Clarification questions

Please ask any question related to this assignment in the class or in the Canvas forum. To be fair to everyone, it is best to ask assignment question openly rather than in email. Do not share your code with anyone. Do not post your code online. Do not ask for help in any public discussion forum or online website. You can search over the internet and read books. Let me know if you find any error in the assignment question as soon as possible. DO NOT POST/SHARE OR DESCRIBE YOUR OUTPUT TO ANY OTHER STUDENT. You can always ask questions about the methods, definitions or approaches to solve a question. DO NOT ASK ME WHETHER MY OUTPUT LOOKS CORRECT.

Assignment 2

What to submit?

Q1: q1.csv, q1.py, q1-report.pdf, q1.jpg (screenshot of the visuzliation)

Q2: mds.html (edit the given code), q2.jpg (screenshot of the visuzliation)

Q3: boxplot.html (edit the given code), q3.jpg (screenshot of the visuzliation)

Do not log any unnecessary values in the console. Do not submit unnecessary files or the files that are given to you.

Q1: GOAL: Goal driven design, Mark: 50

You are given a csv file of food price data in Canada. This has been downloaded as 18100004.csv from Statcan website (`ttps://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1810000401`).

Step 1 [Use of AI-generated code allowed]: Keep only the rows corresponding to 'Saskatchewan', 'Regina, Saskatchewan' and 'Saskatoon, Saskatchewan'. Consider only the data from January 2000 to analyze the recent price trend. Remove all Products that start with 'All-items ...'.

Find the yearly mean price for each product as follows:

```
year,     Product,               Mean_Price
2000,     Alcoholic beverages,   96.0
2001,     Alcoholic beverages,   96.3
---
---
```

Step 2 [Use of AI-generated code allowed]: Create a csv file q1.csv that contains the yearly mean price for each Product as well as a stat column. If there is a Product for which we do not have data for all years, then you can skip that item for simplicity.

```
Product,               2000, 2001, ....,          Stat
Alcoholic beverages,   96.0, 96.3,     ,           x
---
---
```

For each item (equivalently, product), its stat-column value corresponds to the **maximum price change observed over the years compared to its typical value**. Note that this is not a precise definition and you need to design a reasonable statistics that would help reveal the products that experienced high price jump when we color based on the stat column.

Here is an example figure where blue lines show the products that experience high price jump. The picture is a rough sketch without providing you all the details. Your visualization may be different than the one you see here depending on the stat function you use.

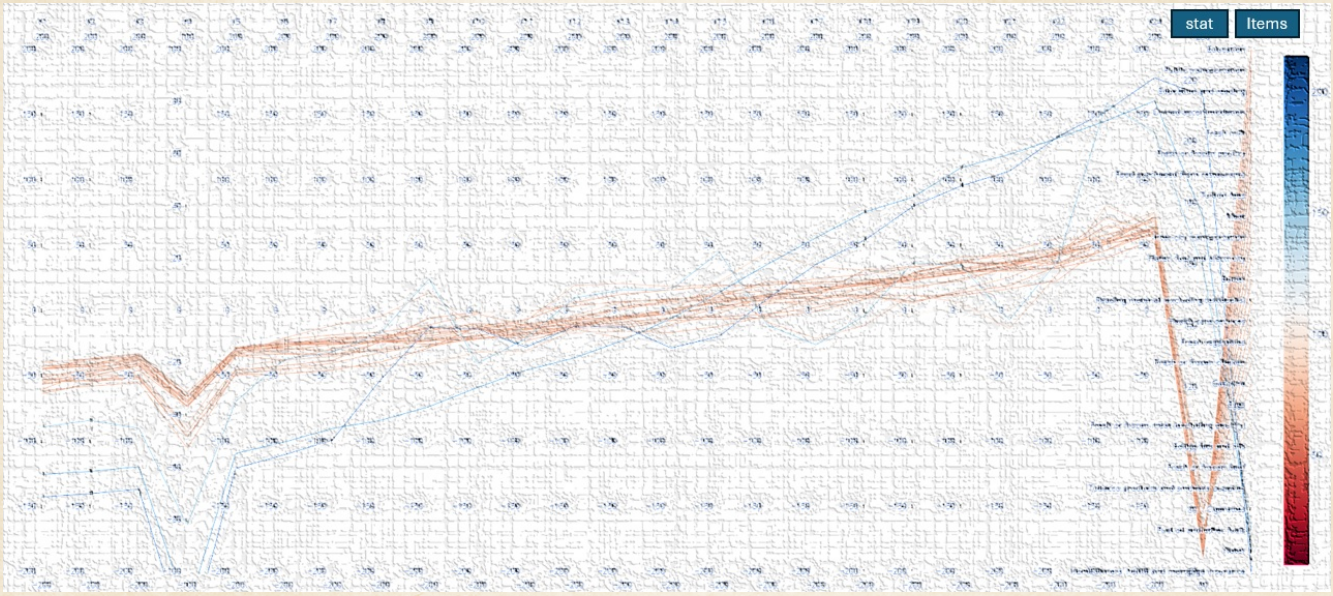

Figure 1: Caption

Step 3 [Use of AI-generated code allowed]: Write a python file q1.py that reads your q1.csv and plots a parallel coordinate view. Only plot the top 25 products that have the highest price jump between 2000–2024 according to the stat value you computed. Use the 'RdBu' color map for your plot. Here is a tutorial: `https://plotly.com/python/parallel-coordinates-plot/`

If you have questions regarding how to plot categorical axis or how to change color scale, you can ask on canvas.

Step 4 [Use of AI NOT allowed, write in your own words]: Explore the parallel coordinate visualization with brushing interaction to find interesting information. You should be able to explore the data using the code q1.py that you wrote. You can also optionally upload q1.csv here (if you were not able to complete Step 3).

`https://syntagmatic.github.io/parallel-coordinates/examples/upload.html`

Submit a report (at most 2 pages). Your report should include the following.

(a) A clear description of the stat function that you used in your visualization.

(b) A picture of your visualization and a description of the information that can be seen from your chart. Your description must be based on the chart that you created. Do not assume that the picture automatically tells us everything — describe explicitly.

(c) Show how interactions can reveal important information. Try to give a few (at least two concrete) interesting observations with pictures as evidence.

Q2. Cluster Visualization by MDS (25 Marks) [AI NOT allowed]

You are given a data file mds.csv that contains grades of the students in four assignments A, B, C, D. The column E indicates the current grade in the course.

Your task is to apply MDS to find a scatterplot of a 4 dimensional dataset [A,B,C,D]. You must edit mds.html to complete this assignment. You are already given the mds.js and numeric.js that compute the MDS-coordinates for you. You dont have to worry about them at all. If you look at the top part of the html file, then you will find out that these files are already included in your project. So keep them along with your html (in the same directory).

You will first need to produce the squared distance matrix (see the lecture slide example). Represent it as a 2D array. Each entry i,j of the distance matrix needs to be filled with a distance between the response of person i and person j.

If everything is correct, you can refresh your browser to see a scatterplot, where the color of each point is assigned based on column E. You will see that MDS automatically separates students of different grades from each other.

Q3: GOAL: Boxplot in D3, Mark : 25 [Use of AI-generated code allowed]

Write a program that reads the data file boxplot.csv and draws a boxplot for each column. Try to thoroughly understand the given code.

Submission Instruction: You must edit the given HTML file to complete the task. **Try to adapt the given code.** Submit the updated file and a screenshot of the visualization. Clicking on the file must generate the output on the browser. Do not log any unnecessary values in the console.