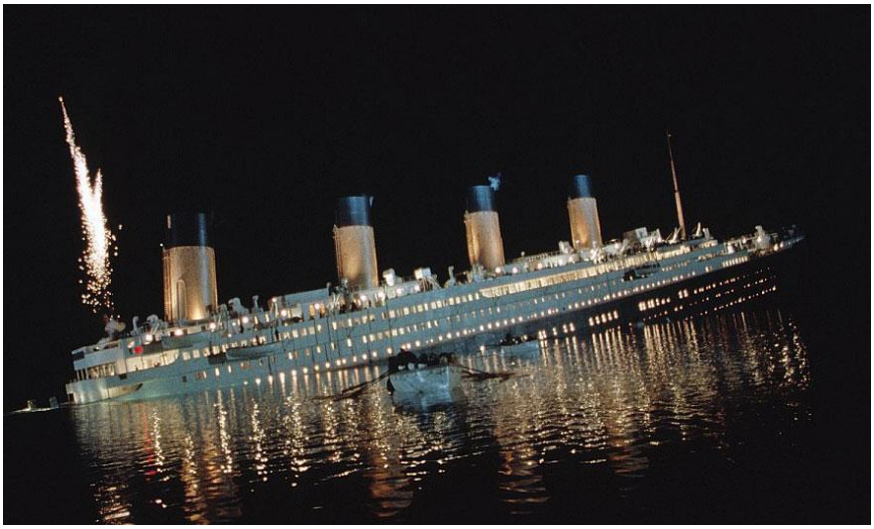
Cairo University
Faculty of Engineering
Computer Engineering Department

CMP4011
Spring 2023

Big Data and Cloud Computing

Lab 1 – RStudio and Introduction to R

* The dataset in this assignment is adopted from the famous Kaggle competition *Titanic: Machine Learning from Disaster*.



The sinking of the RMS Titanic is one of the most famous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.

One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others.

In this lab, we will analyze this disaster, get more insights into it and come up with conclusions about which sorts of people were more likely to survive than others.

Dataset:

Titanic dataset is the most famous dataset for beginners in Data Science. Let's have a view on the data dictionary.

- ***Tip 1: It's very important to spend some time understanding data. Take a closer look at the data available, assess and explore it using tables and graphics.***

Variable	Definition
survival	Survival (0= No, 1=Yes).
pClass	Passenger Class (1=1st, 2=2nd, 3=3rd).
name	Name.
gender	Gender.
age	Age in years.
sibsp	Number of siblings/spouses aboard the Titanic.
parch	Number of parents/children aboard the Titanic.
ticket	Ticket Number.
fare	Passenger Fare.
cabin	Cabin.
embarked	Port of embarkation (C= Cherbourg, Q=Queenstown, S= Southampton).

Variable notes:

pclass: A proxy for socio-economic status (SES)
1st = Upper, 2nd = Middle, 3rd = Lower

age: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5

sibsp: The dataset defines family relations in the following way:
Sibling = brother, sister, stepbrother, stepsister
Spouse = husband, wife (mistresses and fiancés were ignored)

parch: The dataset defines family relations in the following way:
Parent = mother, father
Child = daughter, son, stepdaughter, stepson
*some children travelled only with a nanny, therefore parch=0 for them.

Requirements:

- ***Tip 2: Before asking others for help, it's generally a good idea for you to try to help yourself. R includes extensive facilities for accessing documentation and searching for help.***

1.	First of all, start by cleaning the workspace and setting the working directory.	
2.	Import the dataset **titanic.csv** into a data frame.	
3.	It's time to explore the dataset as a whole.	
	a.	Show the dimensions of the data frame. Hint: Use dim()
	b.	Show the structure of the data frame. Hint: Use str()
	c.	Get more insight into data by exploring the **first** and the **last TEN** rows in the dataset.
	d.	Show summary of all variables in the data frame.

4.	Let's explore some variables in the dataset.	
	a.	Show a summary for the variable *age* only.
	b.	What are the first and third quartile values for this variable? What do these values mean?
	c.	Are there any missing values in the variable *age?* (i.e. written as <NA>) 📌 ***Hint: Read the documentation for is.na() and anyNA() to find out how to know if a certain variable has missing values. What are the differences between them? Which one is better to use in this case?***
	d.	What is the type of the variable *embarked?* Show the levels of this variable. Is that what you were expecting?
	e.	Can you conclude what's needed at this step in the data analysis cycle?
5.	As you probably might have answered in (4.e), preprocessing is needed. Data preprocessing is a very important step in any data analytics project.	
	a.	Remove the rows containing <NA> in the age variable from the data frame.
	b.	Remove the rows containing any unexpected value in the embarked variable from the dataset.
	c.	Now, check that no NA values exist in the age variable. Also, factor the embarked variable and display its levels. Is that what you are expecting?
	d.	Some variables are not very interesting and provide no real indicative value. Remove columns *Cabin* and *Ticket* from the dataset.
6.	An important step also in any data analysis project is **statistical description** and **visualization**. We will now visualize some variables, and try to get insights out of them. In this part, you will practice visualization and slicing/indexing in data frames.	
	a.	Show the number of **males** and **females** aboard the Titanic.
	b.	Plot a pie chart showing the number of males and females aboard the Titanic. *(Hint: use **pie()** function)*.
	c.	Indicate males with a blue color and females with a red color in the above plot. *(Hint: There is a color parameter in any plot function).*
	c.	Show the number of people who **survived** and **didn't survive** from each gender.
	d.	Plot a pie chart showing the number of males and females who survived **only**.
	e.	What do you conclude from that?
	f.	Show the relationship between social class and survival i.e. show how many people survived and how many people didn't survive from each class.
	g.	Plot this relationship as a **stacked bar plot**. (Hint: use barplot() function)
	h.	Indicate survived passengers with a blue color and un-survived passengers with a red color in the above plot.
	i.	What do you conclude from that?
	j.	Plot **a box and whiskers** plot for the variable age (Hint: use boxplot() function) Read about the **box and whiskers** plot and understand it properly.
	k.	What does this plot mean?
	l.	Plot a density distribution for the variable age.
7.	Remove all columns but passenger name and whether they survived or not. Export the new dataset to a file named **"titanic_preprocessed.csv"**	