BIA 385 **Assignment 4 (110 Points)** **Creighton University**

1. Read the nobel dataset in Jupyter. Then answer the following questions:

    a. Select the following columns: (Year, Category, Prize, Motivation, Laureate Type, Full Name, Birth Date, Birth Country, Birth, sex, Death Date). Create a brand new dataframe using the selected columns, called "new_nobel". (5 points)

    b. Considering the "Laureate Type" column, identify how many of those who won any prize were individuals and how many were organizations. (5 points)

    c. In the 18th short video on how to add or insert a column to a dataframe, we learned how to add a column to our nobel dataset. Add a new column, call it "Age" (as we did in the lecture) such that all age values are set to a constant value. (5 points)

    d. After adding the "Age" column, we want to have "Real Age" column added to our dataframe such that it shows the actual age of each nobel winner at the time of winning the prize. To do so, follow the instruction below. (10 points)

Instruction:

you need to extract the year from the "Birth Data" column and then subtract it from the year they won the prize.

First, you need to replace "nan" values with 0. To do so, you can use replace() method on the column of the dataframe you want its "nan" values to be replaced by any other values.

To extract the year "Birth Date", you need to extract first 4 string characters from the "Birth Date" column. To do so, use str[] on the "Birth Date" column.

If you need to convert the column of type string to a numeric column, use pd.to_numeric() method on that column.

    e. Now that the "Real Age" column is added, remove/drop the "Age" column you previously created in step c of this question. (5 points)

    f. Calculate the average age of the people the in the nobel dataset. Report the age column five-number summary (mean, median, minimum, maximum, and standard deviation). (5 points)

        Hint: use describe () method

    g. (**Optional_only for practice**) perform a similar process on the nba dataset and its "Height" column. Create a new column "Height_Centimeter" where it shows the height of each player in centimeter

2. Read the nba (national basketball association) dataset in Jupyter such that a panda series is created with the "Name" column as its index and "Salary" column as its values (5 points). Call the series "nba_series" and answer the following questions using nba_series

    a. How much is the reported salary for "John Holland"? (5 points)

    b. Is the salary value of any person 5000000.0? (5 points)

    c. Write a function that returns "No Salary Reported", "Average salary", "below Average", or "Above Average" where salary is NaN, equal to average value, below the average or above the average, respectively. Then apply the function to nba_series. (10 points)

3. Use the following link to learn about the first four different ways of dealing with missing values mentioned in the link (do nothing, imputing using mean/median values, imputing using most frequent values, and imputing using K-NN). Then answer the following questions:

https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779

    a) Which approaches are better to be used to deal with missing values of a categorical column? (3)

    b) Which method is sensitive to outliers? Which method considers all column's information when dealing with the missing values of one column? (4)

    c) What are the pros and cons of imputation using most frequent values? (3)

4. Read the nba (National Basketball Association) dataset in Jupyter Notebook and answer the following questions.

    a) Take a random sample of the dataset that includes 80% of the data and answer the following questions using the sample. (2)

    b) Find out the number of missing values in each column (5 points)

    c) Drop the rows for which all columns' values are zero. (2 points)

    d) Replace the missing values of the "College" column with the most frequent college name in the dataset and replace the missing values of the "Salary" column with the average salary using imputation. (6 points)

    e) Convert the type of "Salary" and "Number" columns both to integer (2).

    f) What are the unique positions existing within the "Position" column? Is any type conversion needed for this column? If yes, convert the column type to the most appropriate type. (5)

g) How many unique teams does the “Team” column have? If the number of unique values in the “Team” column is at least 10 times less than the number of data points (rows), convert this column type to category. (3)

h) Compare the memory usage in the beginning and at the end and report how much the changes you made impacted the memory usage. (5)

i) Calculate and compare the average salary of players in three different age groups: under 25, between 25 and 35, and over 35. Print all salary values. Which age group has the highest average salary? (10)