Homework 1

Due: March 30, 2020, 11:59 PM (before midnight)

Part 1:

Categorize each plot in The Data Visualisation Catalogue into 8 types of Quantitative messages (for example Bar Plot can be used in Ranking and Nominal Comparison plots) or if it does not fit into the 8 types, disregard them. After that, find what Visual Cues are used in that plot (for example Bar Plot uses Position (the y axis is the same for all the bars) and Length (we are comparing the length of the bars, not their areas)).

Quantitative messages:

- Time-series
- Ranking
- Part-to-whole
- Deviation
- Frequency distribution
- Correlation
- Nominal comparison
- Geographic or geospatial

Visual Cues:

1. Position (common scale)
2. Position (non-aligned scale)
3. Length
4. Direction
5. Angle
6. Area
7. Volume
8. Curvature
9. Shading
10. Color Saturation

Read chapters 3 and 4 from Data Points Visualization That Means Something by Nathan Yau. Book in the drive folder.

Part 2.

Read the description of the data from the source. Understand what each variable shows and their data type (continues, ordinal, nominal, etc.). Data source: https://archive.ics.uci.edu/ml/datasets/Heart+Disease. But download and use data from Drive folder.

1) Merge the heart.csv and target.csv. In the target.csv there are more observations than in heart.csv. Make sure you include all the observations that are both in heart.csv and target.csv. Save the data in data folder with name “final_data.csv”. Make sure you have no additional columns created after saving the data (for example “Untitled: 0”). Read the newly saved data and continue.
2) Separate and write down continuous, ordinal and nominal variables.
3) Create two correlation heatmaps of continuous variables using seaborn. Use the Pearson correlation. The first plot should include only the absolute values of the correlation coefficients, and the plot should have only one color scale (for example shades of blue). Second should include also the sign of the correlation coefficient and should have 2 color scales (all negative correlations should have red color, the higher the correlation the saturated the hue of the red should be; all positive correlations should have blue color (the higher the absolute value of the correlation coefficient, the more saturated should the hue of the blue be).
4) Make subplots of the distribution plots for continuous variables using matplotlib.
5) Use groupby or pivot_table to get the average cholesterol level and the average age in each sex group. Plot two different bar plots.
6) Use pd.cut() (or an alternative method) to group age variable into 0-10, 10-20, 70-80 etc groups. Create a new variable named “age_group” where 0 should be the value for the youngest group, 1 for the second youngest, etc.
7) For each age_group calculate the average cholesterol level and the standard deviation of cholesterol level in that group. Plot the result using error bar plot, where the length of the bar should be the average cholesterol level, and the standard deviation of the cholesterol level should be the error.
8) For each age_group calculate the percentage of people with heart disease in that group. Plot the results as a bar plot.
9) For each age_group calculate the percentage of people with heart disease from all observations. Plot the results as a bar plot.
10) Create a stacked bar plot, where the total length of the bar should be the percentage of people with heart disease in that age_group. And the bars should be separated into sex groups. For example, if in the first age_group there are 100 people, out of which 60 have heart disease, and out of those 60, 10 are female and 50 are male. Then the total length of the bar should be 60 (or 0.6), from which 10 (or 0.1) should have one color and the rest 50 (or 0.5) have another color. Where color represents the sex group.
11) Create a scatterplot for each continuous variable using seaborn and add a regression line with standard error included.

12) Create a scatterplot for each continuous variable and group each point as belonging to either has heart disease or does not have heart disease.

For part two, all your plots must have a plot title, x and y axis labels. And some comments about the plots (what it shows etc.)

Submit your files in one .zip (.rar files will not be accepted). The zip file must include all your jupyter notebooks, saved plots, and the data. In the case of plagiarism, all parties involved will get 0 for the whole assignment. Late submission will not be graded (grade will be 0) but you will get feedback on your assignment.