**Project Report GitHub URL:**

<https://github.com/akabir93/UCDPA_AdnanKabir.git>

**Abstract:**

Analysis of a CSV dataset relating to cereals and their nutritional information and an SQL dataset relating to football players and their stats. The project involves presenting the information within the datasets in a clearer way as well as making new insights based on the existing data within the datasets.

**Introduction:**

I chose this use case as I want to explore how to connect different groups of data together or apply calculations to existing data in order to be able to reach new insights.

**Dataset:**

Cereal csv : <https://www.kaggle.com/crawford/80-cereals>

Football sqlite : <https://www.kaggle.com/hugomathien/soccer>

The Cereal database features data on 80 cereals, from nutritional value to their customer rating and what shelf they’re on in supermarkets. The Football database features footballer and match information collected from over 25,000 matches between 2008 to 2016.

I chose these two datasets due to their differences; one is a smaller dataset with very defined variables whereas the other dataset is much larger with many more variables and values involved. This would allow more varied forms of analysis to be carried out.

**Implementation Process:**

Under the Python file “1. Cereal – Visualising Data” I imported the Pandas, Matplotlib and Seaborn packages, read CSV file and then printed the number of rows, columns as well as the list of table headers used in the dataset. This combined with a histogram of all the data within the dataset being printed gave a feel for what the contents of the dataset were. I then carried out a check to see if there were any null/missing values within the dataset using data.isna().any(), and there wasn’t any. I then queried the relationship between calorie content in a cereal and the customer ratings it got through a scatter plot, assuming that there was going to be correlation between the two which could be illustrated. I then used a countplot to estimate the average high caloric content amongst the cereals available. I produced a heatmap which is another method of illustrating correlation between values, and then I used a boxplot afterwards to check if there were any outliers/anomalies in the data that could throw results off.

Under the Python file “2. Cereal – Analysis” I imported the Pandas, Numpy and Regex packages and read CSV. I had recently read that in supermarkets, items sold on middle shelves attract more customers than items on top or bottom shelves do and sell better1, and wanted to see if I could set up a query identifying items sold on which shelf using this data. Using Regex cereals on different shelves were separated to then store under separate variables. I then set up a loop which would output how many cereals were on the middle shelf and how many were on other shelves.

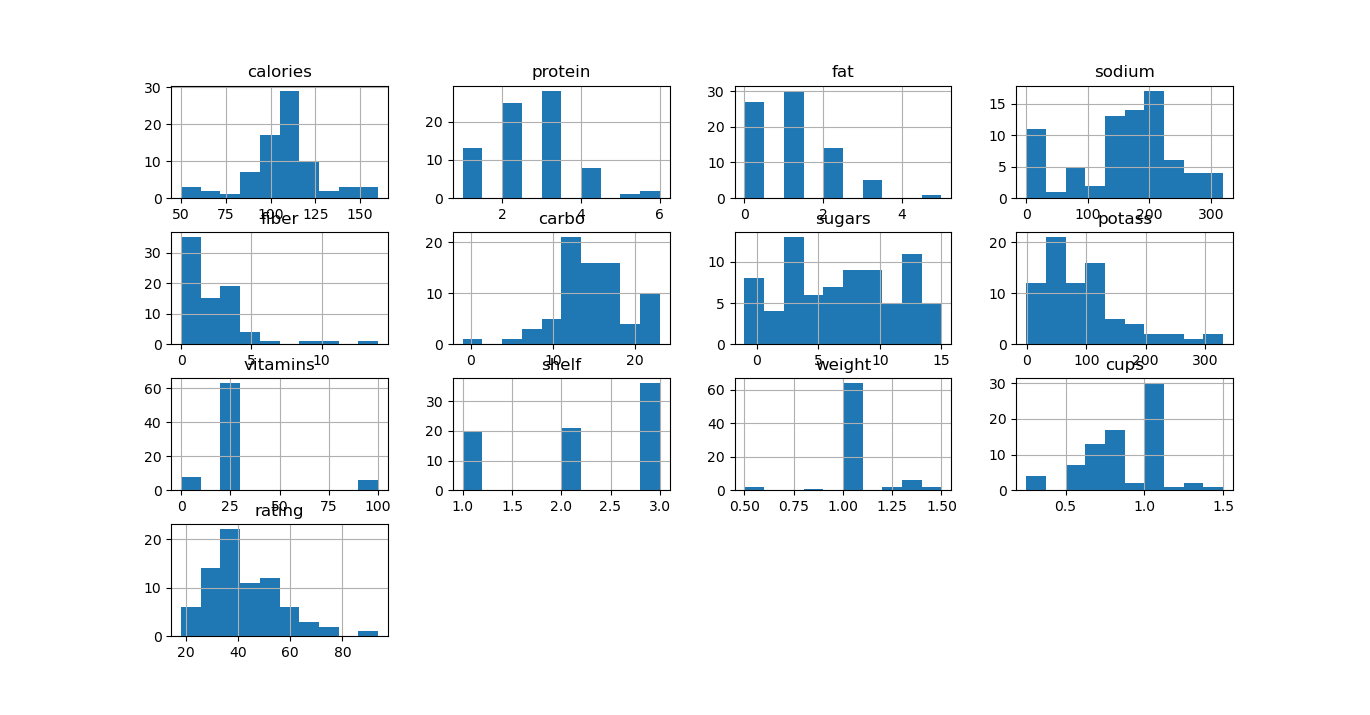
I also recently read that how nutritional value is displayed is often misleading and that in particular the impactful caloric intake can be higher than intended, and one such way of trying to more accurate calculate caloric intake was by working out the “total derived calorie” figure2 by multiplying protein grams by four, carbohydrate grams by four and fat grams by nine. I wanted to estimate a similar value for these cereals to then add onto the dataset. To do this, I stored the dataset variables I identified earlier as lists. I used a list here instead of a dictionary due to wanting to retain the order of the values under each variable. I printed the contents of these lists and converted them to numpy array, as I wanted to carry out mathematical calculations on these values. I carried out the “4-4-9” calculation mentioned above, added the three arrays together and stored it under a single TotalDerivedCalories variable. I then converted this variable to a dataframe and concatenated it with the cereal dataset dataframe. I opted to concatenate instead of appending due to the former being faster especially for smaller dataframes and I wanted to create an entirely new list instead of just modifying the existing one.

Under the Python file “3. Cereal – Machine Learning”, I imported the pandas, numpy, matplotlib and sklearn packages and read the CSV. I wanted to query what nutritional features of a cereal impact it’s rating, and to illustrate this through linear regression showing the relationship between rating and an identified nutritional feature/s. I began by splitting the database into training and test sets to minimize any result discrepancies. I then printed out the nutritional features values of the cereals and decided that I’d focus my efforts on sugars, fiber and proteins as being major factors that impact rating. I plotted a scatter plot showing how present they are at lower ratings and found that sugars stood out. I produced a linear regression with sugar as the independent variable being judged against the rating and plotted the results in a chart.

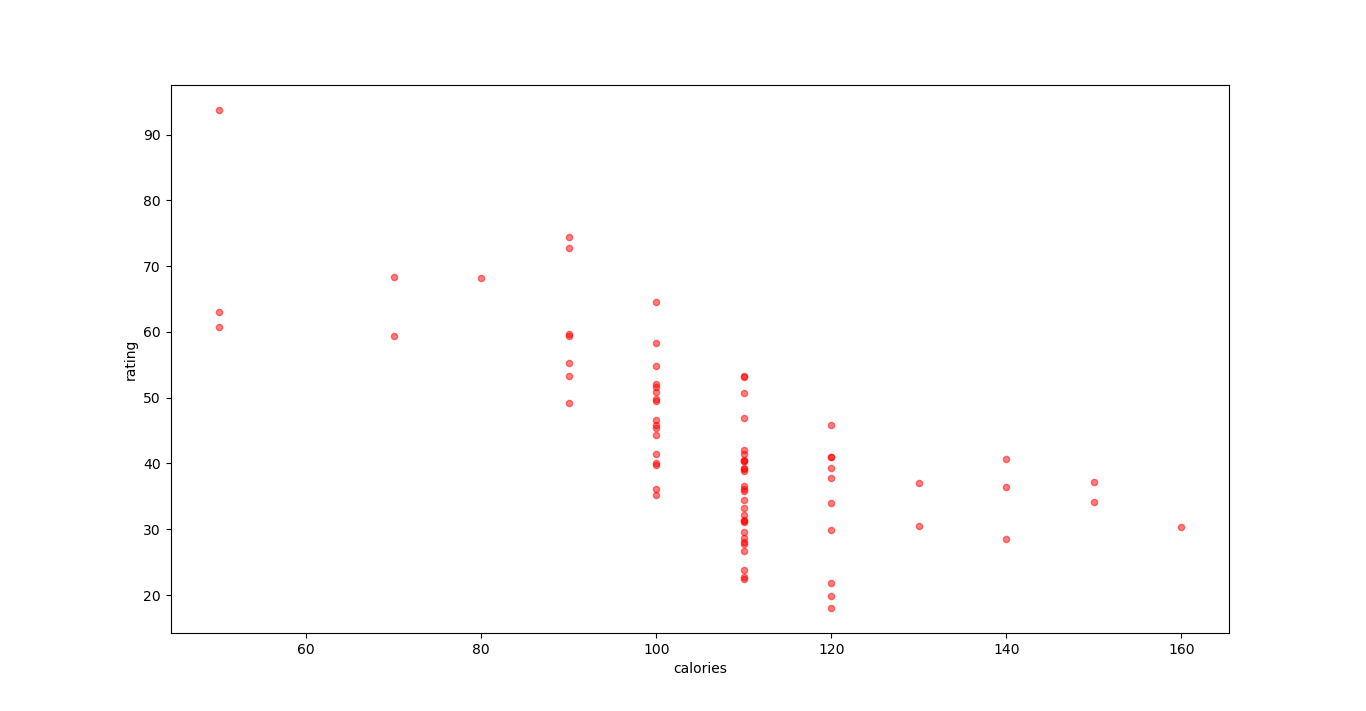
Under the Python file “4. Football – Visualising Data”, I imported the pandas, matplotlib and sqlalchemy and connected to the SQL file by creating a relational database. In order to carry out any further actions on this dataset, I needed to be able to view its contents. I printed the table names, stored each table name as a variable which then allowed me to print the contents of each table name, and I could now view the different values I could query further on. I could now do things such as plot a histogram querying the number of height values within the dataset. I also created a basic function querying what percentage of footballers under a specified height played in the FIFA League between 2008 and 2016.

Under the Python file “5. Football Analysis”, I imported the pandas and sqlalchemy packages and connected to the SQL file by creating a relational database. I wanted to query footballer shot accuracy, and to do I stored the player attributes values (which I was able to identify due to the aforementioned data I printed out in the Python file “4. Football – Visualising Data”) to a variable. From this variable, I extracted the short\_passing, free\_kick\_accuracy and long\_passing values which I would use to calculate average accuracies. I also printed who the best footballer under those individual metrics were. I then calculated an average and then divided the results into right-footed and left-footed outputs, as I wanted to compare the accuracy stats between the two.

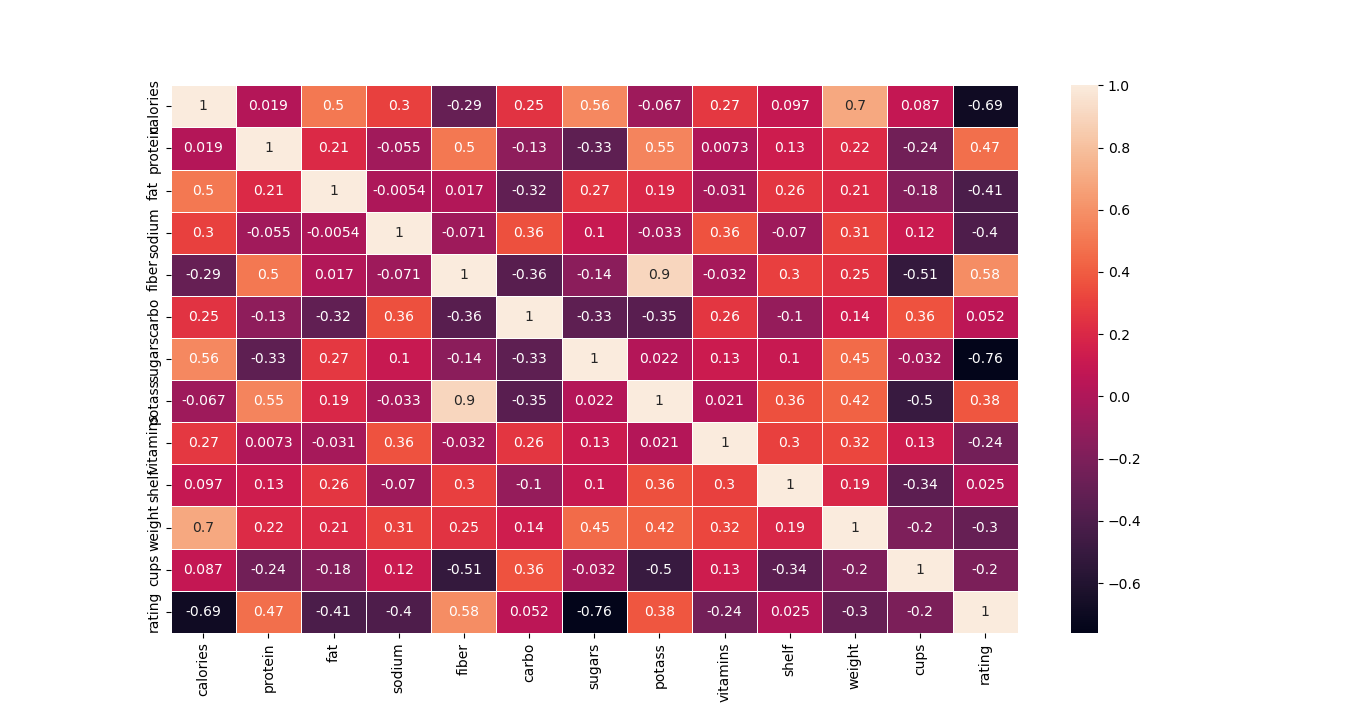
**Results:**



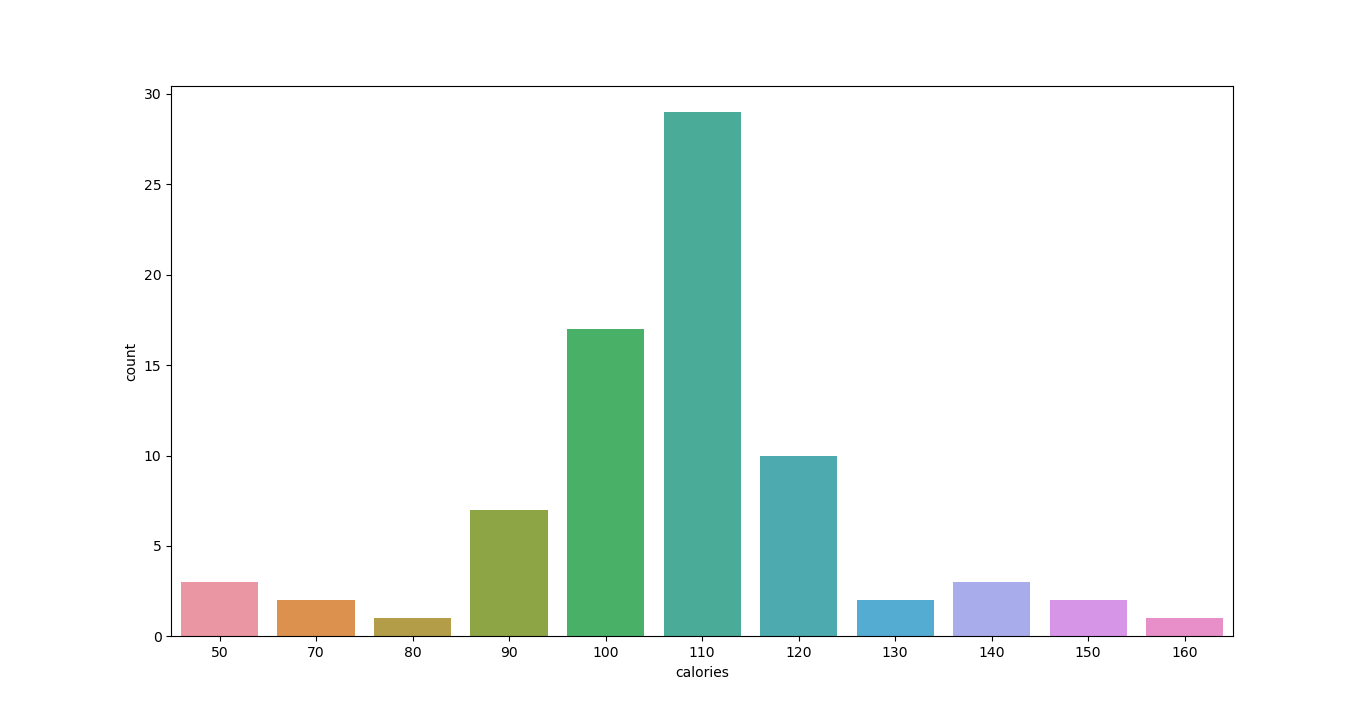
Just a histogram displaying the raw data in the Cereal dataset columns.

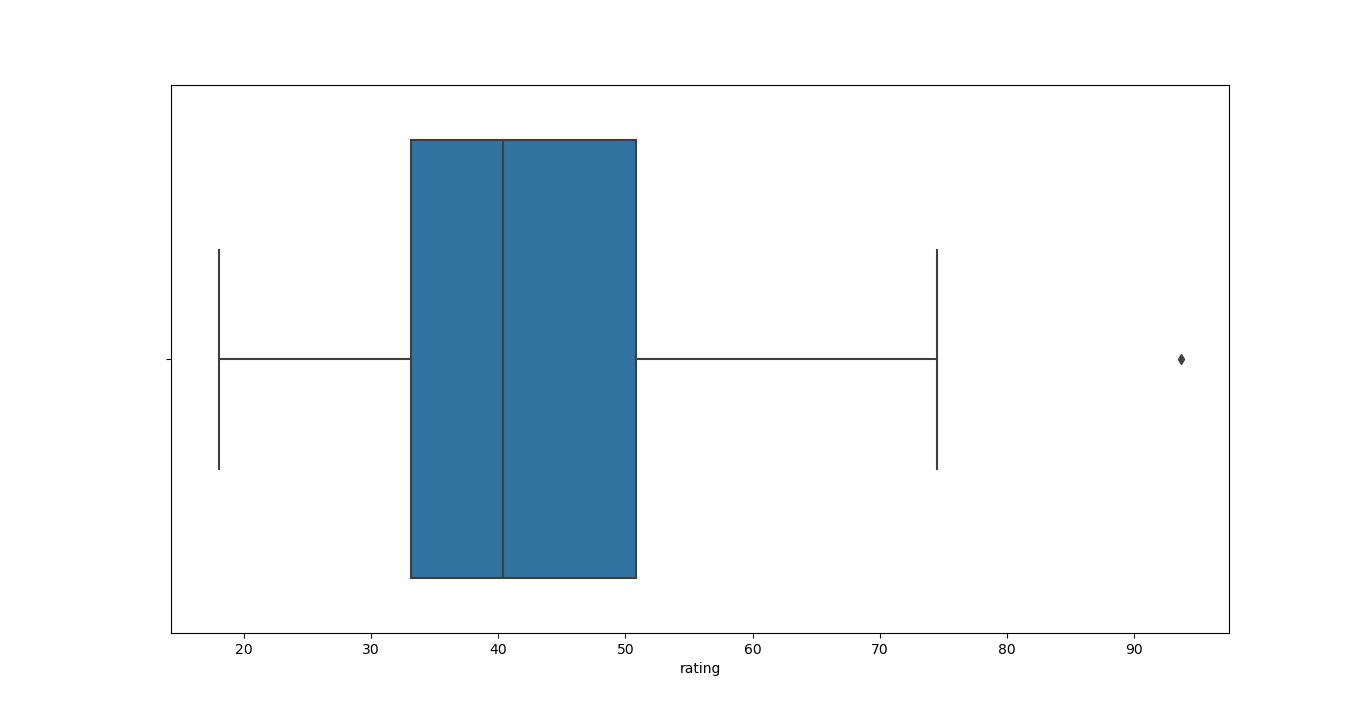


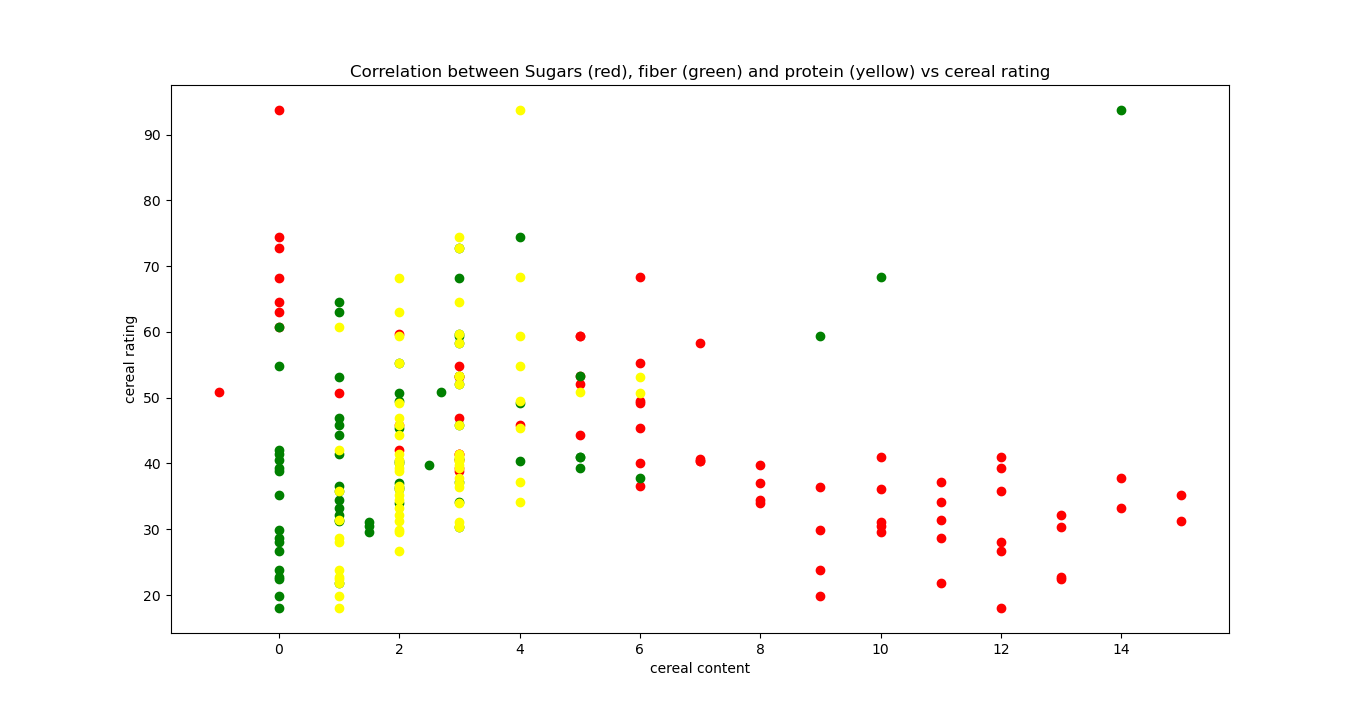
Plotting the level of calories against customer ratings within the Cereal dataset. Negative correlation showing that as calories number increases the ratings decreased.



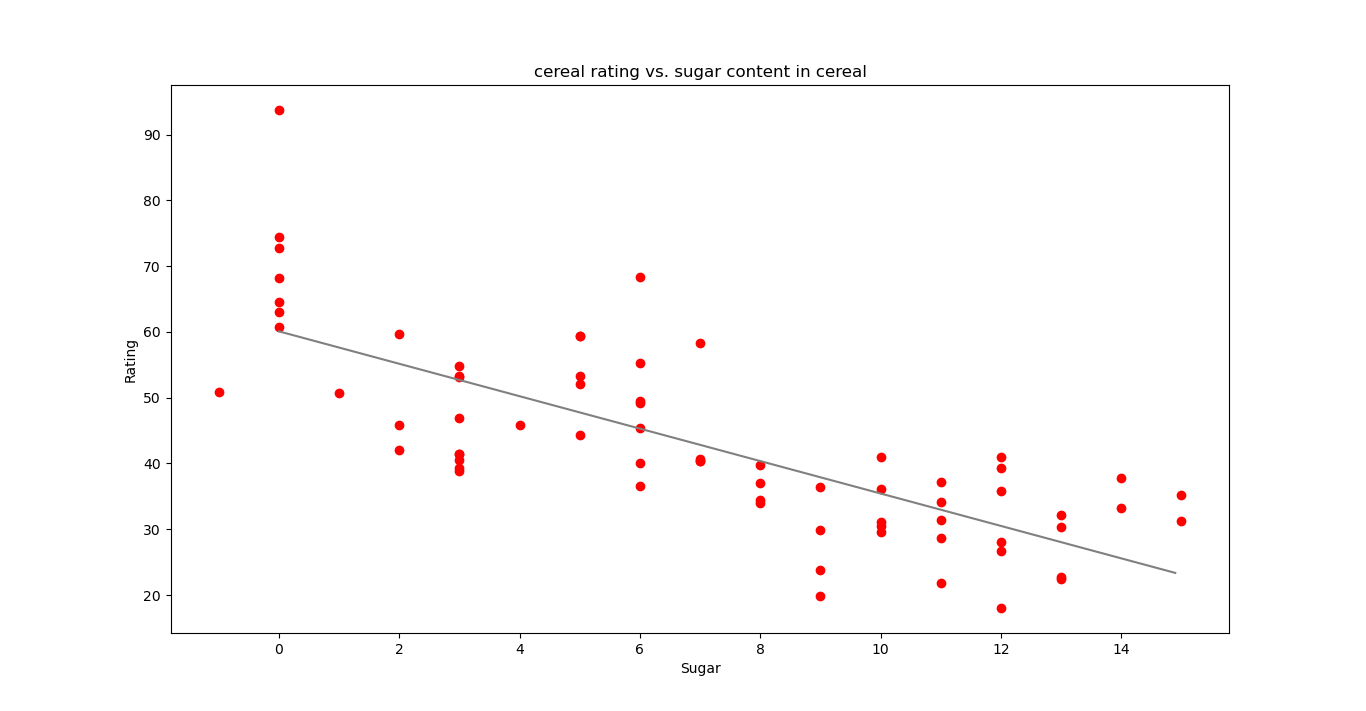
A heatmap showing existing correlation between all Cereal dataset variables. Displays positive correlation between rating and potassium, fibre and protein. Also displays visible negative correlation between rating and fat, sodium and sugars.

A countplot showing how many times cereals with a certain number of calories are present out of the entire sample of cereals.

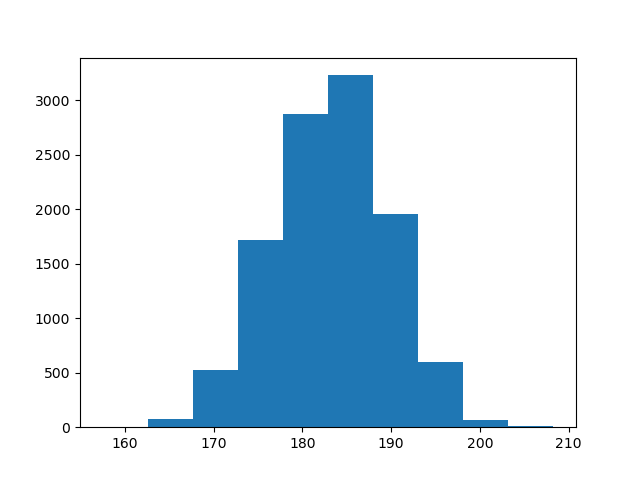
A boxplot relating to the spread of ratings values within the Cereal dataset. Long whiskers indicate a wide array of results, with more variance present in the upper quartile



A scatter plot displaying correlation between sugars, fiber and protein vs the rating in the Cereal dataset. Negative correlation present for sugars.



A scatterplot showing the correlation between sugar and rating within the Cereal database again, this time with that relationship illustrated by a straight line calculated through a linear regression model.



A countplot showing the spread of repeating height values within the Football dataset.

**Insights:**

* Higher sugar and calories nutritional values can result in lower overall customer approval
* Higher potassium, fibre and protein values can result in higher overall customer approval
* Having a lower calories count doesn’t necessarily mean that the Total Derived Calories count will be low
* Less than 40% of players in the FIFA league between 2008 to 2016 were below 6’0 (182cm) tall
* The footballer in the FIFA league between 2008 and 2016 with the most accurate long and short passing was Andrea Pirlo. The player with the most accurate free kicks during this time was Juninho Pernambucano however.
* Left-footed footballers statistically have greater accuracy than right-footed ones

**References:**

1 <https://www.livescience.com/15863-choices-middle-preference.html>

2 <https://nutribase.com/c449.html>