**FORMAN CHRISTIAN COLLEGE (A CHARTERED UNIVERSITY)**

****

**COMP 455 – Data Mining and Warehousing**

**FALL 2022**

**Assignment #1**

**STATISTICAL REPRESENTATION OF DATA**

**Gulraiz Noor Bari (231-525536)**

**Question 1:**

**Calculate Statistical Measures: Find all the statistical measures (Mean Median, Mode, Five number summary) for each of the attributes. If the attribute is categorical just find the mode. Write your report with screenshot of code and output.**

**Mean:** The average of a set of values

**Median:** The middle value when the dataset is arranged in ascending order

**Mode:** The highest occurring value

**Quartile 1:** The value under which 25% of data points are found when they are arranged in ascending order

**Quartile 3:** The value under which 75% of data points are found when they are arranged in ascending order

**Majority of the data points of a attribute lie between Quartiles 1 and 3.**

1. **Graphical user interface, text, application

   Description automatically generatedmfr (Manufacturer):**

The mode of mfr column is K.

1. **Graphical user interface, text, application

   Description automatically generatedType:**

The mode of type column is C.

1. **Calories:**

**Code**

**Text

Description automatically generated**

**Output**

Graphical user interface, text

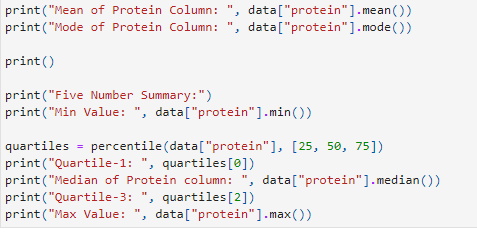
Description automatically generated

The mode/highest occurring value of the calories column is 110.

Quartiles 1 and 3 show that majority values of calories attribute are in the range 100-110.

1. **Protein:**

**Code**



**Output**

Graphical user interface, text, application, chat or text message

Description automatically generated

The mode/highest occurring value of the protein column is 3.

Quartiles 1 and 3 show that majority values of protein attribute are in the range 2-3

1. **Fat:**

**Code**

**Text, letter

Description automatically generated**

**Output**

**Graphical user interface, text

Description automatically generated**

The mode/highest occurring value of the fat column is 1.

Quartiles 1 and 3 show that majority values of fat attribute are in the range 0-2

1. **Sodium:**

**Code**

**Text

Description automatically generated**

**Output**

**Graphical user interface, text

Description automatically generated**

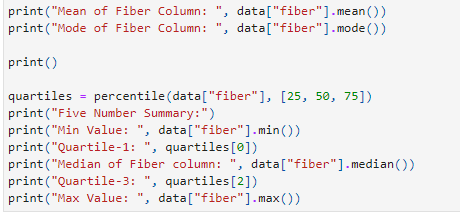
The mode/highest occurring value of the sodium column is 0.

Quartiles 1 and 3 show that majority values of sodium attribute are in the range

130-210

1. **Fiber:**

**Code**

****

**Output**

Graphical user interface, text

Description automatically generated The mode/highest occurring value of the fiber column is 0.

Quartiles 1 and 3 show that majority values of fiber attribute are in the range 1-3

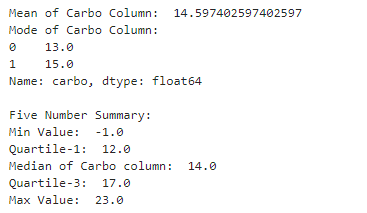
1. **Carbohydrates:**

**Code**

**Text, letter

Description automatically generated**

**Output**

****

The mode/highest occurring values of the carbo column are 13 and 15.

Quartiles 1 and 3 show that majority values of fiber attribute are in the

range 12-17

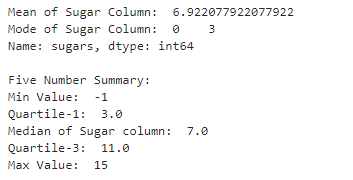
1. **Sugar:**

**Code**

**Text

Description automatically generated**

**Output**

****

The mode/highest occurring value of sugar attribute is 3.

Quartiles 1 and 3 show that majority values of sugar attribute are in the range 3-11

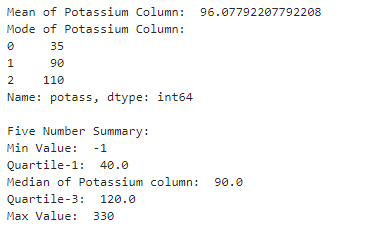
1. **Potassium:**

**Code**

**Text

Description automatically generated**

**Output**

****

The mode/highest occurring values of potassium attribute are 35, 90 and 110.

Quartiles 1 and 3 show that majority values of potassium attribute are in the

range 40-120

1. **Vitamins:**

**Code**

**Text

Description automatically generated**

**Output**

**Text

Description automatically generated**

The mode/highest occurring values of vitamins attribute are 25.

Quartiles 1 and 3 show that majority values of vitamin attribute are at a single point 25.

1. **Shelf:**

**Code**

****

**Output**



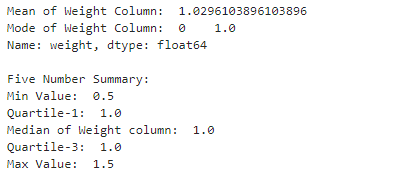
The mode/highest occurring value of shelf attribute is 3.

1. **Weight:**

**Text

Description automatically generatedCode**

**Output**

****

The mode/highest occurring value of weight attribute is 1.

Quartiles 1 and 3 show that majority values of weight attribute are at a single

point 1

1. **Cups:**

**Code**

**Text

Description automatically generated**

**Output**

**Graphical user interface, text

Description automatically generated**

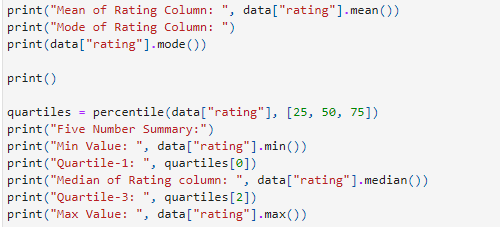
The mode/highest occurring value of cups attribute is 1.

Quartiles 1 and 3 show that majority values of cups attribute are in the

range 0.67-1

1. **Rating:**

**Code**

****

**Output**

**Text

Description automatically generated with medium confidence**

There are no values in rating attribute occurring more than once, so there is no mode.

Quartiles 1 and 3 show that majority values of cups attribute are in the

range 33.174094-50.828392

**Question 2:**

1. **Is there a relation between sugars, calories, carbs, and fat? (Scatter plot for the above-mentioned attributes)**

**Code**

**Chart

Description automatically generated**

**Output**

**Chart, waterfall chart

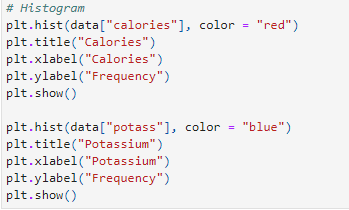
Description automatically generated**

**Analysis**

* Calories and sugar have positive correlation and vice versa.
* Calories and carbohydrates have positive correlation and vice versa.
* Fat has no correlation with any other attribute.
* Sugar and carbohydrates have negative correlation and vice versa.

1. **How are calories and potassium distributed? (Seperate histograms of these attributes with their frequencies)**

**Code**

****

**Output**

**Chart, histogram

Description automatically generatedChart, histogram

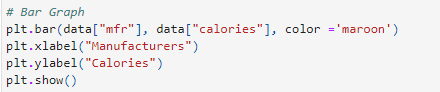
Description automatically generated**

**Analysis**

* Calories is distributed as Normal Distribution
* Potassium is distributed as Dog Food Distribution

1. **Which manufacturers produce cereal with highest calories? (Bar Graph of the concerned attributes)**

**Code**

****

**Output**

**Chart, bar chart

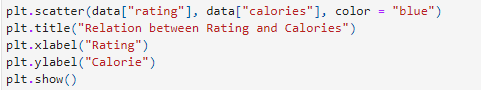
Description automatically generated**

**Analysis**

K, R and G (in descending order) are the manufacturers that produce cereals with highest calories.

1. **How does rating compare to calorie count? (Scatter plot of the two attributes)**

**Code**

****

**Output**

**Chart, scatter chart

Description automatically generated**

**Analysis**

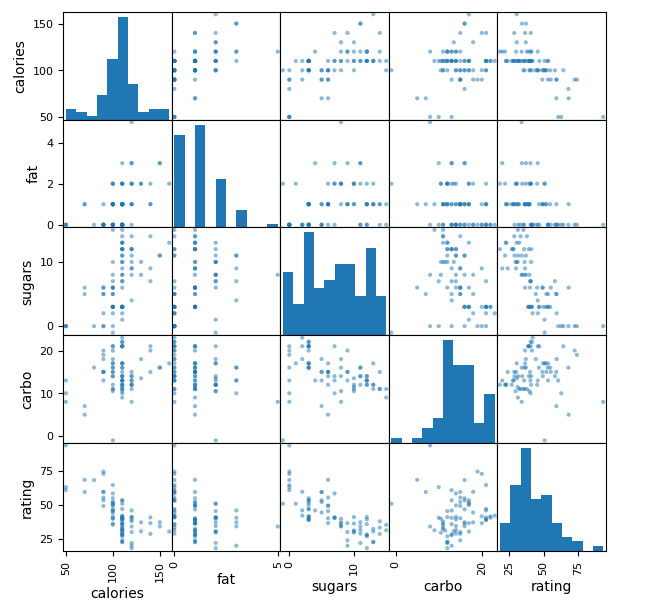
Rating and calories are negatively correlated. If calories increase than the rating decrease and vice versa.

1. **Which nutrients are essential for a good rating for a cereal? (Scatter matrix for nutrients and ratings of the attributes)**

**Code**

****

**Output**

****

**Analysis**

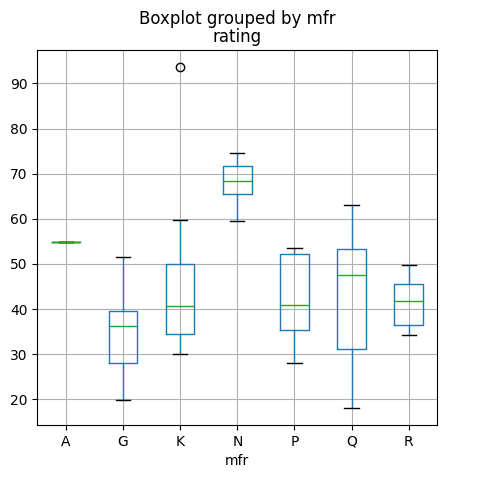
Carbohydrates are essential for a good rating for a cereal.

1. **Is there a relation between manufacturer and rating? (Box plots all manufacturers and ratings)**

**Code**

****

**Output**

****

**Analysis**

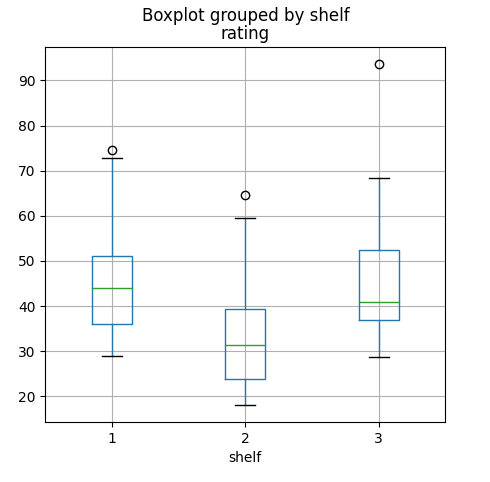
* Manufacturer N has the highest rating
* Manufacturer G has the overall low rating
* Other four manufacturers have average rating between 30-55

1. **Is there a relation between shelf number and rating? (Box plots all shelves and ratings)**

**Code**

****

**Output**

****

**Analysis**

* Items in shelf-1 have average rating between 35-52
* Items in shelf-2 have average rating between 23-40
* Items in shelf-3 have average rating between 36-53