COGS9: Introduction to Data Science
Final Project: Part 2
Submit through Gradescope as a group, do not submit individually

Group Member Information:
https://colab.research.google.com/drive/1HSO3dvCimtgPZaSdWFvJptgcYqYVaWNa#scrollTo=dUmjH0DPaDJG

First Name	Last Name	PID
Garay, Marlon	mjgaray@ucsd.edu	A17648077
Deng, Brian	wudeng@ucsd.edu	A16862564
Lo, Courtney	clo@ucsd.edu	A16892581
Lei, Jingqiao	j4lei@ucsd.edu	A15909014

Question

Clearly state the specific data science question you're interested in answering. This question can be the same as what you submitted for your project proposal. Alternatively, you can edit your original question or change your topic completely. (2 pts)

Is the rating of cereal brands affected by factors such as shelf placement, sugar content, and number of calories per serving?

Hypothesis

Write down your groups hypothesis to your question. Provide justification how you came to this hypothesis. (What background information or instinct led you to that hypothesis?). (2 pts)

We hypothesize that cereals with high amounts of sugar, high calories and that are lowest on the shelf – closest to children's eye levels – correlate to a higher rating of the cereal brand.

We hypothesize this because research indicates that companies' marketing strategies aim to target young adults and children by placing high sugar content cereals at eye-level, and on lower shelves, to make long lasting impressions on buyers. Since cereal is mostly eaten by young adults and children as a breakfast meal, it makes sense to infer that cereals with high sugar contents – which positively correlates with the number of calories – caters more to young kids' taste palettes. Cereal brands that are consumed and purchased more frequently, lead to an overall higher rating, than less popular brands.

Background Information

Include a few paragraphs of background research and information on your topic. This should include at least 2 citations to work from others. Including hyperlinks to reputable sources are fine. (2 pts)

Based on further research, the U.S. Departments of Agriculture (USDA) and Health and Human Services (HHS), suggested in the 2020-2025 authoritative guidance Dietary Guidelines for Americans (DGA), more nutrient dense forms of grains such as breakfast cereals containing less sugar would help form healthier, regulated diet patterns [1]. In addition, while breakfast cereals contain high amounts of free sugar, research did not find a strong association between sugar content and energy or fat content. This indicates that cereals do not provide significant amounts of fat, but instead provide carbohydrates which come with fewer calories [2]. While mentioning healthy diets could help prevent noncommunicable diseases (NCDs) such as diabetes, heart disease, and cancer, the WHO recommends limited free sugar intake for health benefits since free sugars are strongly associated with obesity and other harmful health problems [3].

However, in most supermarkets in order to maximize profits, marketing strategies place the most tempting, sugary cereals on shelves that are easily seen and reachable by young kids. Extensive research has shown that adults tend to buy cereal brands that are placed on eye-level shelves, while children are more likely to be attracted to products at shelf levels they can touch, which are often on the lower shelf (about 0.9m - 1.2m), positioned at their eye level. Once kids are attracted by the eye-catching design of the cereal boxes, which are marketed to appear sugary and tasty, they're prone to asking their parents to purchase that brand of cereal [4]. This is why nowadays, children have become an influential, powerful force on product consumption. Therefore, most marketers began targeting younger audiences, for example, in the breakfast cereal category. Marketers know children like sweet meals, so they tend to put breakfast cereals with higher amounts of sugar on the lower shelf in order to target children. The sugary cereal boxes, typically containing a charming cartoon character, are shelved lower at children's eye level. Moreover, researchers have found that lower shelf placement, induces eye contact between kids and the cartoon characters on cereal boxes. This increase in eye contact leads to an increased trust and connection between the brand and child [5]. The brands that build trust and bond with children, tend to be more popular and likable amongst young consumers.

[1]: Dietary_Guidelines_for_Americans_2020-2025
[2]: Examining the Relationship between Sugar Content, Packaging Features, and Food Claims of Breakfast Cereals
[3]: Healthy Diet from World Health Organization
[4]: Examining the Relationship between Sugar Content, Packaging Features, and Food Claims of Breakfast Cereals
[5]: Why cereal boxes are at eye level with kids

Data

Include a description of the perfect dataset you would need want to answer this question. How many observations would you need? What variables would you collect? Explain the perfect dataset that you would want to answer this question.

Then, look online for available datasets. Find a dataset that could be used to answer this question. Describe how many observations are included and what variables have been collected. Discuss the datasets limitations and how it differs from your ideal dataset. If you collected your own data, explain what information you collected, from whom you collected it, and a link to the data. (3 pts)

The "perfect" dataset needed would be a large dataset with 500 cereal information that includes facts about multiple cereal brands, along with their sugar, calories, nutritional content, and average shelf level placement. Having at least 50 cereal brands would be ideal, however, having more than 50 brands would be ideal, including cereals that are a part of private label brands. Another important data factor that would be needed would be the overall rating of the cereal, as this would allow us to correlate the amount of sugar and consumer rating of the brand.

The dataset we found was on Kaggle and includes information on almost 80 different cereal brands. The main setback is that there is only information on 3 shelf heights, which is not representative of common grocery stores which typically have more than 3 shelf heights, as well as the fact that the information is not up-to-date, as it was not collected in the 2020's decade and hasn't been updated recently. The dataset did include information on the manufacturer of the cereal brand, however, the information is only represented by a letter and not the full name of the manufacturer, making it difficult to use. The dataset contains hot and cold cereals which could be useful if we wanted to see how preferences of hot and cold cereal correlate to rating. The dataset contains other useful information such as cups per serving, sodium content, calories, carbohydrates, weight, shelf height(from the floor), and vitamins.

For a more "perfect" dataset, we would include more shelf options and collect the relative data regarding placement of cereal brands, along with having the manufacturer listed fully instead of being represented by a letter. To collect the data, we would go through multiple big and small supermarkets to record the shelf levels, since larger stores are more likely to have more shelf options. As for the manufacturer list, we would search on the internet to find the comprehensive list as well as look at the information on cereal boxes when we do research in supermarkets.

Ethical Considerations

Data Collection

The data – ratings of different cereals – was collected from participants, leaving their personal identity at risk of being breached, in the case that the location and time of the data collection was found out. Each participant was informed that their votes would be released anonymously, however, they were also asked to sign a consent form in the case that anything were to be leaked. Although the metric used to calculate the final rating for each cereal brand was vague, there is demographic information of the participants who participated in voting for rating of the cereal brand, leaving there to be possibly bias in the rating of the dataset. One such bias is collection bias; for example, there is only information on 3 different shelf heights, which could mean that the authors/data collectors went to specific grocery stores with cereals only placed on 3 shelves to collect the data, which is not representative of all grocery stores. Another example of collection bias in this dataset are the cereal brands chosen to be collected and rated. There is no data that contains ALL the cereal brands sold nationwide, meaning that the dataset only has data of cereals that were available to the data collectors, which would bring bias as they are only collecting data on cereals that are available in areas around them, not representative of what cereals are available nationwide or globally. Also, the data may be outdated, as the Kaggle page containing the dataset shows that the last update made was 5 years ago, which is not representative of current day cereal ratings. In addition, there is also bias in the collection of consumer ratings. It is unclear how the data was collected, and there is a possibility that the participants, when filling out the survey, may not have answered truthfully, either due to forgetfulness or simply lack of motivation to answer all the questions. Also the wording of the survey itself may have persuaded or leaned towards certain answers. To address this, we would suggest researching in various different supermarkets to collect more data, as well as make online interviews with customers, allowing them to share all their opinions, rather than narrowing them down for a short, written survey. To make ratings more accurate, we would like to research more information about the sales tactics of each cereal brand, as we believe sales and marketing strategies also relate to the popularity of the cereal brand.

Data Storage

The data itself does not contain any personal information of an individual; it contains information of inanimate objects. We do not feel that the data could endanger an individual if not completely deleted. The data was collected by a user on Kaggle, and was shared under a CC BY-SA 3.0 license which allows the content to be shared, remixed as long as the data is shared under the same license as the original and must be attributed to the original author.

Data Analysis

One data bias may be that the data lacks more than 3 shelf heights, when many grocery stores have cereal placed on more than 3 shelves. Also, the data may be outdated, as the nutritional content, shelf placement and overall rating of the cereal brand may have changed in the past 5 years, and the dataset was not updated to represent these changes. There is also a possibility that some cereal brands may have been removed or introduced to supermarkets, or are no longer in production. There is no information on where the data was collected, or whether or not this was a nation-wide supermarket chain, in which all the cereals were placed on the same shelves across all locations. The 'rating' provided by the dataset does not specify what demographic the data was collected from, so it is unsure what age range – children, teens, adults – was surveyed. As mentioned in the Data Collection section, we would collect more recent, accurate, and comprehensive data regarding shelf levels and customer ratings, in order to eliminate some bias when analyzing the data.

Modeling (Statistical or Machine Learning)

We feel that the data could be unreliable in the sense that it could have been collected at random stores, or with publicly available information. For example, there is no information about where the data was collected, so there is a chance that the data was collected in upper-income neighborhood supermarkets, which may not sell many cereals that are more common nationwide. Another possibility is that the data was collected through freely available information from many sources, such as online or through other datasets. This provides missing perspectives and only shows information that fits the conditions of where it was collected. The data may lack information about brands available at other stores, or ratings from diverse groups of consumers.

Project or Model Deployment

We plan to use the data to 'uncover' the correlation between sugar content, calorie content, and shelf placement, with cereal brand ratings. We will try to avoid purposeful manipulation of the data, which may misrepresent a cereal brand, and will state that much of the information in the dataset did not include context behind the data collection, demographic of who participated in the survey rating, or location of where the data was collected. The data should not be interpreted as being representative of the demographic who voted in the rating of the cereal, nor is it intended to be applicable to current day ratings of brands, as the data was last updated 5 years ago. We must be cautious as to not assume that the data analyzed is applicable to current-day cereal brands, as sugar content, formulation, and other information could have changed since the last data collection.

Analysis Proposal

Data Collection

The first thing we did was find a dataset that included information on different cereal brands, including the manufacturer, ratings from consumers, sugar content, nutritional content, and shelf placement. We looked for a dataset that contained a large amount of different brands, as it would be more beneficial to finding information that could be used to answer our hypothesis. Since we did not have time to create an ideal dataset by collecting data ourselves, – researching, surveying, and experimenting at local stores – we obtained the data we needed through a dataset that includes 7 unique cereal manufacturers and 77 different cereal brands from Kaggle – publicly available to download from the website.

Data Wrangling

Using Google Colab, we were able to edit our data (the one from Kaggle) with Python along with the Pandas, Matplot and Numpy software libraries.

After downloading the CSV file from the Kaggle website, we first uploaded the CSV file and implemented the code in order to create the first needed dataframe, which contained the original information. The file was cleaned prior to analysis by the dataset providers Petra Isenberg, Pierre Dragicevic and Yconna Jansen, and it was tidy. We were able to create multiple data frames from the dataset, allowing us to add or remove data columns when it was deemed appropriate and necessary. The Pandas library also allowed us to take the mean of the data per brand, sort columns by variables such as the manufacturer or brand, and allowed us to sort information in ascending or descending order, which helped with the visualization of the data. From this data, we created data frames that allowed us to view the rating of the cereal brand, sugar and nutritional content, and the correlation with shelf height.
If the data was originally untidy, we would have first started by tidying the data, doing so by keeping the same variable names across the file to maintain consistency and generalizing the consistent expressions for different manufacturers and shelf placement. The dataset would use the first capitalized letter for each manufacturer of cereals – K for Kelloggs, Q for Quaker Oats – and the shelf columns numbered 1, 2, and 3 to denote shelf levels from the floor up in order to clean up the dataset. If we accessed our data from excel, we would delete unnecessary headings, merge cells, delete empty cells, or clean duplicated content by setting each column to have one variable, and have each variable to be entered in both rows and columns. We would also ensure that the variables are in the correct format needed for analysis – numerical variables in numeric format without the units and categorical variables in factor format.

Descriptive & Exploratory Data Analysis

Basic manipulation of the data, along with data frames with relevant information, were created in order to better understand what would need to be done to answer our hypothesis. We created a data frame that included cereal brand, consumer ratings, and sugar/nutritional content.

- **Summary Stats:** We would calculate summary statistics such as mean, median, and standard deviation for each variable. Understanding these statistics would be helpful in deciding on the appropriate statistical methods to use in our analysis. For example, if our data is normally

distributed, we can use parametric statistical tests such as t-tests or ANOVA, whereas if the data is not normally distributed, we will use non parametric tests.

- **Correlation:** We will compute the correlation coefficients between the independent and dependent variables we are interested in, to determine the strength and direction of their relationships (the approach we chose for Inferential Analysis is correlation, which will be discussed later in more detail).

- **Plots:** We will also create visualizations to explore the relationship between the independent variables (sugar, calories, and shelf placement) and the dependent variable (ratings).
    - **Univariate plots:** We would create histograms for the independent variables and the dependent variable to show the distribution of each single variable. By creating histograms of the sugar content, for example, we will plot the counts on the y-axis against the sugar content bins on the x-axis which can help us identify the shape and central tendency of the distribution and some potential outliers. If the histogram of sugar contents is skewed to the right, we may infer that most cereals have low sugar contents and a few may have high sugar contents.

    - **Bivariate:** We would use a box plot to visualize the distribution of ratings for the cereals in our dataset across different levels of independent variables. For example, we would create a box plot for ratings of cereals with high sugar content and another box plot for cereals with low sugar content. We can compare the median rating, quartiles, and outliers to determine if there is a significant difference in ratings between the two groups. We want to use box plots because it can also be used to identify potential outliers, which are data points that are significantly different from the rest of the data. Outliers can be due to measurement error or actual extreme values. By identifying outliers in the box plot, we can investigate further to determine the cause and decide whether to include or exclude them in our analysis. We also want to use scatter plots to explore the relationship between two continuous variables, in order to identify patterns in the data. For example, we can plot a scatter plot with sugar contents on the x-axis and consumer ratings on the y-axis. The plot can show whether there is a positive or negative linear relationship between sugar content and consumer ratings, or whether there is no relationship at all. We can also plot the regression line on the scatter plot to see the direction and strength of the linear relationship. Similarly, we can plot a scatter plot with calories on the x-axis and consumer ratings on the y-axis to visualize their relationship, and a scatter plot with shelf placement on the x-axis and consumer ratings on the y-axis to see whether or not our hypothesis – cereals with lower shelf placement have higher consumer ratings – holds true.

    - **Multivariate**: After we use the scatter plots for bivariate statistics, we will use a muti-variable scatter plot matrix to show scatter plots of all pairs of variables, making it useful for identifying potential correlations or relationships between multiple variables at once.

Data Visualization

Figure 1: Comparison of cereals with Low Sugar (blue), and cereals with a High Sugar (red). It is visible that among the low sugar content cereals, there are higher ratings.

The “Low Sugar Cereal” were placed in a data frame where they had sugar content lower than the average sugar content (in grams) of all cereals (6.91).

The “High Sugar Cereal” were placed in a data frame where they had sugar content higher than or equal to the average sugar content (in grams) of all cereals (6.92).

The chart below shows the two data frames of cereal separated by sugar content. The graph displays that Low Sugar Cereals have a higher rating.

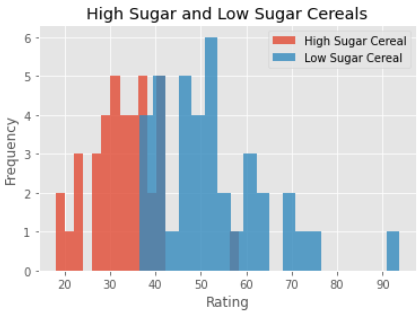


Figure 2: Using the data frame that has data of all cereals, it is visible that there are higher peaks in ratings when the brand of cereal has less sugar content.

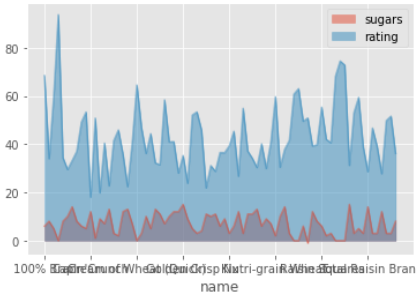


Figure 3. Correlation between sugar content and rating of all cereals in the data set.
It is visible that ratings are higher among cereals with lower sugar content.

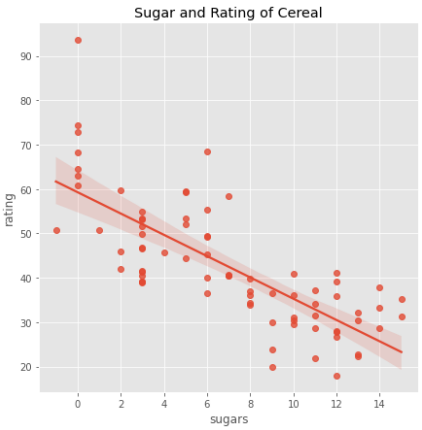


Figure 4. Shelf height and their sugar and rating content.

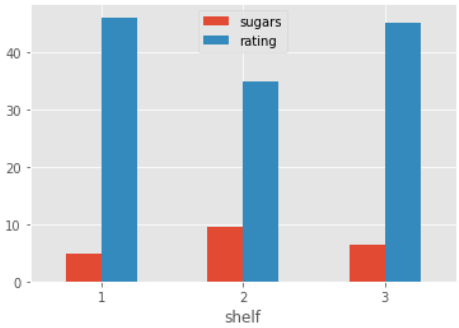


Figure 5. Relationship between sugar content and rating.

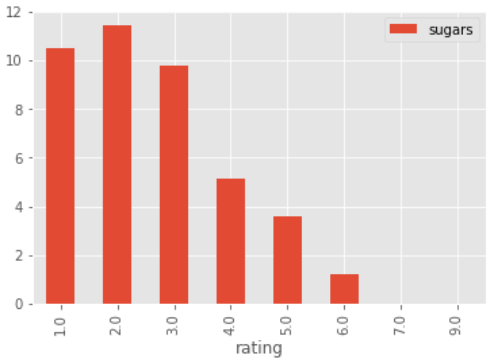


Figure 6. Relationship between shelf placement and rating.

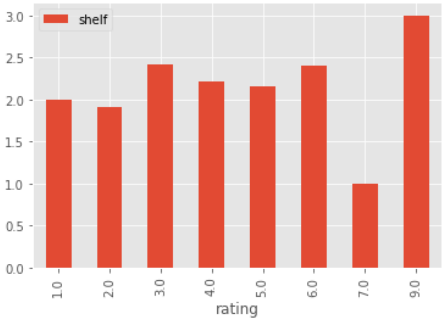


Figure 7. Relationship between calories content and rating.

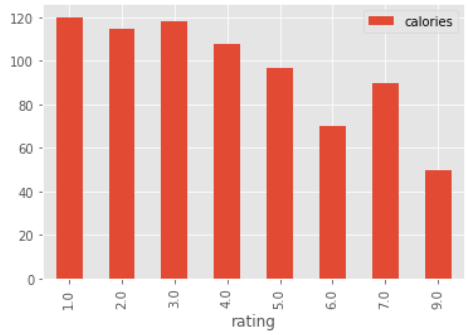
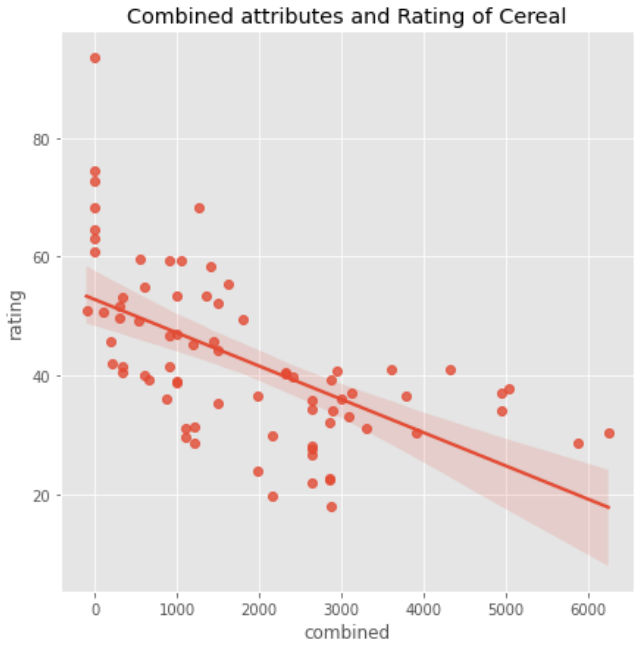


Figure 8. Combined Attributes

Correlation between rating and combined attributes using formula (shelf placement * (sugars + 0.01) * (calories + 0.01))
(0.01 is added to account for 0 calories and 0 grams of sugars which would result in 0 if it weren't there) attributes have different weights so it could be unrepresentative and could present a problem



Analysis Type 1
statistical:Inferential analysis

We want to use **correlation** analysis to **understand the relationship** between the dependent variable – consumer ratings – and the independent variables – sugar content, calories, and shelf placement – allowing us to determine the strength and direction of the linear relationship between the two. In order to calculate a relationship between variables, we could calculate the Pearson correlation coefficient between consumer ratings and sugar content, calories, and shelf placement separately. A positive correlation coefficient indicates a positive linear relationship, meaning that as one variable increases, the other variable tends to increase as well. Oppositely, a negative correlation coefficient indicates a negative linear relationship, meaning that as one variable increases, the other variable tends to decrease.
For example, we used a scatter plot to visualize the linear relationship between consumer ratings, which we picked to be the dependent variable, and sugar content, which we picked to be the independent variable. In figure 3, it shows that the ratings are higher amongst cereals with lower sugar content. The visualization allows us to visually inspect whether there is a linear relationship between the two variables, as well as allows us to identify any potential outliers. We then computed a quantifiable value (Pearson correlation) to see whether there is a significant relationship (< 0.05) between sugar content and consumer ratings in the population of all cereals, based on the sample cereals. If we get a statistically significant computational result, we could conclude there to be a moderate or strong negative correlation/relationship between sugar content and consumer ratings.
We could do this for all independent variables and after we have determined which independent variable has the strongest linear relationship with consumer ratings, we could then perform a **multivariate regression analysis** to determine the combined influence of all independent variables on the dependent variable.

Analysis Type 2
Predictive: regression analysis

To perform the regression analysis, we would first split our dataset into a training set and testing set. The training set would be used to train our regression model, and is split into two subsets, containing a smaller training set and a validation set. The validation set allows us to make sure our model is accurate enough to use, and avoid overfitting for our analysis, while the testing set would be used to evaluate the accuracy of our model.
We would then select the appropriate regression model based on the characteristics of our dataset. In our case, since we are trying to predict a continuous dependent variable – cereal rating – we would use a linear regression model. After selecting the appropriate model, we would fit the model to our training data and evaluate the model's performance using various metrics. In our linear regression analysis, we would use the RMSE cost function – root mean squared error cost function. This measures the average squared difference between the predicted values and the actual values. We chose linear regression as it is convex, meaning that it only has one minimum and the optimization algorithm is guaranteed to find the global minimum. Additionally, the RMSE cost function is sensitive to large errors, meaning that the optimization algorithm will try to minimize the impact of outliers in the data.

Discussion

Pitfalls: The pitfall was that the data was not normalized prior to coming up with the “combined attributes” graph. In order to prevent the “sugars” and “calories” column from becoming “0” when multiplying the columns together, 0.01 was added to both columns. This gave the columns different “weights” which are not representative of finding an actual correlation. To address this issue, there would have to be a normalization function applied to the columns, giving each attribute an equal weight. A societal implication is that people will believe one cereal may be healthier than the other, or may choose a cereal based on the rows if the audience is unaware of the unnormalized weighting. The audience must be informed of the formula used to come up with the weighting.

Limitations, bias, and confounds: Limitations and potential confounds in our analysis exist, for example, our dataset from Kaggle may not be representative of the broader population of cereal consumers, as it only includes a limited number of individual ratings. The ratings themselves may be subject to bias, as they are self-reported and may be influenced by factors like brand loyalty, affordability, and individual taste preferences. To make our dataset more representative of the whole population, we should obtain a larger sample including more brands and consumer ratings of different cereals. In addition, to reduce bias about ratings, we could use different forms of data collection, such as experiments or interviews. Although, working to make our dataset less biased and more representative would consume time, energy, and money out of our budget. We could use a more sophisticated statistical analysis, such as hierarchical linear modeling to control for potential confounding variables and examine the unique contributions of each variable to consumer ratings.

Ethical and social implications: From ethical perspectives, we should consider not to expose personal information while collecting data. In addition to the brief suggestion of how to potentially address ethical issues in part one of the previous section, it is important to think about how the impact of our exposure of people’s PII and their preferences on cereals, may lead to targeted advertising for them. We should also be concerned about the public health issue and consumer behavior, as predicting and inferring the link between high sugar and calorie content and higher consumer ratings of cereals, may contribute to unhealthy dietary habits, particularly amongst children who are more susceptible to marketing messages. We should think about how to not deceive consumers when selling cereals and how to make healthier cereals that also cater to consumers’ preferences and taste. It is important for cereal manufacturers to be transparent about the nutritional content of their products and for regulatory agencies to implement guidelines for marketing towards children. The most important issue we want to address besides the ethical data part is the health cereal problem.

Group Participation

The group contributed by trying to find the best dataset which would later be analyzed for cereal. Marlon contributed by editing and manipulating the dataset in Python and Panda's; he also helped come up with the visualizations/graphs used in the project. Courtney contributed to editing, revising, and writing parts of the sections of the final project. Joanna wrote the background, analysis, and data collection sections.