Fall 2023 DATA 220 Mathematical Methods for Data Analytics – Homework -1
Deadline – 11.59 PM – 9/14/2023

Problem 1 (1.5 pts): The mean length of one-year-old spotted flounder is 126 mm with standard deviation of 18 mm and the mean length of two-year-old spotted flounder is 162 mm with a standard deviation of 28 mm. The distribution of flounder lengths is approximately bell-shaped.

a. Anna caught a one-year-old flounder that was 155 mm in length. What is the z score for this weight?
b. Luis caught a two-year old flounder that was 185 mm in length. What is the z-score for this length?
c. Whose fish is longer relative to fish the same age?

Problem 2: (0.25*6 = 1.5 pts) A real estate agent wants to study the relationship between the size of a house and its selling price. The table below represents the size in square feet and the selling price in thousands of dollars for a sample of houses in a suburban Denver neighborhood. The intercept and coefficients for the regression equation is 160.19 and 0.10 respectively.

1. Find the regression equation.
2. Interpret the slope of the equation.
3. If a home has 2800 square feet, what is the predicted selling price?
4. What is the predicted selling price for a home with 3049 square feet?
5. Would you get reliable results if you predict the price of a 5000 sq. ft home?
6. What is the residual for a home with 2555 square feet?

Size (square feet)	Selling Price ($1000s
2521	400
2555	426
2735	428
2846	435
3028	469
3049	475
3198	488
2820	510
3100	530
2950	850
2250	405
3198	455
2460	270

Problem 3 (0.25*4 = 1pt): A survey was taken of pet owners in a particular area and the results are given below:

Gender of owner	Cats	Dogs	Other pets
Female	100	50	30
Male	50	50	20

a. Find the probability that randomly chosen person owns a cat.
b. Find the probability that randomly chosen person owns a dog or is female.
c. Find the probability that randomly chosen person owns other pets and is female.

d. Find the probability that randomly chosen person owns a cat or a dog.

Problem 4 (0.5*4 = 2pts): Suppose a factory that produces children's books has a 2% defective rate.

a. If two books are randomly selected from the production line, what is the probability that both are defective?
b. If four books are randomly selected from the production line, what is the probability that the first is defective and the next three are not?
c. If three books are randomly selected from the production line, what is the probability that at least one is defective?
d. If five books are randomly selected from the production line, what is the probability that at least one is defective?

Problem 5 - Data Analysis using Titanic Dataset (Coding)

The Titanic dataset is a widely used dataset in the field of data analytics and machine learning. It contains information on the passengers who were on board the Titanic ship when it sank on April 15, 1912. The dataset includes the following information for each passenger:

PassengerId - A unique identifier for each passenger

Survived - Indicates whether the passenger survived (1) or did not survive (0)

Pclass - The passenger class (1 = first class, 2 = second class, 3 = third class)

Name - The name of the passenger

Sex - The gender of the passenger (Male or Female)

Age - The age of the passenger (in years)

SibSp - The number of siblings or spouses the passenger had on board.

Parch - The number of parents or children the passenger had on board.

Ticket - The ticket number of the passenger

Fare - The fare price paid by the passenger.

Cabin - The cabin number of the passenger (if available)

Embarked - The port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

You must use the titanic data that is uploaded in the Canvas.

1. Bar charts are an effective way to compare data between categories and to identify patterns or trends in the data. They're simple and easy to understand, which makes them a great choice for displaying categorical data. Now, draw two bar charts to visualize the distribution of passengers by sex and

distribution of passengers by port of embarkation. Interpret the charts to provide insights into the data? (**2 pts**)

2. Outlier detection is a process of identifying the data points in a dataset that lie far away from the majority of the data. These data points can either be significantly higher or lower than the other observations and can impact the overall results of the analysis. On the other hand, missing value imputation is the process of filling in the missing values in a dataset. It is a common issue in real-world datasets, where some observations might not have complete information. Incomplete data can lead to biased or incorrect results, so imputing the missing values is crucial before conducting any further analysis.
    a. The "Age" variable has missing values. Based on distribution (symmetric / non-symmetric), perform an appropriate missing value imputation technique to fill in the missing values. You can choose from methods such as mean imputation and median imputation. Justify your choice of imputation technique. (**1 + 0.5 = 1.5pts**)
    b. After missing value imputation on the "Age" variable, show the outliers (if any) of the "Age" variable using the – (Consider 3 as threshold for z-score and modified z-score methods) (**3pts**) . (Must not use any direct function in python).
        i. z-score method
        ii. modified z-score method
        iii. interquartile range (IQR) method
3. Based on the IQR method – remove all the observations that contain outlier for Age variable and create a new dataset: **cleaned_data. (0.5 pts)**

Following question should be solved using the cleaned_data

4. A box plot, also known as a box-and-whisker plot, is a data visualization technique used to represent the distribution of a set of continuous or numerical data. The box plot displays the median, quartiles, and outliers of the data, allowing for a quick and easy assessment of the spread and skewness of the data. The box is drawn from the lower quartile to the upper quartile, and the median is represented as a line inside the box. Whiskers are drawn from either end of the box to the minimum and maximum data points, excluding outliers. Outliers are plotted as individual points outside the whiskers. Box plots are commonly used in exploratory data analysis and hypothesis testing.
    a. Create side by side box plots of the variable "Fare" distribution between survivors and non-survivors in the titanic dataset and compare the two box plots in terms of the five number summaries. (**1+1 =2 pt**)
5. Correlation Coefficient is a numerical measurement of the relationship between two variables. It ranges from -1 to 1, where -1 indicates a strong negative correlation, 1 indicates a strong positive correlation, and 0 indicates no correlation. The correlation coefficient can help us determine whether there is a relationship between two variables and the strength of that relationship. (must not use any direct function like corr() in python).
    a. Compute the correlation coefficient between the "Age" and "Fare" variables in the titanic dataset. **(1 pts)**
    b. Use a scatter plot to visually inspect the relationship between the two variables: Age and Fare. Interpret the strength and direction of the relationship based on the computed correlation coefficient. **(1+1=2 pts)**

6. Find mean and standard deviation of the variables: Age and Fare. Based on the shape of the distribution for both variables – (**1+1 = 2 pts**) (must not use any direct function in python).
    a. What percentage of values are between one standard deviations? Give proper explanation. (i.e., between $\mu - \sigma \; and \; \mu + \sigma$;
    $where \; \mu \; and \; \sigma \; are \; mean \; and \; standard \; deviation \; of \; the \; distribution$)

You are required to submit:

1. An MS/PDF/Scanned document:
    a. Include all the steps of your calculations.
    b. Attach screenshots of the code output.
2. Source code:
    a. Python (Jupyter Notebook)
    b. Ensure it is well-organized with comments and proper indentation.

- Failure to submit the source code will result in a deduction of 5 points.
- Format your filenames as follows: "your_last_name_HW1.pdf" for the document and "your_last_name_HW1_source_code.ipynb" for the source code.
- Before submitting the source code, please double-check that it runs without any errors.
- Must submit the files separately.
- Do not compress into a zip file.
- HW submitted more than 24 hours late will not be accepted for credit.