# 3031_A4_Complete_Blank_v2

October 25, 2021

## 1 Assignment 4

Instructions:

a. You may talk to a friend, discuss the questions and potential directions for solving them. However, you need to write your own solutions and code separately, and not as a group activity.

b. Do not write your name on the assignment. (1 point)

c. Please include each question (or question number) followed by code and your answer (if applicable). Write your code in the 'Code' cells and your answer in the 'Markdown' cells of the Jupyter notebook. Ensure that the solution is written neatly enough to understand and grade. (1/2 point value of each question)

d. Export your Jupyter notebook as a PDF file. If you get an error, make sure you have downloaded the MikTex software (for windows) or MacTex (for mac). Note that after installing MikTex/MacTex, you will need to check for updates, install the updates if needed, and re-start your system. Submit the PDF file. (1 point)

**This assignment is due at 11:59pm on Wednesday, November 3rd. Good luck!** *(54 points overall – 52 points for code & answers, 2 points for anonymity and proper formatting)*

### 1.1 Part 1

#### 1.1.1 Use the dataset *movies_cleaned.csv*. We wish to find the 95% confidence interval of Profit for the movies of genre 'Action'. We will use a method known as Bootstrapping to do that.

Bootstrapping is a non-parametric method for obtaining confidence interval. The method is as follows. *(12 points overall)*

(a) Find the profit for each of the Action movies. Suppose there are N such movies. You will have a Profit column with N values.

(b) Randomly sample N values with replacement from the Profit column

(c) Find the mean of the N values obtained in (b)

(d) Repeat steps (b) and (c) 1000 times

(e) The 95% Confidence interval is the range between the 2.5% and 97.5% percentile values of the 1000 means obtained in (c)

Use the following two methods to answer this question:

**(1) Without using NumPy, compute the:**
(a) 95% Confidence interval of profit for 'Action' movies, and (b) Time taken to execute the code for obtaining the confidence interval

*(4 points for code, 2 points for answers)*

**(2) Using NumPy, and without using loops, compute the:**
(a) 95% Confidence interval of profit for 'Action' movies, and (b) Time taken to execute the code for obtaining the confidence interval

*(4 points for code, 2 points for answers)*

Note that Profit = Worldwide gross – Production budget.

### 1.2 Part 2

Download the datasets 'percent-bachelors-degrees-women-usa.csv' and 'percent-bachelors-degrees-women-usa-complete.csv' from Canvas. *(16 points overall)*

**a. Report the number of missing values for each column in 'percent-bachelors-degrees-women-usa.csv'. Which three columns have the most missing values? *(1 point for code, 1 point for answer)***

**b. For which years in the 'Year' column are the values of the 'Biology' column missing? *(1 point for code, 1 point for answer)***

**c. Make a scatter plot showing the values of the 'Biology' column on the vertical axis and the 'Year' on the horizontal axis. Make a trendline representing the points on the scatter plot *(3 points for code)*** Hint: (i) Refer to Lec3_dataViz notes for the trendline, (ii) While making the trendline, do not consider the observations with missing values in the 'Biology' column.

**d. Create a copy of the missing data, and call it "imputed_data". Using the trendline, impute the missing values for the 'Biology' column. Fill in the missing values of the Biology column with the imputed values, in the "imputed_data". *(2 points for code)*** Hint: (i) Refer to Lec3_dataViz notes for the trendline (ii) The function p(x) will provide you an estimate of the value of the 'Biology' column for Year 'x', based on the trendline.

**e. Make a scatter plot as follows:**

(i) Plot the values of the 'Biology' column on the vertical axis and 'Year' on the horizontal axis *(1 point for code)*

(ii) Color the points for the non-missing values of the 'Biology' column as 'lightgrey' *(1 point for code)*

(iii) Plot the trendline (as in part (c) above) *(1 point for code)*

(iv) Plot the imputed values of the Biology column. Color the points as 'red' *(1 point for code)*

(v) Suppose the actual values for the 'Biology' column for the years 1998-2003 are 56.35, 58.23, 59.39, 60.71, 61.89 and 62.17 respectively. Plot these values, and color the points as 'green' *(1 point for code)*

**f. Compute the RMSE (Root mean squared error) and the MAE (Mean absolute error) for the imputed values.** ***(2 points for code)***

### 1.3 Part 3

Find and load the data set "weather_stations.csv" on Canvas. The data is drawn from the city of Chicago's automated sensors at beach weather stations. Each row is a distinct record. Some of the data has been altered to reflect the fact that weather sensor data can sometimes be unpredictable or unreliable.

If a question asks you to alter the data set, please use the updated data set for the following questions. *(24 points overall)*

**1. In this data set the string "-" represents a NaN value. Print all the column names that have at least one "-" as one of their values** ***(2 pts for code)***

**2. Replace all "-" values with np.nan. Drop all rows where "Station Name" is null and report how many rows were dropped.** ***(1 pt for code, 1 pt for answer)***

**3. Wind direction shouldn't be above 360. For records with wind direction values above 360, replace the wind direction values with the median wind direction corresponding to the record's weather station. Plot a histogram of wind direction values.** ***(3 points for code incl. visualization)***

**4. Humidity values should not be negative. Calculate the mean humidity using the original data, then drop all rows with sub-zero humidity records. Recalculate and report mean humidity.** ***(2 points for code)***

**5. Clearly the mean wind speed during a given hour ("Wind Speed" column) shouldn't exceed that hour's maximum wind speed. Where this discrepancy occurs, consider the mean wind speed an outlier and cap wind speed at its hourly maximum value (i.e. if maximum wind speed is 4.0, and the corresponding mean wind speed is 4.2, cap the mean wind speed by setting it to 4.0). Report the new overall mean wind speed at each weather station.** ***(2 points for code)***

**6. Are there any outliers based on Barometric Pressure? Use the interquartile range rule to find and report how many outliers exist.** ***(2 points for code, 1 point for answer)***

**7. We'll take a more conservative approach to finding outliers in terms of air temperature. Identify records where air temperature falls more one standard deviation from its mean value. How many such outliers are there?** ***(2 points for code, 1 pt for answer)***

**8. We're interested in classifying solar radiation, given in Watts per square meter. Define any value at or below 8 as "Low Solar Radiation," any value above 8 and up to 200 as "Moderate Solar Radiation," and any value above 200 as "High Solar Radiation." Use binning to create a new column classifying each record's solar radiation. What percentage of records are in each bin?** ***(2 points for code, 1 point for answer)***

**9. Using the "Measurement Timestamp" column, create new columns based on month, day of month, year, and time.** ***(2 points for code)***

**10. Suppose we're interested in comparing beach weather during different seasons. Let's call March/April/May "Spring," June/July/August "Summer," September/October/November "Fall," and December "Winter." Create dummy variables for the four seasons.** ***(2 points for code)***