Empirical Methods of Data Science

Assignment: Data analysis of infant mortality rates

Amal Alabdulkarim (aa4235), Jing Qian (jq2282)

Step 1. Load data

```
library('readxl')
data = read_excel("table011.xlsx")
```

```
## Warning in strptime(x, format, tz = tz): unknown timezone 'zone/tz/2018i.
## 1.0/zoneinfo/America/New_York'
```

```
names(data)[2]<-"Year"
View(data)
```

In this table, infant mortality rates are shown in different races and years. Five levels of races are discussed: all races, race of child is white, race of mother is white, race of child is black or African American and race of mother is black or African American. The time span of the data is from 1950 to 2016. The mortality rates are shown in four catetories: infant, neonatal under 28 days, neonatol under 7 days and postneonatal. [1]

Step 2: Data exporation

2.1. Difference in Mortality rate due to the age of infants

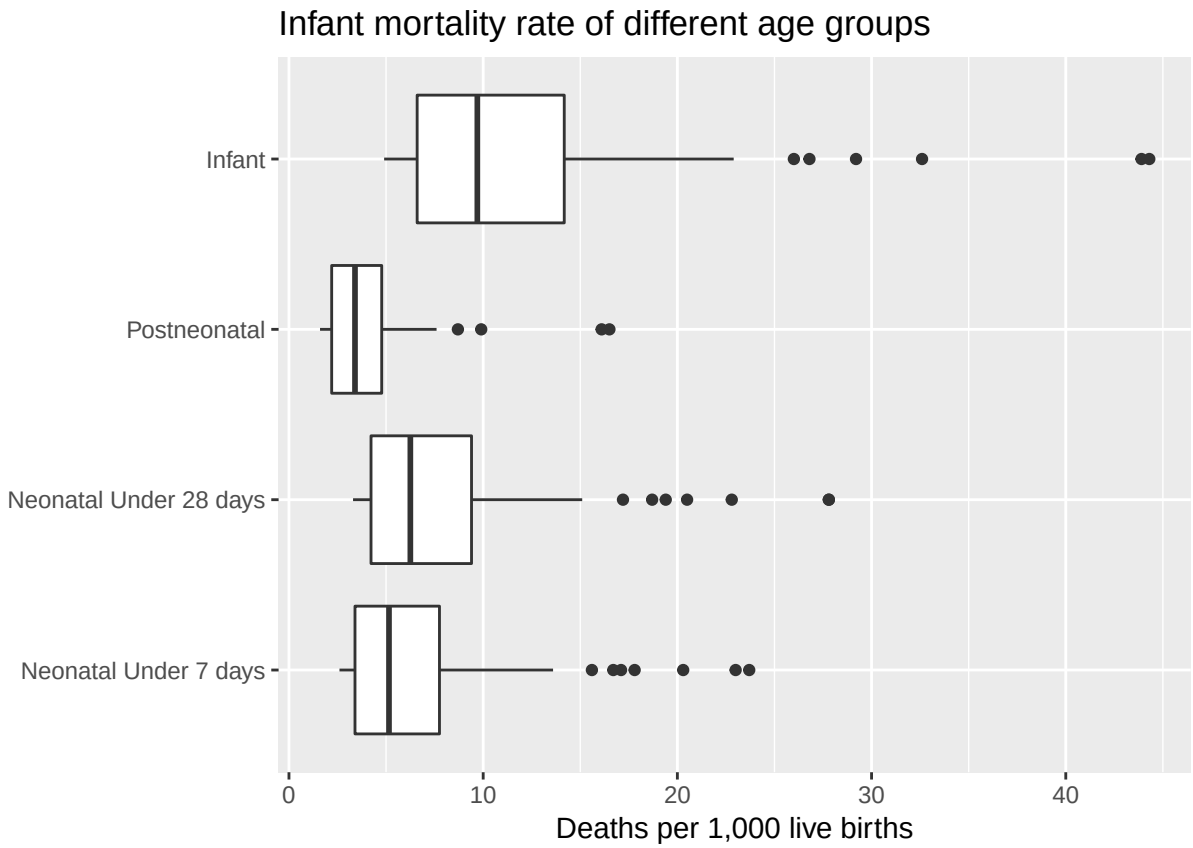
First, let's look at the mortality rate of infants in different age.

```
library("ggplot2")
```

```
## Warning: replacing previous import by 'rlang::dots_n' when loading 'dplyr'
```

```
a = data.frame(group="Infant",value=data$Infant)
b = data.frame(group="Postneonatal", value=data$Postneonatal)
c = data.frame(group="Neonatal Under 28 days", value=data$`Neonatal1 Under 28 days`)
d = data.frame(group="Neonatal Under 7 days", value=data$`Neonatal1 Under 7 days`)
plotdata = rbind(d,c,b,a)
ggplot(data = plotdata, mapping = aes(x=group,y=value))+geom_boxplot()+coord_flip()+
  labs(title="Infant mortality rate of different age groups",y="Deaths per 1,000 live births", x = "")
```

[1] Infant (aged $<$ 1 year), neonatal (aged $<$ 28 days), postneonatal (aged 28 days to 11 months). So the infant death number = the neonatal death number + the postneonatal death.

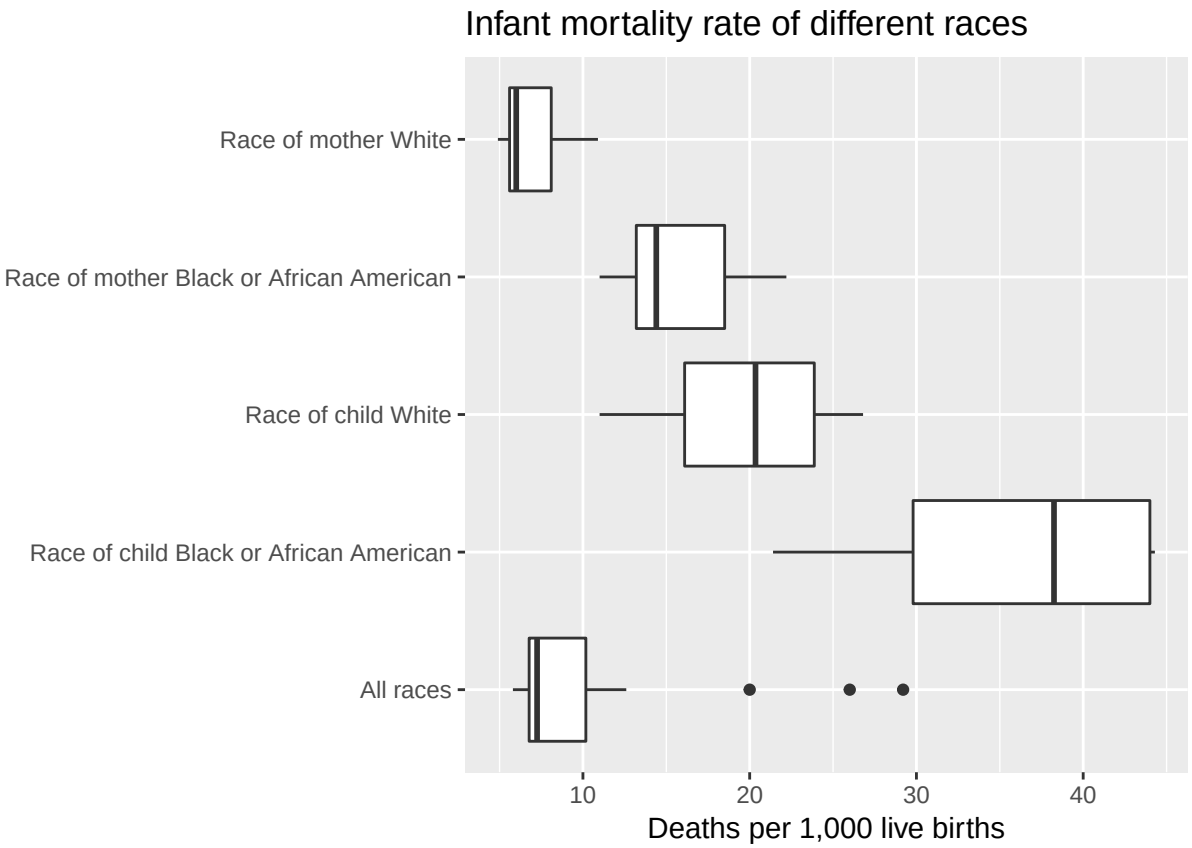



```
# Or use following boxplot command:
#boxplot(data$Infant, data$`Neonatal1 Under 28 days`, data$`Neonatal1 Under 7 days`, data$Postneonatal,
```

From the boxplot of infant mortality rate above, we could see that the death rate of postneonatals has smaller value and narrower distribution than that of neonatals.

2.2. Difference in Mortality rate due to race

```
ggplot(data = data, mapping = aes(x = data$race, y=data$Infant))+geom_boxplot()+coord_flip()+
  labs(title="Infant mortality rate of different races",y="Deaths per 1,000 live births", x = "")
```

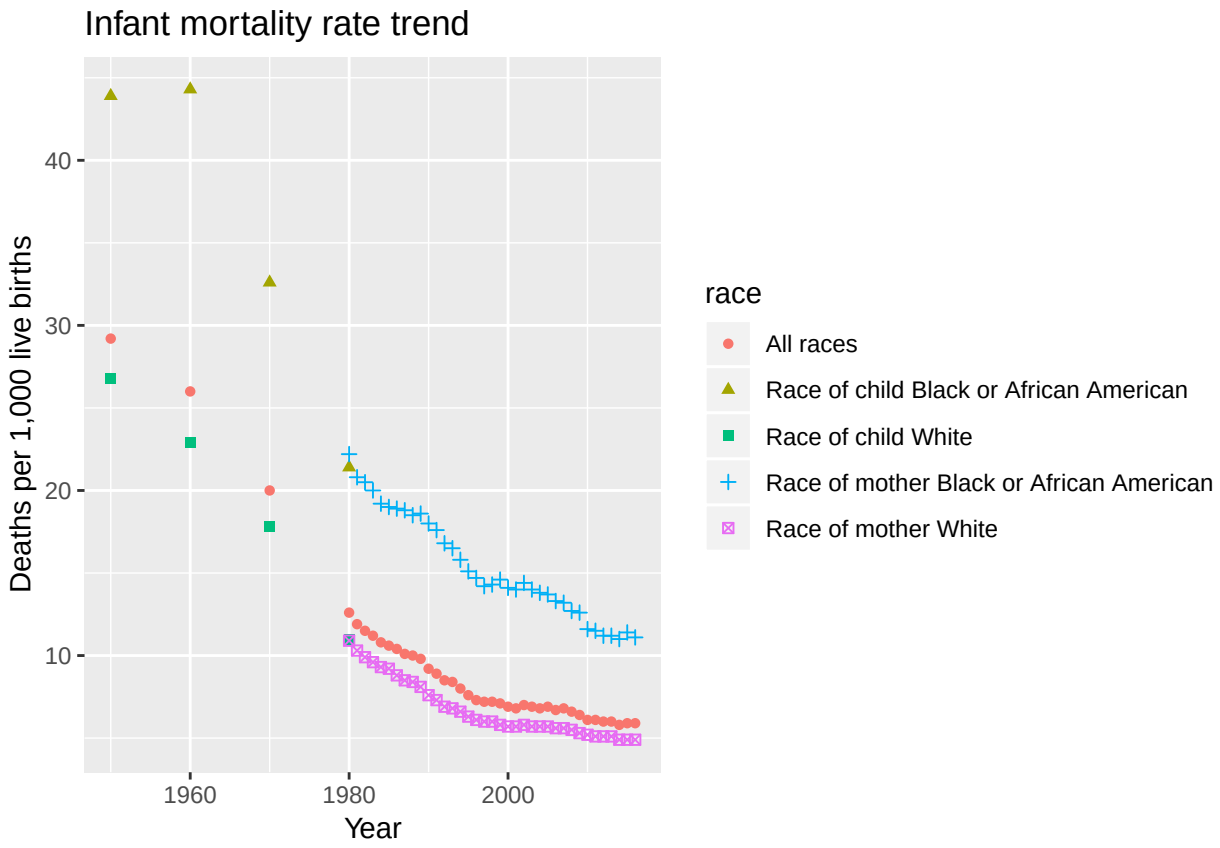


From the boxplot of infant mortality rate of different races, we could find that: 1) The white infants have lower death rate than black or African American infants. 2) The infants born by white mothers have lower death rate than that by black or African American mothers. 3) The death rate of black or African American infants is the highest among the five race groups and has the widest expansion.

It is reasonable to suggest there might be some correlation between the race of mother and that of child. The large difference between the death rate of black or African American infants and that of infants with black or African American mothers seems a little weird. This may due to the fact that data of different race groups are collected in different years.

2.3. Difference in Mortality rate due to time

```
ggplot(data, aes(x=Year, y=Infant, shape=race, color=race)) +
  geom_point()+
  labs(title="Infant mortality rate trend",y="Deaths per 1,000 live births")
```



From the scatter plot above, we could see that the infant mortality rate decreases with time and the decrease trend becomes flatter with time. And the trend is similar for different races while the infant mortality rate of white mother (or child) is less than that of black or African American mother (or child).

2.4. Infant mortality rate correlation between different races

```
cor.test(data[data$race=='Race of mother White',]$Infant, data[data$race=='Race of mother Black or Afri
```

```
##
##  Pearson's product-moment correlation
##
## data:  data[data$race == "Race of mother White", ]$Infant and data[data$race == "Race of mother Black
## t = 24.7592, df = 35, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9470662 0.9859267
## sample estimates:
##       cor
## 0.9726198
```

```
cor.test(data[(data$race=='All races') & (data$Year>1979) ,]$Infant, data[data$race=='Race of mother Bla
```

```
##
##  Pearson's product-moment correlation
##
## data:  data[(data$race == "All races") & (data$Year > 1979), ]$Infant and data[data$race == "Race of
## t = 33.0071, df = 35, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
```

```
## 95 percent confidence interval:
##  0.9695058 0.9919607
## sample estimates:
##       cor
## 0.9843141
```

```
cor.test(data[data$race=='Race of child White',]$Infant, data[data$race=='Race of child Black or Africa
```

```
##
##  Pearson's product-moment correlation
##
## data:  data[data$race == "Race of child White", ]$Infant and data[data$race == "Race of child Black
## t = 5.5312, df = 2, p-value = 0.03117
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1124995 0.9993720
## sample estimates:
##       cor
## 0.9688344
```

In this part, we examine the correlation between infant mortality rate of different races: race of mother white vs. race of mother black or African American, race of mother white vs. all races, race of child white vs. race of child black or African American. From the result above, we could see that all three correlations are significant and high, almost linear. This correlation result agrees with the trend plot we have in Part 2.3.