Data Wrangling

Adhi Roka

11/5/2020

1.a)

```
shoot<-read_csv("fatal-police-shootings-data.csv",
                col_names=TRUE, na=c(".", "NA", "", "?"))
```

```
## Parsed with column specification:
## cols(
##   id = col_double(),
##   name = col_character(),
##   date = col_date(format = ""),
##   manner_of_death = col_character(),
##   armed = col_character(),
##   age = col_double(),
##   gender = col_character(),
##   race = col_character(),
##   city = col_character(),
##   state = col_character(),
##   signs_of_mental_illness = col_logical(),
##   threat_level = col_character(),
##   flee = col_character(),
##   body_camera = col_logical()
## )
```

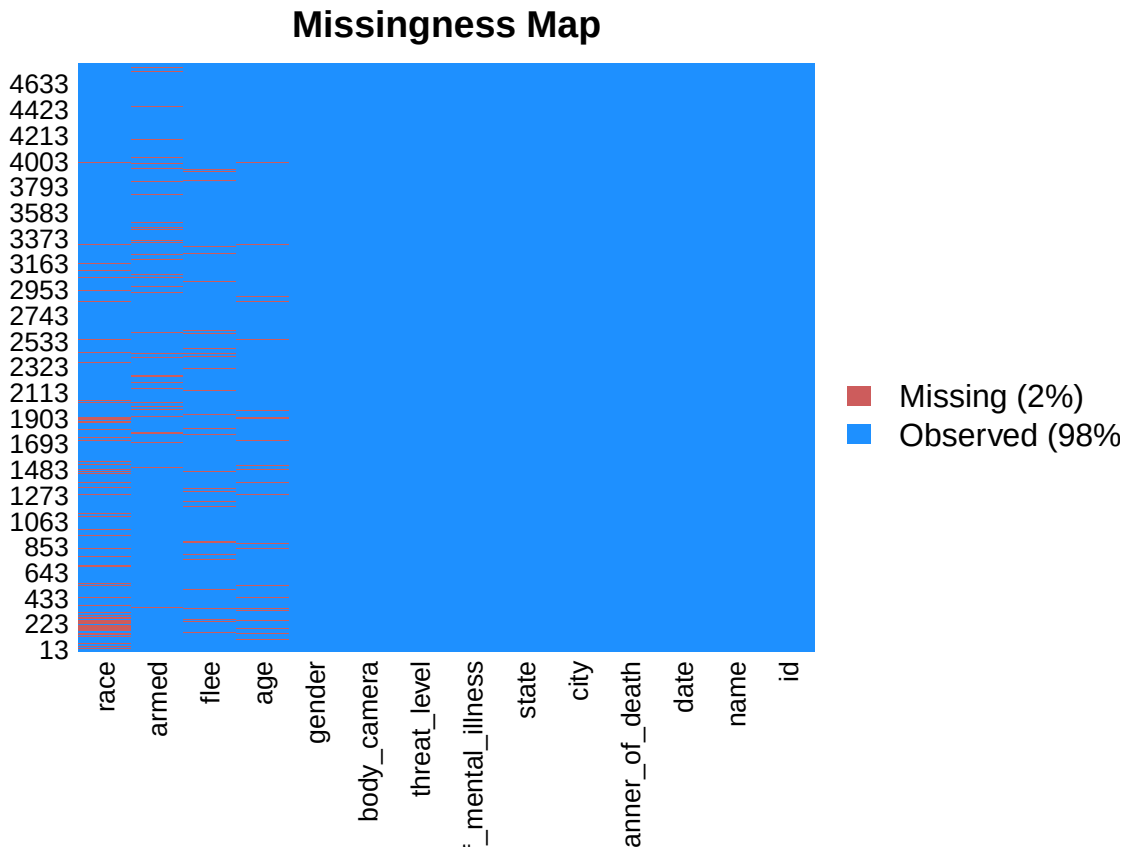
```
#glimpse(shoot)
#Check missing values
missmap(shoot) #Variables race, armed, flee and age having missing values.
```

```
## Warning in if (class(obj) == "amelia") {: the condition has length > 1 and only
## the first element will be used
```

```
## Warning: Unknown or uninitialised column: 'arguments'.
```

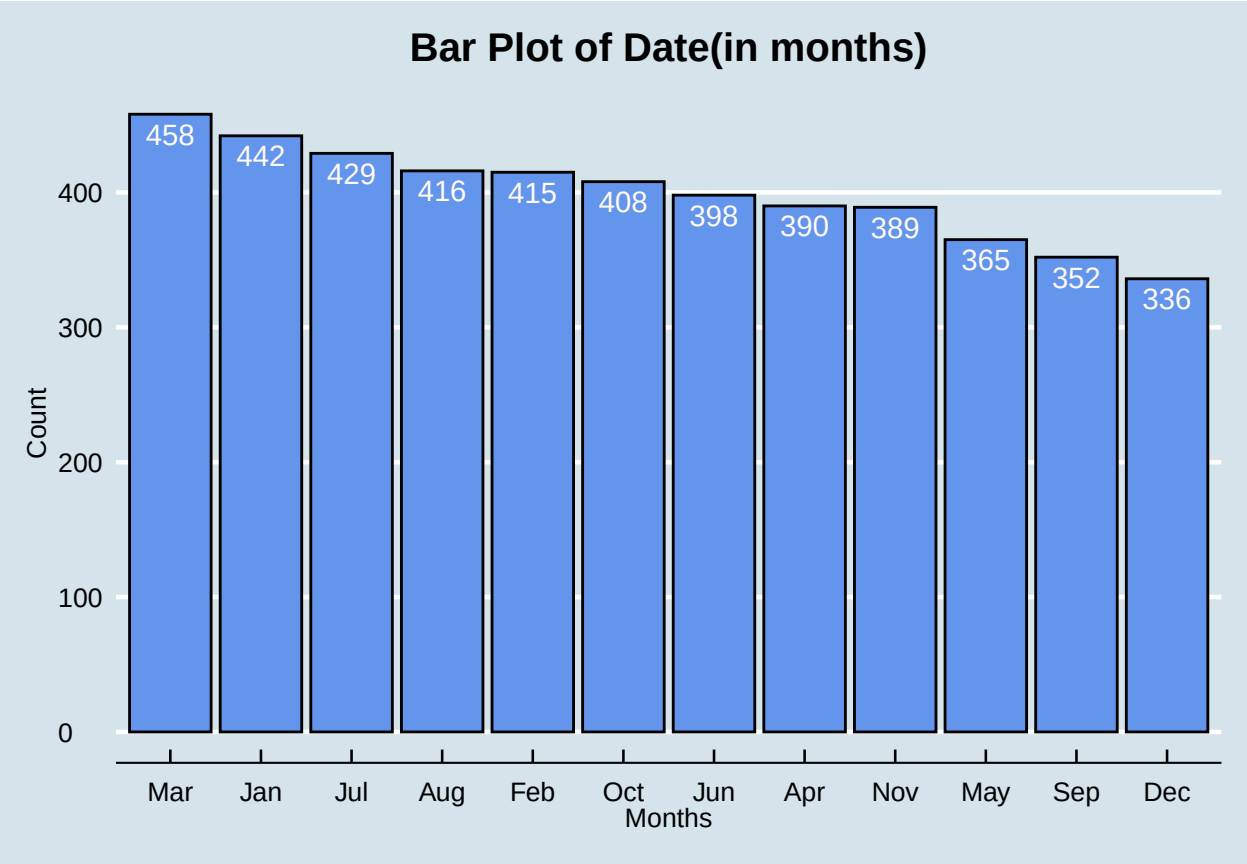
```
## Warning: Unknown or uninitialised column: 'arguments'.
```

```
## Warning: Unknown or uninitialised column: 'imputations'.
```

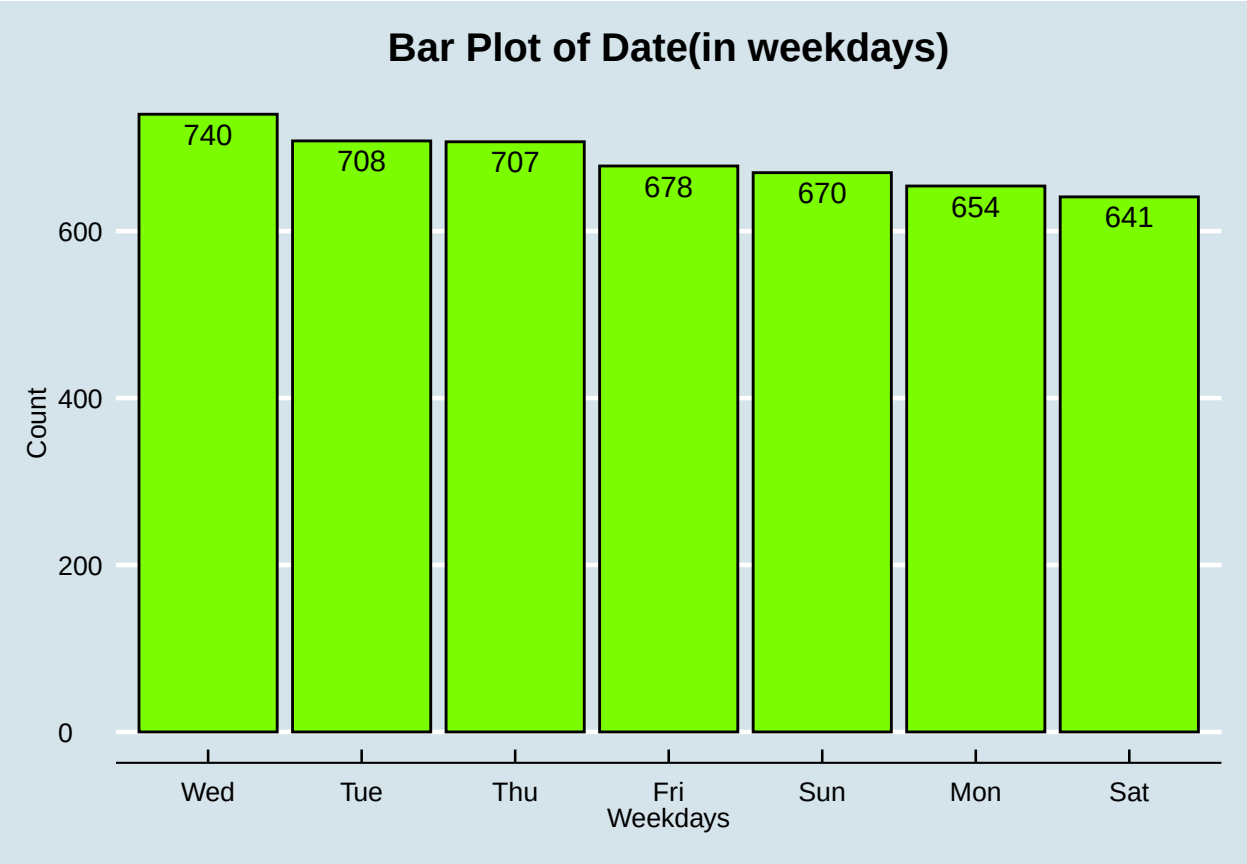



```
#Date
#reorder(month(shoot$date,label=T))

#Months
shoot %>%
  ggplot(aes(fct_infreq(month(date,label=T))))+
  geom_bar(fill="cornflowerblue",color="black")+theme_economist()+
  labs(title = "Bar Plot of Date(in months)",x="Months",y="Count")+
  theme(plot.title = element_text(hjust = .5))+
  geom_text(stat = 'count',aes(label=..count..),
            vjust = 1.45,color="floralwhite")
```

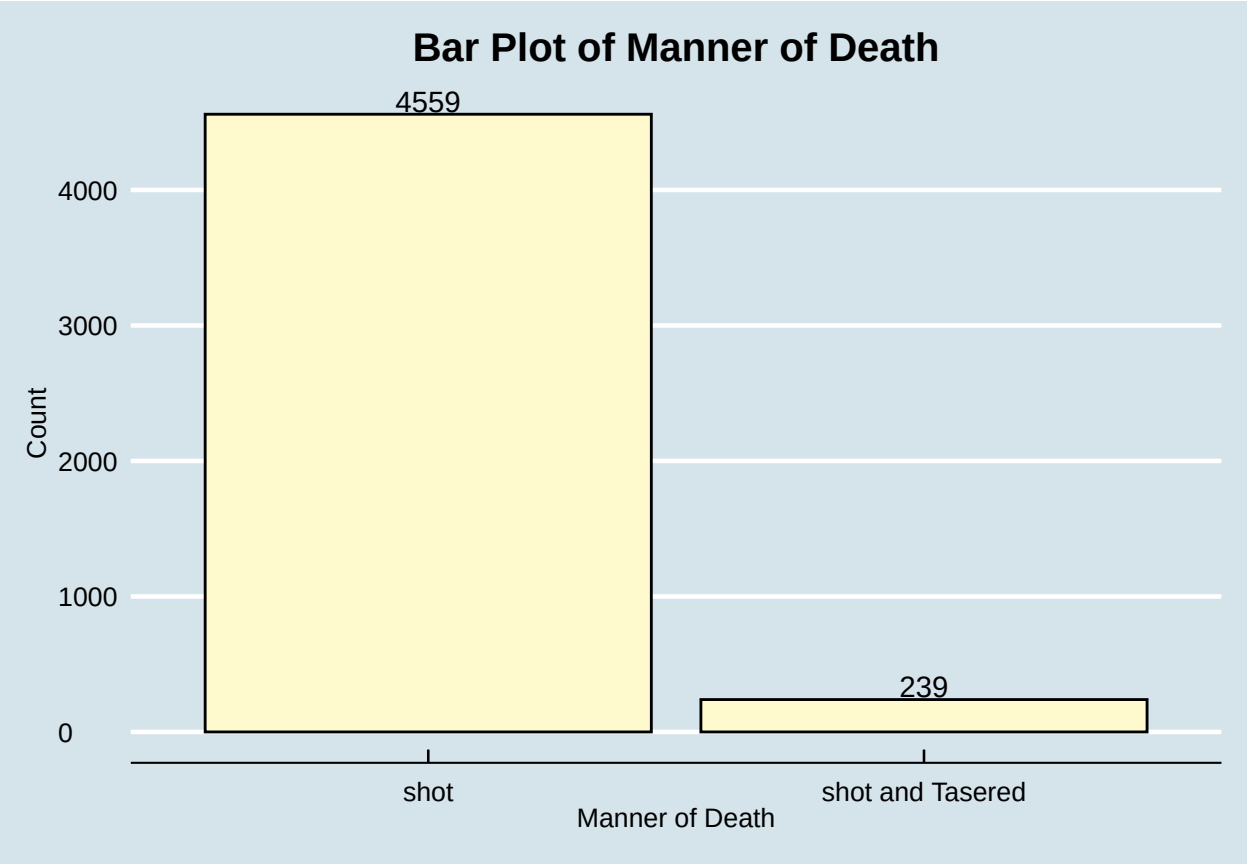



```
#March had the highest count of 458 and December had the lowest count of 336.
#Weekdays
shoot %>%
  ggplot(aes(fct_infreq(wday(date,label=T))))+
  geom_bar(fill="lawngreen",color="black")+theme_economist()+
  labs(title = "Bar Plot of Date(in weekdays)",x="Weekdays",y="Count")+
  theme(plot.title = element_text(hjust = .5))+
  geom_text(stat = 'count', aes(label=..count..),
            vjust = 1.45,color="black")
```



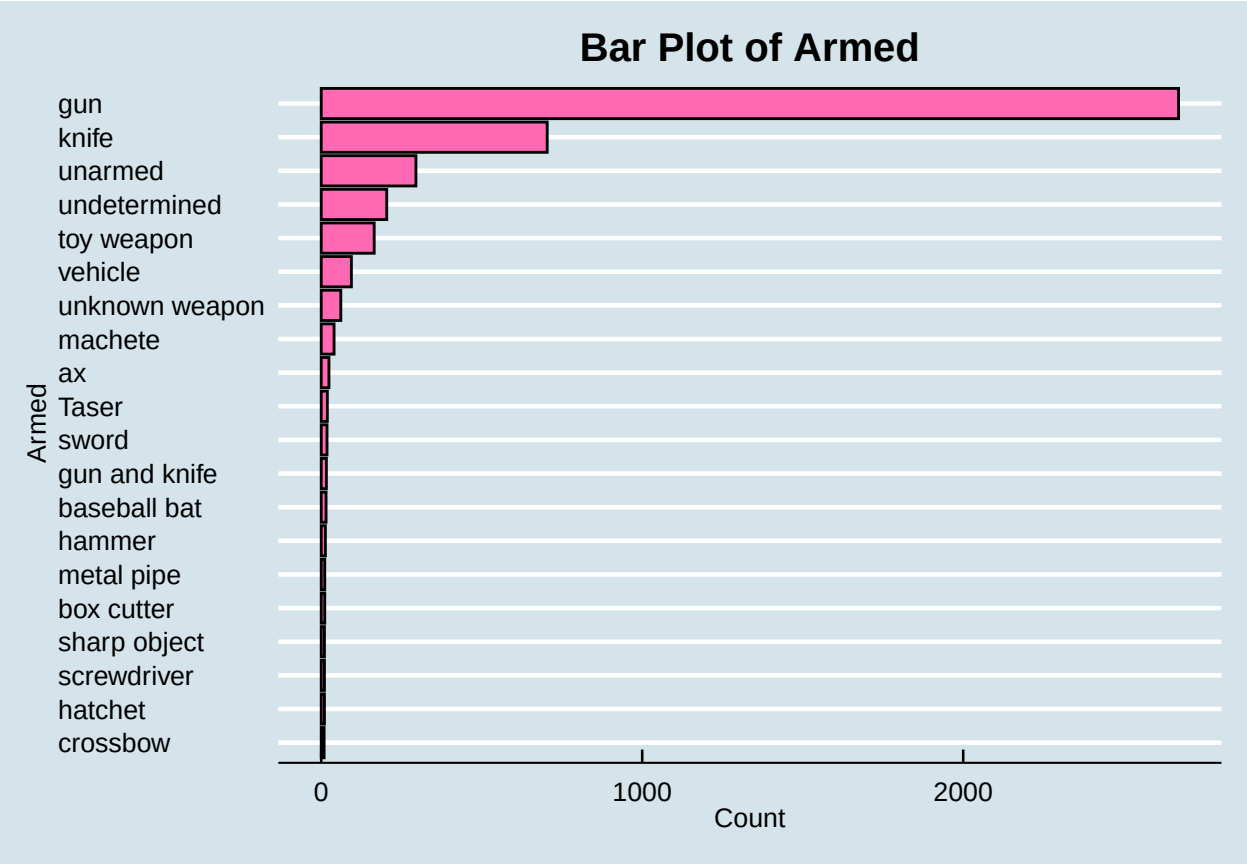

```
#Wednesday had the highest count of 740 and saturday had the lowest count of 641.

#Manner of Death
shoot %>%
  ggplot(aes(manner_of_death))+
  geom_bar(fill="lemonchiffon1",color="black")+theme_economist()+
  labs(title = "Bar Plot of Manner of Death",
       x="Manner of Death",y="Count")+
  theme(plot.title = element_text(hjust = .5))+
  geom_text(stat = 'count', aes(label=..count..),
            vjust = -0.1,color="black")
```




```
#Out of the 2 categories, manner of death=shot was the larger one.

#Armed
#Dropped NAs. Displayed only the top 20 categories with highest counts.
shoot %>%
  drop_na(armed) %>%
  group_by(armed) %>%
  summarise(count=n()) %>%
  arrange(desc(count)) %>%
  slice(1:20) %>%
  ggplot(aes(reorder(armed,count),count))+
  geom_col(fill="hotpink",color="black")+theme_economist()+
  labs(title = "Bar Plot of Armed",
       x="Armed",y="Count")+
  theme(plot.title = element_text(hjust = .5))+
  coord_flip()
```



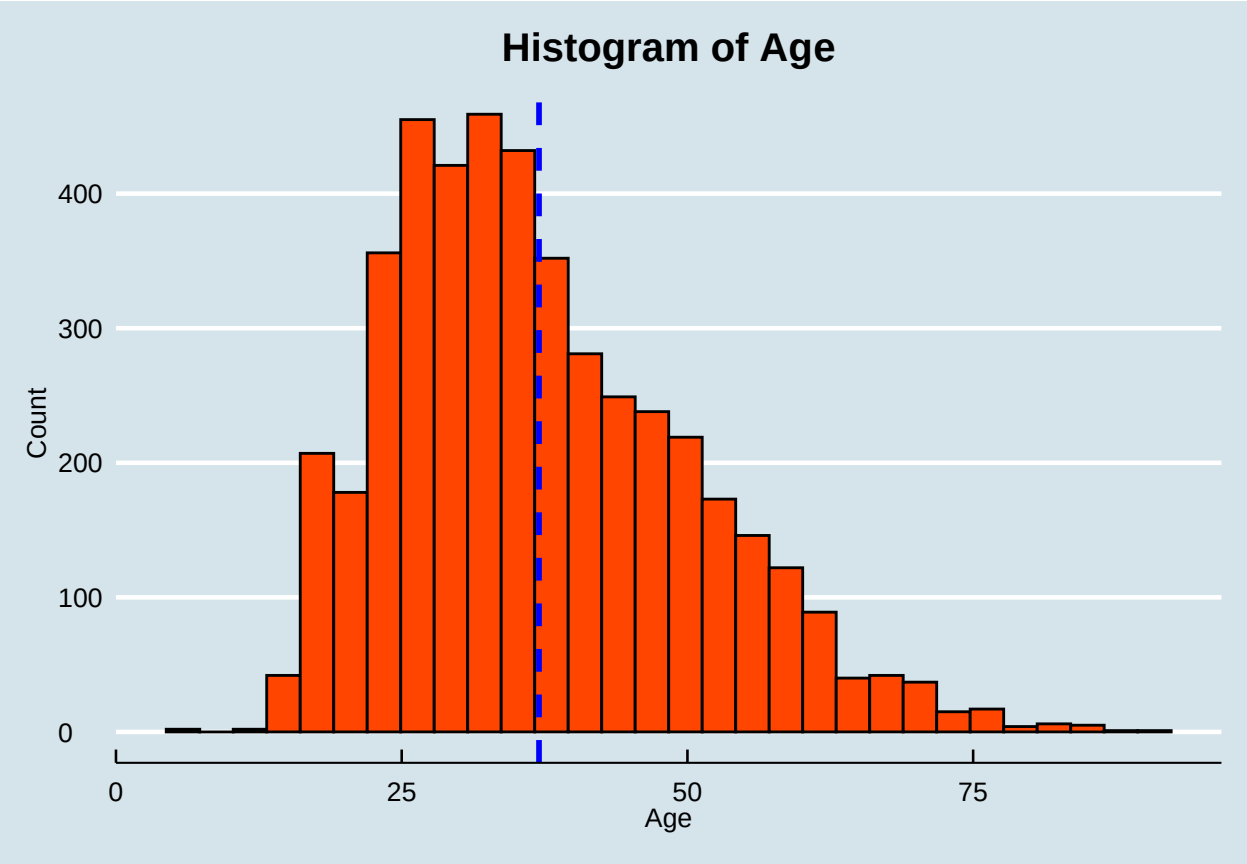

```
#Clearly out of all the categories, gun has the highest count, followed by knife. Most of the categorie

#summary(as.factor(shoot$armed))
#count(shoot$gender,summary)

#Age
shoot %>%
  ggplot(aes(age))+
  geom_histogram(fill="orangered",color="black")+
  theme_economist()+
  labs(title = "Histogram of Age",
       x="Age",y="Count")+
  theme(plot.title = element_text(hjust = .5))+geom_vline(aes(xintercept=mean(age, na.rm=T)),
            color="blue", linetype="dashed", size=1) #blue vertical line is mean age
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 207 rows containing non-finite values (stat_bin).
```




```
boxplot(shoot$age, plot=T,col = "orangered")$out
```

```
##  [1] 75 77 83 76 86 76 77 76 76 80 82 86 76 83 91 79 75 82 76 78 84 84 81 89 77
## [26] 76 84 81 80 77 76 76 77 75
```

```
skewness(shoot$age,na.rm = T)
```

```
## [1] 0.7062273
## attr(,"method")
## [1] "moment"
```

```
#Age is positively skewed. It has outliers which are possibly making the right tail longer.
```