Name: Samriddhi Verma
Reg. No.: 16BCE1375
Slot: L49+L50

WEEK-2

EXERCISE: PRACTICE GGPLOT2 BASICS.

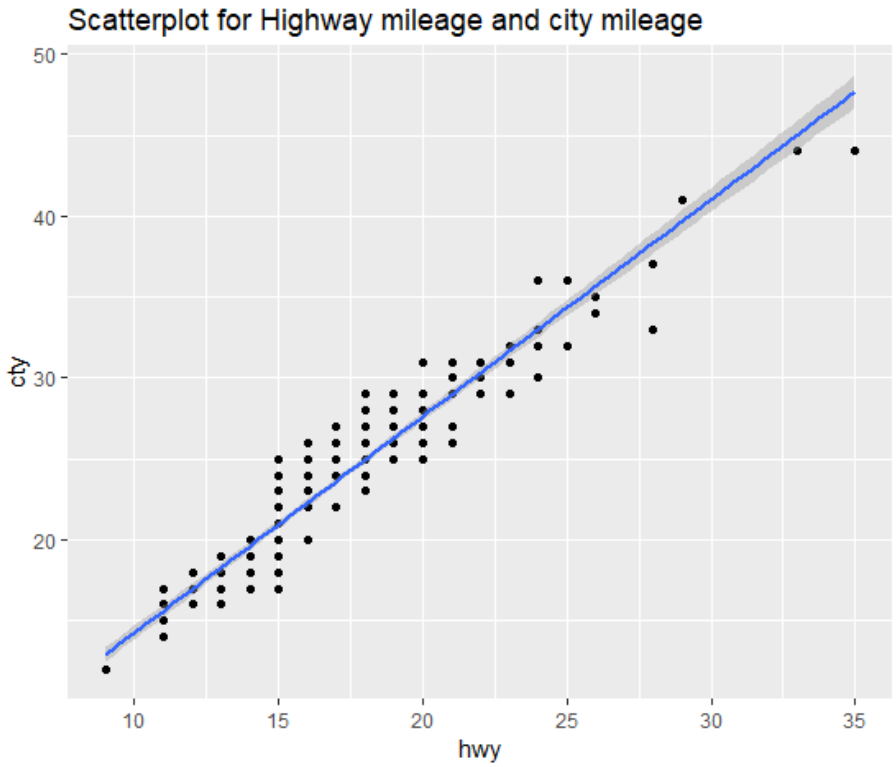
CODE:

```
library(ggplot2)    #imported the ggplot2 package/library

data(mpg, package = "ggplot2")      #loaded the dataset mpg
```

p = ggplot(mpg, aes(cty, hwy)) #initialised p to hold the data to be visualised using ggplot. Aes is the aesthetic mappings describe how variables in the data are mapped to visual properties (aesthetics) of geoms. Here we find the relation between the city and highway mileage of the dataset.

1. Scatter plot:

```
p+ geom_point() + geom_smooth(method = "lm")+
 labs(x="hwy", y="cty", title="Scatterplot for Highway mileage and city mileage")
```



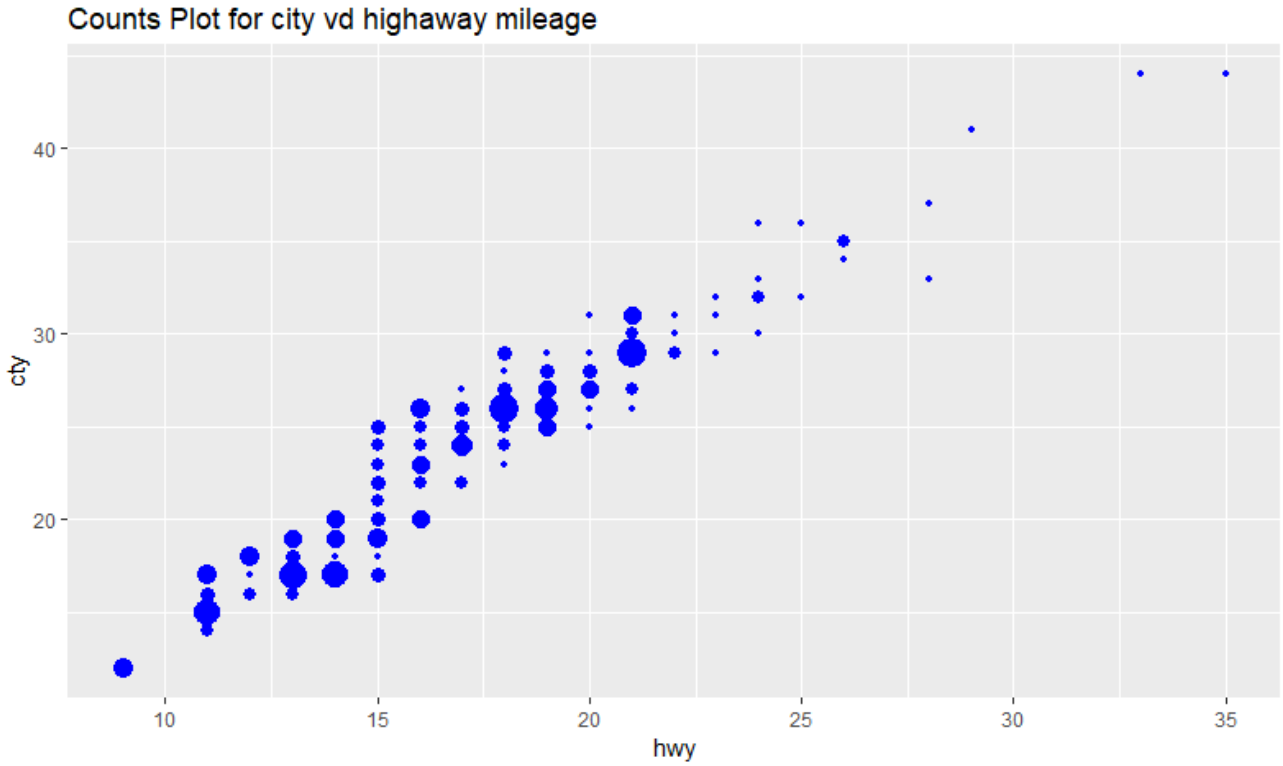
This type of plot is used to find the correlation between the features or attributes of the dataset. Here we can see a linear relation between the two. The original data has 234 points but we see that there are fewer points displayed in this chart. To overcome this , we use jittered plots to see less overlapping points.

Name: Samriddhi Verma
Reg. No.: 16BCE1375
Slot: L49+L50

2. Counts chart:

p + geom_count(col="blue", show.legend = F) + labs(x="hwy", y="cty", title="Counts Plot for city vs highway mileage")

➔ The labs() function is used to label the plot ,i.e., set the title and subtitle of the chart/visual.

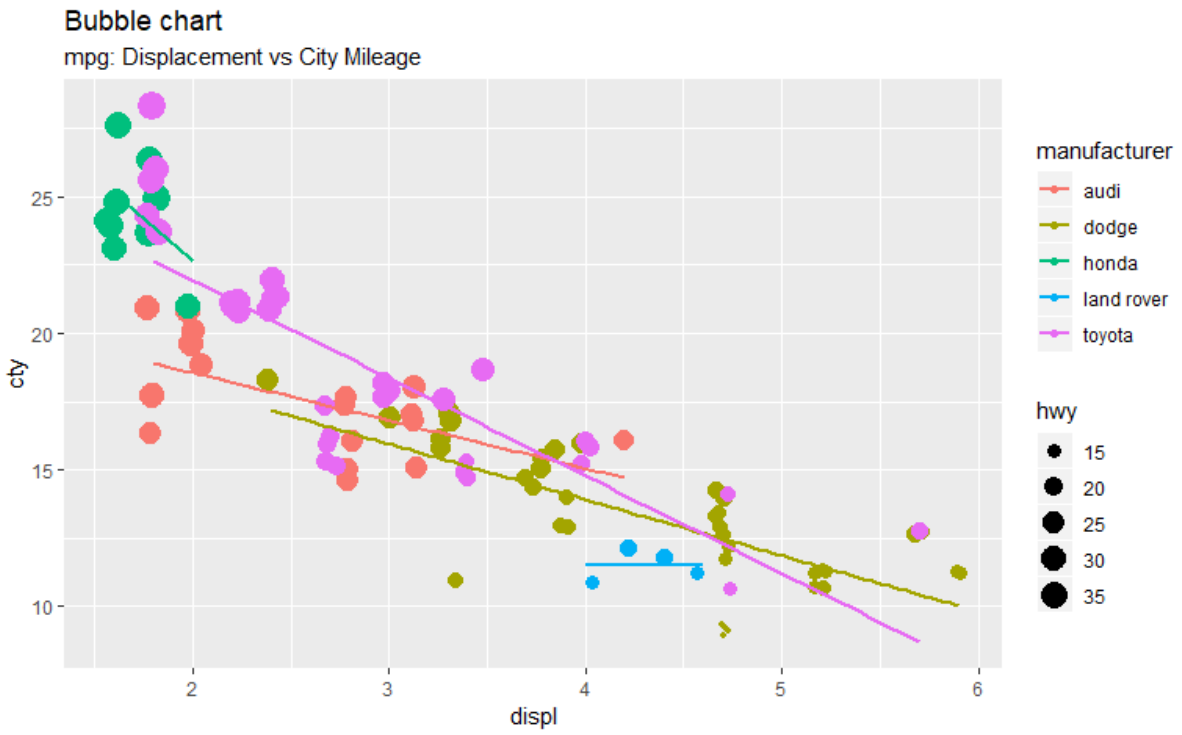


To overcome the problem of overlapping points, we use counts chart. The more the overlapping, bigger is the size of the circle.

Name: Samriddhi Verma
Reg. No.: 16BCE1375
Slot: L49+L50

3. Bubble plot:

```
select = mpg[mpg$manufacturer %in% c("audi", "dodge", "honda", "land rover" , "toyota"), ]

# Intialised select to hold  the data of manufactures named “audi”, “honda”…

p = ggplot(select, aes(displ, cty)) +

 labs(subtitle="mpg: Displacement vs City Mileage",

     title="Bubble chart")

# p holds the mapping of the features of displacement of engines in litres vs city mileage being held in the variable select which is sliced from main dataset containing the details of the manufacturers.

p + geom_jitter(aes(col=manufacturer, size=hwy)) +

 geom_smooth(aes(col=manufacturer), method="lm", se=F)
```



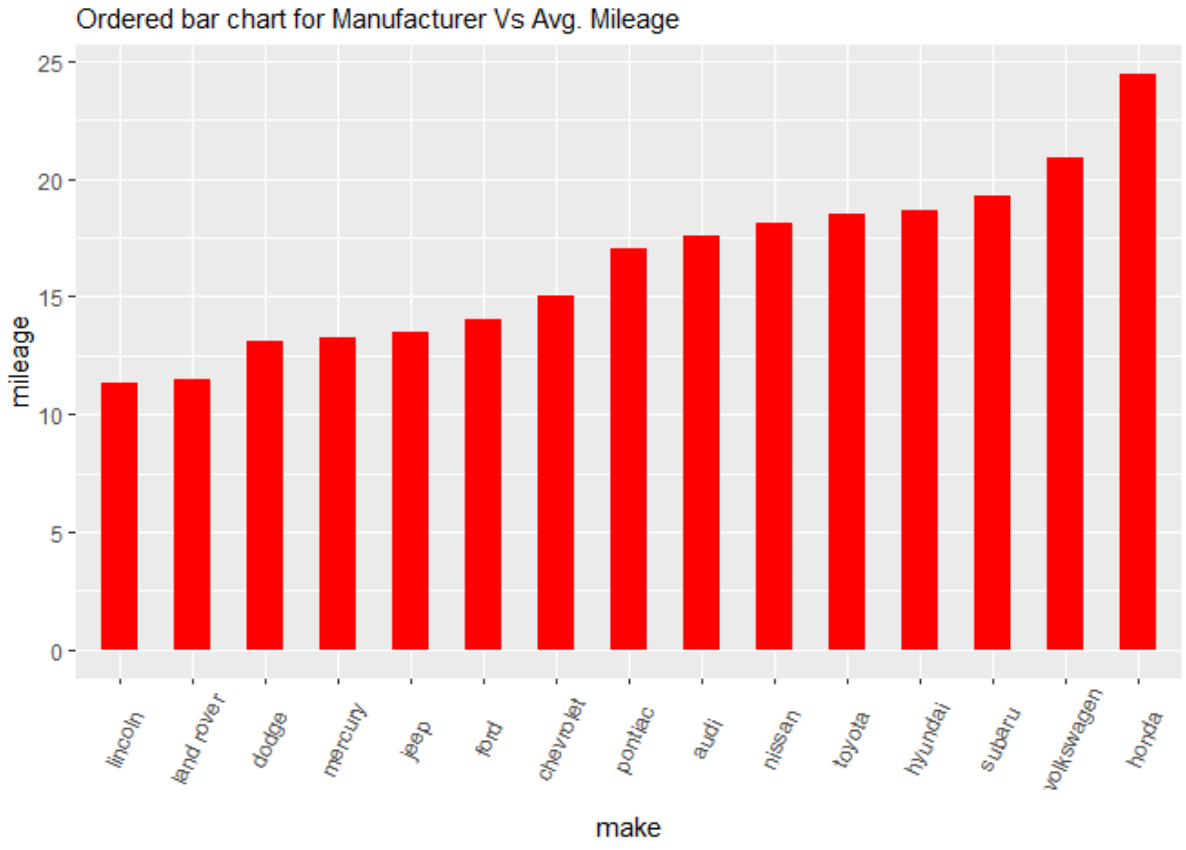
Using the bubble chart, we can distinguish the range of hwy between manufacturers and how the slope of the best fit varies. This plot helps to visualise the data which contains numerical values (hwy here) and the categorical ones too using color (manufacturer here).

4. Ordered Bar chart:

```
cty_mpg <- aggregate(mpg$cty, by=list(mpg$manufacturer), FUN=mean)

# We create a variable which stores the group mean city mileage by manufacturer. Aggregate() function serves the role to find mean.

colnames(cty_mpg) <- c("make", "mileage")  # changes column names of manufacturer&city mileage
```

Name: Samriddhi Verma
Reg. No.: 16BCE1375
Slot: L49+L50

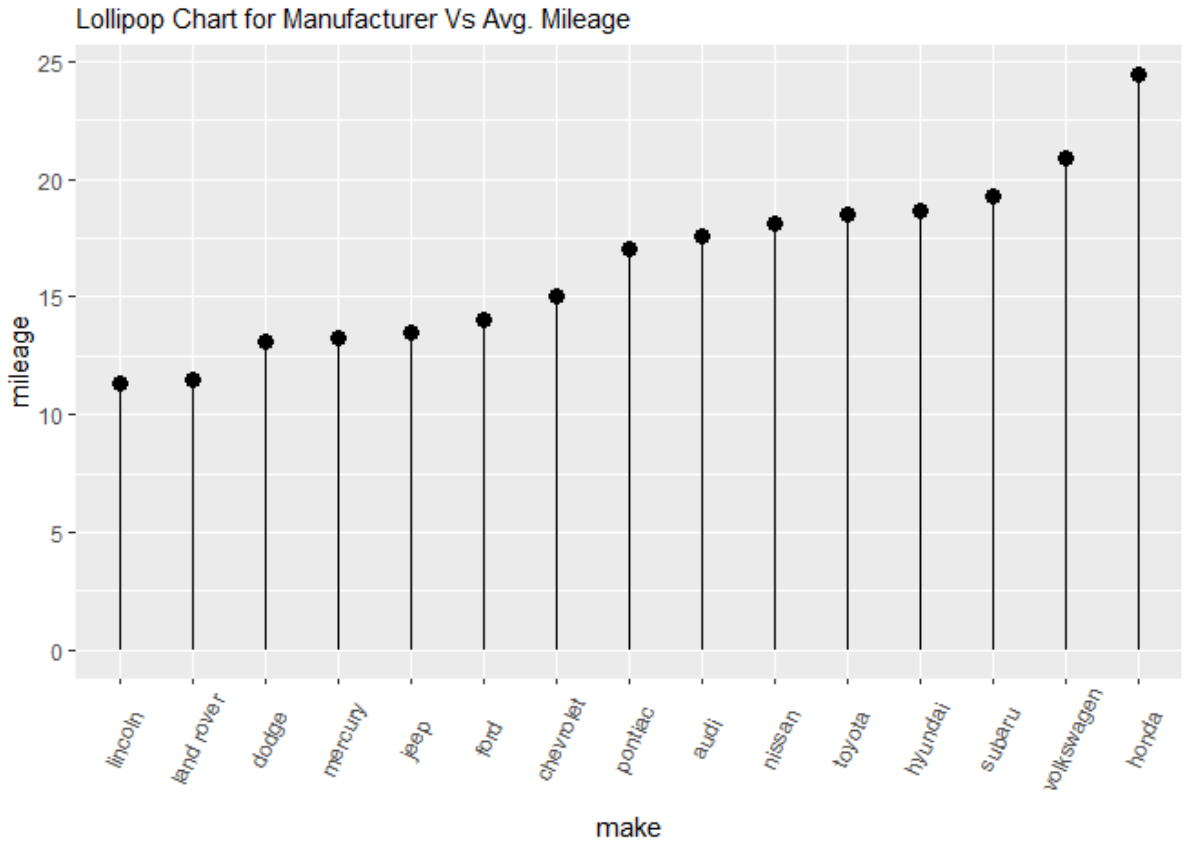
```
cty_mpg <- cty_mpg[order(cty_mpg$mileage), ]
# sort the data according to the mileage which is the city mileage we assigned to before.It is stored in cty_mpg variable.
cty_mpg$make <- factor(cty_mpg$make, levels = cty_mpg$make)  # to retain the order in plot.
head(cty_mpg, 4) # display the top 4 rows of the cty_mpg column


ggplot(cty_mpg, aes(x=make, y=mileage)) +
 geom_bar(stat="identity", width=.5, fill="red") +
 labs(subtitle="Ordered bar chart for Manufacturer Vs Avg. Mileage") +
 theme(axis.text.x = element_text(angle=65, vjust=0.6))
```



Name: Samriddhi Verma
Reg. No.: 16BCE1375
Slot: L49+L50

5. **Lolipop Chart:**

```
ggplot(cty_mpg, aes(x=make, y=mileage)) +  geom_point(size=3) + geom_segment(aes(x=make, xend=make,  y=0, yend=mileage)) +   labs(subtitle="Lollipop Chart for Manufacturer Vs Avg. Mileage") + theme(axis.text.x = element_text(angle=65, vjust=0.6))
```

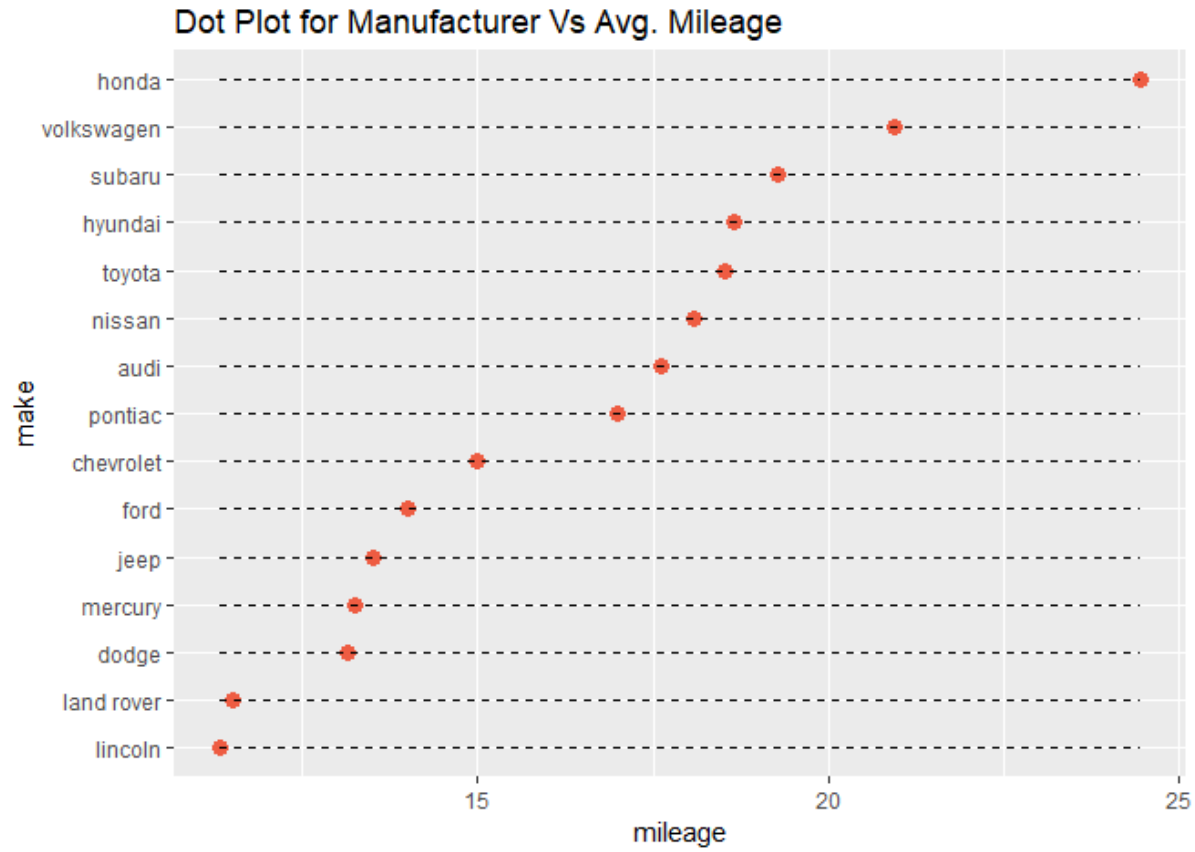


Lollipop charts conveys the same information as in bar charts. By reducing the thick bars into thin lines, it reduces the clutter and lays more emphasis on the value.

Name: Samriddhi Verma
Reg. No.: 16BCE1375
Slot: L49+L50

6. DOT PLOT:

```
library(scales)

ggplot(cty_mpg, aes(x=make, y=mileage)) +

 geom_point(col="tomato2", size=3) +   # Draw points

 geom_segment(aes(x=make,

                  xend=make,

                  y=min(mileage),

                  yend=max(mileage)),

              linetype="dashed",

              size=0.1) +   # Draw dashed lines

 labs(title="Dot Plot for Manufacturer Vs Avg. Mileage") +

 coord_flip()
```

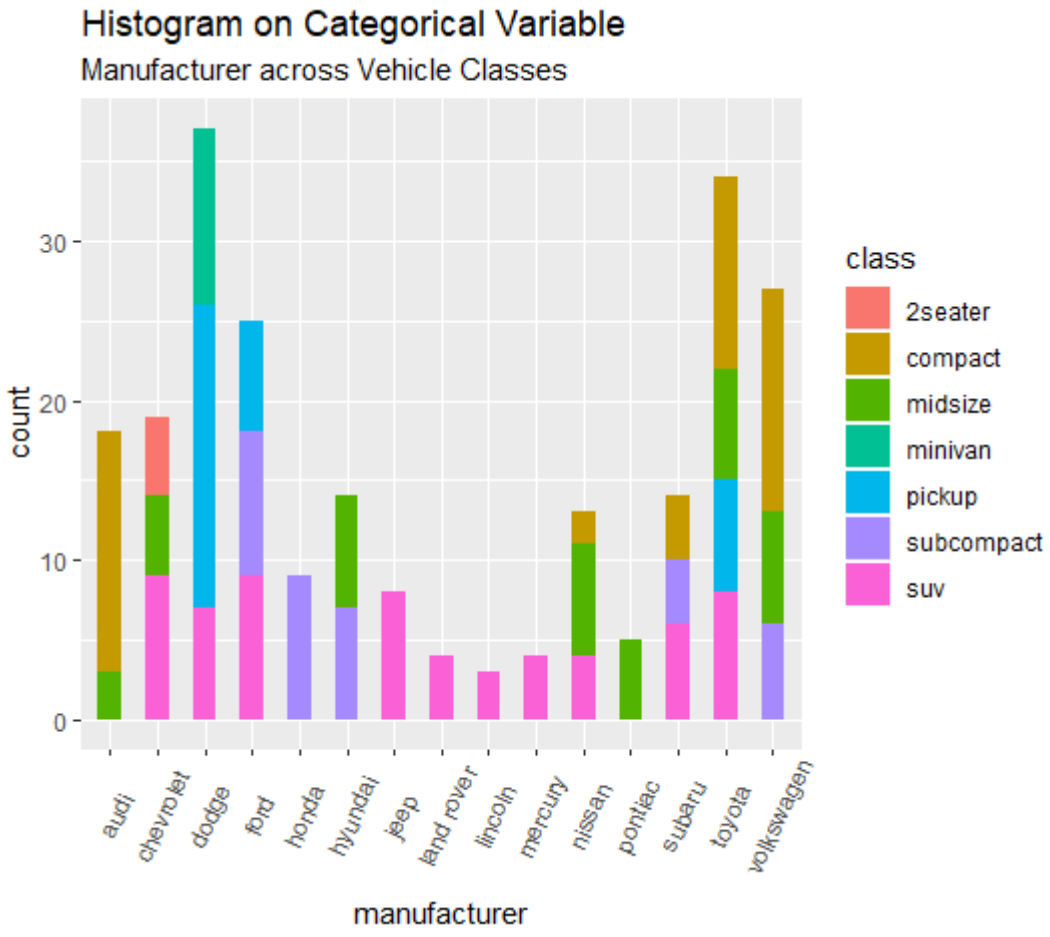


Dot plots are very similar to lollipops, but without the line and is flipped to horizontal position. It emphasizes more on the rank ordering of items with respect to actual values and how far apart are the entities with respect to each other. Here we see the average mileage among the manufacturers.

Name: Samriddhi Verma
Reg. No.: 16BCE1375
Slot: L49+L50

7. HISTOGRAM:

```
g <- ggplot(mpg, aes(manufacturer))

g + geom_bar(aes(fill=class), width = 0.5) +

 theme(axis.text.x = element_text(angle=65, vjust=0.6)) +

 labs(title="Histogram on Categorical Variable",

     subtitle="Manufacturer across Vehicle Classes")
```

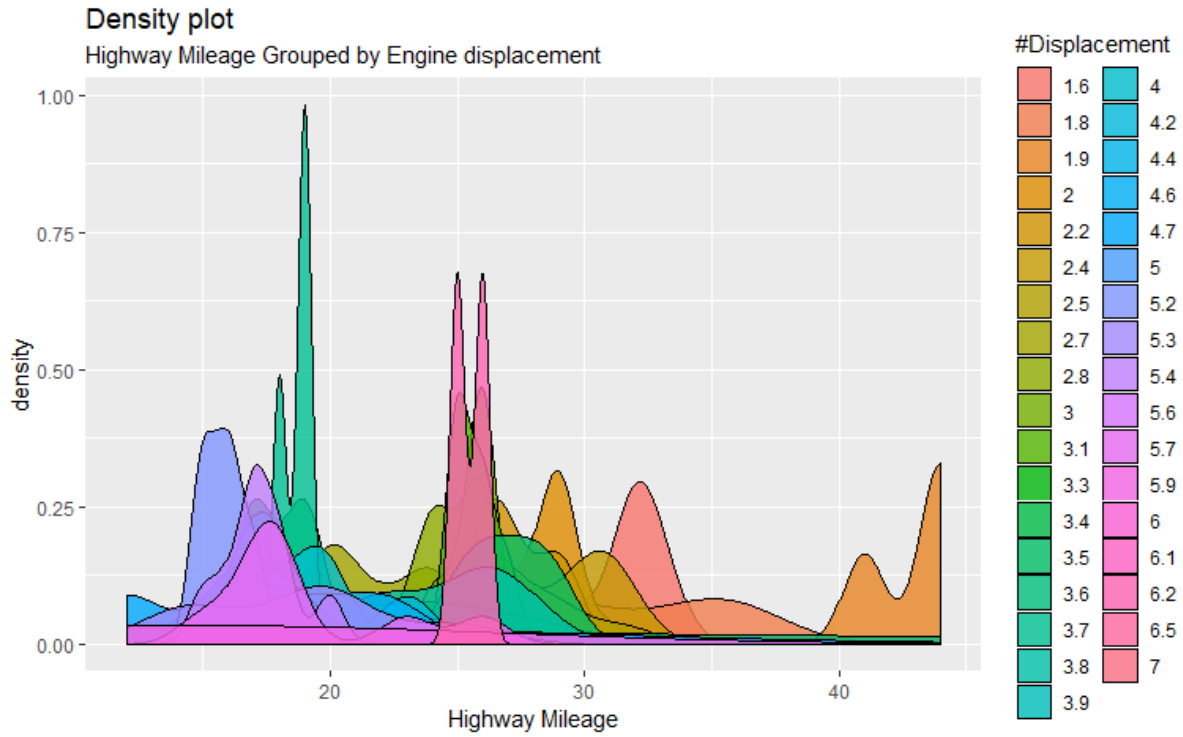


To understand the distribution of the data, we use histograms. Here we have a categorical variable, i.e., manufacturer. So, histogram on a categorical variable would result in a frequency chart showing bars for each category.

Name: Samriddhi Verma
Reg. No.: 16BCE1375
Slot: L49+L50

8. DENSITY PLOT:

```
g <- ggplot(mpg, aes(hwy))

g + geom_density(aes(fill=factor(displ)), alpha=0.8) +

 labs(title="Density plot",

     subtitle="Highway Mileage Grouped by Engine displacement",

      x="Highway Mileage",

     fill="#Displacement")
```

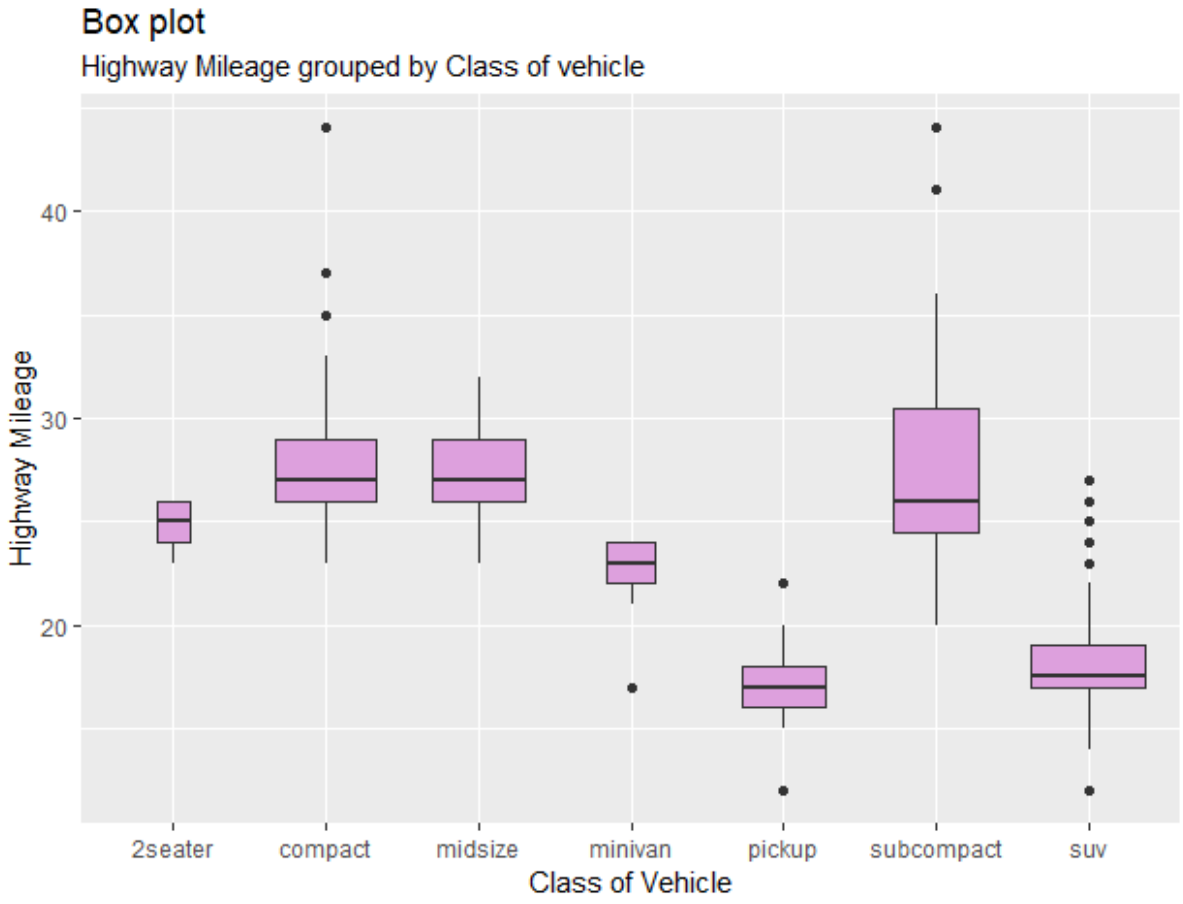


Here the density plot visualises the distribution of data over a continuous interval or time period. We can see the distribution of highway mileage grouped by engine displacement. The distribution shape helps us understand the concentration of plot values.

Name: Samriddhi Verma
Reg. No.: 16BCE1375
Slot: L49+L50

9. BOX PLOT:

```
g <- ggplot(mpg, aes(class, hwy))

g + geom_boxplot(varwidth=T, fill="plum") +

 labs(title="Box plot",

     subtitle="Highway Mileage grouped by Class of vehicle",

     x="Class of Vehicle",

     y="Highway Mileage")
```



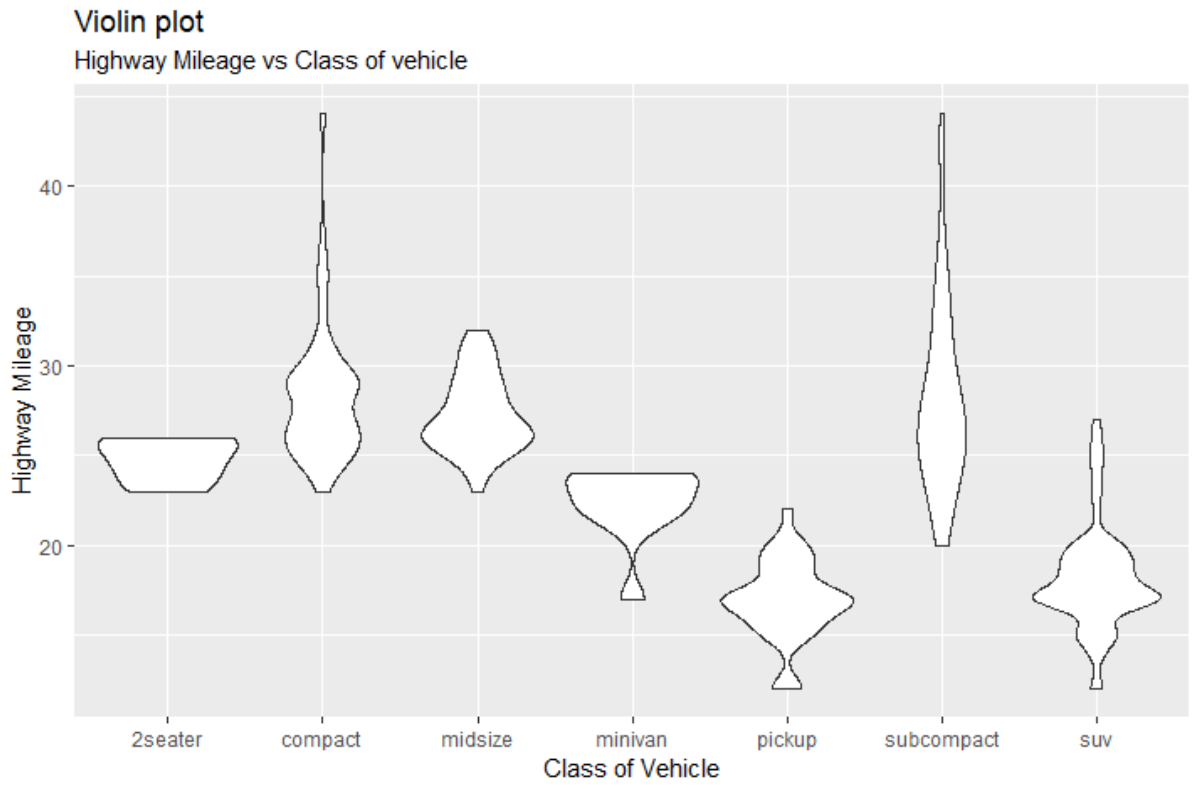
Box plot is an excellent tool to study the distribution. It can also show the distributions within multiple groups, along with the median, range and outliers if any.

The dark line inside the box represents the median. The top of box is 75%ile and bottom of box is 25%ile. The end points of the lines (aka whiskers) is at a distance of 1.5*IQR, where IQR or Inter Quartile Range is the distance between 25th and 75th percentiles. The points outside the whiskers are marked as dots and are normally considered as extreme points.

Name: Samriddhi Verma
Reg. No.: 16BCE1375
Slot: L49+L50

10. VIOLIN PLOT:

```
g <- ggplot(mpg, aes(class, hwy))

g + geom_violin() +

 labs(title="Violin plot",

     subtitle="Highway Mileage vs Class of vehicle",

     x="Class of Vehicle",

     y="Highway Mileage")
```



A violin plot is similar to box plot but shows the density within groups. Here we can see the density of class vehicles distributed according to highway mileage.

Name: Samriddhi Verma
Reg. No.: 16BCE1375
Slot: L49+L50