Problem Set 5

Apurva Gupta, Shailendra Patil, Surbhi Paithankar

February 27, 2018

PART A

Firstly we have created a new variable called death rate. Later, we constructed a new data frame with all the said parameters. We picked up PctBlack as our parameter of choice.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.2
```

```
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 3.3.2
```

```
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 3.3.2
```

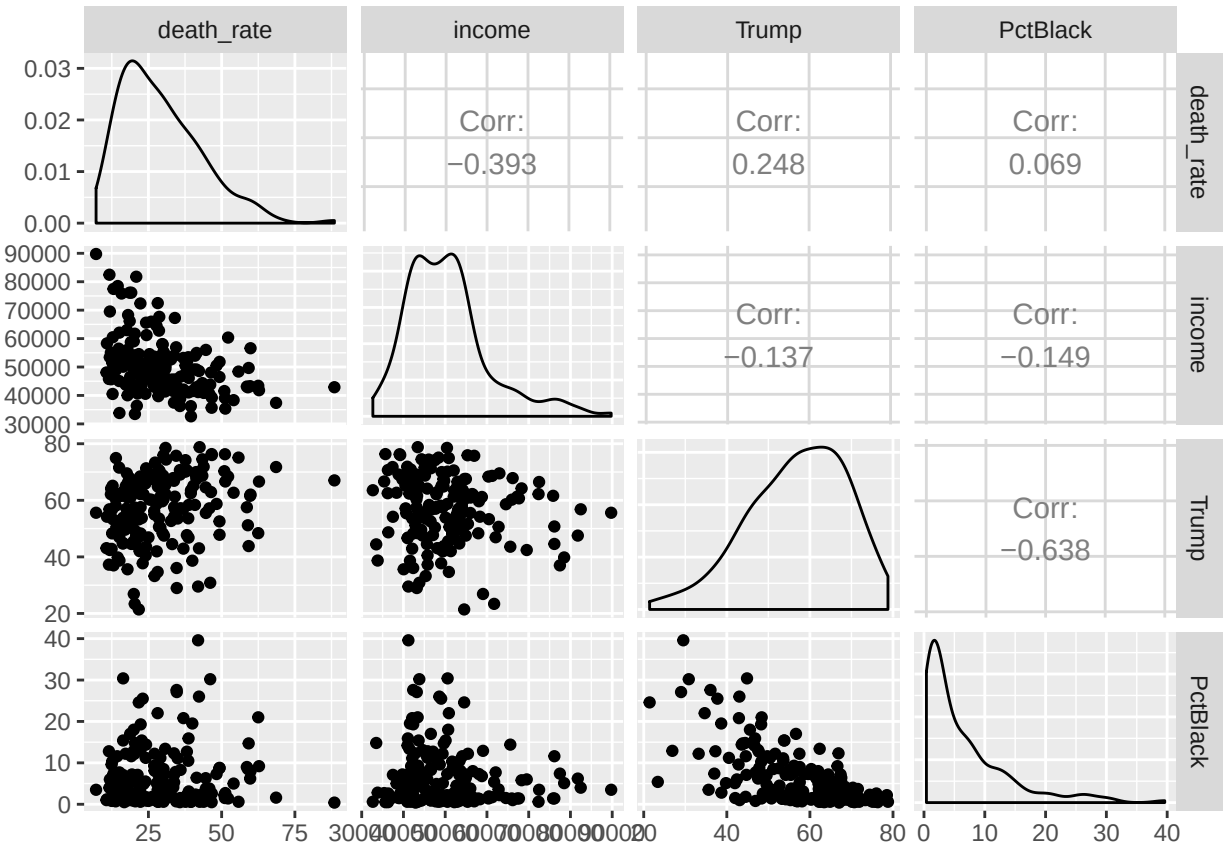
```
death.data = read.table('rustdrugs2016.txt',header = TRUE)

death_rate = (death.data$Deaths / death.data$Population)* 100000
deathrate.df = data.frame(death_rate = death_rate, income = death.data$Income,
                    Trump = death.data$Trump*100 , PctBlack = death.data$PctBlack)
head(deathrate.df)
```

```
##   death_rate income    Trump PctBlack
## 1   15.01998  45073 71.55178      3.7
## 2   11.51526  45808 37.26064     12.8
## 3   30.02192  45145 68.72632      1.6
## 4   21.71616  54548 21.41932     24.6
## 5   16.26358  53375 44.71541      7.1
## 6   14.52600  78487 39.80224      5.1
```

Lets draw GGpairs to look into the inter-relationships between variables.

```
ggpairs(deathrate.df)
```



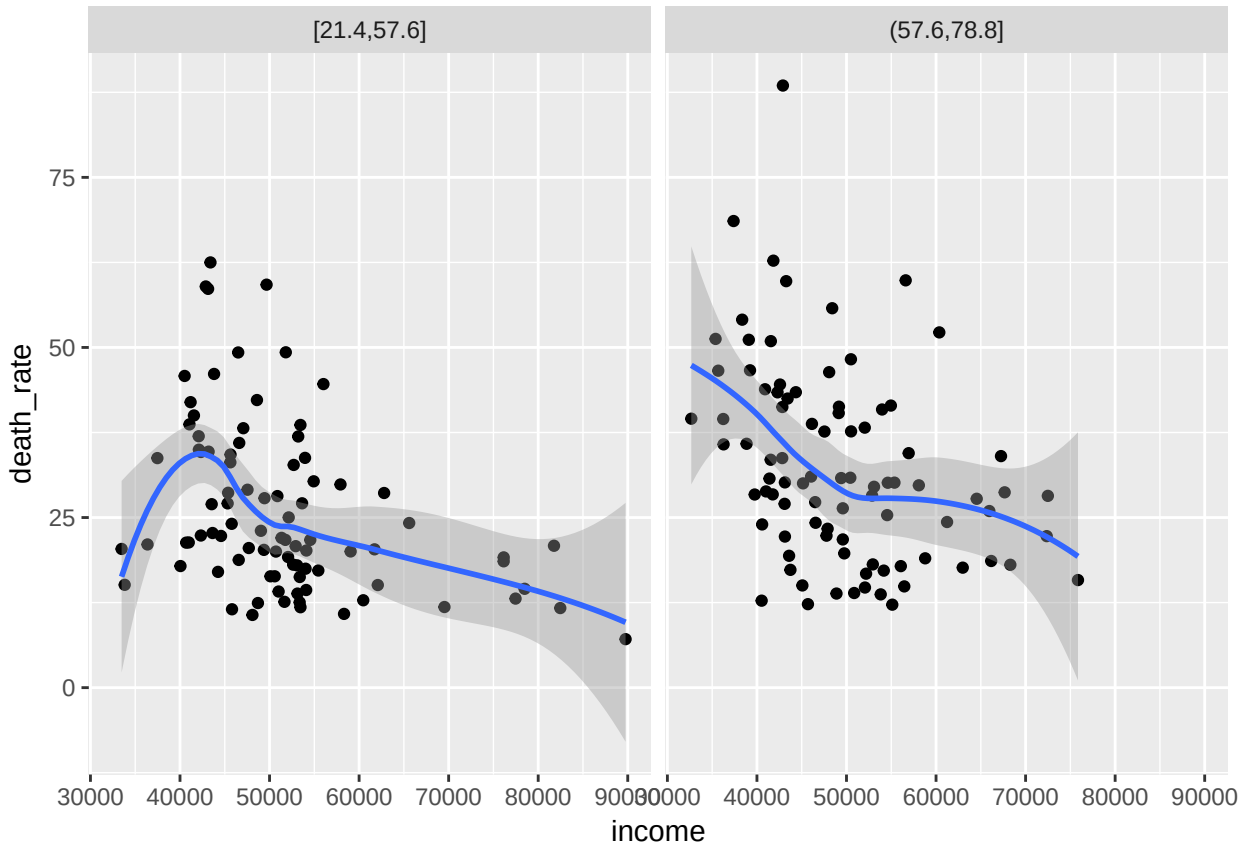
Here it is evident that variables Trump percentage and Black percentage have relatively high collinearity.However it can be ignored(<70%). Also the black percentage is extremely skewed towards right. This might potentially be a problem in building a model.

PART B

Lets go ahead and fit a loess model to check if we need interactions

```
gg = ggplot(deathrate.df,aes(income,death_rate))
gg + geom_point()+geom_smooth() + facet_grid(~cut_number(Trump,2) )
```

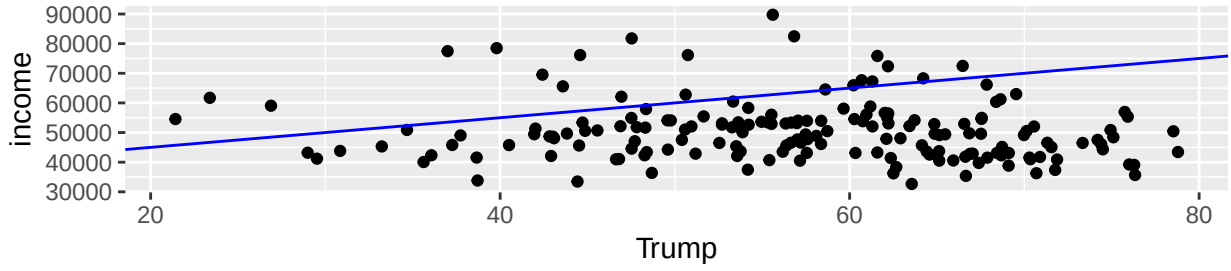
```
## `geom_smooth()` using method = 'loess'
```

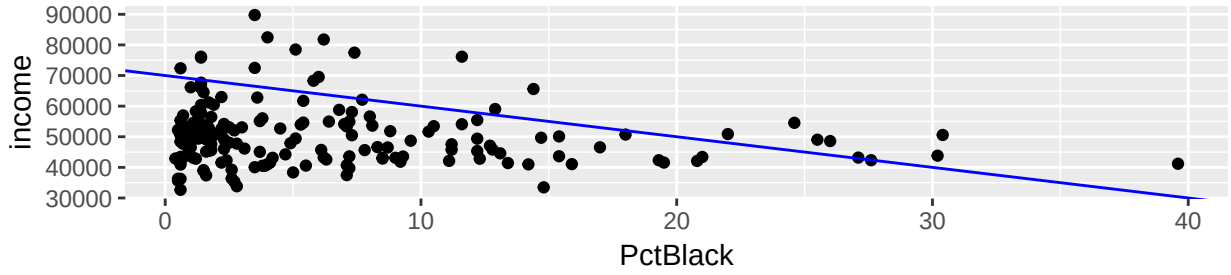


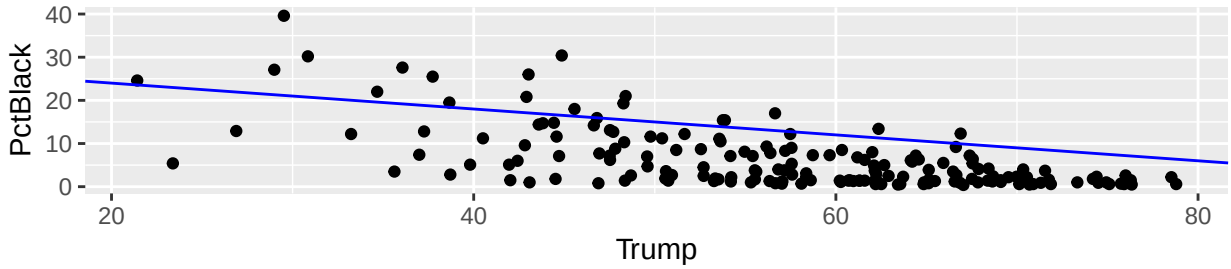
The curves seem to follow a non-monotonic relationships over different Trump proportions. This indicates that we need an interaction between the income and Trump proportion for creating the model. Also we can see a wide confidence band for high income region. The ggplot predictions in these regions aren't credible

Lets look at the scatterplot between income and Trump proportion.

```
g1 = ggplot(deathrate.df,aes(x = Trump, y = income))+ geom_point()+
  geom_abline(intercept = 35000,slope =500, color = "blue")
g2 = ggplot(deathrate.df,aes(x = PctBlack, y = income))+ geom_point()+
  geom_abline(intercept = 70000,slope =-1000, color = "blue")
g3 = ggplot(deathrate.df,aes(x = Trump, y = PctBlack))+ geom_point() +
    geom_abline(intercept = 30,slope =-0.3, color = "blue")
grid.arrange(g1,g2,g3)
```







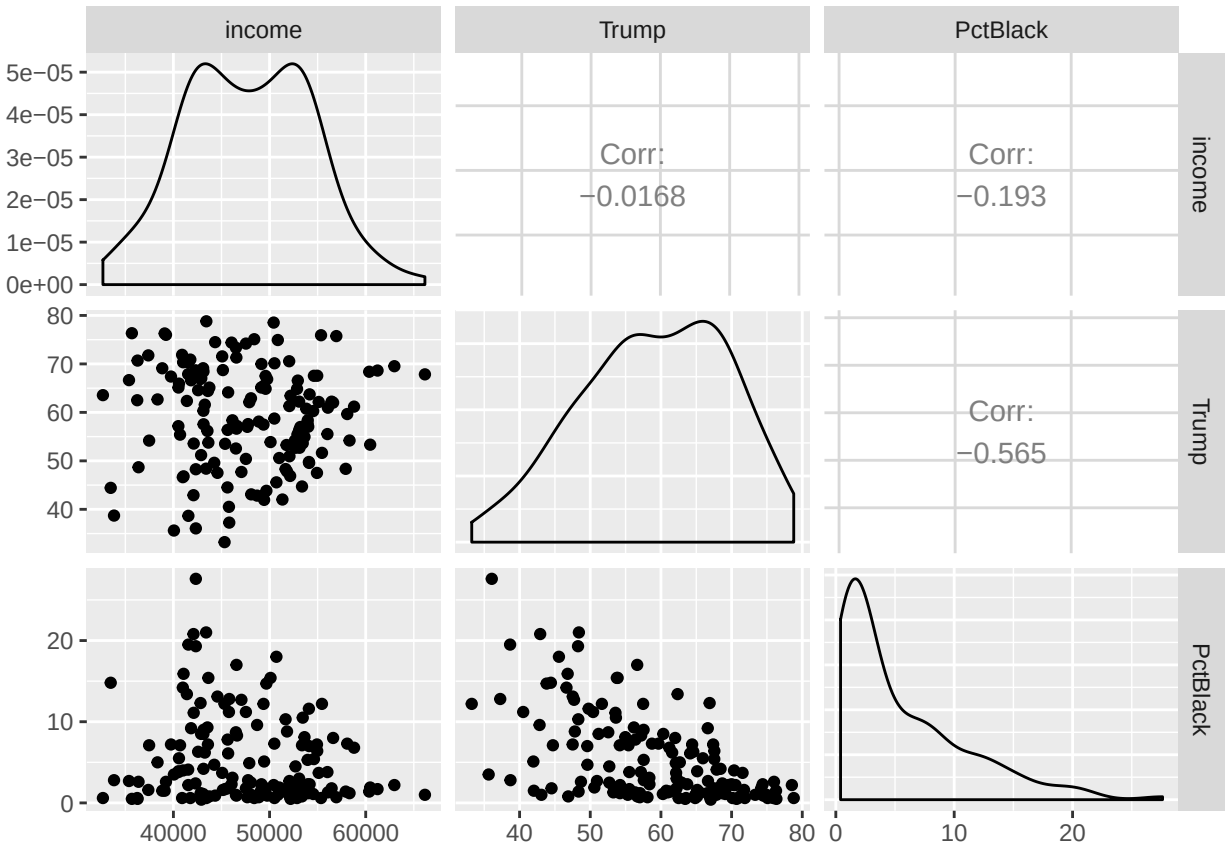
We will crop out the space where we do not have sufficient predictor data available.

```
income=deathrate.df$income
Trump = deathrate.df$Trump
PctBlack = deathrate.df$PctBlack

crop = (income>32000) & (income<75000) & (income<(500*Trump+35000)) &
(income<(-1000*PctBlack+70000))& (Trump >23) & (Trump<80)
deathrate.df = deathrate.df[crop,]
```

Lets check the ggpairs.

```
ggpairs(deathrate.df,columns = 2:4)
```



##PART C

Lets go ahead and fit the lm model with interactions. Later using AIC we will compare various models using diffferent interactions and choose the best of them.

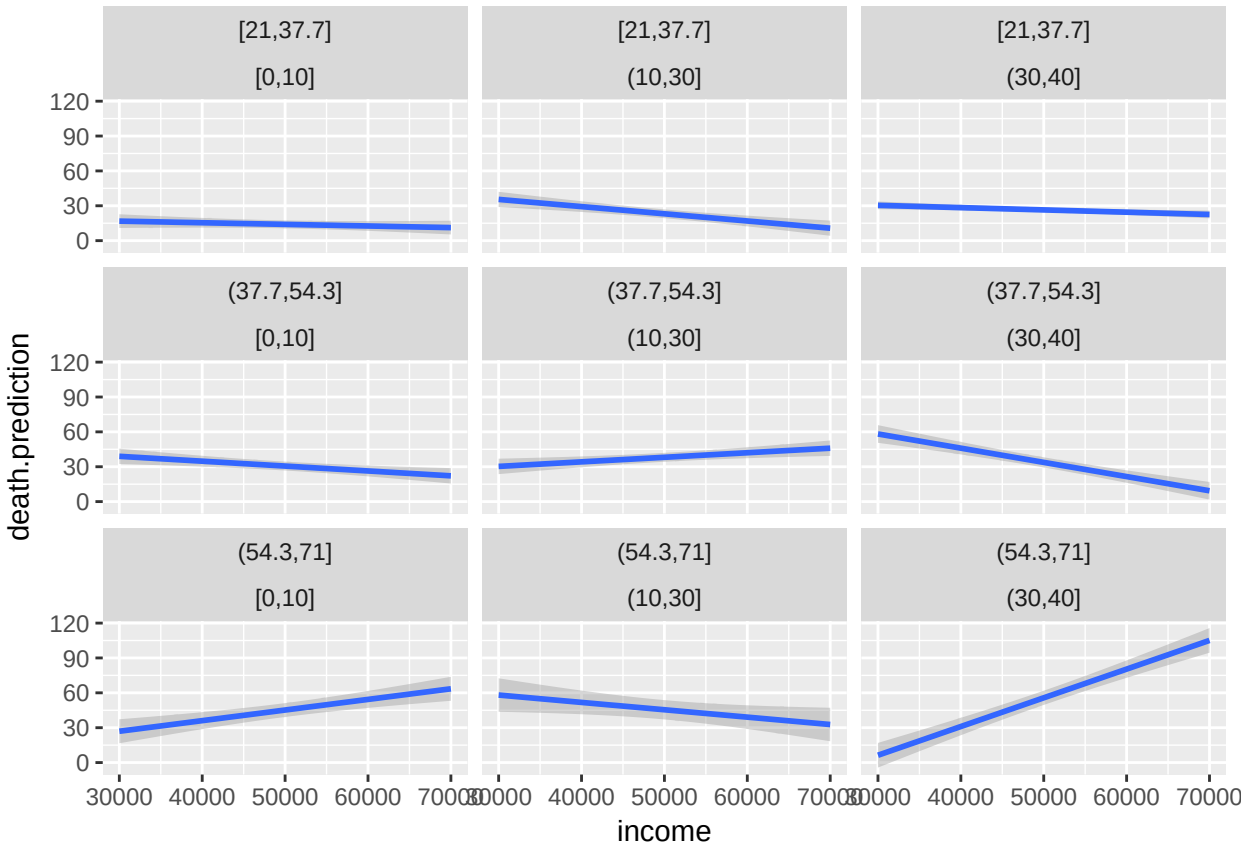
```
model1.lo = lm(death_rate~ income*Trump,data = deathrate.df)
model2.lo = lm(death_rate~ income*Trump*PctBlack,data = deathrate.df)
model3.lo = lm(death_rate~ income*PctBlack,data = deathrate.df)

AIC(model1.lo,model2.lo,model3.lo)
```

```
##           df      AIC
## model1.lo  5 1188.055
## model2.lo  9 1187.676
## model3.lo  5 1198.399
```

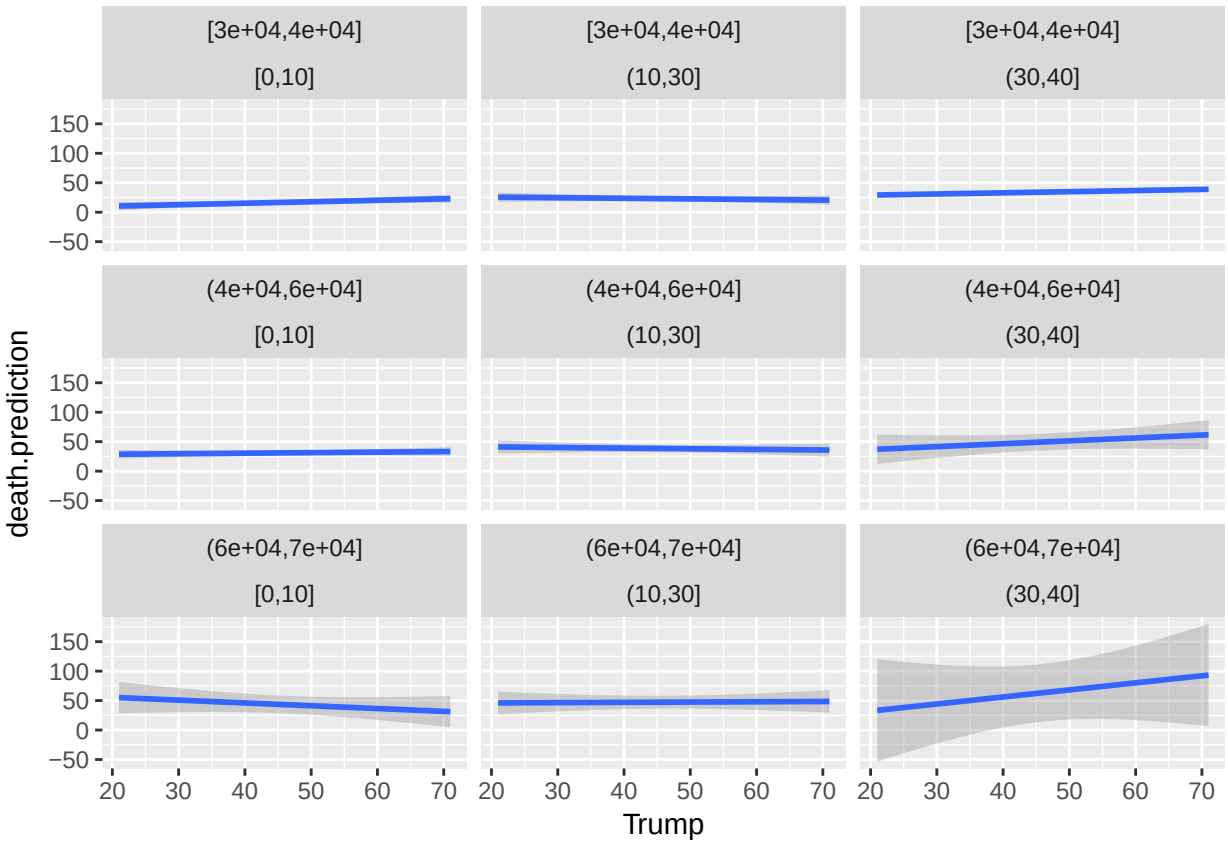
We can see that the second model has lowest AIC. Hence we select it for our further analysis.

```
death.grid = expand.grid(income= seq(30000,75000,10000),Trump=seq(21,80,10),
                         PctBlack = seq(0,40,10))
death.prediction = predict(model2.lo,newdata = death.grid)
death.prediction.df = data.frame(death.grid, as.vector(death.prediction))
gg = ggplot(death.prediction.df,aes(income,death.prediction))
gg +  geom_smooth(method = 'lm') + facet_wrap(~cut_number(Trump, n = 3)+ cut_number(PctBlack,n = 3))
```



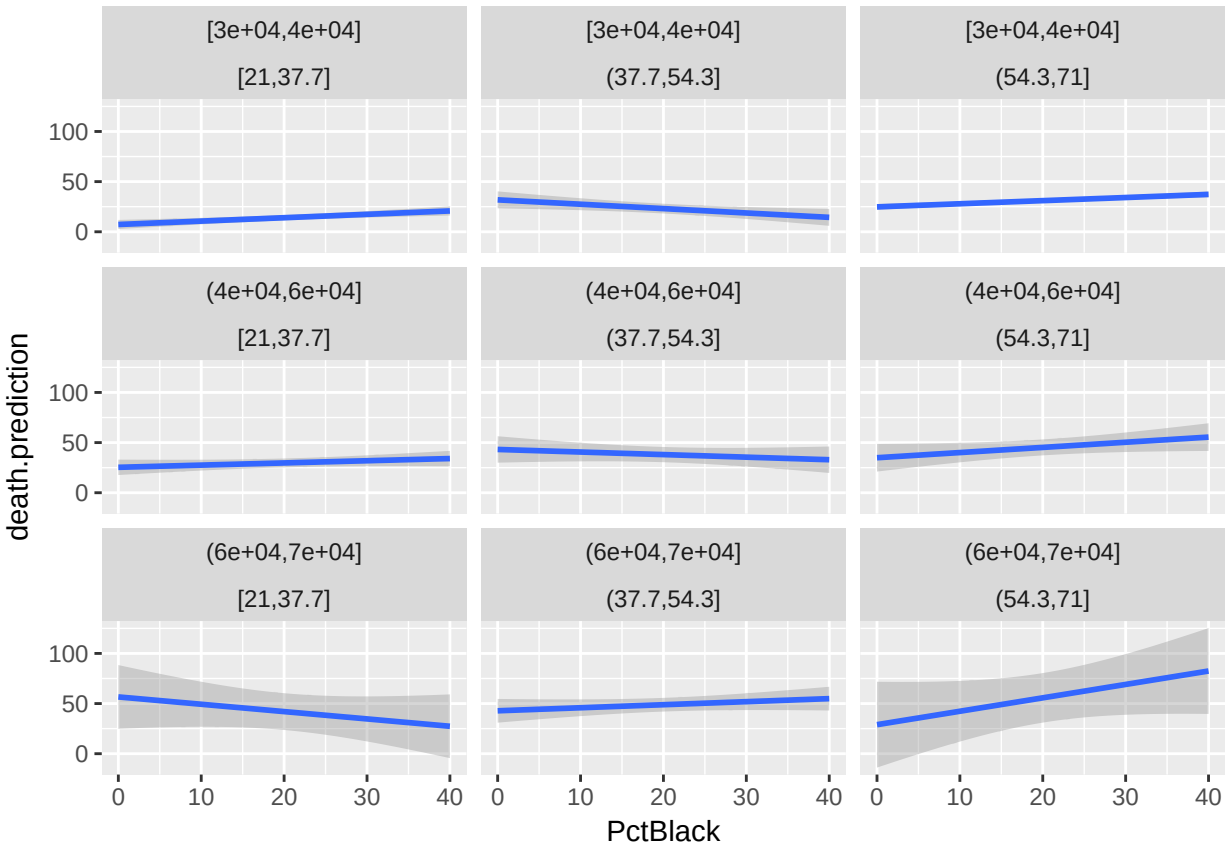
The slopes seems to change as we go downwards. This indicates the income:black percentage interaction.

```
gg = ggplot(death.prediction.df,aes(Trump,death.prediction))
gg +  geom_smooth(method = 'lm') + facet_wrap(~cut_number(income, n = 3)+ cut_number(PctBlack,n = 3))
```



Here the slope changes slightly when we go downwards. This shows Trump and Black percentage interaction.

```
gg = ggplot(death.prediction.df,aes(PctBlack,death.prediction))
gg +  geom_smooth(method = 'lm') + facet_wrap(~cut_number(income, n = 3)+ cut_number(Trump,n = 3))
```



Finally in this graph, we can see a positive slope in the last column. This shows that there is an interaction black percentage and Trump percentage.