Titanic Lab

Please read all of the guidelines carefully before submitting the lab. ☺ There are ***150 points*** in total.

Due date: Wednesday, Feb. 17, 11:59 PM. Late submissions will be accepted with a penalty!

Deliverables:

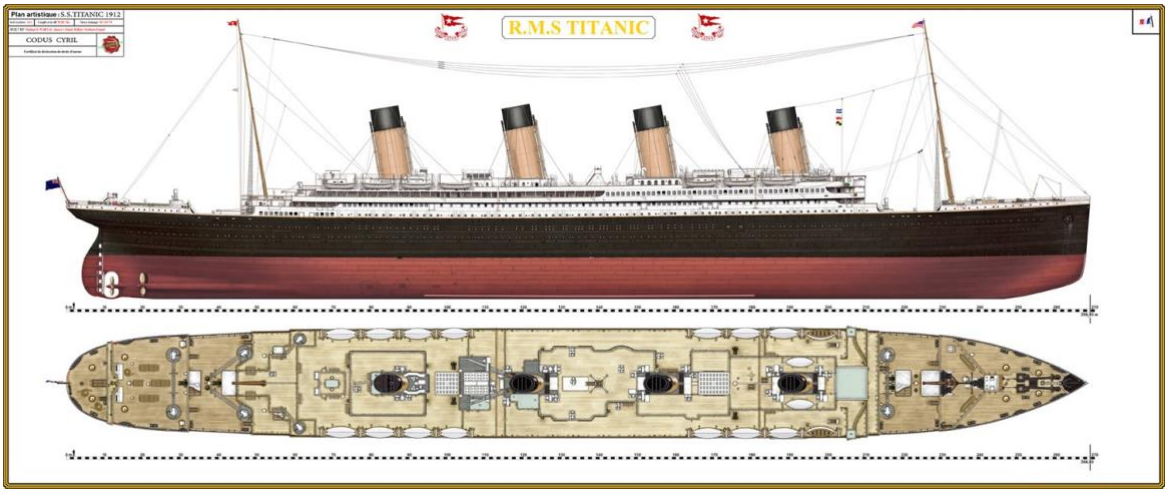
1) The code of the project in **.py** or **.ipynb** format (in several files, as explained below)
2) The lab report written with ***LaTeX*** and exported in **.pdf** format (one file)

Guidelines – Before You Start

1) You will be using the **`Python`** programming language for this lab. You need to write your codes in an empty **.py** OR **.ipynb** file. Don't use the files we used for labs to submit your homework or other assignments (those submissions will not be graded.).
2) Make sure that you provide many comments to describe your code and the variables that you created.
3) Please use the ***IEEE*** journal template on ***overleaf.com***. Here is the link: https://www.overleaf.com/latex/templates/preparation-of-papers-for-ieee-sponsored-conferences-and-symposia/zfnqfzzzxghk
   To be able to work on ***overleaf.com***, you will need to register first (you can also compile your ***LaTeX*** file locally.)
4) For some of the code, you may need to do a little bit of "Googling" or review the documentation.

What is RMS Titanic?

The RMS Titanic was a British passenger liner that sank in the North Atlantic Ocean in the early morning hours of 15 April 1912, after it collided with an iceberg during its maiden voyage from Southampton to New York City. There were an estimated 2,224 passengers and crew aboard the ship, and more than 1,500 died, making it one of the deadliest commercial peacetime maritime disasters in modern history. The RMS Titanic was the largest ship afloat at the time it entered service and was the second of three Olympic-class ocean liners operated by the White Star Line. The Titanic was built by the Harland and Wolff shipyard in Belfast. Thomas Andrews, her architect, died in the disaster.



Lab - Coding

In this assignment, you will be working with the Titanic Dataset. This dataset provides information on the passengers of the famous cruise ship Titanic. The data has been provided in the assignment folder (**titanic_data.xlsx**). Open the Excel file and take a look at it before starting. Individual features (columns) of the dataset have been described below:

PassengerId: The ID number given to each passenger by the creator of the dataset

Survived: Indicates whether the passenger survived (=1) or not survived (=0) in the Titanic disaster

Pclass: Indicates the passenger class that the passenger belongs to. There are three classes, 1 is the luxury class, 2 is the middle class, 3 is the lower class

Name: The name of the passenger

Sex: Gender of the passenger

Age: Age of the passenger

SibSp: Number of siblings/spouses of the passenger who are aboard

Parch: Number of parents/children of the passenger who are aboard

Ticket: Ticket number of the passenger

Fare: The price paid by the passenger for the ticket

Cabin: The room that the passenger is staying in (only relevant for some passenger classes)

Embarked: Port of embarkation (C = *Cherbourg*; Q = *Queenstown*; S = *Southampton*)

Before you start with the questions below, create a new and empty folder called **data_350_titanic_lab**. Call the file where you will write your code **data_350_titanic_lab.py**, as well.

Please make sure that you go over the lab Material on *Pandas* before you start with this assignment.

Part I: Setting Up Your Dataset (20 points)

Q1: Import the dataset by using the ***read_excel( )*** function of *Pandas*. Name the dataset that you imported as **titanic_df.** Add brief comments to your code and explain what your code is doing.

There are many passengers in the dataset for whom there is no age information (age feature is **NA**). Drop all of the passengers from the dataset who have no age information. How many missing values are there? What is the percentage of missing values in the whole dataset? Add brief comments to your code and explain what your code is doing (**10 points**).

Hint: A possible solution is to find the **PassengerId** for the passengers who do not have age information using a for loop and drop these rows. Or you can use a built-in function from *Pandas*.

Q2: Create a new column that is called **NotAlone**. The columns **SibSp** and **Parch** show the number of relatives passengers have on board. The value in the new column should be equal to ***0*** if the values in both **SibSp** and **Parch** are equal to zero, and ***1*** otherwise (**10 points**).

Make sure that all of your code is running!

Open an empty file in a text editor (like *Notepad++* on Windows or *Atom* on Mac). Copy and paste the code (along with your comments) into the empty file. Go to the "Save As…" option and save the code file you have created as "**data_350_titanic_lab.ipynb**" in the **data_350_titanic_lab** folder you have created at the beginning.

Part II: Mean and Median (20 points)

You will be using Python for the questions below. You can use *Pandas* and *numpy* libraries.

Q3: There are three different passenger classes in the dataset. Using *Pandas*, split the dataset into three parts and store these datasets in memory separately:

titanic_df_passenger_class_1 : Dataset that only includes information on passengers from class 1

titanic_df_passenger_class_2 : Dataset that only includes information on passengers from class 2

titanic_df_passenger_class_3 : Dataset that only includes information on passengers from class 3

Indicate how many observations there are in each dataset. Which is the biggest passenger class? Provide your answers in the form of comments.

Using *Pandas* or *numpy*, calculate the **mean** and **median** values for *Age* and *Fare* columns. You will be calculating **12 values** in total (**mean** and **median** value from each of the datasets that you created above). You can provide your answers in the form of comments.

Indicate which passenger class has the highest **mean** and which passenger class has the lowest **mean**. Also indicate which passenger class has the highest **median** and which passenger class has the lowest **median** for *Age* and *Fare* columns.

Add brief comments to your code and explain what your code is doing (**10 points**).

Q4: Take the dataset from **Q2**. Create two different datasets using the *Survived* feature:

titanic_df_survived : Dataset that includes the passengers who survived (*Survived* = 1)

titanic_df_not_survived: Dataset that includes the passengers who didn't survive (*Survived* = 0)

Indicate how many observations there are in each dataset. How many people survived? How many people did not survive?

Add brief comments to your code and explain what your code is doing.

Using the datasets you created in **Q3** and the methods you used in previous questions, calculate how many people survived and how many people died in each passenger class. **You will be reporting six values in total**. You can provide the answers in the form of comments. Add brief comments to your code and explain what your code is doing (**10 points**).

Make sure that all of your code is running!

Save the code in the "**data_350_titanic_lab.ipynb**" file in the **data_350_titanic_lab** folder you have created at the beginning.

Part III: Bivariate Statistics and Variance (30 points)

Q5: Write (hand-code) a function for variance, standard deviation, and correlation. **Do the calculations in the questions below by using the functions that you coded** (**10 points**).

Q6: Take the dataset from **Q2**. Using the hand-coded version of correlation that you wrote in **Q8**, please provide a correlation matrix that shows the correlation between all numerical values possible (***Survived***, ***Age***, ***SibSp***, ***Parch***, ***Fare***, ***NotAlone***). What are your observations? What are some of the closely correlated values? (**10 points**).

For reference:

> 0.8:	*strong* correlation
0.6 – 0.8:	*medium* correlation
< 0.6:	*weak* correlation

Add brief comments to your code and explain what your code is doing.

Q7: Take the dataset from **Q2**. Calculate the standard deviation for *Age* and the standard deviation for *Fare*. Report these two values in the form of comments.

Again, take the dataset from **Q2**. Calculate the interquartile range for the *Age* and *Fare* columns. What can you say about the results? What do you think the shape of your data looks like? Is the majority of observations below the mean or above the mean (**10 points**)?

Add brief comments to your code and explain what your code is doing.

Make sure that all of your code is running!

Save the code in the "**data_350_titanic_lab.py**" file in the **data_350_titanic_lab** folder you have created at the beginning.

Part IV: Conditional Probabilities (20 points)

Q8: Write a program that finds the answers to the following conditional probability questions:

i. Calculate the conditional probability that a person *survives* given their *sex* and *passenger class*:

P(Survived = true | Gender = female, Class = 1)
P(Survived = true | Gender = female, Class = 2)
P(Survived = true | Gender = female, Class = 3)
P(Survived = true | Gender = male, Class = 1)
P(Survived = true | Gender = male, Class = 2)
P(Survived = true | Gender = male, Class = 3)

What are your observations? Who survived the most? Who survived the least?

ii. What is the probability that a child who is in third class is 10 years old or younger survives?

iii. How much did people pay to be on the ship? Calculate the expectation of fare (X) conditioned on passenger class.

E[X | Class= 1]
E[X | Class= 2]
E[X | Class= 3]

Report the results and write your observations (**10 points**).

Q9: Another way to calculate the rates of survival in different groups of passengers is by looking at the titles (Mr., Miss, Mrs., Capt., Sir., Dr., Jonkheer etc.). For this question please do the following:

Hint: You may need to use some sort of loop mechanism and the .**split()** function used for strings.

i. Find all titles of social status/nobility and provide a list of them.
ii. Calculate the survival rate for each category.
iii. Report the results and write your observations (**10 points**).

Make sure that all of your code is running!

Repeat the steps above to save your code. Go to the "Save As…" option and save the code file you have created as "**data_350_titanic_lab.ipynb**" in the **data_350_titanic_lab** folder you have created at the beginning.

Part V: Visualization (20 points)

Q10: For this part of the analysis, you can use any visualization tool within the Python realm. Some useful packages are *matplotlib* and *seaborn*. You can find some helpful documentation here:

https://matplotlib.org/
https://seaborn.pydata.org/

i. Create a bar graph OR pie chart showing the number of people who survived and who lost their lives (**5 points**).
ii. Create a heatmap that shows the correlations in color between the numerical variables you have in **Q8** (**5 points**).
iii. According to what you have found in **Q13**, choose the top 5 titles who survived the most and compare their survival rates using a bar graph (**5 points**).
iv. Draw a line chart that looks at the relationship between ticket fare and survival rate (**5 points**).

Part VI: Creating the lab report (40 points)

Q11: Write a report (2 pages) that includes all of your findings and the visuals that you created. The report that you will write should use the *IEEE format* and include the following sections:

Abstract: A short summary of your report

Introduction: A summary of what you expected and did, and two-three of your most significant findings

Data: A description of your dataset (number of observations, what kind of information is there etc.)

Results: A summary of the analyses you completed for the lab + a discussion on ***causal inference*** that talks about these subjects:

- What are some of the potential causes that helped passengers survive the accident?
- Do you think "women and children first" was applied in this disaster?
- Do you think the correlation analysis that you did to uncover some potential relationships is statistically satisfactory? Do you see any potential room for improvement?
- Do you see any typical causal inference problems (***selection bias***, ***simultaneity***, ***omitted variable bias***) in this analysis?

Final step:

Compress your submission folder **"data_350_titanic_lab"** and send the compressed folder through *Notebowl*.