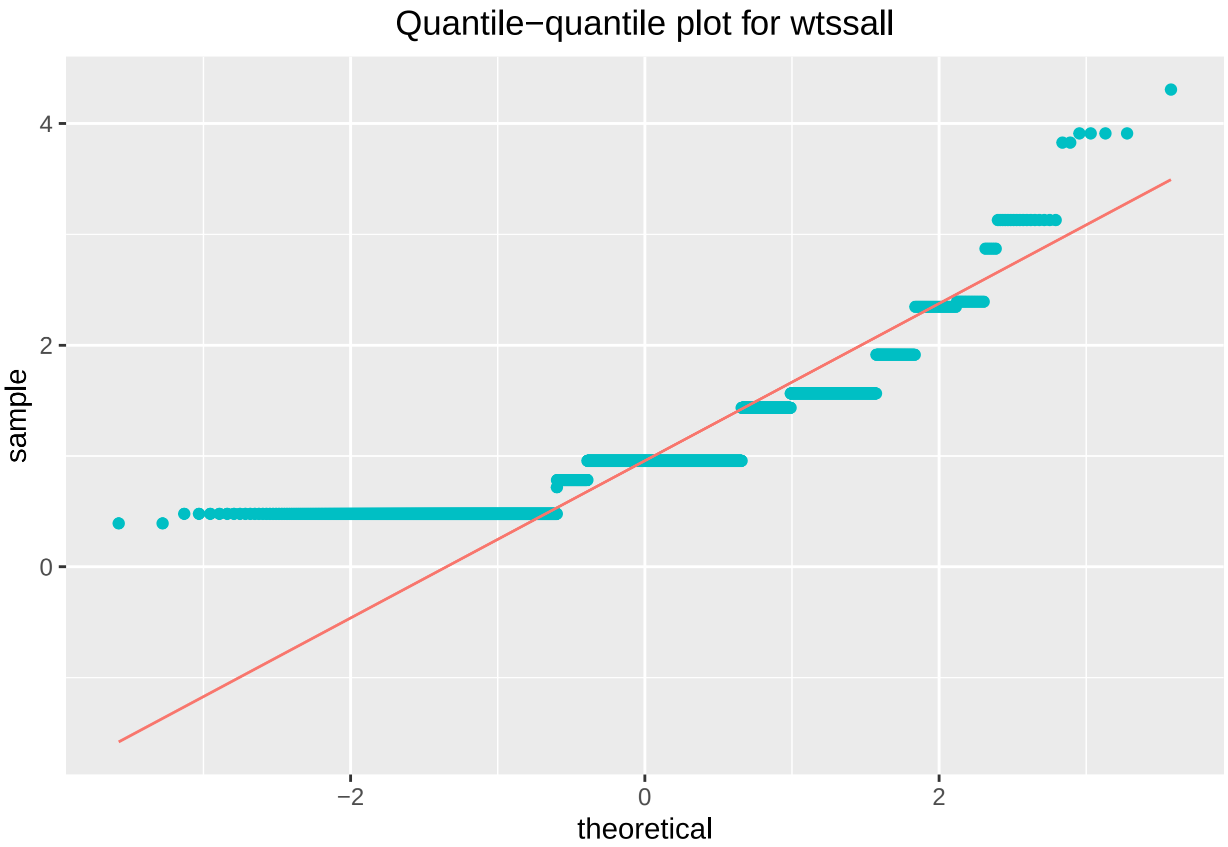
194161016

Q1)wtsall is one of the variable in the gss_sm. Using the appropriate plots discuss whether this variable follows normal distribution?

```
library("tidyverse")
```

```
## -- Attaching packages --------------------------------------------------------------------------- ti

## v ggplot2 3.3.0     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.4
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts ------------------------------------------------------------------------------ tidyvers
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library("gapminder")
library("socviz")
library("ggplot2")

#loding data
data.gss <- data.frame(gss_sm)

#plotting a qqplot using stat_qq and stat_qq_line
ggplot(data.gss,aes(sample=wtssall))+
  stat_qq(mapping=aes(color="red"),show.legend = FALSE)+
  stat_qq_line(mapping=aes(color="blue"),show.legend = FALSE)+labs(title="Quantile-quantile plot for wt
```

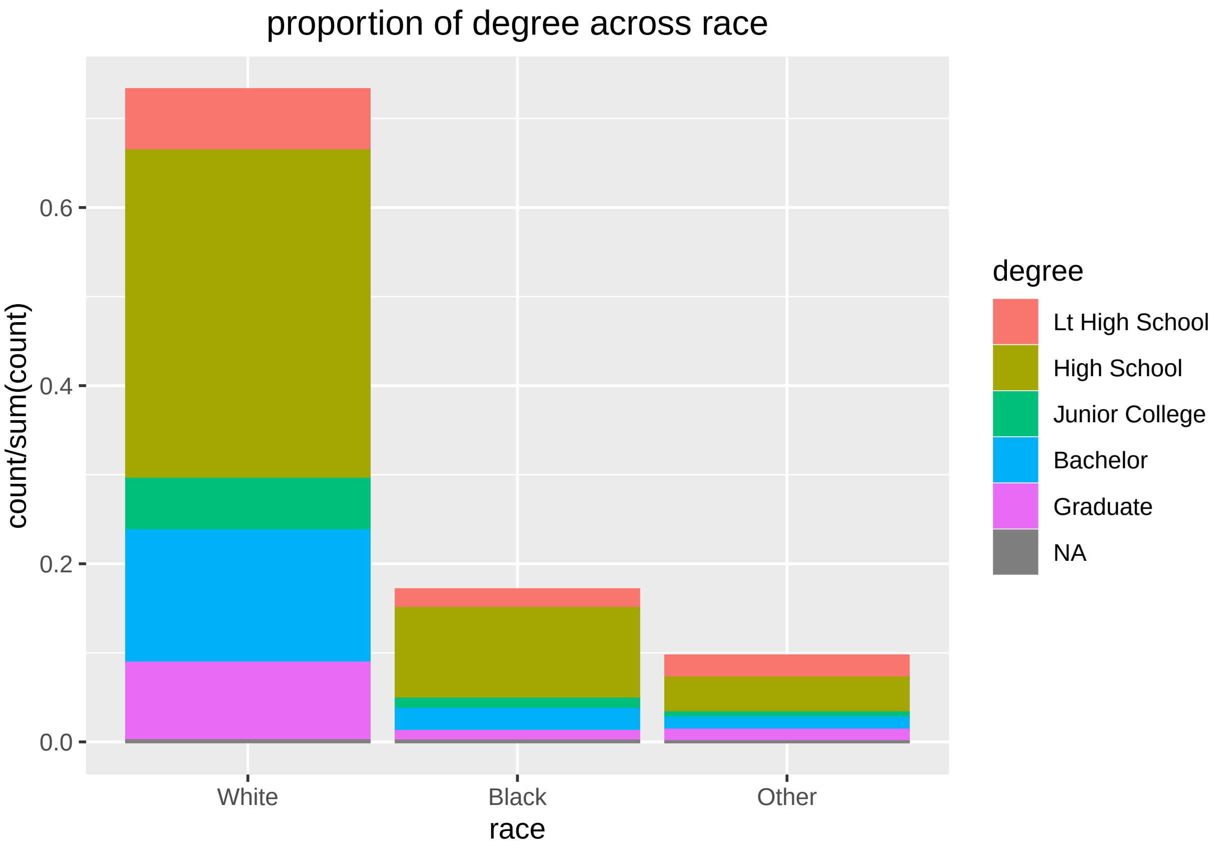


Ans:1)To check the normality of any variable we can use quantile-quantile plot.It uses theoretical quantiles(sampled from standard normal distribution) on x-axis and actual samples on y-axis.
2)If any variable is following normal distribution then when we plot them against theoretical quantiles, the points should allign on the straight line.
3)In our case the points are strongly deviating from the straight line.So we can say that wtssall is not following normal distribution.

Q2)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.

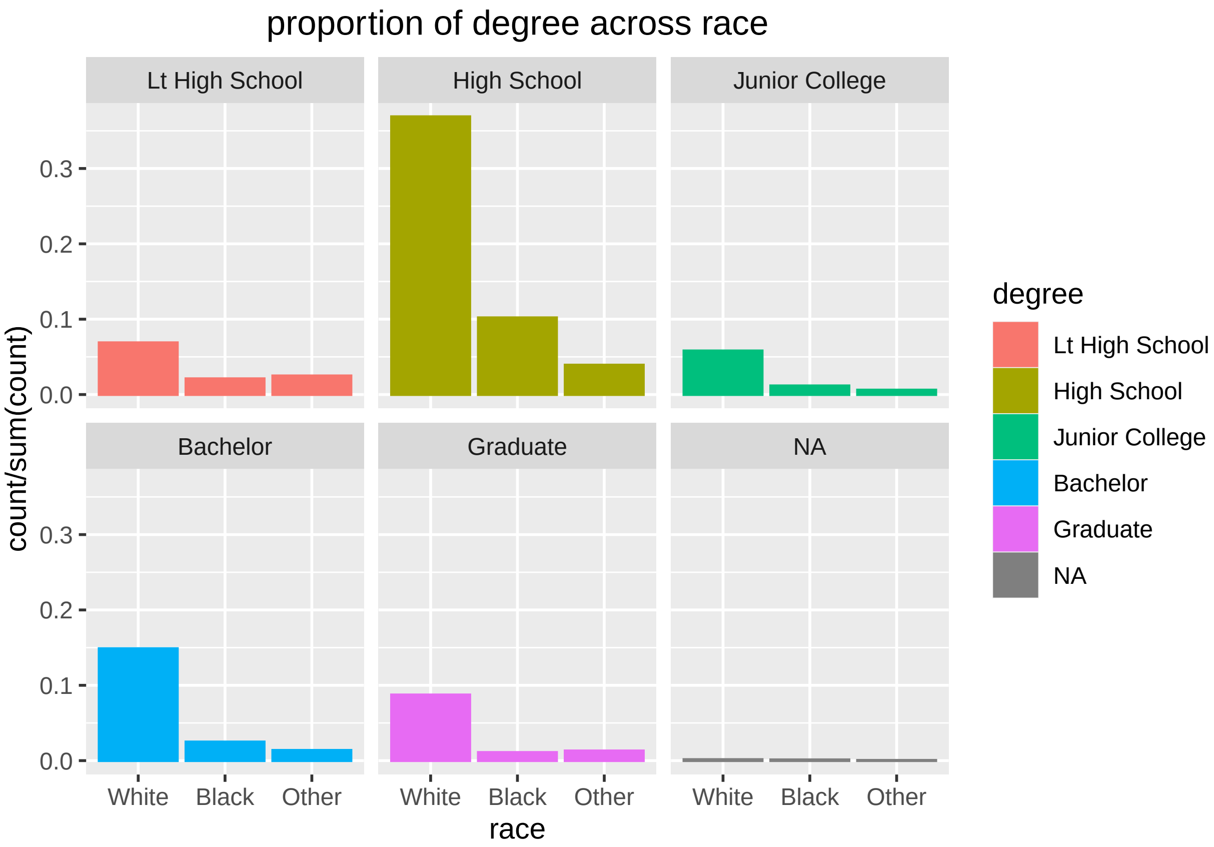
```
#plotting
ggplot(data=data.gss,mapping=aes(x=race,y=..count../sum(..count..)))+
  geom_histogram(mapping=aes(color=degree,fill=degree),stat="count")+labs(title="proportion of degree ac
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



```
ggplot(data=data.gss,mapping=aes(x=race,y=..count../sum(..count..)))+
  geom_histogram(mapping=aes(color=degree,fill=degree),stat="count")+facet_wrap(~degree)+labs(title="pr
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

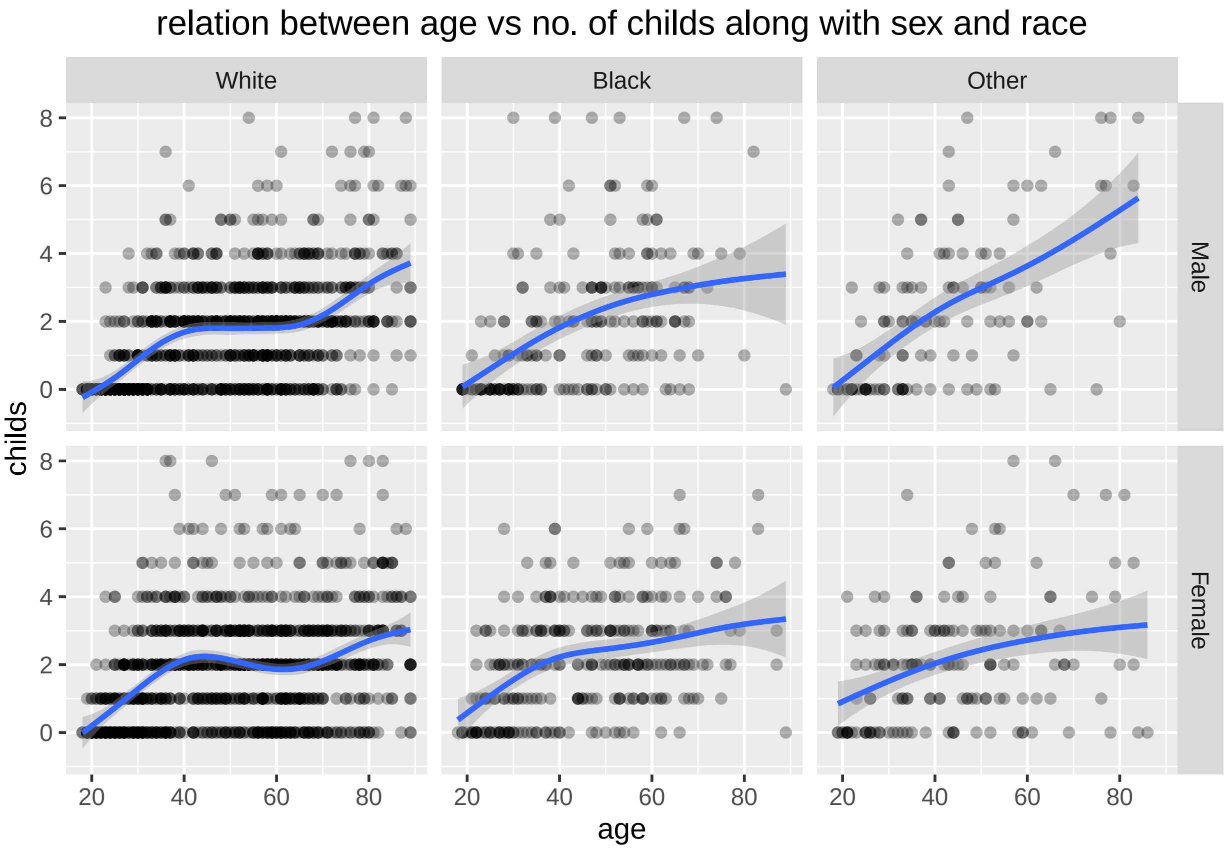


Ans:1)As the race and degree are categorical variables, a bar plot or histogram will give us the perfect information of proportion of degree across all the races.
2)Here the x axis has different races and y axis gives us different proportions colored according to degree.
3)By observing the plot we can say that there are more no of white people with a high school degree.
4)We can also see that there are comparitively more number of people with high school degree in all the races.
5)There are significantly very less people with no degree across all the races.
6)we can also using facet_wrap to check for different proportions of degrees across different races.

Q3)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.

```
ggplot(data=data.gss,mapping=aes(x=age,y=childs))+
  geom_point(alpha=0.3)+
  facet_grid(sex~race)+geom_smooth()+labs(title="relation between age vs no. of childs along with sex ar
```

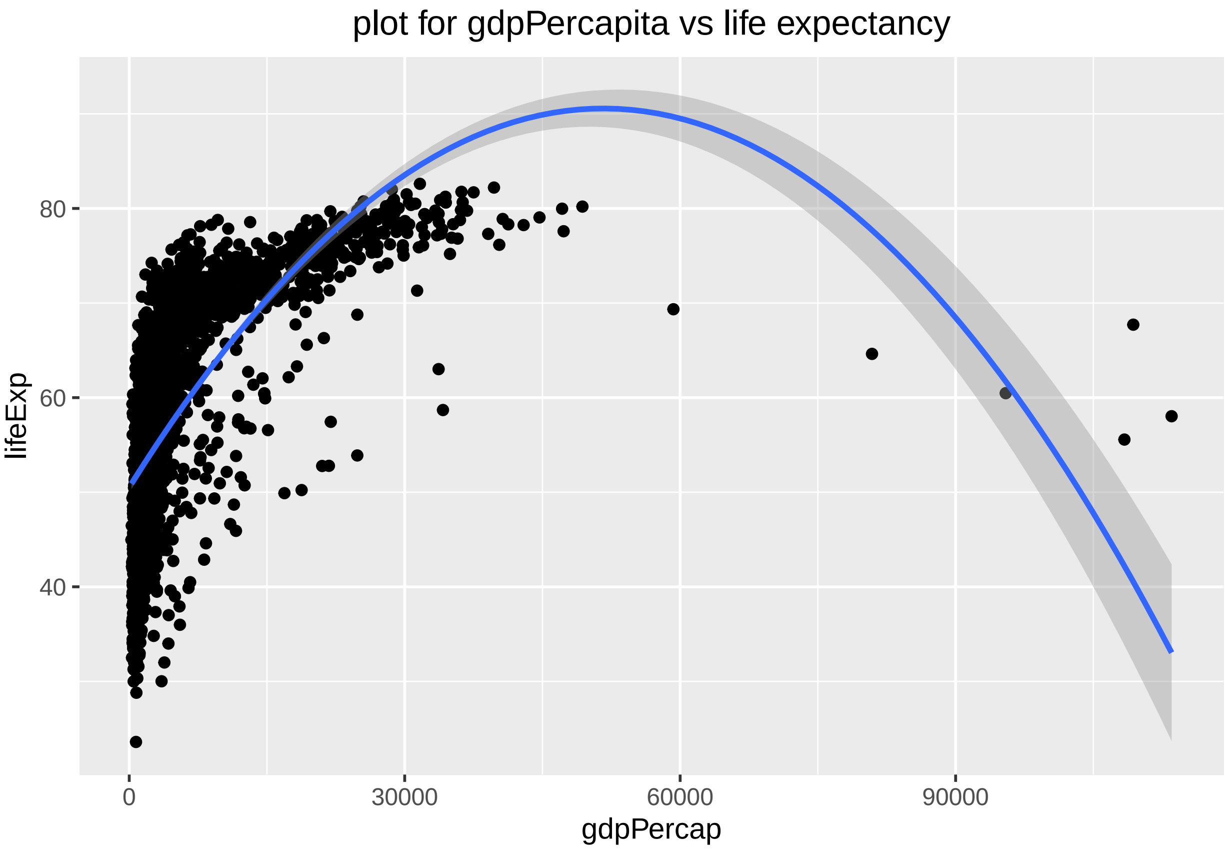
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'

## Warning: Removed 18 rows containing non-finite values (stat_smooth).

## Warning: Removed 18 rows containing missing values (geom_point).
```



Ans:1)We can see a grid containing age(x-axis) and no of children(y-axis).The grid is divided according to different combinations of sex and race.
2)we observe that white people has comparitively more no of children and other race people has less aongst all the three races.
3)We can also observe that there are more people from all the races with no children(last line of every subplot).

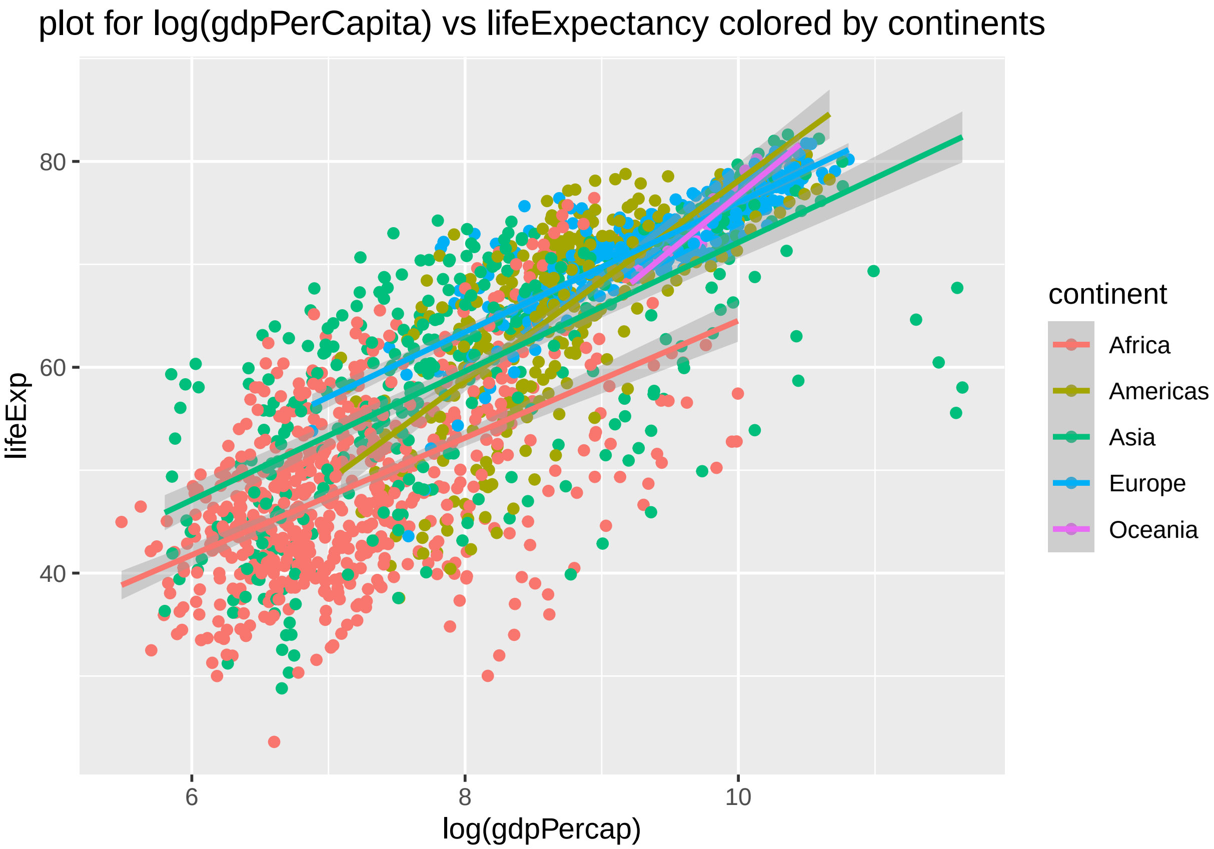
Q4)This question is based on the gapminder dataset. We want to understand the relation between gdpPercap (x-axis) and lifeExp. Compare the following: 1. Two smoothed scatterplot considering the two variables. In one case, use the raw values of both variables. In the second case, use log-transformation 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.

```r
#loading data
data.gapminder <- data.frame(gapminder)

#plotting gdppercap vs lifeexpectancy using geom_point, geom smooth
ggplot(data=data.gapminder,mapping=aes(x=gdpPercap,y=lifeExp))+
  geom_point()+
  geom_smooth(method="lm",formula = y~poly(x,2))+labs(title="plot for gdpPercapita vs life expectancy")
```



```
#plotting log(gdppercap) vs lifeexpectancy using geom_point, geom smooth
ggplot(data=data.gapminder,mapping=aes(x=log(gdpPercap),y=lifeExp))+
  geom_point(mapping=aes(color=continent))+
  geom_smooth(method="lm",mapping=aes(color=continent))+labs(title="plot for log(gdpPerCapita) vs lifeE
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Ans:1)In this I have fitted two smoothing methods lm(quadratic in x terms) and lm(linear in x terms).
2)If we observe the plot gdpPercap vs lifeExpectancy plot, the data points are not following linearity(neither strictly increasing nor decreasing).So fitting a quadratic linear model will give us a better smoothing.
3)We can also clearly see the trend of lifeExpectancy is strictly increasing for lesser gdpPercap and becoming almost constant as we increase gdpPercap.
4)In the second plot, we are considering log(gdpPercap) vs lifeexpectancy.
5)If we observe the plot, the data points are follwoing a linear pattern(strictly increasing).Hence a linear model will give us a better smoothing.