CS 412: Fall'23
Introduction To Data Mining

Assignment 1

(Due Monday, September 25, 11:59 pm)

- The homework is due on Monday, September 25, at 11:59 pm. We will be using Gradescope for homework assignments. If you still having issues with joining Gradescope, you may use the code shared on Canvas Announcements date September 5. Please do NOT email a copy of your solution. Contact the TAs if you are having technical difficulties in submitting the assignment. Unfortunately, We will NOT accept late submissions without a reasonable justification.

- Please use Slack or Canvas first if you have questions about the homework. You can also join office hours and/or send us emails. If you are sending us emails with questions on the problems, please start the subject with "CS 412 Fall'23: " and send the email to *all of us* (Arindam, Ruby, Hyunisk, Rohan, Kowshika, Sayar) for faster response.

- Please write down your solutions entirely by yourself and make sure the answers are clear. The homework should be submitted in pdf format; there is no need to submit source code about your computing. You are expected to typeset the solutions, i.e., **handwritten solutions will not be graded**. We encourage you to use Latex.

- For each question, you will NOT get full credit if you only give a final result. Please show the necessary steps and details, including any formulae you use, what the variables mean in the formulae, any derivation or calculation steps, and explanations as appropriate. Consider the following two examples:

Example 1 **Q:** Given a dataset $\mathcal{X} = \{3.1, 4.2, -1\}$, compute the mean.
A: For any set of n numbers $\mathcal{X} = \{x_1, \ldots, x_n\}$, the mean can be computed as $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$. For the given dataset $\mathcal{X}$, the mean is $\mu = \frac{3.1+4.2-1}{3} = 2.1$

Example 2 **Q:** A coin claimed to be unbiased has been tossed 100 times, with 54 heads and 46 tails. What is the χ^2 statistic?
A: For a categorical variable taking k possible values, if the expected values are $e_i, i = 1, \ldots, k$ and the observed values are $o_i, i = 1, \ldots, k$, then the χ^2 statistic can be computed as: $S = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{o_i}$. For the problem, since the coin is claimed to be unbiased, the expected values are $50, 50$. Further, the observed values are $54, 46$. Then, the chi-squared statistic is given by $S = \frac{(54-50)^2}{50} + \frac{(46-50)^2}{50} = 0.64$.

- Please make sure to clearly mark each question. You may use as many pages as needed. Please do not change the order of the questions and answers.

- When submitting Assignment 1, Gradescope will ask you to select pages for each problem; please do this precisely!

- All data for the assignment can be downloaded from Canvas (`https://canvas.illinois.edu/courses/37840/assignments`) Assignment 1.

1. (25 points) A wellness clinic tested the age and body fat of 20 randomly selected adults. The results are summarized in the following table:

age	23	23	27	27	39	41	47	49	50	52
%fat	8.7	27.2	7.4	17.9	31.8	24.9	27.4	27.2	31.2	34.6
age	54	54	54	56	57	58	58	60	61	65
%fat	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7	37.4	36.2

Table 1: The **age** and **%fat** for 20 randomly selected adults.

Based on the dataset, compute the following statistical descriptions for each **age** and **%fat**.. If the result is not an integer, round it to 3 decimal places.

Note: you may write a script to perform the computations. Moreover, you do not need to submit the Python script.

(a) (5 points) Mean and Standard Deviation.

(b) (9 points) First quartile Q1, median, and third quartile Q3.

(c) (4 points) Maximum and minimum.

(d) (3 points) Mode.

(e) (4 points) Using `matplotlib` in `Python`, draw the boxplot for **age** and **%fat**. *Note that you do not need to submit the Python script.*

2. (8 points) Consider the histogram representing a population's age in a small town (Figure 1). Approximately compute the median age in the town using the histogram. Show the details of how you are doing the computation and clearly define any intermediate variables you use.

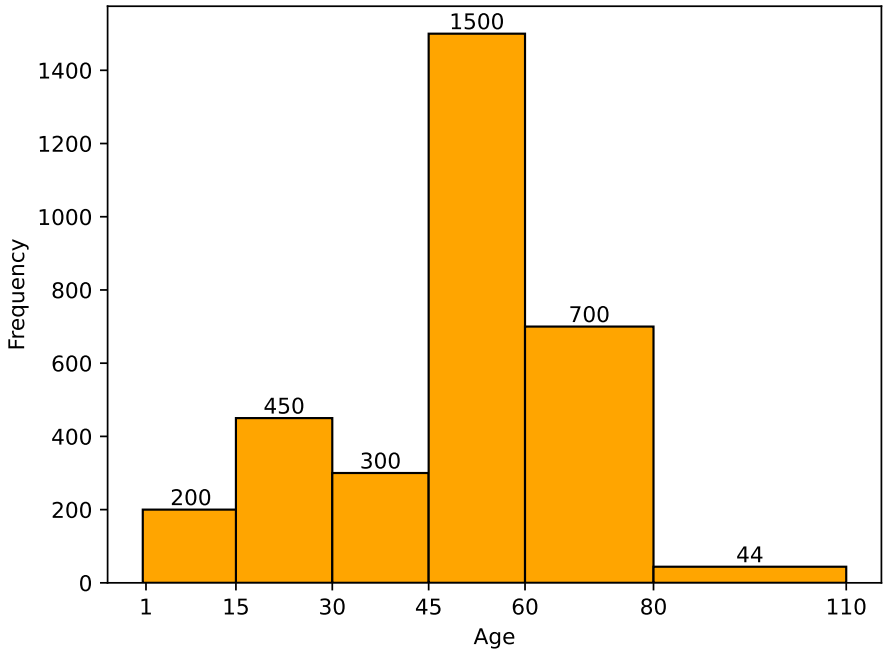

Figure 1: Histogram representing a population's age in a small town.

3. (17 points) Consider the dataset of the average temperatures in 1000 cities in January 1st, 1980 and January 1st, 2020 (file: `temperatures.csv`). The first column city represents the city name. The second column Jan-80 is the average temperature in January 1st, 1980, and the third column Jan-20 is the average temperature in January 1st, 2020.

   Please normalize the Jan-20 column using z-score normalization. We will refer to the original Jan-20 as Jan-20-original and the normalized temperatures scores as Jan-20-normalized. Moreover, we will refer to the original Jan-80 temperatures as Jan-80-original.

   Note: you may write a script to perform the computations. Moreover, you do not need to submit the Python script

   (a) (4 points) Compute and compare the variance of Jan-20-original and Jan-20-normalized, i.e., the temperatures before and after normalization.

   (b) (4 points) Compute the Pearson's correlation coefficient between Jan-80-original and Jan-20-original.

   (c) (4 points) Compute the Pearson's correlation coefficient between Jan-80-original and Jan-20-normalized.

   (d) (5 points) Compute the covariance between Jan-80-original and Jan-20-original.

4. (26 points) Given the number of positive COVID-19 cases in Monroe County, IN, and Tippecanoe County, IN, between 11/01/2020 - 02/08/2021 (file: covid-data-county.csv), we will compare the similarity between the two counties by using different proximity measures. The data for each county is for 100 days and contains information on how many cases per day each county had. When computing a similarity, if the result is not an integer, then round it to 3 decimal places.

   Note: you may write a script to perform the computations. Moreover, you do not need to submit the Python script

   (a) For each day between 11/01/2020 - 02/08/2021, the number of positive `COVID-19` cases are reported. Based on all the days (treat the 100 days data per county as a feature vector), compute the Minkowski distance of the vectors for `Monroe` and `Tippecanoe` concerning different h values:

      (i) (3 points) $h = 1$.

      (ii) (3 points) $h = 2$.

      (iii) (3 points) $h = \infty$.

   (b) (5 points) Compute the `Cosine` similarity between `Monroe` County and and `Tippecanoe` with regard to the feature vector.

   (c) (12 points) Using `matplotlib` in `Python`, draw a Bar Chart (similar to the example in Slide 37, Lecture 4) to visualize the total number of `COVID-19` positive cases in November 2020, December 2020, and January 2021 in Monroe County, IN, and Tippecanoe County, IN. The following are the chart's details:

      - `Y-axis` The month: November 2020, December 2020, and January 2021.
      - `X-axis` Number of `COVID-19` positive cases
      - For each month, there will be two bars. The first bar represents the number of positive `COVID-19` cases in Monroe County, IN and the second bar represents the number of positive `COVID-19` cases in Tippecanoe County, IN.
      - Chart title: The Number of `COVID-19` Positive Cases in Monroe County, IN and Tippecanoe County, IN from November 2020 until January 2021.

      Note that you do not need to submit the Python script

5. (24 points) Suppose that you are studying the relationship between (Not) Drinking Coffee and (Not) Eating Chocolate. You have distributed questionnaires, and the following results are collected from 2450 students on campus, as shown in Table 2. For the problem, we will treat both Drink Coffee and Eat Chocolate as binary attributes. Be sure to include necessary intermediate steps, e.g., formulas, variable references, and calculation results (report all values rounded to 4 places of decimal).

   Note: you may write a script to perform the computations. Moreover, you do not need to submit the Python script

	Drink Coffee	Do Not Drink Coffee
Eat Chocolate	205	20
Do Not Eat Chocolate	25	2200

Table 2: Contingency table for Drink Coffee and Eat Chocolate.

   (a) (5 points) Calculate the distance between the binary attributes Drink Coffee and Eat Chocolate by assuming they are symmetric binary variables.

   (b) (5 points) Calculate the Jaccard coefficient between Drink Coffee and Eat Chocolate.

   (c) (6 points) Compute the χ^2 statistic for the contingency table.

   (d) (8 points) Consider a hypothesis test based on the χ^2 statistic where the null hypothesis is that Drink Coffee and Eat Chocolate are independent. Can you reject the null hypothesis at a significance level of $\alpha = 0.05$? Explain your answer and mention the degrees of freedom used for the hypothesis test.