# Creating Reports, Dashboards, and Visualizations using Tableau with Titanic Disaster

## Introduction

### The Story and the Statistics

The Titanic, a luxury ship considered indestructible due to its advanced engineering (Stone, D., 2022), met a disastrous end when it hit an iceberg and sank in the North Atlantic Ocean on the fateful night of April 15, 1912. This disaster was responsible for more than 1500 deaths. The Titanic story is marvelous because of the calamity that it was, the loss of lives of innocent people, and the drama that ensued of the rescue and or lack of it.

### A Dataset for Discovery

Fortunately, the passenger list of the Titanic has been recorded in detail for those who had the misfortune of being on board. This dataset enables us to analyze more on the disaster and the variables that could have led to the varying results on those who survived.

* Who was entailing the highest risk? Were the opportunities of survival within the region influenced by means such as social status, gender or age?
* Is there a relationship between passenger characteristics and survival rates? The goal here is to identify whether specific groups of passengers had higher or lower chances of survival.
* Data analysis: Recounting the human aspect can be effective to use in charts and graphs to tell a great story about the disaster

Realizing that there is a wealth of information in this data, one is able to not only find out more about the occurrences on that fatal night but comprehend gender roles, disaster response, and other aspects of historical disasters.

## Data Preparation

The Titanic data set was sourced from Kaggle and includes different attributes regarding the passengers (Gao, L., 2024)such as class survival, age, siblings/spouse aboard, parents/children aboard, fare, cabin, and other features, embark town among others. To prepare the dataset for analysis in Tableau, the following data cleaning and preprocessing steps were undertaken:

### 1. Conversion of Survival Status:

- To encode the Survival information, the column “survived” included only two possible values – 0 and this was then changed to a new column using survival\_status as the name with 1 for ‘Survived’ and 0 for ‘Dead’.

Example:

*Import pandas as pd*

*titanic\_data=pd.read\_excel("data/titanic3.xls")*

*titanic\_data["survival\_status"]=titanic\_data["survived"].map({1:"Survived",0:"Dead"})*

### 2. Handling Missing Values:

- The age column for passengers was having some empty values; these were imputed using median age of passengers so that there isn’t much distortion in the data.

Example:

*titanic\_data ['age']. fillna(titanic\_data ['age']. median(), inplace=True)*

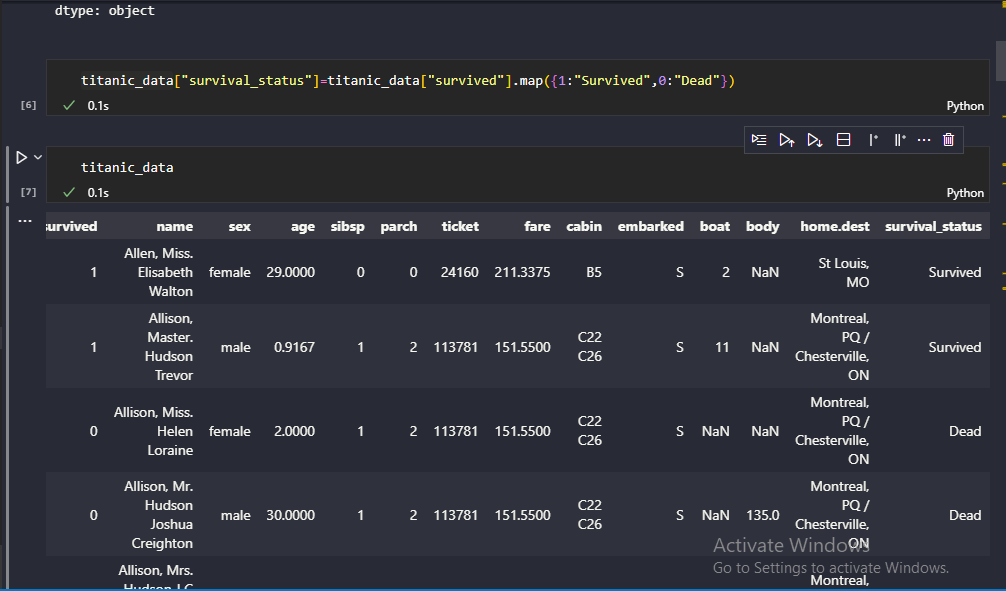
### 3. Dropping Unnecessary Columns:

- Specifically the following pre-processing steps were followed: Don’t need all the columns for analysis hence some of them were deleted. These included name, ticket, cabin, sex, boat, body, home. dest, and fare,

### 4. Encoding Categorical Variables:

- For titanic dataset machine learning integration (Ai, Y., 2023.), it is necessary to involve categorical data such as: sex and embarked which were encoded into numerical form. Sex being coded into a binary variable where 0 represented female and 1 represented male while embark was categorical variable encoded into a binary form using the one hot method.

To do this, use the ‘get\_dummies’ utility of the pandas library at hand, specifying the columns to make nominal and excluding one category of each of them. These preprocessing steps guaranteed that the data was clean and well preprocessed for analysis and visualization (Milani, A.M.P., Paulovich, F.V. and Manssour, I.H., 2020in tableau and to be blended with machine learning models for predictive modeling.



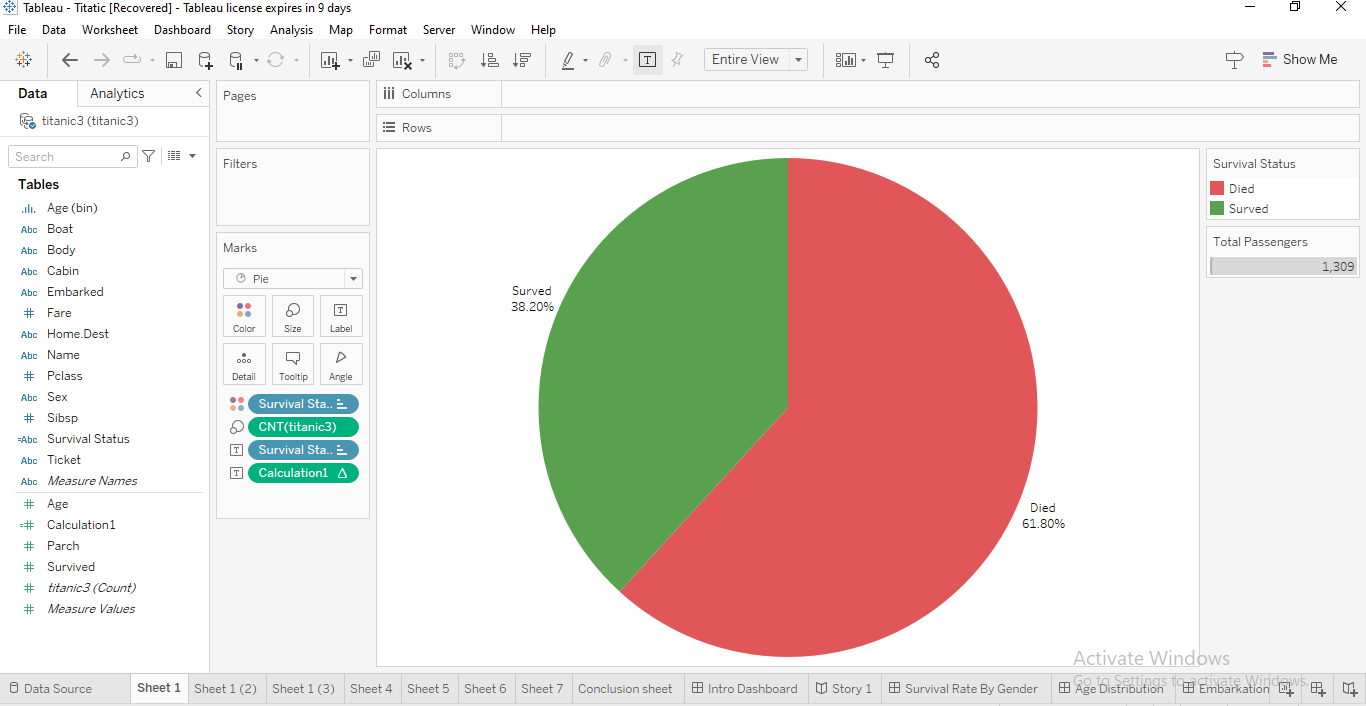
## Visualization Design

### Visualization 1: Pie chart depicting the survival status

This category of pie chart therefore seeks to compare the ratio of passengers who survived as against those who died. Among 1,309 total passengers, 500 (38%) were alive, and 809 (61. 80%) died.

**Justification**: Pie chart can clearly demonstrate the average rate of survival at a single glance and thereby draws the attention to the difference between survivor and non-survivor groups. It also paints an envisaged picture of much causality, an aspect that is crucial in highlighting the magnitude of the disaster and the inadequacy of lifeboats.

**Insights**: There were chances of survival if one belonged to a certain demographic, specifically women, children, and the upper class but it also probably depended on fate.



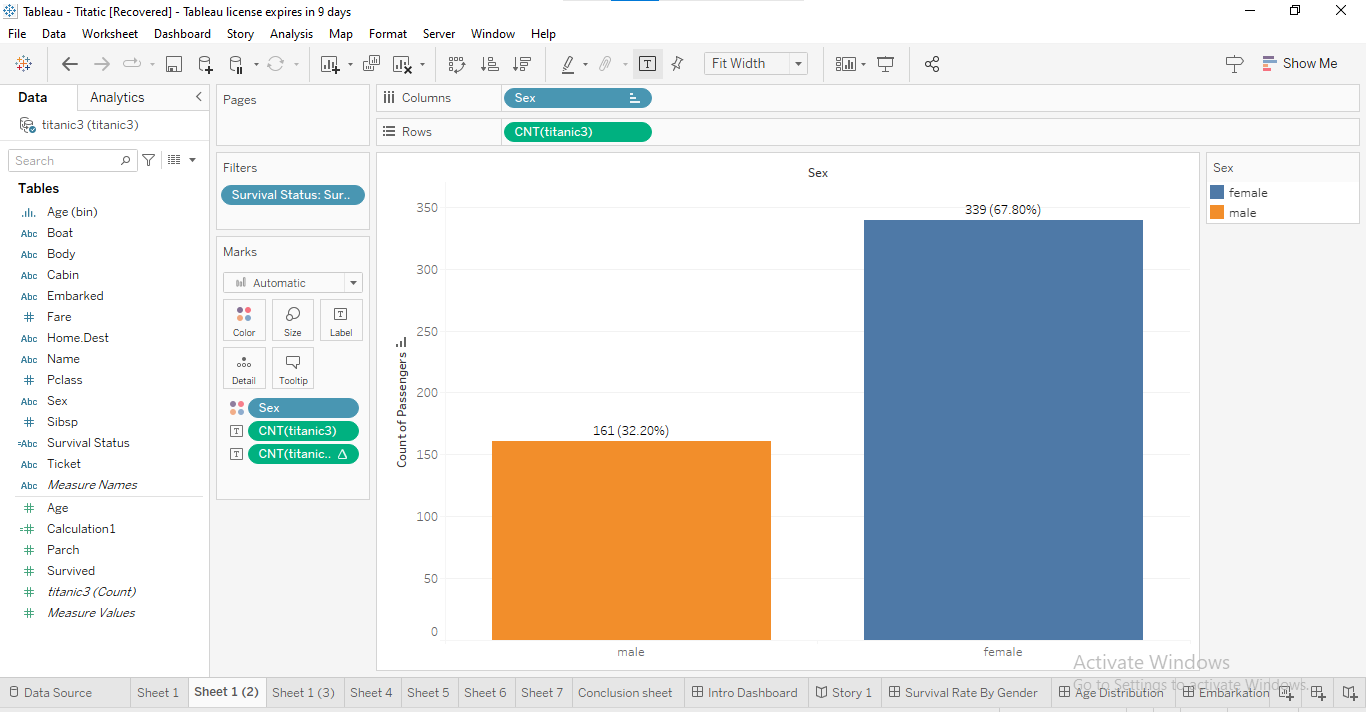
Title: Pie chart of survival status

### Visualization 2: Survival Status Distribution by Gender: Histogram

This histogram gives survival status of the passengers by their gender; it shows that there were 161(32.20%) males and 339(67.80%) females who survived.

**Justification**: The number of the female survivors is depicted as higher which is historically supported by evidence that women and children were given a priority.

**Insights**: More than three-fourths of the survivors were women below 30 years of age and more than three fourths of the perished were young men below 30 years of age.



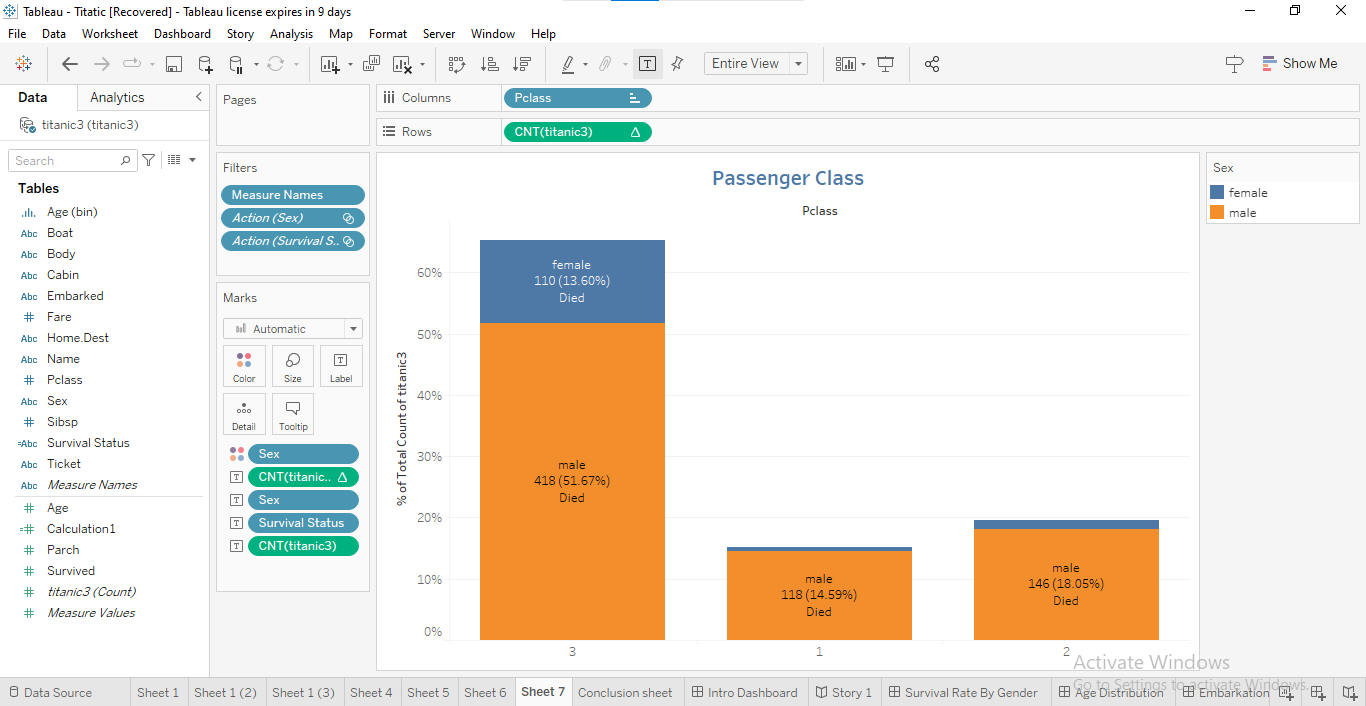
Title: Survival status by Gender

## Visualization 3: Survival by Passenger Class (Pclass)

This histogram represents the survival rates of passengers based on their class of service. In the third class 418 (51.67%) of males and 110 (13. 60 %) females population died. In the first class, the number of males who died was 118 (14.59%) and 5(0.62%) were females. They were 146 males (18.05%) and 12 (1.48%) females in second class who perished.

**Justification**: Through visually comparing survival rates between different classes, this histogram shows the effect of social class on survival. Ladies and first-class passengers had higher chances of survival; this shows that in the escape strategies and availability of lifeboat; there was prejudice.

**Insights**: Priced and class of tickets: Tickets’ price and class were other important factors because the first and second class passengers had higher chances of surviving.

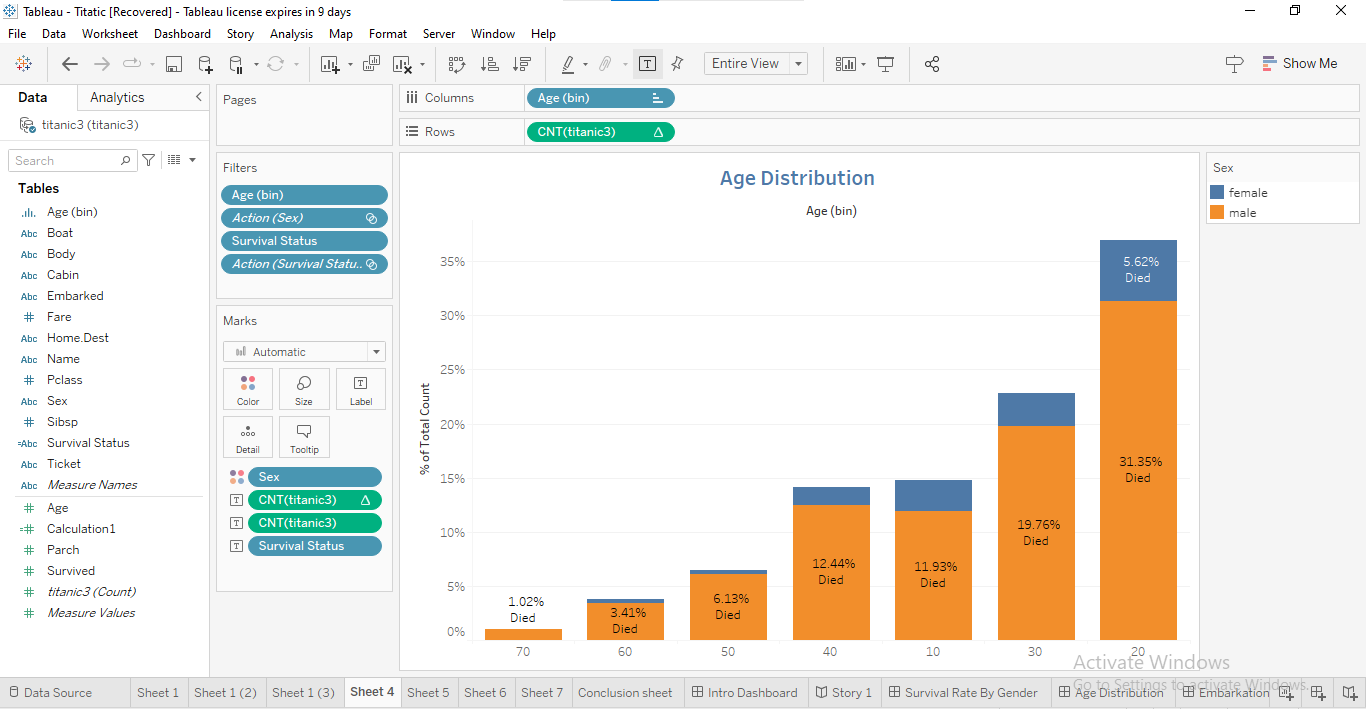


Title: Survival status by Pclass

### Visualization 4: Survival status based on age distribution.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Age | 10 | 20 | 30 | 40 | 50 | 60 | 70 |
| Males | 70(11.93%) | 184(31.35%) | 116(19.76%) | 73(12.44%) | 36(6.13%) | 20(3.41%) | 6(1.02%) |
| Females | 17(2.90%) | 33(5.62%) | 18(3.07%) | 10(1.70% | 2(0.34%) | 2(0.34%) | 2(0.34%) |

Table showing survival rate by age distribution, highlighting those who died



Title: Histogram showing survival rate by age distribution, highlighting those who died

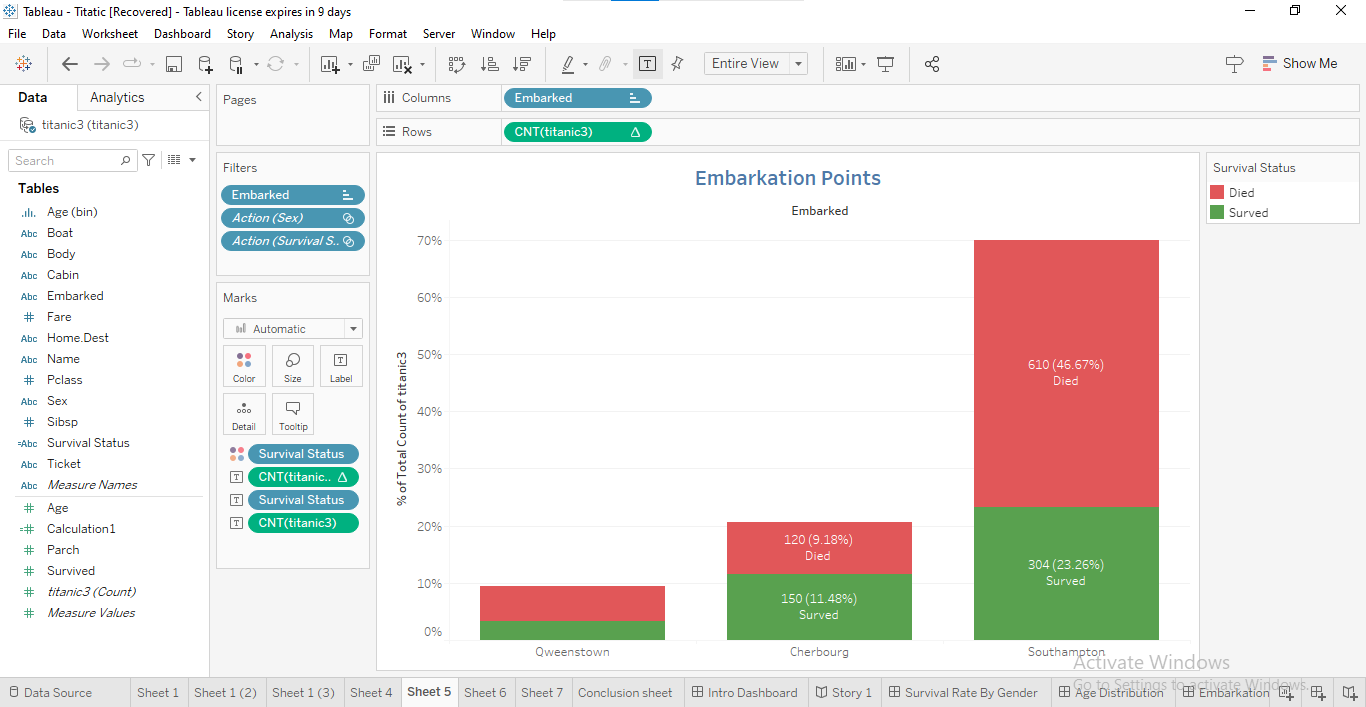
**Justification**: The age histogram helps to understand the difference in the age of the survived and non-survived people. It reveals which age categories were at the risk of perishing and which had higher probabilities of survival.

**Insights**: In most of the casualties, the fatalities were young men aged between 20 and 29 years, while most of the survivors were females below 30 years.

### Visualization 5: Survival Status by Port of Embarkation Histogram

|  |  |  |  |
| --- | --- | --- | --- |
| Port of Embarkation | Queenstown(Q) | Cherbourg (C) | Southampton (S) |
| Died | 79(9.77%) | 120(14.83%) | 610(75.04%) |
| Survived | 44(3.37%) | 150(11.48%) | 304(23.26%) |

Title: Table showing survival rate by port of embarkation, highlighting those who died and survived.



Title: Histogram showing survival rate by port of embarkation, highlighting those who died and survived.

**Justification**: This chart displays survival probability with respect to embarkation factors. The particular ports had different survival rates implying the possible demography of passengers and ship amenities.

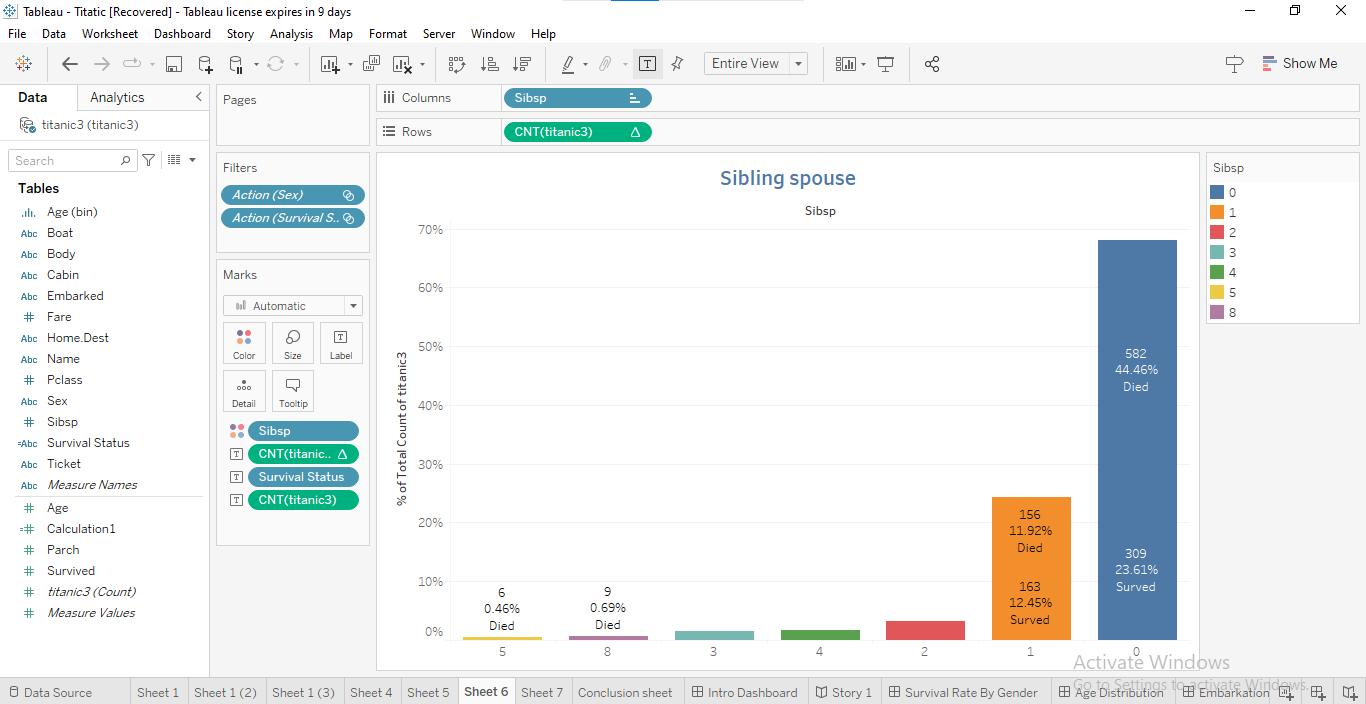
**Insights**: However, the data presented shows that passengers that embarked from the port of Southampton had the lowest chance of survival with a mere 23% surviving.

### Visualization 6: Survival Status by Sibling-Spouse Relationship (SibSp) Distribution

This histogram is to illustrate the verdict rates of the dying persons with brothers/sisters on the ship or mothers/fathers on board or wives/husbands on board.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Sibsp | 0 | 1 | 2 | 3 | 4 | 5 | 8 |
| Died | 582(44.46%) | 156(11.92%) | 23(1.76%) | 14(1.07%) | 19(1.45%) | 6(0.46%) | 9(0.69%) |
| Survived | 309(23.60%) | 163(12.45%) | 19(1.45%) | 16(0.46% |  |  |  |

Title: Table showing survival rate by sibling spouse relationship, highlighting those who died and survived.



Title: Histogram showing survival rate by sibling spouse relationship, highlighting those who died and survived.

**Justification**: The histogram enlightens on the rate of survival, should there be presence of family members. It also worries about the correlation of the number of family members taking a trip on the ship with the chances of survival.

**Insights**: Most of the dead persons were not accompanied by any other person; this is an implication that persons who were travelling alone may have suffered a loss.

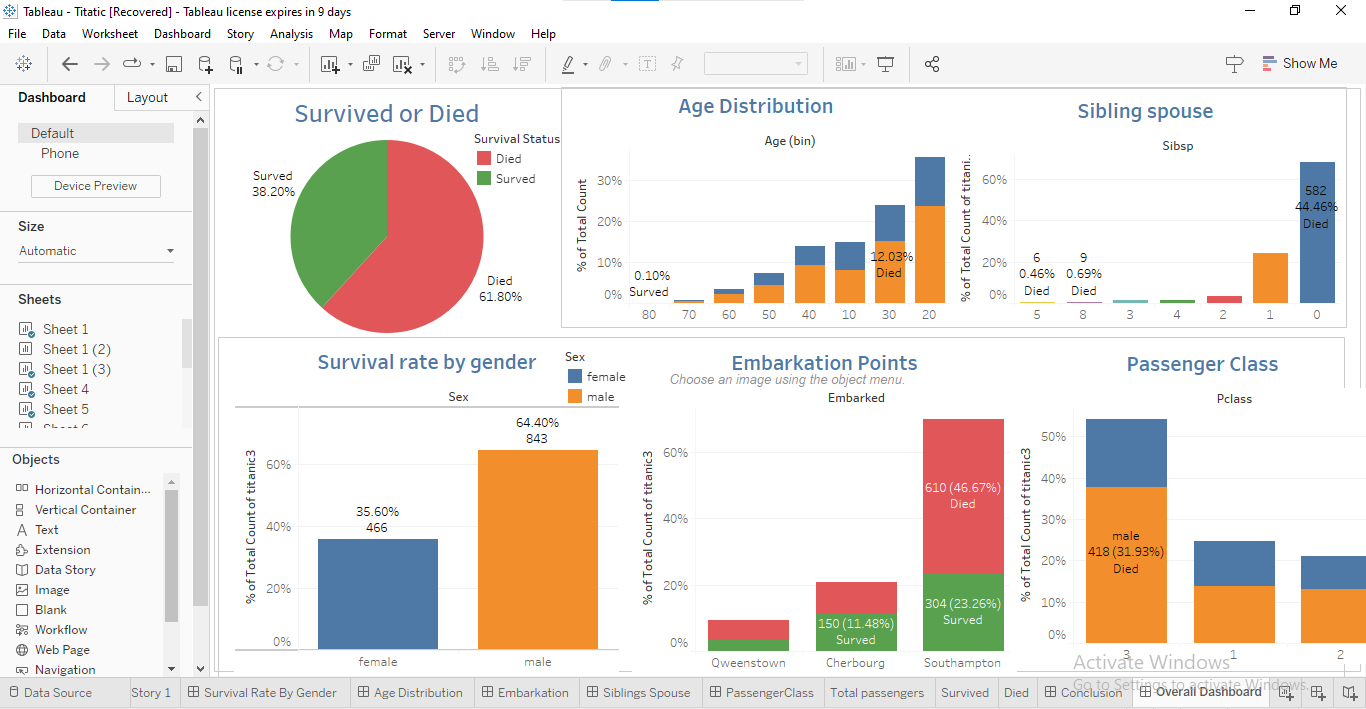
## Dashboards

### Explaining the Dashboard Design

The dashboard contains an overview of the survival status of Titanic passengers in the form of various graphics and tables (Haque, M.A., Shivaprasad, G. and Guruprasad, G., 2021.). It comprises a pie chart and several histograms that compare the survival of passengers through gender, passenger’s class, age, the port of embarkation, and whether the passenger is a sibling or spouse of another on board the ship. All the visualizations are arranged in such a manner that one can easily make comparisons and draw conclusions.

### Interactive Features and Their Use

* Filter Controls: There are filter options presented on the top of the dashboard which enable users to work with particular subsets of data by such variables as gender, passenger class, and the port of embarkation. This makes it easier for the user to navigate through the data and focus on a particular area of interest.
* Tooltips: Hovering over any area of the pie chart or bars in the histograms shows tooltips with the values and percentages. This feature is useful in giving the user accurate information without this making the display area congested.
* Dynamic Legends: Legend tab is especially engaging because users can click on it to emphasize or filter out some types of data in the visualizations.



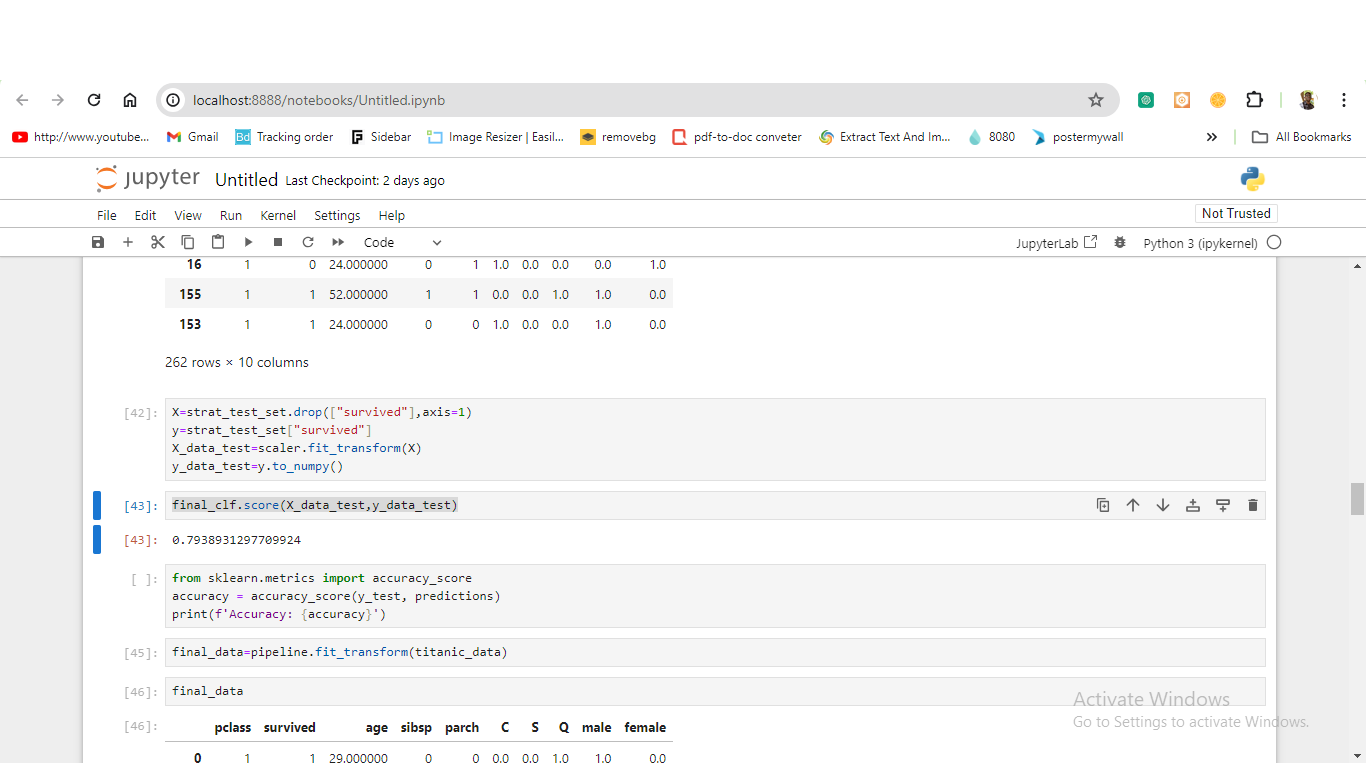
Title: Dashboard featuring a pie chart of survival status and histograms for gender, passenger class, age, port of embarkation and sibling/spouse relationships etc.

## Machine Learning Model

### Model Specification and Evaluation

The problem has always been to predict if the Titanic passengers would survive (Tabbakh, A., Rout, J.K. and Rout, M., 2021) or not, depending on their age, gender, passenger class, and number of siblings/spouse on board. The chosen model is a **Random Forest Classifier**, it is a strong algorithm with high accuracy in classification.

* Data Preparation: First of all, the data set was cleaned by imputing missing values, encoding categorial variables and excluding redundant fields.
* Model Training: The Random Forest Classifier was used to model the data, where seventy percent of the data was used for training while twenty percent was used for testing. The model under test produced the accuracy of 79%.
* Performance Metrics: This included the confusion matrix, accuracy score, precision, and recall as the metrics for testing the entity. These metrics showed that the proposed model provided a significantly better differentiation between people who survived and those who did not.



Title: Model Performance Metrics: Random Forest Classifier

### Steps Taken for Integration with Tableau

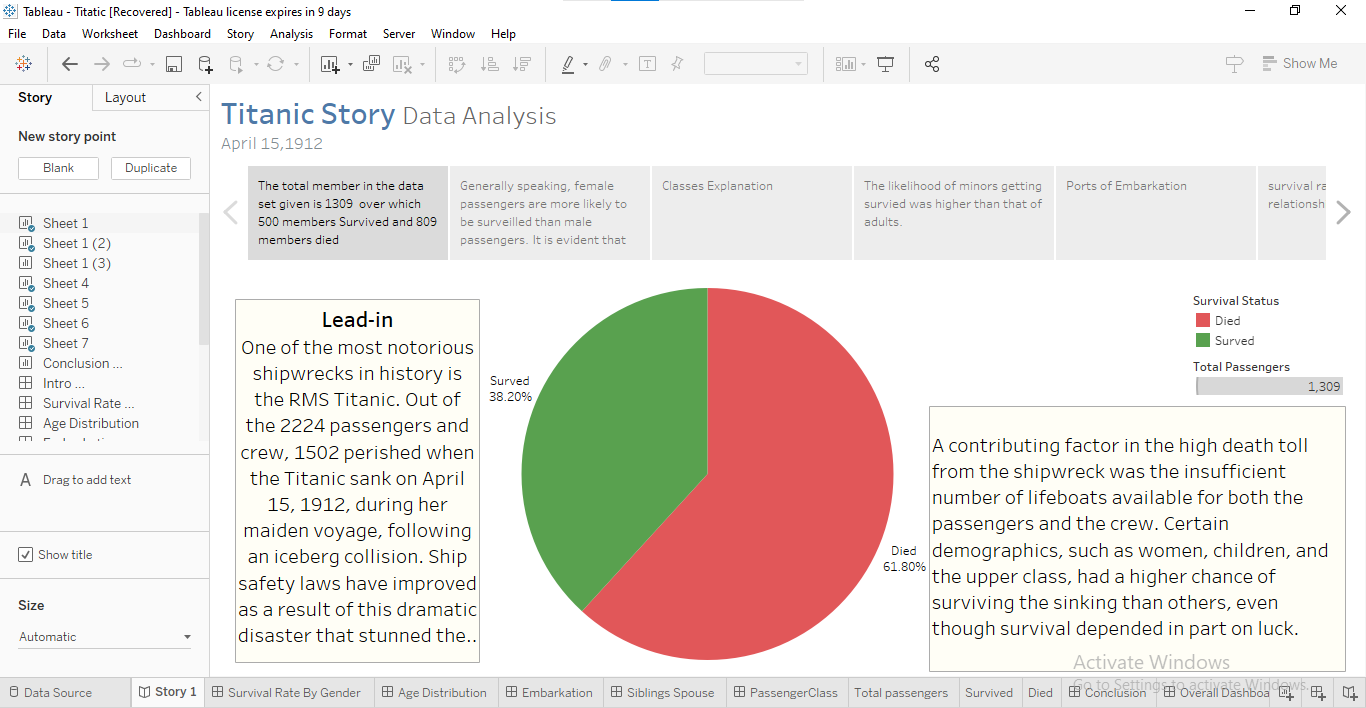
* Installing TabPy: The dependency package called TabPy (Tableau Python Server) was installed using pip.
* Starting TabPy Server: Execute the bat file to begin the tabpy server at port 9004.
* Used the application Tableau and clicked on the “Help” button located in the upper-right corner of the application window.
* Click the tab called “Settings and Performance’ and then select ‘Manage External Service Connection’.
* As the service,selected “TabPy/External API”, in Server ,wrote ‘localhost’ and for port number wrote ‘9004’.
* Right-clicked on each of the queries and clicked “Test Connection” to confirm that Tableau was able to establish a proper connection with the TabPy server.
* Creating Calculated Fields: In Tableau, developed calculated fields that use the existing Python functions on the TabPy server to make the solution based on the machine learning model. The factors that will be used are the passenger class, the age of the passenger, the number of siblings/spouses aboard, and the number of parents/children aboard, labeled as [Pclass], [Age], [SibSp], [Parch] respectively.
* Visualizing Predictions: For this purpose, utilized the manually computed fields in Tableau to develop visualizations that indicate the predicted survival status of a patient along with vital information. This integration enables users to apply dynamic and real-time analysis of model predictions and other elements within the Tableau dashboards.

When the machine learning model is embedded into Tableau, the respective forecast data can be shared back to the dashboard layer of Tableau, improving the application of storytelling and analysis in the visualization environment.

## Storytelling Narrative

### Overview of the Storytelling Narrative

It was found that the use of storytelling in Tableau libraries controls the narrative and directs the viewer’s attention to the aspects of the data related to the events that happened during the Titanic disaster and their impact on passengers’ chances of survival. The first shot in the film is a called all survived a video that contains graphic content that gives an indication of the survival rates where the pie chart shows the difference between the survivals and non-survivals. It then proceeds to an expert assessment accompanied by histograms highlighting survival rates based on gender, passenger class, age, and region of embarkation as well as by index compiled in regards to the relation between passengers’ siblings or spouses. The story also uses the machine learning forecast to present a theoretical-digital environment where various key factors of influence can be analyzed and demonstrated in depth, thus making prediction of influences toward survival more comprehensible.



Title: Visual Storytelling of Titanic Disaster in Tableau

This structure of the narrative again puts together an extensive and highly memorable explanation of the Titanic tragedy, augmented by visualization and artificial intelligence.

## Conclusion

To sum up, the findings highlight several factors that will increase the probability of passenger survival based on the Titanic dataset analysis. From the explorations made in tableau, higher survivability was detected in female, children, and the higher class passengers. Furthermore, likelihood of survival depended on different factors such as age, the class of passengers, gender and the port of embarkation.

The combination of machine learning with visualization also proved very effective in how the titanic data analysis (Gupta, A., Arora, D. and Tiwari, S., 2023) was done since it offered what-if analysis and meant that multiple scenarios could be examined concurrently. This integration made it possible for the analyst to take a closer look at the results obtained and find out details that could not have been discovered when using other techniques of analyzing the data. In conclusion, one can note that both machine learning and data visualization are effective at creating the titanic story (Cao, E.Y., Xie, W., Dong, C. and Qiu, J., 2020) and for analyzing big data such as the event described in the dataset – the Titanic disaster.

## References

1. Stone, D., 2022. *Sinkable: Obsession, the Deep Sea, and the Shipwreck of the Titanic*. Penguin.
2. Gao, L., 2024. Data set analysis of Titanic distress data. *Highlights in Science, Engineering and Technology*, *92*, pp.323-329.
3. Milani, A.M.P., Paulovich, F.V. and Manssour, I.H., 2020. Visualization in the preprocessing phase: Getting insights from enterprise professionals. *Information Visualization*, *19*(4), pp.273-287.
4. Gupta, A., Arora, D. and Tiwari, S., 2023, May. Exploratory Data Analysis of Titanic Survival Prediction using Machine Learning Techniques. In *2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)* (pp. 418-422). IEEE.
5. Cao, E.Y., Xie, W., Dong, C. and Qiu, J., 2020. Titanic Machine Learning Study from Disaster.
6. Tabbakh, A., Rout, J.K. and Rout, M., 2021. Analysis and prediction of the survival of titanic passengers using machine learning. In *Advances in Distributed Computing and Machine Learning: Proceedings of ICADCML 2020* (pp. 297-304). Springer Singapore.
7. Haque, M.A., Shivaprasad, G. and Guruprasad, G., 2021, February. Passenger data analysis of Titanic using machine learning approach in the context of chances of surviving the disaster. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1065, No. 1, p. 012042). IOP Publishing.
8. Ai, Y., 2023. Predicting Titanic Survivors by Using Machine Learning. *Highlights in Science, Engineering and Technology*, *34*, pp.360-367.