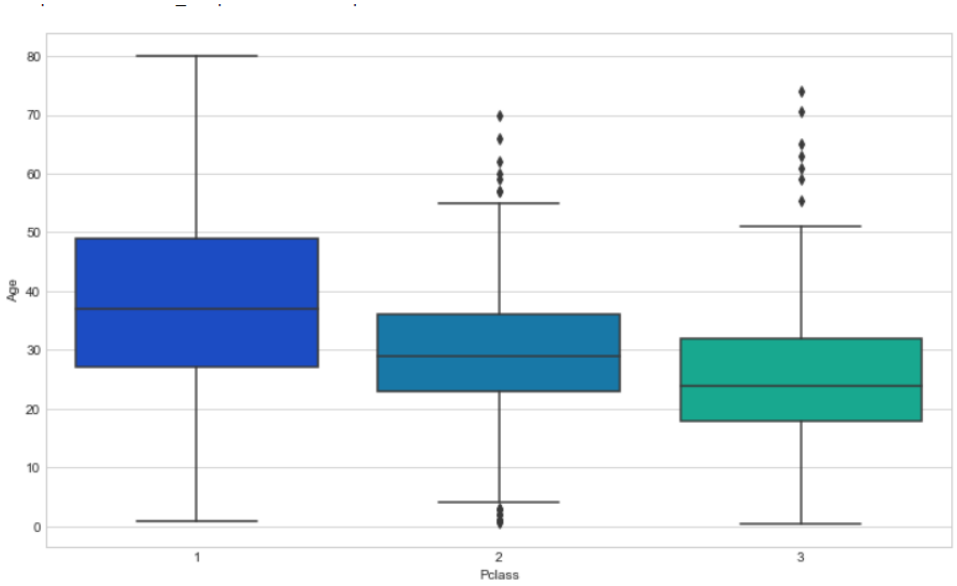
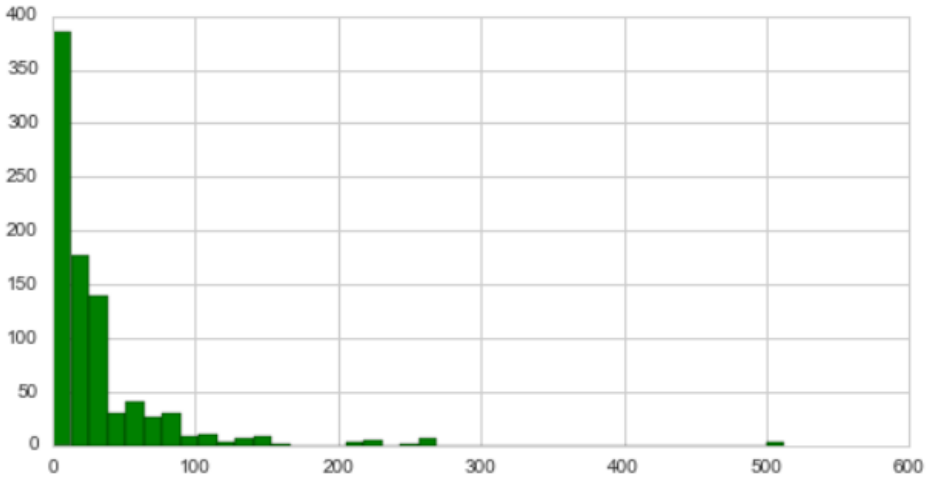
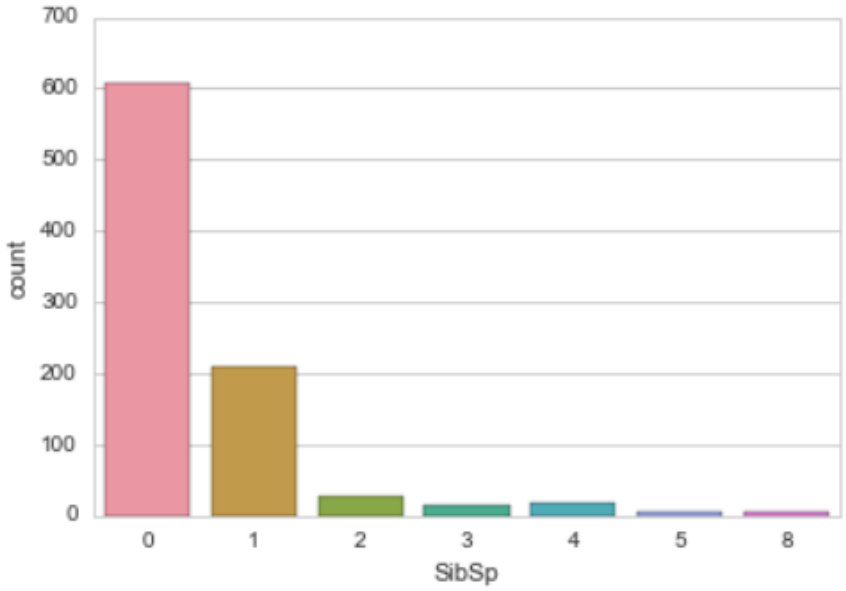
Project Name: EDA With Visualization Using Titanic Dataset Manually

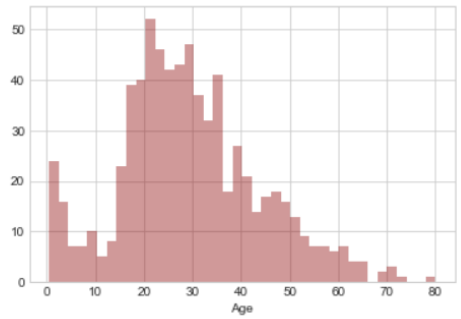
Table of Content

Demo



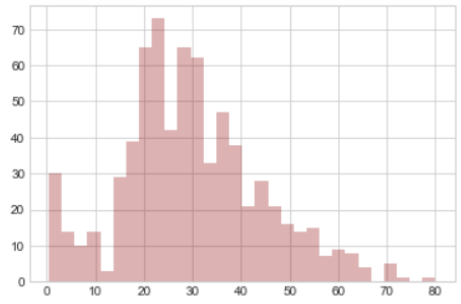


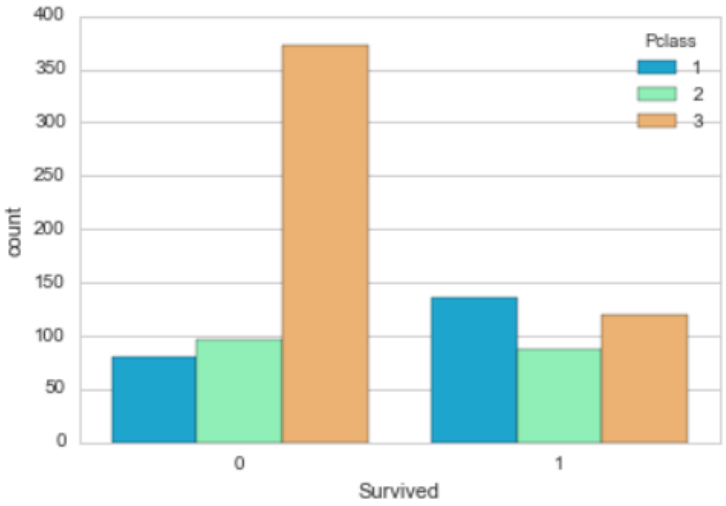





```
In [25]: train['Age'].hist(bins=30,color='darkred',alpha=0.3)
```

```
Out[25]: <matplotlib.axes._subplots.AxesSubplot at 0xe2d8978>
```





Overview

This is an Exploratory Data Analysis with Visualization using Titanic Dataset.
EDA is essentially a type of storytelling for statisticians.
It allows us to uncover patterns and insights, often with visual methods, within data.
EDA is often the first step of the data modelling process.
This repository contains the code for EDA along with Visualization using python's various libraries.
It used numpy, pandas, matplotlib and seaborn libraries.
These libraries help to perform individually one particular functionality.
Numpy is used for working with arrays. It stands for Numerical Python.
Pandas objects rely heavily on Numpy objects.
Matplotlib is a plotting library.
Seaborn is data visualization library based on matplotlib.
The purpose of creating this repository is to gain insights into EDA.
These python libraries raised knowledge in discovering these libraries with practical use of it.
It leads to growth in my ML repository.
These above screenshots will help you to understand flow of output.

Motivation

The reason behind building this is, to maximize I as analyst's insight into a data set and into the underlying structure of a data set, while providing all of the specific items that an analyst would want to extract. It is a way of visualizing, summarizing and interpreting the information that is hidden in rows and column format. EDA is understanding the data sets by summarizing their main characteristics often plotting them visually. This step is very important for me especially when I arrive at modelling the data in order to apply Machine learning. Another motive is, as a master's student I have learnt Data Mining Subject which has somewhere led me to also learn about EDA. Though, EDA and Data Mining has difference. EDA distinguishes itself from data mining, even though the two are closely related, as many EDA techniques have been adopted into data mining. Also, the goals of the two are very similar: EDA indeed makes sure that you

explore the data in such a way that interesting features and relationships between features will become clearer. In EDA, you typically explore and compare many different variables with a variety of techniques to search and find systematic patterns. Data mining, on the other hand, is concerned with extracting patterns from the data. Those patterns provide insights into relationships between variables that can be used to improve business decisions. Also, in both cases, you have no a priori expectations or expectations that are not complete about the relations between the variables. However, in general, Data Mining can be said to be more application-oriented, while EDA is concerned with the basic nature of the underlying phenomena. In other words, Data Mining is relatively less concerned with identifying the specific relations between the involved variables. As a result, Data Mining accepts a "black box" approach to data exploration and doesn't only use techniques that are also used in EDA but also techniques such as Neural Networks to generate valid predictions but don't identify the specific nature of the relationships between the variables on which the predictions are based. Exploratory Data Analysis (EDA) is used on the one hand to answer questions, test business assumptions, generate hypotheses for further analysis. On the other hand, you can also use it to prepare the data for modelling. The thing that these two probably have in common is a good knowledge of your data to either get the answers that you need or to develop an intuition for interpreting the results of future modelling. There are a lot of ways to reach these goals: you can get a basic description of the data, visualize it, identify patterns in it, identify challenges of using the data, etc. Hence, I continue to gain knowledge while practicing the same and spread intellectual wings in tech-heaven.

Technical Aspect

Numpy contains a multi-dimensional array and matrix data structures. It works with the numerical data. Numpy is faster because is densely packed in memory due to its homogeneous type. It also frees the memory faster.

Pandas module mainly works with the tabular data. It contains Data Frame and Series. Pandas is 18 to 20 times slower than Numpy. Pandas is seriously a game changer when it comes to cleaning, transforming, manipulating and analyzing data.

Matplotlib is used for EDA. Visualization of graphs helps to understand data in better way than numbers in table format. Matplotlib is mainly deployed for basic plotting. It consists of bars, pies, lines, scatter plots and so on. Inline command display visualization inline within frontends like in Jupyter Notebook, directly below the code cell that produced it.

Seaborn provides a high-level interface for drawing attractive and informative statistical graphics. It provides a variety of visualization patterns and visualize random distributions.

Need to train_test_split - Using the same dataset for both training and testing leaves room for miscalculations, thus increases the chances of inaccurate predictions.

The train_test_split function allows you to break a dataset with ease while pursuing an ideal model. Also, keep in mind that your model should not be overfitting or underfitting.

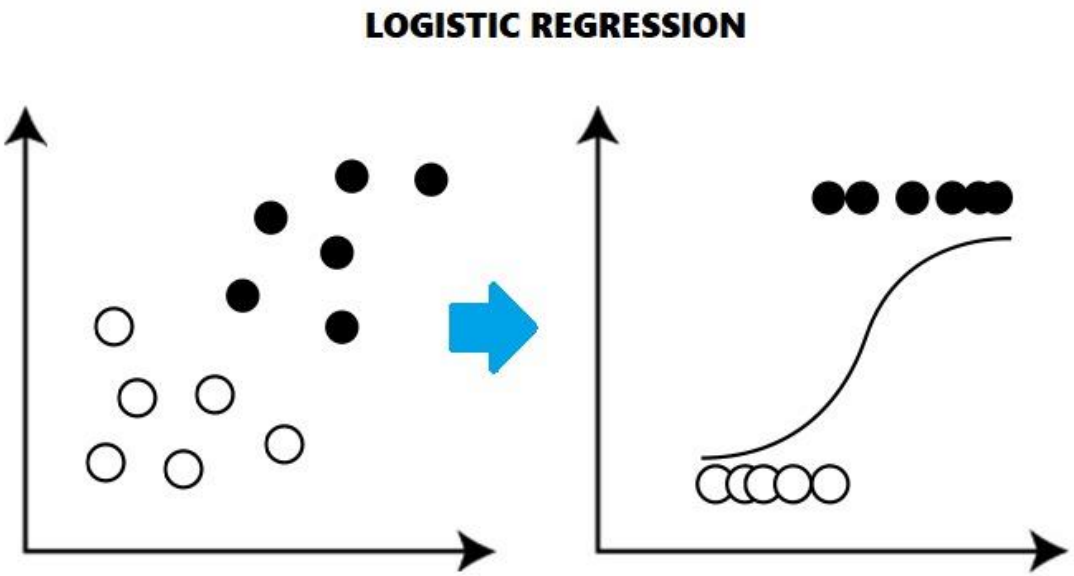
Logistic regression is appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). I Used simple logistic regression here because when you have one nominal variable and one measurement variable, and you want to know whether variation in the measurement variable causes variation in the nominal variable.

Confusion matrix, accuracy_score, classification_report – sklearn.matrix – Refer to

https://scikit-learn.org/stable/modules/model_evaluation.html

https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html

A confusion matrix is a summary of prediction results on a classification problem.
The classification report visualizer displays the precision, recall, F1, and support scores for the model. In order to support easier interpretation and problem detection, the report integrates numerical scores with a color-coded heatmap. All heatmaps are in the range (0.0, 1.0) to facilitate easy comparison of classification models across different classification reports.



Installation

Using intel core i5 9th generation with NVIDIA GFORECE GTX1650.
Windows 10 Environment Used.
Already Installed Anaconda Navigator for Python 3.x
The Code is written in Python 3.8.
If you don't have Python installed then please install Anaconda Navigator from its official site.
If you are using a lower version of Python you can upgrade using the pip package, ensuring you have the latest version of pip, *python* -m pip install --*upgrade pip and press Enter*.

Run/How to Use/Steps

Keep your internet connection on while running or accessing files and throughout too.
Follow this when you want to perform from scratch.

Open Anaconda Prompt, Perform the following steps:

cd <PATH>
pip install numpy
pip install pandas
pip install matplotlib
pip install seaborn

pip install sklearn

You can also create requirement.txt file as, pip freeze > requirements.txt

run files.

Follow this when you want to just perform on local machine.
Download ZIP File.
Right-Click on ZIP file in download section and select Extract file option, which will unzip file.
Move unzip folder to desired folder/location be it D drive or desktop etc.
Open Anaconda Prompt, write cd <PATH> and press Enter.
eg: cd C:\Users\Monica\Desktop\Projects\Python Projects 1\11)EDA_and_Visualization\titanic_dataset_manually

In Anconda Prompt, pip install -r requirements.txt to install all packages.
Open in Jupyter Notebook, <filename>.ipynb
That is,
Open in Jupyter Notebook, 1)EDA_With_Visualization_manually.ipynb
This takes titanic_train.csv file as input dataset.
Please be careful with spellings or numbers while typing filename and easier is just copy filename and then run it to avoid any silly errors.
Note: cd <PATH>
[Go to Folder where file is. Select the path from top and right-click and select copy option and paste it next to cd one space <path> and press enter, then you can access all files of that folder] [cd means change directory]

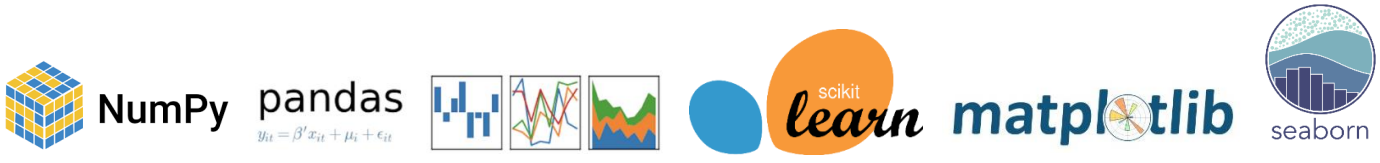
Directory Tree/Structure of Project

Folder: 11)EDA_and_Visualization > titanic_dataset_manually
1)EDA_With_Visualization_manually.ipynb

To Do/Future Scope

Can try with other datasets.

Technologies Used/System Requirement/Tech Stack



Credits

Krish Naik Channel