RESTAURANT AND HEALTH

Group-1 BI- ISQS 6339

Team Members:

Saranya Murugan

Madhu Atyam

Introduction

In today’s world everyone needs things with ease and quick. So, most of the people consider cooking as waste of time in kitchen rather spend some few dollars and get the tasty food so that they can utilize that time for other activities. Restaurant helps lot of people such as office crowd, school crowd and people who doesn’t know cooking! But we need to understand that everything is at the cost of our health.

I. Analysis of data:

In this project we are going to analyze the restaurant and health dataset to infer the below

- Type of cuisine people in each state likes
- Health issues of people in that state which could be due to food habits
- Which state tops high in health issues such as High cholesterol and diabetes
- Does people like US chain restaurants

In these datasets, we have data items such as restaurant name, city, state, stars, level of the health issues such as high cholesterol, obesity, diabetes, cancer, asthma which would be used in analysis

Direct Business Application:

- Based on the number of restaurants for the cuisine type in each state and city we can understand favorite food type of people in each city/state and provide consultation to set up restaurant or we can start our own restaurant. However, from the findings it is understood that few states had really high proportion of obesity and Cholesterol issues. These issues could be a possible chance of fast-food type of restaurants. So, we can suggest or using healthy food habits like “non gmo”, organic and less fat based products an alternative to mayonnaise sauces and etc., which are mostly found in all kind of foods in startup restaurant business. Products and preparing healthy food in this restaurant knowing the fact and statistics of health issues which could also be used for business marketing.

- Based on this analysis we could provide suggestions to hospitals in the city/state to have more equipped for the health issue wide spread in that city/state.

- We can provide consultation for student by encouraging them to study the particular health issue or research on it which could provide more opportunity in career.

- Doctors can specialize on this and invent preventive measures for the future generation.

Indirect Business Applications:

- This analysis can be useful for fitness centers to advertise and design their exercises accordingly.

- This analysis would be useful for grocery shop and supermarkets as they would know what people prefer more and what is good for people in the city.

Potential Usages:

- The few variables in the analysis which are not used are correlated with other variable, like population count and a data value of percent of health issues. To understand the use of health issues data value, population count is necessary. Other variables are directly used for analysis.

II. Data Cleaning:

- The overall quality of the datasets is above average, we have null values and misrepresented data in few columns.

- **Business dataset**: This dataset has missing values in columns such as address, attributes, city, hours and postal code. As part of data cleaning we considered the observations which has city and saved it into the data frame, then filtered the observations which contain the word "restaurant" in the categories column. We had few na's in the attributes column which was replaced with "other".

- **500_Health dataset:** In this dataset, however there are few missing data for few columns; the dataset is filtered initially with city without NULL values. Variables like 'StateAbbr', 'CityName', 'MeasureId' (duplicates), these duplicates are removed from the data frame. For GeographicLevel' (mismatch data) the data supposed to be city, but there was some mismatch in the column. The whole column data is replaced with variable "city".

- **500_City Level dataset:** There is a missing data for Data value column, these values are associated with Data Value Type. We replaced the missing data with median value of the specific data value by grouping the Data Value Type variable.

- **Bi project file dataset:** Before loading this dataset to csv file, with the final merged dataset, we dropped few columns which are repetitive and also which are not required for the analysis. The variables are can be identified in the code file submitted along with this report.

III. Data Merging:

- As there are 3 datasets in this analysis, initial data merging was performed with 500_Health dataset and 500_City level dataset using inner join with City FIPS of the health dataset and Place FIPS of City level dataset. These are the common unique keys to merge the datasets. These two datasets are also sourced from the same website.

- Second data merging was performed with first merged dataset and business dataset from yelp website. The common elements between the two datasets is the State and City. So inner join is performed between the two datasets to complete the merging process.
- Multilevel measurement issue occurs as we are joining the city and state which is not unique in all the datasets, but it is always better to secure the data without losing most of the details as data is important and useful for analysis. In this project we have used city and state which is repetitive but gives meaning when combined with review and other health related records.
- Variables that are more valuable after combined from the datasets are:

  I. **Business**: - State, City, Name (restaurant name) and Categories (restaurant categories)

  II. **Health**: - Year, Category (health category), Data Value Unit, Data Value, Population Count, MeasuredId (different type of health issues), Geolocation, State and City

  III. **City level data:** - State, Place name (City), Arthritis, BPHigh, BPMed, Cancer, Asthma, Diabetes, High Cholesterol and Obesity.

  IV. Variables such as High cholesterol level, obesity level and other health issues doesn't provide the meaning unless combined with the type of cuisine or food they consume (eg: Pizza, Burger)

- Data becomes more valuable after merging as it could be used for the below:

  I. After merging three datasets related to restaurants, types of health issues data and category of health issues data which made a valuable insight to understand the type of restaurants could impact the what type of health issues. E.g., a fast food restaurant could impact a person for his unhealthy habits and causes Obesity.

  II. Provide consultation to hospital as in which area doctors should be more specialized according to the state/region

  III. Business consultation for startup restaurants, fitness center and supermarket

  IV. Educational consultation for students as this would be one of the career opportunities and also to prevent the health issues in their region

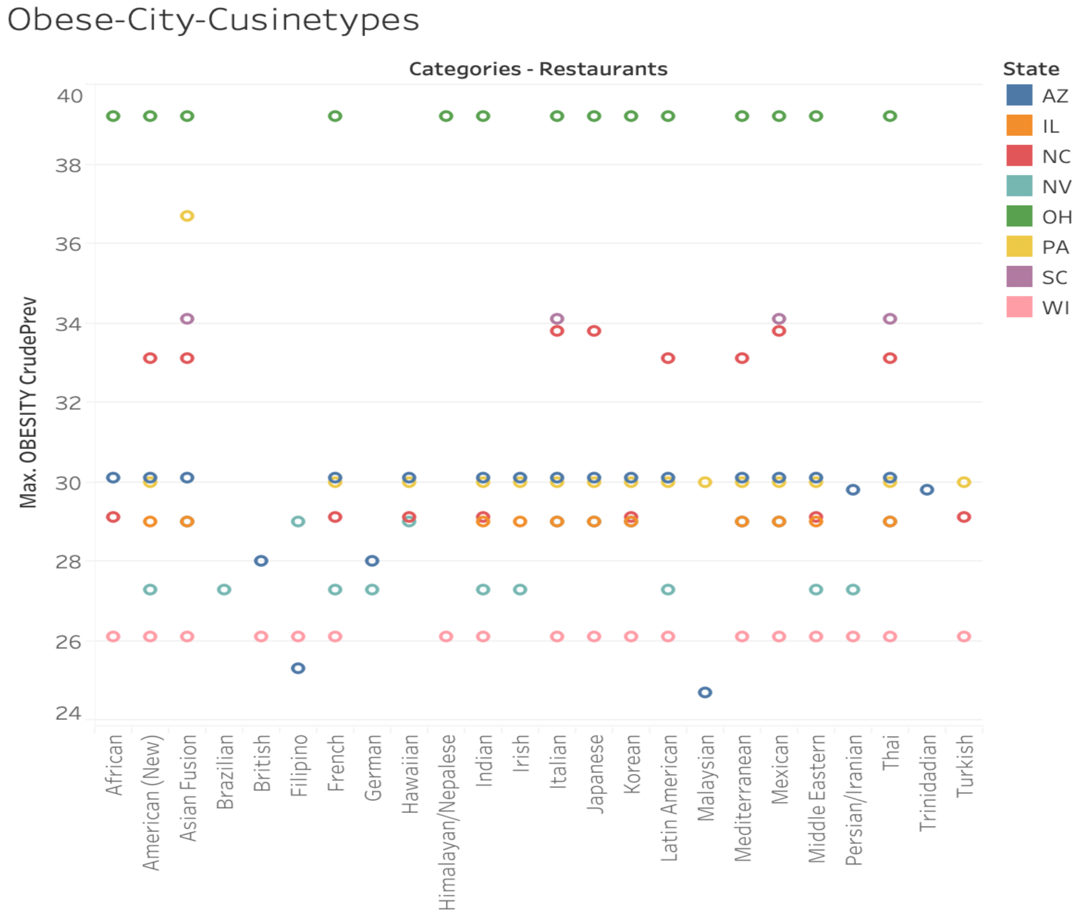
  V. Non-GMO and organic products can sell more in these regions as people would be prone to buy due to the statistics of health issues

- To find these meaningful insights it is necessary to merge different datasets to bring value to the table.

IV. Analysis of Visualizations

- In the visualizations, the similar items are grouped together, and dissimilar items are not grouped. The other characteristics of visualizations Ordered perception and Quantitative Perception are maintained while performing the visualizations. This analysis doesn't contain any sort of visualizations related to natural processing.

Figure-1 State with Obese index and type of Cuisine

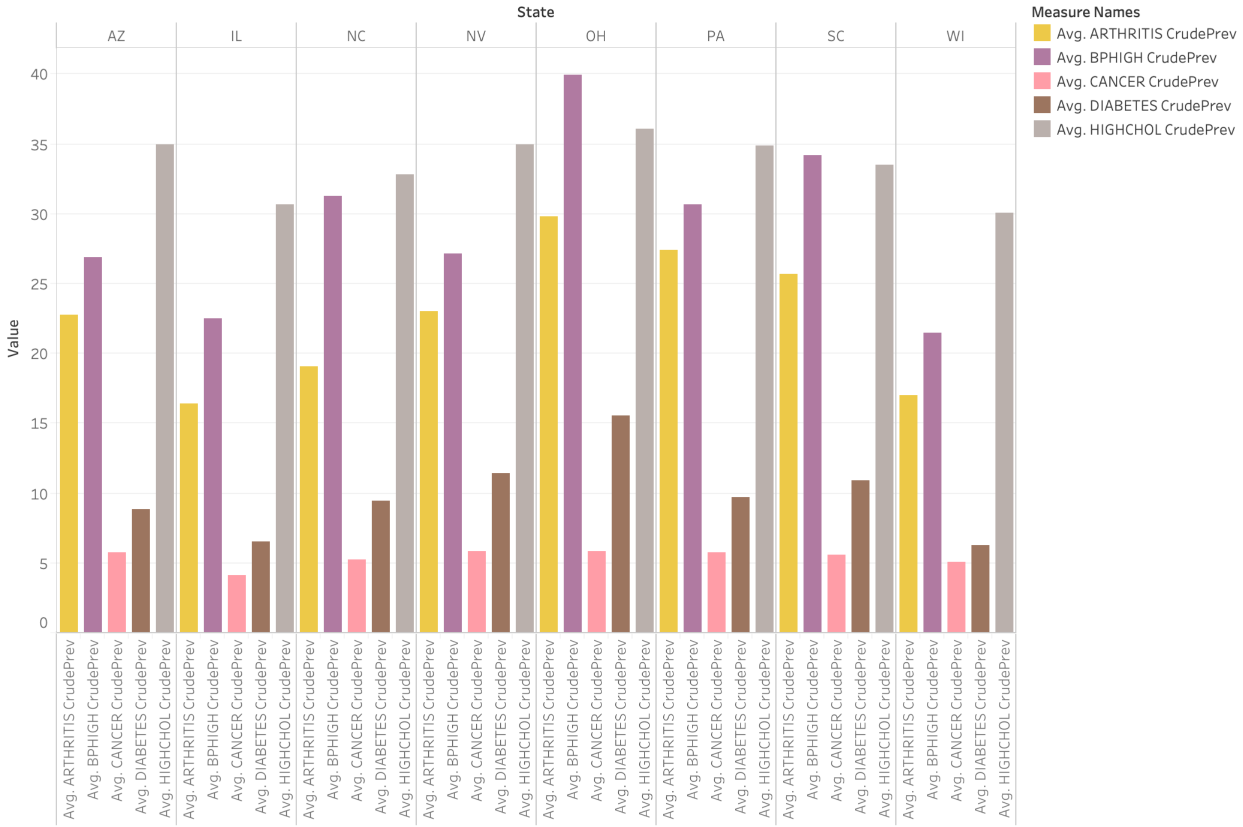


Maximum of OBESITY CrudePrev for each Categories - Restaurants. Color shows details about State. The view is filtered on maximum of OBESITY CrudePrev and Categories - Restaurants. The maximum of OBESITY CrudePrev filter ranges from 24.00 to 39.20. The Categories - Restaurants filter keeps 24 members.

The graph is plotted against the cuisine type, Obese value and across the states in USA. The above graph states that, Ohio state has highest obese value w.r.t most of the cuisine types. Wisconsin and Nevada have low obese score compare to other states. For American, Asian fusion and Italian cuisine, 5 out of 8 states have high obese scores and Mexican has 6 out of 8 states with high obese scores. It could be understood that population in Ohio has bad unhealthy food habits. The state of Ohio has the opportunity for healthy food startup restaurant and fitness center.

Figure-2 Population with health issue in various states in USA

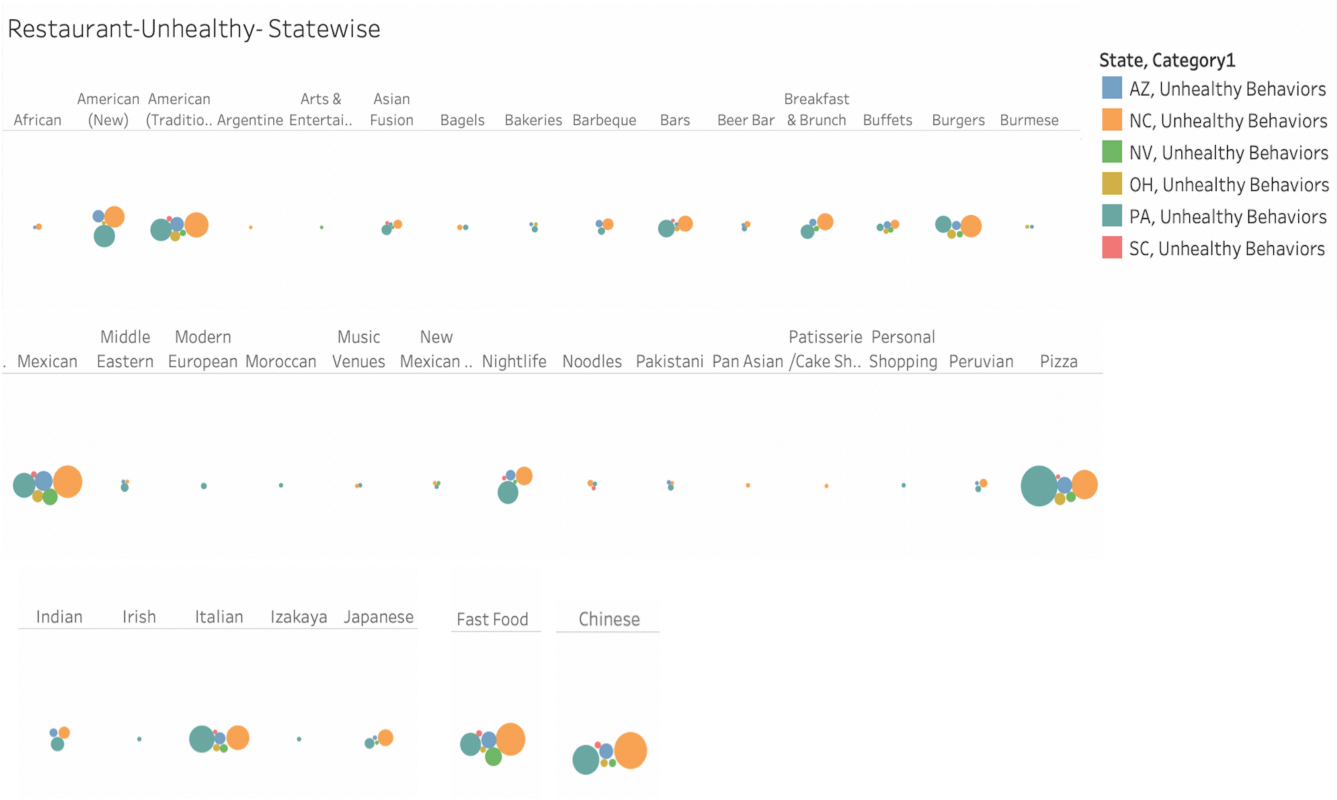
State - Avg disease



Avg. ARTHRITIS CrudePrev, Avg. BPHIGH CrudePrev, Avg. CANCER CrudePrev, Avg. DIABETES CrudePrev and Avg. HIGHCHOL CrudePrev for each State. Color shows details about Avg. ARTHRITIS CrudePrev, Avg. BPHIGH CrudePrev, Avg. CANCER CrudePrev, Avg. DIABETES CrudePrev and Avg. HIGHCHOL CrudePrev. The data is filtered on Restaraunt - category and Group - American. The Restaraunt - category filter keeps 49 of 210 members. The Group - American filter keeps 48 members.

The above graph is plotted with states against the health issues. From the above graph, the cancer disease is less than other health issues across the state. While in every state expect Ohio, major population suffer from Cholesterol and next Blood Pressure. The population of Ohio tops in all the health issues compare to other states. As discussed in the business application section, there could be an opportunity to startup healthy choice restaurant, fitness center, hospital and health research activities.

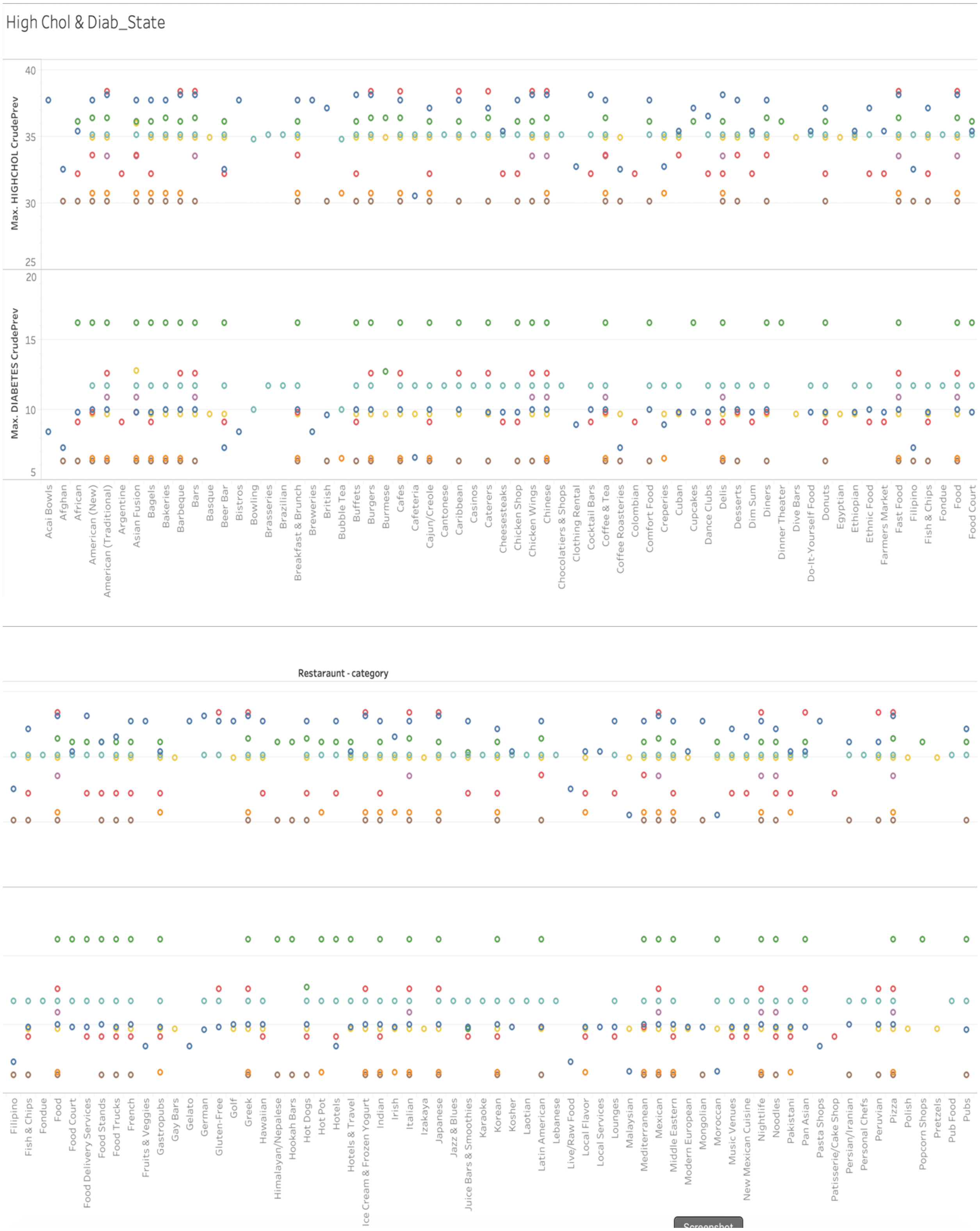
Figure-3 Unhealthy Choice (caused Obesity) for food type in various states in USA

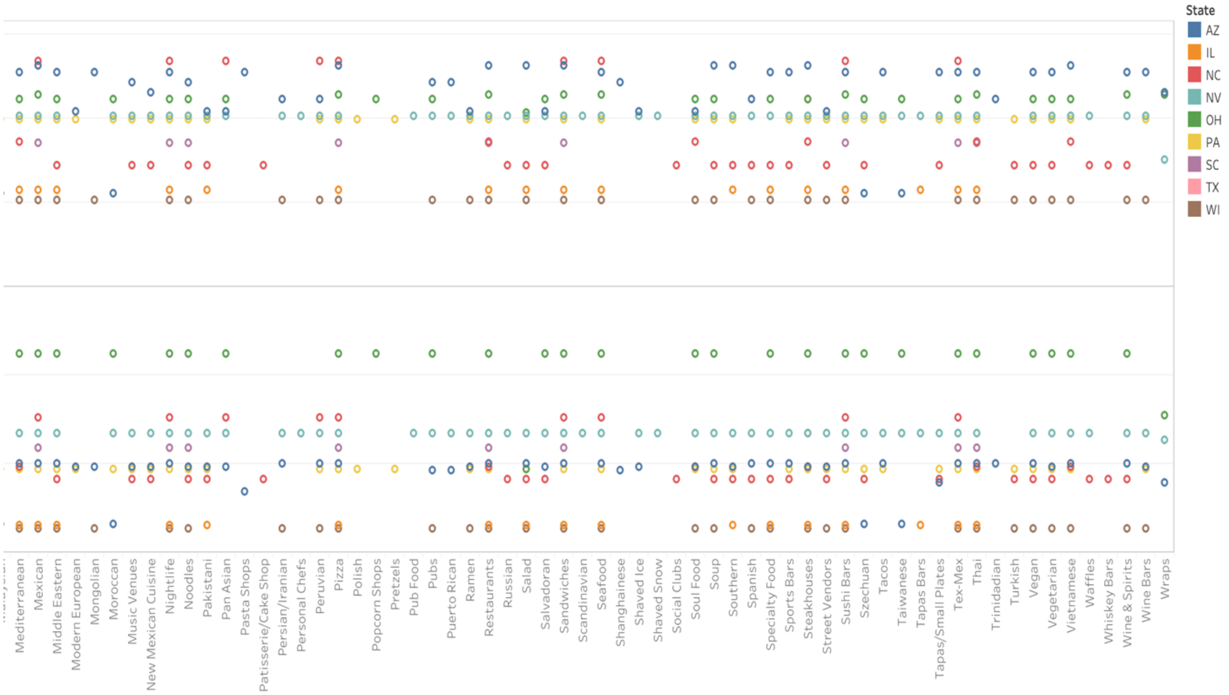


The above graph is plotted with States, Restaurant types and Unhealthy activity (obesity). This above analysis in bubble plot shows that people eating at American traditional food, Chinese, Pizza, Mexican, Italian, Fast Food, Burgers could cause obesity. Generally, every restaurant has healthy food options and fast food options, but few food types at these restaurants which might be preferred by most of the population causes obesity. For eg., food with more cheese, mayo and other sauces which generally has lots of fat should be have less options in their menu.

For Pennsylvania state, the unhealthy habits have high value for Pizza and low for personal shopping. The personal shopping could be a grocery store, supermarket and etc., where people have choice to choose the good food they wanted to cook at home, but at pizza shop major among the food ingredients is cheese. It could be possible evident from the above analysis; which food type could cause obesity.

Figure-4 Analysis of High Cholesterol and Diabetes for food type in various states in USA

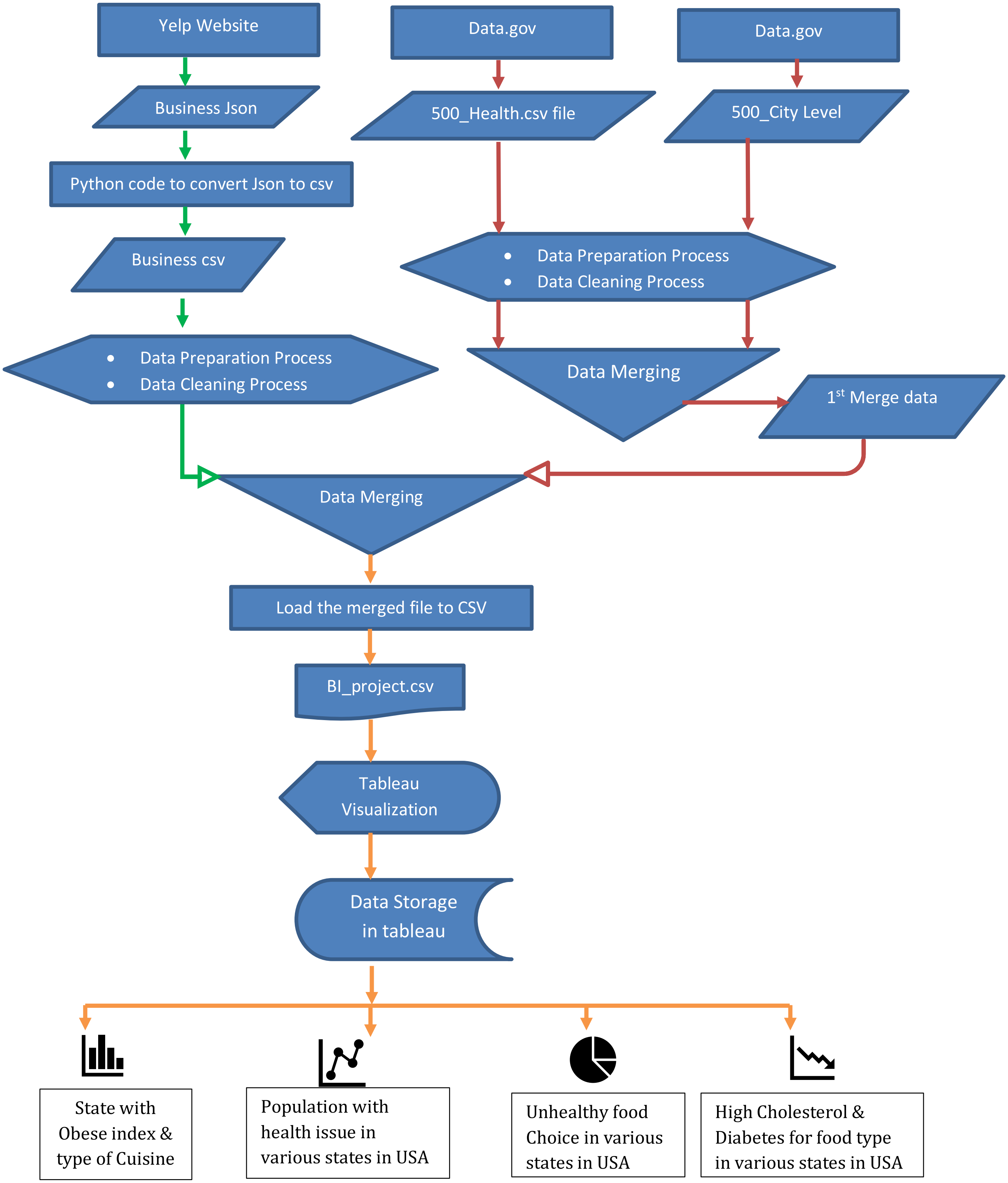




The graph is plotted along States, Diabetes, Cholesterol and categories of restaurants. This graph analyzes that Ohio tops for highest Cholesterol and second in Diabetes across all restaurants. Arizona state has highest Cholesterol value and average Diabetes. While Wisconsin and Illinois have the lowest value in both the health issues. It can be analyzed that, most of the population from Ohio and Arizona are dependent on restaurant food and while states with low health issues are less dependent. This analysis could be providing an insight to start healthy restaurant chains in Arizona and Ohio since the population might be depended on restaurants.

However, further research on health and choice of food at particular restaurants could be researched along with this dataset to unlock more insights.

V. Flow diagram: (Every Step is in Python Language)



VI. Instructions for code:

Step 1: Place the business json file, 1 excel file, all 3 csv files and two tableau workbook files in a folder

Step 2: In the code give the path link where all the files are kept

Step 3: Run the code

Step 4: bi_proj_grp1.csv would be created in the same folder path which would be used for analysis

Step 5: Import this csv file or excel file in the folder into tableau for further analysis

VII. Dataset Sources

- **Business Json:** Yelp website(https://www.yelp.com/dataset/download)
- **500_Cities__Local_Data_for_Better_Health__2018_release csv**
- **500_Cities__City-level_Data__GIS_Friendly_Format___2018_release csv**

  https://catalog.data.gov/dataset?q=500+local+cities+&sort=views_recent+desc&organization=hhs-gov&as_sfid=AAAAAAWqOHBRHNVdFbdLtqOXaM-S93f90-62IWs7MmyfrWmXAtIlJQHoe0piXsD_r99sK-8wEcF4_bk-d9LiOwpcVEINrCerX_KZWoE0XhsULLMWrFx8IeI9oPafUNuWPXJBHeY%3D&as_fid=2d8c295ffe986b9e62c798d767ae210d836ca67f&ext_location=&ext_bbox=&ext_prev_extent=-142.03125%2C8.754794702435618%2C-59.0625%2C61.77312286453146