Greeting…

The Titanic was a huge passenger ship going from England to America in 1912. It hit a large block of ice in the ocean and badly damaged the ship. Most people on board died because there were not enough lifeboats for everyone. Only about 700 people survived the tragedy, it was one of the worst large ship accidents in history.

In simple terms:

* Titanic was a passenger ship travelling in 1912
* It hit a large iceberg that damaged the ship
* Approximate number of passengers (2,200)
* Over 1500 passengers died
* Not enough lifeboats for all people
* One of the worst shipping accidents ever

The first step in any data analysis project is importing the Python libraries we will need for working with data and visualizing the analysis.

For the Titanic dataset, some key libraries we import are:

**Pandas** - For loading, manipulating and analyzing dataset. It allows us to store data in DataFrames similar to Excel tables and has many built-in methods.

**NumPy** - Provides support for mathematical and numerical operations on arrays and matrices which is useful for statistical analysis.

**Matplotlib and Seaborn** - Primary Python data visualization libraries that allow us to create plots, charts, histograms to visualize trends and patterns in data.

**Data Transformation Libraries** - Libraries like scikit-learn can be used for data transformations at later stages of analysis.

**Loading Data:**

Import pandas and use pd.read\_csv() to load titanic.csv dataset into a DataFrame. Use .head() method to view first few rows and get a sense of the structure.

The Titanic dataset provides information about each passenger on the ship such as:

PassengerId - A unique ID number for each passenger

Survived - Whether the passenger survived (1) or died (0)

Pclass - The passenger's class (1st, 2nd, 3rd class)

Name - The passenger's name

Sex - Male or female

Age - Age of passenger

SibSp - # of siblings/spouses aboard

Parch - # of parents/children aboard

Ticket - Ticket number

Fare - Amount passenger paid

Cabin - Cabin number

Embarked - Where passenger boarded

Our goal is to analyze the above passenger attributes to understand what factors influenced a person's chance of surviving the disaster. Things like passenger class, gender, age and fare paid may reveal interesting insights into survival odds.

Further, we looked into the shape of the dataset, which shows the number of columns and rows we have: 891 rows and 12 columns, respectively.

With this much data available, let's inspect the top and bottom rows of the dataset. To display the topmost rows, we can use the code **df.head()**, and similarly, **df.tail()** shows the bottommost rows from the dataset.

Furthermore, we looked into the information of the dataset using **df.info()**. This method displays the number of columns, number of rows, null values, data types of the columns, and the memory location it occupies.

Additionally, **df.describe()** provides useful statistical summary (count, mean, std dev etc) for numerical columns. This quickly conveys valuable insights into distribution of features like Age, Fare etc. **df.dtypes** returns the data types for each column in a DataFrame.

**Example: -**

Now that we have gathered all the raw building materials, our next step is to prepare and process them for high quality construction. Just like how stones and metal rods can't be used as is from a quarry or foundry to make a house, real-world data also needs cleaning and processing.

We need to polish the stones, trim unnecessary edges, check them for cracks or defects. Metal rods may need to be cut, welded, bent into shape first before framework and structure building can begin. Sand and cement mix needs right moisture levels, proper ratios and consistent texture for concreting.

**Data Cleaning:**

In this step, we understand the presence of null or missing values and address duplicate rows. We then proceed to handle missing values, typically through one of three methods: Acceptance, Deletion, or Imputation. Imputation involves replacing null values with predetermined values based on the category of the data.

If we're not utilizing a machine learning model, we may simply opt to delete rows with null or missing values, as well as duplicate rows.

We observed that one columns have 77% null values, and the PassengerId column has No missing values but doesn’t required for analysis.

As part of the data reduction step, we remove unwanted columns from the dataset. Certain columns or variables can be dropped if they do not contribute value to our analysis.

Given there are 177 missing values in the Age column, we need to select an appropriate imputation strategy for filling them in.

We have missing age values we need to fill in. I want to use the middle-most age or median age to fill them.

Using middle age is better than average age or most common age. Because -

* Middle not affected much by very high or low ages.
* Middle age considers all ages equally not just most common one.
* Find the middle age is easy to do.
* Filling with middle reasonable guess for missing.

**Feature Engineering on Age:**

* Imagine you have a list of people's ages, like 25, 30, 40, etc.
* Instead of using each specific age, we group them into categories like Child, Adult, Middle Age, Senior using the hot encoded.

**Feature Engineering for Embarked:**

* Suppose we have information about where people boarded a ship, but it's encoded in a way that's not very easy to understand: C, Q, S.
* We translate these codes into names of cities: Cherbourg, Queenstown, Southampton.

Up to this stage, just like builders refine raw materials for building.

**EDA Analysis: -**

Finally, we enter into EDA, which stands for Exploratory Data Analysis.

Using **df.describe().T**, we gain insight into the statistical description of the data.

The need to find mean, standard deviation, max, min, and count (measures of central tendency) serves several purposes:

* Summarizing data to get a quick overview of its distribution and characteristics.
* Handling missing values by understanding their impact on the overall dataset.
* Identifying extreme values that deviate significantly from the rest of the data. These outliers can indicate data quality issues, such as data entry errors or measurement inconsistencies.

After describing the data, we proceed to separate numerical and categorical columns based on their data types.

Moving on to analysis, we start with univariate analysis, which focuses on describing or summarizing one variable at a time.

Specifically, we begin by analyzing numerical variables. All numerical columns are iterated through using a for loop.

Following this, we plot a histogram or distplot and a boxplot. The histogram displays the distribution of the data, while the boxplot defines the outliers in the data.

Next, we analyze categorical variables using countplot or barplot, which display the count of each unique value.

Further, we move on to bivariate analysis, which involves describing or summarizing two variables and exploring the relationship between them.

Here, we plot a heatmap or pairplot to visualize the relationship between variables. It's better to use a heatmap because it clearly denotes the values of correlation.

From this plot, we select the dependent and feature engineering columns to make decisions. In the titanic data, we select "Survival" as dependent variables. This helps us significant disparities in survival outcomes based on gender, age group, passenger class, and travel companionship.

Finding the count of the Survival and plot a pie chart and bar chart for the same.

Similarly finding the gender count and plot a pie chart and bar chart for the same.

Imagine we have a bunch of people from the Titanic, and we want to see who survived and who didn't, based on their age groups.

Then, we count how many people from each age group survived and how many didn't. Similarly, we do the same for gender and class.

We want to represent different groups of people on the Titanic, such as passengers of different ages or genders. Our aim is to plot a graph showing how many passengers survived based on class and gender.

Finally, we conclude by summarizing the overall analysis and the insights gained.

Finally, we are ready with very beautiful Building to leave.