615 assignment 3 data cleaning

Si Chen,Grace Xie, Kaiyu yan, Siwei Hu

September 30, 2018

Introduction

We collect the data from he Data.gov, named Community Health Status Indicators (CHSI) to Combat Obesity, Heart Disease and Cancer(Chsi_dataset).It is imperative to understand that behavioral factors such as obesity, tobacco use, diet, physical activity, alcohol and drug use, sexual behavior and others substantially contribute to these deaths. After data cleaning, we choose obesity, uninsurence, diabetes, smkoing as our variables.

```
#import datasets from MEASURESOFBIRTHANDDEATH and RISKFACTORSANDACCESSTOCARE
tableM <- read.csv("MEASURESOFBIRTHANDDEATH.csv")
tableR <- read.csv("RISKFACTORSANDACCESSTOCARE.csv")
tableD <-read.csv("DEMOGRAPHICS.csv")

# select the variable total death in RISKFACTORSANDACCESSTOCARE
death <- select(tableM, Total_Deaths)

# select the variable obesity, smoker, diabetes, uninsured, Dentist_Rate,CHSI_State_Name from RISKFACTO
risk <- select(tableR,Obesity,Smoker,Diabetes,Uninsured,Dentist_Rate,CHSI_State_Name)
#selec the variable population_size and build a new table
population <- select(tableD, Population_Size)

# combine to a new table
tablenew <- cbind.data.frame(death,risk,population)

# remove all useless inputs
tablenew[tablenew < 0] <- NA
```

```
## Warning in Ops.factor(left, right): '<' not meaningful for factors
```

```
# create the final data frame
tablenew1<-tablenew[complete.cases(tablenew),]
```

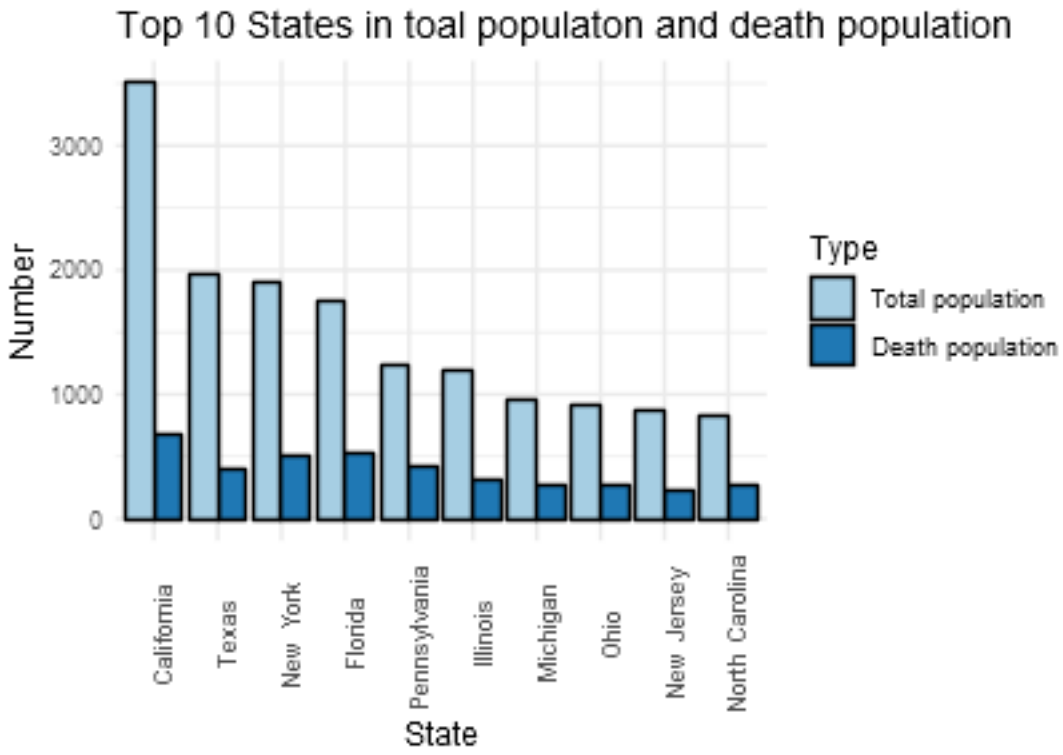
We choose Top 10 total population states and filter their relation between population and total deaths to make more sense of the data.

```
# total population and death in every state

T_population<-tablenew1 %>%
 group_by(CHSI_State_Name) %>%
 summarise(total_p = sum(Population_Size)/10000,total_d=sum(Total_Deaths)/1000,
           death_rate=total_d/total_p) %>%
 arrange(desc(total_p))
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

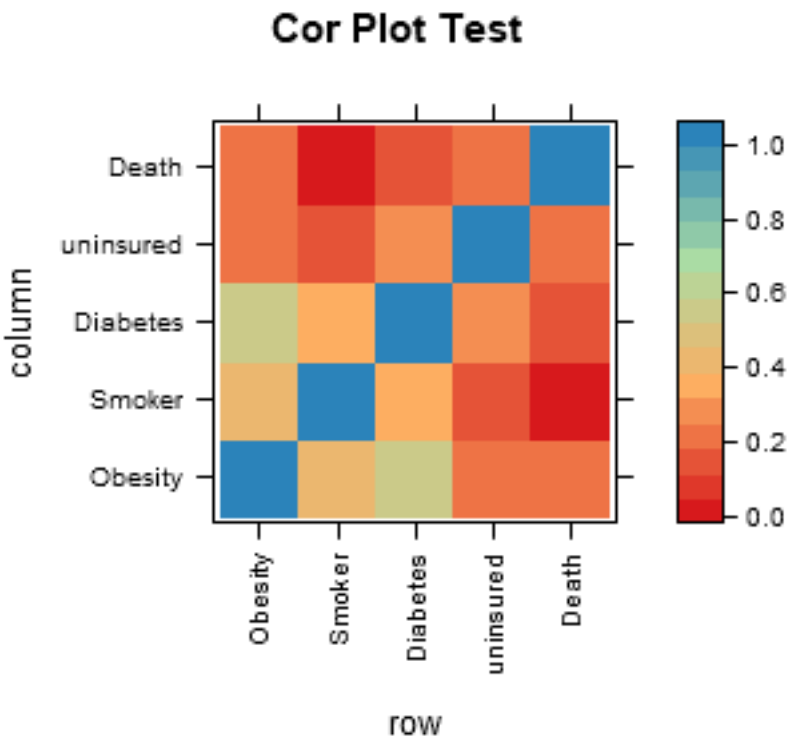
```
T_p<-data.frame(T_population[c(1:10),])
T_p1<-T_p %>% select(CHSI_State_Name,total_d)%>% mutate(type=rep("1",10)) %>% arrange(desc(total_d))
T_p2<-T_p %>% select(CHSI_State_Name,total_p) %>% mutate(type=rep("0",10))
names(T_p1) <- c("State","Number","Type")
```

```
names(T_p2) <- c("State","Number","Type")
T_p_f<-merge(T_p1,T_p2,all=T)


ggplot(data=T_p_f, aes(x=State, y=Number, fill=Type)) +
geom_bar(stat="identity",position=position_dodge(),color="black")+
theme_minimal()+
theme(axis.text.x = element_text(angle = 90))+
scale_fill_brewer(palette="Paired",labels=c("Total population", "Death population"))+
scale_x_discrete(limits=c("California","Texas","New York", "Florida","Pennsylvania","Illinois","Michigan
ggtitle("Top 10 States in toal populaton and death population")
```



we select all variables which may impact the death rate and scale our uninsured and death data into percent. Then we build a chart about correlation between each two variables.

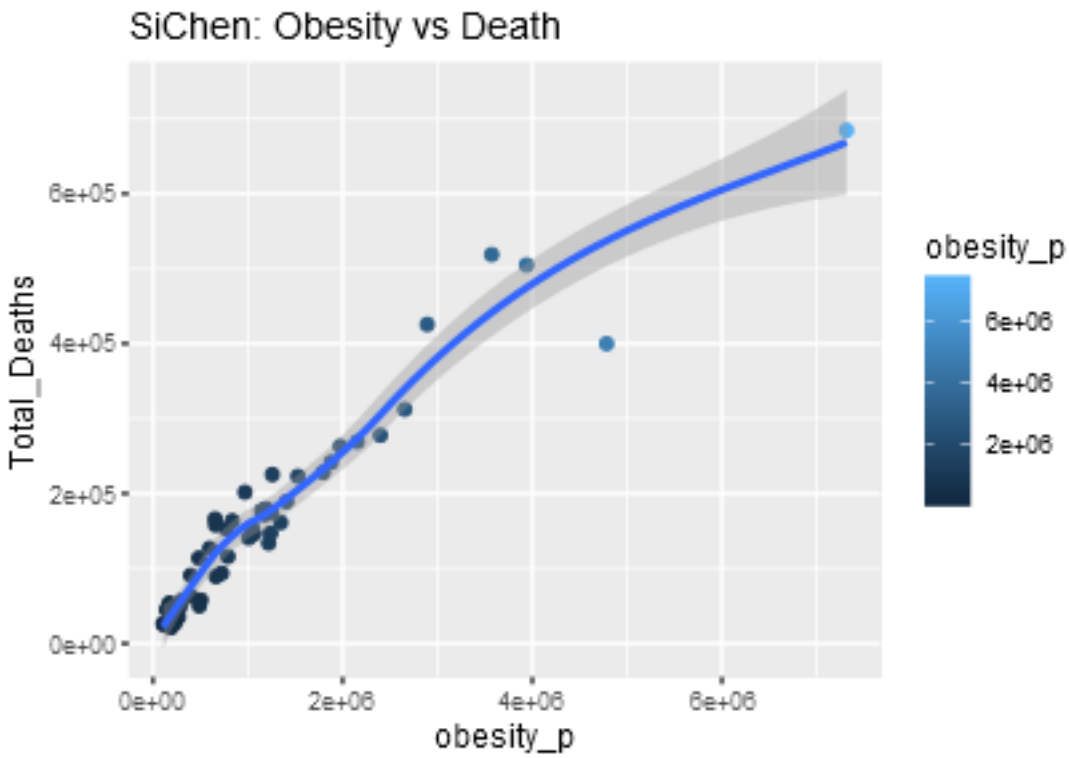
```
# heat plot
cols<-brewer.pal(4,"Spectral")
col<-colorRampPalette(cols)
heat_data<-tablenew1 %>% select(Obesity,Smoker,Diabetes,Uninsured,Total_Deaths,Population_Size) %>%
mutate(uninsured=Uninsured/Population_Size*100,Death=Total_Deaths/Population_Size*100)
heat_data<-heat_data[,-c(4,5,6)]
vc<-cor(heat_data)
levelplot(vc,col.regions=col(100),main="Cor Plot Test",scales=list(x=list(rot=90)))
```



We visualize the datasets with different units: percentage and counts, have seperately created the ggplot for each variable. More details are listed below.

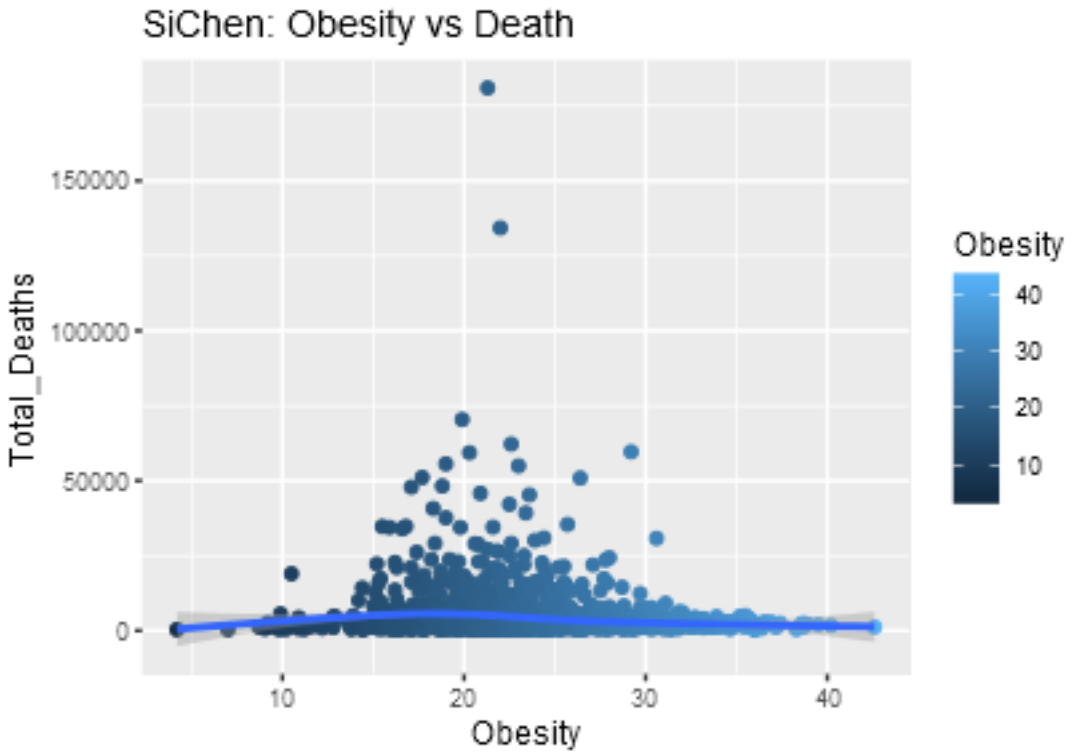
```
Obesityp <- tablenew1 %>%
  select(Obesity,Population_Size,Total_Deaths,CHSI_State_Name) %>%
  mutate(obesity_p = Obesity/100*Population_Size)

table.state1 <- Obesityp %>% group_by(CHSI_State_Name) %>% summarise(Total_Deaths = sum(Total_Deaths),ob
#tablenew2 <- cbind.data.frame(Obesityp, tablenew1)

ggplot(table.state1) + geom_point(aes(x = obesity_p, y = Total_Deaths, color = obesity_p)) + geom_smoothl
```

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



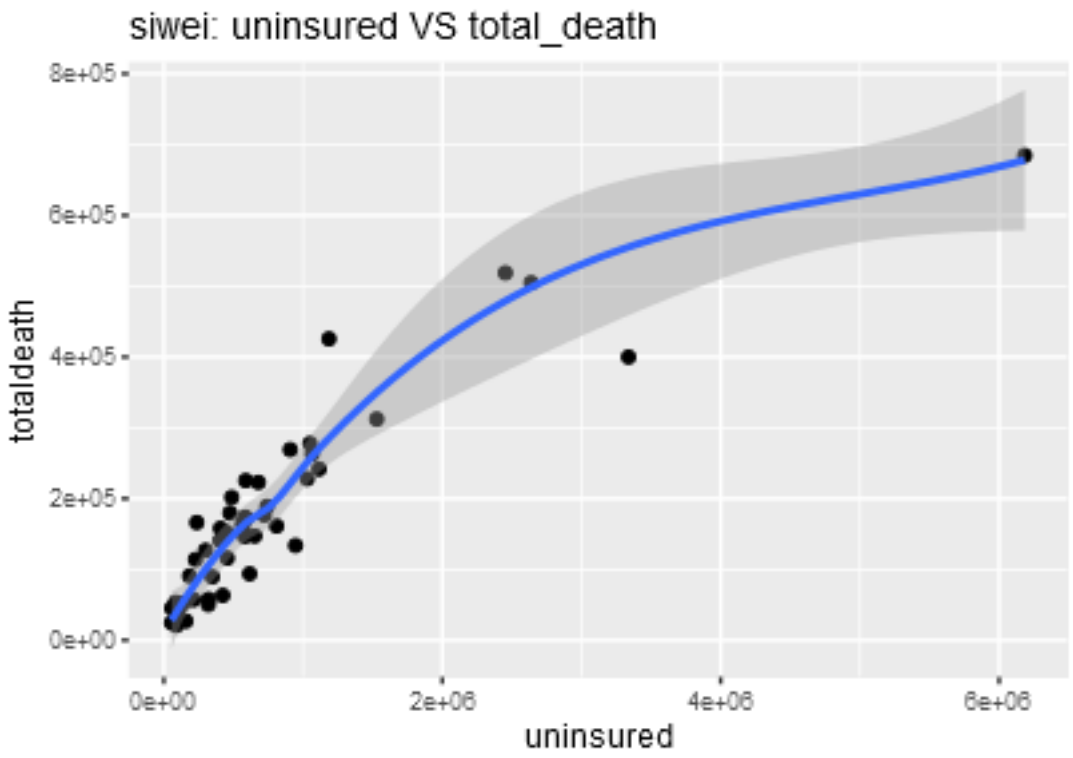

```
ggplot(tablenew1) + geom_point(aes(x = Obesity, y = Total_Deaths, color = Obesity)) + geom_smooth(aes(x
```

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



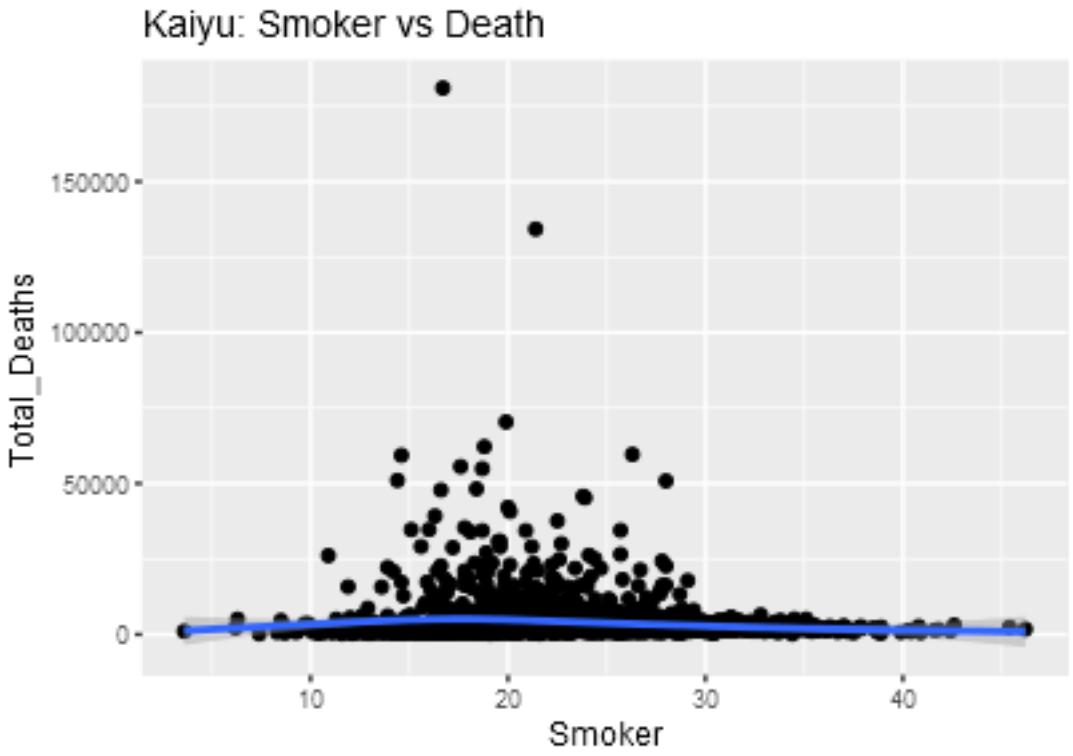
```
table.state2 <- tablenew1 %>% group_by(CHSI_State_Name) %>% summarise(totaldeath = sum(Total_Deaths),uni

ggplot(table.state2)+
  ggtitle("siwei: uninsured VS total_death")+
  geom_point(mapping = aes(x = uninsured,y = totaldeath))+
  geom_smooth(mapping = aes(x = uninsured,y = totaldeath))
```

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



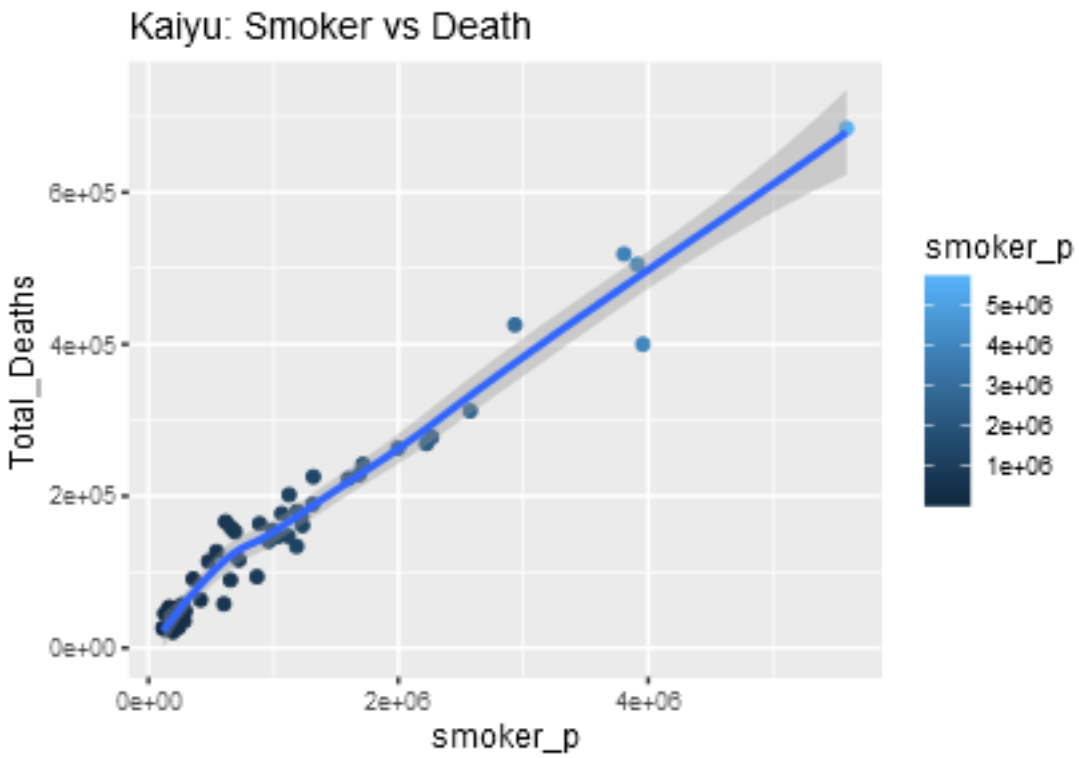

```
#plot smoker with death
smokerp <- tablenew1 %>%
  select(Smoker,Population_Size,Total_Deaths,CHSI_State_Name) %>%
  mutate(smoker_p = Smoker/100*Population_Size)

table.state3 <- smokerp %>% group_by(CHSI_State_Name) %>% summarise(Total_Deaths = sum(Total_Deaths),smo

ggplot(data = tablenew1) +
  geom_point(mapping = aes(x = Smoker, y = Total_Deaths))+ geom_smooth(mapping = aes(x = Smoker, y = To
```

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

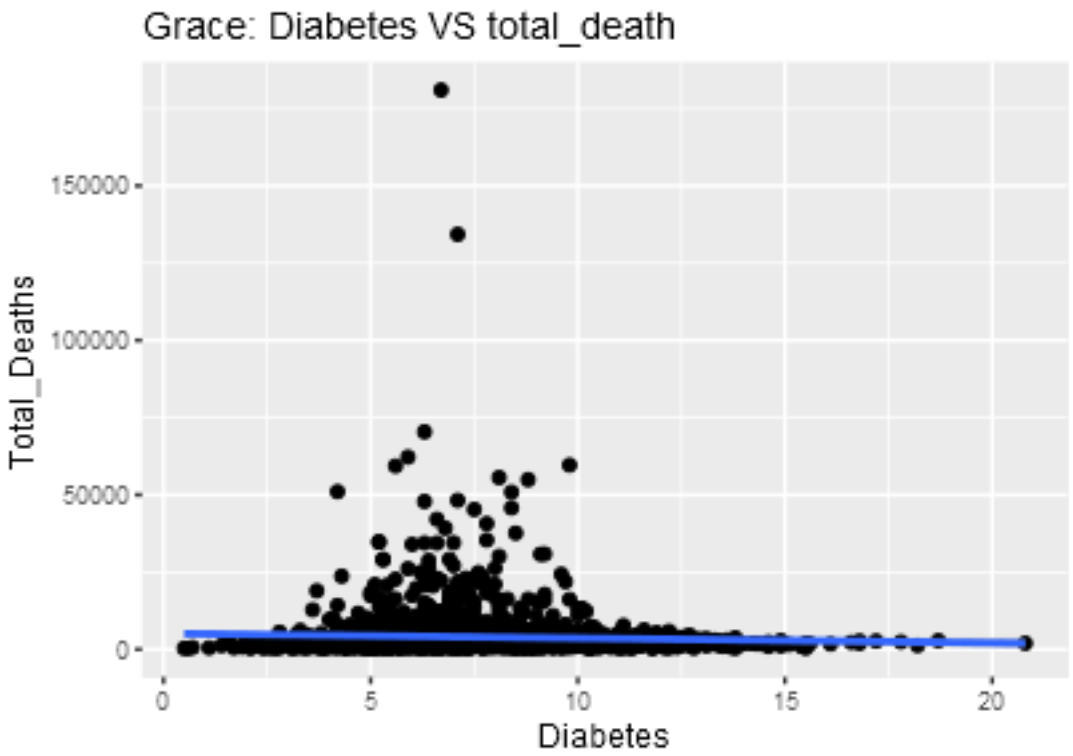



```
ggplot(table.state3,aes(x = smoker_p, y = Total_Deaths)) +
  geom_point(aes(color = smoker_p)) +
  geom_smooth() + ggtitle( "Kaiyu  Smoker vs Death")
```

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

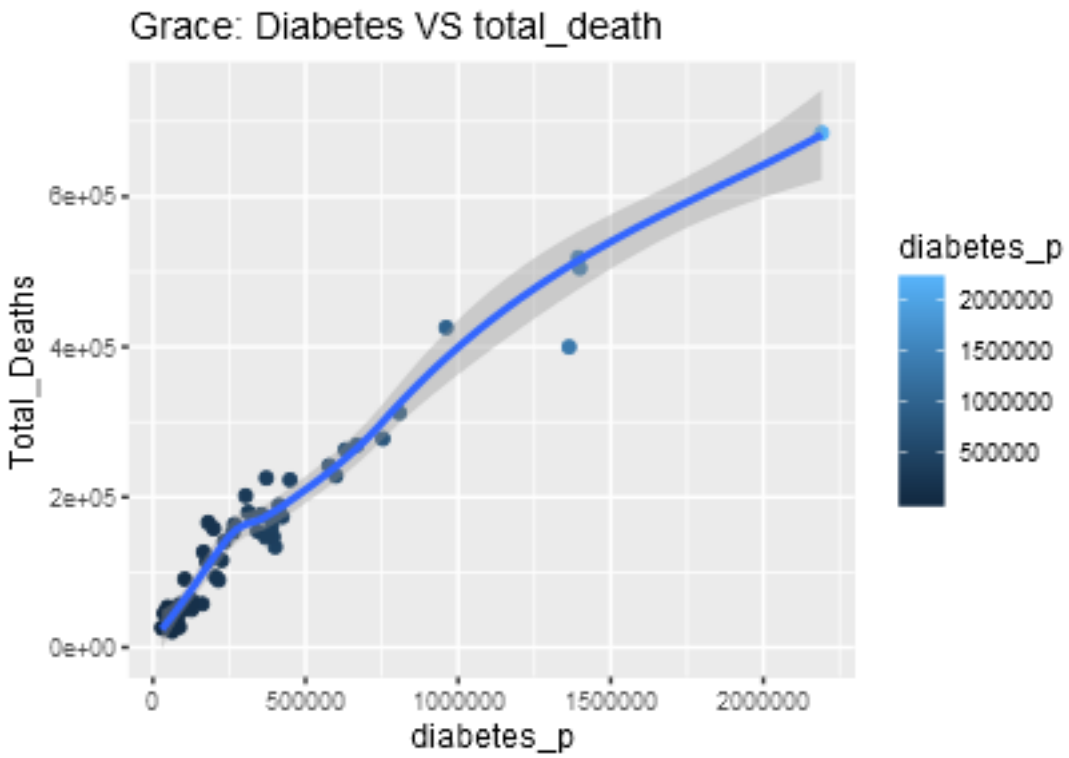



```
#plot diabetes with death
diabetesp <- tablenew1 %>%
  select(Diabetes,Population_Size,Total_Deaths,CHSI_State_Name) %>%
  mutate(diabetes_p = Diabetes/100*Population_Size)

table.state4 <- diabetesp %>% group_by(CHSI_State_Name) %>% summarise(Total_Deaths = sum(Total_Deaths),

ggplot(data = tablenew1,mapping = aes(x = Diabetes, y = Total_Deaths)) +
  geom_point()+ geom_smooth(method = "gam")+
  ggtitle("Grace: Diabetes VS total_death")
```




```
ggplot(table.state4,aes(x = diabetes_p, y = Total_Deaths)) +
  geom_point(aes(color = diabetes_p)) +
  geom_smooth() +ggtitle("Grace: Diabetes VS total_death")
```

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



In conclusion, uninsurence, diabetes, smoking and obesity have positive relation with total death counts. Among all the variables, smoking is the most obvious factor that affect the death rate. More smoking causes higher death counts.