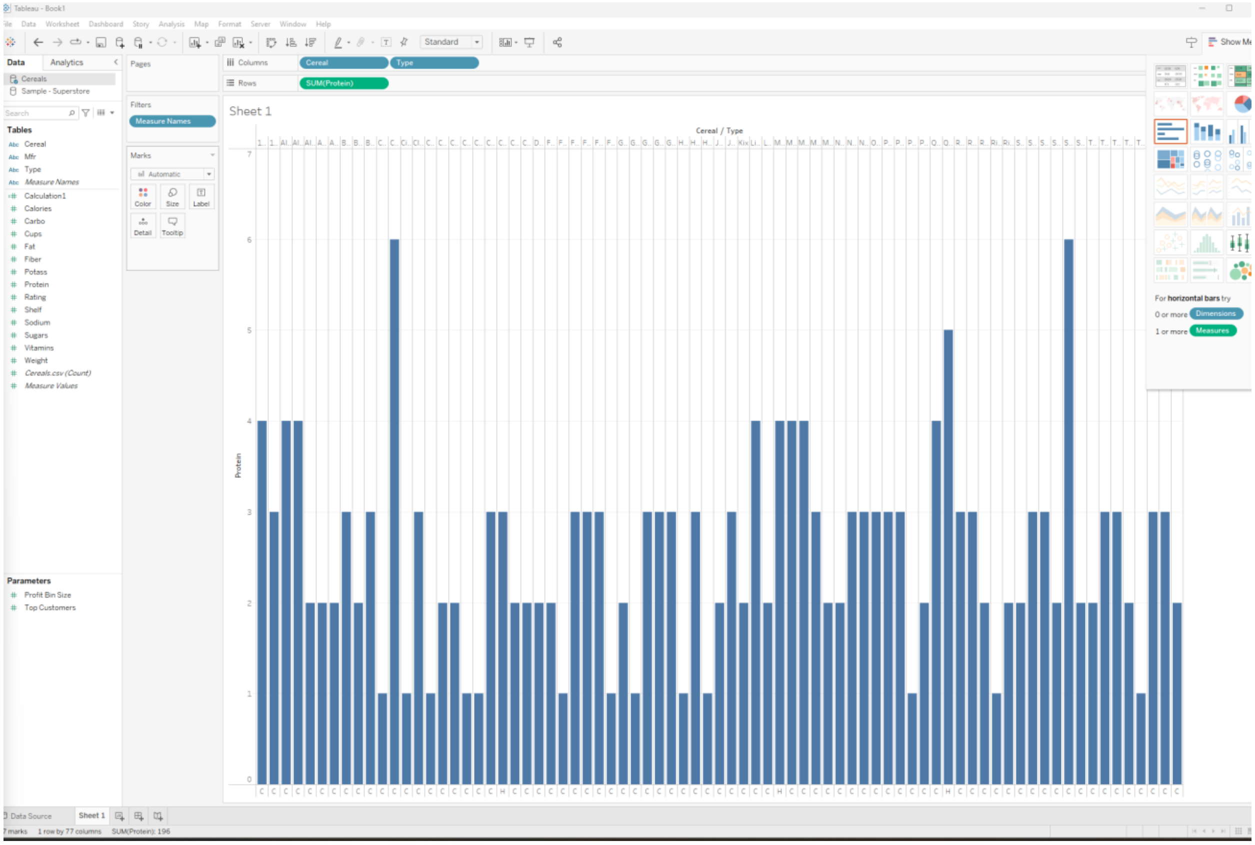
<u>DS 544: Homework #1</u>
Jacob Edwards

Question 1: (2 points) Install the Tableau (free for students), and work through the tutorials specified in class. Download the data sets, Cereals.csv and Cereals.xlsx, use Tableau to plot the 13 histograms like the samples below.

Answer: The plot below matches the diagram in the sample shown for the homework with Protein on our y-axis and the Cereal name/ type in the x-axis.



Columns: Cereal, Type
Rows: SUM(Protein)
Filters: Measure Names

7 marks 1 row by 77 columns SUM(Protein): 196

Question 2: Download the dataset cars.xls for Cars in the shared folder. Use Tableau Desktop/Laptop to answer the following questions.

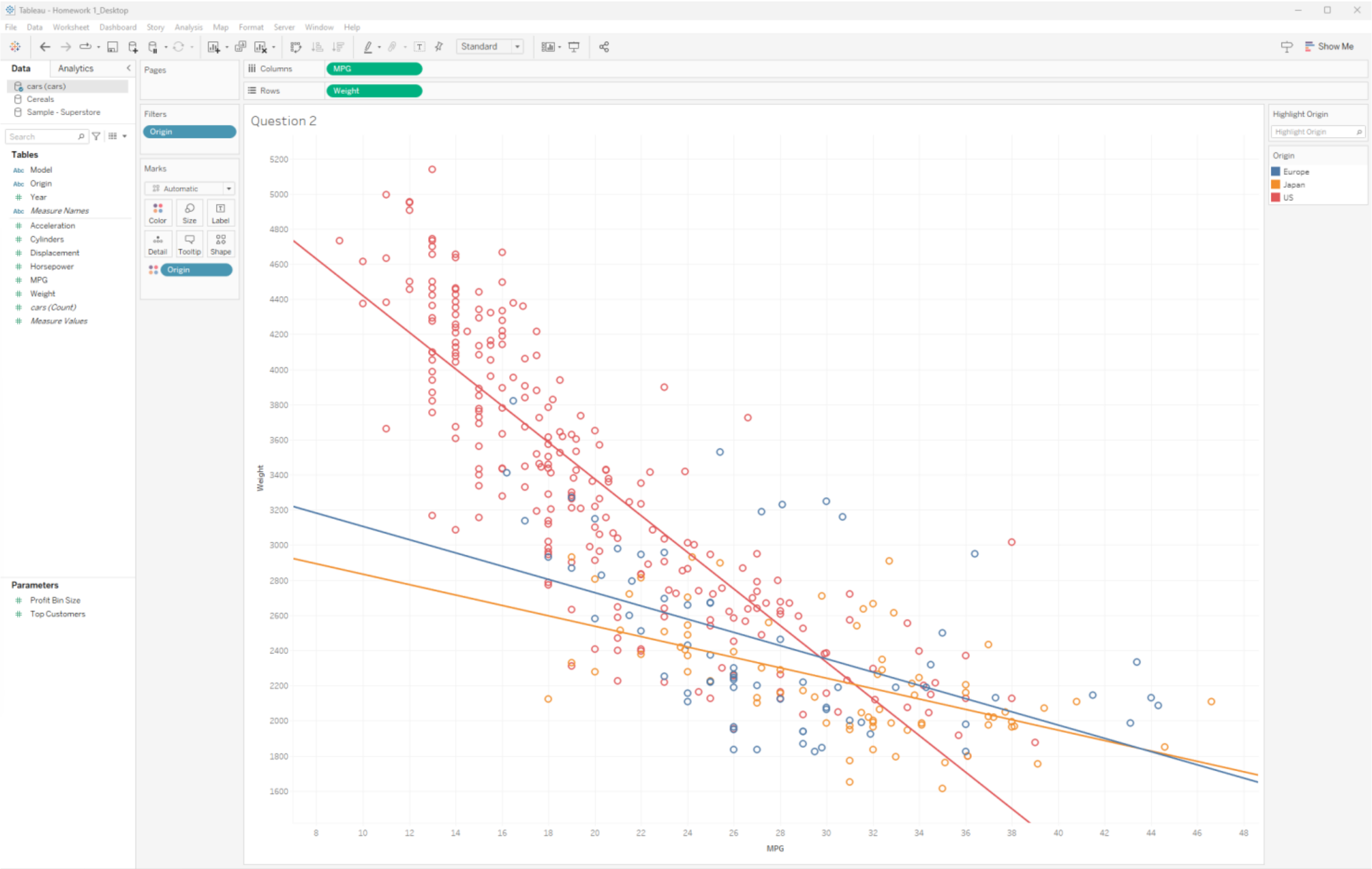
a. What are the types of each variable?

Answer: Variables: Model (Qualitative), MPG (Quantitative), Cylinders (Quantitative), Displacement (Quantitative), Horsepower (Quantitative), Weight (Quantitative), Acceleration (Quantitative), Year (Quantitative), Origin (Qualitative).

DS 544: Homework #1
Jacob Edwards

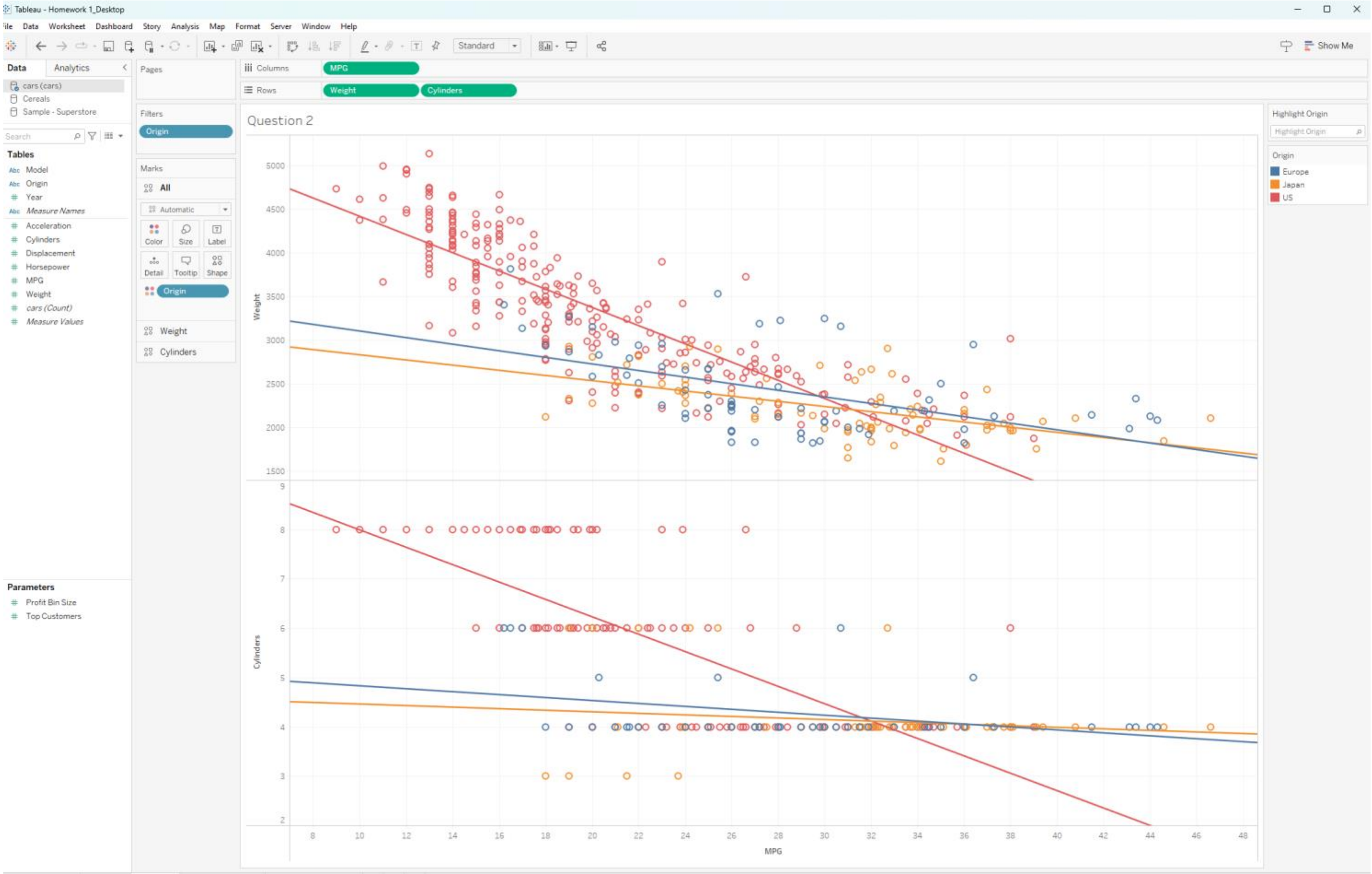
b. Sketch the graph representing the regression relationship between Weight of Cars and MPG, with the color of dots to encode the third variable the Origin of Cars

Answer: y-axis represents Weight of vehicles while x-axis represents MPG of vehicles. We have a downward sloping regression line showing that as the weight of cars decrease the miles per gallon of vehicles increase.



c. Use small multiples to encode the fourth variable Number of Cylinders of the Cars; print three similar figures to the samples below

<u>DS 544: Homework #1</u>
Jacob Edwards



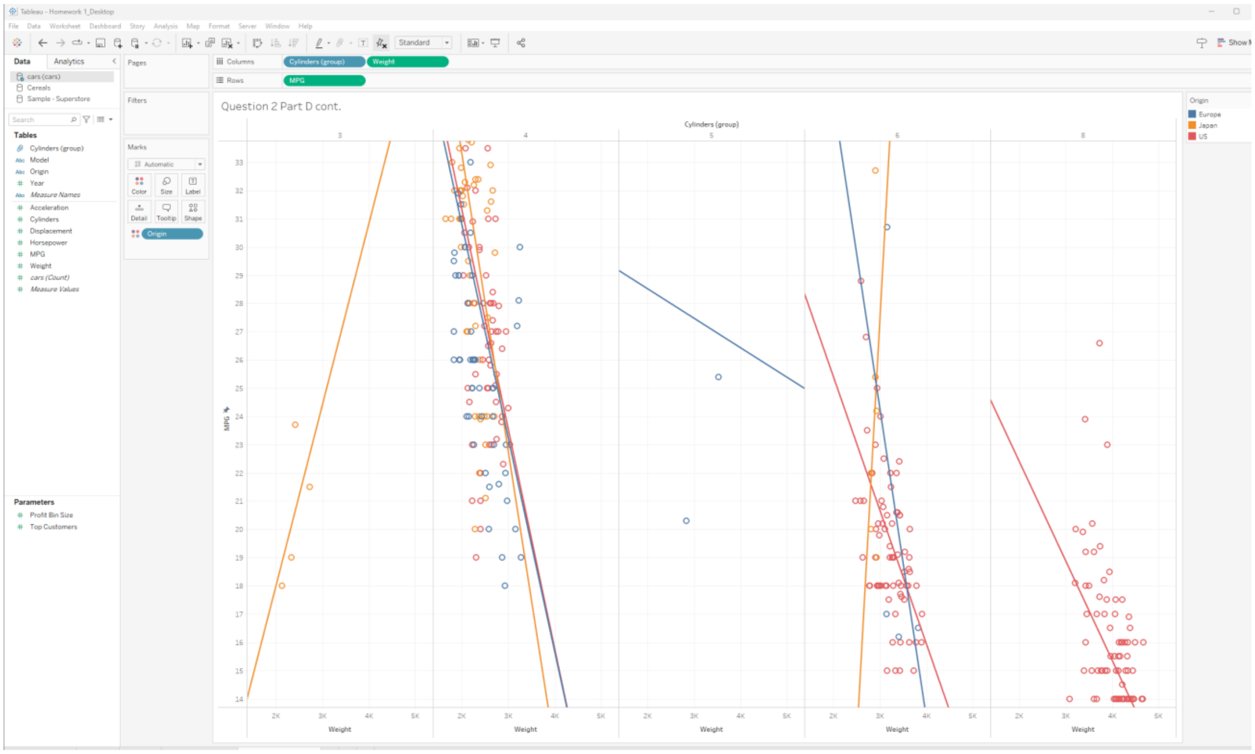
Columns: MPG
Rows: Weight, Cylinders
Filters: Origin
Marks: Origin
Weight
Cylinders
Origin: 🟦 Europe, 🟧 Japan, 🟥 US

Answer: The figure representing the Cylinders of each of the cars are represented below the first chart. It compares the cylinders on the y-axis to the MPG on the x-axis. The trend line or regression line also shows the direct comparison between car origins and the different axis that are represented in each model.

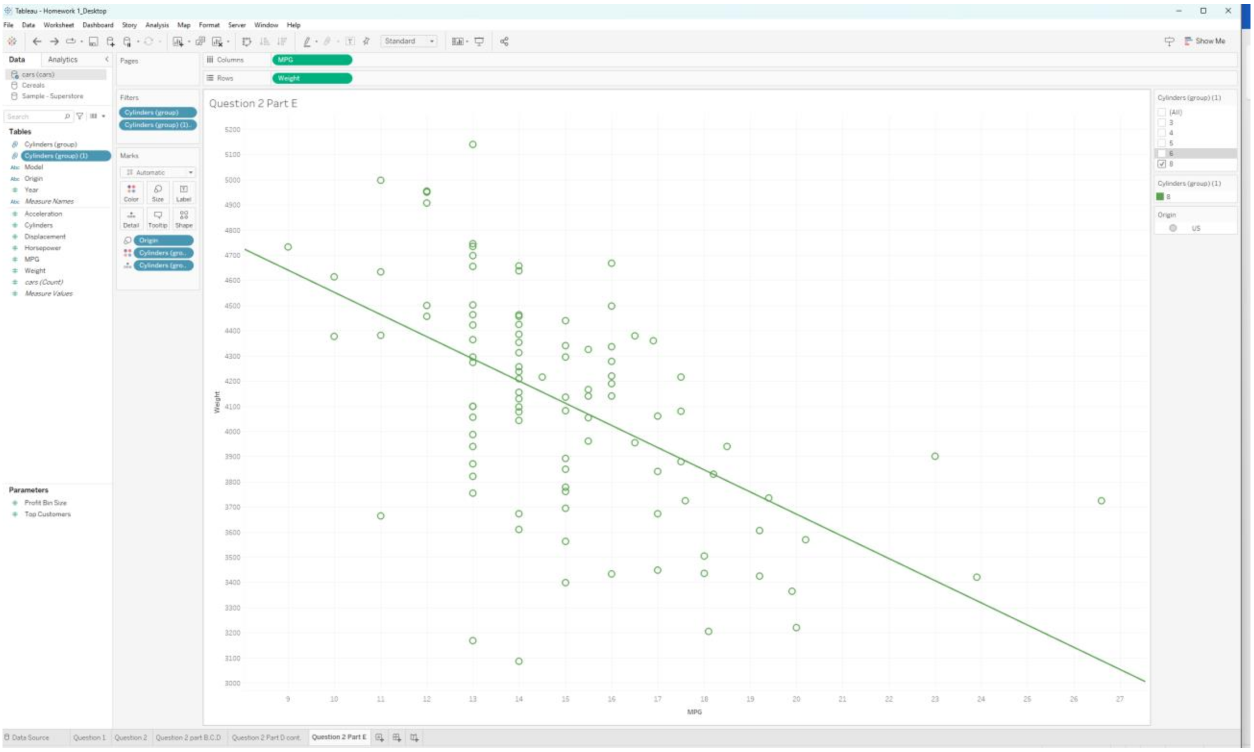
d. Select all four variables above and use the hint on the right pane from Tableau and use the suggested diagrams to represent the relationship among the variables and print the results.

Answer: We went over this question at the end of class, and at first, I was not correctly grouping the columns by the Cylinders field. However, with the Cylinders field grouped (image below), you can see that the weight is being shown directly in each section split up per Cylinder input: 3,4,5,6,8.

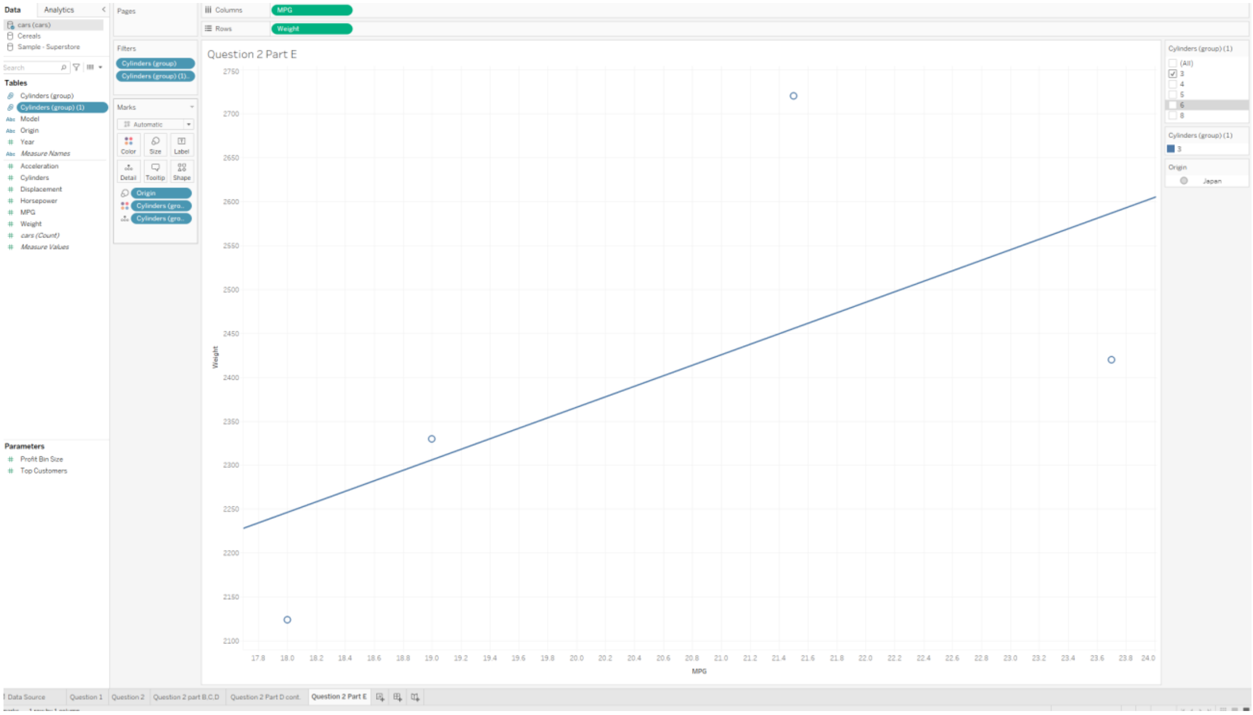
DS 544: Homework #1
Jacob Edwards


Columns: Cylinders (group) Weight
Rows: MPG
Question 2 Part D cont.

e. Set up a filter to hide blue and red colors but only show the green color variable in the second panel. Next, print the screenshot showing only the blue color variable but hiding the red and green colors


Columns: MPG
Rows: Weight
Filters: Cylinders (group), Cylinders (group) (1)
Question 2 Part E

DS 544: Homework #1
Jacob Edwards



Question 3: 2 points) Register Tableau Cloud and use the same dataset above to answer the same questions of problem 2 on the Tableau Cloud. Please summarize the difference in functionalities between Tableau Desktop and Tableau Cloud.

Answer: The biggest differences between the Cloud version and Desktop version of Tableau are some of the limitations of the features available via the cloud version vs. the desktop version. Specifically, the desktop version has more functional features for advanced users than the cloud version. Users are responsible for the updates regarding the Desktop version, whereas the cloud version updates live. Obviously, in terms of functionality, the Desktop version is limited to the downloaded installation on someone's local machine vs. the web-based cloud version that be accessed via any network connection. It's much easier for teams to use the web-based version for sharing and collaboration. This source provides an even further description as to what the fundamental functionality differences are between Tableau Desktop and Cloud.

1. Data source editing is limited in Tableau Server. You can only edit some information. For example, in Tableau release 2020 you still cannot edit joins or relationships within Tableau Server's web authoring environment. As an alternative,

you can use Tableau Prep (Prep is a Tableau ETL tool similar to Alteryx). Tableau Prep is available within the browser in versions 2020.4 and above.

2. Tableau Desktop provides some analytic capabilities not found in Tableau Server's web authoring environment. What is available depends on the version of Tableau Server. For a complete list of capabilities by version see Tableau's Web Authoring and Tableau Desktop Feature Comparison.

3. As of Tableau version 2020.4, formatting options are more limited in Tableau Server. You cannot format at the item or worksheet level, only at the worksheet level. (Kumar, 2024).

Question 4: (4 points) This question is based on the following videos:
The Data Detective by Tim Harford: https://www.youtube.com/watch?v=rmqi09ldF_E
Ten Rules for Thinking Differently About Numbers: https://youtu.be/FkvdtqtL1aM
How to Lie With Statistics by Darrell Huff:
https://www.youtube.com/watch?v=Vu7WKcWbg24
a. Use your words to describe what are the top three misconceptions about the role of Statistics.

Answer: I think the first misconception about statistics that is widely misunderstood is that just because you are seeing a graph doesn't make it "correct" or accurate. Likely, there is inherent bias as to what you are being shown for someone's job or perspective. Secondly, I think another big one is that correlation is equal to causation (Learn Statistics Easily, 2024). Most people might believe that if you can show them a trend line or regression analysis, in other words, that it will directly show the correlation and provide you with the cause of such trend between those two features. It's not always accurate as to what is causing a certain metric to move in a particular direction. Typically, there are other features that might have a stronger more foundational impact on why a graph looks a certain way that isn't being shown. With that in mind, there is a level of domain knowledge that can be extremely insightful when identifying truly causational & correlated features for a particular dataset. Thirdly, in my opinion, Statistics are misunderstood due to the implicit nature of them being mathematically involved. I think most people want to look at an image or metric and believe it. I don't think they want to put in the legwork to understand how it's analyzed or measured. Often in the first two situations described, the third can play a large factor in influencing the audience and why certain statistics are being accepted as factual evidence.

DS 544: Homework #1
Jacob Edwards

b. Give two examples of how you use statistics or computational thinking in your daily life to make decisions. Do you estimate your uncertainty in statistical terms?

Answer: I think one way that I use statistics in my daily life is thinking about the likelihood or probability of me getting a parking spot on campus when driving to this class for instance. I have experienced the lots being full in past semesters as well as I know the number of incoming students. The number of classes that are in the morning outweigh the number of classes that are in the afternoon/ evening as well. All those factors combined lead me to believe that it's going to be very hard to acquire a parking spot on campus during the 10am time block. Therefore, I choose to spend the extra money to park at the parking garage as to mitigate my chances of being late to class. Another example, I use the weather forecast to plan my day sometimes because there are higher probabilities on certain days or times that forecast inclement weather. I would argue that I do estimate my uncertainty in statistical terms. Often, in conversation, I will use percentages as a measure of confidence.

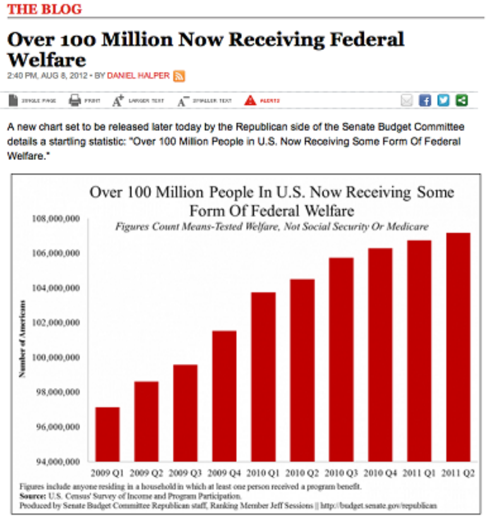
c. Can you give two examples of how Statistics are used to mislead the public in media and political debates? It is OK to Google it, but you need to provide the link to the source information.

Answer: One example comes directly from the slides that were shown on the first day of class. It was a representation of a couple different pie charts that don't add up to 100%. This false metric to statisticians is easy to see. However, to the public eye, it can be confusing or misleading due to them seeing the same portion size on each of the slices of the pie chart to be around the same size, and each portion measures up to be around the same. We know that it isn't possible for the chart to represent an amount over 100%, so it sticks out as misinformation used to persuade one party vs. the other regarding political campaigns. A second example can be seen below in the screenshot I found when researching this topic. (Statistics How To, 2024) The graph shows how the chart has continued to grow, but due to the size of the chart in comparison to the actual measurements being used, it seems like a larger trend is happening than the reality of the situation. If it started out at 0 or even 10,000,000 instead of 94,000,000, then it would show a more accurate representation of the increase, which is marginal in relation to the size of the metrics we are looking at. In two years, it actually represents an increase of less than 10%.

DS 544: Homework #1
Jacob Edwards

USA Today

USA today is notorious for fussy graphs that have too much information and mislead. This graph makes our welfare problem look like it's spiraling out of control. But note where the y-axis starts…at 94 million!



THE BLOG

Over 100 Million Now Receiving Federal Welfare

2:40 PM, AUG 8, 2012 • BY DANIEL HALPER

A new chart set to be released later today by the Republican side of the Senate Budget Committee details a startling statistic: "Over 100 Million People in U.S. Now Receiving Some Form Of Federal Welfare."

Figures include anyone residing in a household in which at least one person received a program benefit.
Source: U.S. Census' Survey of Income and Program Participation.
Produced by Senate Budget Committee Republican staff, Ranking Member Jeff Sessions || http://budget.senate.gov/republican

Answer:

d. Can you offer your opinion about the opinion from the claims of the source information above?

> For example, you may find the original source is from a liberal-lean Journalist or more conservative point of view and which parts are facts and which are simply opinions.

Answer: This is a great question because inherently there will almost certainly be bias on either side. There is bias in this question because my answer may coincide with what the writer/ information source provided, or it may differ entirely based on beliefs that I have about the economy, politics, etc. To answer the actual question, I believe that the person writing this article is most likely a liberal since they are picking on Fox news quite a bit in this article. Traditionally, in my experience, Fox tends to run more conservative from the conversations that I've had about it. That said, I don't think that these statements are wrong in the article, but it does have a target that's being used repeatedly. They are using their power as an author to represent Fox in a negative light; however, it is not wrong with what they are providing. I think it would provide a more equitable argument if they provided some misleading statistics from other news sources as well that lean more towards the liberal side like CNN for example. It would provide a less-biased opinion on the topic of statistics being misleading overall.

e. You may suggest what type of data source (you do not need to actually collect them) may help to validate your opinion.

Answer: The data source that would help validate my opinion would be to find a more conservative author who is bashing on CNN for instance. It would at least provide some equity in the argument that all news broadcasting stations have some inherent bias and tailor make their charts/ visualizations a certain way to persuade their viewers towards their side of the argument. Frankly, I don't have much of an opinion on what's a better news station, but I do think that the truth matters. Truth being portrayed a certain way to sway opinions can very well be labeled as misinformation at a certain point of exaggeration. I think it would be best if we could provide an equal number of examples in one article to show that news stations are providing biased information overall.

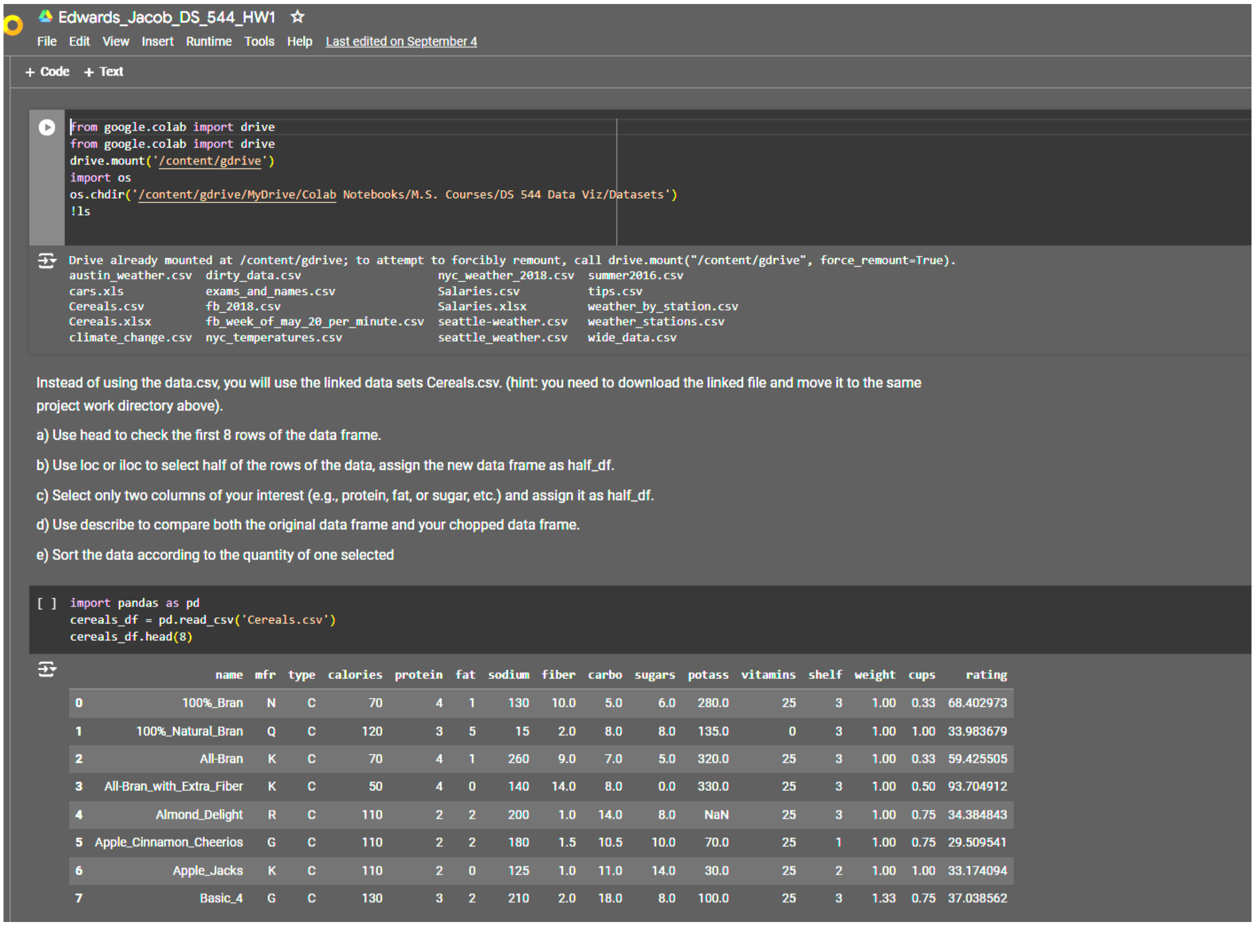
Question 5:

(2 points) When you read files to your Google Colab, it is critical that you work in the right project directory or use the exact right path to open the data files. Before you run the codes of Molin's book, insert a cell on top and run the following codes to mount your Google drive with Colab, then set your project work directory (pwd) as the folder you put your unzipped file. Make sure to change the … below as the folder where you extract the zip files.
Instead of using the data.csv, you will use the linked data sets Cereals.csv. (hint: you need to download the linked file and move it to the same project work directory above).

a) Use head to check the first 8 rows of the data frame.

DS 544: Homework #1

Jacob Edwards


Edwards_Jacob_DS_544_HW1 ☆
File Edit View Insert Runtime Tools Help Last edited on September 4
+ Code + Text

```
from google.colab import drive
from google.colab import drive
drive.mount('/content/gdrive')
import os
os.chdir('/content/gdrive/MyDrive/Colab Notebooks/M.S. Courses/DS 544 Data Viz/Datasets')
!ls
```

```
Drive already mounted at /content/gdrive; to attempt to forcibly remount, call drive.mount("/content/gdrive", force_remount=True).
austin_weather.csv  dirty_data.csv                   nyc_weather_2018.csv  summer2016.csv
cars.xls            exams_and_names.csv              Salaries.csv          tips.csv
Cereals.csv         fb_2018.csv                      Salaries.xlsx         weather_by_station.csv
Cereals.xlsx        fb_week_of_may_20_per_minute.csv seattle-weather.csv   weather_stations.csv
climate_change.csv  nyc_temperatures.csv             seattle_weather.csv   wide_data.csv
```

Instead of using the data.csv, you will use the linked data sets Cereals.csv. (hint: you need to download the linked file and move it to the same project work directory above).

a) Use head to check the first 8 rows of the data frame.

b) Use loc or iloc to select half of the rows of the data, assign the new data frame as half_df.

c) Select only two columns of your interest (e.g., protein, fat, or sugar, etc.) and assign it as half_df.

d) Use describe to compare both the original data frame and your chopped data frame.

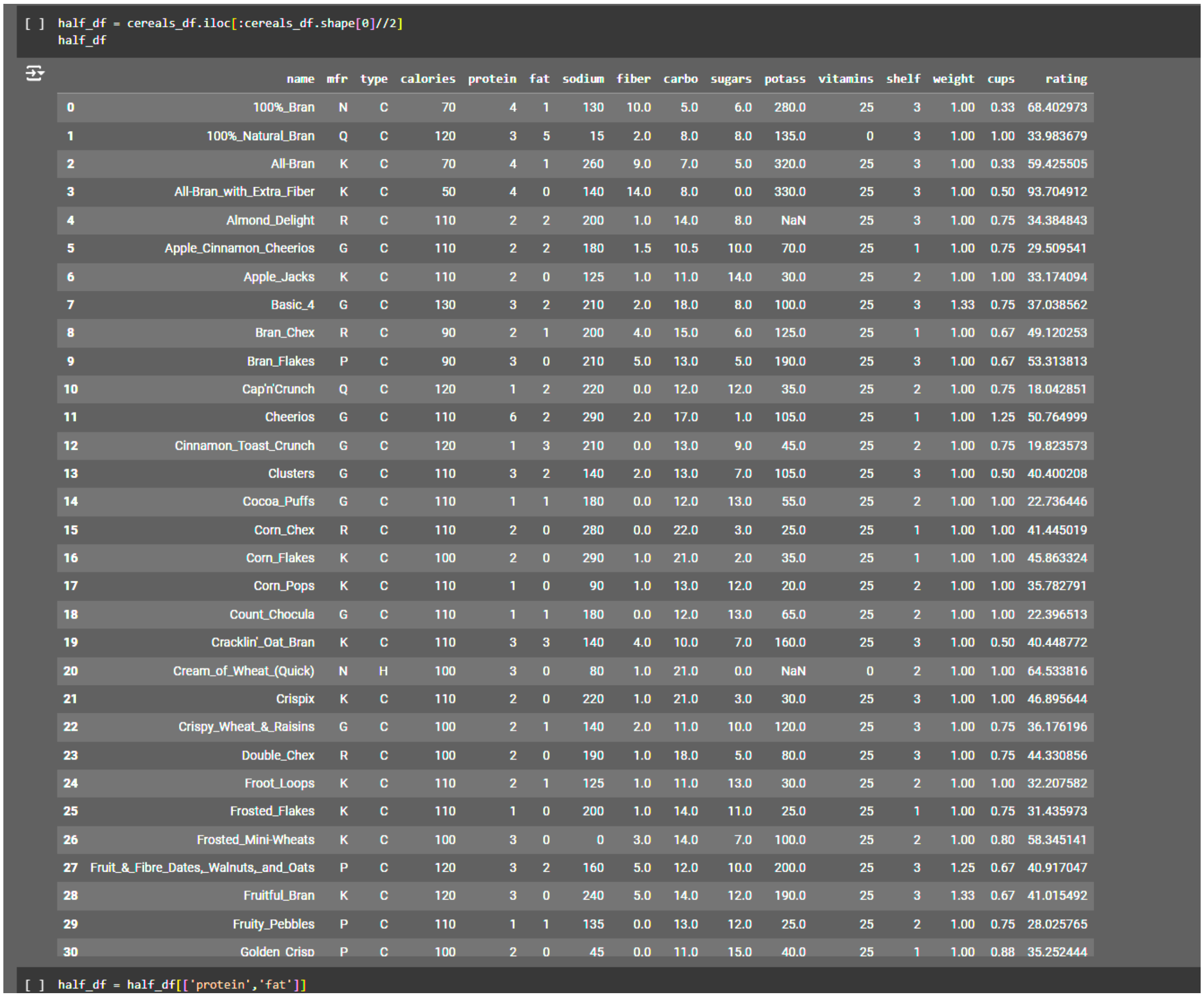
e) Sort the data according to the quantity of one selected

```
[ ] import pandas as pd
    cereals_df = pd.read_csv('Cereals.csv')
    cereals_df.head(8)
```

	name	mfr	type	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	weight	cups	rating
0	100%_Bran	N	C	70	4	1	130	10.0	5.0	6.0	280.0	25	3	1.00	0.33	68.402973
1	100%_Natural_Bran	Q	C	120	3	5	15	2.0	8.0	8.0	135.0	0	3	1.00	1.00	33.983679
2	All-Bran	K	C	70	4	1	260	9.0	7.0	5.0	320.0	25	3	1.00	0.33	59.425505
3	All-Bran_with_Extra_Fiber	K	C	50	4	0	140	14.0	8.0	0.0	330.0	25	3	1.00	0.50	93.704912
4	Almond_Delight	R	C	110	2	2	200	1.0	14.0	8.0	NaN	25	3	1.00	0.75	34.384843
5	Apple_Cinnamon_Cheerios	G	C	110	2	2	180	1.5	10.5	10.0	70.0	25	1	1.00	0.75	29.509541
6	Apple_Jacks	K	C	110	2	0	125	1.0	11.0	14.0	30.0	25	2	1.00	1.00	33.174094
7	Basic_4	G	C	130	3	2	210	2.0	18.0	8.0	100.0	25	3	1.33	0.75	37.038562

DS 544: Homework #1
Jacob Edwards

b) Use loc or iloc to select half of the rows of the data, assign the new data frame as half_df.



```
[ ] half_df = cereals_df.iloc[:cereals_df.shape[0]//2]
    half_df
```

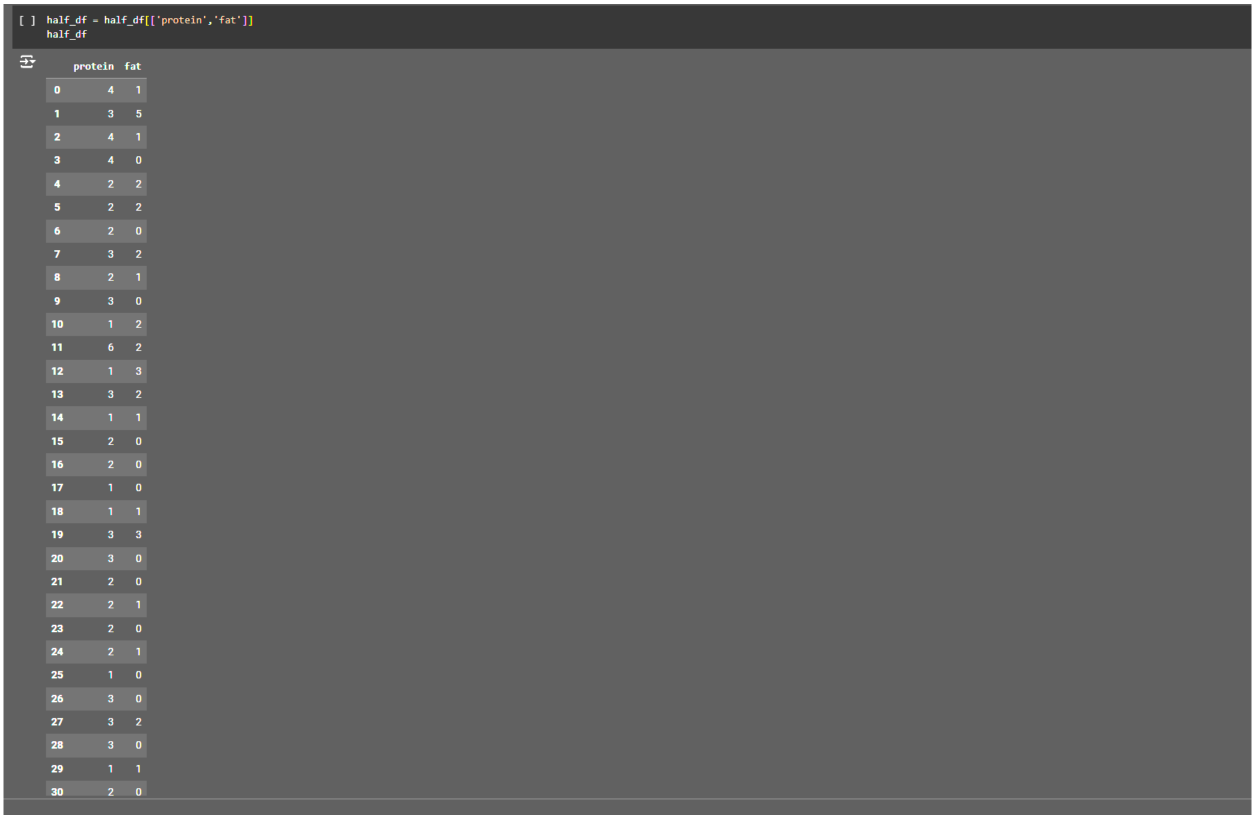
	name	mfr	type	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	weight	cups	rating
0	100%_Bran	N	C	70	4	1	130	10.0	5.0	6.0	280.0	25	3	1.00	0.33	68.402973
1	100%_Natural_Bran	Q	C	120	3	5	15	2.0	8.0	8.0	135.0	0	3	1.00	1.00	33.983679
2	All-Bran	K	C	70	4	1	260	9.0	7.0	5.0	320.0	25	3	1.00	0.33	59.425505
3	All-Bran_with_Extra_Fiber	K	C	50	4	0	140	14.0	8.0	0.0	330.0	25	3	1.00	0.50	93.704912
4	Almond_Delight	R	C	110	2	2	200	1.0	14.0	8.0	NaN	25	3	1.00	0.75	34.384843
5	Apple_Cinnamon_Cheerios	G	C	110	2	2	180	1.5	10.5	10.0	70.0	25	1	1.00	0.75	29.509541
6	Apple_Jacks	K	C	110	2	0	125	1.0	11.0	14.0	30.0	25	2	1.00	1.00	33.174094
7	Basic_4	G	C	130	3	2	210	2.0	18.0	8.0	100.0	25	3	1.33	0.75	37.038562
8	Bran_Chex	R	C	90	2	1	200	4.0	15.0	6.0	125.0	25	1	1.00	0.67	49.120253
9	Bran_Flakes	P	C	90	3	0	210	5.0	13.0	5.0	190.0	25	3	1.00	0.67	53.313813
10	Cap'n'Crunch	Q	C	120	1	2	220	0.0	12.0	12.0	35.0	25	2	1.00	0.75	18.042851
11	Cheerios	G	C	110	6	2	290	2.0	17.0	1.0	105.0	25	1	1.00	1.25	50.764999
12	Cinnamon_Toast_Crunch	G	C	120	1	3	210	0.0	13.0	9.0	45.0	25	2	1.00	0.75	19.823573
13	Clusters	G	C	110	3	2	140	2.0	13.0	7.0	105.0	25	3	1.00	0.50	40.400208
14	Cocoa_Puffs	G	C	110	1	1	180	0.0	12.0	13.0	55.0	25	2	1.00	1.00	22.736446
15	Corn_Chex	R	C	110	2	0	280	0.0	22.0	3.0	25.0	25	1	1.00	1.00	41.445019
16	Corn_Flakes	K	C	100	2	0	290	1.0	21.0	2.0	35.0	25	1	1.00	1.00	45.863324
17	Corn_Pops	K	C	110	1	0	90	1.0	13.0	12.0	20.0	25	2	1.00	1.00	35.782791
18	Count_Chocula	G	C	110	1	1	180	0.0	12.0	13.0	65.0	25	2	1.00	1.00	22.396513
19	Cracklin'_Oat_Bran	K	C	110	3	3	140	4.0	10.0	7.0	160.0	25	3	1.00	0.50	40.448772
20	Cream_of_Wheat_(Quick)	N	H	100	3	0	80	1.0	21.0	0.0	NaN	0	2	1.00	1.00	64.533816
21	Crispix	K	C	110	2	0	220	1.0	21.0	3.0	30.0	25	3	1.00	1.00	46.895644
22	Crispy_Wheat_&_Raisins	G	C	100	2	1	140	2.0	11.0	10.0	120.0	25	3	1.00	0.75	36.176196
23	Double_Chex	R	C	100	2	0	190	1.0	18.0	5.0	80.0	25	3	1.00	0.75	44.330856
24	Froot_Loops	K	C	110	2	1	125	1.0	11.0	13.0	30.0	25	2	1.00	1.00	32.207582
25	Frosted_Flakes	K	C	110	1	0	200	1.0	14.0	11.0	25.0	25	1	1.00	0.75	31.435973
26	Frosted_Mini-Wheats	K	C	100	3	0	0	3.0	14.0	7.0	100.0	25	2	1.00	0.80	58.345141
27	Fruit_&_Fibre_Dates,_Walnuts,_and_Oats	P	C	120	3	2	160	5.0	12.0	10.0	200.0	25	3	1.25	0.67	40.917047
28	Fruitful_Bran	K	C	120	3	0	240	5.0	14.0	12.0	190.0	25	3	1.33	0.67	41.015492
29	Fruity_Pebbles	P	C	110	1	1	135	0.0	13.0	12.0	25.0	25	2	1.00	0.75	28.025765
30	Golden_Crisp	P	C	100	2	0	45	0.0	11.0	15.0	40.0	25	1	1.00	0.88	35.252444

```
[ ] half_df = half_df[['protein','fat']]
```

c) Select only two columns of your interest (e.g., protein, fat, or sugar, etc.) and assign it as half_df.

DS 544: Homework #1

Jacob Edwards

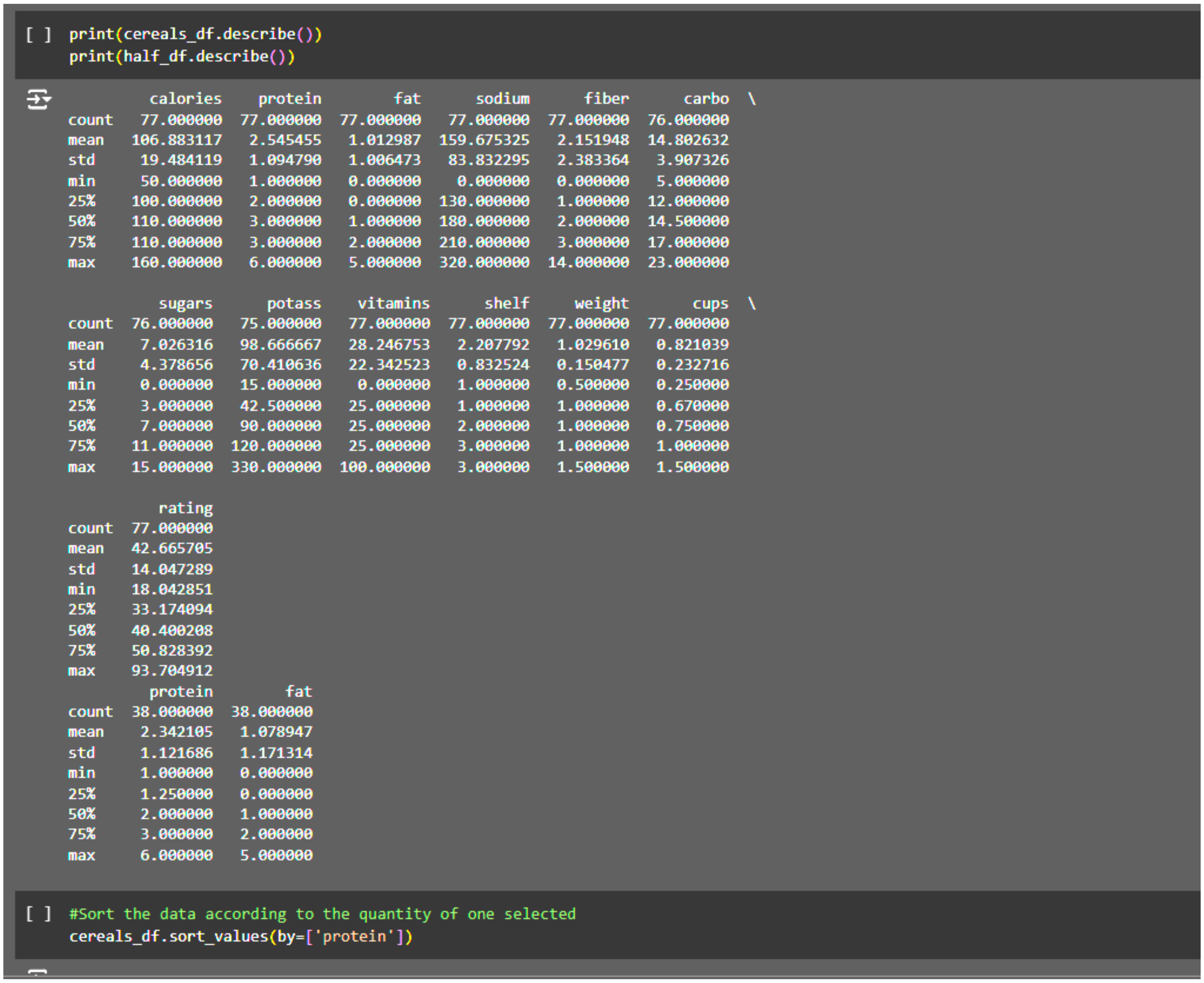


```
[ ] half_df = half_df[['protein','fat']]
    half_df
```

	protein	fat
0	4	1
1	3	5
2	4	1
3	4	0
4	2	2
5	2	2
6	2	0
7	3	2
8	2	1
9	3	0
10	1	2
11	6	2
12	1	3
13	3	2
14	1	1
15	2	0
16	2	0
17	1	0
18	1	1
19	3	3
20	3	0
21	2	0
22	2	1
23	2	0
24	2	1
25	1	0
26	3	0
27	3	2
28	3	0
29	1	1
30	2	0

DS 544: Homework #1
Jacob Edwards

d) Use describe to compare both the original data frame and your chopped data frame.



```
[ ] print(cereals_df.describe())
    print(half_df.describe())
```

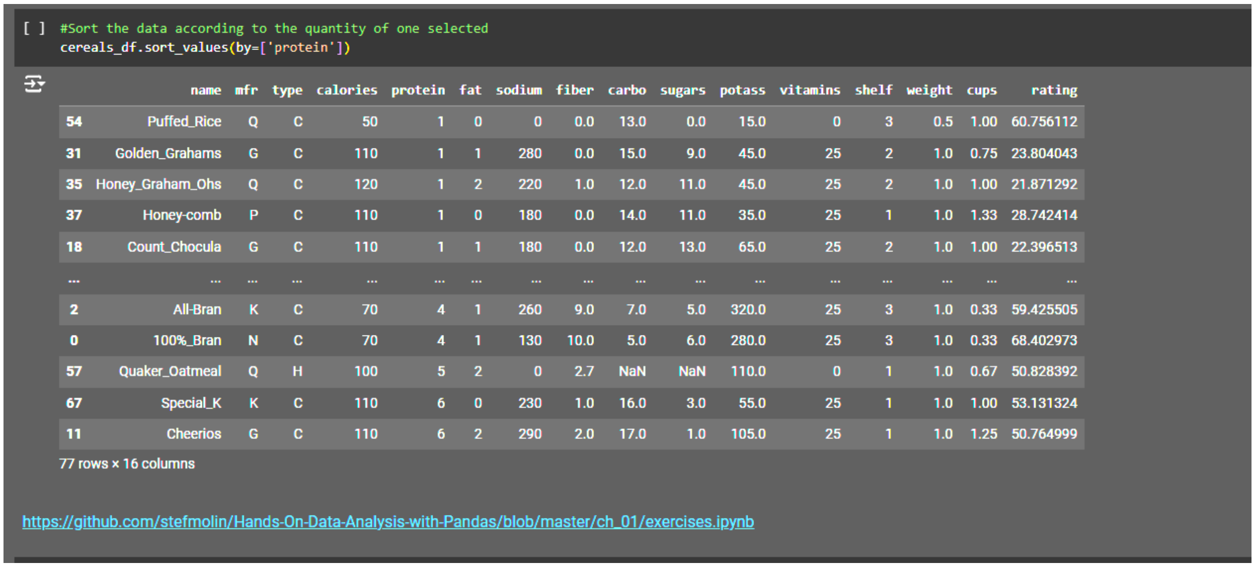
```
         calories    protein        fat      sodium      fiber      carbo  \
count   77.000000  77.000000  77.000000   77.000000  77.000000  76.000000
mean   106.883117   2.545455   1.012987  159.675325   2.151948  14.802632
std     19.484119   1.094790   1.006473   83.832295   2.383364   3.907326
min     50.000000   1.000000   0.000000    0.000000   0.000000   5.000000
25%    100.000000   2.000000   0.000000  130.000000   1.000000  12.000000
50%    110.000000   3.000000   1.000000  180.000000   2.000000  14.500000
75%    110.000000   3.000000   2.000000  210.000000   3.000000  17.000000
max    160.000000   6.000000   5.000000  320.000000  14.000000  23.000000

          sugars      potass    vitamins      shelf     weight       cups  \
count  76.000000   75.000000   77.000000  77.000000  77.000000  77.000000
mean    7.026316   98.666667   28.246753   2.207792   1.029610   0.821039
std     4.378656   70.410636   22.342523   0.832524   0.150477   0.232716
min     0.000000   15.000000    0.000000   1.000000   0.500000   0.250000
25%     3.000000   42.500000   25.000000   1.000000   1.000000   0.670000
50%     7.000000   90.000000   25.000000   2.000000   1.000000   0.750000
75%    11.000000  120.000000   25.000000   3.000000   1.000000   1.000000
max    15.000000  330.000000  100.000000   3.000000   1.500000   1.500000

          rating
count  77.000000
mean   42.665705
std    14.047289
min    18.042851
25%    33.174094
50%    40.400208
75%    50.828392
max    93.704912
         protein        fat
count  38.000000  38.000000
mean    2.342105   1.078947
std     1.121686   1.171314
min     1.000000   0.000000
25%     1.250000   0.000000
50%     2.000000   1.000000
75%     3.000000   2.000000
max     6.000000   5.000000
```

```
[ ] #Sort the data according to the quantity of one selected
    cereals_df.sort_values(by=['protein'])
```

e) Sort the data according to the quantity of one selected



```
[ ] #Sort the data according to the quantity of one selected
    cereals_df.sort_values(by=['protein'])
```

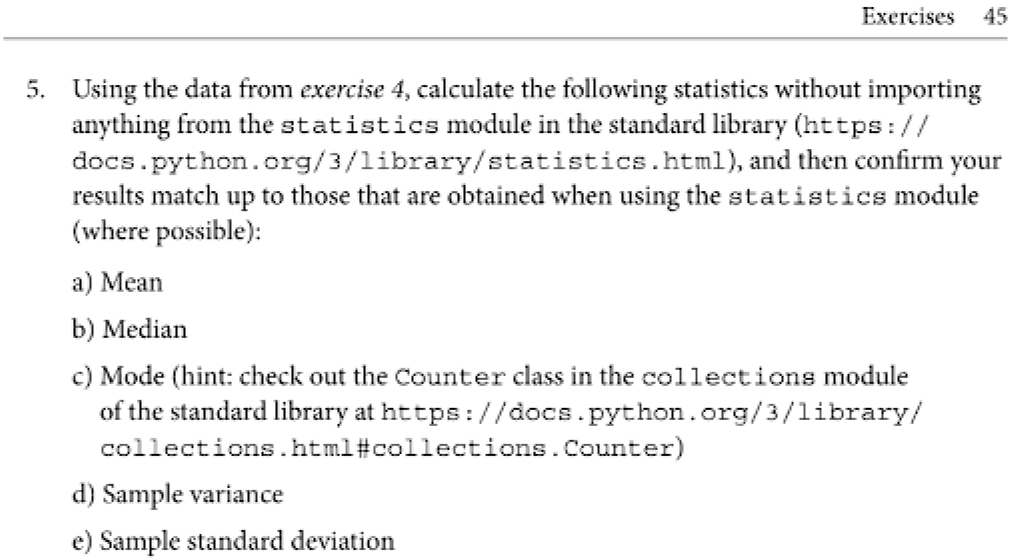
	name	mfr	type	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	weight	cups	rating
54	Puffed_Rice	Q	C	50	1	0	0	0.0	13.0	0.0	15.0	0	3	0.5	1.00	60.756112
31	Golden_Grahams	G	C	110	1	1	280	0.0	15.0	9.0	45.0	25	2	1.0	0.75	23.804043
35	Honey_Graham_Ohs	Q	C	120	1	2	220	1.0	12.0	11.0	45.0	25	2	1.0	1.00	21.871292
37	Honey-comb	P	C	110	1	0	180	0.0	14.0	11.0	35.0	25	1	1.0	1.33	28.742414
18	Count_Chocula	G	C	110	1	1	180	0.0	12.0	13.0	65.0	25	2	1.0	1.00	22.396513
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2	All-Bran	K	C	70	4	1	260	9.0	7.0	5.0	320.0	25	3	1.0	0.33	59.425505
0	100%_Bran	N	C	70	4	1	130	10.0	5.0	6.0	280.0	25	3	1.0	0.33	68.402973
57	Quaker_Oatmeal	Q	H	100	5	2	0	2.7	NaN	NaN	110.0	0	1	1.0	0.67	50.828392
67	Special_K	K	C	110	6	0	230	1.0	16.0	3.0	55.0	25	1	1.0	1.00	53.131324
11	Cheerios	G	C	110	6	2	290	2.0	17.0	1.0	105.0	25	1	1.0	1.25	50.764999

77 rows × 16 columns

https://github.com/stefmolin/Hands-On-Data-Analysis-with-Pandas/blob/master/ch_01/exercises.ipynb

Answer: The above questions should be completed correctly in this notebook.

Question 6:



Exercises 45

5. Using the data from *exercise 4*, calculate the following statistics without importing anything from the `statistics` module in the standard library (`https://docs.python.org/3/library/statistics.html`), and then confirm your results match up to those that are obtained when using the `statistics` module (where possible):

   a) Mean

   b) Median

   c) Mode (hint: check out the `Counter` class in the `collections` module of the standard library at `https://docs.python.org/3/library/collections.html#collections.Counter`)

   d) Sample variance

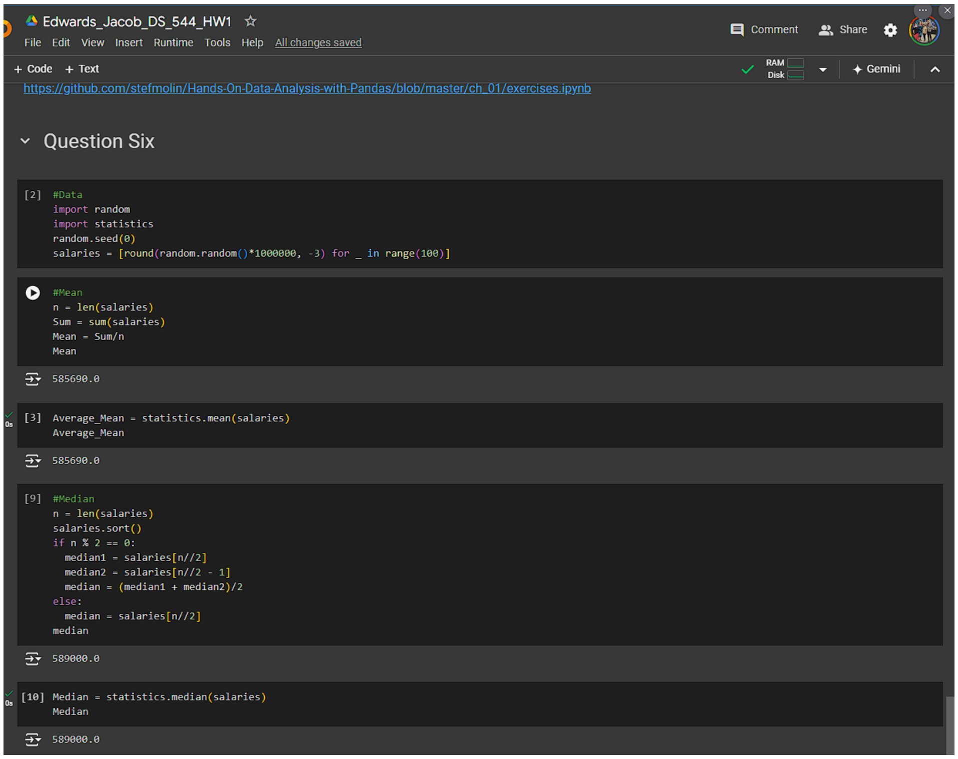
   e) Sample standard deviation

Answer:

A&B)

DS 544: Homework #1

Jacob Edwards



https://github.com/stefmolin/Hands-On-Data-Analysis-with-Pandas/blob/master/ch_01/exercises.ipynb

Question Six

```
[2] #Data
import random
import statistics
random.seed(0)
salaries = [round(random.random()*1000000, -3) for _ in range(100)]
```

```
#Mean
n = len(salaries)
Sum = sum(salaries)
Mean = Sum/n
Mean
```

```
585690.0
```

```
[3] Average_Mean = statistics.mean(salaries)
Average_Mean
```

```
585690.0
```

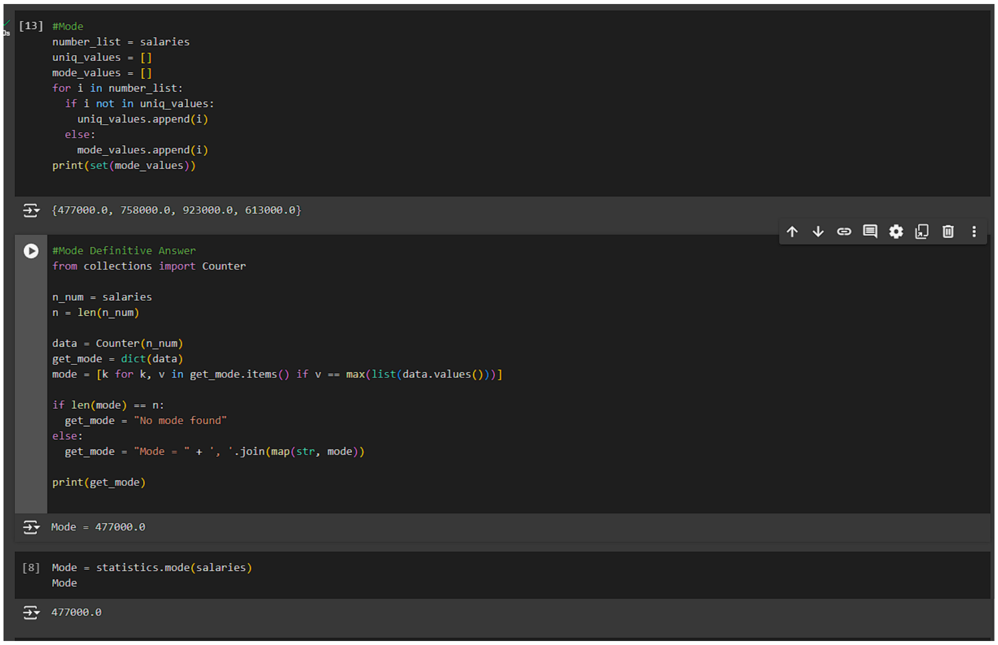
```
[9] #Median
n = len(salaries)
salaries.sort()
if n % 2 == 0:
  median1 = salaries[n//2]
  median2 = salaries[n//2 - 1]
  median = (median1 + median2)/2
else:
  median = salaries[n//2]
median
```

```
589000.0
```

```
[10] Median = statistics.median(salaries)
Median
```

```
589000.0
```

C)



```
[13] #Mode
number_list = salaries
uniq_values = []
mode_values = []
for i in number_list:
  if i not in uniq_values:
    uniq_values.append(i)
  else:
    mode_values.append(i)
print(set(mode_values))
```

```
{477000.0, 758000.0, 923000.0, 613000.0}
```

```
#Mode Definitive Answer
from collections import Counter

n_num = salaries
n = len(n_num)

data = Counter(n_num)
get_mode = dict(data)
mode = [k for k, v in get_mode.items() if v == max(list(data.values()))]

if len(mode) == n:
  get_mode = "No mode found"
else:
  get_mode = "Mode = " + ', '.join(map(str, mode))

print(get_mode)
```

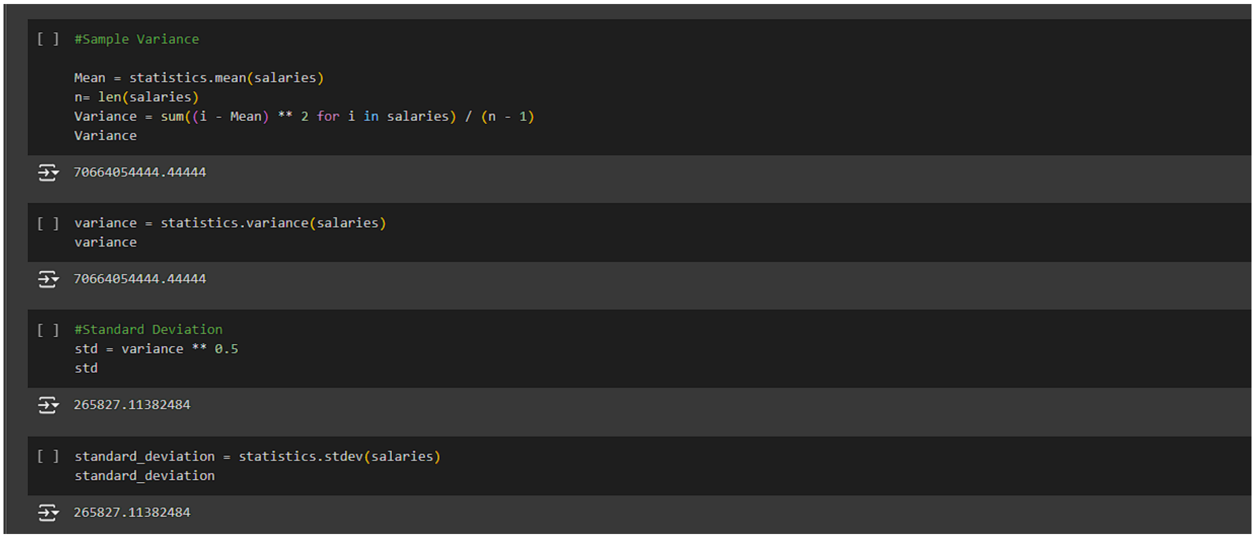
```
Mode = 477000.0
```

```
[8] Mode = statistics.mode(salaries)
Mode
```

```
477000.0
```

D & E)

DS 544: Homework #1
Jacob Edwards



```
#Sample Variance

Mean = statistics.mean(salaries)
n= len(salaries)
Variance = sum((i - Mean) ** 2 for i in salaries) / (n - 1)
Variance
```

```
70664054444.44444
```

```
variance = statistics.variance(salaries)
variance
```

```
70664054444.44444
```

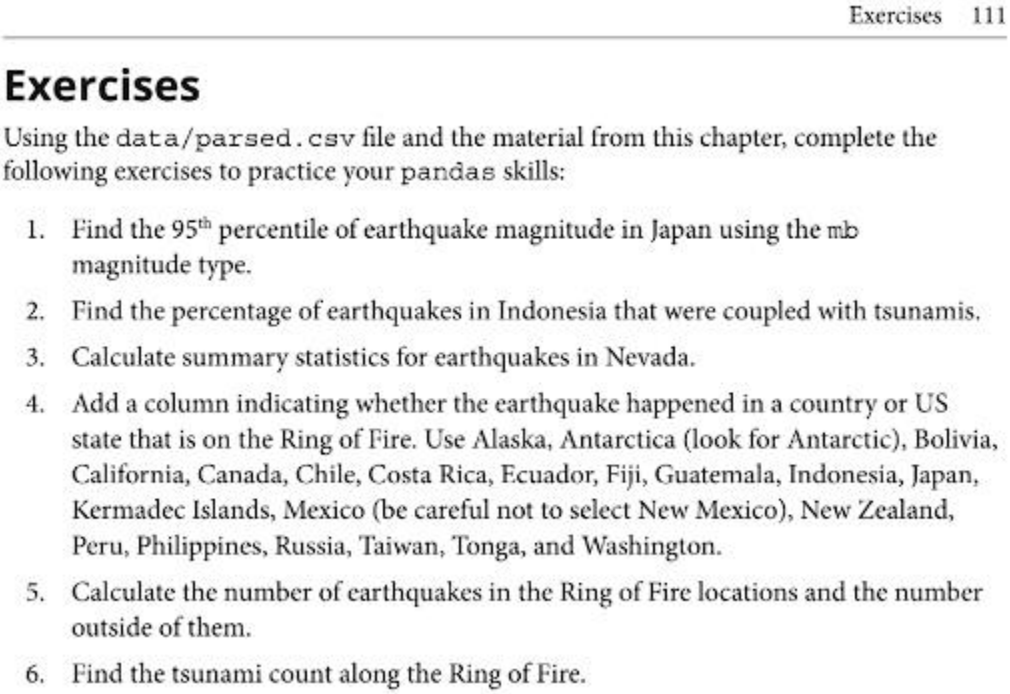
```
#Standard Deviation
std = variance ** 0.5
std
```

```
265827.11382484
```

```
standard_deviation = statistics.stdev(salaries)
standard_deviation
```

```
265827.11382484
```

Question 7:



Exercises 111

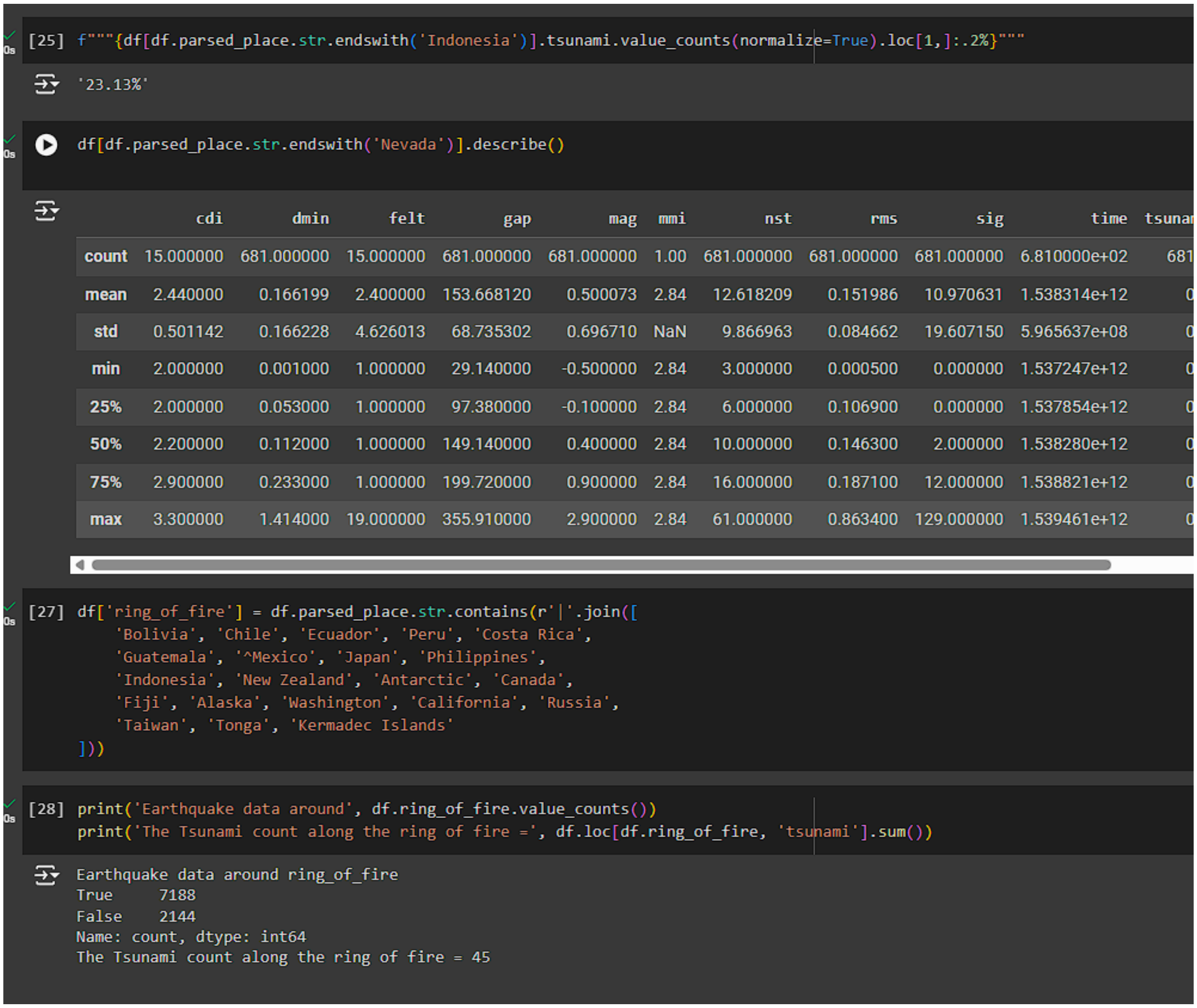
Exercises

Using the `data/parsed.csv` file and the material from this chapter, complete the following exercises to practice your `pandas` skills:

1. Find the 95th percentile of earthquake magnitude in Japan using the `mb` magnitude type.
2. Find the percentage of earthquakes in Indonesia that were coupled with tsunamis.
3. Calculate summary statistics for earthquakes in Nevada.
4. Add a column indicating whether the earthquake happened in a country or US state that is on the Ring of Fire. Use Alaska, Antarctica (look for Antarctic), Bolivia, California, Canada, Chile, Costa Rica, Ecuador, Fiji, Guatemala, Indonesia, Japan, Kermadec Islands, Mexico (be careful not to select New Mexico), New Zealand, Peru, Philippines, Russia, Taiwan, Tonga, and Washington.
5. Calculate the number of earthquakes in the Ring of Fire locations and the number outside of them.
6. Find the tsunami count along the Ring of Fire.

Answer:

DS 544: Homework #1
Jacob Edwards



```
[25] f"""{df[df.parsed_place.str.endswith('Indonesia')].tsunami.value_counts(normalize=True).loc[1,]:.2%}"""
```

```
'23.13%'
```

```
df[df.parsed_place.str.endswith('Nevada')].describe()
```

	cdi	dmin	felt	gap	mag	mmi	nst	rms	sig	time	tsunar
count	15.000000	681.000000	15.000000	681.000000	681.000000	1.00	681.000000	681.000000	681.000000	6.810000e+02	681
mean	2.440000	0.166199	2.400000	153.668120	0.500073	2.84	12.618209	0.151986	10.970631	1.538314e+12	0
std	0.501142	0.166228	4.626013	68.735302	0.696710	NaN	9.866963	0.084662	19.607150	5.965637e+08	0
min	2.000000	0.001000	1.000000	29.140000	-0.500000	2.84	3.000000	0.000500	0.000000	1.537247e+12	0
25%	2.000000	0.053000	1.000000	97.380000	-0.100000	2.84	6.000000	0.106900	0.000000	1.537854e+12	0
50%	2.200000	0.112000	1.000000	149.140000	0.400000	2.84	10.000000	0.146300	2.000000	1.538280e+12	0
75%	2.900000	0.233000	1.000000	199.720000	0.900000	2.84	16.000000	0.187100	12.000000	1.538821e+12	0
max	3.300000	1.414000	19.000000	355.910000	2.900000	2.84	61.000000	0.863400	129.000000	1.539461e+12	0

```
[27] df['ring_of_fire'] = df.parsed_place.str.contains(r'|'.join([
         'Bolivia', 'Chile', 'Ecuador', 'Peru', 'Costa Rica',
         'Guatemala', '^Mexico', 'Japan', 'Philippines',
         'Indonesia', 'New Zealand', 'Antarctic', 'Canada',
         'Fiji', 'Alaska', 'Washington', 'California', 'Russia',
         'Taiwan', 'Tonga', 'Kermadec Islands'
     ]))
```

```
[28] print('Earthquake data around', df.ring_of_fire.value_counts())
     print('The Tsunami count along the ring of fire =', df.loc[df.ring_of_fire, 'tsunami'].sum())
```

```
Earthquake data around ring_of_fire
True     7188
False    2144
Name: count, dtype: int64
The Tsunami count along the ring of fire = 45
```

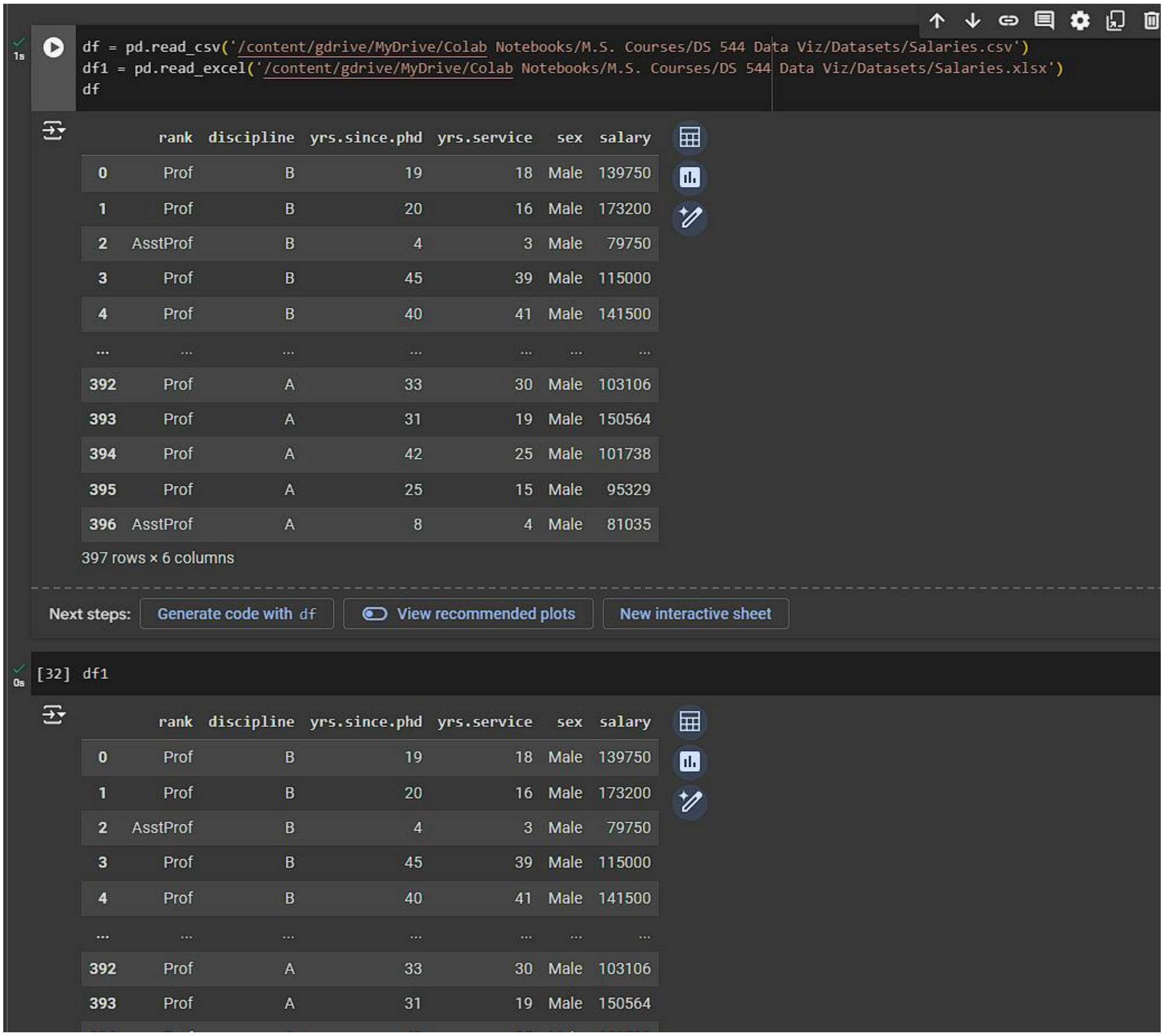
Question 8:

8. (3 points) Complete the following exercise using the skills you gained above. a. The two datasets Salaries.csv is stored in your shared folder. Download the file into your project work directory (pwd, e.g. chapter 2 of Molin's subfolder). b. (Use Pandas to read_csv to read the Salaries file into your working environment and assign to data frame as df (make sure the directory and file names are correctly spelled). c. Check the first 8 record of the Salaries file. d. Find the median and means of Assistant Professors, Associate Professors, and Full Professors. e. Find the median and means of the female and male professors.

DS 544: Homework #1

Jacob Edwards

A&B)



```
df = pd.read_csv('/content/gdrive/MyDrive/Colab Notebooks/M.S. Courses/DS 544 Data Viz/Datasets/Salaries.csv')
df1 = pd.read_excel('/content/gdrive/MyDrive/Colab Notebooks/M.S. Courses/DS 544 Data Viz/Datasets/Salaries.xlsx')
df
```

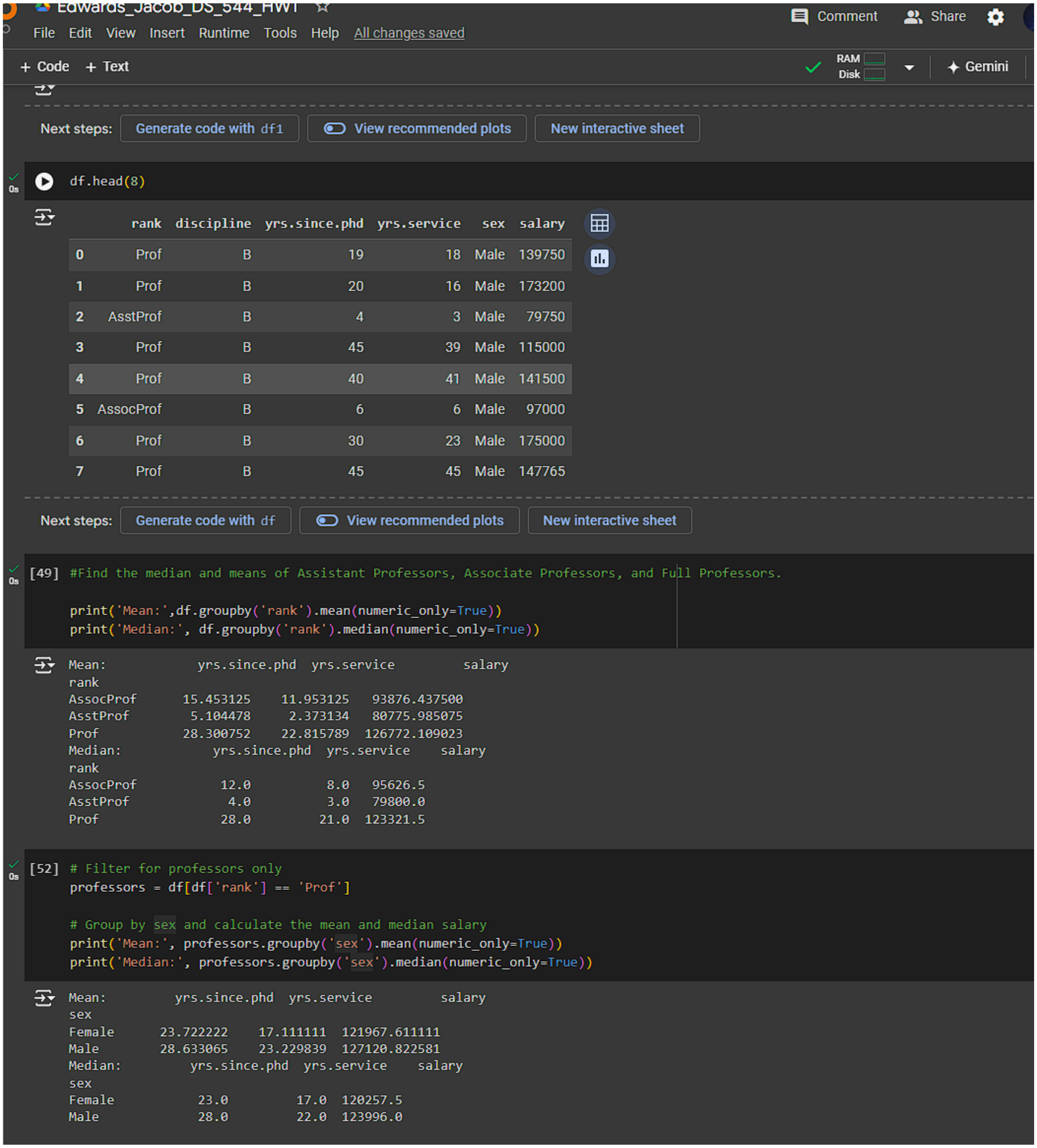
```
          rank  discipline  yrs.since.phd  yrs.service   sex  salary
  0       Prof           B             19           18  Male  139750
  1       Prof           B             20           16  Male  173200
  2   AsstProf           B              4            3  Male   79750
  3       Prof           B             45           39  Male  115000
  4       Prof           B             40           41  Male  141500
...        ...         ...            ...          ...   ...     ...
392       Prof           A             33           30  Male  103106
393       Prof           A             31           19  Male  150564
394       Prof           A             42           25  Male  101738
395       Prof           A             25           15  Male   95329
396   AsstProf           A              8            4  Male   81035
397 rows × 6 columns
```

```
[32] df1
```

```
          rank  discipline  yrs.since.phd  yrs.service   sex  salary
  0       Prof           B             19           18  Male  139750
  1       Prof           B             20           16  Male  173200
  2   AsstProf           B              4            3  Male   79750
  3       Prof           B             45           39  Male  115000
  4       Prof           B             40           41  Male  141500
...        ...         ...            ...          ...   ...     ...
392       Prof           A             33           30  Male  103106
393       Prof           A             31           19  Male  150564
```

C&D&E)

DS 544: Homework #1
Jacob Edwards



```
df.head(8)
```

	rank	discipline	yrs.since.phd	yrs.service	sex	salary
0	Prof	B	19	18	Male	139750
1	Prof	B	20	16	Male	173200
2	AsstProf	B	4	3	Male	79750
3	Prof	B	45	39	Male	115000
4	Prof	B	40	41	Male	141500
5	AssocProf	B	6	6	Male	97000
6	Prof	B	30	23	Male	175000
7	Prof	B	45	45	Male	147765

```
[49] #Find the median and means of Assistant Professors, Associate Professors, and Full Professors.

print('Mean:',df.groupby('rank').mean(numeric_only=True))
print('Median:', df.groupby('rank').median(numeric_only=True))
```

```
Mean:            yrs.since.phd  yrs.service         salary
rank
AssocProf      15.453125    11.953125   93876.437500
AsstProf        5.104478     2.373134   80775.985075
Prof           28.300752    22.815789  126772.109023
Median:            yrs.since.phd  yrs.service    salary
rank
AssocProf           12.0          8.0   95626.5
AsstProf             4.0          3.0   79800.0
Prof                28.0         21.0  123321.5
```

```
[52] # Filter for professors only
professors = df[df['rank'] == 'Prof']

# Group by sex and calculate the mean and median salary
print('Mean:', professors.groupby('sex').mean(numeric_only=True))
print('Median:', professors.groupby('sex').median(numeric_only=True))
```

```
Mean:         yrs.since.phd  yrs.service         salary
sex
Female      23.722222    17.111111  121967.611111
Male        28.633065    23.229839  127120.822581
Median:         yrs.since.phd  yrs.service    salary
sex
Female           23.0         17.0  120257.5
Male             28.0         22.0  123996.0
```

Question 9:

(2 points) Read the article by Gelman and Unwin, Infovis and Statistical Graphics: Different Goals, Different Looks.

DS 544: Homework #1
Jacob Edwards

Answer the following questions:
Use a few pairs of sentences or a table to compare the different goals and different views between statisticians and information visualization designers. What type of attitudes does the author suggest that we should take in regard to the differences?

Answer: The different views that are discussed in this article for the statistician's side of things really boil down to conveying data accurately and enabling comparisons to understand patterns. It is prioritizing precision and clarity, overall. While the Infovis side is mainly focused on telling a story and making the data interesting for the viewer. It's mainly focused on leveraging different design principles to create visually appealing graphics. As for the type of attitude the author suggests the reader should take in regarding the differences is that both sides have value and their respective approaches can be complementary of one another; however, there are also some potential shortcomings. Such shortcomings could be described as unappealing to a wider audience for the statistician's side, and the Infovis side might provide nice visually aesthetic designs that could potentially mislead the viewer. Overall, communication is key, and we want to provide accurate and clear visualizations. There certainly is an art to finding the balance between the two perspectives.

Question 10:

After reading the above article, answer the following questions:

a. What are your thoughts on the pros and cons of Wordle, which main purpose does the Wordle graph in Figure 3 serve?
b. Comparing figure 5 FLORENCE NIGHTINGALE'S COXCOMB and line plot figure 6.
• Which one do you prefer, and why?
• Who are the potential readers? What are the purposes of the graphic designers of Figures 5 and 6?
c. Compare Figure 9 THE BABY NAME WIZARD, and histogram Figure 10,
• Which one do you prefer, and why?
• Who are the potential readers? What are the purposes of the graphic designers of Figures 5 and 6?
d. Why the author concluded that "the progress in InfoVis important—should be important—to statisticians." Alternatively, what should statistician gain from the progress in InfoVis?

Answer:

DS 544: Homework #1
Jacob Edwards

A) The primary purpose of the Wordle graph in Figure 3 is to provide a quick and visually appealing overview of the most frequent words in the text. Like how a Word Cloud works.

B) I absolutely prefer the line plot in figure 6 over figure 5's graph. The figure 5 graph is hard to read what's happening, plus each of the two figures are different sizes. I can't make out or distinguish what it's wanting to show as easily. The purpose for figure 5 is to make it more visually appealing because it's unlike most graphs I've ever seen; however, the figure 6 plot shows a real representation that is able to be read and interpreted easier than 5. The potential readers are those doing research on this time period and the mortality rates associated with those wars.

C) A very similar answer to part B quite honestly. The figure 9 visualization shows a very appealing graph with a color-coded graph that isn't very easily readable for accuracy; however, it does provide eye catching information. With that being their goal, they do an exceptional job of it. I think the histograms shown in figure 10 are specific. They are showing growth in comparison to the last letter of boys' names in two charts separated by 60 years' time. It is very easily distinguishable what the data represents, and the accuracy of it is easy to read, for the most part. The only thing that I would critique is the y-axis are not the same. It does provide a little misleading information with the y-axis not being the same and the height of the bars. At first glance, someone who isn't used to reviewing charts may not catch this difference. As such, it could be interpreted as a little misleading. I think the purpose of each of these graphic designers for these figures is simply to convey information in each of their respective perspectives. The potential readers are parents who are looking to name their children in the future most likely if they are planning on having children.

D) In summary, statisticians can gain from the progress in InfoVis by:
Improving their ability to communicate their findings effectively. Discovering new insights in their data. Enhancing their data exploration capabilities. Fostering collaboration with other fields. (Andrew Gelman & Antony Unwin, 2013)

The code in my .ipynb file can be found below. I have also attached a link here to the Colab notebook for reference.

Edwards_Jacob_DS_544_HW1 (1).ipynb

DS 544: Homework #1
Jacob Edwards

References

Andrew Gelman & Antony Unwin (2013) Infovis and Statistical Graphics: Different Goals, Different Looks, Journal of Computational and Graphical Statistics, 22:1, 2-28, DOI: 10.1080/10618600.2012.761137

Kumar, A. (2024, March 20). *Tableau Server vs. Tableau Desktop: Comparison*. Senturus. https://senturus.com/blog/tableau-server-vs-desktop-comparison-2/

Learn Statistics Easily. (2024, February 6). *7 myths about statistics you need to stop believing*. LEARN STATISTICS EASILY. https://statisticseasily.com/7-myths-about-statistics/#:~:text=Highlights%201%20Correlation%20%E2%89%A0%20causation.%202%20Low%20p-value,matters.%207%20Not%20all%20stats%20are%20universally%20applicable.

Agrawal, Rishabh. (2024, January 28). *Finding mean, median, mode in python without libraries*. GeeksforGeeks. https://www.geeksforgeeks.org/finding-mean-median-mode-in-python-without-libraries/

Statistics How To. (2024, July 20). *Misleading graphs: Real life examples*. https://www.statisticshowto.com/probability-and-statistics/descriptive-statistics/misleading-graphs/