Assignment No.9

Name - Kuldharan Sumit Dattatraya
Class - TE
Div - 4
Subject - DSBDAL

Problem Statement -

1. Use the inbuilt dataset 'titanic' as used in the above problem. Plot a box plot for distribution of age with respect to each gender along with the information about whether they survived or not. (Column names: 'sex' and 'age')
2. Write observations on the inference from the above statistics.

Theory -

① Explain box plot with example.

→ A Box plot is also known as Whisker plot is created to display the summary of the set of data values having properties like minimum, first quartile, median, third quartile & maximum. Here x-axis denotes the data to be plotted while the y-axis shows the frequency distribution.

Syntax to create box plot using matplotlib -

matplotlib.pyplot.boxplot(data, notch = None, vert = None, patch_artist = None, widths = None)

Example -

```
import matplotlib.pyplot as plt
import numpy as np
np.random.seed(10)
data = np.random.normal(100,20, 200)
fig = plt.figure(figsize =(10,7))
plt.boxplot(data)
plt.show()
```

2) How to read box plot ?

→ ① find the minimum.
② find the first quartile.
③ find the median.
④ find the third quartile.
⑤ find the maximum.

3) Explain Boxplot() with different attributes.

→ following attributes are used with boxplot()-

① data - array or sequence of array to be plotted.
② notch - optional parameter accepts boolean values.
③ vert - optional parameter accepts boolean values false & true for horizontal and vertical plot respectively.
④ bootstrap - optional parameter accepts int specifies intervals around notched boxplots.
⑤ usermedians - optional parameter accepts array or sequence of array dimension compatible with data
⑥ positions - optional parameter accepts array & sets the position of boxes.
⑦ widths - optional parameter accepts array & sets the width of boxes
⑧ labels - sequence of strings sets label for each dataset.
⑨ order - optional parameter sets the order of the boxplot.

4) What is the difference between boxplot & histogram ?

→ Histograms are sometimes called frequency plots while boxplots are reffered to as Box-and-Whisker Plots.
A histogram is normally used for continuous data while a bar chart is a plot of count data.

5) How do you compare data in a BoxPlot?

→ step1 - compare the medians of box plots.

compare the respective medians of each box plot. If the median line of a box plot lies outside of the box of a comparision box plot then there is likely to be a difference between the two groups.

step2 - compare the interquartile ranges & whisker of box plots.

compare the interquartile ranges to examine how the data is dispersed between each sample.

step3 - Look for potential outliers.

When reviewing a box plot an outlier is defined as a data point that is located outside the whiskers of the box plot.

step4 - Look for signs of skewness

If the data do not appear does each sample show the same kind of asymmetry

6) Does a box plot show standard deviation&

→ No, box plot cannot show standar deviation.

7) What are outliers in boxplot&

→ An outlier is an observation that is numerically distant from the rest of the data.

When reviewing a box plot an outlier is defined as data point that is located outside the whiskers of the box plot.

8) How do you remove outliers from boxplot&

→ For removing the outlier, one must follow the same process of removing an entry from the dataset using its exact position in the dataset because in all above methods of detecting the outliers

end result is the list of all those data items that statisfy the outlier definition according to the method used

```
dataframe.drop(row_index, inplace=True)
```

Conclusion -

Hence, Plotted a box plot for distribution with age with respect to gender along with information whether they survived or not.