The sinking of the Titanic. Statistics of the disaster

Introduction to the group project

[i]The assignment

For this project, you will form teams of 3 students belonging to the same class[1]. With your team, you will complete the project presented here.

The assignment subject and the purpose of the project.

The event

The sinking of the ocean liner *Titanic*, during its maiden voyage, at the beginning of the 20th century, is still a vivid event, in the collective memory of a large part of the western population. A large proportion of the Titanic passengers and crew lost their lives as the consequence the sinking.

Publications

Many articles and books, some of them fallacious, some trustworthy, have been devoted to the circumstances of, and the lessons to be drawn from the sinking.
In 1997, a film named *Titanic*, by James Cameron[2], with actors Leonardo DiCaprio and Kate Winslet playing the main characters, has been released. The film was a worldwide success (worldwide gross above $1 billon, 20 758 841 spectators in France at the first release). Although the focus of the film is not the catastrophe itself but rather a (fictitious, and very unlikely) love story between the two main characters, the general opinion at the time of the release was that the framework of the story (the circumstances of the sinking) was well researched, thanks to the presence of historians in the film team.

The data

The web site *Kaggle*, (https://www.kaggle.com) proposes many prediction challenges, including a beginner's challenge that focuses on the prediction of survival among passengers during the sinking of the *Titanic* (https://www.kaggle.com/c/titanic).
For the present project, we will use the dataset published by Kaggle but not with the same intent. If you are interested in *Machine Learning*, you are encouraged to join Kaggle and to try your luck predicting survival, although that would require you to learn a lot about prediction methodology and techniques.
The dataset and its codebook will be posted in this repo.

[1] If you wish to work with a team of less than 3 persons, you are free to do so, although a group size of three is recommended. If you think that the circumstances dictate a larger team, first ask for the approval of the professor in charge of your TD group. He will make the final decision.
[2] https://en.wikipedia.org/wiki/Titanic_(1997_film)

Purpose of the project

The purpose of the project is to give you the opportunity to perform *data preprocessing* and *exploratory data analysis* on a real dataset.
You will be required to go through these steps and produce a preliminary report and, most importantly, a final report presenting your findings.

The project

Objective

During this project, you will analyze the dataset proposed by Kaggle and made accessible on the Ipagora site of the course, in the case study folder. You will be guided during the first part of the analysis.
You are required to analyze the dataset and to form hypotheses about how to understand and explain the sailent features of the data: most importantly, to describe and explain the survival of the passengers.
You will summarize your findings in two successive written reports, to be handed in to the professor in charge of your TD group.

Requirements and Deadlines

Each report will be typed, and printed on paper. Each report will list, on the front page, the TD group and the names of the team members. The form of these reports is described later in the document.

The deadlines are:

- **Preliminary report:** You will hand in your preliminary report during the class of April 25th
- **Final report:** You will hand in your final report to the procter, during your final exam.

Part 1 - Preliminary Information gathering

Find and read as much information you can find about the Titanic and its maiden voyage.

Questions to consider:

Information sources

- Can you identify some reliable information sources? Are these sources in agreement with one another?

The ship

- What ship was the *Titanic*? Who was the owner of the ship? What was its intended function? (cruise ship? Passenger transport ship? Cargo ship?...)
- Who were its crew? Its passengers?
- How was the ship organized? What is the list of its decks? What was the use of each deck?
- Much has been said about the lifeboats. How many lifeboats did the *Titanic* carry? How many individuals could fit in them? Did the ship have enough lifeboats, according to the regulations of the time?

- What other properties did the Titanic supposedly possess, which would justify the small number of lifeboats?

The maiden voyage

- When did the voyage begin? When did the accident occur?
- What was the route followed by the ship? What ports did it stop at and in what order? Which of these ports had its name changed since that time? Why?
- How many people were on board? How many passengers? How many crew members?
- What were the circumstances of the accident?

The accident

- What happened? How was the evacuation organized? How were the survivors rescued?
- *"Women and children first".* What does it mean? How is it related to the sinking of the Titanic?

The Investigations

- What investigations have been made after the accident? What were their conclusions?

Part 2 - Discovering, and preprocessing the data- Getting to know the passengers.

Preparation

Read the document named "Document complémentaire – Analyse exploratoire des données", on the Ipagora site for the course, week 4.
Read the document about using XLStatistics, in the folder of the case study. The use of Xlstatistics for this analysis is recommended. Install and try XLStatistics. You may use the data of the TD sessions for that.

First contact with the dataset

Take a good look at the codebook (also called data dictionary). Read the data into your analysis software (Excel). Does the data appear to match the codebook? Do you understand everything? Check.

Questions to consider:

- How many variables are there? What are the types of these variables? What is the meaning of each variable? What are the units used?
- What is an individual here? How many individuals do we have? Does the dataset describe a sample or the whole population? If it is a sample, how was the sample selected?
- Are there missing values? Many? What variables are most affected?
- Given the mission, are there facts that would be interesting, but are not recorded in the dataset?

Data: Exploration and preprocessing

The Passenger's demographic variables.

1. Draw a graphical representation for each of the variables ***Sex***, ***Age***, ***SibSp***, ***Parch***. Comment.

The age

2. The age might be an extremely important element, in order to understand the survival rates of the passengers (remember *"Women and children first"?)* Create a new qualitative variable ***AgeStatus***, with three levels:
   - Child (12 years of age or less)
   - Teenager (from 13 to less than 20 years of age)
   - Adult (20 years of age and more)
3. Plot the new variable. Comment. Are there many missing values?
4. Let us try to discover some of the missing ***AgeStatus*** values. Have a good look at the variable ***Name***. Do you see the title of the passenger? (Mister, Miss, Mrs, etc…). Create a new variable ***Title***, which extracts the title from the ***Name*** field (you might use the text functions of Excel - CHERCHE(), STXT() and the like…).
   Then make a contingency table with the variables ***Title*****,** and ***AgeStatus*** (You can use XlStatistics for that, or a pivot table, i.e. Excel's "Tableau croisé dynamique"). What do you think of the age status of the persons called "Master"? Can you guess their age status if it is missing? Can you guess other age classes using the title information? Complete the ***AgeStatus*** information as much as you can.

The family size

5. Consider the variables ***SibSp***, ***Parch*** . Do you understand them? Let us combine the information they offer, by creating a variable named ***Fsize***, as the sum ***SibSp***+***Parch***.
6. We might have another information related to the family size. Create a new variable ***Numtick***, which is the number of passengers of the sample who share the same ticket number (You may use the Excel function NB.SI() here). There might be a good chance that in most cases, the people who share the same ticket number belong more or less to the same family (siblings, spouse, parents, cousins, in-laws… etc). Try to verify the likeliness of this hypothesis:
   a. Compute the difference ***Fdiff*** = ***Numtick*** - ***Fsize*** .
   b. Compute (with XlStatistics, or with a pivot table) the relative frequency distribution of the new variable ***Fdiff*** . What is the mode of ***Fdiff***? Does this result support our hypothesis? Can you think of a possible explanation for the other values?
   c. Modify your table to treat separately the passengers of each class (variable ***Pclass***). What does the table reveal?
   d. Confirmation: Get the list of the first class passengers on Wikipedia (https://en.wikipedia.org/wiki/Passengers_of_the_RMS_Titanic). Do you understand the lines of the table that begin with the word "and"? Do you find the same kind of lines in the list for other passenger's classes?
   e. What is your conclusion regarding the family size?
7. Using these results, create a new variable ***FamilySize*** , which values will be your estimations of the size of the family group (onboard *Titanic*) for each passenger of our dataset.

The Passenger's Class

8. Make a frequency table and a graph for the variable ***Pclass***. Comment.

Passenger's class, sex, age and family size

9. Is there a relationship between the ***FamilySize*** variable you have just created and the passenger's class ***Pclass***?
10. Is there a relationship between the ***Sex*** variable and the passenger's class ***Pclass***?
11. Is there a relationship between the ***AgeStatus*** variable and the passenger's class ***Pclass***? Explain.

Variables relating the passengers to the ship and the voyage

Embarkation port

12. The variable ***Embarked*** lists the port of embarkation of the passengers. Present a frequency table for this variable. Are there missing values?

13. Is there a relationship between the port of embarkation and the class? Between the port of embarkation and the sex of the passengers? Between the port of embarkation and **both** the class **and** the sex of the passengers?

Situation in the ship

14. Investigate the ***Cabin*** variable. Are there missing values? Is there a relationship between the presence of a cabin number and the class? What is the meaning of the first letter of the cabin number? (See https://www.encyclopedia-titanica.org/cabins.html)
15. Create a variable ***Deck***, the value of which being the first letter of the cabin number.
16. Make a joint frequency table (or graph) of the variables ***Deck*** and ***Pclass***. Comment.
17. Is there a relationship between ***Deck*** and ***Embarked***?

Fare

18. Consider the variable ***Fare***. Do you understand it? Make a boxplot for this variable.
19. Considering only the passengers travelling alone, make a table (or graph) of the mean fare, for each class. Compare with the fare listed for passengers who travel with a larger family group.
20. Temporarily arrange the data by ***Numtick*** (largest to smallest value), and ***ticket number***. Look at the values of ***Fare***, for the passengers who share the same ticket number. What do you see? How could you explain this fact? Make a scatterplot of the fare versus the number of persons concerned by the ticket price. What do you think?
21. Make a new variable ***FPP***, equal to the **Fare per person** paid for the trip. You may try different versions of ***FPP***, depending on how you compute the number of persons concerned by the ticket. For each version, make a multiple boxplot showing FPP for each Passenger's class. (Use XLStatistic's 1num1cat) What calculation of FPP do you prefer? Comment on the Fare per person, for each Class.
22. Does the fare ***FPP*** appear to be dependent on the port of embarkation?

Conclusion of the first part of your exploration.

You can now write your preliminary report: It will be very short (1-2 pages), and consists in two parts:

1. *Introduction*
   Keep it concise. (5 to 20 lines of text)

2. *The Passengers*
   You will present a **synthetic** account of everything the dataset taught you about the passengers. Be complete, but only tell the essential (1-3 pages). Include graphics if they are useful.

Part 3 - Passenger's survival

You are now ready to analyze the passenger's survival. What are the variables that best explain the survival of the passengers? Has the *"Women and children first"* policy been effective?
Try this analysis on your own, select the results you find most important, and draw your conclusions.

Part 4 - Final report

You can now write your full report. It will typically have the following structure:

1. Executive Summary

Here, you present the gist of you work for the reader in a hurry: From the problem statement to your conclusions, through the methods you used. Keep it **very** short. (Rem: of course, it comes first in the report, but it is the last thing you write)

2. Body of the report

Problem statement

What is it all about? What is the problem? What is your mission definition? (Rem: this is also best rewritten when everything else is completed)

The data

Briefly present your data. Population or sample? Origin? Structure?

The analysis

Your most significant and interesting findings.
In this case, interesting means:

- The description of the passengers (i.e. what you included in your preliminary report)
- The description of the passenger's survival, and how it depends on some selected variables.

The conclusion(s)

As the name implies: what lessons should be learned from the analysis?

3. Appendix

You put here:

- An account of the preprocessing of the data:
    - missing values treatment
    - data transformations
    - New variables created
    - Other decisions you made,
    - Etc.
- The technical details of your analysis which are important, but not worth presenting in the main body of the report.
- Every analysis you tried (even if it was not interesting, after all).

Of course, even if it is the lesser-read part of your report, the structure of the appendix should make access to the material as easy as possible, should the need arise.
