Do the complete EDA in details to explore the insights of data and write the detailed observations of each analysis .

EDA - Exploratory Data Analysis: Using Python Functions

Well, first things first. We will load the titanic dataset into python to perform EDA.

#Load the required libraries

import pandas as pd

import numpy as np

import seaborn as sns

#Load the data

df = pd.read_csv('titanic.csv')

#View the data

df.head()

1. Basic information about data - EDA

The df.info() function will give us the basic information about the dataset. For any data, it is good to start by knowing its information. Let's see how it works with our data.

#Basic information

df.info()

#Describe the data

df.describe()

2. You can use the df.duplicate.sum() function to the sum of duplicate value present if any. It will show the number of duplicate values if they are present in the data.

#Find the duplicates

df.duplicated().sum()Duplicate values

3. Unique values in the data

You can find the number of unique values in the particular column using unique() function in python.

#unique values

df['Pclass'].unique()

```
df['Survived'].unique()
```

```
df['Sex'].unique()
```

```
array([3, 1, 2], dtype=int64)
```

```
array([0, 1], dtype=int64)
```

```
array(['male', 'female'], dtype=object)
```

4. Visualize the Unique counts

```
#Plot the unique values
```

```
sns.countplot(df['Pclass']).unique()
```

5.Find the Null values

Finding the null values is the most important step in the EDA. As I told many a time, ensuring the quality of data is paramount. So, let's see how we can find the null values.

```
#Find null values
```

```
df.isnull().sum()
```

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

6. Replace the Null values

Hey, we got a replace() function to replace all the null values with a specific data. It is too good!

#Replace null values

df.replace(np.nan,'0',inplace = True)

#Check the changes now

df.isnull().sum()

PassengerId 0

Survived 0

Pclass 0

Name 0

Sex 0

Age 0

SibSp 0

Parch 0

Ticket 0

Fare 0

Cabin 0

Embarked 0

dtype: int64

7. Know the datatypes

Knowing the datatypes which you are exploring is very important and an easy process too. Let's see how it works.

#Datatypes

df.dtypes

PassengerId int64

Survived int64

Pclass int64

Name object

Sex object

Age object

SibSp int64

Parch int64

Ticket object

Fare float64

Cabin object

Embarked object

dtype: object

8. Filter the Data

Yes, you can filter the data based on some logic.

#Filter data

df[df['Pclass']==1].head()

9. A quick box plot

You can create a box plot for any numerical column using a single line of code.

#Boxplot

df[['Fare']].boxplot()

Eda Boxplot

10. CFinally, to find the correlation among the variables, we can make use of the correlation function. This will give you a fair idea of the correlation strength between different variables.

#Correlation

df.corr()orrelation Plot - EDA

#Correlation plot

sns.heatmap(df.corr())

Ending Note - EDA

EDA is the most important part of any analysis. You will get to know many things about your data. You will find answers to your most of the questions with EDA. I have tried to show most of the python functions used for exploring the data with visualizations. I hope you got something from this article.

That’s all for now! Happy Python :)