Mid Sem: Data Visualization

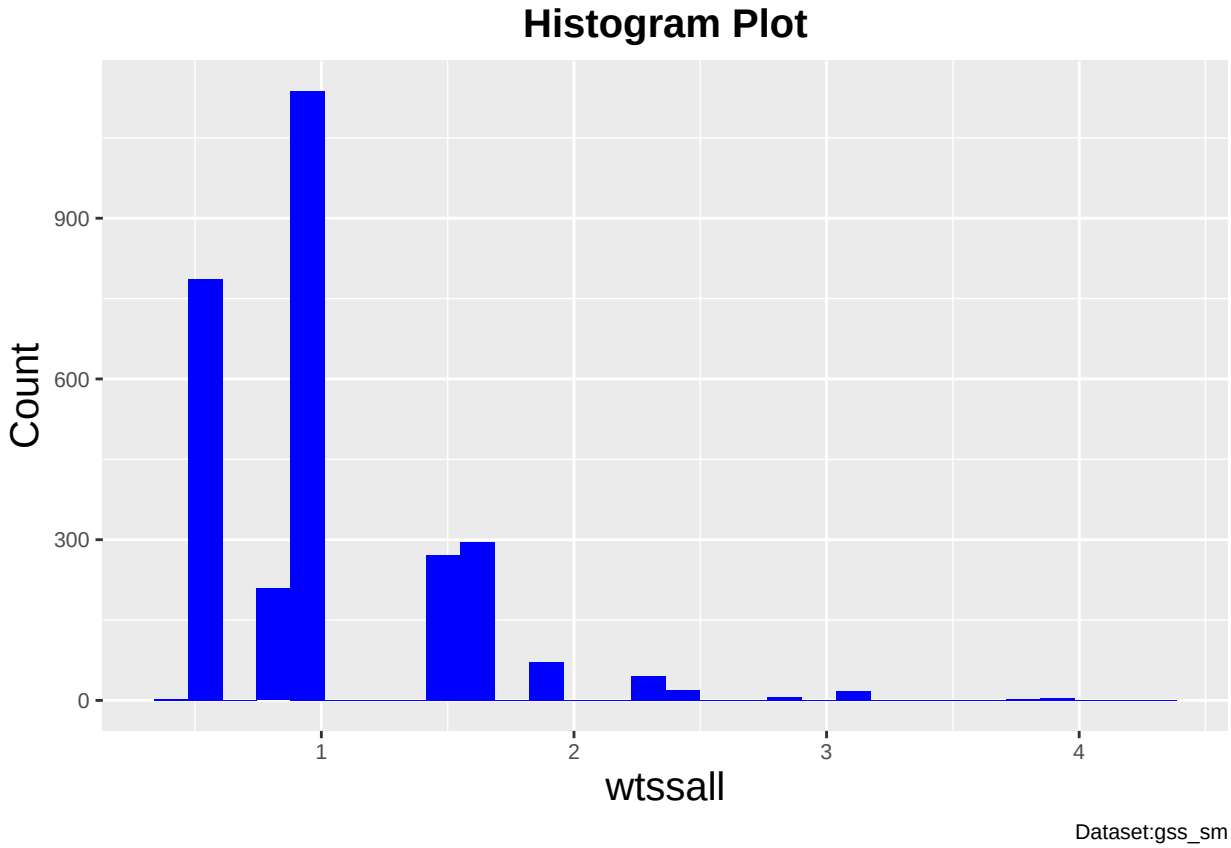
Sagar Kumar | 194161013

Question 1

wtsall is one of the variable in the gss_sm. Using the appropriate plots discuss whether this variable follows normal distribution?

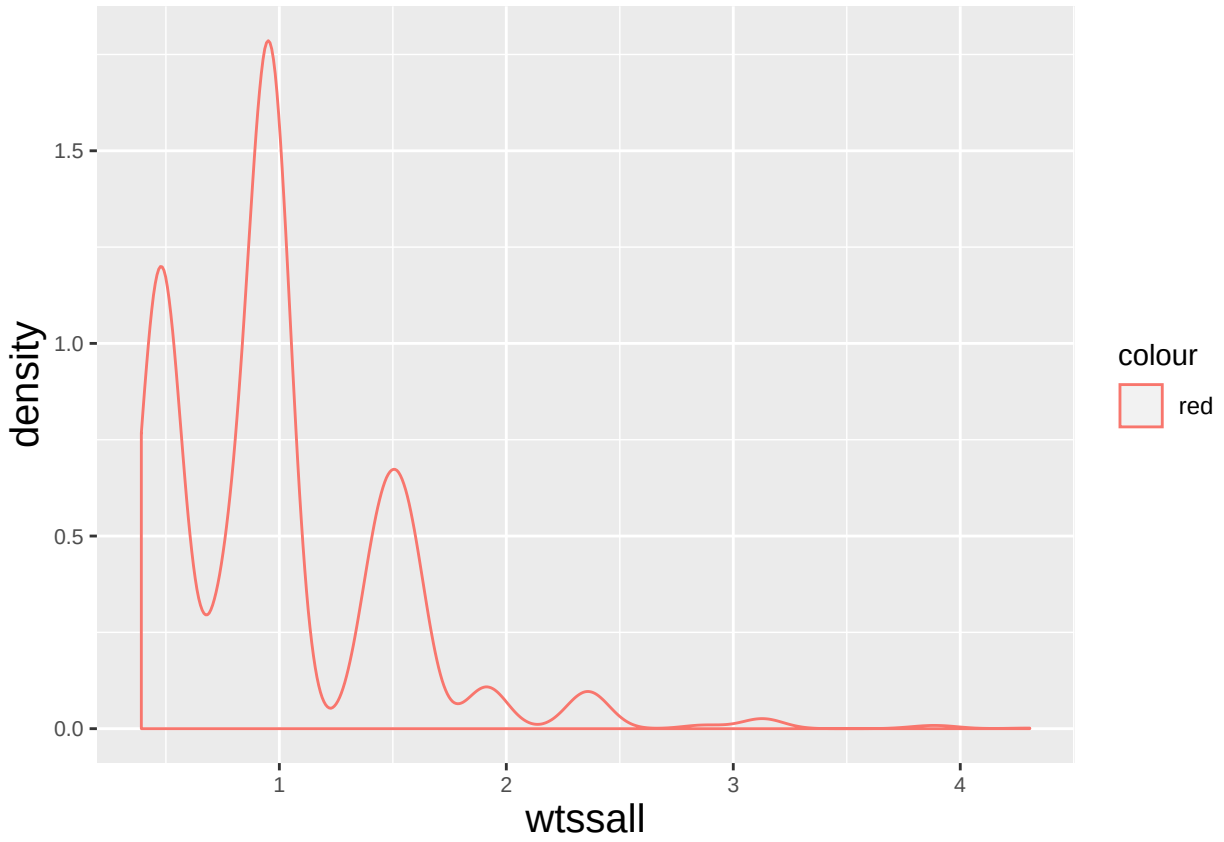
```
ggplot(data=gss_sm)+geom_histogram(aes(wtssall),fill="blue")+theme+labs(x="wtssall", y="Count",
       caption="Dataset:gss_sm",
      title="Histogram Plot")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



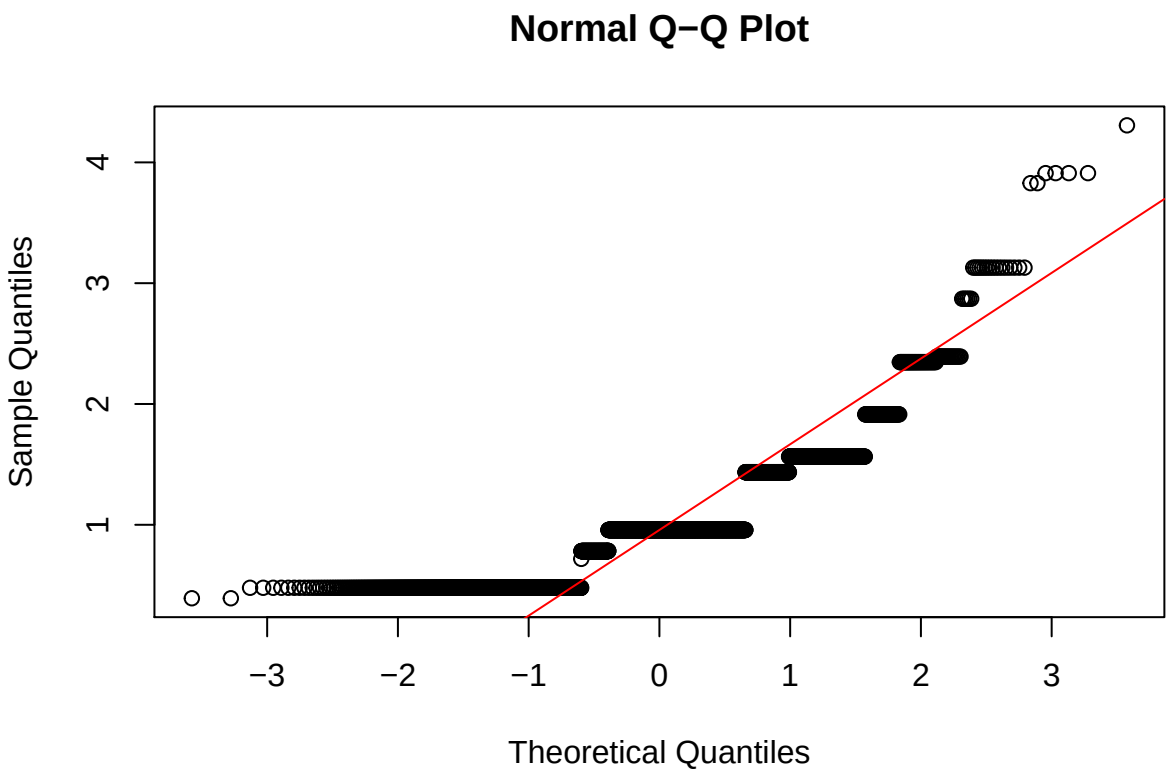
Histograms are much preferred to analyze this type of data .But from histogram we don't see pattern like normal distribution. Normal Distribution is symmetric around mean.

```
ggplot(data=gss_sm)+
  geom_density(aes(wtssall,col="red"))+theme
```



This is not a normal distribution curve.It is multimodal function. This curve has many maxima and minima. So we can say that this curve is a combination of Normal distribution

```
qqnorm(gss_sm$wtssall)
qqline(gss_sm$wtssall,col="red")
```



Ans:

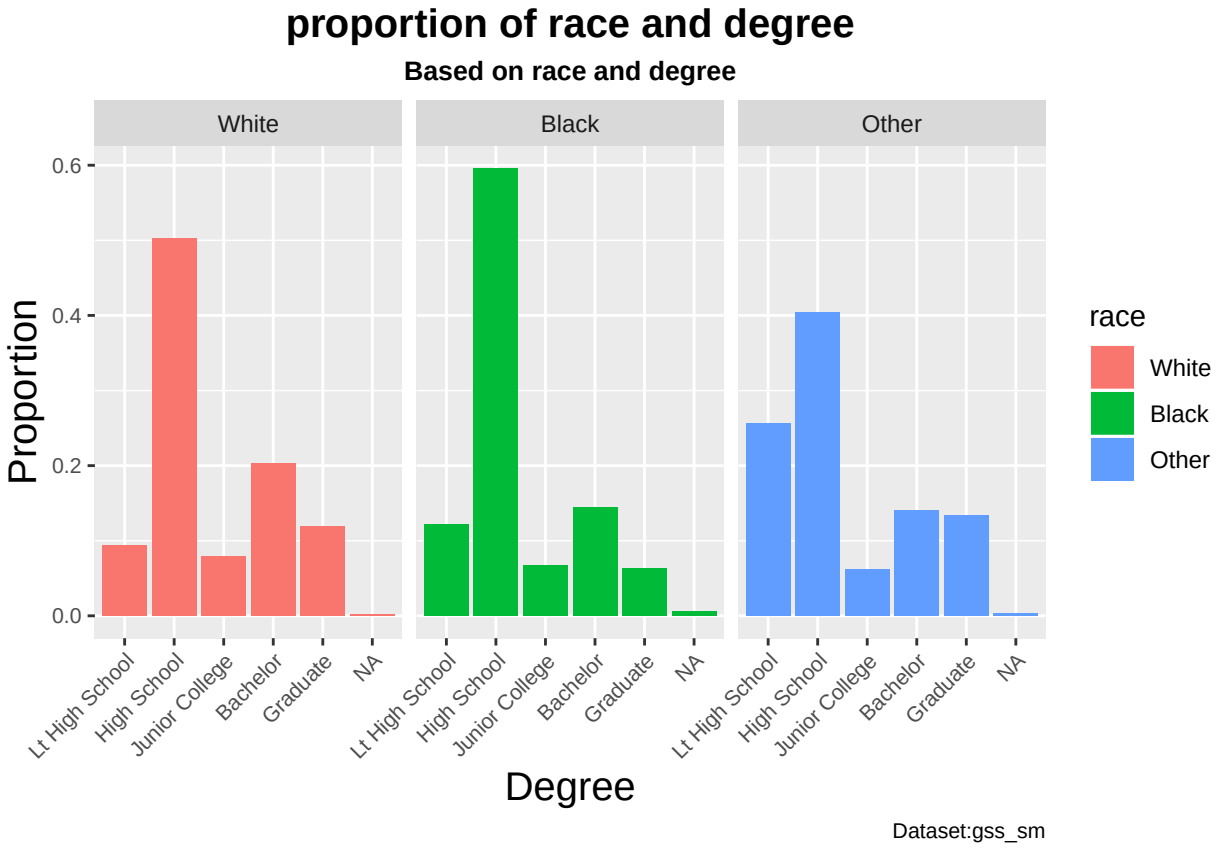
from QQPlot we can test the normality of curve.Clearly we can see that

Normal Q-Q Plot" provides a graphical way to determine the level of normality. The red line indicates the values your sample should adhere to if the distribution was normal. The dots are your actual data. If the dots fall exactly on the red line, then your data are normal

Question2

We want to understand proportion of different degree across race. Visualize using appropriate plot. Discuss if one could have used different plots to show the same information.

```
ggplot(data=gss_sm, aes(x=degree,fill=race))+geom_bar(aes(y=..prop..,group=1))+facet_wrap(~race)+theme+l
       caption="Dataset:gss_sm",
       subtitle="Based on race and degree",
       title="proportion of race and degree")+theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



This plot shows the porportion of race and degree. yes we can use different plots to show the same information like pie chart. But Barplot is best to visualize categorical variable And from this plot we can infer that proportion of high school student is more. and from this plot we can infer that approximately same no of students in Black and white . and less no of students in other compare to white and black race.

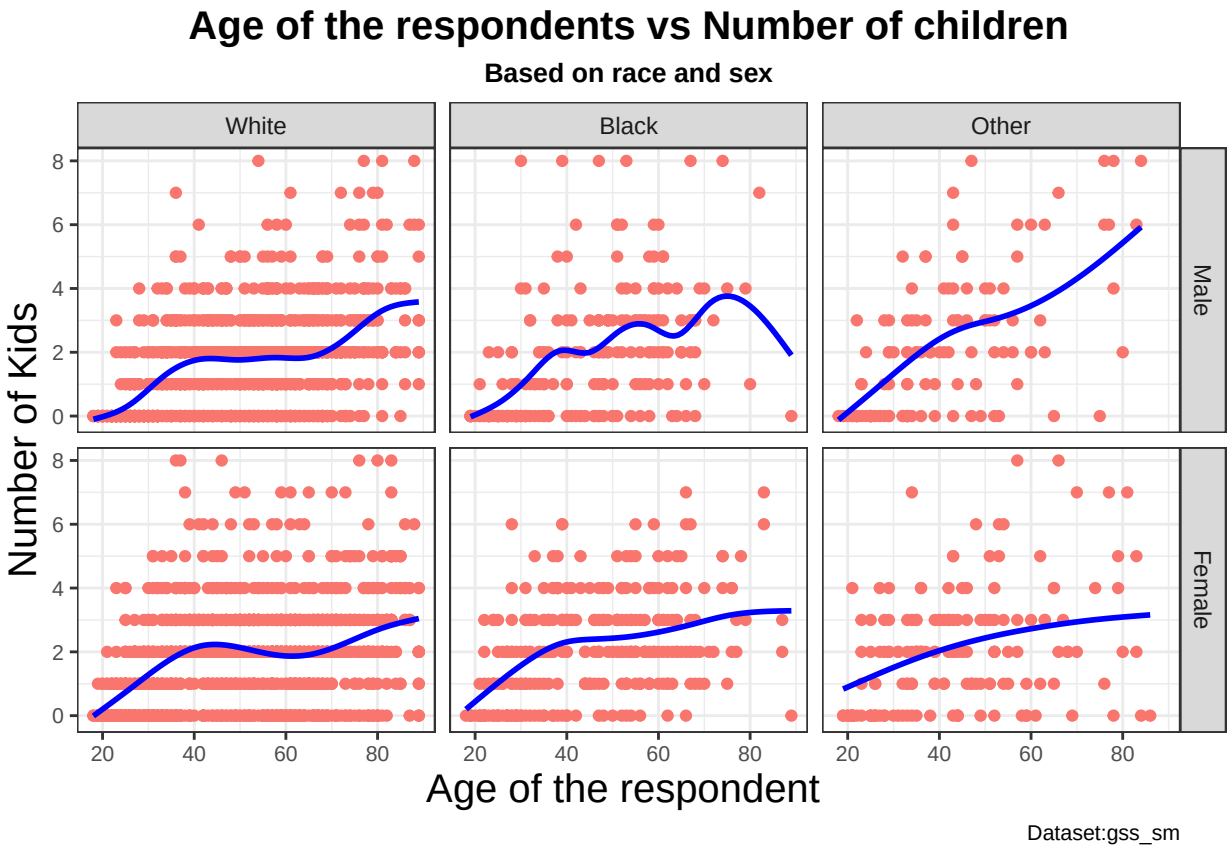
Question 3

We want to understand the relation between age of respondents and the number of kids (childs variable) they have. However, we want to understand this relationship along with the two other variables sex and race. Draw a single scatterplot to visualize this relationship. Remark: A faceted plot with several sub-plots will be considered a single plot.

```
p<-ggplot(data=gss_sm, aes(x=age, y=childs))

p+ geom_point(aes(col="blue"))+
  geom_smooth(se=FALSE, col="blue")+
  facet_grid(sex~race)+
   guides(col=FALSE) +
  theme_bw()+
  theme+
  labs(x="Age of the respondent ", y="Number of Kids",
       caption="Dataset:gss_sm",
       subtitle="Based on race and sex",
       title="Age of the respondents vs Number of children")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



Ans: age. age of respondent childs. number of children.

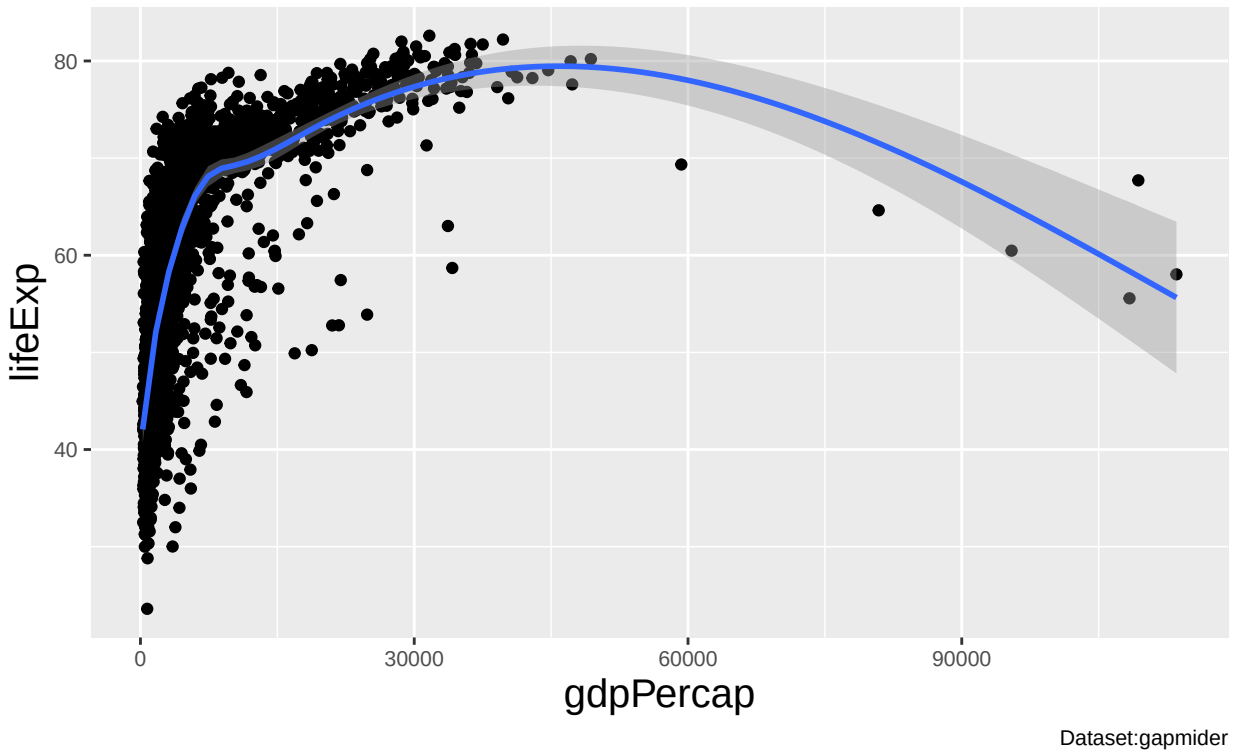
Blue line in this plot shows the trend of the Data. ans red dot is Scatter Plot between Age of the respondent and Number of kids Here six subplots where in every plot blue line shows the no.of kids. And clearly we can see that data follows linear realtionship. As Age of the respondent is more number of kids is more. And we can see that at the Age of 20 (age of respondent)No.of kids is almost zero for all the races. And at the Age of 80 (age of respondent) No.of kids is more.

Question 4

This question is based on the gapminder dataset. We want to understand the relation between gdpPercap (x-axis) and lifeExp. Compare the following: 1. Two smoothed scatterplot considering the two variables. In one case, use the raw values of both variables. In the second case, use log-transformation on gdbPercap. Further, we want to reflect continent in the second case. In other words, we want points on the plot should be colored based on continent. The smoothening with standard error should also reflect the same continent color. Finally, the x-ticks should be labelled with $. For example, values like 1e+03 on the x-tick should be written as $1000 or $1,000.

```
ggplot(data=gapminder,aes(x=gdpPercap,y=lifeExp))+geom_point()+geom_smooth()+
  labs(x="gdpPercap", y="lifeExp",
       caption="Dataset:gapmider",
       subtitle=" ",
       title=" Scatter Plot of gdpPercap v/s lifeExp")+theme
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



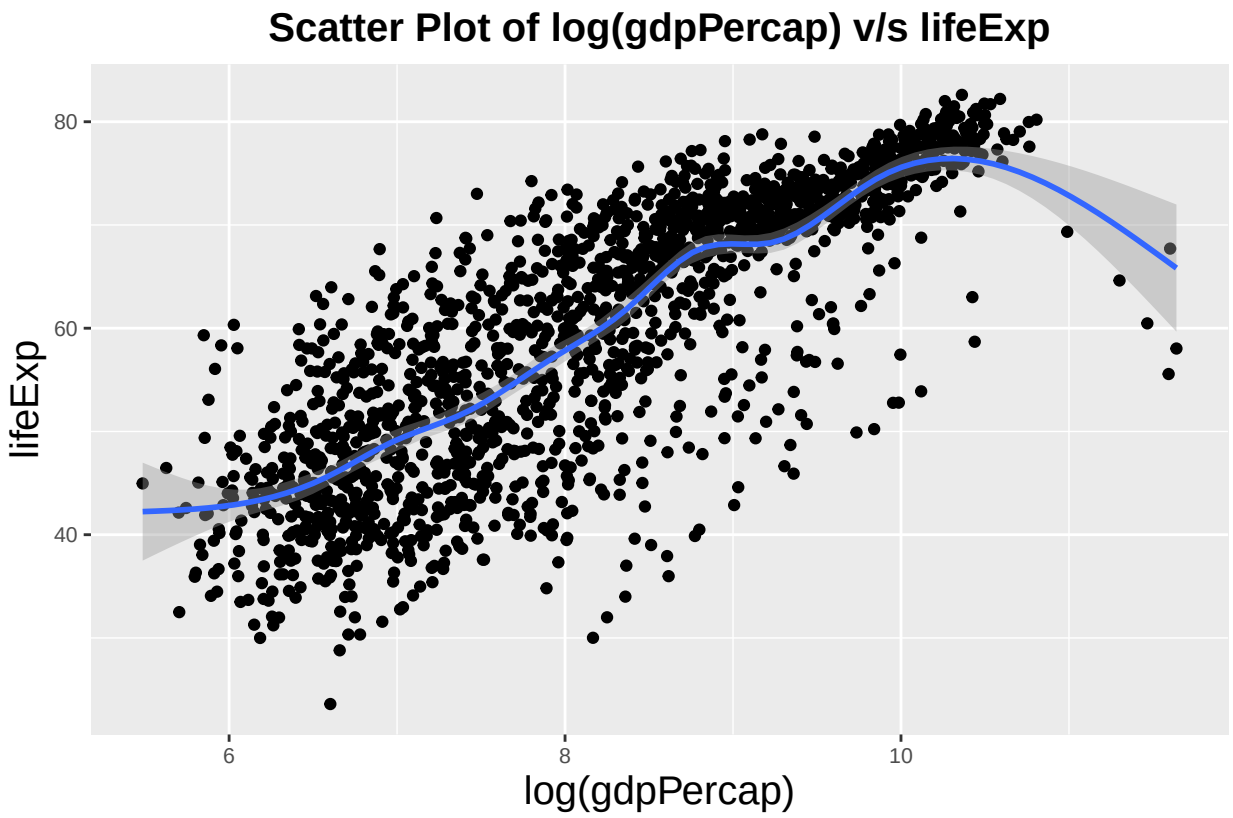
Nonlinear relationship between gdpPercap and LifeExp.
In this plot blue curve shows smoothness

As clearly we can see that Most of the data is concentrated around Zero. And most of the data lie in between 0 to 30000

```
ggplot(data=gapminder,aes(x=log(gdpPercap),y=lifeExp))+geom_point()+geom_smooth()+theme+  labs(x="log(g
       caption="Dataset:gapmider",

       title="Scatter Plot of log(gdpPercap) v/s lifeExp")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

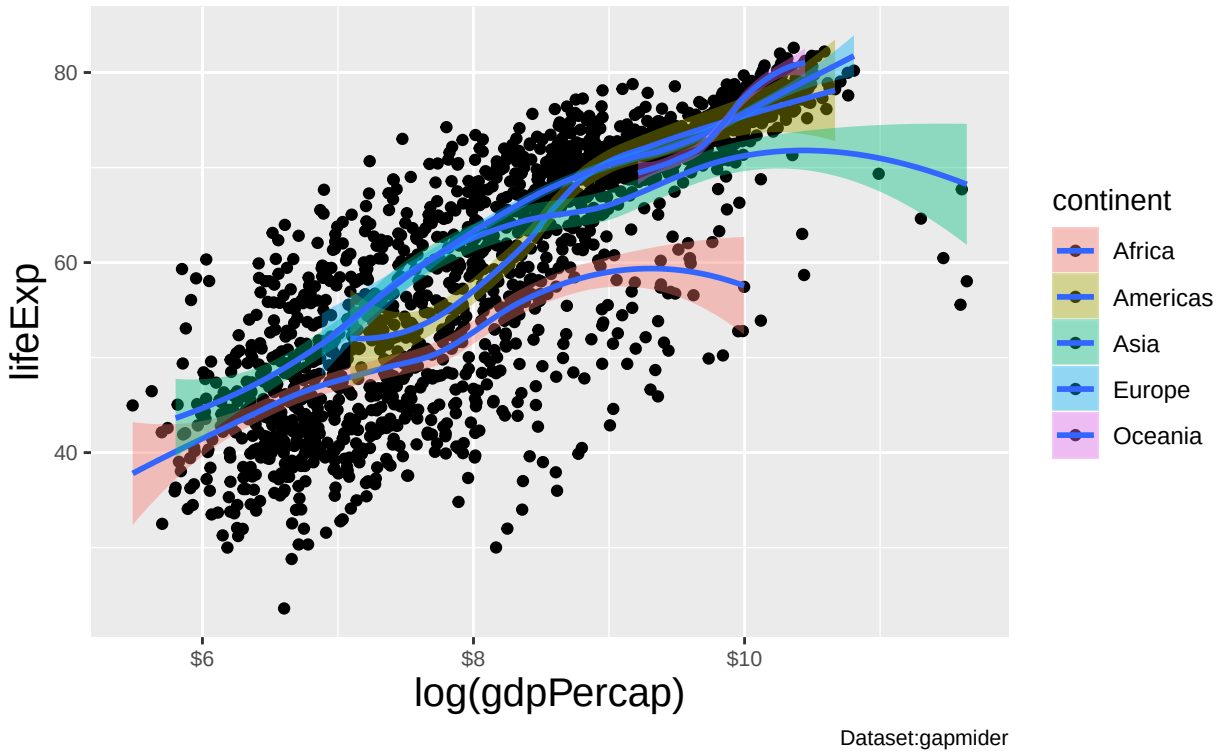


In this plot By default method is used gam. GAM: Generalized Additive Models. The term 'GAM' being taken to include any quadratically penalized GLM and a variety of other models estimated by a quadratically penalised likelihood type approach. The degree of smoothness of model terms is estimated as part of fitting.

Nonlinear relationship between gdpPercap and LifeExp.
In this plot blue curve shows smoothness

```
ggplot(data=gapminder,aes(x=log(gdpPercap),y=lifeExp,fill=continent))+geom_point()+geom_smooth(se=TRUE)+
       caption="Dataset:gapmider",
       subtitle=" ",
       title=" ")+scale_x_continuous(labels=scales::dollar)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



In this we have five different smootheing line. because we have five continent. and all smootheing line shows the trend between lifeExp and log(gdpPercap). And we can see that that there is linear relationship between lifeExp and log(gdpPercap). if we use method="lm"(linear models) to fit the data then we will see linear trends between these variables.

So we can infer that irrespective of the continnent lifeExp is increases as log(gdpPercap) is more