CSCI 200 Jterm 2026

# Homework 1: Exploring Data

Complete the following exercises. Remember to explain your answers.

1. For this exercise, you will use the titanic dataset (*see* : *https* : *//www.kaggle.com/competitions/titanic*). You may have to sign-in using your Google account to view the data. Use the train.csv dataset on the Data tab to inform your answers below.

   (a) Give an example of a record. What does it represent?

   (b) Suppose you are using this dataset to train a model that will use a passenger class, sex, age, and fare to predict whether or not they will survive.

      i. Is this a classification task or a regression task? Explain your answer.

      ii. What are the features? Explain your answer.

      iii. What are the targets? Explain your answer.

      iv. We call a model a null model if it predicts all observations (regardless of any features selected) as the target variable. For this dataset, the null model is a model that would label all records in our data as survived. What would be the accuracy of this null model's predictions?

      v. Would you expect a model using passenger class, sex, age, and fare to predict whether or not a passenger would survive to give a better accuracy than the null model? Explain why or why not.

2. In your own (non-technical) words, explain what each of the following summary statistics measures.

   (a) mean:

   (b) median:

   (c) standard deviation:

   (d) mode:

3. You are given the following sets of five numbers, each with a mean and median of 3.

   SetA: $\{1, 2, 3, 4, 5\}$ SetB: $\{3, 3, 3, 3, 3\}$ SetC: $\{1, 1, 3, 5, 5\}$

   For each of the examples below, give a brief explanation for each of the choices made.

   (a) What is true about the sum of all numbers for each set? Which measure tells us why this is occurring?

   (b) Order the sets from smallest standard deviation to largest. Explain (non-technical, no math) why this is.

   (c) Create a set where the mean is 3, but the median is 4. Keep all numbers in the set between 0 and 5.

(d) Create a set where the mean is 3, but the median is 0. Why must this set contain numbers higher than 5?

4. For this exercise, you will use the titanic dataset to understand the relationships between different variables and draw meaningful conclusions using statistical tests. Your task is to use the given information to understand what each test is doing. You will then fill in the missing information for selecting a feature pair for each statistical test, forming hypotheses, and interpreting the results.

   (a) t-test

      i. Features used:

      ii. What the test is doing: It's checking if the average age of survivors differs significantly from the average age of non-survivors

      iii. Null Hypothesis ($H_0$):

      iv. Alternative Hypothesis ($H_1$):

      v. How to interpret this test:

      vi. What to do if test is significant: If the result is significant, you could explore further by examining the age distribution for each group (survivors vs. non-survivors). You could create visualizations to check if the distribution is skewed or if there are specific age ranges that were more likely to survive.

   (b) chi-square test

      i. Features used:

      ii. What the test is doing:

      iii. Null Hypothesis ($H_0$):

      iv. Alternative Hypothesis ($H_1$): There is a significant association between passenger class and survival status

      v. How to interpret this test:

      vi. What to do if test is significant:

   (c) correlation test

      i. Features used: age and fare

      ii. What the test is doing:

      iii. Null Hypothesis ($H_0$):

      iv. Alternative Hypothesis ($H_1$):

      v. How to interpret this test: If the p-value is small, you reject the null hypothesis and conclude that there is a significant linear relationship between age and fare. If the correlation coefficient is close to 1 or -1, the relationship is strong, while a coefficient near 0 indicates no significant relationship.

      vi. What to do if test is significant: