Impact of Food Types on Health

Shreyas Suhas Yeolekar
School of Computing
National College of Ireland
Dublin, Ireland
yeolekar.shreyas@gmail.com

Karan Mahendra Singh
School of Computing
National College of Ireland
Dublin, Ireland
karanmsingh10@gmail.com

Richard Saldanha
School of Computing
National College of Ireland
Dublin, Ireland
saldanharichard6@gmail.com

***Abstract*— This report describes the process and findings of a study about the impact of different types of food on people's health. Links between eating outdoor food and being ill or getting hospitalized due to it, were primarily sought. The project has dealt with three semi-structured datasets in pursuit of the primary aim. The dataset of food borne diseases was implied to check whether there was a relation between outdoor food and the rate of hospitalization due to illnesses. Visualization, along with a correlation matrix of the data was done to see if there was any correlation between the two. Thereafter, the alcohol consumption and adult mortality data retrieved from the GHO repository was used to analyse the effect of alcohol on the mortality rate in a particular region. The Correlation Coefficient was calculated to find the extent of the relationship between alcohol consumption and mortality rates. The results obtained revealed that alcohol had a nearly negligible negative correlation. Furthermore, the Zomato Restaurant data was analysed to compare the popularity of restaurants that serve alcohol with the restaurants that do not serve alcohol in a specific region.**

Keywords*— *food, disease, alcohol, restaurants, mortality

I. Introduction

Although food is a necessity, yet for many foodies it is also a hobby. People enjoy eating cross-culture cuisines and visiting fancy restaurants and as a result, an entire sophisticated industry known as the "restaurant and food service industry" has thrived over a couple of decades. Subsequently, numerous hygiene norms were introduced that concern about hygiene began to surface. However, according to what many individuals believe, despite all efforts, food prepared at homes rank higher in hygiene and therefore result into relatively fewer cases of illness than those caused due to outdoor food. As food became one of the top professions and people's taste buds demanded tastier food, innovative recipes were introduced to mark the beginning of a new era. Furthermore, as people's lives got busier in offices, leaving behind little time for food to be cooked at homes, easier options such as take away, home- delivery and street food became popular – so much that all above means became a part and parcel of their lives. Therefore, it was necessary to assess if the habit of exploring and consuming outdoor food was safe at all and thereby, we ourselves being foodies, felt connected to the topic. The project would thus settle once and for all, the debate on whether outdoor food really proves to be harmful to our health or it is fit for consumption as an easier option.

It is commonly seen that people share a bond of love with drinks all across the globe. However, there are similar number of individuals who do not possess a taste for drinking, while some even prefer not to dine at places where alcohol is served. As a result, there are fewer number of restaurants serving alcohol while others do not. Thus, we decided to find out whether there was any correlation between the ratings of restaurants received from the customers and whether they serve alcohol or not. We tried to find out if restaurants received lower ratings if they did not have alcohol on their menu. In addition to this, we also related the alcohol consumption rates with the mortality rates in a particular region using the data acquired from GHO Data Repository.

II. Literature Review

Shina, S. Sharma, and A. Singla[7], in their research intend to classify the restaurants in India into different categories which are based on the online reviews of the customers and the services they have received from the different restaurants that are registered with Zomato using the two machine learning algorithms which are Decision Tree and Random Forest. The results of the analysis proved that Decision Tree Classifier gave good performance results with an effective accuracy rate of around 63.5% as compared to Random Forest which gave an accuracy result of only 56%. A. Taneja, P. Gupta, A. Garg, A. Bansal, K. P. Grewal and A. Arora [8], in their research, focused on building a recommendation system to understand, analyze, visualize and suggest locations as well as restaurants based on user behavior. In this analysis the authors have used datasets of Facebook check-in and Zomato location-wise restaurant review and predominantly used PHP scripts and Neo4J to create interactive graphs of different cities' restaurant rating comparison among defined features-based rating and Zomato rating. A. Pisutaporn, B. Chonvirachkul and D.Sutivong[9], have done a very fruitful research on the impact of student's grades in the exam due to consumption of alcohol using Machine Learning Classification models such as decision tree algorithm and random forest algorithm. The observation suggested that the percentage of male students consuming alcohol is more than female students and it is seen that there exists an inverse correlation between the grades obtained by students in their exams and alcohol consumption i.e. both the variables under observation are negatively correlated also the performance of random forest is better than the decision tree algorithm.

III. Methodology

Four Datasets were used in this analysis. For more than 1,5 million restaurants across 10,000cities around the world, Zomato APIs provide you with access to the freshest and most accessible information. The Zomato data[1] allows you to search by name, cuisine, or place for restaurants. Nearby Restaurants Display comprehensive data including ratings, location and cuisine. The dataset for Food Borne Diseases[2] emphasizes on the diseases spread through consumption of food being contaminated covering a time period from 1998 to 2015 which is reported to CDC (Centers

for Disease Control and Prevention) and where was the observation found like restaurants, catering services and many other locations. The reason we have taken this dataset is to understand the impact on the health of the people through the consumption of food at different locations. The volume of the dataset is around 19000 records with data fields consisting from the period year and month, to the Venue like restaurants, Banquet Facility, catering services and many other locations. The WHO Global Drug and Health Information System (GISAH) [3][5] offers easy and quick access to a wide range of health indicators related to alcohol. It is an essential tool in countries to assess and monitor the health situation and trends in alcohol consumption, alcohol-related damage, and policy responses.

The data description table for each of the datasets is given below.

Databases	Variables	Data Types
Food Borne Diseases	Food Location	String
	Illness	Number(int)
	Hospitalization	Number(int)
Alcohol and Mortality Dataset	Country	String
	Average Alcohol Consumption	Rate(int)
	Mortality Rate	Rate(int)
Zomato	Location	String
	Votes	Number(int)
	Alcohol	Number (Categorical)

Fig 1: Data Description Table

The workflow for each of the dataset can be given below.

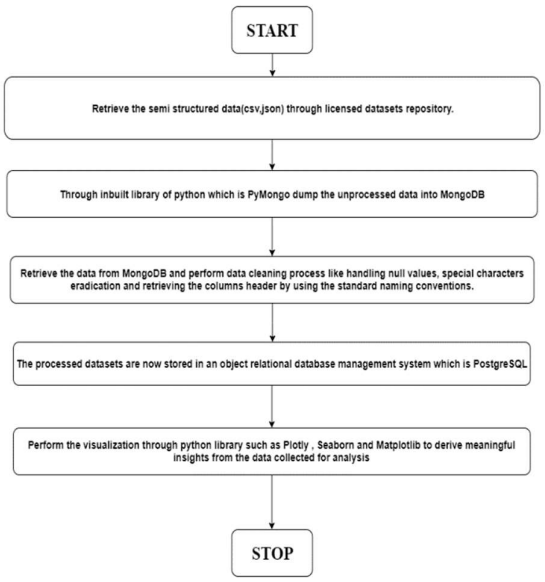

Fig 2: Process flow Diagram

Choice of Technologies:

Python: It has tools such as Pandas, NumPy, SciPy and so on which help in easy data manipulation and analysis.

MongoDB: This platform is used for dealing with documents with schema such as NoSQL and JSON.

PostgreSQL: Is a write-ahead logging database which ensures atomicity and durability in databases. Further, it has parallel read queries, thereby supporting concurrency.

A. Food Born Diseases Data

This dataset contained columns such as number of Illnesses, number of hospitalizations, fatalities, locations at which food was consumed and so on which coincided directly with one of the objectives of the finding which was confirming if home-made food or outside food is safe for consumption.

Data Gathering: This data [2] was in csv format. It was loaded in Python and converted to json for saving into MongoDB as it is used to store semi-structured data

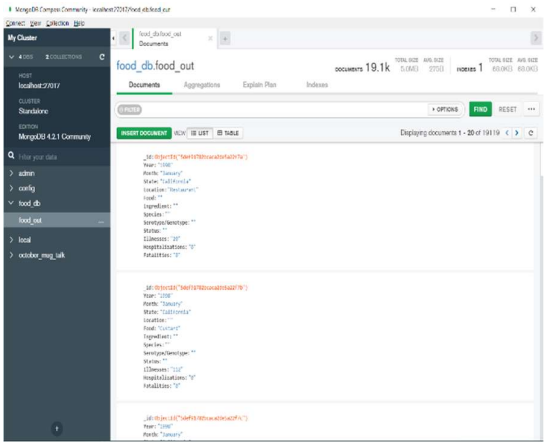

Fig 3: Food Borne Diseases Data

Data Cleaning: The data contained a number of missing values which were discarded for smooth analysis. Thereafter, few unwanted columns were removed for simplicity.

Data Storage: After cleaning it, the data was stored on PostgreSQL from Python, for later use.

Data Analysis: The cleaned data was analyzed and visualized to find trends in the data and create visualizations to depict what the data proves and meet the research objective.

B. *Alcohol and Mortality Data*

Data Gathering: - The data [3][5] was retrieved from the GHO repository in a semi structured format. The data was then stored as a MongoDB collection.

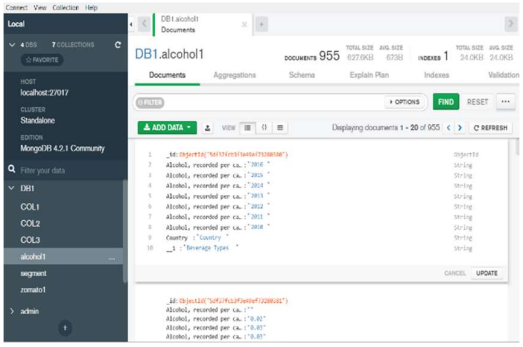

Fig 4: Alcohol Consumption and Mortality data

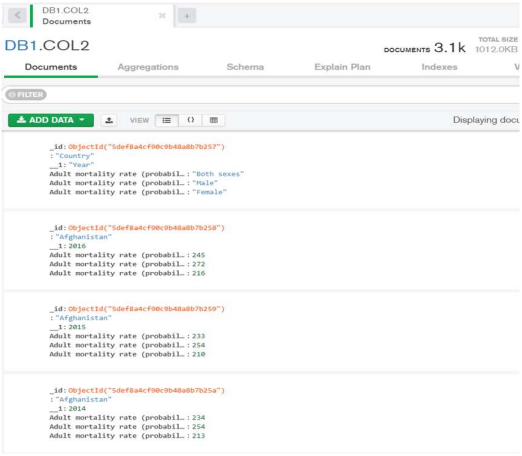

Fig 5: Adult Mortality Data

Data Cleaning: - The dataset was then imported from MongoDB as a python data frame. The dataset for alcohol consumption and mortality rates had missing values. The null values were dropped from the final data frame. The data types of the variables were checked for and transformed if required to avoid any errors in the analysis.

Data Processing and Transformation: - The alcohol and mortality dataset were joined and grouped by country. The data was transformed such that the result data frame aggregated the average values of each year to total average value.

Data Storage: - The data was stored in a PostgreSQL database for further use using Python.

Data Analysis: - The data was visualized using various python libraries plotly and seaborn. The Pearson's Correlation Coefficient for the two variables was calculated. The Mathematical equation for the Pearson Correlation Coefficient is:

$$r = \frac{\sum(x - \bar{x})(y - \overline{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}}$$

where,

$r = measure\ of\ linear\ correlation$

$between\ two\ variables\ x\ and\ y,$

$\bar{x}\ and\ \bar{y}\ \ are\ the\ mean\ values$

C. *Zomato Data*

Data Gathering: - The data set was obtained from the Zomato API provided by the developer website

[]. The data was then converted to JSON format. The JSON data was then parsed to acquire the essential attributes from the semi structured data. The json data was flattened and then stored as a MongoDB collection.

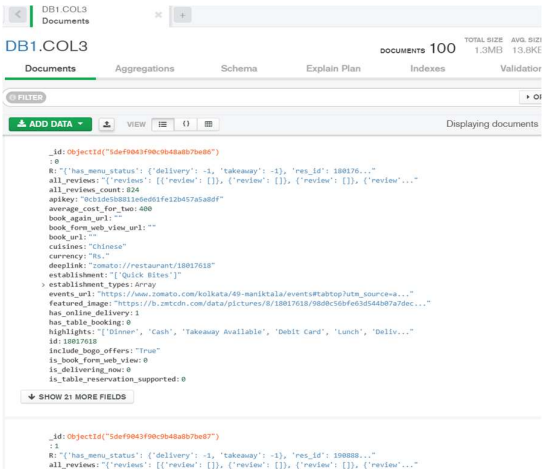

Fig 6: Zomato Data

Data Cleaning: - After importing the data set from MongoDB it was converted to a data frame and tested for null values. If found, the null values were dropped.

Data Processing and Transformation: - The Zomato data had a country code attribute. The location attributes had multiple values such as country id, city, region. The country id was the only string required from the column. To achieve this the string was split into multiple columns. The location attribute had multiple string values. The values were split into multiple columns and only the country_id value was joined to the data frame. The Country names were mapped to the country codes with the help of Country Codes data. The highlights column was split into multiple columns to check if the restaurant was serving alcohol. The data was categorized into two binary values 0 and 1, where, 0 means the restaurant doesn't serve alcohol and 1 means the restaurant serves alcohol. The data frame was then filtered and the only columns that were included in the final data frame.

Data Storage: The data was then stored in the PostgreSQL for further use.

Data Analysis: - The data was then analyzed to find if restaurants serving alcohol are more popular than restaurants that don't. The libraries pandas were used to plot a grouped bar chart.

IV. Results

A. *Food Borne Diseases*

The aim of the project was to relate the source of food consumed by people and the number of cases of illnesses of hospitalizations. Thus after relating the above said variables, the following results were achieved:

1. Illnesses caused by eating in locations
   As is apparent from the graph, those who ate food from Catering, suffered the greatest number of illnesses whereas, food from Child Daycares saw the least amount of illnesses. Interestingly, even those who ate in Private Homes suffered a lot of illnesses, higher than those seen after eating in regular Restaurants and Fast-food Restaurants.

Fig 7: Illnesses by eating in Locations

2. Hospitalizations by eating in places
   According to the plot, those who ate in Banquets as well as Private Homes showed the highest amounts of hospitalizations. Closely followed by Fastfood Restaurants and Catering. Thereafter, Child Daycare showed least amounts of hospitalizations.

Fig 8: Hospitalizations by eating in location

	Home	Banquet	Catering	Child_Daycare	Fast_Food_Restaurant	Illnesses
Home	1.000000	-0.027388	-0.030997	0.000023	-0.361956	-0.018422
Banquet	-0.027388	1.000000	-0.013245	-0.006747	-0.162269	0.062183
Catering	-0.030997	-0.013245	1.000000	-0.007360	-0.288089	0.085479
Child_Daycare	0.000023	-0.006747	-0.007360	1.000000	-0.042883	0.002439
Fast_Food_Restaurant	-0.361956	-0.162269	-0.288089	-0.042883	1.000000	-0.138966
Illness	-0.018422	0.062183	0.085479	0.002439	-0.138966	1.000000

Fig 9: Correlation Matrix

B. *Adult mortality and alcohol consumption*

The objective pertaining to this part was to find whether the alcohol consumption rates in various countries equated with mortality rates in them.

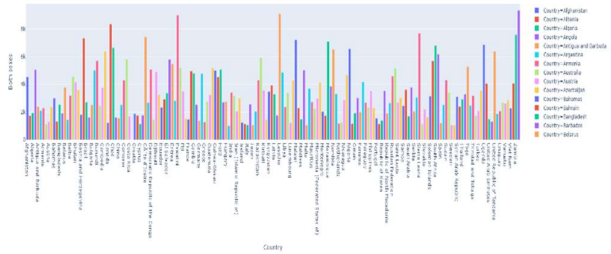

Fig 10: Deaths by country

Fig 10 shows average deaths of years 2000-2016 of both sexes by Countries.

C. *Adult Mortality and Alcohol Consumption*

The correlation coefficient and the trendline for the adult mortality and alcohol consumption can be seen in Fig 11.

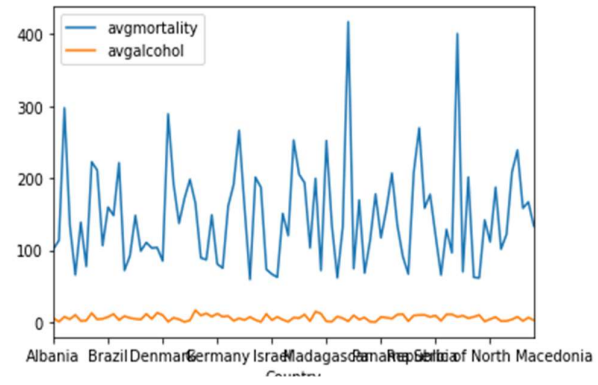

Fig 11: Adult Mortality and Alcohol Consumption

	avgalcohol	avgmortality
avgalcohol	1	-0.164873
avgmortality	-0.164873	1

Fig 12: Adult Mortality Data

The correlation between alcohol consumption and average mortality is 16.4% and is a negative correlation. This correlation can be almost deemed as insignificant.

D. *Zomato Dataset*

The Zomato dataset was used to compare the popularity of restaurants serving alcohol with restaurants that do not serve alcohol. The analysis produced the results as shown in Fig 12. The bar chart shows that in most of the countries people voted more for restaurants that served alcohol, except for Australia, United States, South Africa and Sri Lanka. But even those countries there wasn't much of a difference in the upvotes to consider it to be significant.

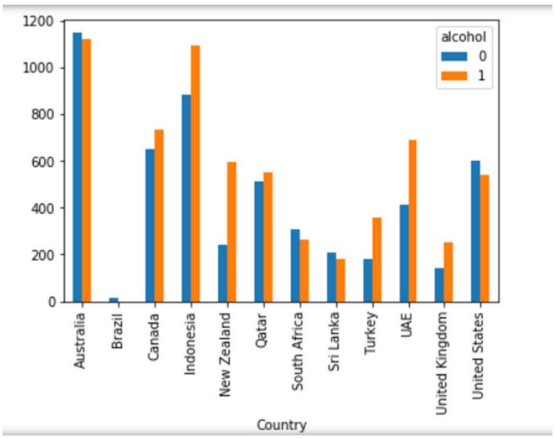

Fig 12: Restaurants Serving Alcohol

V. Conclusion and Future Scope

Thus, it can be said that eating outdoor food is a risky job and that in spite of maintaining hygiene standards, there could still be some flaws as many individuals fell sick after eating food from catering. However, illnesses and hospitalizations seen in Private Homes are around equally high and it cannot be said that home cooked food is safe to consume or perhaps there are more factors working to make people sick after consuming food. Also, from the comparison between alcohol consumption and mortality rate in a specific region, it was observed that the former had little or no effect on the later. Thus, is it safe to say that alcohol has no effect on the health of an individual? The answer to this question is 'no', as there are various other factors to be considered. The Zomato dataset produced results which indicate that people in most countries prefer a restaurant that serves alcohol. Even though this does not assume that each person upvoting the restaurant drinks alcohol, yet we can conclude that alcohol serving restaurants are more popular in most of the countries across the globe.

In the future prospect of the project, we can consider other factors such as type of an alcoholic beverage to analyze if a specific type of an alcohol has any effect on health or mortality rate. Moreover, we could predict if the high ratings

of the restaurant vary on food quality, ambience or price range of the restaurant.

REFERENCES

[1] "Zomato API", *Zomato*, 2019. [Online]. Available: https://developers.zomato.com/api. [Accessed: 10- Dec- 2019].

[2] "Foodborne Disease Outbreaks, 1998-2015", *Kaggle.com*, 2019. [Online]. Available: https://www.kaggle.com/cdc/foodborne-diseases. [Accessed: 10- Dec- 2019].

[3] "GHO | By category | Adult mortality - Data by country", *Apps.who.int*, 2019. [Online]. Available: http://apps.who.int/gho/data/view.main.1360?lang=en. [Accessed: 10- Dec- 2019].

[4] "Indicator Metadata Registry Details", *Who.int*, 2019. [Online]. Available: https://www.who.int/data/gho/indicator-metadata-registry/imr-details/64. [Accessed: 10- Dec- 2019].

[5] "GHO | By category | Recorded alcohol per capita consumption, from 2010 - Updated May 2018", *Apps.who.int*, 2019. [Online]. Available: http://apps.who.int/gho/data/node.main.A1039?lang=en. [Accessed: 10- Dec- 2019].

[6] "Indicator Metadata Registry Details", *Who.int*, 2019. [Online]. Available: https://www.who.int/data/gho/indicator-metadata-registry/imr-details/462. [Accessed: 10- Dec- 2019].

[7] Shina, S. Sharma and A. Singla, "A Study of Tree Based Machine Learning Techniques for Restaurant Reviews," *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, Greater Noida, India, 2018, pp. 1-4. doi: 10.1109/CCAA.2018.8777649.

[8] A. Taneja, P. Gupta, A. Garg, A. Bansal, K. P. Grewal and A. Arora, "Social graph based location recommendation using users' behavior: By locating the best route and dining in best restaurant," *2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC)*, Waknaghat, 2016, pp. 488-494. doi: 10.1109/PDGC.2016.7913244.

[9] A. Pisutaporn, B. Chonvirachkul and D. Sutivong, "Relevant factors and classification of student alcohol consumption," *2018 IEEE International Conference on Innovative Research and Development (ICIRD)*, Bangkok, 2018, pp. 1-6. doi: 10.1109/ICIRD.2018.8376297