Homework 2

**Breakfast Cereals**​. Use the data *cereal.xlsx* to explore and summarize the data as follows: (note that a few records contain missing values; since there are just a few, a simple solution is to remove them first.)

a. Which variables are quantitative/numerical? Which are ordinal? Which are nominal? (10 point)

b. Create a table with the average, median, min, max, and standard deviation for each of the quantitative variables. This can be done through Excel’s functions or Excel’s *D*​*ata* → *Data Analysis* → *Descriptive Statistics*​ menu. (5 point)

c. Use RapidMiner to plot a histogram for each of the quantitative variables (5 pt). Based on the histograms and summary statistics, answer the following questions:

i. Which variables have the largest variability? (5 pt)

ii. Which variables seem skewed? (5 pt)

iii. Are there any values that seem extreme? (5 pt)

d. Plot a side­by­side boxplot comparing the calories in hot vs. cold cereals. What does this plot show us? (5 pt)

e. Plot a side­by­side boxplot of consumer rating as a function of the shelf height. If we were to predict consumer rating from shelf height, does it appear that we need to keep all three categories of shelf height? (5 pt)

f. Compute the correlation table for the quantitative variable

i. Which pair of variables is most strongly correlated? (2.5 pt)

ii. How can we reduce the number of variables based on these correlations? (2.5 pt)

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**ANSWERS:**

**a. *Which variables are quantitative/numerical? Which are ordinal? Which are nominal? (10 point)***

|  |  |  |  |
| --- | --- | --- | --- |
| **Order** | **Variable** | **Type** | **Description** |
| 1 | Name | Nominal | Name of cereal |
| 2 | mfr | Nominal | Manufacturer (A = American Home Food Products, G = General Mills, etc.) |
| 3 | type | Nominal | Type of cereal (cold or hot) |
| 4 | calories | Quantitative | Calories per serving |
| 5 | protein | Quantitative | Grams of protein |
| 6 | fat | Quantitative | Grams of fat |
| 7 | sodium | Quantitative | Milligrams of sodium |
| 8 | fiber | Quantitative | Grams of dietary fiber |
| 9 | carbo | Quantitative | Grams of complex carbohydrates |
| 10 | sugars | Quantitative | Grams of sugars |
| 11 | potass | Quantitative | Milligrams of potassium |
| 12 | vitamins | Ordinal | Vitamins and minerals (0, 25, or 100% of FDA recommendation) |
| 13 | shelf | Ordinal | Display shelf position (1 = bottom, 2 = middle, 3 = top) |
| 14 | weight | Quantitative | Weight in ounces of one serving |
| 15 | cups | Quantitative | Number of cups in one serving |
| 16 | rating | Quantitative | Consumer Reports rating of the cereal |

**-------------------------------------------------------------------------------------------------------------**

**b. *Create a table with the average, median, min, max, and standard deviation for each of the quantitative variables. This can be done through Excel’s functions or Excel’s Data → Data Analysis → Descriptive Statistics menu. (5 point)***

Data Analysis wasn’t accessible at first and so I had to do the following to make it available:

File → Options → Add-ins → Manage: Excel Add-ins, and check Analysis ToolPak  
  
To use Descriptive Analysis, I had to move the quantitative variable next each other as the data needed to be contiguous for the Data Analysis tool to work. This created a new sheet with a summary statistics of all the quantitative rows.

**A screen shot of a computer

Description automatically generated**

**A close-up of a computer screen

Description automatically generated**

After removing the other statistical values, we have the answer for the question below.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ***calories*** |  | ***protein*** |  | ***fat*** |  | ***sodium*** |  |
| Mean | 107.027 | Mean | 2.513514 | Mean | 1 | Mean | 162.3649 |
| Median | 110 | Median | 2.5 | Median | 1 | Median | 180 |
| Standard Deviation | 19.84389 | Standard Deviation | 1.075802 | Standard Deviation | 1.006826 | Standard Deviation | 82.76979 |
| Minimum | 50 | Minimum | 1 | Minimum | 0 | Minimum | 0 |
| Maximum | 160 | Maximum | 6 | Maximum | 5 | Maximum | 320 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ***fiber*** |  | ***carbo*** |  | ***sugars*** |  | ***potass*** |  |
| Mean | 2.175676 | Mean | 14.72973 | Mean | 7.108108 | Mean | 98.51351 |
| Median | 2 | Median | 14.5 | Median | 7 | Median | 90 |
| Standard Deviation | 2.423391 | Standard Deviation | 3.891675 | Standard Deviation | 4.359111 | Standard Deviation | 70.87868 |
| Minimum | 0 | Minimum | 5 | Minimum | 0 | Minimum | 15 |
| Maximum | 14 | Maximum | 23 | Maximum | 15 | Maximum | 330 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ***weight*** |  | ***cups*** |  | ***rating*** |  |
| Mean | 1.030811 | Mean | 0.821622 | Mean | 42.37179 |
| Median | 1 | Median | 0.75 | Median | 40.25309 |
| Standard Deviation | 0.153416 | Standard Deviation | 0.235715 | Standard Deviation | 14.03371 |
| Minimum | 0.5 | Minimum | 0.25 | Minimum | 18.04285 |
| Maximum | 1.5 | Maximum | 1.5 | Maximum | 93.70491 |

**-------------------------------------------------------------------------------------------------------------**

**c. *Use RapidMiner to plot a histogram for each of the quantitative variables (5 pt). Based on the histograms and summary statistics, answer the following questions:***

A screenshot of a computer

Description automatically generated

A screenshot of a graph

Description automatically generated

A screenshot of a computer

Description automatically generated

1. ***Which variables have the largest variability? (5 pt)***

Variability is represented by the standard deviation. The larger the standard deviation, the more variability exists in the data. From the data we can observe high standard deviation from the following variables.

* Sodium: Standard deviation of 82.77
* Potass: Standard deviation of 70.88
* Rating: Standard deviation of 14.03
* Sugars: Standard deviation of 4.36

So we can say that Sodium, Potassium, and Rating have the largest variability.

1. ***Which variables seem skewed? (5 pt)***

For skewness, we compare the mean and median of each variable. If the mean is significantly higher or lower than the median, this suggests skewness.

We can also observe this with the histograms as in a normal distribution, both tails are symmetrical. In left (negative) skew, the left tail is longer, in right (positive) skew, the right tail is longer, and the mean is greater than the mode. This is marked in the histograms to compare the skew and highlighted with **red** boxes.

So we can say **Sodium** is showing *slight negative skew***,** and **Rating and Potass** showing *positive skew*.

***iii. Are there any values that seem extreme? (5 pt)***

We may be able to identify extreme values with maximum and minimum. Using this we can see that Sodium, Potassium and Rating have maximums that are larger than the mean and having high standard deviation can suggest that they have extreme values.

But by looking at the histograms we can see that isolated bars in Fat, Fiber, Weight and Rating which suggests outliers as the rest of the bars are crammed together.  
  
Using the above logic we can say **Fat, Fiber, Weight** and **Rating** have values that seem extreme.

**-------------------------------------------------------------------------------------------------------------**

**d. *Plot a side­by­side boxplot comparing the calories in hot vs. cold cereals. What does this plot show us? (5 pt)***

A graph with blue lines and a rectangle

Description automatically generated

For cold cereals the lower whisker extends to 50 and the upper whisker extends to 160, with a median of 110, indicating the full range of data. The interquartile range, that is the middle 50% of values lie between 100 and 115. There's some variation but no extreme outliers.

For hot cereal values we can observe that all calorie values are exactly 100 with no variability at all. This tells us that there isn’t a good distribution for hot cereals or hot cereals tend to be more consistent.  
  
But since we observed the data before comparing the boxplot, we know this is because only 1 hot cereal entry exits, and the observations made above are due to that.

**-------------------------------------------------------------------------------------------------------------**

**e. *Plot a side¬by¬side boxplot of consumer rating as a function of the shelf height. If we were to predict consumer rating from shelf height, does it appear that we need to keep all three categories of shelf height? (5 pt)***

A diagram of a graph

Description automatically generated with medium confidence

Shelf 3 has a wide range of ratings, from approximately 30 to 90, with a median around 45. The IQR (middle 50% of data) spans from around 40 to 60, indicating a diverse spread of consumer ratings.

Shelf 2 ratings are slightly more concentrated, with a range from around 30 to 75, and a median around 40. It has a narrower IQR compared to Shelf 3.

Shelf 2 has the smallest range, with ratings from approximately 30 to 70 and a median just above 35. The IQR is tighter compared to Shelves 1 and 3, indicating more consistency in ratings for products placed on this shelf.

The differences in ranges, medians, and spread suggests that removing any of the three categories may result in a loss of important information. So, it's mostly beneficial to keep all three shelf heights If we were to predict consumer rating from shelf height.

**-------------------------------------------------------------------------------------------------------------**

**f. *Compute the correlation table for the quantitative variable***

***A screenshot of a graph

Description automatically generated***

***i. Which pair of variables is most strongly correlated? (2.5 pt)***

The pair of variables that are most strongly correlated are potass and fiber.

***ii. How can we reduce the number of variables based on these correlations? (2.5 pt)***

A screenshot of a computer

Description automatically generated

After examining the correlation matrix, we can choose only the attributes with above 0.3 and below -0.3 correlation to the label, so all attributes with correlation between **-0.3** and **+0.3** can be rejected. Then to prevent any multicollinearity we can discard one of the variables of a pair of variables which have a correlation of above **+0.8** or below **-0.8**.

We can see in the image above that using the strategy above reduces the number of variables.