Report

Task 1

Part 1 Question

> What kind of data set is this?

This dataset is an overview of the passengers who died on the Titanic. The data contains information about each person, such as their age, name, gender, and which cabin on the ship they were assigned.

To begin analyzing the data, I will first take a look at the overall structure and content of the dataset. The data includes columns of different types, including integers, floats, and objects (which are strings). By understanding the types of data that are included in the dataset, it will be possible to determine which methods and techniques are most appropriate for analyzing the data.

The 'PassengerId' column is an integer that is used as an index for the passengers in the dataset. This information is not particularly useful for analyzing the data, and can be dropped from the dataset during the analysis process. It is a continuous numerical data type.

The 'Survived' column is a boolean that indicates whether or not a passenger survived the Titanic disaster. This is an important piece of information that will be used to analyze the data and identify any patterns or trends. It is a categorical data type, and Binary.

The 'PClass' column is an integer that indicates the socio-economic class of the passenger. There are three possible values for this column: 1 (for higher class), 2 (for middle class), and 3 (for lower class). This information can be used to understand the demographics of the passengers on the Titanic, and how this may have affected their chances of survival. It is a categorical data type.

The 'Name' column is a string that contains the name of the passenger. This information can be used to identify the gender and marital status of the passenger, based on the title that is included in their name (e.g. "Miss", "Mrs", "Mr", etc.). It is a categorical data type.

The 'Sex' column is a string that indicates the gender of the passenger. This information can be used to understand the distribution of males and females. It is a categorical data type.

The 'Age' column is a float that indicates the age of the passenger. Fractional ages below 1 are represented as decimal values in this column. This information can be used to understand the distribution of ages on the Titanic. It is a continuous numerical data type.

The 'SibSp' column is an integer that indicates the number of siblings and spouses that a passenger had on board the Titanic. This information can be used to understand the demographics of the passengers on the Titanic. It is a continuous numerical data type.

The 'Parch' column is an integer that indicates the number of parents and children that a passenger had on board the Titanic. This information can be used to understand the demographics of the passengers on the Titanic. It is a continuous numerical data type.

The 'Ticket' column is a string that contains the ticket number for each passenger. This information could potentially be used as a unique identifier for each passenger, since it is unlikely that two passengers would have the same ticket number. However, it is possible that multiple passengers within a family may have been assigned the same ticket number. It is a categorical data type.

The 'Fare' column is a float that indicates the price that a passenger paid for their ticket on the Titanic. It is likely that the prices were similar for passengers in the same socio-economic class, but there may be some differences based on factors such as where the ticket was purchased or the tax laws in the passenger's home state. It is a continuous numerical data type.

The 'Cabin' column is a string that contains a unique identifier for the cabin that a passenger was assigned on the Titanic. This identifier is not unique for each passenger, as multiple passengers may have been assigned the same cabin. It is a categorical data type.

The 'Embarked' column is a character that indicates the port from which a passenger boarded the Titanic. There are three possible values for this column: 'C' (for Cherbourg), 'S' (for Southampton), and 'Q' (for Queenstown). This information can be used to understand the distribution of passengers on the Titanic. It is a categorical data type.

One interesting aspect of the data is the size of each passenger's family. To better understand this, I will create a new column called 'Family' that combines the 'SibSp' and 'Parch' columns to give a single numeric value for the number of siblings, spouses, parents, and children that a passenger was traveling with.

Another interesting aspect of the data is whether or not a passenger was traveling alone. To explore this, I will create a new column called 'Traveling_Alone' that will be a binary value indicating whether a passenger was traveling without any family members. This can be determined by checking if the 'Family' column has a value of 0.

Task 2

2a

> What can you use this data set for? Name at least 2 different applications, or examples of getting value out of the data set.

Application / Example 1

One potential use for the data in this dataset is to analyze the socio-economic status of passengers based on their passenger class ('Pclass'), the price of their ticket ('Fare'), and their point of embarkment ('Embarked'). By examining these factors, we can gain insight into the wealth of each city's population and make more informed decisions about what products to offer and how to target potential customers. For example, we could use this information to determine if we should be more exclusive in our onboarding process to better accommodate wealthier families, or if we should offer a wider range of products at different price points to cater to a broader range of customers. This type of analysis could also be useful in other contexts, such as cities using it to determine how to price goods and services.

Application / Example 2

One potential analysis of the data is to examine the relationship between the size of a passenger's family and their chances of survival. By comparing the number of parents and children (as indicated by the 'Parch'

column) with the 'Survived' column, we can determine if larger families had a higher or lower death rate than passengers who were traveling alone. This information could be used to make more informed decisions about where to place larger families during an evacuation, such as ensuring that they have access to more escape boats or other safety measures.

2b

To improve the quality of the data, some processing will be necessary. The 'Cabin' column has a large number of missing values (over 60%), and it is difficult to find values for this column. One possible solution would be to map persons by name and ticket to try to make sense of the cabin data, and then fill in the missing values using this information. However, since the applications that will be using the data do not require cabin information, it may be more efficient to simply exclude the cabin column from the dataset.

The 'Age' column also has a large number of missing values, but these can be filled in using mean calculations. The 'Embarked' column is missing only two values, which can be filled in with the most common values in that column.

Task 3

3a

> What methods would I apply to the dataset?

To begin the analysis of the given dataset, the first step would be to familiarize myself with the data by exploring the variable names, data types, and any missing or invalid values, as done above. This would help me understand the structure and content of the data, and identify any potential issues or inconsistencies that may affect the analysis.

After that, I would perform any necessary data cleaning and preprocessing steps, such as removing missing or invalid values, transforming or scaling the data, or combining multiple variables into a single variable and / or doing encoding. This would ensure that the data is in a suitable form for analysis and that any potential issues or biases are minimized.

Once the data is prepared, I would begin the actual analysis by applying appropriate statistical methods and visualizations to explore and analyze the data. This would typically involve calculating descriptive statistics, such as mean, median and mode.

The chosen application aims to save lives by analyzing the death rate of families and using this information to improve evacuation planning. By analyzing the number of parents and children (Parch) in the dataset, and comparing this information to the survival rates of passengers, it may be possible to identify patterns that can be used to plan more effective evacuations. This could include placing larger families in areas of the ship with more escape boats, or ensuring that there are enough lifeboats available to accommodate all passengers, regardless of their family size. By applying these methods to the data, the chosen application can help to improve the safety of passengers in the event of an emergency.

First of all, I would first calculate the number of passengers who had a certain number of parents or children on board, lets call this column familiy, as well as the number of passengers who survived in each of those groups.

Next, I will calculate the survival rate for each group of passengers with a certain number of family members on board. This could be done by dividing the number of passengers who survived in a group by the total number of passengers in that group.

Once the survival rates have been calculated, I will plot them on a graph to visualize any patterns or trends in the data. For example, if the survival rate was higher for larger families, this could suggest that larger families had a better chance of surviving the disaster.

One potential approach to analyzing the data would be to use statistical methods and visualizations to understand the relationship between family size and survival. This approach would allow us to examine the strength and significance of the relationship between these variables, without the need to build a predictive model. Instead of using a machine learning algorithm, we could compute measures of association such as the Pearson correlation coefficient or use a chi-square test to determine whether there is a significant relationship between family size and survival. We could also create visualizations of the data, such as bar charts or scatterplots, to help us identify trends or patterns in the data.

While building a machine learning model may also be a useful way to analyze the data, it may not be the best approach in this case. Since the goal is to understand the relationship between family size and survival in order to inform decisions about evacuation procedures, a more descriptive approach may be more suitable. By using statistical methods and visualizations, we can gain a better understanding of the characteristics of the data and identify any patterns or trends that may be present. This information can then be used to inform decisions about how to best protect larger families during an evacuation.

Task 4

4a / b

> What did you get out of the data? Show concrete numbers, figures and graphs.

My interpretation of the question is to show my findings of the analysis in task 3b. There are my findings.

When examining the scatter plots of age and family size, it is clear that the age of the deceased group is substantially higher than that of the survived group. Additionally, the deceased group has larger families and is younger on average. This can be seen in the histograms for both groups, where there are more younger individuals in the survived group. When looking at the mean, skew, and standard deviation of age for both groups, it is apparent that the numbers are lower for the survived group. A t-test of the correlation between these two variables shows a significant relationship between them.

When examining the Pclass of both groups, it is clear that there are more individuals from the lower class in the deceased group. The mean Pclass of the deceased group is above 2.5, indicating that there were mostly middle to lower class individuals. The standard deviation of Pclass for the deceased group is 0.7, indicating that the spread is not as large. In comparison, the mean Pclass of the survived group is below 2, indicating that it was mostly higher to middle class individuals that survived. The histogram for Pclass for both groups shows that there were mostly lower class individuals that died, while both middle and higher class had fewer individuals. In the survived group, higher class individuals had the highest survival rate, with 62% surviving. These percentages of Pclass survival in the survived group are 47% for middle class and 24% for lower class.

When examining the sex of the passengers, it is clear that sex played a significant role in survival rates. In the deceased group, there were fewer than 100 females and over 400 males. In the surviving group, there were just over 100 males and more than 200 females. In terms of percentages, 74% of females survived, while only 18% of males survived. It is possible that this data is biased due to the higher proportion of male staff members and the fact that they were all in the lower class. However, it is worth noting that historically, males have often prioritized the rescue of women and children. Nevertheless, males had a lower survival rate.

Based on the analysis, it can be concluded that age, family, sex, and pclass had a significant correlation with survival rate.

TOOLS

You'll find all libraries used in 'requirements.txt'.