CE605A Reading Assignment

Name ***Shumet Getahun*** *Roll No.* ***22203262***

1. *The main motivation for developing the violin plot, despite the existence of a box plot?*

Answer

i. The shape of the violin plot is more descriptive than box plot to compare multiple groups and variables. It is easier to compare multiple variables using violin plot. It can provide detailed insights into the distribution and trends using peaks, valleys and bumps.

ii. Violin plot clearly shows the density and distribution of data along the axis in an easier way to describe and understand which cannot not shown by a box plot. Violin plot also shows the skewness of the distribution.

iii. Violin plot clearly shows the two peaks of bimodal distribution, whereas Comparing bimodal and uniform data distribution is difficult using a box plot since the shape of the distribution is difficult to verify peaks.

iv. In the Violin plot, mild and severe outliers are part of the shape, whereas in boxplot they are represented by individual dots separately. In violin plot circle represents median line which is easier for comparison compared to the boxplot line.

v. Violin plot provides an overall picture of data by including summarized statistics and density shape using box plot and density. (Hintze & Nelson, 1998)

2. *What are the essential ingredients for constructing a violin plot?*

Answer

a. **Sample Size (n)** number of observations.
b. **Box plot** consist of data that are being analyzed including median, lower values, upper values, first quartile, third quartile, spreads and outliers. The median line is represented by a circle and outliers are part of the shape in violin plot.
c. **Density trace (d)** defines the shape of the plot. It depends on the interval width and number of data.
d. **Interval width (h)**, defines the density of a given data for analysis. It should be between 10 to 40% of the range value.
e. **Grid** equally spaced x values.
f. **Opening the box of the box plot** to add information and make a comparison.
g. **Computer** to plot the violin plot using Statistical software.

(Hintze & Nelson, 1998)

3. *Is the violin plot a miniature histogram? Give reasons and examples for your answer*

Answer

No, the violin plot is not a miniature histogram

Violin plot is more inclusive than histogram. Violin plot used to show the comparison of multiple group data. Histogram is used to show data using bins or intervals. Violin plot combines both a smoothed histogram and boxplot.

Example

The following table shows the distribution of total bills and tips received per day. The table is partially showing "tips" data. Source (Bhatt, Dec 29, 2018)

Table 1 Distribution of total bill and tips

total bill	Tip	Sex	smoker	Day	Time	Size
16.99	1.01	Female	No	Sun	Dinner	2
10.34	1.66	Male	No	Sun	Dinner	3
21.01	3.5	Male	No	Sun	Dinner	3
23.68	3.31	Male	No	Sun	Dinner	2
24.59	3.61	Female	No	Sun	Dinner	4

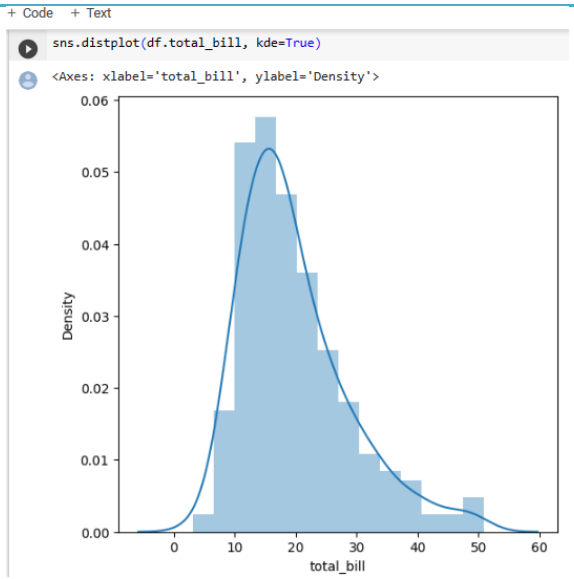

Figure 1: Histogram

The histogram in Fig. 1 shows the total bill distribution and density of the given data.

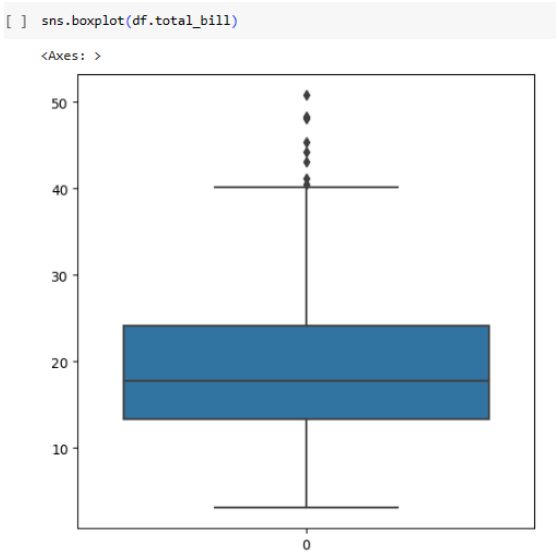

Figure 2: Box Plot

The box plot in Fig.2 shows the total bill distribution and density of the given data.

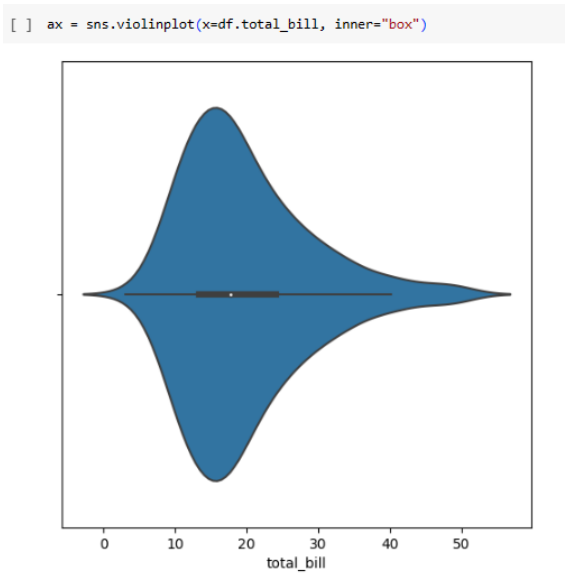

Figure 3: Violin Plot

This violin plot in Fig.3 shows the total bill distribution of the given data. This plot shows the comprehensive effect of both the histogram and box plot. The bump part shows the distribution of data (shown in the histogram) and the internal values (asymmetry, spreads, median and outliers) are part of the box plot.

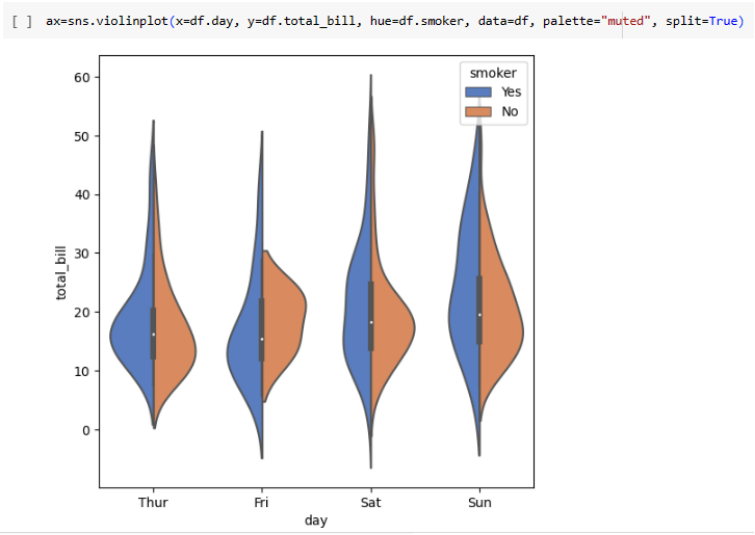

Figure 4: Violin Plot for Comparing Different Days Total Bill Distribution

As shown in Fig.4, a violin plot is used to compare multiple groups of data easily and clearly. As shown in Fig.4, the total bill distribution for each day and the distribution of smokers and nonsmokers people per day are shown by different violin plots. Therefore, violin plot is more descriptive than histogram.

4. *What are the scenarios in which a violin plot fails? Correspondingly, when should the violin plot be avoided?*

Answer

Violin plot fails for too small data sets since it is not enough to plot data and is difficult to define shape.

Violin plot should not be used in the following cases;

a. In the case of small data sets, the interval value is small as a result, it will create difficulty in understanding the shape of density and difficult to conclude. As a general rule, data observations should be at least 30.
b. Too large an interval width (h) is not comfortable to draw violin plot and causes the density race over smoothness. It should be 10% to 40% of the range of data.
c. Violin plot regulates the density traces maximum heights of bimodal and normal distributions are equal despite the sample size of both distributions being different. This will lead to a misunderstanding of the sample size of both distributions and difficult to compare.
(Hintze & Nelson, 1998)

References

Bhatt, B. (Dec 29, 2018). *Simplest Interpretation on Violin Plots - YouTube*. https://www.youtube.com/@bhattbhavesh91

Hintze, J. L., & Nelson, R. D. (1998). Violin plots: a box plot-density trace synergism. *The American Statistician*, *52*(2), 181-184.