A close up of a sign

Description automatically generated

**Lebanese American University**

**School of Business**

BDA211 Final Project

Titanic Tragedy: In Terms of Data Analysis

Submitted by:

Yousef Atrach (leader)

*Under the guidance of*

*Dr. Manal Younes*

Table of Contents

[Data Analysis 3](#_Toc119250875)

[Introduction 3](#_Toc119250876)

[Purpose 4](#_Toc119250877)

[Project Topic and Research Questions 4](#_Toc119250878)

[Gender Vs Ticket Class Choice 5](#_Toc119250879)

[Ticket Class Vs. Survival Rate 6](#_Toc119250880)

[Age Vs Survival Rate 7](#_Toc119250881)

[Passenger Fare and Ticket Class 8](#_Toc119250882)

[Conclusion 9](#_Toc119250883)

# Data Analysis

## Introduction

Data is the base factor and the major contributor in making the decisions that involve major changes in a company where it either make or break. Due to the increase demand on technology and introducing new terms like big data, the thing that required a new technology that allows the normal user derive insights from the available data. Business data analytics along with its outstanding impacts have changed the world. Thanks to data mining, companies became able to detect certain patterns and derive numerous insights coming from huge data. There are five data mining techniques: classification, regression, clustering, anomaly detection and association learning. Classification is used to classify data into categories. Regression is used to predict numbers and values based on a unique equation. Clustering is used to group variables based on common characteristics. Anomaly detection is used to identify suspicious activities to fight fraud and detect abnormalities. Furthermore, business data analytics introduced the concept of data aggregation. It is a process of gathering data and presenting it in a summarized format. Finally, data analytics is also considered as a powerful tool because it introduced companies to the world of data types: descriptive, diagnostic, predictive and prescriptive analytics. Descriptive analytics is used on past as well as present data to show them in the form of simple visualized reports. Diagnostic analytics is concerned with studying problems and reveal the reasons of their occurrence. Predictive analytics is used for forecasting purposes to predict what could happen in the future. Prescriptive analytics is used to prescribe solutions to the companies’ problems. It is worth mentioning that descriptive analytics is the base of all the other types of data analytics. Our project will be a significant example of how data analytics can be used to make the world a better place. We will be using data mining techniques as well as aggregate methods to conduct our analysis. We further will be using data types to present our data in a neat simple way that is easy for reviewers to perceive. Finally, we will diagnose the issue related to the death incidents and prescribe solutions.

In our study, we have chosen the titanic data set. It is composed of 12 columns as well as 892 rows of data. The survival variable. The Passenger ID. It is the unique ID that each passenger must have. The pclass. It shows the ticket class chosen by each passenger. The sex variable that defines the gender of each passenger. The age of each passenger. The number of siblings and spouses each passenger has abroad on the Titanic. The number of parents and children each passenger has abroad on the Titanic. The ticket numbers. The passenger fares. The number of Cabins in the ship. The port of embarkation.

Using this wide set of data, we believe that we can study the effect of each variable on the survival rates of the passengers. We are looking to identify the variables that could cause the survival rate of passengers to increase and those variables that might have contributed to the death of the passengers. We will further study the correlation of variables with each other’s to validate our research questions. We will be using plenty of data analytics tools such as: Tableau, Excel, and SPSS Statistics. The tableau will be used to include better visualization of the relations that we have among variables. Moreover, Excel and SPSS will be used to validate whether the variables are correlated or not. For that, we will be using the correlation and Chi-square techniques.

## Purpose

The main purpose of our project is to introduce our research questions and come up with significant findings. Our main research question is concerned specifically with the relation between the demographic factors and the survival rates of the passengers. To achieve this goal, a set of sub-questions was introduced to enforce the main topic, and thus, derive useful insights. We also intend to visualize our findings to fuel the understanding of the present data.

## Project Topic and Research Questions

To analyze the Titanic tragedy in the light of the data set we have, a research question was raised: *How does the demographic factors affect the survival rates of passengers?* Several tools and sub-questions were also raised to favor the suggested topic. Mainly, the main interest is to study the relation between separate demographic factors and link them to the survival rate. To further explain that, theoretically, it is believed that the gender, age, ticket class, fare, availability of siblings/ relatives on board are factors that can be linked together, and a conclusion can be drawn out on whether they affect the survival rate or not. Both the numerical and the visual results will help proving or disproving our theory.

Below presented the data analysis of the sub-questions in both numerical and visual methods. Excel and SPSS were used to find the descriptive statistics the data set, and Tableau was used to visualize them.

### Gender Vs Ticket Class Choice

Table

Description automatically generated

The Chi-Square table shows a significance which is less than 0.01. This implies that there is a strong relationship between gender and the ticket class choice.

Graphical user interface, application

Description automatically generatedThis Chart represents the number of male and female passengers in each ticket class category. It is shown that the number of females is significantly less than males in the third class. Both genders are almost equally presented in the other two classes.

### Ticket Class Vs. Survival Rate

Table

Description automatically generated

The Chi-Square table reveals a significance which is less than 0.01. This shows that there is a strong relationship between the ticket class and the survival rate. To elaborate, the higher the ticket class is, the higher the survival rate of passengers was recorded. This was shown in the graph.

Graphical user interface, application

Description automatically generatedWe applied a filter to the previous chart to show only the survivors

### Age Vs Survival Rate

Table

Description automatically generated

The results show a significance level of 90% which is slightly acceptable but can’t be taken into consideration.

A screenshot of a computer

Description automatically generated

The chart shows that both the lines of the survivors and the dead are similar in shape. This signifies that the age and the survival rate are not correlated.

### Passenger Fare and Ticket Class

Table

Description automatically generated

The Chi-Square table reveals a significance which is less than 0.01. This shows that there is a strong relationship between the ticket class and the passenger fare. To illustrate, the higher the passenger fare is, the higher the chosen ticket class was.

A screenshot of a computer

Description automatically generated

It is clear that the higher the class, the higher the fare in the chart above.

## Conclusion

In conclusion, we believe that our theory was proved. Out of all the examined factors, we found that there is a correlation between the survival rate and the other demographic factors except in one example which is the relation of age and survival rate. Furthermore, our results are presented in a well-organized Dashboard that will be further discussed in the presentation. Below is an example of the representation and connecting the dataset and the research question, all in one screen:

**Present the results of the dataset of the passengers that survived and were embarked from Cherbourg Port:**

A screenshot of a computer

Description automatically generated