ggplot2

Gowthamkumar

12 October 2017

How to create a Scatterplot in ggplot2

Visualization is a key in data science to understand how the data is distributed, how the data points are related to each other and so on. **R** has a package called **ggplot2**, it is a great tool for visualization.

Let's see how can we create visualization's using ggplot package

Step One: Load the ggplot2 package

Load the ggplot2 package inside our working environment by executing the following command in the R console window.

```r
#Load the ggplot2 package
library('ggplot2')
```

If ggplot2 is not installed in the R studio, First install the ggplot2 pacakge by using the following command in the R console window.

```r
# Install the ggplot2 package
install.packages('ggplot2')
```

Note: Then use *library('ggplot2')* to load the package in to the working environment.

Step Two: Loading the Data

Here, In this tutorial we are going to use **mtcars** dataset that is comes with R studio.

To access the dataset execute the following command.

```r
# Loading the dataset
data("mtcars")
```

After loading the dataset, then examine the **mtcars** dataset by executing **head()**, **summary()** and **str()** methods.

- *head()* will give the first 6 rows of the dataset
- *str()* will give the structure of the dataset
- *summary()* will give the summary statistics of the dataset

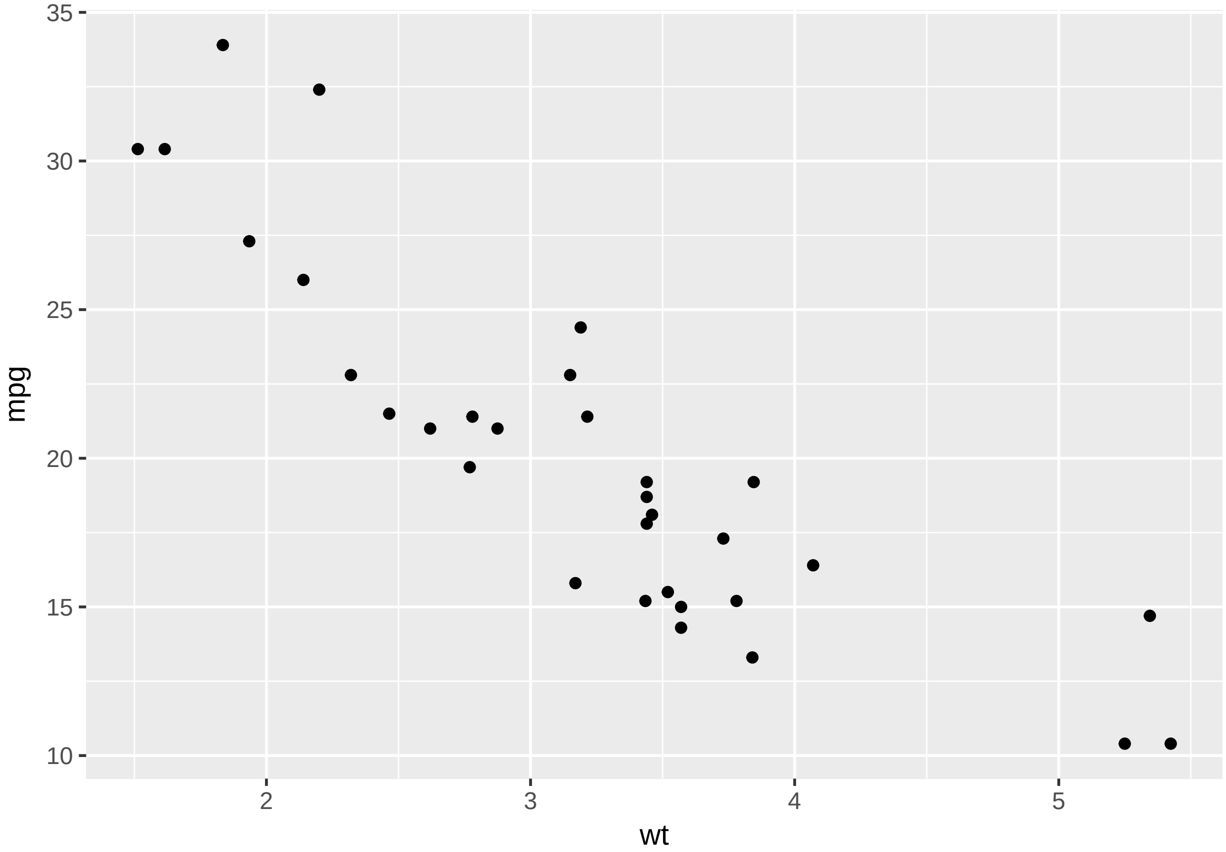
For example, if we execute the *summary()* method we get the following result.

```r
# summary for the mtcars dataset
summary(mtcars)
```

```
##       mpg             cyl             disp             hp       
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0  
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7  
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
##       drat             wt             qsec             vs        
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000  
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000  
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000  
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375  
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000  
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000  
##        am              gear            carb      
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000  
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000  
##  Median :0.0000   Median :4.000   Median :2.000  
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812  
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000  
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000  
```

Step Three: Create a basic scatterplot using ggplot2

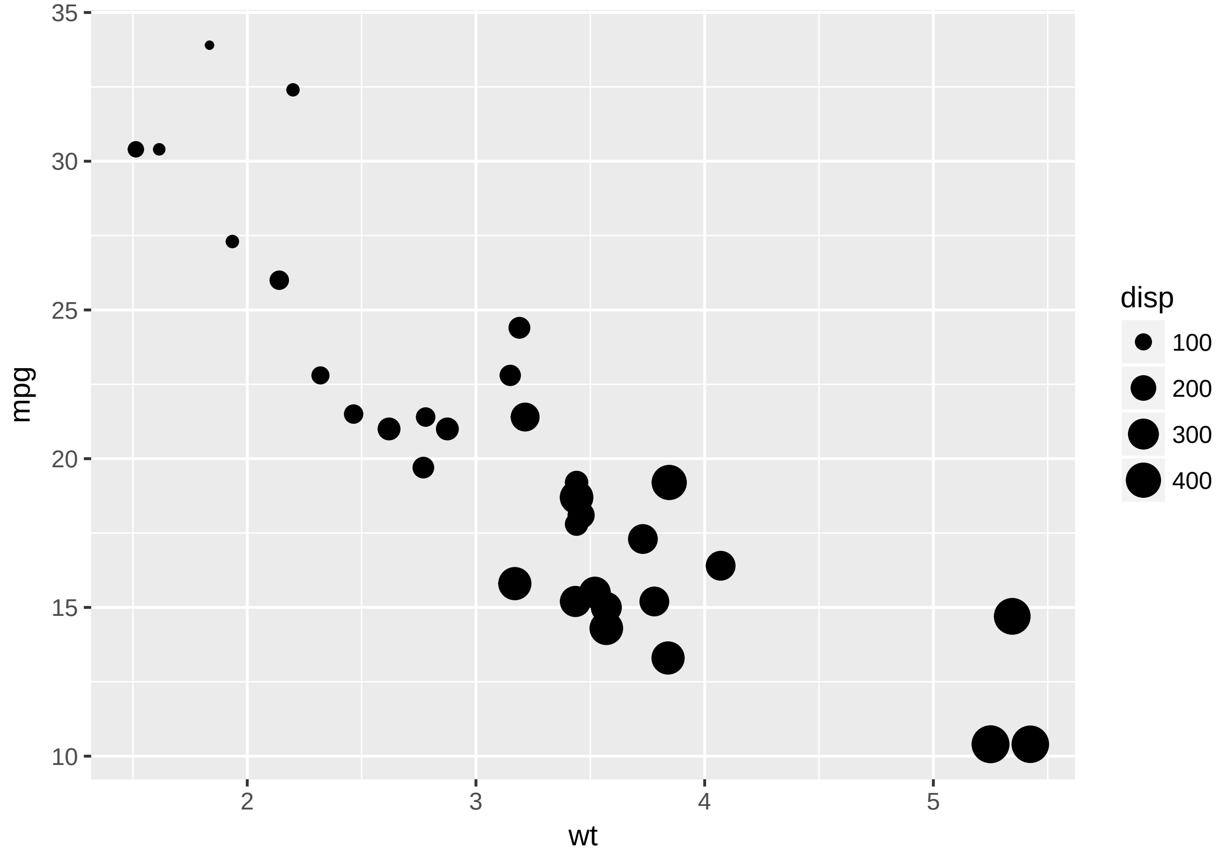
```r
# A scatter plot has been made  for mpg (miles per galon) against the weight
# (in thousands of pounds)
ggplot(mtcars, aes(x = wt, y = mpg)) +
  geom_point()
```



Look the above scatter plot, we have create the plot for mpg (miles per galon) against the weight (in thousands of pounds). From this we can easily find out how the weight of the car infulence the mpg (miles per galon). We can easily spot the negative trend, like if the weight increases the miles per galon is decreases. So you can tell the car with less weight can go more miles per galon.

Step Four: Make some aesthetic changes:

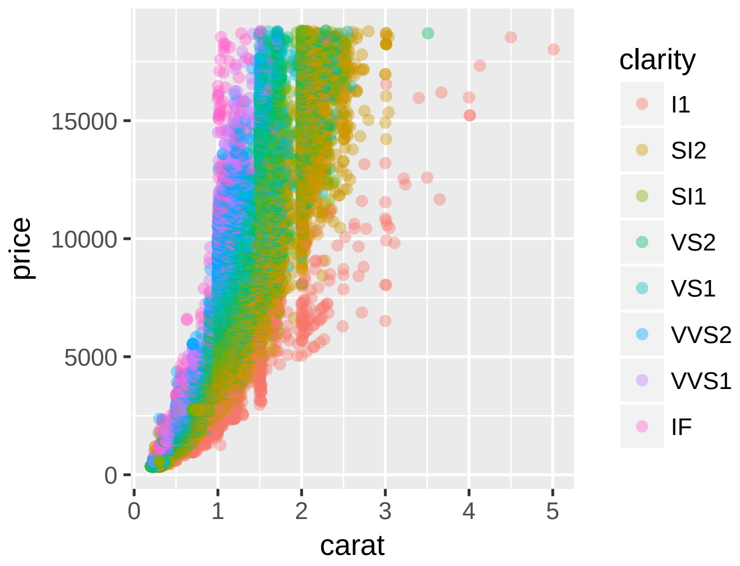
```
# Add the disp in the size attribute
ggplot(mtcars, aes(x = wt, y = mpg, size = disp)) +
  geom_point()
```



Here you can see, the *disp* value is set the *size* attibute and you can find how the scatter plot is used to differentiate the *disp* values using *size*.

Let's, see another dataset called **diamonds**, how the *color*, *alpha* attributes are used to make your scatterplot in more meaningful, and easy to differentiate the values in the dataset.

```
# Diamond Dataset wih color and alpha
ggplot(diamonds, aes(x = carat, y = price, color = clarity)) +
  geom_point(alpha = 0.4)
```

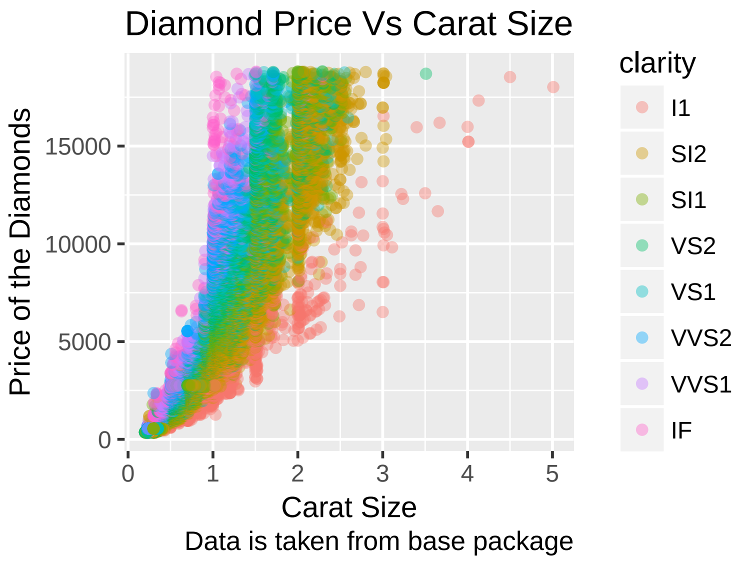


Adding Title, Labels, Captions

The chart without a title and labels are difficult to interpert and we don't know for what purpose the chart has been created, So we have to create the Title and Label's for the chart.

- *labs()* is used to add the Title, Subtitle, Labels, Captions.

```
# Diamond Dataset wih Title and label
ggplot(diamonds, aes(x = carat, y = price, color = clarity)) +
  geom_point(alpha = 0.4) + labs(x = "Carat Size", y = "Price of the Diamonds",
  title = "Diamond Price Vs Carat Size", caption = "Data is taken from base package")
```



Conclusion

There are lof of options to create a scatter plot in **R**, you can try out those options and find out how those options are useful to findout the insights of the dataset. First you have decide how to deliver your data in a useful manner using scatterplot in R and findout the insight of the data. You have to train well using the ggplot2 package inorder to findout which set options are more suitable for present your data.

For those people, who really want to start the data visualization this above scatter plot example is a starting point for your visualization using ggplot2 package.

Tips: Click the link for the complete documentation of ggplot2 package: **ggplot2 Documentation**