1

DSBDAL

Assignment No:- 09

* Data Visualization II

1) use the input dataset 'titanic' as used in the above problem. plot a box plot for distribution of age with respect to each gender along with information about whether they survived or not.

2) Write observations on the inference from the above stastics.

Theory:-

What is Data Visualization?

- Data visualization represents the text or numerical data in visual format, which makes it easy to group the information the data express.

We humans, remember the pictures more easily than readable text, so python provides us various libraries for data visualization like matplotlib, seaborn, ploty.

* Exploratory Data Analysis:-

Creating Hypothesis, testing various business assumptions while dealing with any m/c learning problems stostements. is very important & this is about EDA.

We will use a very popular Titanic dataset with which everyone is familire with & you can download it.

- univariate Analysis -

univariate analysis is the simplest form of analysis where we explore a single variable.

univadiate analysis is performed to describe the data into better way, we perform univariate analysis of numerical & the categorical variables differently because plotting uses different plots.

- categorical Data:-

A variable that has text-based information is reffered to as categorical variables.

Various plot's we can use for visualizing categorical data.

1) count plot:-

countplot is basically a count of frequency plot in form of a bar graph.

It plots the count of each categorical data.

Bivariate / Multivariate Analysis.

We have study about various plot to explore single categorical & numerical data.

And when we analyze more than 2 variables together then it is known as multivariate analysis.

- Numerical & Numerical.

1) Scatter plot:-

- To plot the relationship betn two numerical variables, scatter plot is a simple plot to do.
- Relationship betn the total bill & tip
sns. scatterplot (tips {"total -bill"}, tips ['tip'])

- multivariate analysis with scatter plot:-

- We can also plot 3 variable or 4 variable relationship with scatter plot.
- suppose we want to find the separate ratio male & female with the total bill & tip provided.

Now, along with gender I also want to know whether the customer was a smoker or not so we can do this,

```
sns. scatterplot (tips ['total_bill"], tips
["tip"], hue = tips ["sex"], style = tips
['smoker"])
   plt. show
```

• Numerical & categorical:-

1) Bar plot:-

- Bar plot is a simple plot which we can use to plot categorical variable on the x-axis and numerical variable on y axis and explore the relationship betn Both the variables.

```
sns. barplot (data ['pclass'], data ['Age'])
   plt. show()

sns. barplot (data ['pclass'], data
['fare'], hue = data ["sex"])
       plt. show ()
```

When we use the pandas value counts function or any column, it is the same visual form of the value counts functions.

2) Distplot:-

- Distplot is also known as the second Histogaam because it is a slight.

3) Boxplot:-

- Boxplot is a very interesting plot that basically plots a 5 number summary.

• median - middle value in series after sorting.

* categorical & categorical:-

1) Heatmap
2) cluster map
3) Boxplot

```
sns. boxplot (data ('sex'), data ("Age"] )

sns. boxplot (data ("sex"), data ("Age"), data ("survived"))
    plt. show ()
```

4) Distplot:-

- Distplot explains the PDF function using kernel desity estimation.

sns. distplot (data [data ('survived']=0] ['Age'] ,hist =false , color ="blue")