CSC380 **Homework 5** **Fall 2024**

This homework is due on Sunday, November 10^{th} at 11.59pm. Late submissions are not accepted.

Submission Guidelines

- You must make two submissions:
  - Your complete homework as a SINGLE PDF file by the stated deadline to the gradescope. **Include your code and output of the code as texts in the PDF.**
  - Your codes to a separate submission: a single notebook file including codes for questions 2 through 6.
- For your PDF submissions:
  - Select the page number for the answer to each question in the Gradescope.
  - You may submit typed or handwritten/scanned answers. If you decide to submit handwritten answers then please ensure that it is easily readable.
  - You can easily scan and upload your answers as a PDF using the Gradescope mobile app.

1. **[12pts]** Visit http://www.wtfviz.net. Find two visualizations to critique. Provide the visualizations you selected and your comments on them. You need to make use of the jargon employed in class and in Chapter 6 of the textbook in your discussions. Your comments on each visualization should answer the following questions:

   - What data is represented in this visualization? Be specific.
   - What questions does the visualization answer?
   - Describe one aspect of the visualization that is effective.
   - Describe one aspect of the visualization that is not effective.
   - Why do you like/dislike the visualization?

2. **[12pts]** We will make use of the ted_main data set for this problem. Use Matplotlib to construct scatter plots for four sets of data points, where each data point consists of log transformed value of the *comments* (x-coordinate) and log transformed value of the *views* (y-coordinate) fields. The difference in the sets of data points is the filtering; one is filtered with respect to talks with at least 100 comments, one with at least 500 comments, one with at least 1000 comments and one with no restriction at all (all the rows of the data frame).

   Experiment with the point size parameter of the scatter plot function to find the most revealing value for each data set.

3. **[12pts]** In this problem you will use Matplotlib to experiment with different color scales to construct scatter plots for a particular set of (x, y, z) points, where color is used to represent the z dimension. Use the log transformed values of the *comments*, the *views* fields, and the *duration* fields of the ted_main dataset as the x, y, z coordinates respectively, where a filtering is applied to work on the data points with at least 750 comments. Which color schemes work best? Which are the worst? Explain why.

4. **[12pts]** Create a violin plot of the *duration* field using the seaborn library. Comment on the information provided by the visualization including the major changes in the distribution of the values, median value, where the middle 50% of the values are located, and at what value do we start seeing the outliers.

5. **[12pts]** Download the 2012 London Summer Olympics Data from here. Create a 2D heatmap visualization of the data using the seaborn library. More specifically one axes of the 2D heat map should consist of the rows (countries) of the data set, whereas the other axes consists of the columns (GDP, population etc.) of the dataset. You may want to apply some type of normalization such as z-score normalization on the columns of the dataset to get the same color gradient scale for different columns.

6. **[12pts]** Using the matplotlib library create a grouped bar chart for the olympics data where each country is shown at the x-axis with the names of the countries as the labels and each country has a group of three bars depicting its number of gold, silver, and bronze medals. Check here for an example of grouped bar charts.

7. **[6pts]** Suppose you build a classifier that answers *yes* on every possible input. What precision and recall will this classifier achieve?

8. **[10pts]** Suppose $f \leq 1/2$ is the fraction of positive elements in a classification. What is the probability p that the blind classifier should guess positive, as a function of f, in order to maximize the specific evaluation metric below? Report both p and the expected evaluation score the blind classifier achieves.

(a) Accuracy.

(b) Precision.

(c) Recall.

(d) F-score.

9. **[6pts]** Suppose we want to train a binary classifier where one class is very rare. Give an example of such a problem. How should we train this model? What metrics should we use to measure performance?

10. **[6pts]** What is cross-validation? How might we pick the right value of k for k-fold cross validation?