EDA ASSignment-1

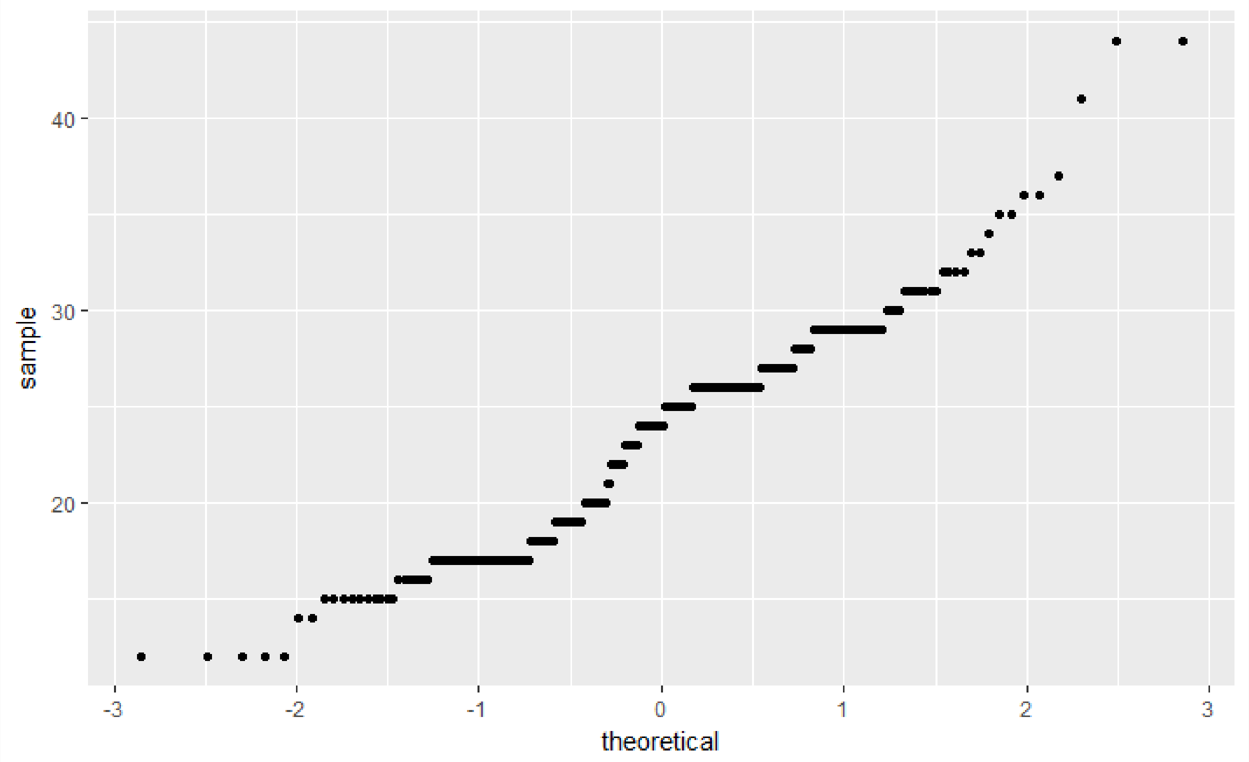
vinoth

August 31, 2017

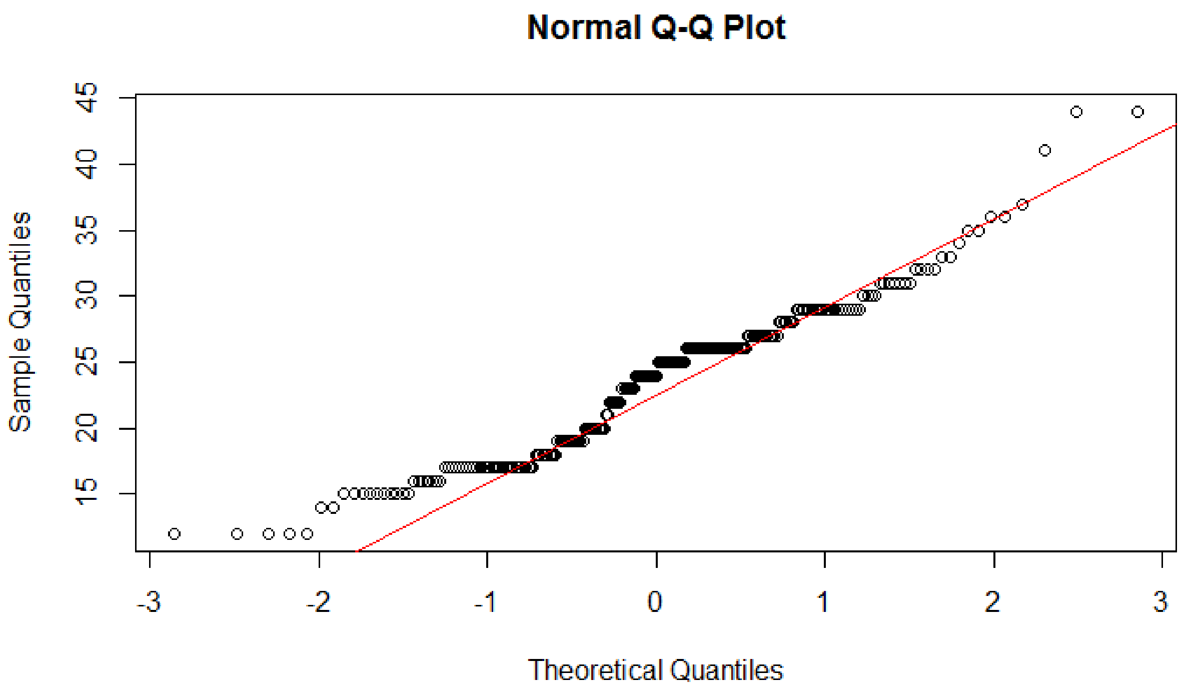
Q1. Does highway miles per gallon follow a normal distribution? If not, how does the data differ from a normal distribution ?

Here is the first plot which whether the given hwy is normal distribution

```
#ggplot(mpg,aes(sample=hwy))+stat_qq()
```



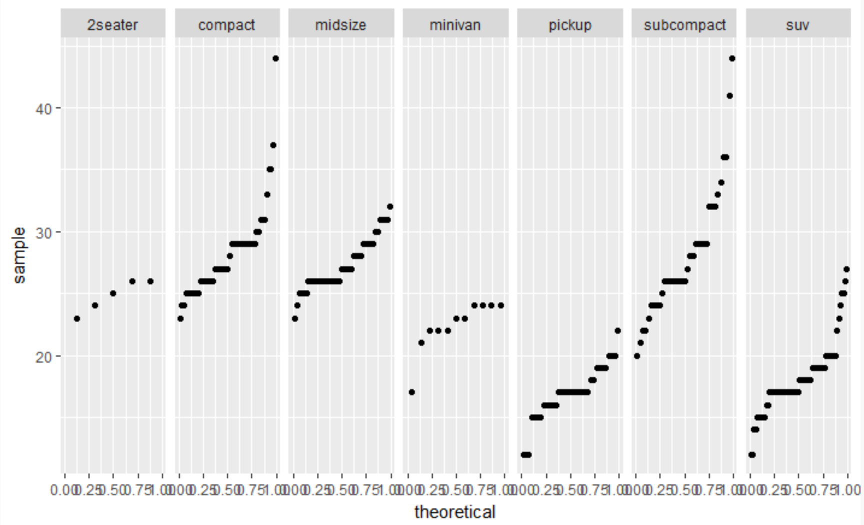
```
#x1=mpg$hwy
#qqnorm(x1)
#qqline(x1, col = "red")
```



By seeing the above given graph the data "hwy" does not follow the normal distribution because curve which formed by given data does notlook like a straight line.

Q2. How does the distribution of "hwy" change with "class"?

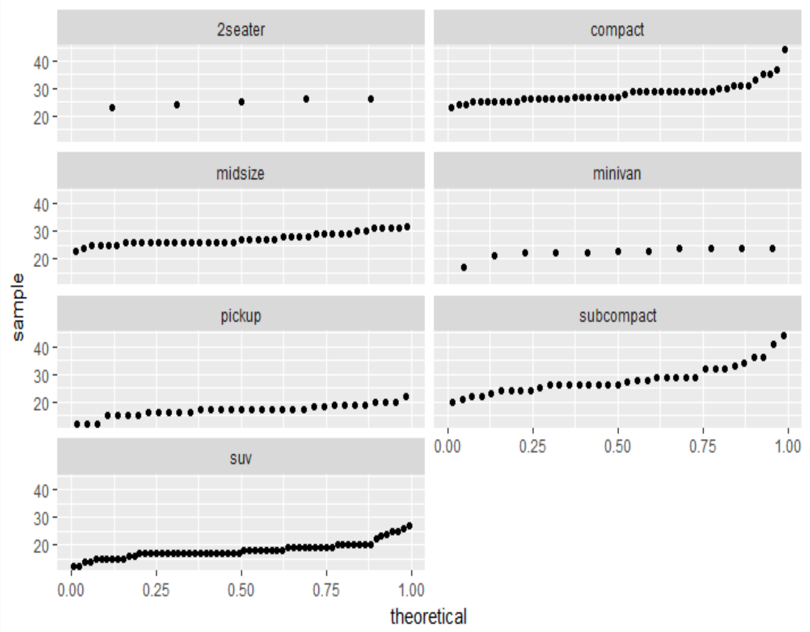
```
#ggplot(mpg,aes(sample=hwy))+stat_qq(distribution = qunif)+facet_grid(~class)
```



For more clarification I am using other step

here we have seven different class of car models the graph is cramped.the picture will be more effiecent when used 4 by 2 graphs.

```
#ggplot(mpg,aes(sample=hwy))+stat_qq(distribution=qunif)+facet_wrap(~class,nc
ol=2)
```

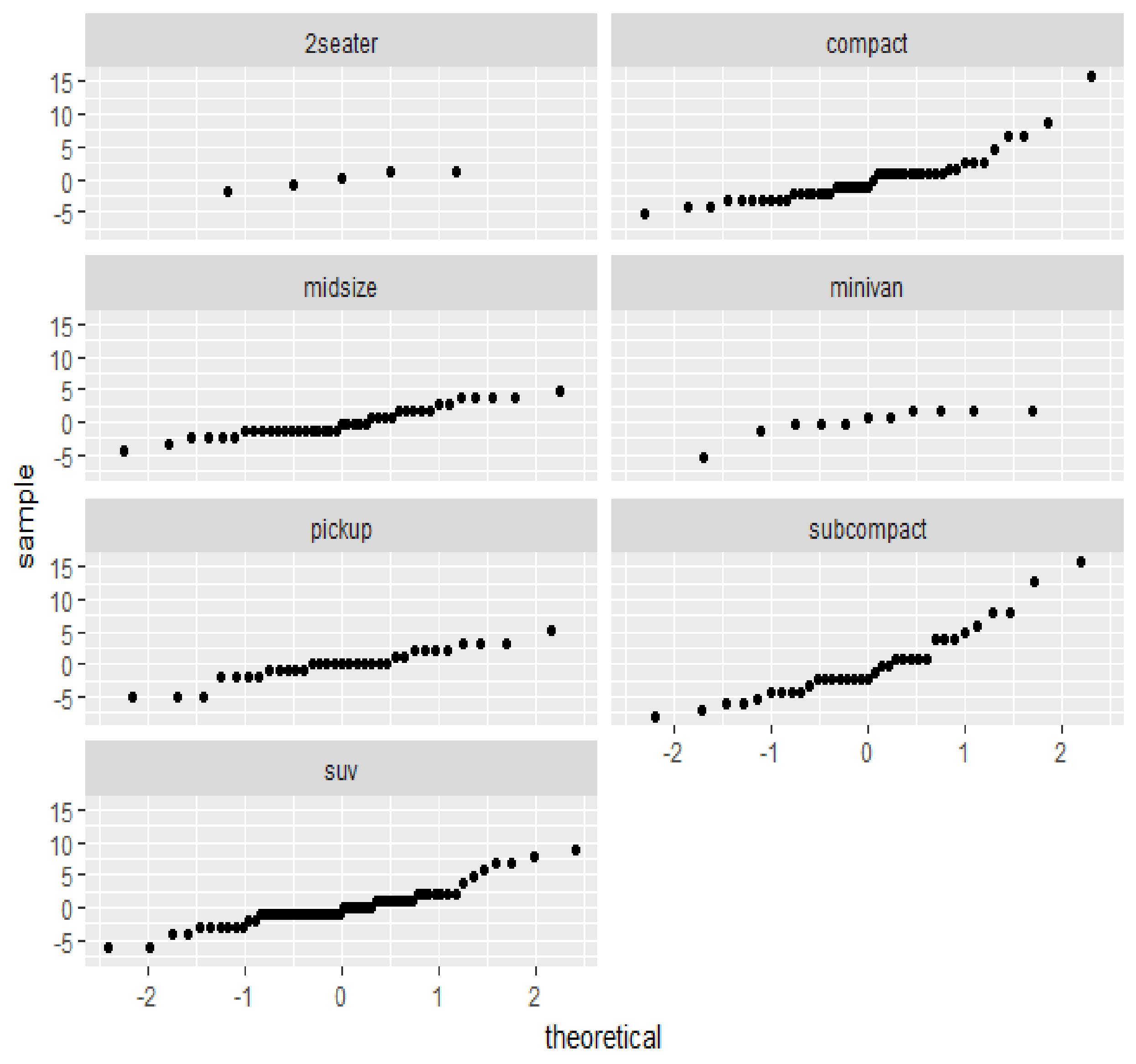


In this two sample distribution its clear that midsize class only form normal distribution.other class car type does not form normal distribution.

problem 3 Using lm(), fit a simple linear model for highway miles per gallon:

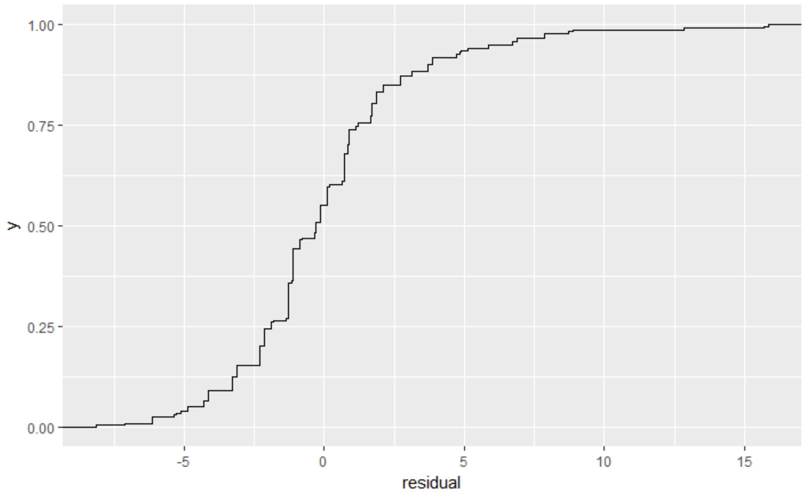
To check the given data Does it look like we can pool the residuals

```
mpg.lm=lm(hwy~class,data = mpg)
mpg.res=data. Frame(class=mpg$class,residual=residuals(mpg.lm))
ggplot(mpg.res,aes(sample=residual))+stat_qq()+facet_wrap(~class,ncol=2)
```



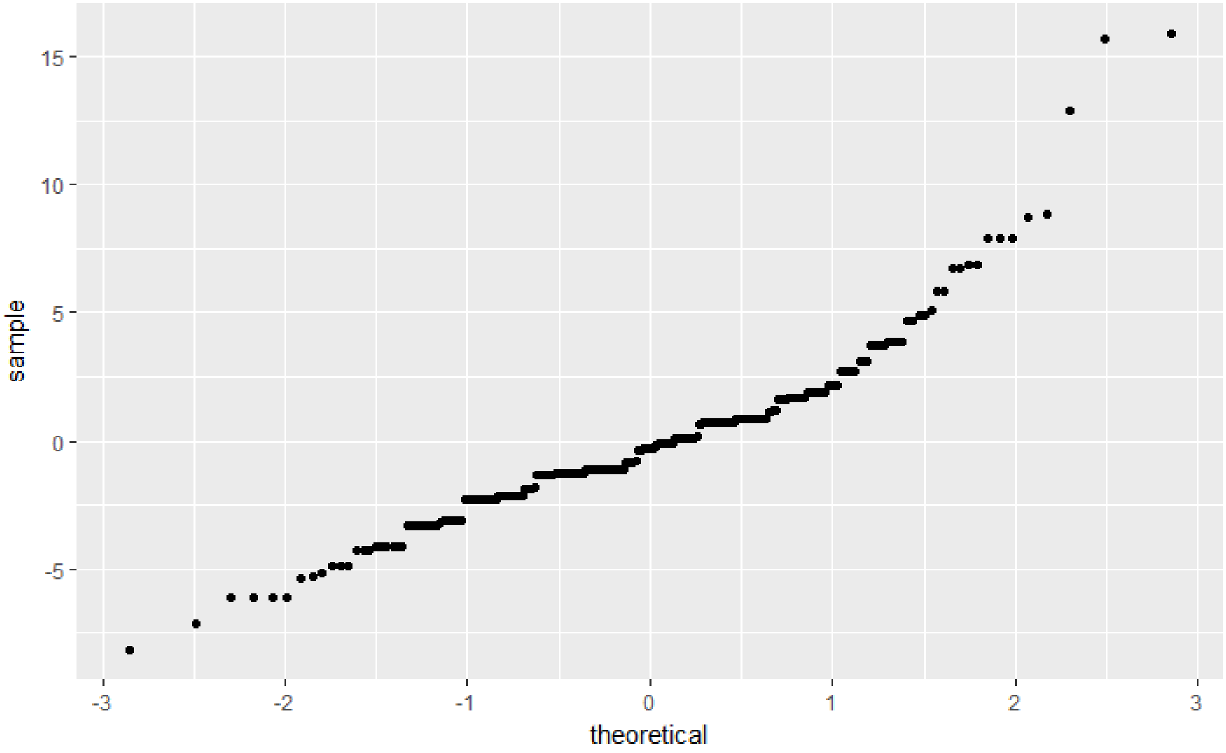
The above given data does not look a reasonably straight and the scales does not look similar for all seven class. To check more normality distribution

```
ggplot(mpg.res,aes(x=residual))+stat_ecdf()
```



Other step check with normality by using following commend

```
ggplot(mpg.res,aes(sample=residual))+stat_qq()
```

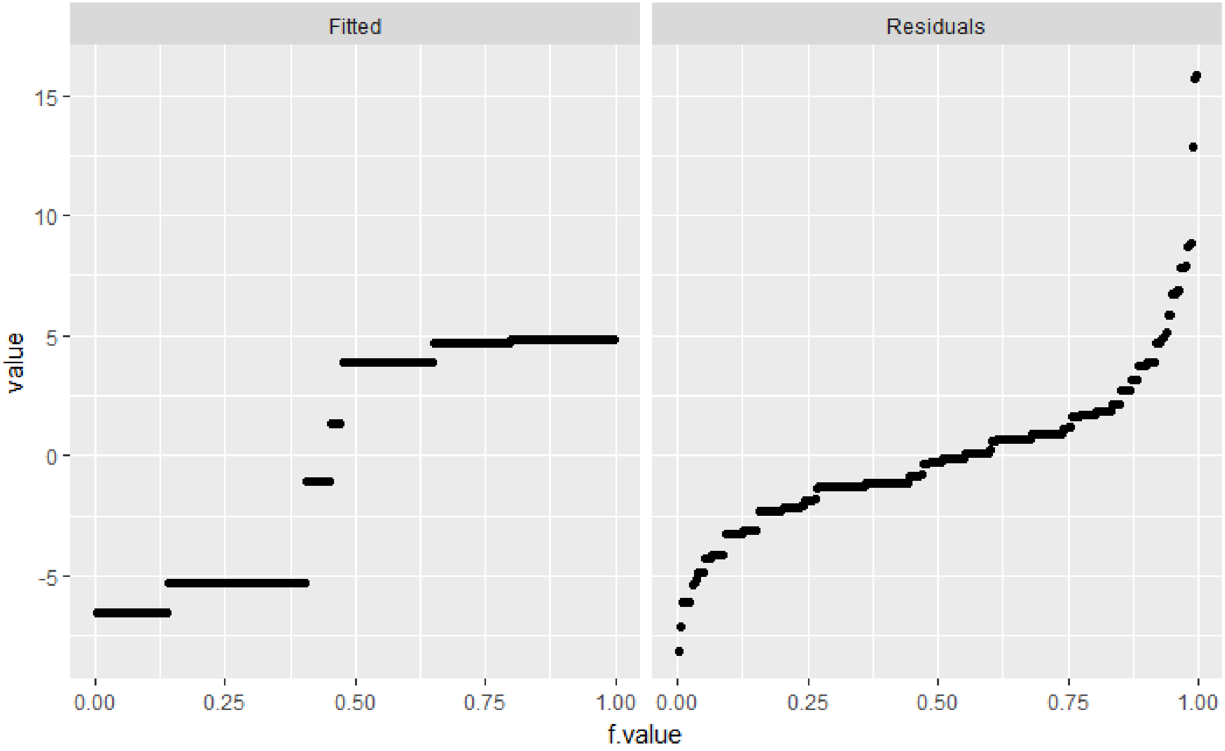


If we look at the above graph its clear that its does not form a straight line so we cannot pool the residuals.

3.b

```
mpg.fitted=sort(fitted.values(mpg.lm))-mean(fitted.values(mpg.lm))
mpg.residuals=sort(residuals(mpg.lm))

n=length(mpg.residuals)
f.value=(0.5:(n-0.5))/n
mpg.fit=data.frame(f.value,Fitted=mpg.fitted,Residuals=mpg.residuals)
mpg.fit.long = mpg.fit %>% gather(type, value, Fitted:Residuals)
ggplot(mpg.fit.long, aes(x = f.value, y = value)) + geom_point() + facet_wrap(~type)
```



If we see above given its clear that fitted values not closer to residuals. So its clear that which residual values are greater than the fitted values.