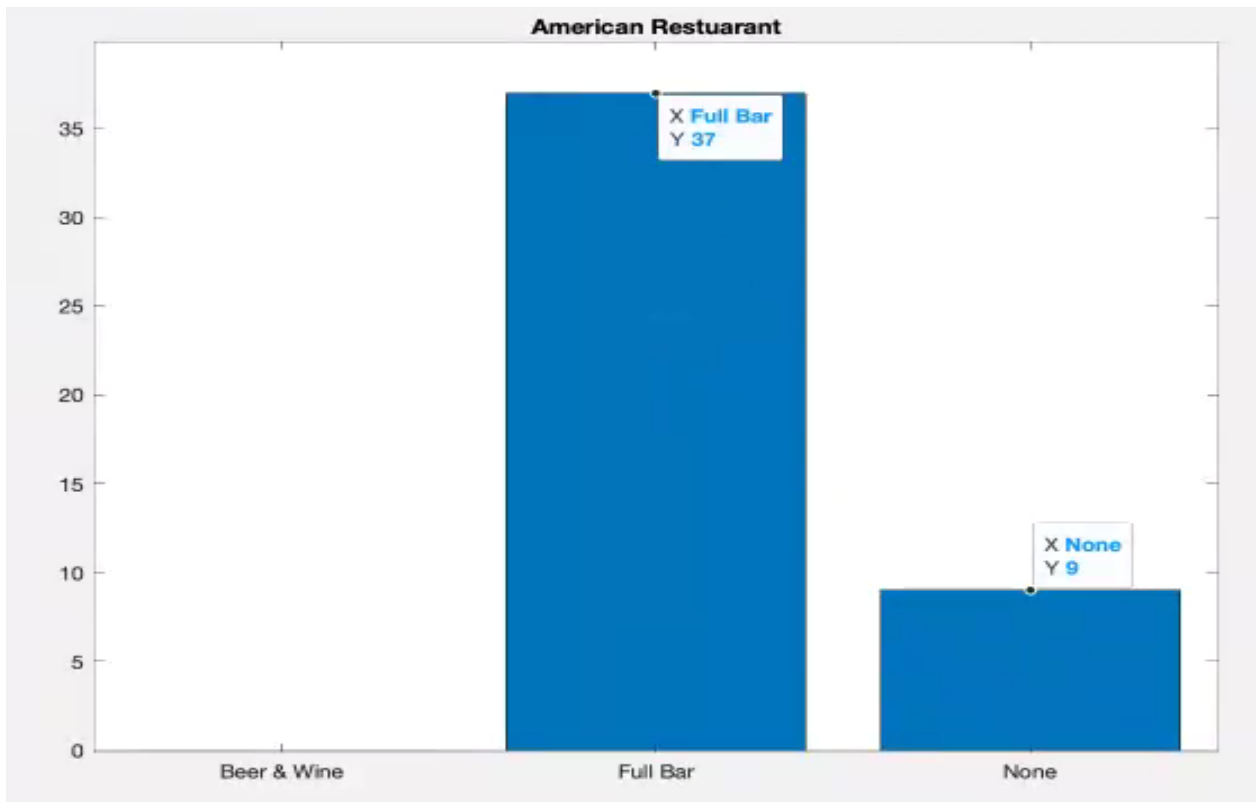
## 5. Alcohol

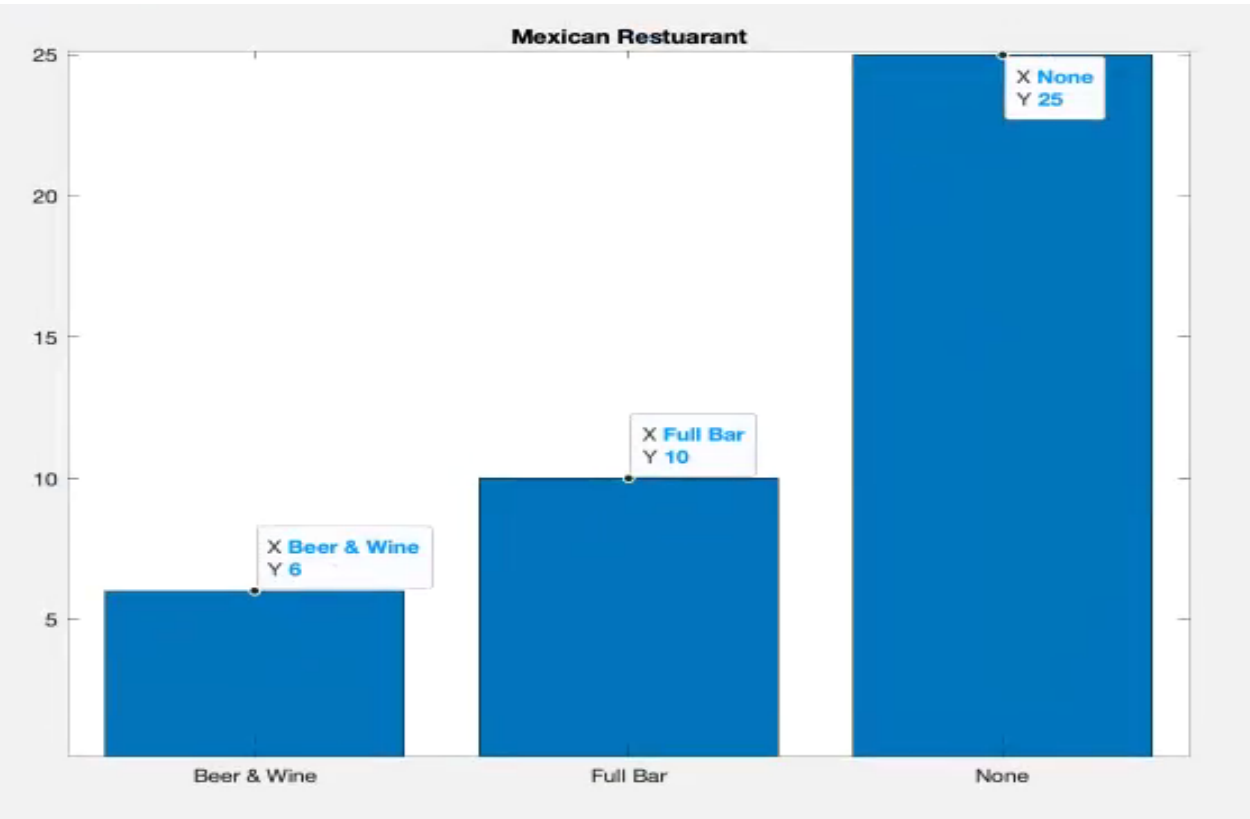
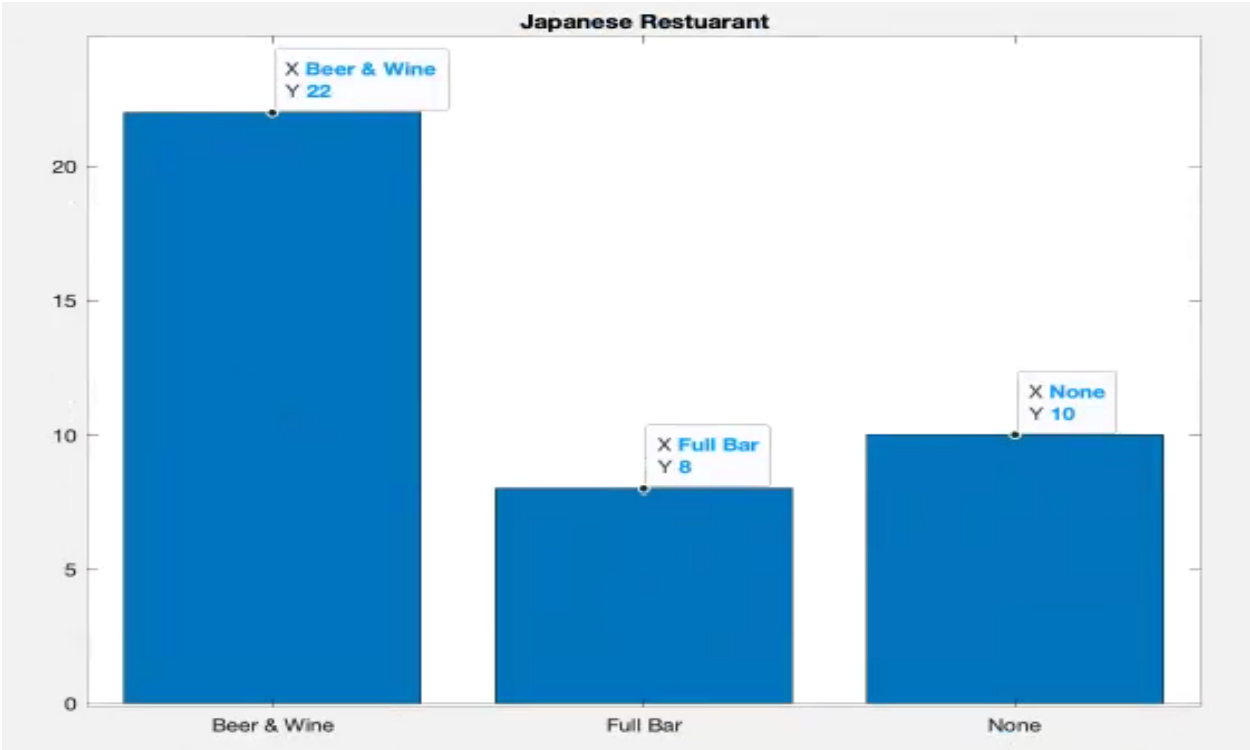
At this point there is not much difference in the attribute offered by the top starred restaurants.

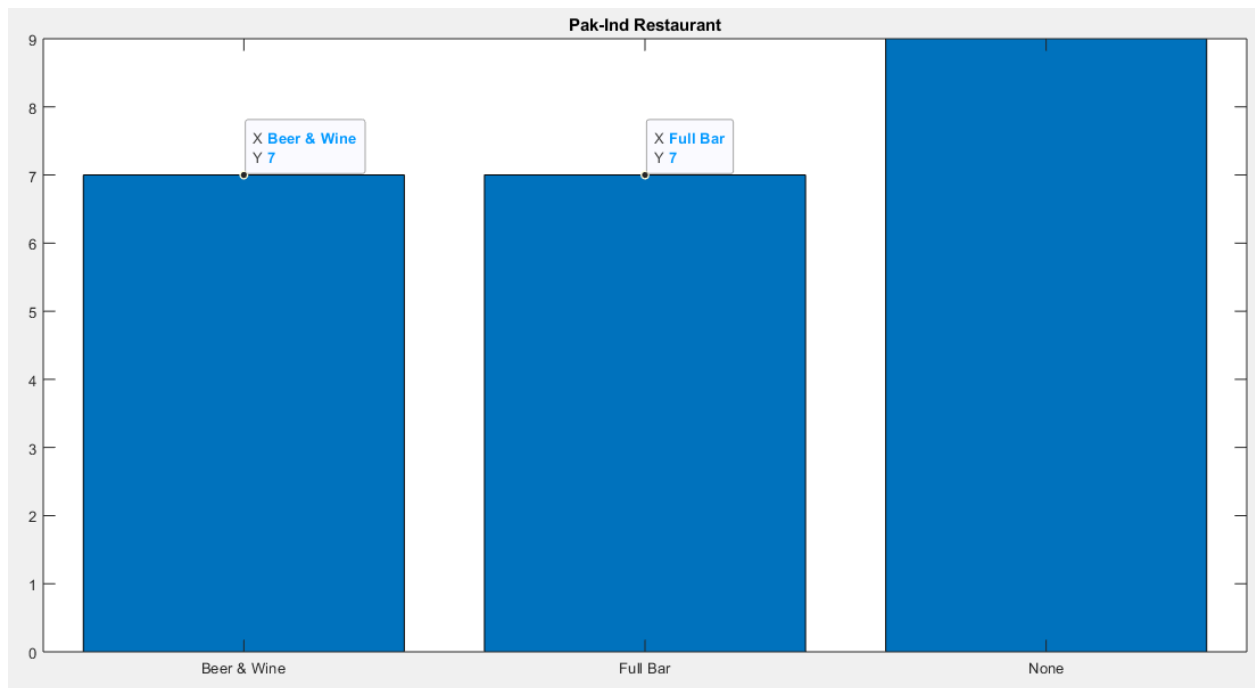
The top 50 restaurants when examined show that more than half have full bars. None of them have just beer & wine.

This alone gives us proof that a good restaurant should have a full bar and should not limit itself to a smaller selection like just offering beer & wine.

Below you'll find difference restaurant and what type of bars they have.







This is the code for getting Alcohol Data

```
% alcohol Method

function alcoresults = alcoholfind1(resturant)
bnw = 0;
none = 0;
full_bar = 0;

for i = 1:50
    if (char(resturant(i)) == "beer_and_wine")
        bnw = bnw + 1;
    end
    if (char(resturant(i)) == "full_bar")
        full_bar = full_bar + 1;
    end

    if (char(resturant(i)) == "none")
        none = none + 1;
    end
end

x = [full_bar,bnw,none]
names= {'Full Bar','Beer & Wine','None'};
y= categorical(names);
alcoresults = bar(y,x) end
```

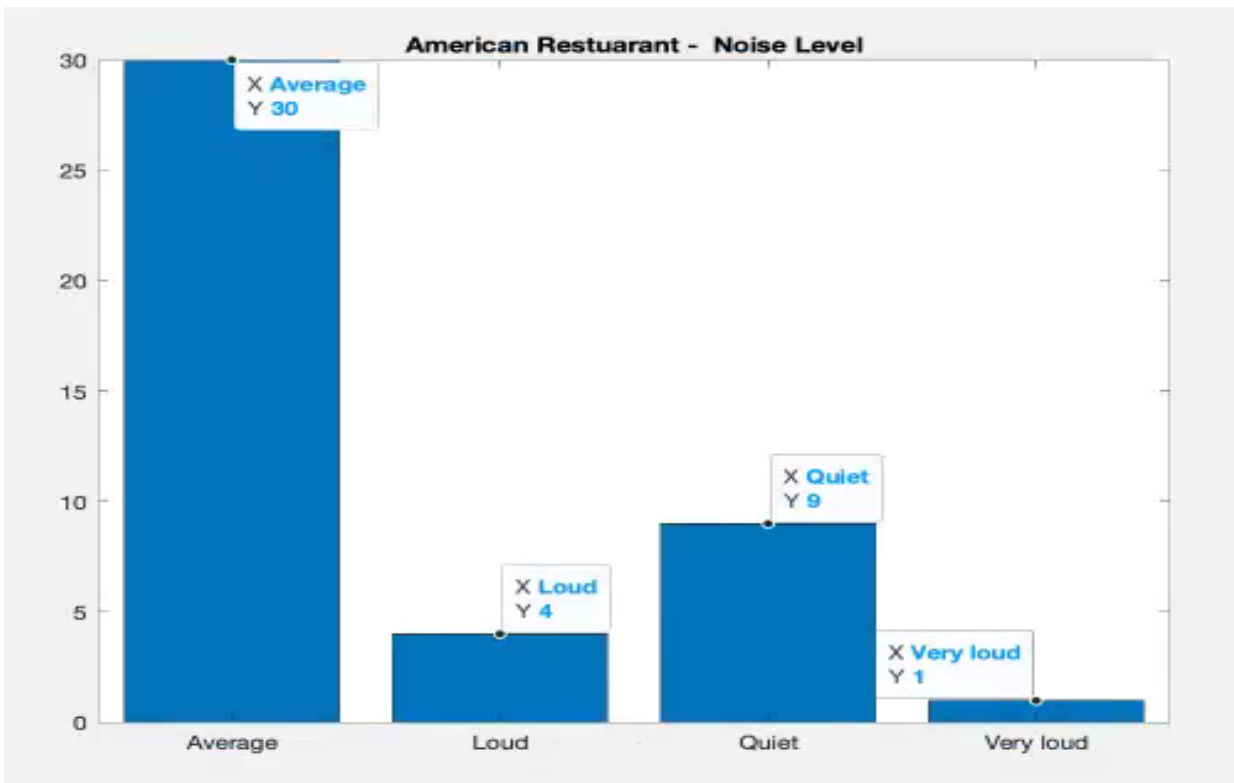
Below is the code for Plotting Alcohol Data:

```
% Plotting Comprehensive Alcohol
alcoholfind1(Italian_Top_Star.attributesAlcohol)
hold on;
alcoholfind1(Japanese_Top_Star.attributesAlcohol)
hold off;
hold on;
alcoholfind1(Mexican_Top_Star.attributesAlcohol)
hold off;
hold on;
alcoholfind(Pak_Ind_Top_Star.attributesAlcohol)
hold off;
hold on;
alcoholfind1(Chinese_Top_Star.attributesAlcohol)
hold off;
title('Comprehensive Alcohol Serving Restaurants')
legend({'Italian','Japanese','Mexican','Pak-Ind','Chinese'},'Location','northwest')
```

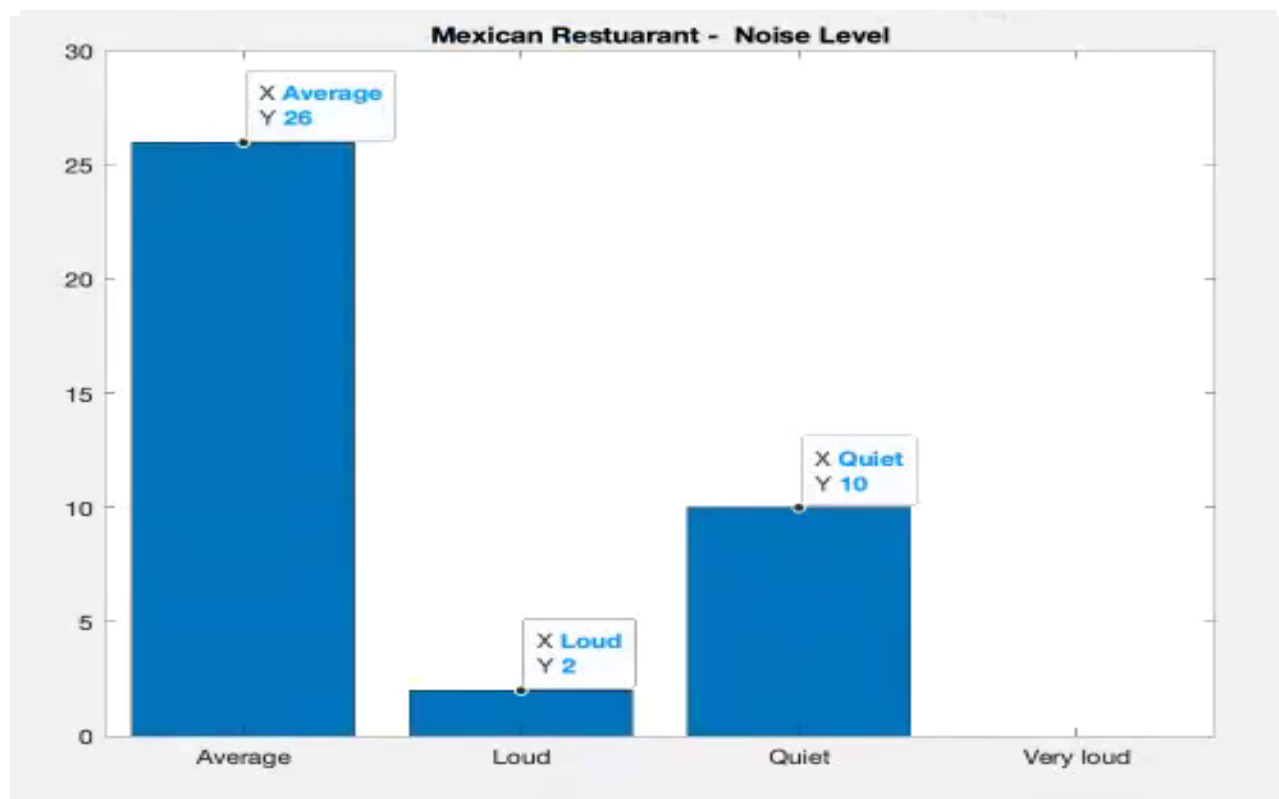
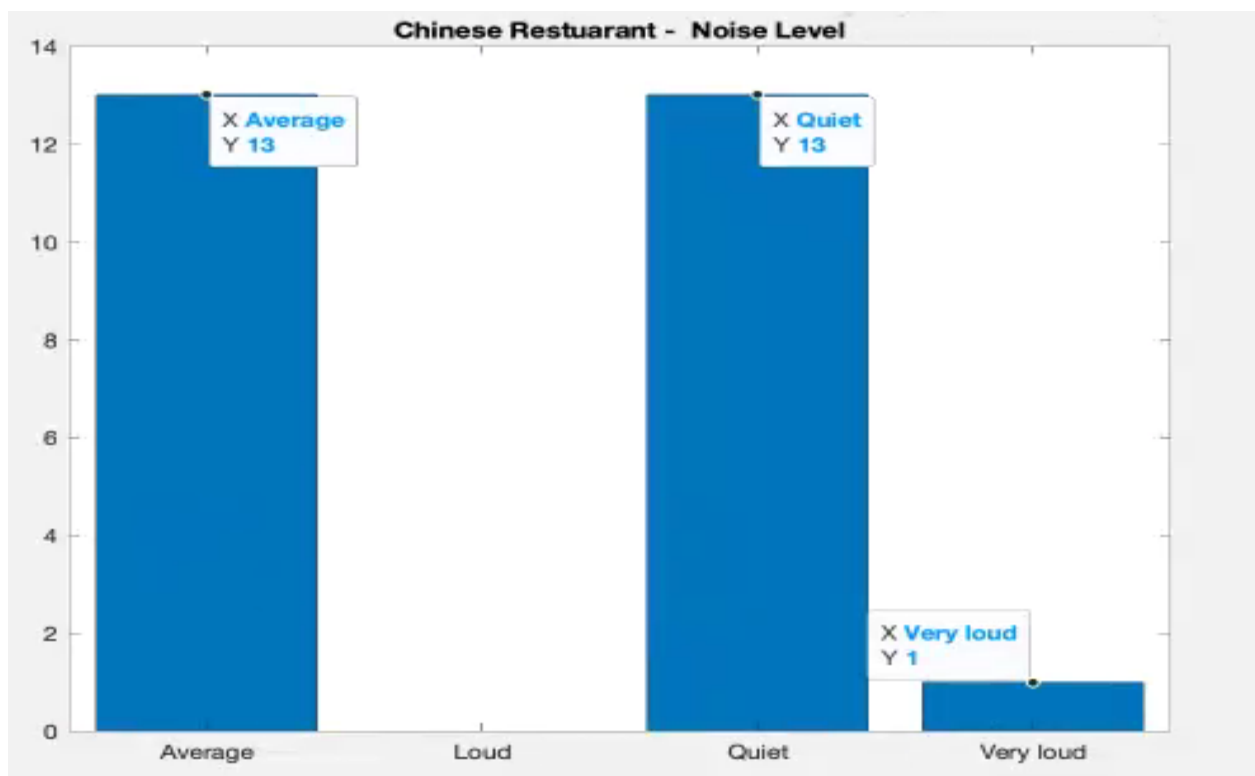
## 5. Noise Levels

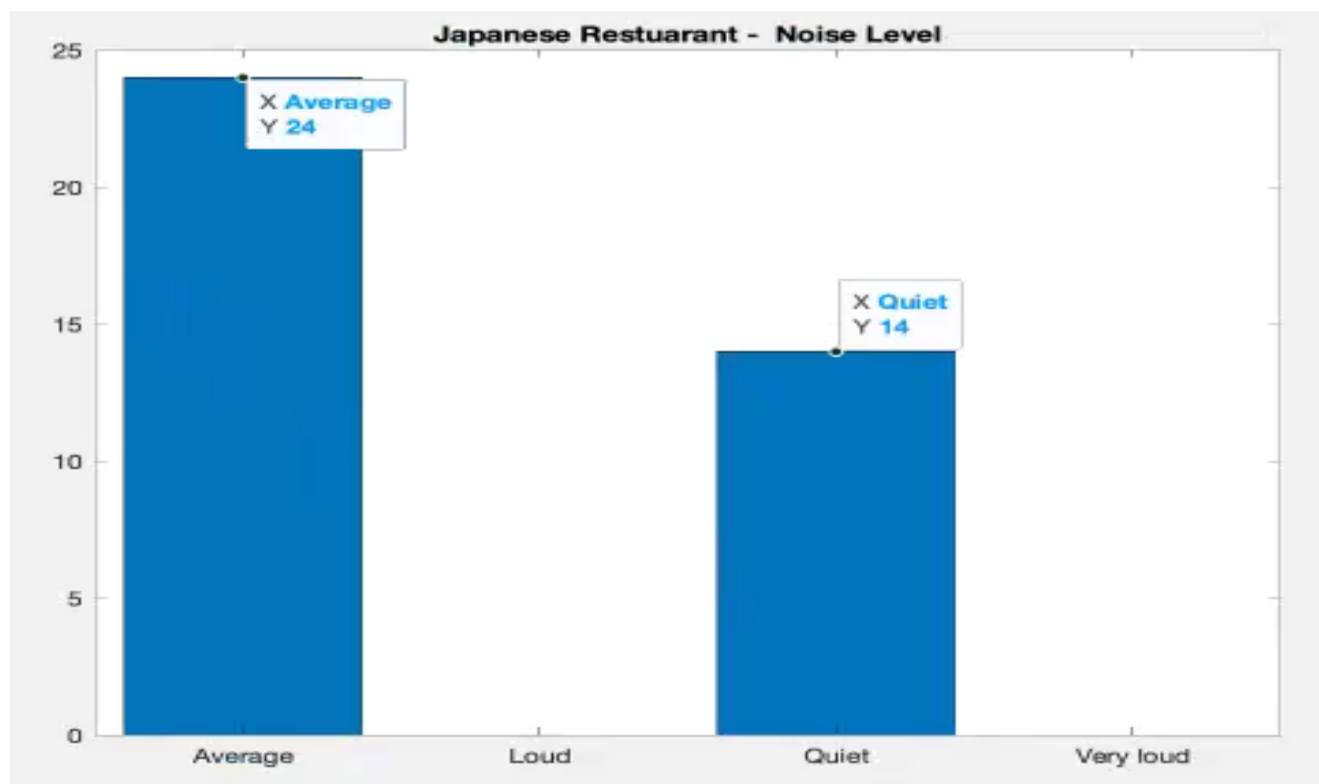
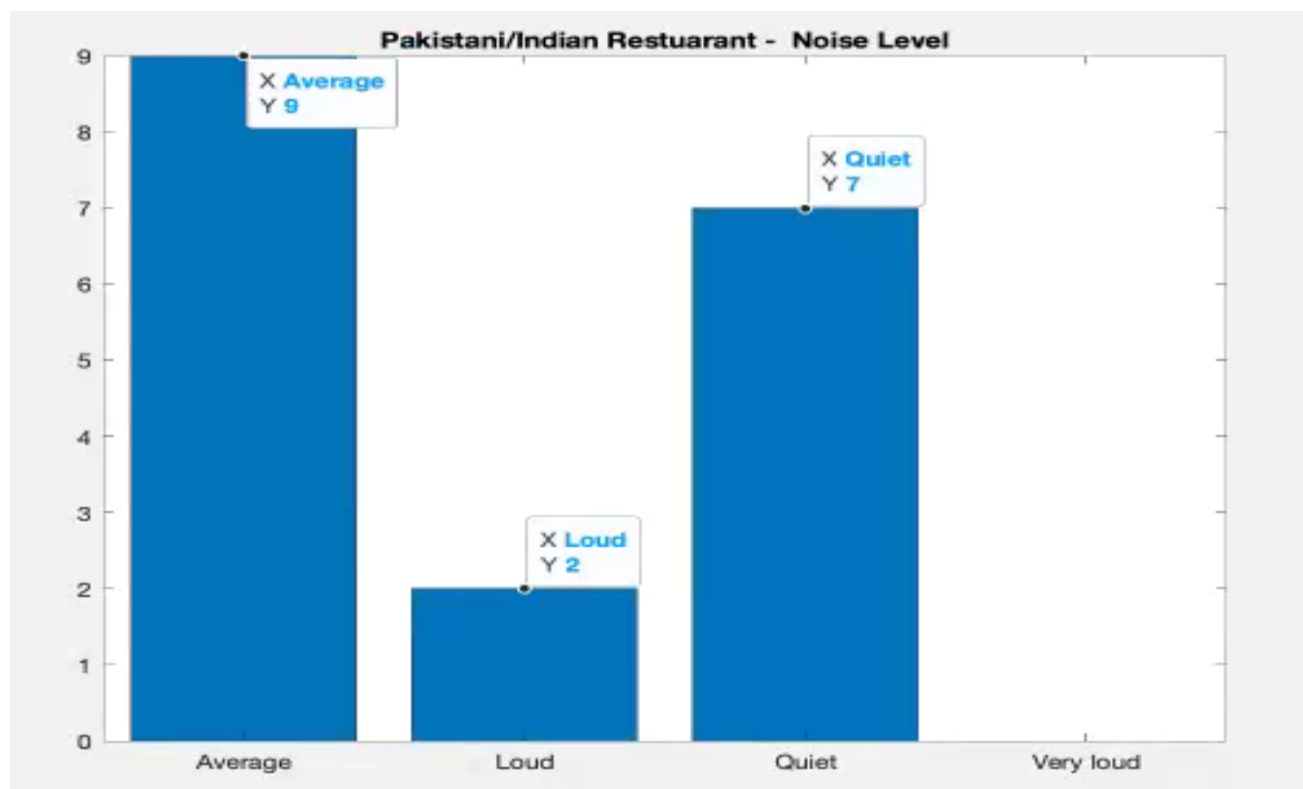
The same study was carried out for the Noise levels in restaurants.

Noise isn't something people admire, the graphs below show us what noise levels are maintained in different types of restaurants.

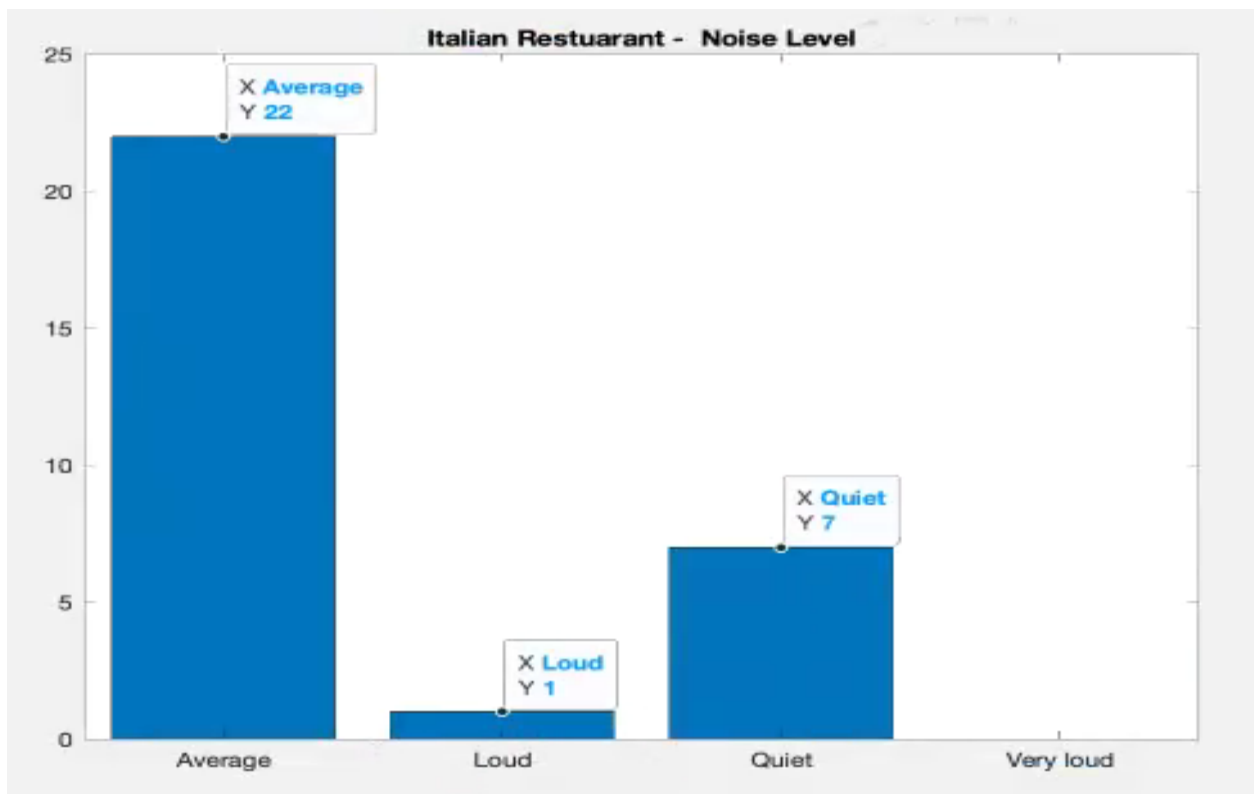












Code for Noise Levels:

```
function noise_plot = plotnoise4(noise)
avg = 0;
loud = 0 ;
quiet = 0 ;
vloud = 0;
for i = 1:size(noise)
    if (char(noise(i))=="loud")
        loud = loud + 1;
    end
    if (char(noise(i))=="quiet")
        quiet = quiet + 1;
    end
    if (char(noise(i))=="average")
        avg = avg + 1;
    end
    if (char(noise(i))=="very_loud")
        vloud = vloud + 1;
    end
end
x = [loud, quiet, avg, vloud]
names = {'Loud','Quiet','Average','Very loud'};
y=categorical(names);
y=reordercats(y,names);
```

```
plot(y,x)
end
```

### Plotting Comprehensive Noise

```
%% Comprehensive Plot
plotnoise4(Italian_Top_Star.attributesNoiseLevel)
hold on;
plotnoise4(Japanese_Top_Star.attributesNoiseLevel)
hold off;
hold on;
plotnoise4(Mexican_Top_Star.attributesNoiseLevel)
hold off;
hold on;
plotnoise4(Pak_Ind_Top_Star.attributesNoiseLevel)
hold off;
hold on;
plotnoise4(Chinese_Top_Star.attributesNoiseLevel)
hold off;
title('Comprehensive Noise Level')
legend({'Italian','Japanese','Mexican','Pak-Ind','Chinese'},'Location','northwest')
```

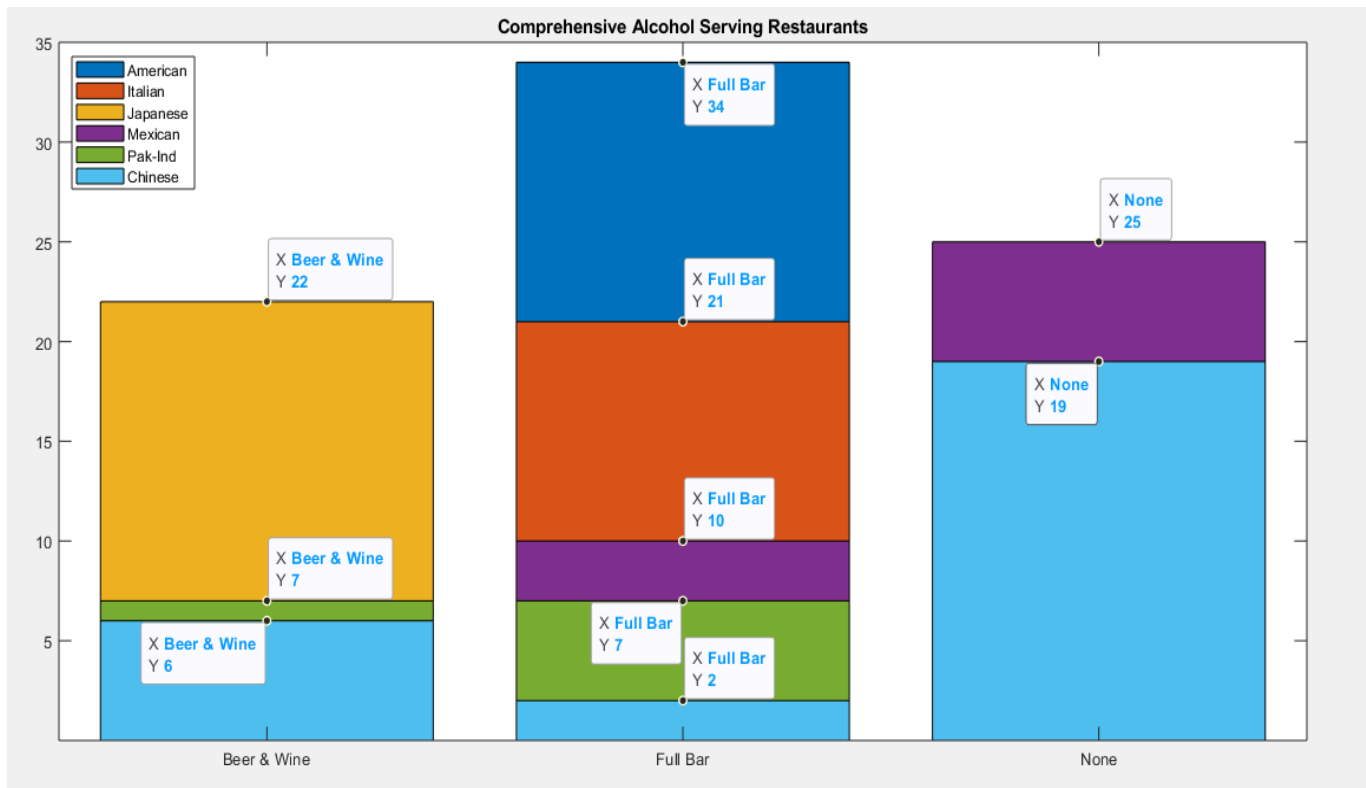
## Discussion :

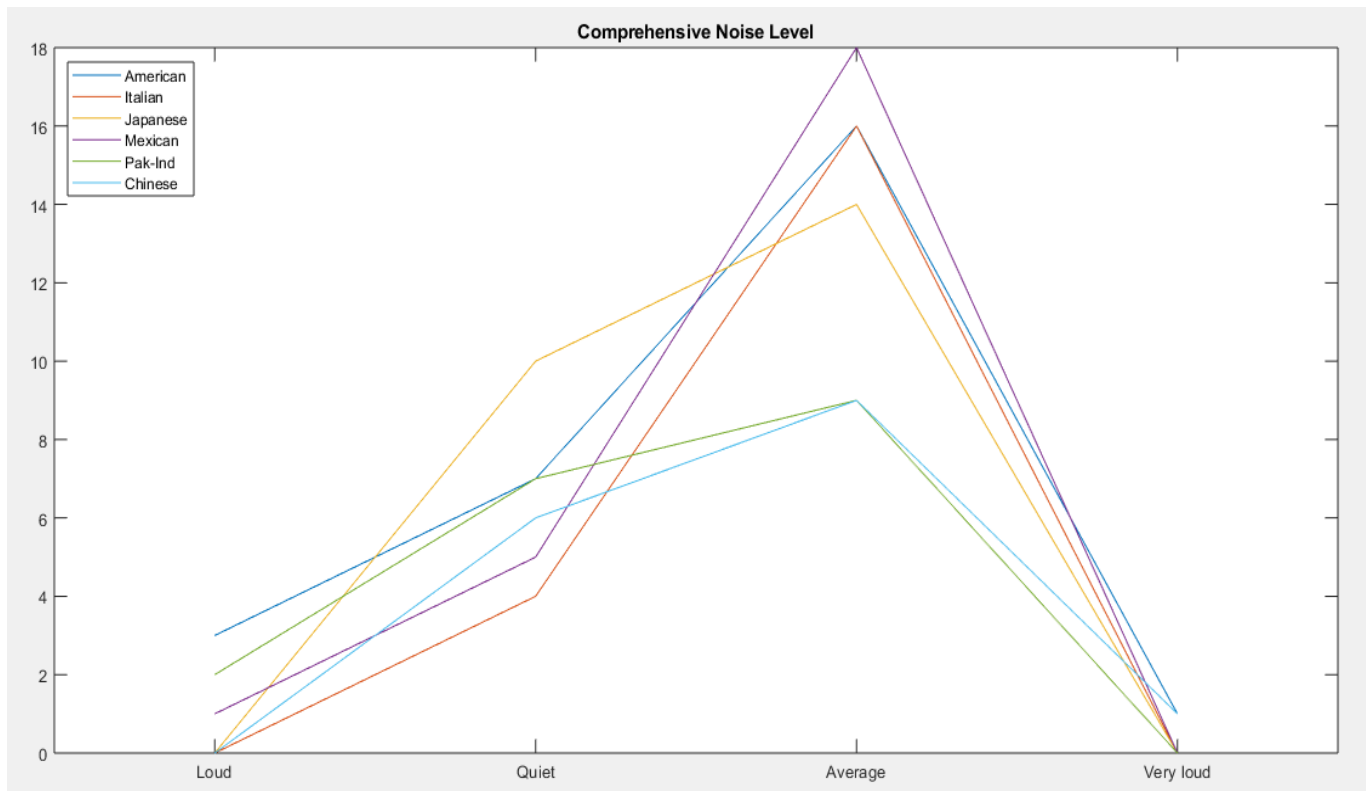
Based on our analysis, we found out that for most restaurant types, friendly ranks first before all the other positive comments, indicating that service might weigh more than taste when people are judging a restaurant.

In addition, different characteristics are shown for different restaurant categories. Such as alcohol and noise level in restaurants.

Although the performance of our model is decent, there are still a lot of space for improvement.

By comparing both these levels we can come to a conclusion.





## Conclusion

Since our analysis may help to extract specific features from any set of reviews, restaurant owners can make good use of it for essential information once they received a certain amount of Yelp reviews. From those reviews they can understand why customers love or dislike their restaurants, maybe great reviews primarily due to fresh food, or perhaps unsatisfied reviews caused by a too high price. Meanwhile they can also compare the restaurant with similar restaurants within the same type.

The noise levels in restaurants should be average or below, along with this alcohol should be served in the form of a full bar.