DCSI 510: PRINCIPLES OF PROGRAMMING FOR DATA SCIENCE

FALL 2023 SEMESTER PROJECT REPORT

NAME: AKSHAYMANI SORNALINGAM

USC-ID: 8467195123

SEMESTER: FALL 2023

INTRODUCTION

PROJECT NAME: Will nutrient data analysis of fast-food items help us understand its actual nutritional composition or will it just be the usual "junk-food" that adds body fat?

DESCRIPTION: We often avoid analyzing the nutritional composition of fast-food items in depth, because we are accustomed to a preconceived notion that fast food items have only fat, and they are not good for health. This project aims to break the shackles of this notion by doing a comparative study between the nutritional data of two leading fast-food eateries – Chick-Fil-A and KFC. The data of about 120-180 food items from each eatery were taken and pre-processed. The data was reduced to 30 food items, and it was used for a statistical analysis. By analyzing various food component parameters (such as Calories, Carbohydrates, Proteins, Sugar, Fat, Cholesterol) of the items provided by the two eateries, we will be able identify the better fast-food outlet among the two. We will also be able to understand intricacies in composition of popular fast food and the relationship between each of the food components/nutrients. The results of the analysis were represented in various formats using matplotlib and seaborn.

To extend the application of the pre-processed data, a section has been included in the code where the user can interact with the final set of pre-processed data of the two eateries and check if the items, he/she selected are within the recommended range of major nutrients for his/her weight, body stature, activity, goal (weight loss, weight gain, maintenance of weight). The recommended range is generated using fitness module tool of python. The choice of items by the user will tell us which eatery would be the best for the user based on the choices made.

CHANGE FROM ORIGINAL PROJECT PLAN: I initially planned on working with calorie count data of general food items (fruits, vegetables, bread, basic junk food) from calories.info and do a comparative study between the calorie levels of healthy and junk food options of the page. Then the collected data will be used to develop different types of ideal diet plans based on users' needs and goals. A comparative study between the user's current diet and the developed ideal diet plan was planned as the final segment of the previous idea. The idea seemed very ambitious considering the short time frame as it needed a lot of background work. Choosing an ideal meal plan based on user's needs and the diet plan type needed extra research work (more to do with nutritional science in healthcare). The scope for comparative study between healthy and junk food in this case was very less as the data had only calories data and moreover the results after analysis will not be very significant or eye-opening. The idea of bringing a second data for calories data set to give more weightage to the analysis made the project more complicated as the two datasets varied a lot. So, I decided to replace the two general calorie count data sets with specific nutritional data set of items provided by two eateries. With this the data analysis became more meaningful as there were many parameters with which the two eateries can be compared. The concept of "diet-plan" was removed from the project, however the concept of user interacting with the selected datasets was still there.

PROJECT WORKFLOW

DATA COLLECTION: **2 Nutritional data sets** (amount of Fat, Proteins, Calories, Carbohydrates, Fiber, Sodium etc. of food items) are used in the project- One for the food items of Chick-Fil-A and the other one for food items of KFC. **Web-Scrapping using beautiful soup was implemented to obtain the data of Chick-Fil-A**. The URL of the webpage is: https://www.chick-fil-a.com/nutrition-allergens .Using requests package a request is made to access the required data from the website. The response is then passed to the object created for beautiful soup which is characterized by a HTML parser. The static website had around 18 sub-divided tabulations out of which only 4 tabulations had significant data for comparative analysis. The Table IDs of the chosen tables were obtained after inspecting the elements of the HTML file of the webpage. The 4 IDs were passed find_all() method of the beautiful soup's object and the data was scrapped. Each row of each table was written together in a CSV file. The final CSV file was used for pre-processing. **A PDF of all food items and its nutritional data was obtained from the webpage** https://www.kfc.com/full-nutrition-guide **for KFC**. The PDF was converted to CSV format using online software's and the data was preprocessed.

DATA PREPROCESSING: Following steps were followed for both the data sets to clean the CSV file

- Dropping the column/row that is not needed for analysis using drop()
- Checking null/NaN values from all rows by using .isna().sum() and removing the specific row
- Dropping repetitive data using drop_duplicates()
- Removing data with escape characters
- Removing special characters from item_names
- Converting numbers in string data type to integer, so statistical operations can be carried out
- Converting the units of specific columns to bring uniformity in data
- Grouping of items with very similar names using membership functions & contains() and finding the common mean value for all parameters to find the new set of values for the whole group
- Re-indexing the dataframe after removal of all the redundant data

DATA ANALYSIS: Various functions of Pandas and numpy python library were used for data analysis. To make the analysis more meaningful a variable was assigned to keep a count of the points an eatery receives for every positive factor. The basis for +1 points will be explained below.

- The Pre-Processed data which is stored as a csv was read as a Data Frame using pd.read_csv.
- This Data Frame was used to calculate the mean, minimum and maximum value of each parameter using df['Column name'].mean(), df['Column name'].min(), df['Column name'].max() respectively. The range was calculated using the difference of maximum and minimum values.
- np.percentile(data,25) and np.percentile(data,75) were used to calculate the first and third quartile range respectively. The difference between Q1 and Q3 were stored as IQR in a separate list.
- 4 common items in the two eateries were shortlisted and grouped using membership functions. The data obtained after this operation was used to compare and find which eatery is better for each of the 4 items.
- The calculated data was compared using basic assignment operators. The basis of +1 points will be based on the results of the assignment operations.
  - ✓ +1 points if the range of all the food component/nutrients is higher: Higher the IQR, Net Range more will be the spread of data for analysis/variety.
  - ✓ +1 points if the average values of carbohydrates, proteins, fiber, sodium are higher and average values of calories, fat, trans fat, saturated fat and sugar are lower.
  - ✓ +1 points if the point above applies for the 4 common items. The greater the number of items complying to that above mentioned condition, the more the chances for +1 points.

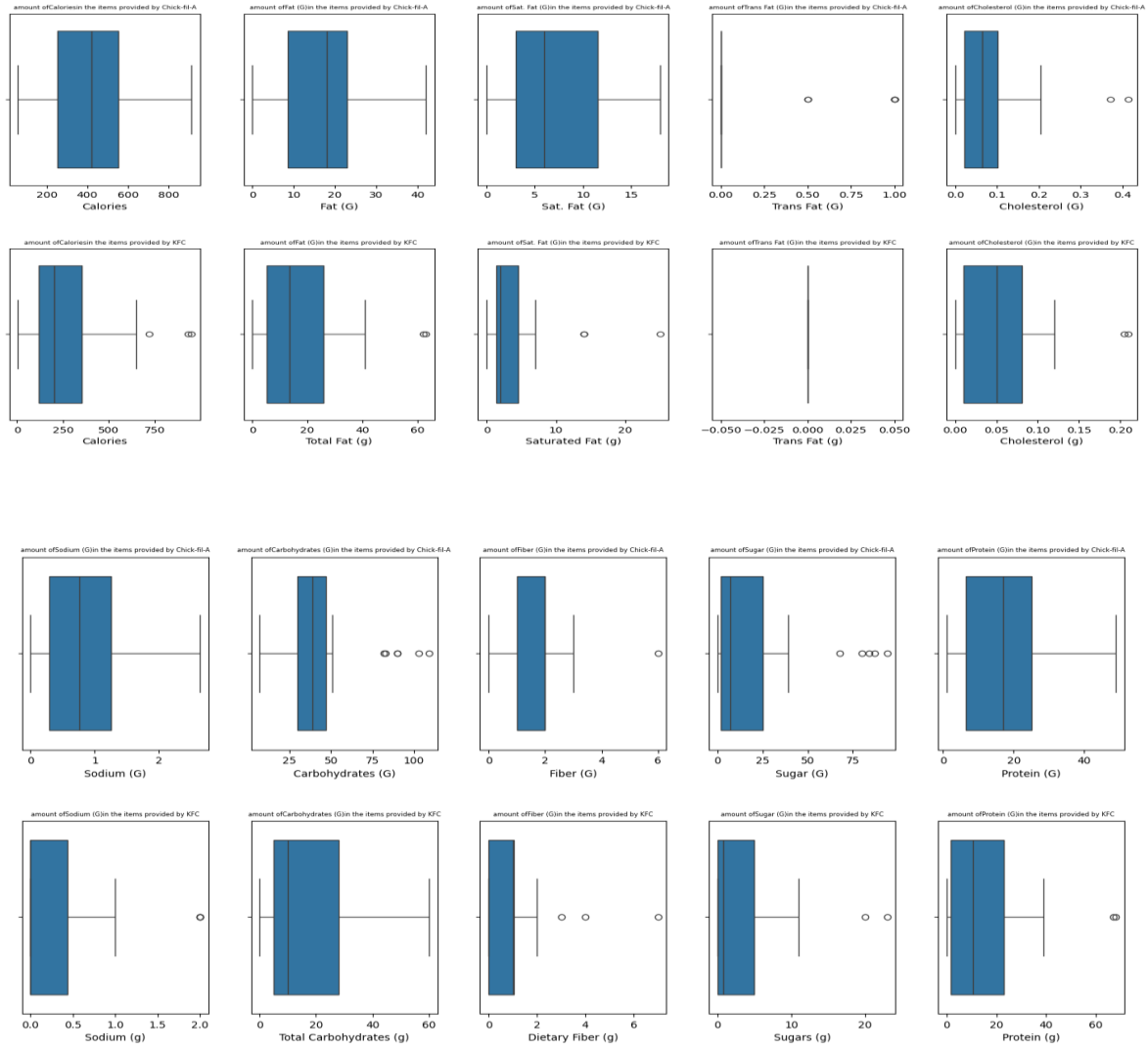
- User interaction with data: The user will be made to input the item from the displayed menu of the two eateries as it is. The nutritional data of the items selected by the user will be extracted from the analyzed data sets of the two eateries and the nutritional intake amounts of the user will be calculated based on the selected quantity used the extracted data. On the other hand, the recommended range of nutrients will be calculated using fitness tools module based on the input parameters such as weight, body stature, activity and goal. If the calculated nutritional parameters (based on the user's choice from an eatery) is within the recommended range of nutrients, then the eatery will get +1 points and these points will be stored as a new variable (points after user's choice). This is a prelude to mass cohort studies which are done to find popular eateries among the public

DATA VISUALISATION:

Whisker plots, Bar graphs, Joint bar plots, stacked bar graphs, pie charts and correlation matrices were used to represent various results of the analyzed data

Visualization 1

Visualization type: Box Plot



The question that is answered through the visualization: Which eatery is better in terms of providing food items with higher range?

Data that is visualized: The inter quartile range/net range of values of each food component in 30 food items of both the eateries is visualized.

Visualization set up and description: Set up- Seaborn uses sns.boxplot() method to represent data in the form of whisker plot. The column of the data frame that should be visualized is passed a variable x (as it must be plotted along the x axis) along with data frame from which the variable must be extracted (as data=df). Subplots are implemented in this case as there are 20 whisker plots (10 plots for each eatery) that must be represented. Each plot is given a specific title based on the parameter visualized using ax[row,column].set_title().

Description- The information that we can obtain from reading the whisker plots are: Q1(first quartile), Q3(third quartile), IQR (Inter Quartile Range), The whisker extensions (indicating the maximum and minimum value within a certain range only), the outliers (therefore the actual minimum and maximum value that lie beyond the negative and positive spectrum of the whisker) and the net range of values (minimum minus maximum). For skewed data that does not follow normal distribution IQR is used to comment on the distribution, therefore variability of data. Along with IQR, even the net range is used to check the approximate spread of the whole data. The following tabulation describes the results of the visualization.

IQR:

Food component parameter	Chick-Fil-A	KFC
Calories	Higher	Lower
Fat	Lower	Higher
Saturated Fat	Higher	Lower
Trans Fat	Zero	Zero
Cholesterol	Higher	Lower (By a small margin)
Sodium	Higher	Lower
Carbohydrates	Lower	Higher
Fiber	Lower	Higher (By a small margin)
Sugar	Higher	Lower
Protein	Lower	Higher

Net Range:

Food component parameter	Chick-Fil-A	KFC
Calories	Lower	Higher
Fat	Lower	Higher
Saturated Fat	Lower	Higher
Trans Fat	Higher	Lower
Cholesterol	Higher	Lower
Sodium	Higher	Lower
Carbohydrates	Higher	Lower
Fiber	Lower	Higher
Sugar	Higher	Lower
Protein	Lower	Higher

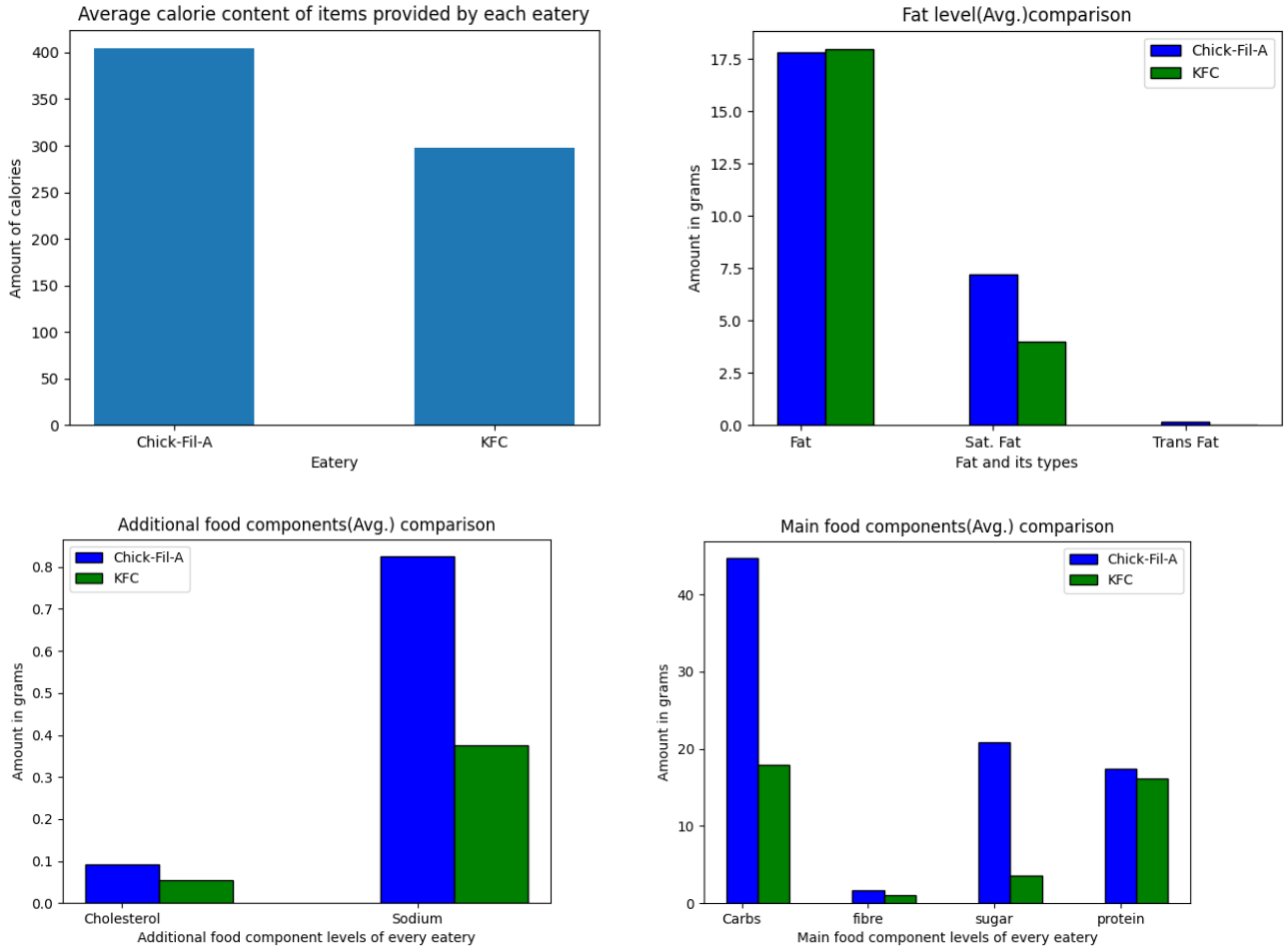
Understanding from the visualization: For statistical analysis purposes both Interquartile range and net range was taken to describe the spread of data. The information we get from IQR regrading spread and variability is comparatively more concrete than the information we get from the net range because, IQR is the measure of statistical dispersion that represents the range of the middle 50% of the data in a dataset which tells us about the points that are distributed/clustered in a certain quartile. The net range on the other hand also gives us an overview

of the spread of data but there might be changes in the variability of the data. To make the analysis holistic both parameters are chosen. The IQR is higher for 5 parameters in Chick-Fil-A and for 4 parameters in KFC (IQR is 0 for Trans Fat). In the cases where IQR is higher the clustering of data points is higher in the third quartile and in the cases where IQR is less the clustering is higher in the first quartile. The Net range is higher for 5 parameters in both the eateries.

Conclusion: In this case IQR comparison is the tie breaker in Net range comparison, moreover the veracity of IQR for a comparative study is higher, therefore we can conclude that the range of food component parameters in the food items provided by Chick-Fil-A is higher, which fetches Chick-Fil-A 1 point.(Higher the range, better the variability, therefore more the choices- making the eatery a better choice)

Visualization 2

Visualization type: Bar charts and Joint bar plots



The question that is answered through the visualization: After holistically analyzing the data of all 30 items of the two, which eatery provides higher amount of calories/protein/sugar/carbs/fat/cholesterol?

Data that is visualized: The mean values of every parameter of both the eateries are visualized as a joint bar graph.

Visualization set up and description: Set up- The number of bars that should be represented together as a joint bar is first decided. Here, it is two as we are dealing with 2 eateries and the snippet plt.bar(r, data, label) is written that many times .r+width of the bar is used for setting up the second bar. r which is equal to np.arange(len(data)) is the number of full joint bars we need in one graph. For Calories normal bar charts are used.

Description- The Average calorie value is represented as a normal bar chart and the rest of the mean values are represented as joint bar graphs where the parameters are grouped into 3 based on their range of values and type of parameters. The following are the **approximate mean values** that we obtain from the chart representation.

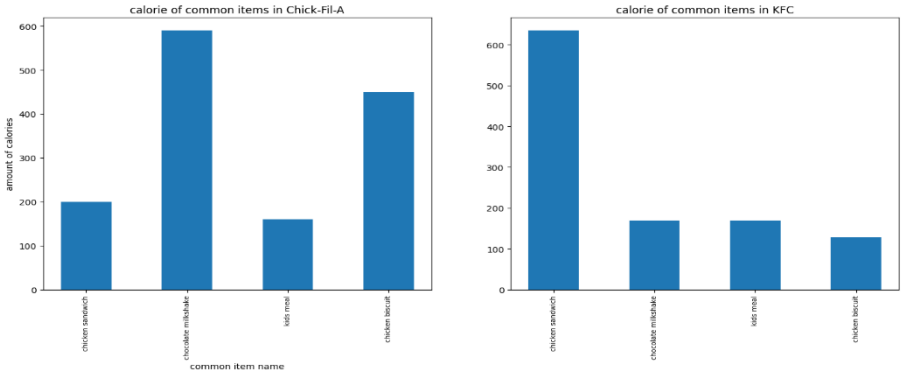
Food component parameter	Chick-Fil-A (in g)	KFC (in g)
Calories	400.0 Cal	200.0 Cal
Fat	17.6	17.9
Saturated Fat	7	3.75
Trans Fat	0.1	0.0
Cholesterol	0.09	0.06
Sodium	0.8	0.35
Carbohydrates	45.0	19.0
Fiber	3.0	1.0
Sugar	20.0	5.0
Protein	18.0	16.0

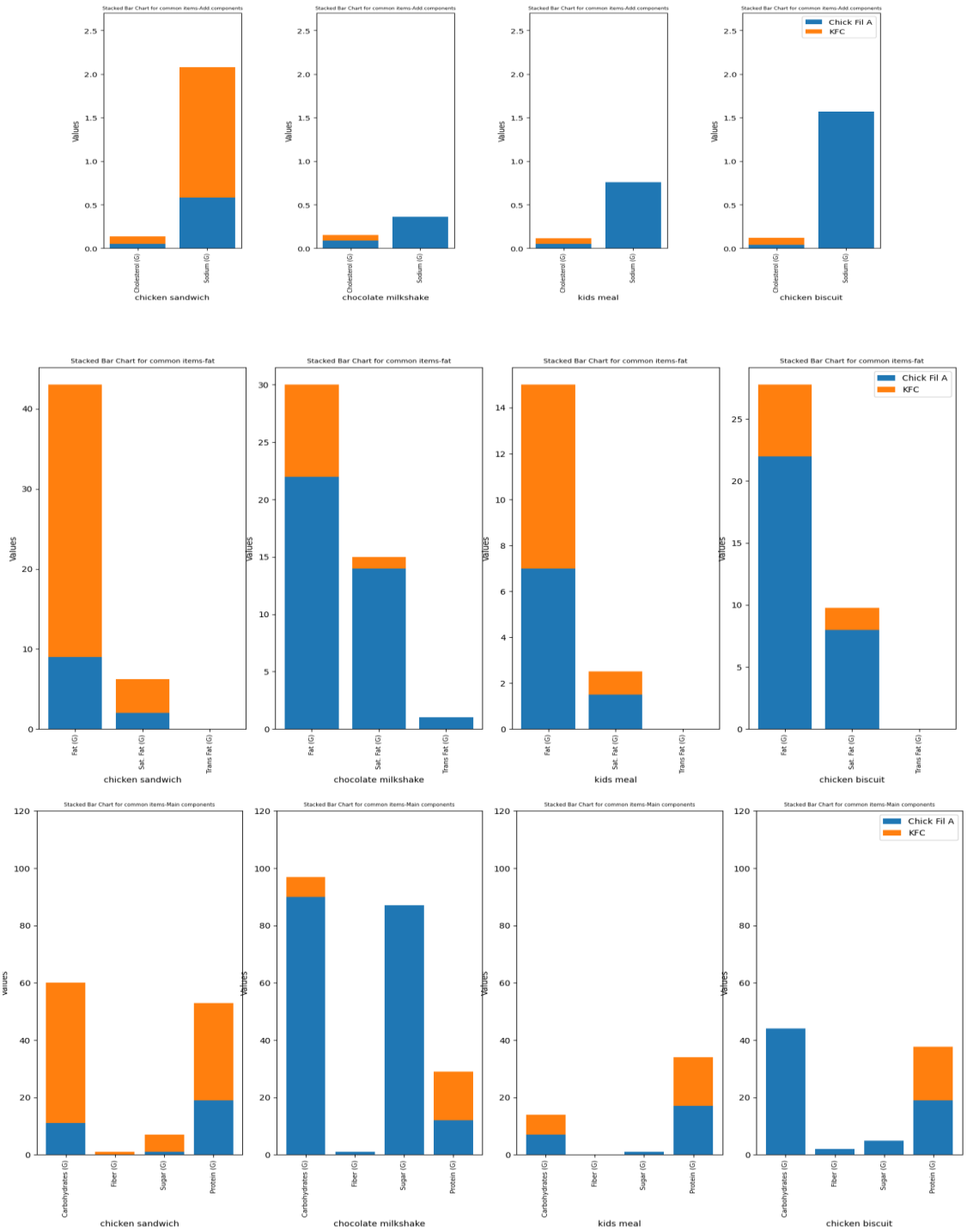
Understanding from the visualization: From the graphs it is evident that calories, protein, sugar, Fiber, Carbohydrates, Cholesterol, Trans Fat, Saturated fat are higher in items provided by Chick-Fil-A and the rest of the parameters are higher in the items provided by KFC. For an eatery to be considered over the other (therefore receive points) the items provided by them must be statistically higher in carbohydrates, protein, fiber, Sodium and lower in the remaining food components. (therefore, considered as good parameters)

Conclusion: Since both the eateries have equal number of good parameters both Chick-Fil-A and KFC get equal number of points

Visualization 4

Visualization type: Bar charts and Stacked bar graphs





The question that is answered through the visualization: Which eatery is preferred for each of the common food items of Chick-Fil-A and KFC?

Data that is visualized: The numerical values of all 10 parameters of the 4 common food items are visualized as a stacked bar graph.

Visualization set up and description: Set up- The same syntax as used for representing bar charts is used for stacked bar charts. An extra line (following the same syntax) must be included to account for the bottom bar after choosing which data should be considered as the bottom segment of the stacked bar. Bottom=data is included as

a parameter in the second line. Chick-Fil-A is stacked at the bottom and KFC is present at the top. For calories alone normal bar chart is used for visualization

Description- Following are the **Approximate values** of the 10 parameters for the 4 items

Chicken Sandwich:

Food component parameter	Chick-Fil-A (in g)	KFC (in g)
Calories	200.0 Cal	630.0 Cal
Fat	9.0	43.0
Saturated Fat	2.0	4.5
Trans Fat	0.0	0.0
Cholesterol	0.1	0.3
Sodium	0.5	2.0
Carbohydrates	10.0	60.0
Fiber	0.0	1.0
Sugar	1.0	5.0
Protein	16.0	50.0

Chocolate Milk:

Food component parameter	Chick-Fil-A (in g)	KFC (in g)
Calories	580.0 Cal	180.0 Cal
Fat	22.5	29.0
Saturated Fat	12.5	13.5
Trans Fat	1.0	0.0
Cholesterol	0.1	0.15
Sodium	0.35	<0.35
Carbohydrates	90.0	97.0
Fiber	1.0	0.0
Sugar	83.0	<83.0
Protein	8.0	28.0

Kids Meal:

Food component parameter	Chick-Fil-A (in g)	KFC (in g)
Calories	170 Cal	180 Cal
Fat	7.0	15.0
Saturated Fat	1.7	2.1
Trans Fat	0.0	0.0
Cholesterol	0.09	0.11
Sodium	0.7	<0.7
Carbohydrates	5.0	11.0
Fiber	0.0	0.0
Sugar	1.0	<1.0
Protein	15.0	36.0

Chicken Biscuit:

Food component parameter	Chick-Fil-A (in g)	KFC (in g)
Calories	410 Cal	120 Cal
Fat	22.0	27.5
Saturated Fat	7.5	9.5
Trans Fat	0.0	0.0
Cholesterol	0.05	0.10
Sodium	1.5	<1.5
Carbohydrates	41.0	<41.0
Fiber	2.0	<2.0
Sugar	4.0	<4.0
Protein	19.0	36.0

Understanding from the visualization:

- From the graphs it is evident that for Chicken sandwich- Calories, Sodium, protein, sugar, Fiber, Carbohydrates, Cholesterol, Fat, Saturated fat and protein are higher in items provided by KFC and the rest of the parameters are higher in the items provided by Chick-Fil-A. Trans Fat is 0.
- For Chocolate milk- Calories, Protein, Sugar, Fiber and Sodium are higher in items provided by Chick-Fil-A and the rest of the parameters are higher in items provided by KFC. Trans Fat is 0.
- For Kids Meal - Protein and Sugar are higher in items provided by Chick-Fil-A and the rest of the parameters are higher in the items provided by KFC. Trans fat and fiber are 0.
- For Chicken Biscuit- Calories, Protein, Sugar, Fiber and Carbohydrates are higher in items provided by Chick-Fil-A and the rest of the parameters are higher in the items provided by KFC. Trans Fat is 0.
- For an eatery to be considered over the other (therefore receive points) the items provided by them must be statistically higher in carbohydrates, protein, fiber, Sodium and lower in the remaining food components. (therefore, considered as good parameters)

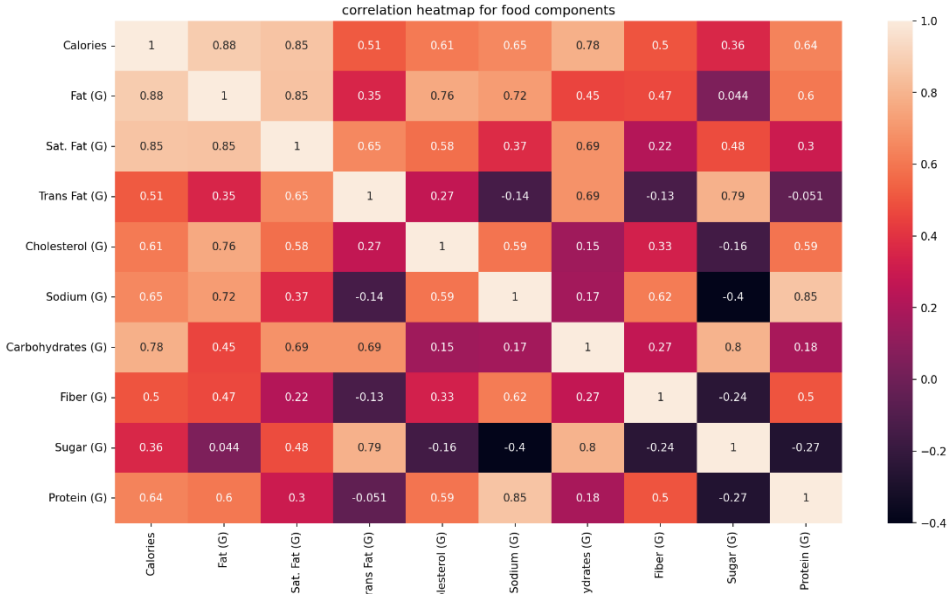
Conclusion:

Item Name	Best eatery for the item
Chicken sandwich	KFC
Chocolate milk	Chick-Fil-A
Kids Meal	Chick-Fil-A
Chicken biscuit	Chick-Fil-A

Since Chick-Fil-A is chosen for majority of items, Chick-Fil-A gets a point and becomes more preferred than KFC.

Visualization 5:

Visualization type: Correlation matrix



The question that is answered through the visualization: How are each food component parameter levels related to each other?

Data that is visualized: The basic relationship between the food component parameters of all items in Chick-Fil-A for examples using the correlation values present in the grid.

Visualization set up and description: Set up-Correlation of the data frame is obtained using corr() function and it is visualized using sns.heatmap(). Description- Since there are 10 parameters a 10x10 grid is created where each grid will have a numerical value that tells us about the relationship of the two parameters. Heatmap gives the multicolor effect so the color gradient also will help us understand the relationship between the two parameters (the color gradient is linked to a scale –0.4 to 1.0: the correlation values).

Understanding from the visualization and conclusion: If the values are positive, the two parameters are directly proportional, and the direct proportionality becomes lesser with decrease in the positive value, and it becomes inverse proportionality with negative values. The correlation values are the highest (I.e 1) along the diagonal.

Top 3 Highest positive values and the parameters involved:

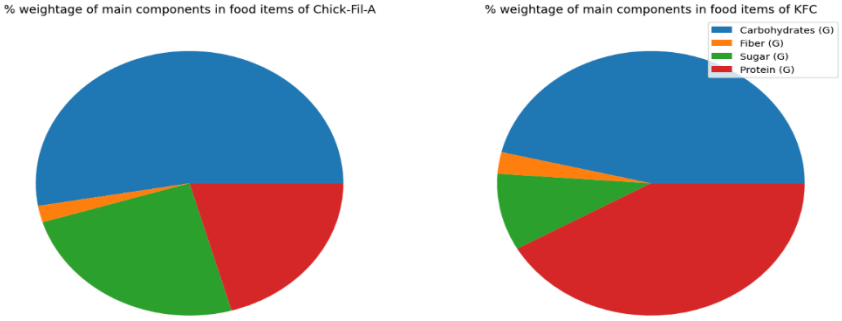
Correlation value	Parameters involved (directly proportional)
0.88	Calories-Fat
0.85	Calories- Saturated Fat
0.80	Carbohydrates- Sugar

Top 3 lowest negative values and the parameters involved:

Correlation values	Parameters involved (Inversely Proportional)
-0.40	Sodium and Sugar
-0.27	Protein and Sugar
-0.24	Fiber and Sugar

Visualization 6

Visualization type: Pie Chart



The question that is answered through the visualization: What is the weightage of the main food components (carbs, fiber, sugar, protein) in the food items provided by KFC and Chick-Fil-A?

Data that is visualized: percentage of each nutrient (sum of specific nutrient values of all items/sum of all nutrients values for all items) was appended in a list and used for visualizing its weightage.

Visualization set up and description: Set Up- In general pie charts accept list of data in any numerical format (not just percentage). Plt.pie(list_of_data,labels=labels) method automatically calculates the percent share of segment in the pie and assigns the individual labels. Each segment is represented using different colors for which a legend is included using plt.legend(). In the above case, since two pie charts are represented side by side, subplots are used and ax[column_no.] is used to place the pie chart in the right axis. To avoid the crowding of labels, the list of labels is passed as a parameter to plt.legend().

Description- Following are the percentage weightages of the main food components of all food items in the tow eating outlets than are obtained on studying the pie chart

Major food component/nutrient	Chick-Fil-A (%)	KFC (%)
Carbohydrates	52.74	46.74
Fiber	2.02	2.63
Sugar	24.63	9.38
Protein	20.59	41.76

Understanding from the visualization: The weightage of carbohydrates and sugar are higher in the Chick-Fil-A, the weightage of protein in KFC is double the weightage of the protein in Chick-Fil-A and the weightage of fiber is marginally higher in KFC as compared to Chick-Fil-A

Conclusion: This result says that the major nutrients that are considered as health promoters are more in items of KFC than Chick-Fil-A. This visualization was included just to show the weightage of the main food parameters. The other parameters should be taken into consideration holistically before a concrete conclusion is drawn (which was done in the previous visualizations)

PROJECT CONCLUSION

On qualitatively analyzing the data, the results obtained were not stable. While there were results which said the main food components such as protein is higher and sugar is lesser in KFC, there are results which negated it. Carbohydrates are more in the food items provided by Chick-Fil-A, The IQR of majority of the important parameters are more in Chick-Fil-A. Considering the fact that this is a statistical analysis-based project, I decided to stick to the quantitative results that I obtained. After a detailed statistical analysis of the complete nutritional data of Chick-Fil-A and KFC where mean, IQR, Net Range for each parameter are calculated and points were assigned for every positive factor in the two eateries, **we come to a conclusion that Chick-Fil-A is better than KFC (As Chick-Fil-A has more points than KFC).**

FUTURE WORKS

- Integrate data of more eateries to the current project.
- Develop an ML based robust recommendation engine for diet plans with the currently used data.
- An attempt to develop a Graphical user interface (GUI) in future as the current idea is a prelude to a user-interaction based program.