Northeastern University

EECE5642: Data Visualization
Homework1

Rudra Patel
Feb 02, 2022

Answer 1

The dataset that I have chosen is **Breakfast Cereal**. It contains nutritional information for **77 different breakfast cereals**. It was used for the 1993 Statistical Graphics Expositions a challenge data set. The data is from the nutritional labels and is in a **CSV** format.

The **variables** are:
Cereal name
manufacturer (e.g., Kellogg's)
type (cold/hot)
calories (number)
protein (g)
fat (g)
sodium (mg)
dietary fiber (g)
complex carbohydrates (g)
sugars (g)
display shelf (1, 2, or 3, counting from the floor)
potassium (mg)
vitamins and minerals (0, 25, or 100, respectively)
weight (in ounces) of one serving (serving size)
cups per serving.
rating

Manufacturers are represented by their **first initial**: A=American Home Food Products, G=General Mills, K=Kellogg's, N=Nabisco, P=Post, Q=Quaker Oats, R=Ralston Purina.

The data set can be found at this source:
https://perso.telecom-paristech.fr/eagan/class/igr204/datasets

Below Shown is the sample image of the data set being used for the assignment.

cereal.csv — Open with Microsoft Excel

name	mfr	type	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	weight	cups	rating
String	Categorical	Categorical	Int	Int	Int	Int	Float	Float	Int	Int	Int	Int	Float	Float	Float
100% Bran	N	C	70	4	1	130	10	5	6	280	25	3	1	0.33	68.402973
100% Natural Bran	Q	C	120	3	5	15	2	8	8	135	0	3	1	1	33.983679
All-Bran	K	C	70	4	1	260	9	7	5	320	25	3	1	0.33	59.425505
All-Bran with Extra Fiber	K	C	50	4	0	140	14	8	0	330	25	3	1	0.5	93.704912
Almond Delight	R	C	110	2	2	200	1	14	8	-1	25	3	1	0.75	34.384843
Apple Cinnamon Cheerios	G	C	110	2	2	180	1.5	10.5	10	70	25	1	1	0.75	29.509541
Apple Jacks	K	C	110	2	0	125	1	11	14	30	25	2	1	1	33.174094
Basic 4	G	C	130	3	2	210	2	18	8	100	25	3	1.33	0.75	37.038562
Bran Chex	R	C	90	2	1	200	4	15	6	125	25	1	1	0.67	49.120253
Bran Flakes	P	C	90	3	0	210	5	13	5	190	25	3	1	0.67	53.313813
Cap'n'Crunch	Q	C	120	1	2	220	0	12	12	35	25	2	1	0.75	18.042851
Cheerios	G	C	110	6	2	290	2	17	1	105	25	1	1	1.25	50.764999
Cinnamon Toast Crunch	G	C	120	1	3	210	0	13	9	45	25	2	1	0.75	19.823573
Clusters	G	C	110	3	2	140	2	13	7	105	25	3	1	0.5	40.400208
Cocoa Puffs	G	C	110	1	1	180	0	12	13	55	25	2	1	1	22.736446
Corn Chex	R	C	110	2	0	280	0	22	3	25	25	1	1	1	41.445019
Corn Flakes	K	C	100	2	0	290	1	21	2	35	25	1	1	1	45.863324
Corn Pops	K	C	110	1	0	90	1	13	12	20	25	2	1	1	35.782791
Count Chocula	G	C	110	1	1	180	0	12	13	65	25	2	1	1	22.396513
Cracklin' Oat Bran	K	C	110	3	3	140	4	10	7	160	25	3	1	0.5	40.448772
Cream of Wheat (Quick)	N	H	100	3	0	80	1	21	0	-1	0	2	1	1	64.533816
Crispix	K	C	110	2	0	220	1	21	3	30	25	3	1	1	46.895644
Crispy Wheat & Raisins	G	C	100	2	1	140	2	11	10	120	25	3	1	0.75	36.176196
Double Chex	R	C	100	2	0	190	1	18	5	80	25	3	1	0.75	44.330856
Froot Loops	K	C	110	2	1	125	1	11	13	30	25	2	1	1	32.207582
Frosted Flakes	K	C	110	1	0	200	1	14	11	25	25	1	1	0.75	31.435973
Frosted Mini-Wheats	K	C	100	3	0	0	3	14	7	100	25	2	1	0.8	58.345141

Figure 1: Breakfast cereal dataset

Visualizations and Evaluations:

As seen above we have a total of 12 variables which can be used to visualize and find various correlations between these nutritional values. In the following section we will analyze some data using various plots and see what fits best.

1. **Stacked bar chart.**

   A bar chart or bar graph is a chart or graph that uses rectangular bars with heights or lengths proportional to the values they represent to convey categorical data. Stacked bar graphs are used to highlight how a bigger entity is made up of into entities and how each element affects the total amount.

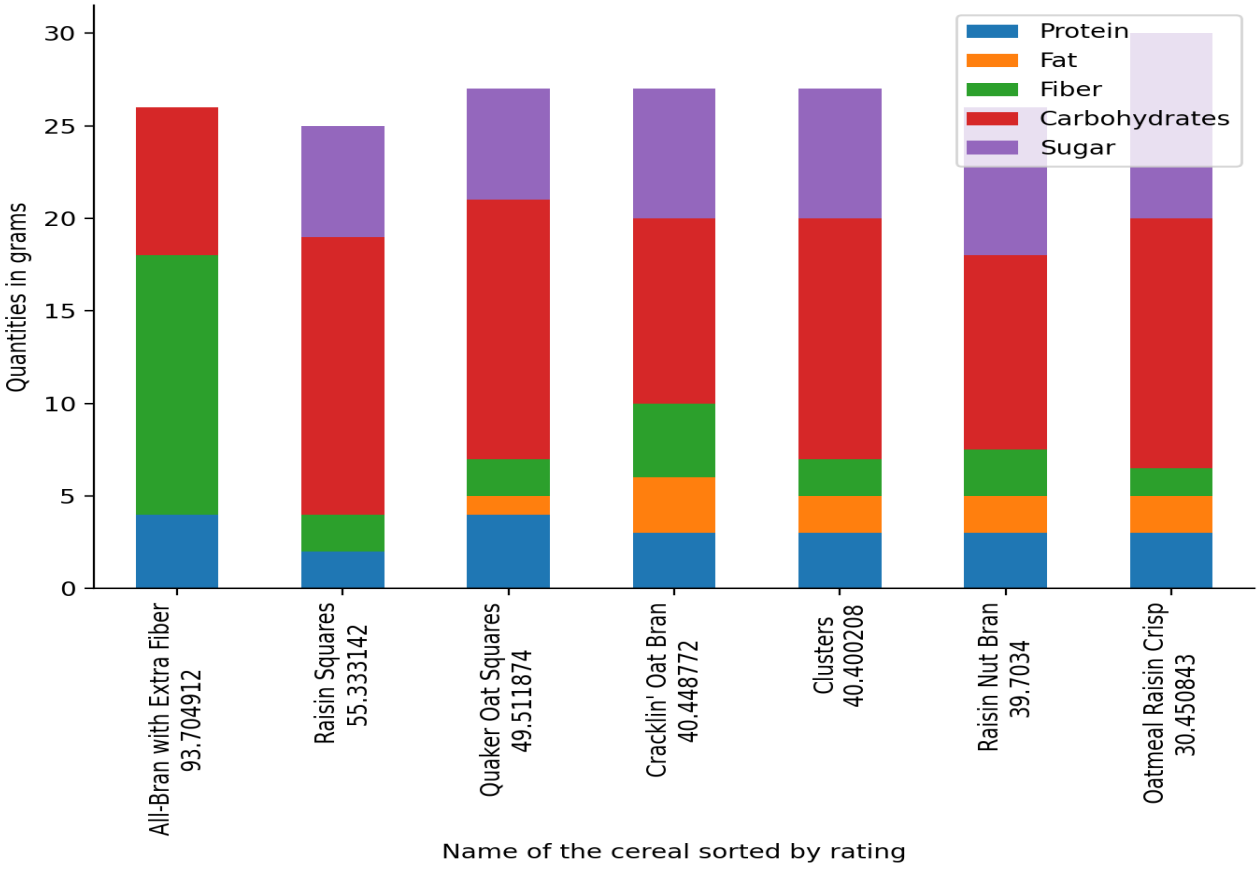


Figure 2: Stacked Bar chart showing different nutrition content for a group of 7 cereals with same serving size sorted by their rating value.

2. Multiple Bar Chart

A multiple bar graph plots the relationship between various data variables. A column in the graph represents each data value. The horizontal (x-axis) lists the classifications of various types of data. Along the vertical (y-axis) the quantity or amount of data is listed. As demonstrated below, the same information is recreated using many bar charts.

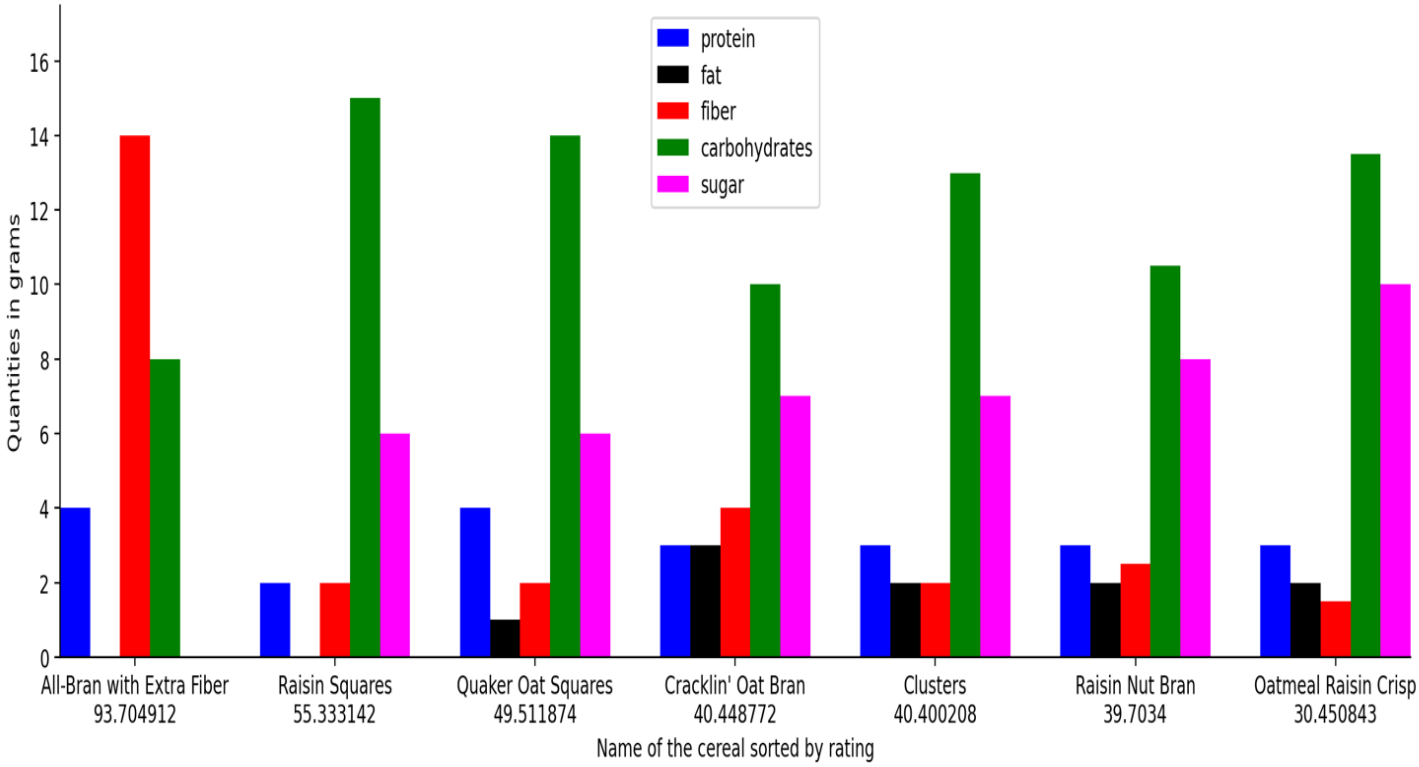


Figure 3: Multiple Bar chart showing different nutrition content for a group of 7 cereals with same serving size sorted by their rating value.

3. Line Chart

A line chart, also known as a line plot, line graph is a style of chart that shows data as a succession of' markers' connected by straight line segments. The same data that we plotted before can be illustrated using a line plot, as illustrated below.

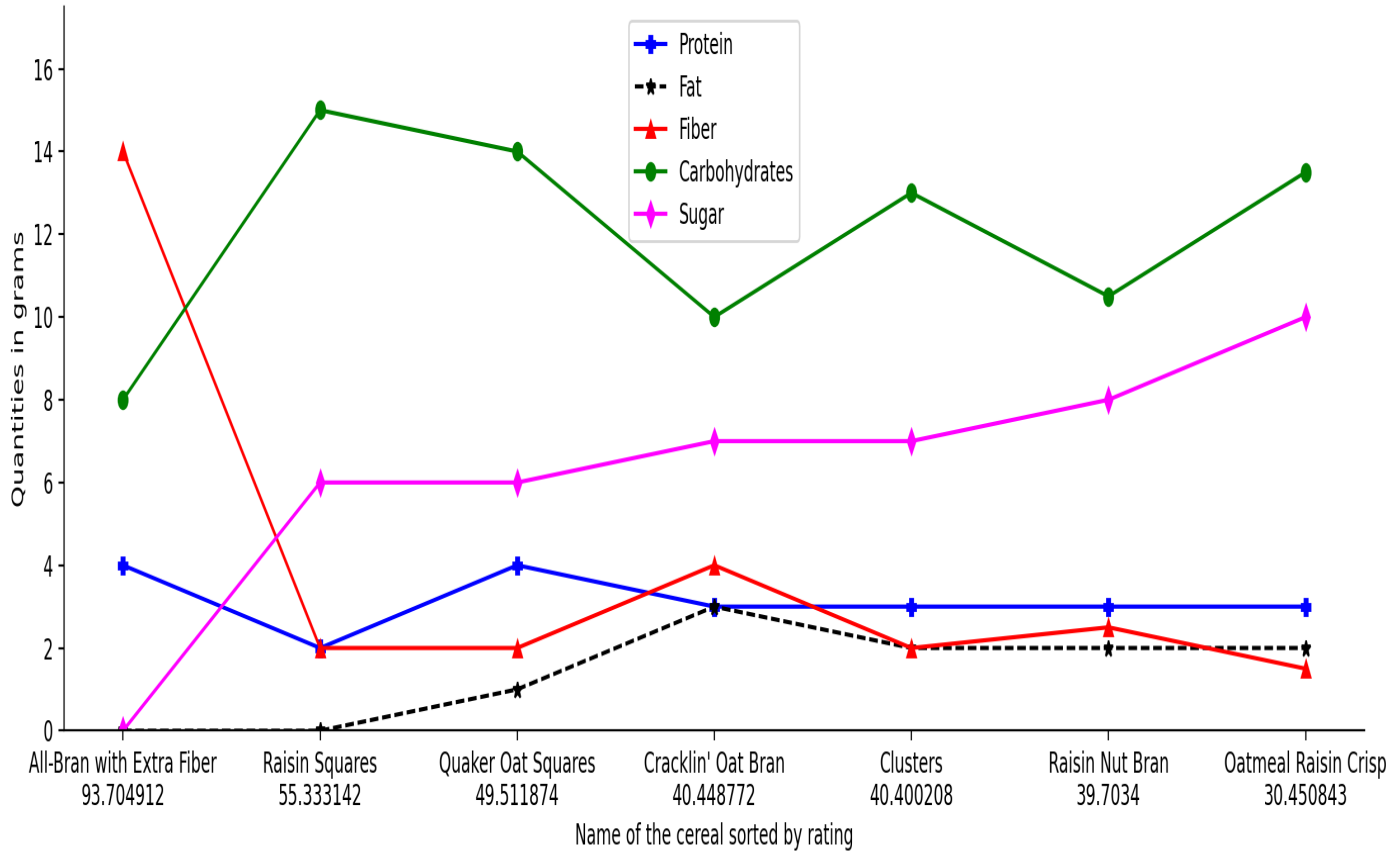


Figure 4: Line chart showing different nutrition content for a group of 7 cereals with same serving size sorted by their rating value.

4. Donut Pie Chart

The donut chart is a special type of pie chart that has a hole in the middle for removing unnecessary redundancies and presents categories as arcs instead of slices. Both make it simple to understand part-to-whole linkages right away. They can't illustrate changes over time, unlike line charts, area charts, column charts, and bar graphs.

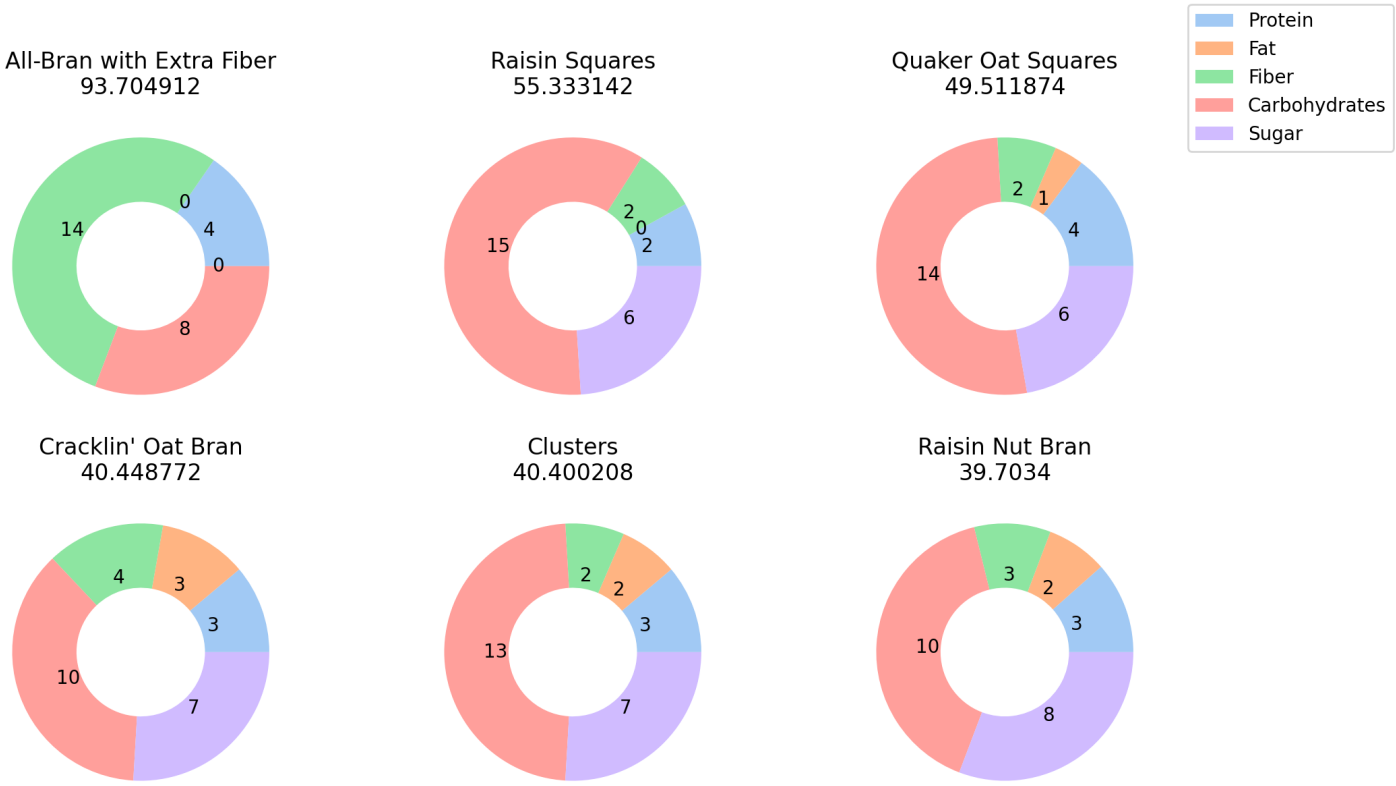


Figure 5: Pie chart showing different nutrition content for a group of 6 cereals with same serving size sorted by their rating value.

Analysis:

All the above produced plots can be evaluated and discussed to find out which ones works the best for this data set.

- The task or goal here is to find out the different nutrition contents in a cereal. To be unbiased about the observations the cereals have been picked in such a manner as they have the exact same serving size which gives us the opportunity to compare them. The contents help us obtain a picture of what factors is the overall rating dependent on.

- The dimension of the data here is 7x7 where the rows depict the 7 different cereal and the 7 columns are for ‘Protein’, ‘Fat’, ‘Fiber’, ‘Carbohydrates’, ‘Sugar’, ‘Serving cup size’ and the ‘rating’ value.

- The data types of the variables are as follows:
  - Name of the cereal: string
  - Protein: int
  - Fat: int
  - Fiber: float
  - Carbohydrates: float
  - Sugar: int
  - Cup size: float
  - Rating: float

- The information depicted by all these visualizations tells us the exact amount of nutrition contents in grams present in each cereal. It further makes it possible for us to know what factor is affecting in a positive or negative manner. Like more amounts of sugar leads to a decrease in rating whereas more amounts of fiber lead to a better rating. This information accounts for an unbiased comparison between the cereal as the serving size for all of them is the exact same. Furthermore, when some manufacturer thinks about creating a new cereal, they can see the amounts of nutrients that their competitors are offering for a serving size and what amounts of each nutrient would lead them to a better rating.

- According to me for this particular task the best chart here would be the Multiple Bar charts. It gives us a clear visualization for the amounts constituted by each variable against the scale. The color coding allows us to distinguish the different variables. There is a con associated with this type of chart making it very difficult to accommodate large sets of data in a single plot frame.

- The second visualization of stacked bars does account for this disadvantage where we can have more observations on the horizontal scale leading to a greater visualization dataset. However, the disadvantage here is that we need to calculate and evaluate for each observation since it shows us a continuous reading so to get the amount of second variable, we need to subtract the total of that variable from the previous reading making it difficult to visualize immediately.

- For the third type we have line charts. Looking at it we can say they look more congested and scattered as compared to the other two plots. There is a small advantage for this type of technique which connects the markers for the variables among different cereals which gives us a trend of increasing or decreasing quantities.

- The fourth and final chart is a donut pie chart. The pro we have here is that we have a clear representation for the share of all the variables for a cereal which leads to the consequent rating whereas the con is we need to look back and forth when we are comparing between two or more types.

- A suggestion here would be to show the trends of increasing and decreasing values of these variables and the effect it has on the rating of the cereal. A few other plots have been produced which all gives us some insights and comparisons which are shown in the following section.

An attempt is made to remove all redundant entities as discussed in the lecture like removing the unwanted grid lines, removing the unrequired border frames, keeping minimum ticks on both the axis. Labelling of the axis have been clearly made which eliminates the need of a title for the plot thereby further reducing ink which is a good visualization principle.

5. Box Plot

A boxplot is a standardized method of depicting data distributions using a five-number summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum"). It can give you information about your outliers and their values.

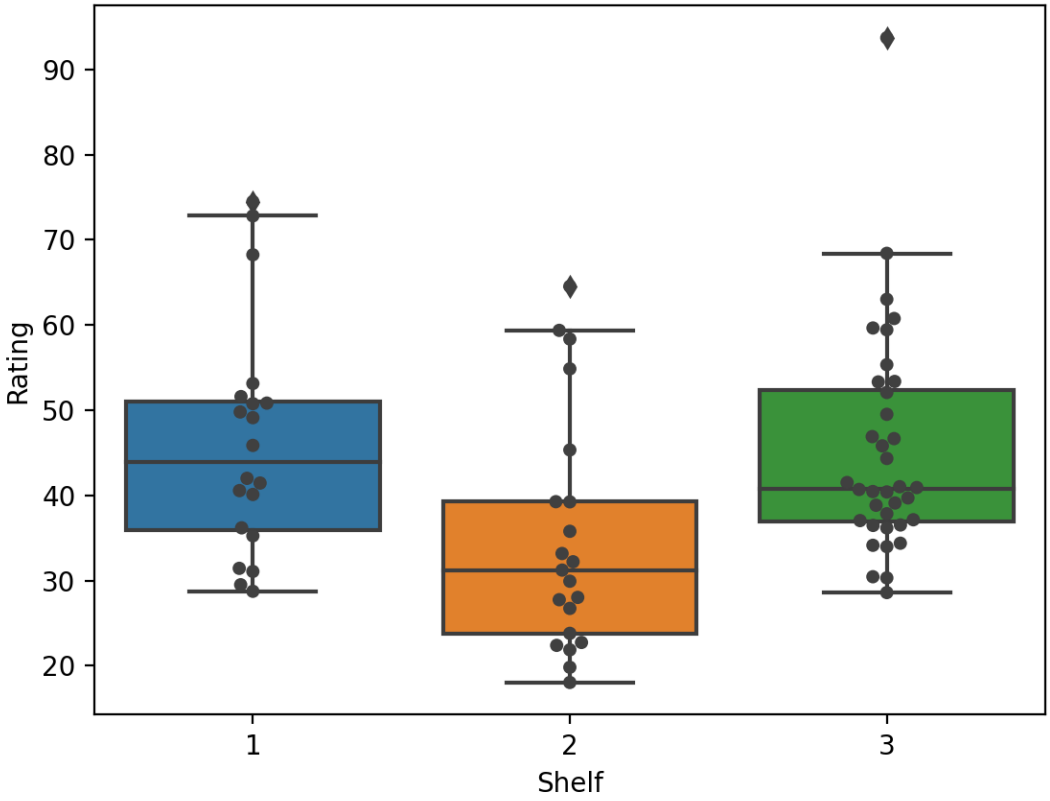


Figure 6: Boxplot for Shelf versus Rating.

Although this plot doesn't give us any significance information, we can judge those cereals kept in shelf 1 have the highest median reading for rating telling us that they are usually the healthy ones. However, among all the 77 cereals the one with the highest rating is kept in shelf 3 also it can be seen that there are a greater number of cereals with higher than median ratings present in shelf 3.

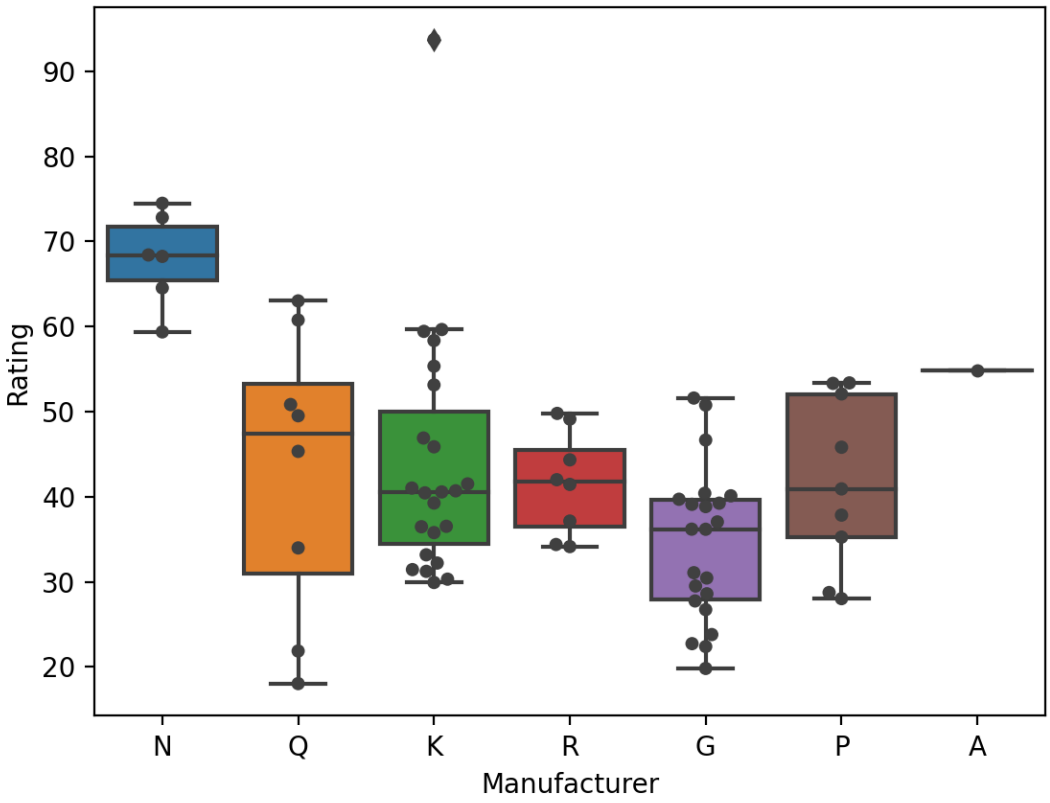
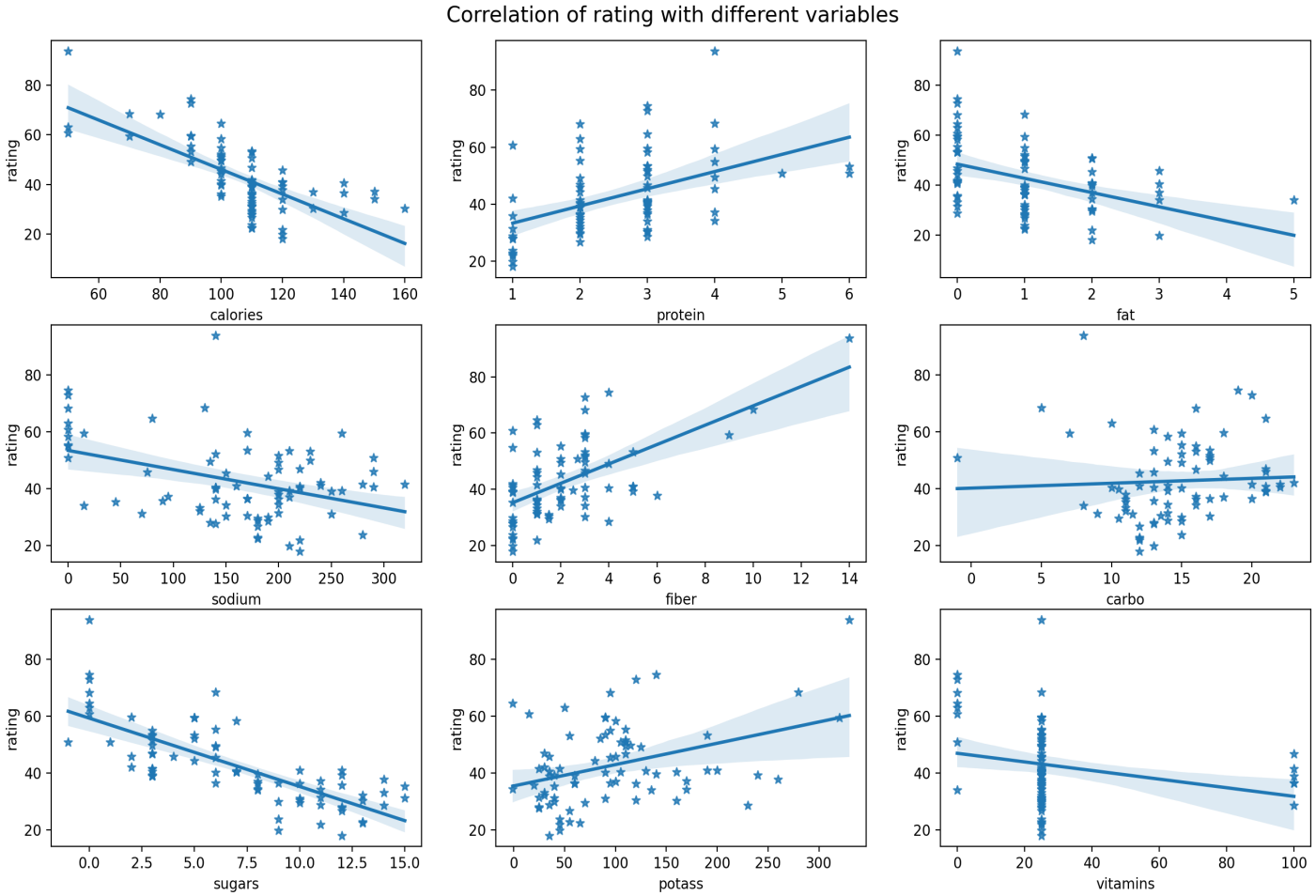


Figure 7: Boxplot for manufacturer versus Rating.

We can assess that the manufacturer with highest mean rating is Nabisco (N). Although this couldn't be the only dependency because the total number of cereals manufactured by it are only 6 which is way less than that of Kellogg's (K) and General Mills (G). The highest cereal is manufactured by Kellogg's which is shown by the outlier marked in diamond shape. It can also be seen that the average rating for all the 77 cereals should be something around 45.

6. Scatter Plots



The information that can be taken out from the above plot is that there is a strong positive relation between the protein and fiber content with the rating value. This is shown by and increasing exponential line as seen in the plot. The nature of curve for calories and sugars with that of rating is negative with a big negative slope which tells us that the ratings are degraded as the contents of these variables are increased.

7. Density Distribution Plots

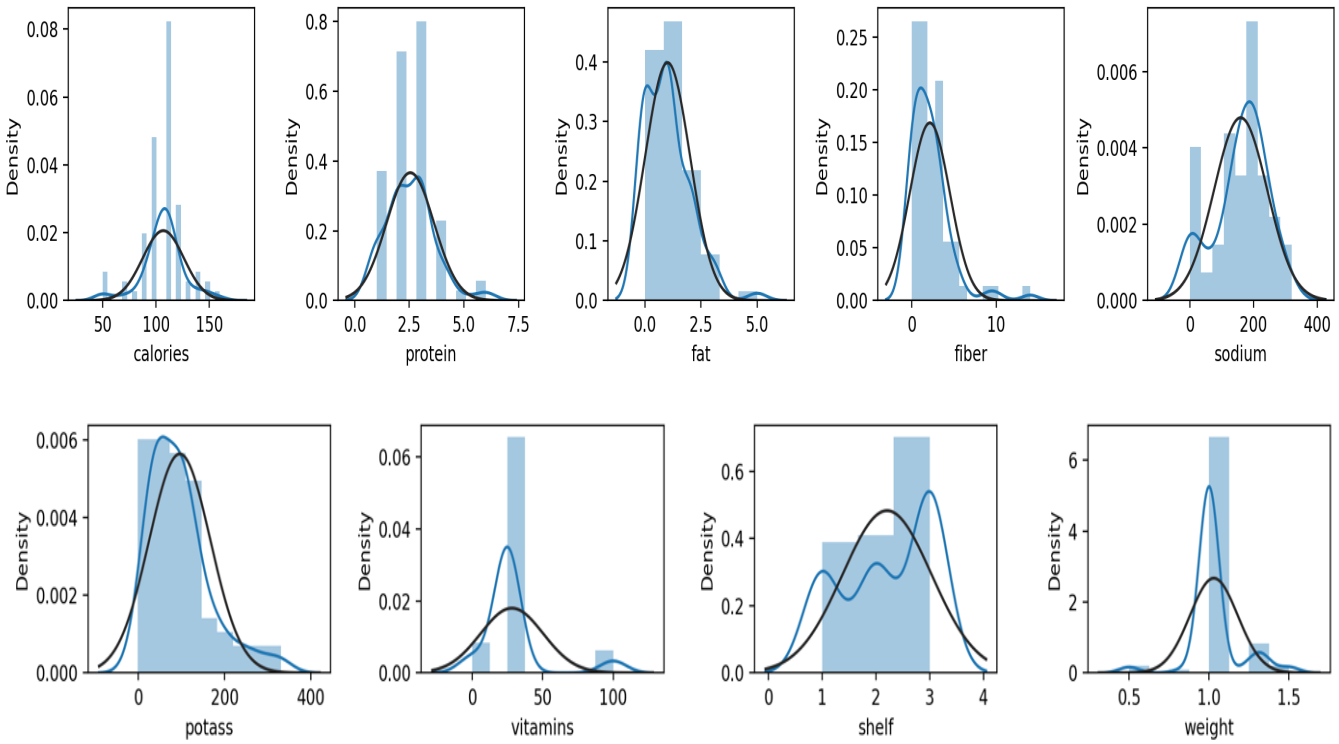


Figure 9: Density distribution Plots for various variables

These various distribution plots tell us about how different variables are spanned among the entire data. We can make out that the values of calories and sodium have little variance whereas the fat and potassium values are right skewed meaning that most of them have less amounts of each quantity.

8. Line plot for entire data

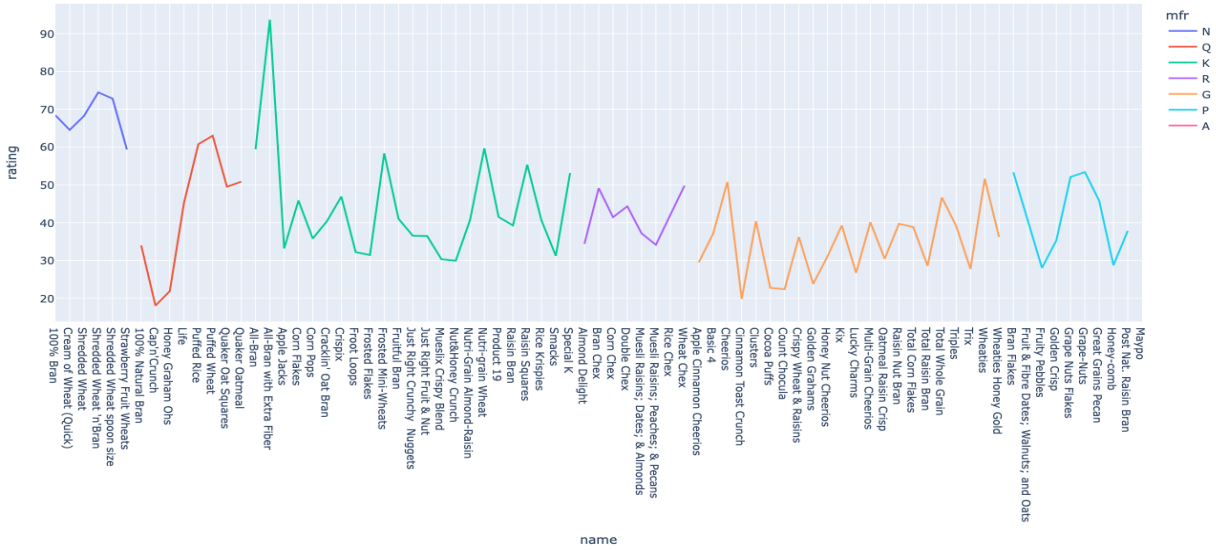


Figure 10: Line Plot showing all cereals with their rating and grouped by manufacturer.

This grouping allows us to visualize how each brand is doing based on their rating values. It also tells us what cereals are above a given threshold value by looking at the grid in the graph.

The width of the green and orange lines tells us that there are a greater number of cereals manufactured by them. The highest peak gives us the value of the cereal with highest rating which is ***All Bran with extra fiber produced by Kellogg's.***

9. Heatmap

Finally let us have a look at a heatmap. It calculates a correlation matrix between all the variables and gives us a rating for each variable against a target. It gives us information between the correlation factor between 2 variables.

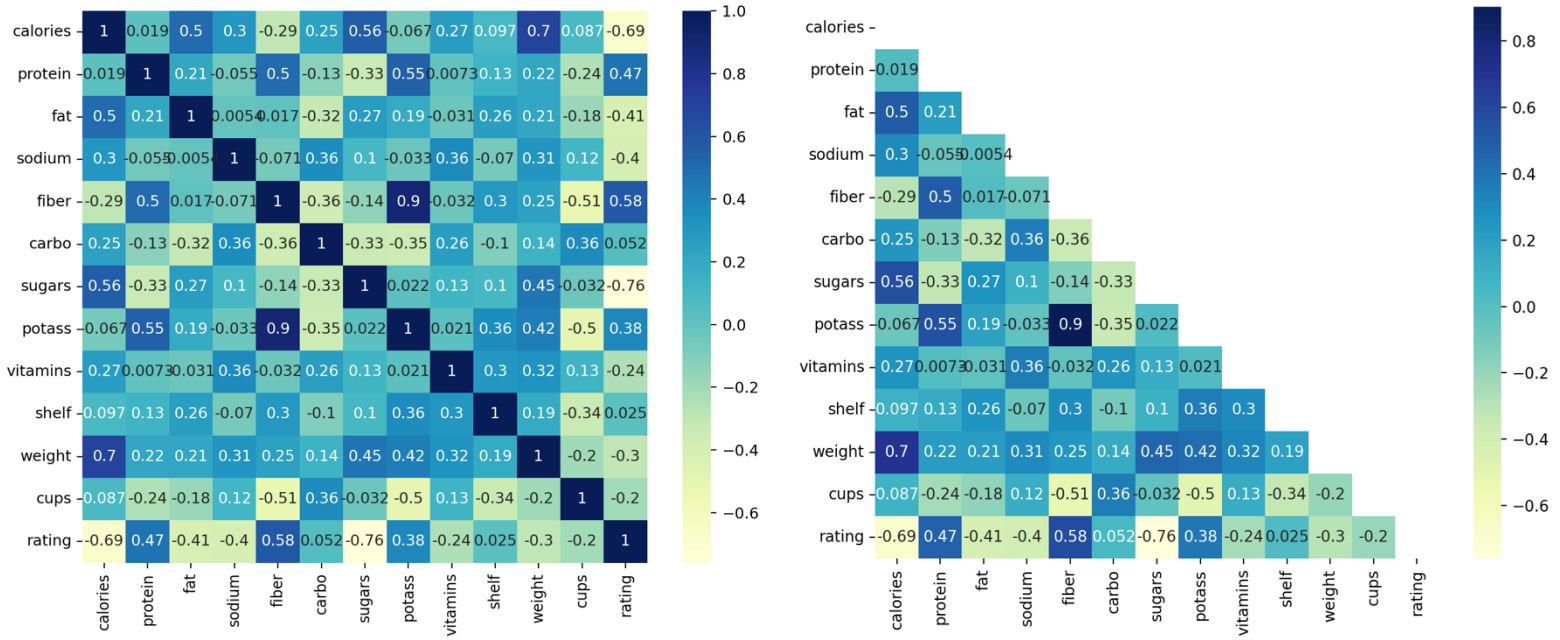


Figure 11: Heatmap showing the correlation matrix between all variables

The above representation tells us that there is a very strong correlation between (rating and sugars) and (rating and fiber) which was also seen from the scatter plots which were obtained earlier. The second figure on the right is just a modification of the left one while masking out the top triangle to decrease the ink used to generate the plot as a good practice of visualization.

Answer 2

Zoom into the Human Bloodstream:

1. The topic of this visualization poster is the internal anatomy of the human body typically the blood stream intended audience for this visualization is all people in the age group 11 years and adult.

2. The design principle used in this visualization technique is fitting multiple scales into an image for in a landscape fitting. It is just like an example professor showed us in the lecture where a road was shown, and lines were drawn across that looked like an end-horizon picture. The picture is depicted by using an increasing magnitude scale of 10 folds.

3. According to me this visualization does a good job at passing the information of what its intended audience should seek. It points out the different elements like starting from the heart to the red blood cell and then so on to the tiniest oxygen atom. It tells us about the significance of atoms as they are building blocks of all matter. It is responsible for living of all organs and the human body as a whole. Although the visualization depicts the information in a good way, I feel the lie factor here would be reasonably high which is not a good design principle.

4. It tells us about the relationship and interaction between macro and micro elements in the bloodstream The things that can be imagined from the visualization is that all matter is made up of atoms. The molecule chains shown in the picture depict the oxygen atoms that are the key aspect to the bloodstream. It is hard to identify tell for a layman to notice clearly the different elements present without the appropriate labels.

5. The color used in the work is appropriately fitting as it depicts the real color of the various difference elements present in the bloodstream. Using the color shades of red helps to create a strong environment for the visualization. The depictions of veins are also clear as they use a color different than red which distinguishes it properly.

6. As given in the article this picture was awarded the best in the year 2008 by NSF. However, there are certain things that make it difficult to read the image like why the size of red blood cells are changing in the depictions. Same elements should have the same size. The need of using labels is extremely useful and necessary to make it easy to understand and identify the function all the elements for more variably people from a non-biology background.

Answer 3

The given data can be re-written as follows considering the bleeding as the boundary conditions as provided in the question.

DATA =

8	8	8	6	-2	3	3	3
8	8	8	6	-2	3	3	3
8	8	8	6	-2	3	3	3
1	1	1	6	4	5	5	5
3	3	3	2	-4	11	11	11
10	10	10	-1	7	1	1	1
10	10	10	-1	7	1	1	1
10	10	10	-1	7	1	1	1

The highlighted portion shows us the original data

KERNEL =

-2	3	-1
4	-1	2
0	5	3

C_{11} data = $\begin{matrix} 8 & 8 & 6 \\ 8 & 8 & 6 \\ 1 & 1 & 6 \end{matrix}$ kernel = $\begin{matrix} -2 & 3 & -1 \\ 4 & -1 & 2 \\ 0 & 5 & 3 \end{matrix}$

$C_{11} = -16 + 24 - 6 + 32 - 8 + 12 + 5 + 18$

$C_{11} = 61$

C_{12} data = $\begin{matrix} 8 & 8 & -2 \\ 8 & 6 & -2 \\ 1 & 6 & 4 \end{matrix}$ kernel = $\begin{matrix} -2 & 3 & -1 \\ 4 & -1 & 2 \\ 0 & 5 & 3 \end{matrix}$

$C_{12} = -16 + 18 + 2 + 32 + 30 - 6 - 4 + 12$

$C_{12} = 68$

C_{13} data = $\begin{matrix} 6 & -2 & 3 \\ 6 & -2 & 3 \\ 6 & 4 & 5 \end{matrix}$

$C_{13} = -12 - 6 - 3 + 24 + 2 + 6 + 20 + 15$

$C_{13} = 46$

C_{14} data = $\begin{matrix} -2 & 3 & 3 \\ -2 & 3 & 3 \\ 4 & 5 & 5 \end{matrix}$

$C_{14} = 4 + 9 - 3 - 8 - 3 + 6 + 25 + 15$

$C_{14} = 45$

C_{21} data = $\begin{matrix} 8 & 8 & 6 \\ 1 & 1 & 6 \\ 3 & 3 & 2 \end{matrix}$

$C_{21} = -16+24-6+4-1+12+15+6$

$C_{21} = 38$

C_{22} data = $\begin{matrix} 8 & 6 & -2 \\ 1 & 6 & 4 \\ 3 & 2 & -4 \end{matrix}$

$C_{22} = -16+18+2+4-6+8+10-12$

$C_{22} = 8$

C_{33} data = $\begin{matrix} 6 & -2 & 3 \\ 6 & 4 & 5 \\ 2 & -4 & 11 \end{matrix}$

$C_{23} = -12-6-3+24-4+10-20+33$

$C_{23} = 22$

C_{24} = $\begin{matrix} -2 & 3 & 3 \\ 4 & 5 & 5 \\ -4 & 11 & 11 \end{matrix}$

$C_{24} = 4+9-3+16-5+10+55+33$

$C_{24} = 119$

C_{31} data = $\begin{matrix} 1 & 1 & 6 \\ 3 & 3 & 2 \\ 10 & 10 & -1 \end{matrix}$

$C_{31} = -2+3-6+12-3+4+50-3$

$C_{31} = 55$

C_{32} data = $\begin{matrix} 1 & 6 & 4 \\ 3 & 2 & -4 \\ 10 & -1 & 7 \end{matrix}$

$C_{32} = -2+18-4+12-2-8-5+21$

$C_{32} = 30$

C_{33} data = $\begin{matrix} 6 & 4 & 5 \\ 2 & -4 & 11 \\ -1 & 7 & 1 \end{matrix}$

$C_{33} = -12+12-5+8+4+22+35+3$

$C_{33} = 67$

C_{34} data = $\begin{matrix} 4 & 5 & 5 \\ -4 & 11 & 11 \\ 7 & 1 & 1 \end{matrix}$

$C_{34} = -8+15-5-16-11+22+5+3$

$C_{34} = 5$

C41 data = 3 3 2
10 10 -1
10 10 -1

$C_{41} = -6+9-2+40-10-2+50-3$
$C_{41} = 76$

C42 data = 3 2 -4
10 -1 7
10 -1 7

$C_{42} = -6+6+4+44+1+14-5+21$
$C_{42} = 75$

C43 data = 2 -4 11
-1 7 1
-1 7 1

$C_{43} = -4-12-11-4-7+2+35+3$
$C_{43} = 2$

C44 data -4 11 11
7 1 1
7 1 1

$C_{44} = 8+33-11+28-1+2+5+3$
$C_{44} = 67$

RESULT =

61	68	46	45
38	8	22	119
55	30	67	5
76	75	2	67

From the above found resultant matrix we found out to be as follows :

A = 45
B = 22
C = 55
D = 76

Answer 4

Until date, this course has been really good. Right from the first lecture where we had introductions from the professor and other students. This helps us gain an understanding of what professor is working on currently and his field of research. Student interactions help us know about the backgrounds they have prior to taking this course and helps in making connection and team building for the upcoming project work and paper representation.

A topic that I would be keen on learning apart from the ones included in the regular curriculum would be some regression methods that could come handy in analyzing variations in the data. This could be with help of some linear models or machine learning methods that could possibly be used to find the correlations to a target factor.

The kind of skills that I would be interested to learn over the course would be having hand on experience on various techniques and tool that can be used to read, analyze and evaluate data and learn about creating and presenting impactful presentations. All the visual techniques can be used to read, understand and interpret the data to learn about the trends, outliers and fit in the data sets.

A thing that could maybe be incorporated is some sort of lab session wherein we are exposed to some high-level visualization tool that otherwise would be difficult to access or learn from other resources. Also, if possible, we could visit professor's lab and learn more about the kind of work that happens and the research that he is working on and how that could be helpful to students like us who are new to the field.

So far, the homework seems to be nicely written with detailed instructions given so have no suggestions on that.

Finally, as a robotics engineer this course could be useful to me in all sense. Robotics is an extremely inter-disciplinary field which links closely between different branches of engineering. This course would promote the visualization and analysis of the data that could be generated by a robot. There is a huge amount of data generated which is often overlooked which proves to be significant. So, this kind of a course shall be useful in gaining the right knowledge and skills needed to make obedient use of the data.