**Data Visualization I**

1. Use the inbuilt dataset 'titanic'. The dataset contains 891 rows and contains information about the passengers who boarded the unfortunate Titanic ship. Use the Seaborn library to see if we can find any patterns in the data.
2. Write a code to check how the price of the ticket (column name: 'fare') for each

passenger is distributed by plotting a histogram.

**Seaborn** is a library mostly used for statistical plotting in Python. It is built on top of Matplotlib and provides beautiful default styles and color palettes to make statistical plots more attractive.

## Different categories of plot in Seaborn

Plots are basically used for visualizing the relationship between variables. Those variables can be either be completely numerical or a category like a group, class or division. Seaborn divides plot into the below categories – 

* **Relational plots:** This plot is used to understand the relation between two variables.
* [**Categorical plots:**](https://www.geeksforgeeks.org/seaborn-categorical-plots/)This plot deals with categorical variables and how they can be visualized.
* [**Distribution plots:**](https://www.geeksforgeeks.org/seaborn-distribution-plots/)This plot is used for examining univariate and bivariate distributions
* [**Regression plots:**](https://www.geeksforgeeks.org/seaborn-regression-plots/)The regression plots in seaborn are primarily intended to add a visual guide that helps to emphasize patterns in a dataset during exploratory data analyses.
* [**Matrix plots:**](https://www.geeksforgeeks.org/ml-matrix-plots-in-seaborn/) A matrix plot is an array of scatterplots.
* **Multi-plot grids:**It is an useful approach is to draw multiple instances of the same plot on different subsets of the dataset.

Titanic dataset is downloaded by default with the Seaborn library.

Or

Source: <https://www.kaggle.com/c/titanic/data>

**The dataset consists of 891 rows and 12 columns .**

1. PassengerId
2. Survived
3. Pclass
4. Name
5. Sex
6. age
7. Sib Sp
8. Parch
9. Ticket
10. Fare
11. Cabin
12. Embarked

The Seaborn library is built on top of Matplotlib and offers many advanced data visualization capabilities.

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

print(sns.get\_dataset\_names()) # to check inbuild dataset

dataset = sns.load\_dataset('titanic')

dataset.head()

The dataset contains 891 rows and 15 columns and contains information about the passengers who boarded the unfortunate Titanic ship.

**Features:**The titanic dataset has roughly the following types of features:

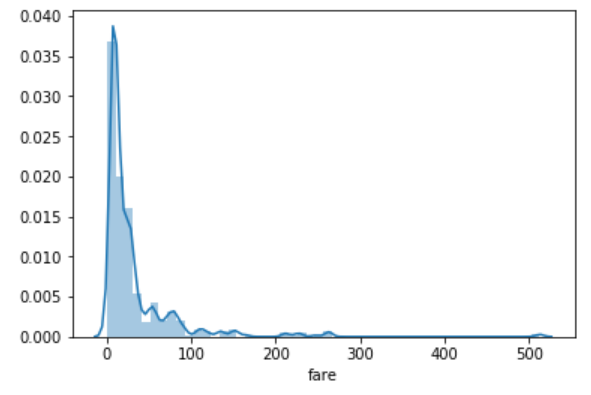
* **Categorical/Nominal**: Variables that can be divided into multiple categories but having no order or priority.   
  Eg. Embarked (C = Cherbourg; Q = Queenstown; S = Southampton)
* **Binary**: A subtype of categorical features, where the variable has only two categories.   
  Eg: Sex (Male/Female)
* **Ordinal**: They are similar to categorical features but they have an order(i.e can be sorted).   
  Eg. Pclass (1, 2, 3)
* **Continuous**: They can take up any value between the minimum and maximum values in a column.   
  Eg. Age, Fare
* **Count**: They represent the count of a variable.   
  Eg. SibSp, Parch
* **Useless**: They don’t contribute to the final outcome of an ML model. Here, *PassengerId, Name, Cabin* and *Ticket* might fall into this category.

**Distributional Plots**

Distributional plots, as the name suggests are type of plots that show the statistical distribution of data.

The distplot() shows the histogram distribution of data for a single column. The column name is passed as a parameter to the distplot() function.

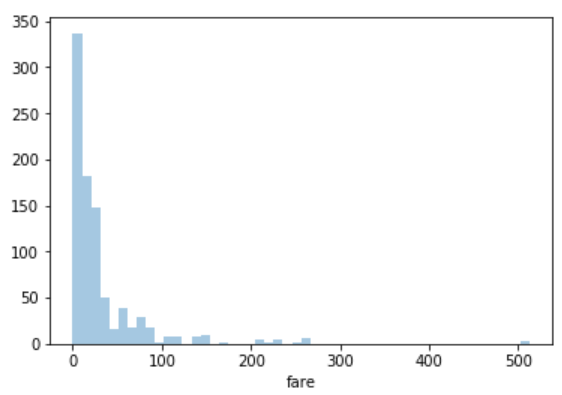
sns.distplot(dataset['fare'])



You can see that most of the tickets have been solved between 0-50 dollars.

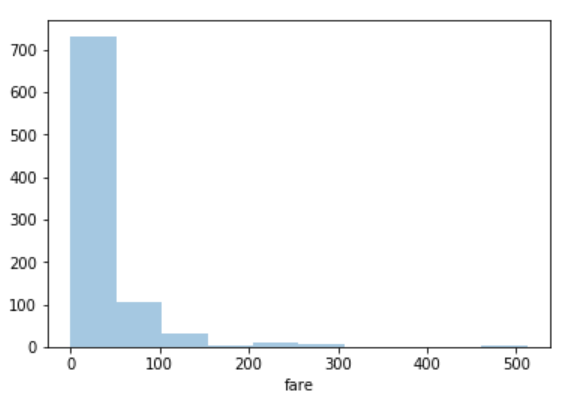
The line that you see represents the [kernel density estimation](https://en.wikipedia.org/wiki/Kernel_density_estimation). You can remove this line by passing False as the parameter for the kde attribute as shown below:

sns.distplot(dataset['fare'], kde=False)



sns.distplot(dataset['fare'], kde=False, bins=10)

Here we set the number of bins to 10. In the output, you will see data distributed in 10 bins as shown below:



**Histogram :**

**Histograms** are visualization tools that represent the distribution of a set of continuous data. In a histogram, the data is divided into a set of intervals or **bins** (usually on the x-axis) and the count of data points that fall into each bin corresponding to the height of the bar above that bin. These bins may or may not be equal in width but are adjacent (with no gaps).

A**density plot** (also known as kernel density plot) is another visualization tool for evaluating data distributions. It can be considered as a smoothed histogram. The peaks of a density plot help display where values are concentrated over the interval. There are a variety of smoothing techniques. **Kernel Density Estimation** (KDE) is one of the techniques used to smooth a histogram.

***Syntax:****seaborn.histplot(data, x, y, hue, stat, bins, binwidth, discrete, kde, log\_scale)*

***Parameters:-***

* ***data:****input data in the form of Dataframe or Numpy array*
* ***x, y****(optional): key of the data to be positioned on the x and y axes respectively*
* ***hue****(optional): semantic data key which is mapped to determine the color of plot elements*
* ***stat****(optional): count, frequency, density or probability*

***Return:****This method returns the matplotlib axes with the plot drawn on it.*

[Python | Titanic Data EDA using Seaborn - GeeksforGeeks](https://www.geeksforgeeks.org/python-titanic-data-eda-using-seaborn/)