



STATISTICS FOR DATA SCIENCE

TITANIC SURVIVAL ANALYSIS





ABSTRACT:

Our dataset, Titanic Survival Analysis describes the survival status of individual passengers. The dataset has 890 rows and 12 columns.

The variables on our extracted dataset are pclass, survived, name, age, sex, ticket, embarked, etc.

Visualization is done for survival based on gender and embarkment and also about total and survived population.

The goal of our team was to find out if the class of the people affected their survival rate by performing hypothesis testing.

INTRODUCTION:

The sinking of the Titanic is one of the most infamous shipwrecks in history.

On April 15, 1912, during her maiden voyage, the widely considered “unsinkable” RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren’t enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew.

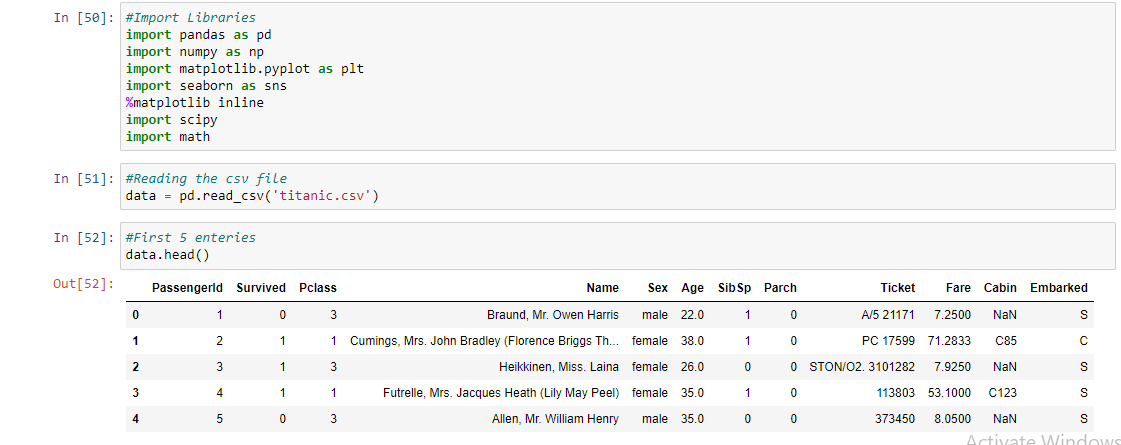
While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others. In order to determine this, we chose this dataset.

DATASET:

This ‘TITANIC SURVIVAL ANALYSIS’ dataset is taken from Kaggle.

It is a popular machine learning dataset.

INITIAL DATASET:



ATTRIBUTES OF THE DATASET:

• **Survival**: Indicates whether the person survived or not. The data in this column is of Boolean type. 1 indicates that the person survived and 0 for vice versa

• **Pclass**: corresponds to the ticket class of the passenger who boarded the titanic. The three classes that exist are 1st, 2nd and 3rd

• **Sex**: corresponds to the gender of the passenger. Stores string type values as ‘female’ or ‘male’

• **Age**: Integer type data storing the age of the passenger

• **Sibsp**: Indicates the number of siblings/spouses of the passenger aboard on the ship

• **Parch**: Indicates the number of parents/children of the passenger aboard on the ship

• **Ticket**: Denotes the ticket number of the passenger aboard

• **Fare**: Indicates the ticket fare of the passenger

• **Cabin**: Denotes the cabin number of the passenger aboard

• **Embarked**: Port from which the passenger has embarked. The three ports are C for Cherbourg, Q for Queenstown, S for Southampton.

WITHOUT DATA CLEANING:



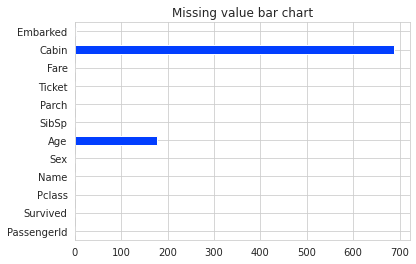
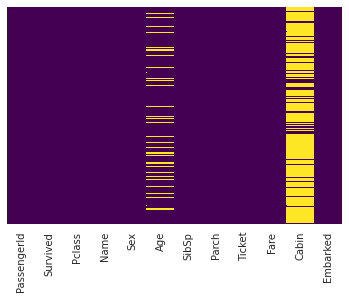
PREPROCESSING OR DATA CLEANING:

First, we find the sum of null values in each attribute of the dataset and plot a missing value Bar Chart. We also plot a missing value Heat map.

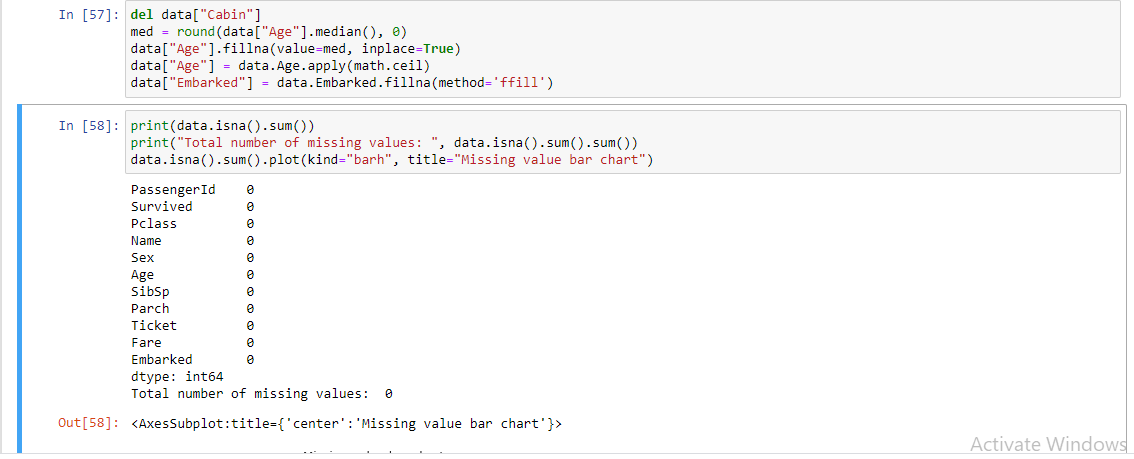
Since the Cabin attribute has the highest percentage of null values (more than 75% null values), we delete that entire attribute. We replace the null values in age attribute with the median age value.

Then we fill the embarked attribute with a method called forward fill which replaces the null value with the previous non-null value.

MISSING VALUES DEMONSTRATED IN HEAT MAP AND BAR CHART



AFTER DATA CLEANING:

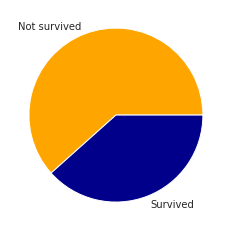


GRAPH VISUALIZATION:

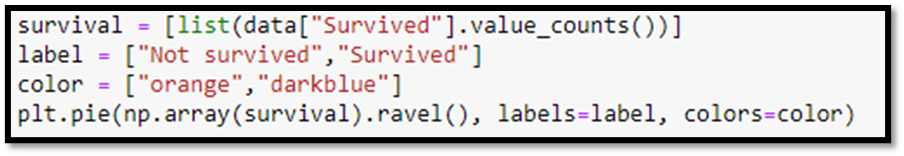
The process of graphical representation of information and data is known as visualization of data. Data visualizations make big and small data easier for the human brain to understand, and visualization also makes it easier to detect patterns, trends, and outliers in groups of data.

PIE CHART DEPICTING SURVIVAL RATE:

Here is a pie chart depicting the percentage of survival of the passengers aboard on the titanic. The pie function from the matplot library is used to plot this. Thus, we can conclude that a higher population of people did not survive the shipwreck

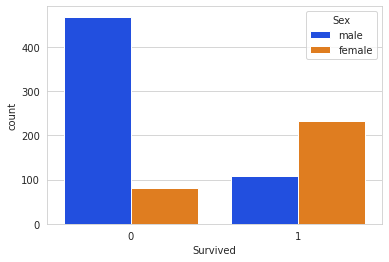


Code for the same:



COUNT PLOT FOR GENDER

A comparison between the male survival rate and that of the female is depicted by the count plot. The count plot function belongs to the seaborn library. From the count plot it is inferred that male mortality rate was higher.

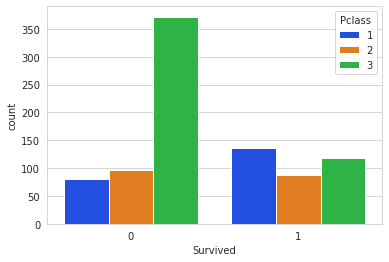


Code for the same:



COUNT PLOT FOR TICKET CLASS:

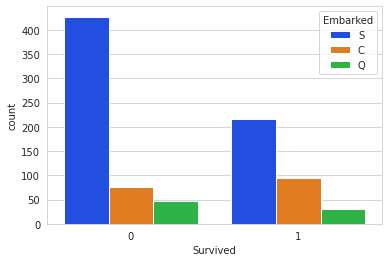
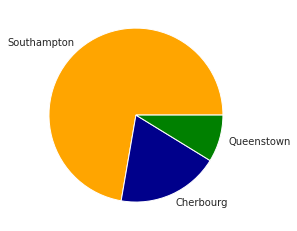
A similar plot comparing the relationship between survival rate and ticket class of the passenger is drawn. The plot shows that amongst all the three classes of passengers, the class 3 passengers had more victims.



Code for the same:



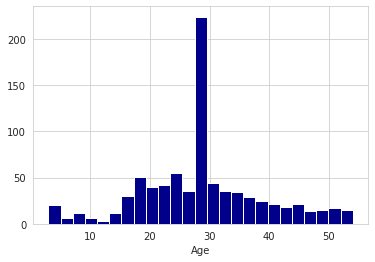
Analysis of passenger embarkment with pie chart and count plot



The pie chart depicts the number of people who embarked from a particular port. We can conclude that the highest number of passengers embarked from the Southampton port.

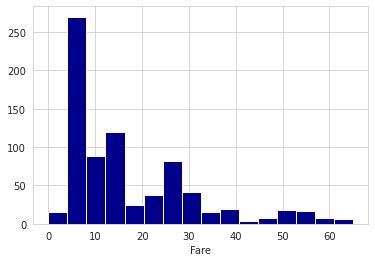
The count plot depicts the survival rate of the passengers depending upon the port of embarkment. We notice that Southampton port has both higher survival and death rate owing to its highest passenger embarkment

HISTOGRAM FOR AGE OF THE PASSENGERS:



A histogram depicting the age of the passengers is plotted. We observe that the ship had passengers majorly of ages around 30

HISTOGRAM FOR THE TICKET FARE:



A histogram depicting the fares paid by the passengers is plotted. We observe that majority of the passengers paid a fare of about 10 dollars

REMOVAL OF OUTLIERS:

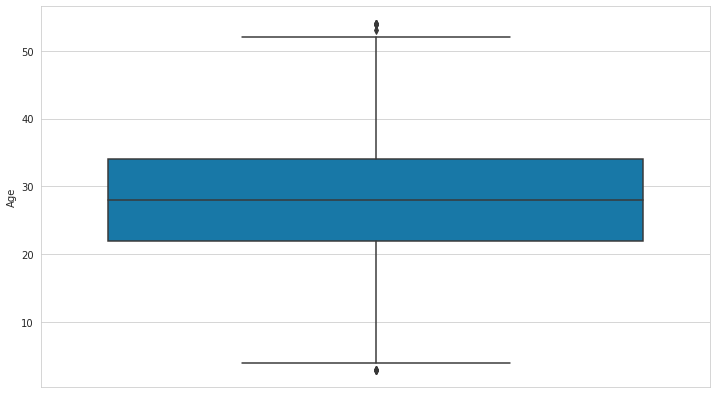
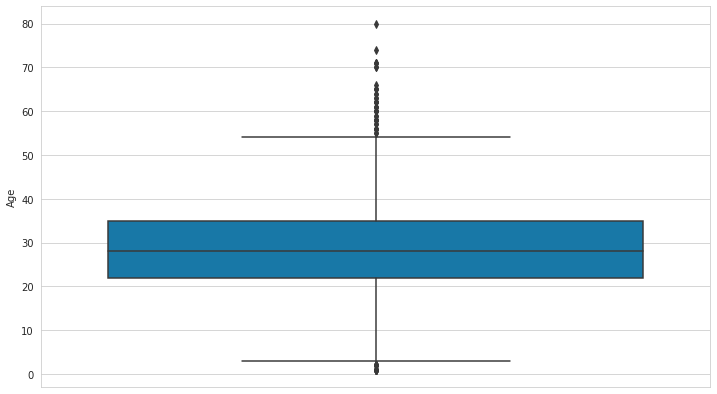
Outliers are basically too large or too small data points in a dataset. In other words, these are noticeably different from most of the data points. An outlier indicates variability or error in measurement. Outliers can have a disproportionate effect on the statistical results and the conclusions we make. They can have a significant effect on mean and standard deviation and they can skew the dataset and can cause serious problems in statistical analysis.

So, it is important that we remove such data points from the dataset. Outlier removal techniques involve Box-plot and scatter-plot method. We have adopted Boxplot method for the removal. The boxplot is a useful graphical display for describing the behavior of the data in the middle as well as at the ends of the distributions. The box plot uses Median (Q2), upper quartile (Q3) and lower quartile (Q1). (Q3-Q1) is called IQR (Interquartile range). Upper bound and lower bound are calculated using UB=Q3+1.5\*IQR and LB=Q1-1.5\*IQR. If a data point falls out of either of the bounds, it is removed. Outliers should be investigated carefully.

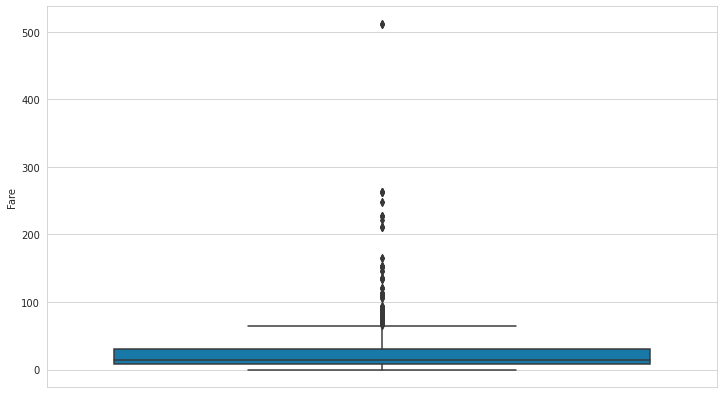
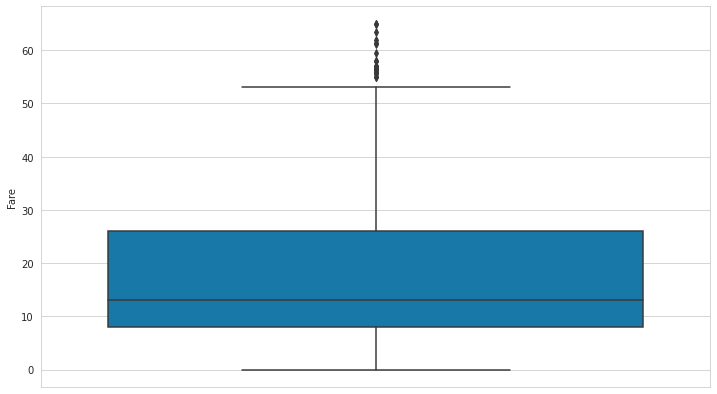
Often, they contain valuable information about the process. Before considering the elimination of these points from the data, one should try to understand why they appeared and whether it is likely that similar values will continue to appear. Most of the time, Outliers are often bad data-points.

BOXPLOTS BEFORE AND AFTER REMOVAL OF OUTLIERS:

FOR FARE COLUMN:



FOR AGE COLUMN:

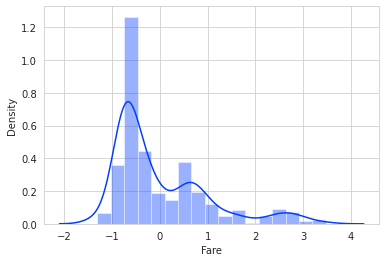
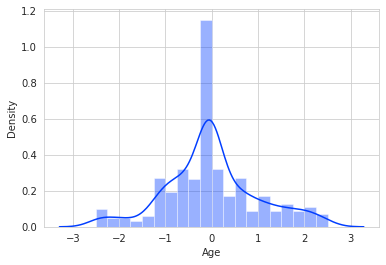
STANDARDISATION AND NORMALISATION:

Standardization is useful when the data in our dataset has varying scales in numeric columns and the techniques we use makes assumptions about our data having a Bell/Gaussian distribution such as in linear regression. It helps in multivariate analysis when you want all the variables to be in comparative units.

The result of standardization is that the features/variables will be rescaled so that they have the properties of a standard normal distribution with μ=0 and σ=1 when we use the transformation Z=(x-μ)/σ. X~N (0,1) where μ and σ are mean and standard deviation of the particular attribute.

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling. Normalization is a good technique to use when you do not know the distribution of your data or when you know the distribution is not Gaussian (a bell curve). Normalization is useful when your data has varying scales and the algorithm you are using does not make assumptions about the distribution of your data, such as k-nearest neighbors and artificial neural networks. The goal of normalization is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values. For machine learning, every dataset does not require normalization. It is required only when features have different ranges. Hence, standardization and normalization are very important techniques in statistical analysis.

CURVES AFTER NORMALISATION:

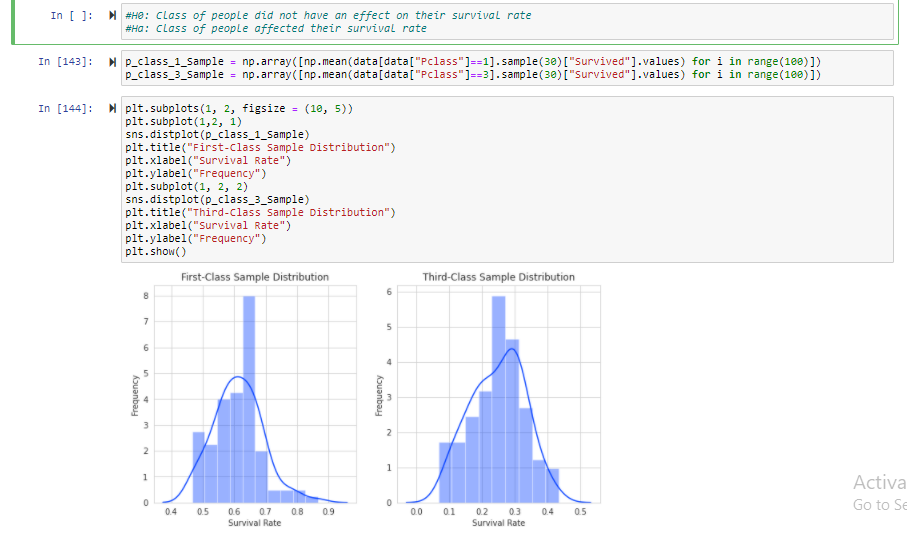
HYPOTHESIS TESTING:

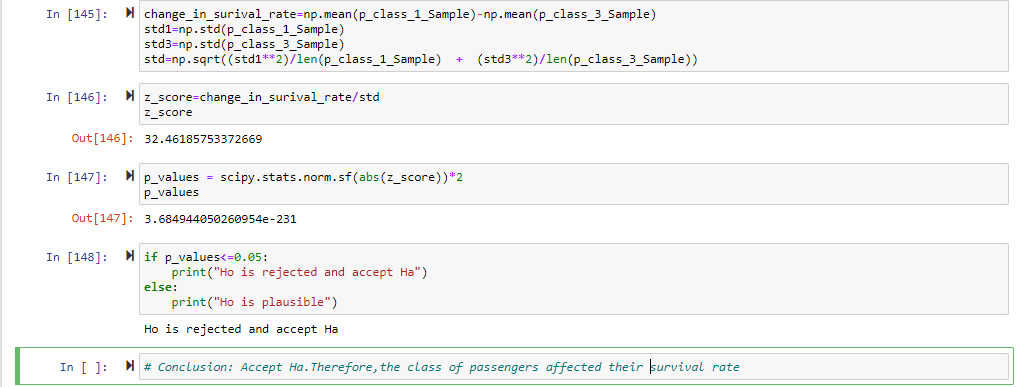
The Null and Alternate Hypothesis are stated are follows:

H0: The class of people didn’t affect their survival rate.

H1: The class of people affected their survival rate.

The hypothesis test is performed at 5% significance level. We perform Z-test for a sample (for passengers of class-1 and class-3) from the population. Z-statistic and corresponding p-value are found out. p-value is approximately zero. Since p<=0.05, this gives strong evidence against the Null hypothesis. We reject H0 and accept H1.



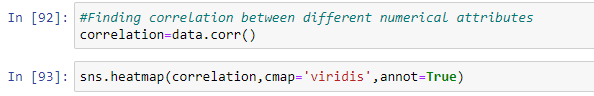


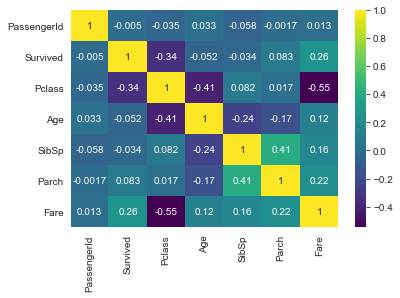
**Therefore, we conclude that class of people affected the survival rate of the passengers onboard.**

CORRELATION USING HEATMAP

When two sets of data are strongly linked together, we say they have a High Correlation. Correlation is Positive when the values increase together, and. Correlation is Negative when one value decreases as the other increases.

A heatmap is a graphical representation of data in which data values are represented as colors. That is, it uses color in order to communicate a value to the reader.





Each square shows the correlation between the variables on each axis. Correlation ranges from -1 to +1. Values closer to zero means there is no linear trend between the two variables. The closer to 1 the correlation is the more positively correlated they are; that is as one increases so does the other and the closer to 1 the stronger this relationship is.

A correlation closer to -1 is similar, but instead of both increasing one variable will decrease as the other increases. The diagonals are all 1/yellow because those squares are correlating each variable to itself (so it's a perfect correlation). For the rest the larger the number and brighter the color the higher the correlation between the two variables. The plot is also symmetrical about the diagonal since the same two variables are being paired together in those squares.

CONCLUSION

We started with the data exploration where we got a feeling for the dataset, checked about missing data and learned which features are important. During this process we used seaborn and matplotlib to do the visualizations. During the data preprocessing part, we computed missing values, performed data cleaning, and created a few new features.

From this analysis, we would be inclined to conclude that passengers had higher chances of survival if:

* they had a high-class ticket

We have applied the Z-test using Z-statistic and thus from the results obtained from the test, we thus concluded that the choice of the class made by the people did affect the survival rate.

* they were women

We utilized the visualization techniques to prove this

We have also observed that the rate of embarkment from Southampton was higher, resulting in both higher mortality and death rate for the passengers embarking from that port

We also inferred the correlation between various attributes of the dataset through the use of heatmap

Our team is thankful to Uma Ma'am for giving us this exciting opportunity of performing tests and analysis on datasets of the real world. This project proved to be a great learning experience, and has helped us to truly understand the practical implications of the various processes performed by us.