

615 assignment 3 data cleaning

Si Chen, Grace Xie, Kaiyu yan, Siwei Hu

September 30, 2018

We collect the data from the 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, uninsured, diabetes, smoking 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 RISKFACTORSANDACCESSTOCARE
risk <- select(tableR, Obesity, Smoker, Diabetes, Uninsured, Dentist_Rate, CHSI_State_Name)
# select 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 visualize the datasets with different units: percentage and counts, have separately created the ggplot for each variable. More details are listed below.

```
library(ggplot2)

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

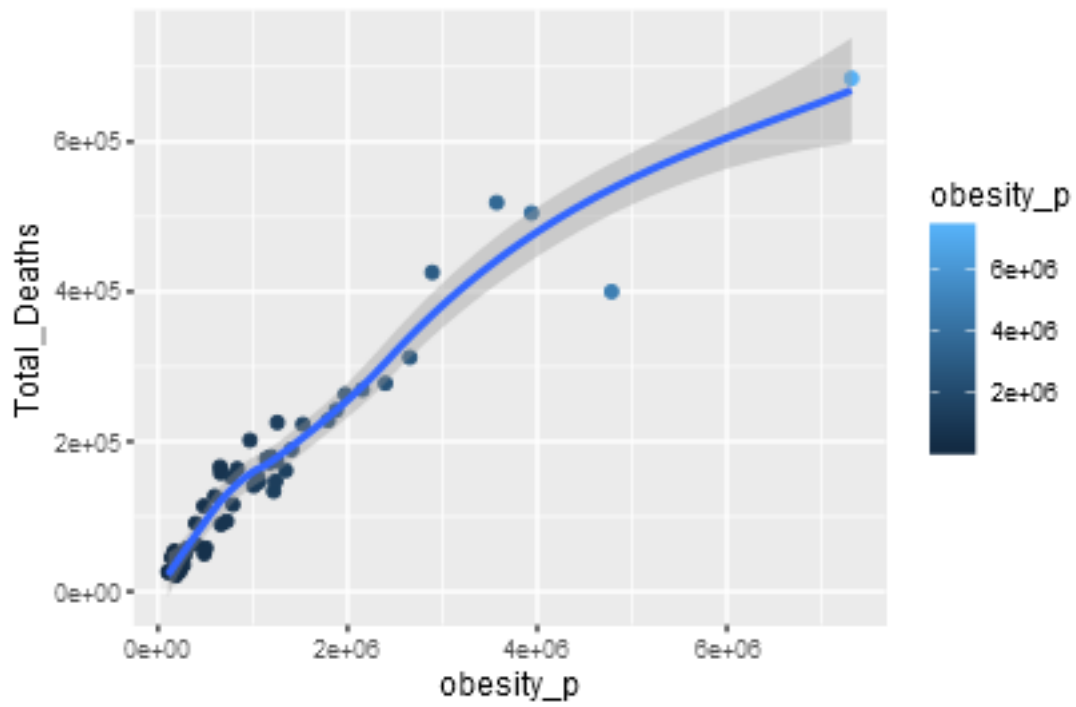
## Warning: package 'bindrcpp' was built under R version 3.4.4

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

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

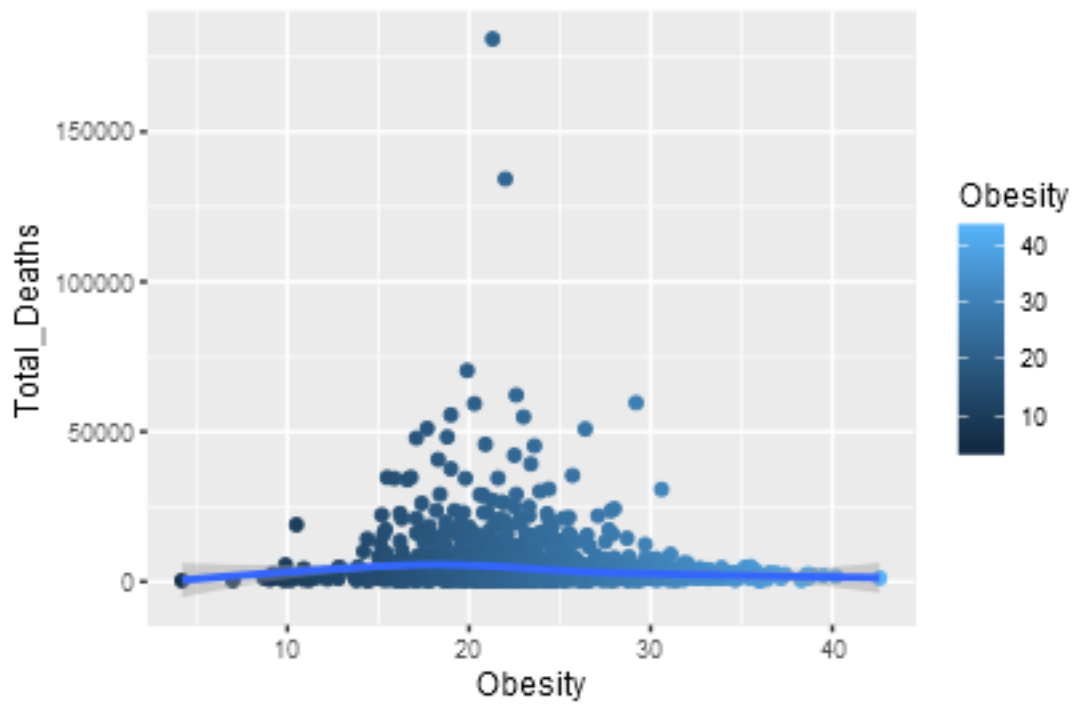
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

SiChen: Obesity vs Death



```
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")'
```

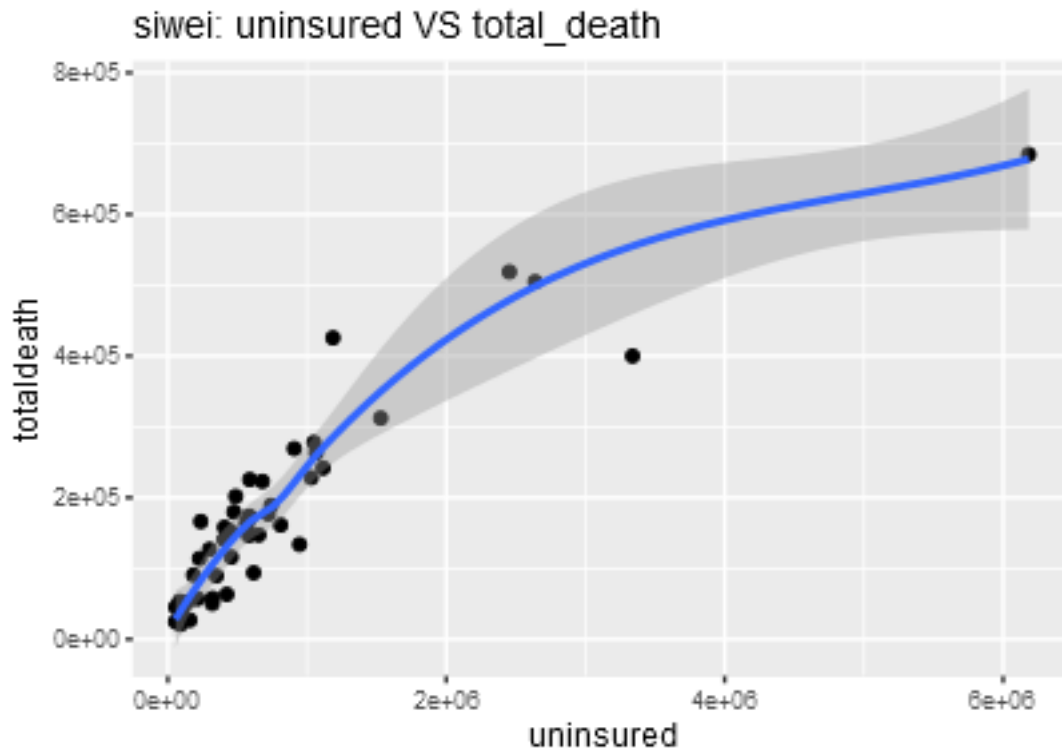
SiChen: Obesity vs Death



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

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'
```

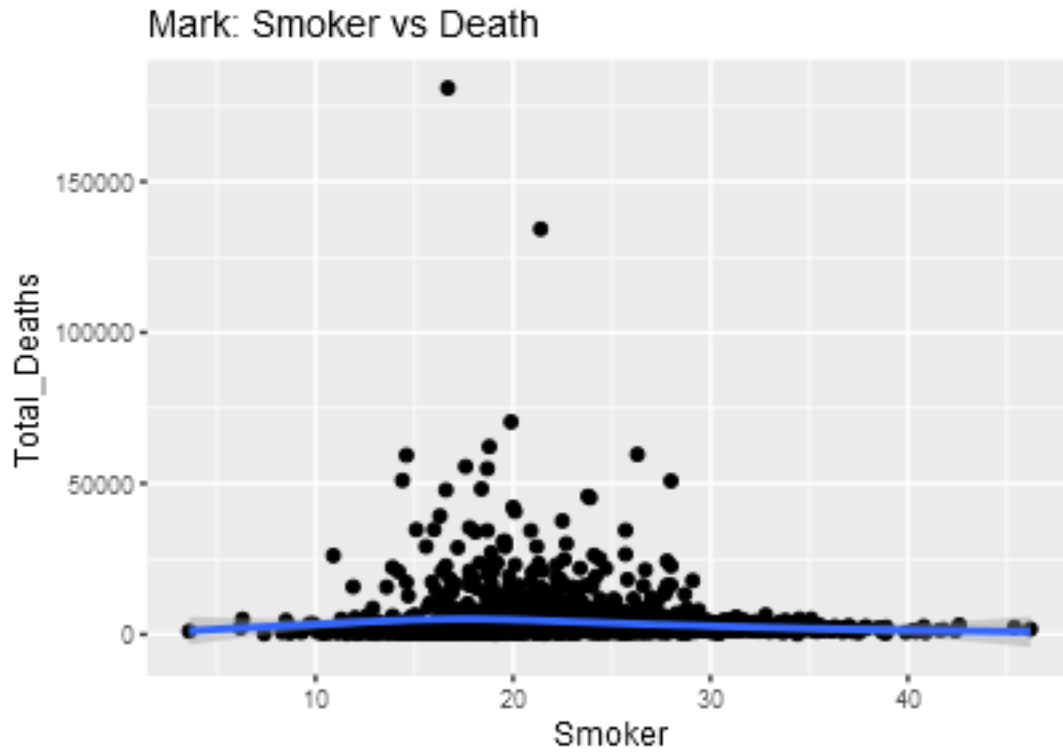


```
#plot smoker with death
smokerp <- tablenew1 %>%
  select(Smoker,Population_Size,Total_Deaths,CHSI_State_Name) %>%
  mutate(smoker_p = Smoker/100*Population_Size)

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

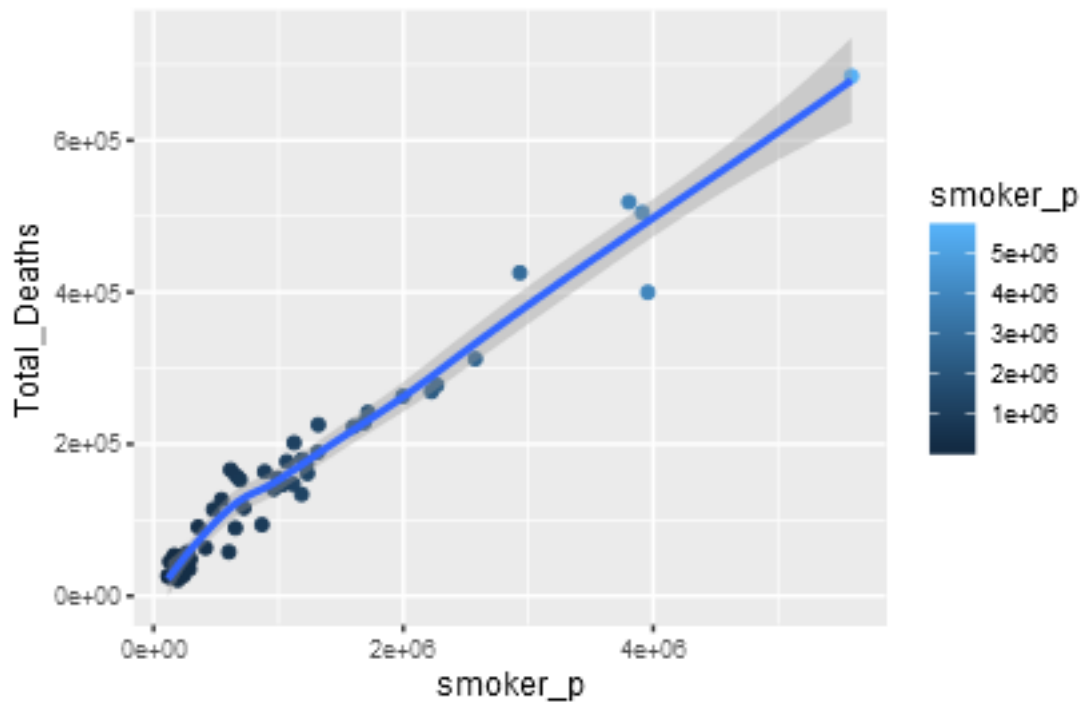
ggplot(data = tablenew1) +
  geom_point(mapping = aes(x = Smoker, y = Total_Deaths))+ geom_smooth(mapping = aes(x = Smoker, y = Total_Deaths))

## `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( "Mark Smoker vs Death")  
  
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

Mark: Smoker vs Death

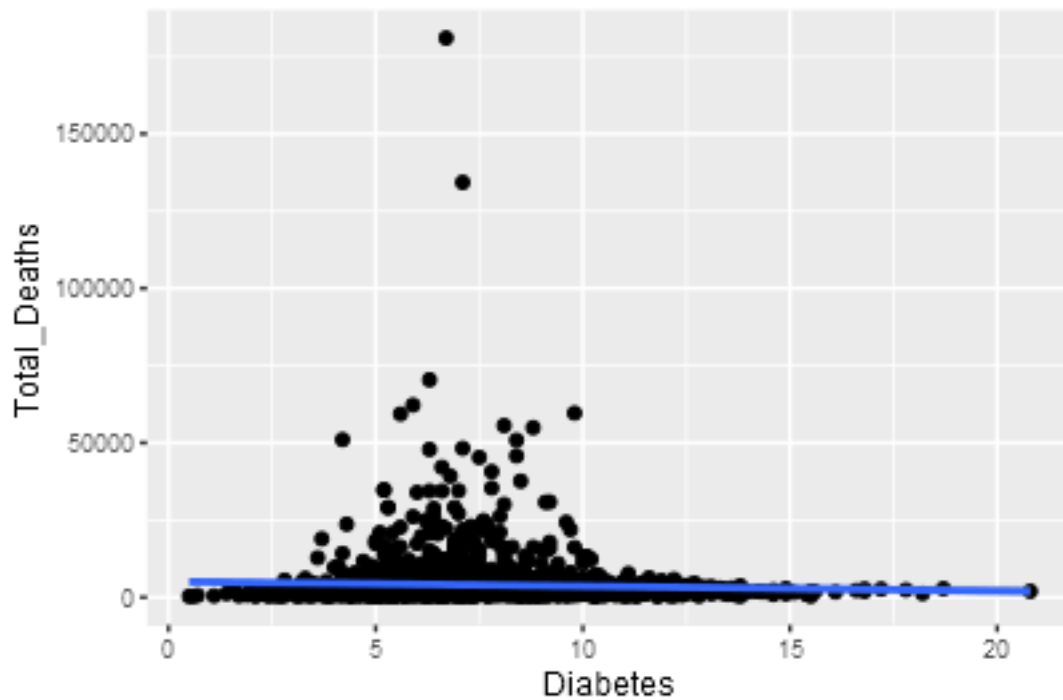


```
#plot diabetes with death
diabetesp <- tablenew1 %>%
  select(Diabetes,Population_Size,Total_Deaths,CHSI_State_Name) %>%
  mutate(diabetes_p = Diabetes/100*Population_Size)

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")
```

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, uninsurance, 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.