lab0

Omer Ronen

9/1/2020

- Loading USArrests data

```
data("USArrests")
```

- Loading coords

```
fileName <- 'data/stateCoord.txt'
coords_txt <- readChar(fileName, file.info(fileName)$size)

extract_city <- function(s){
  city <- strsplit(s, ' ')[[1]][1]
  city <- gsub('-', ' ', city)
  return(city)
}

extract_long <- function(s){
  str_split <- strsplit(s, ' ')[[1]]

  return(as.numeric(str_split[length(str_split)-1]))
}

extract_lan <- function(s){
  str_split <- strsplit(s, ' ')[[1]]

  return(as.numeric(str_split[length(str_split)]))
}


lines <- strsplit(coords_txt, '\n')[[1]]


cities <- unlist(lapply(lines[2:length(lines)], FUN = extract_city))
long <- unlist(lapply(lines[2:length(lines)], FUN = extract_long))
lan <- unlist(lapply(lines[2:length(lines)], FUN = extract_lan))

coords = data.frame(long=long, lan=lan)
rownames(coords) = cities
```

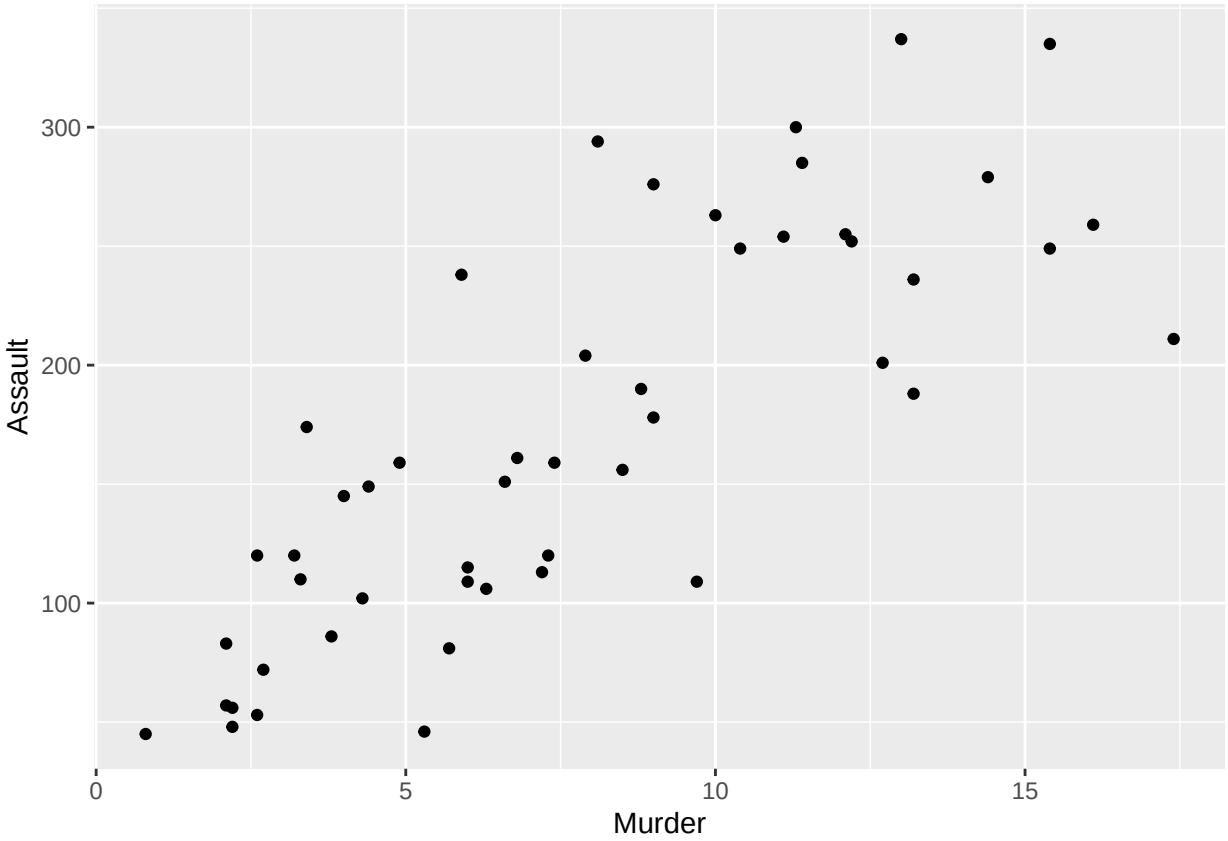
Manipulating the data

- Merging datasets

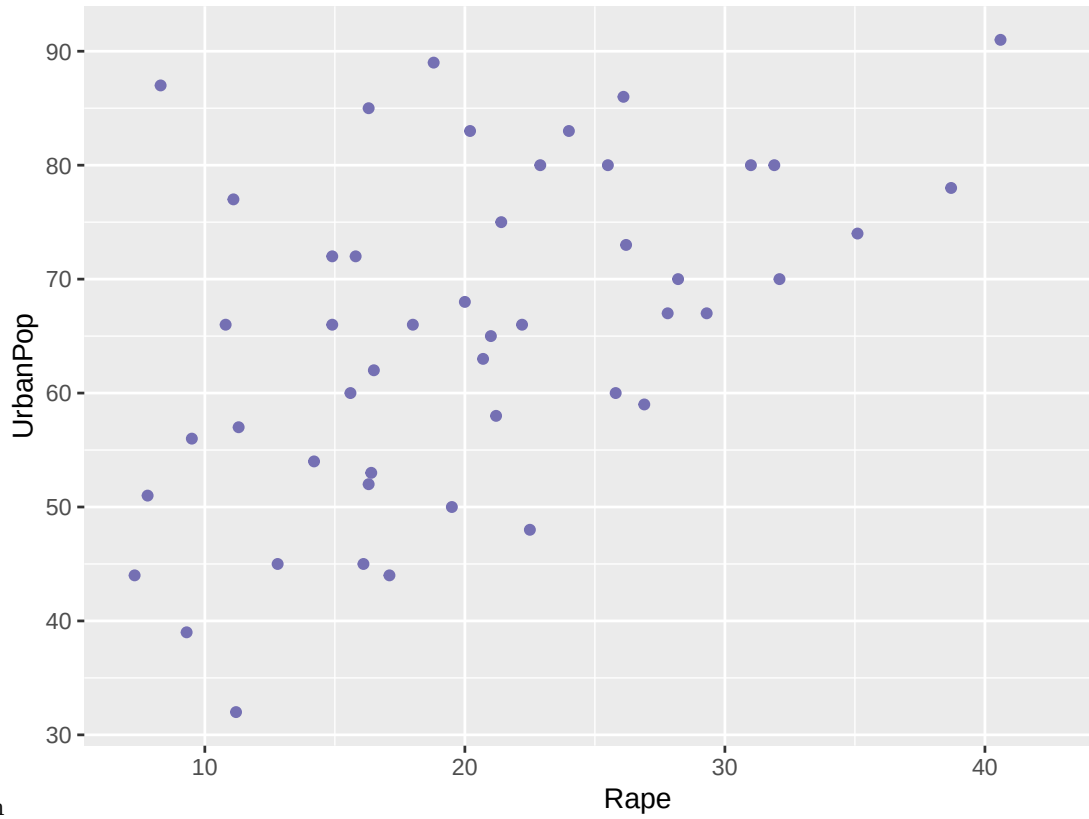
```
arr <- tibble::rownames_to_column(USArrests, "region")
coo <-  tibble::rownames_to_column(coords, "region")


arrests = dplyr::full_join(arr, coo,by="region")
```

Visualizing the data

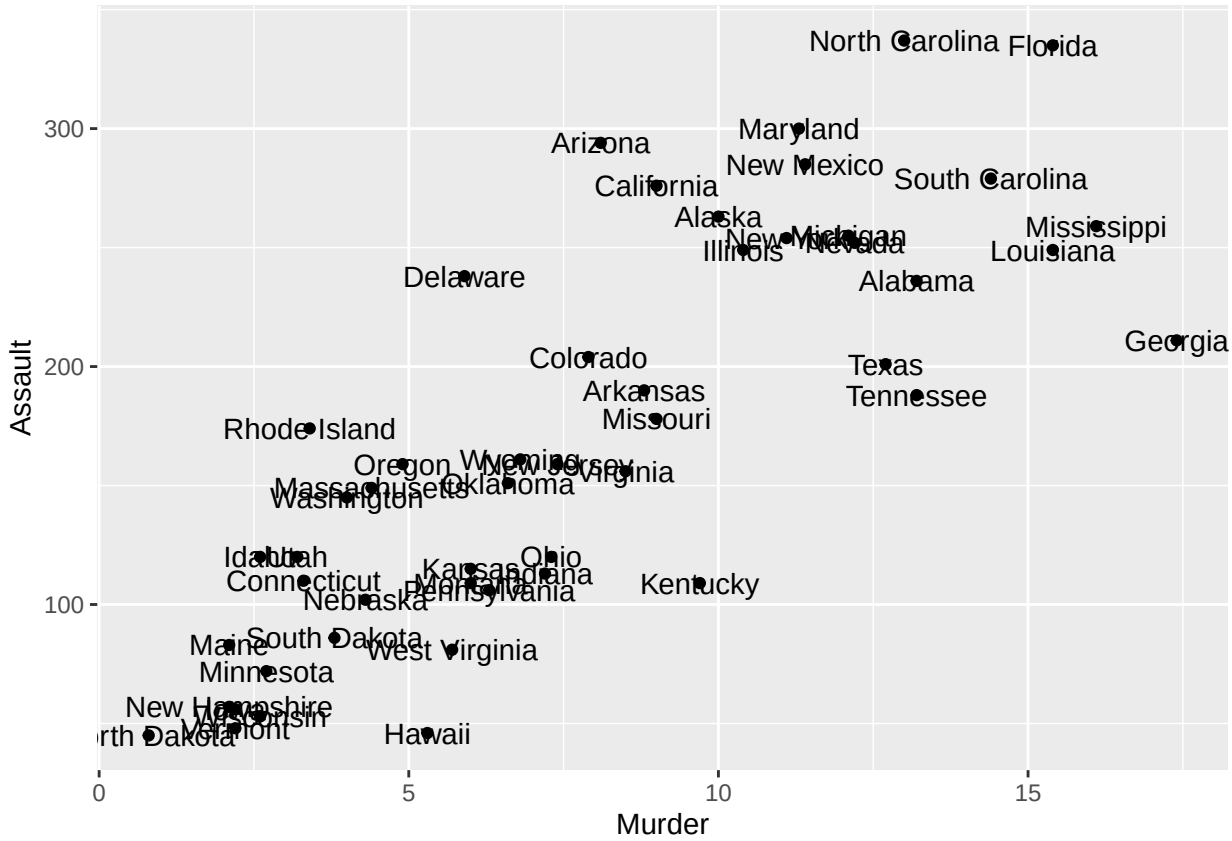
- Murder vs Assault

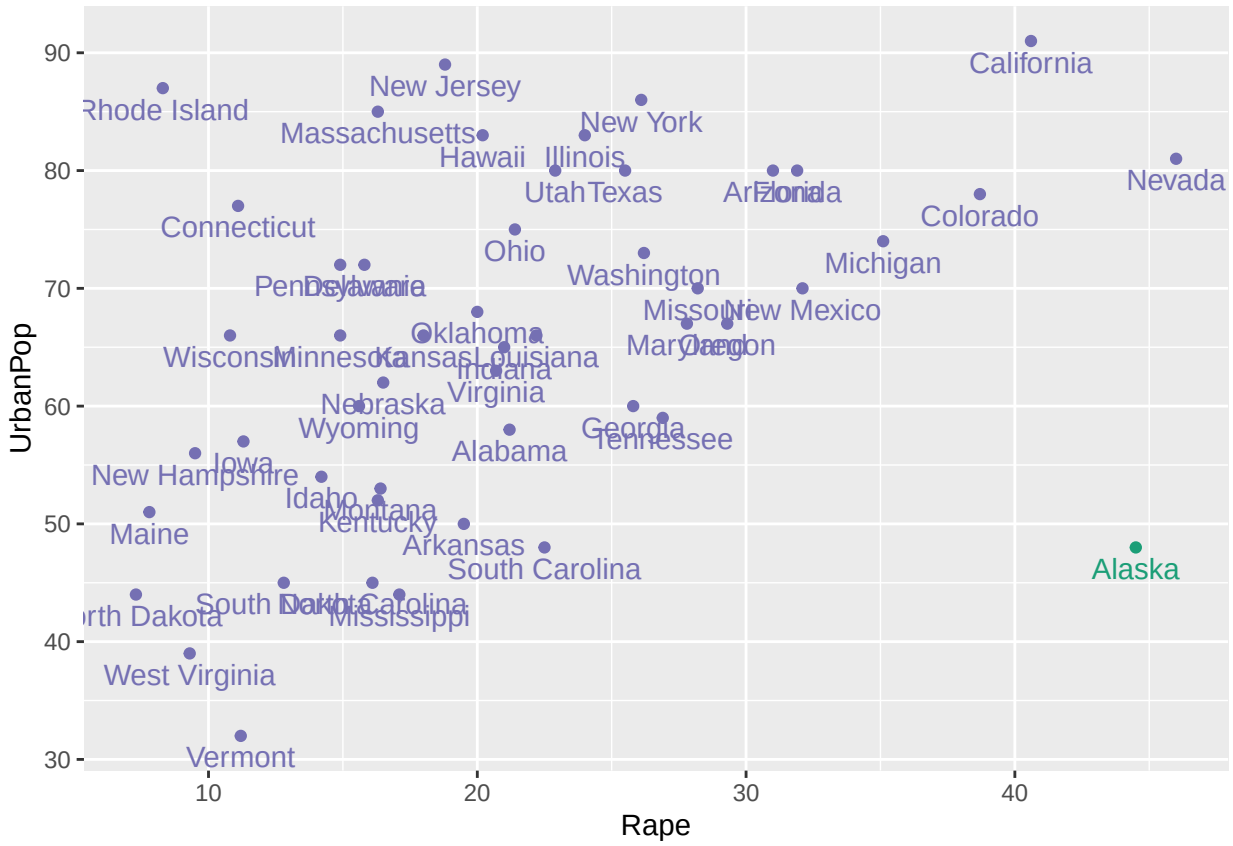


It seems like there is a linear connection between murder and assault

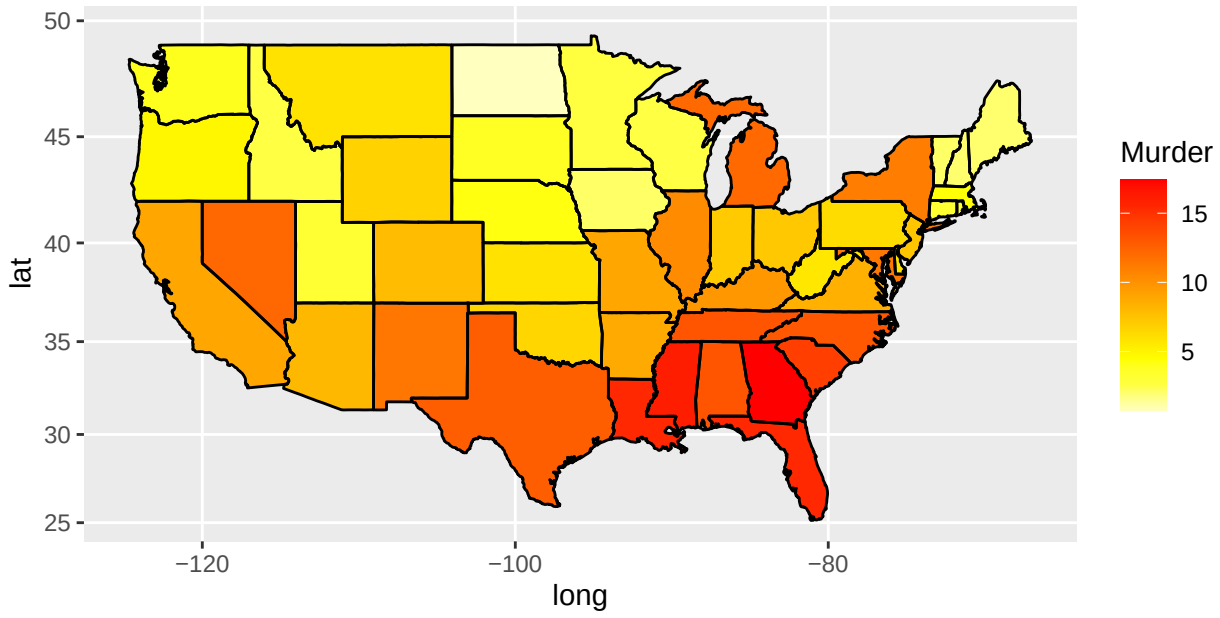


- Rape vs urban population
- Now with names



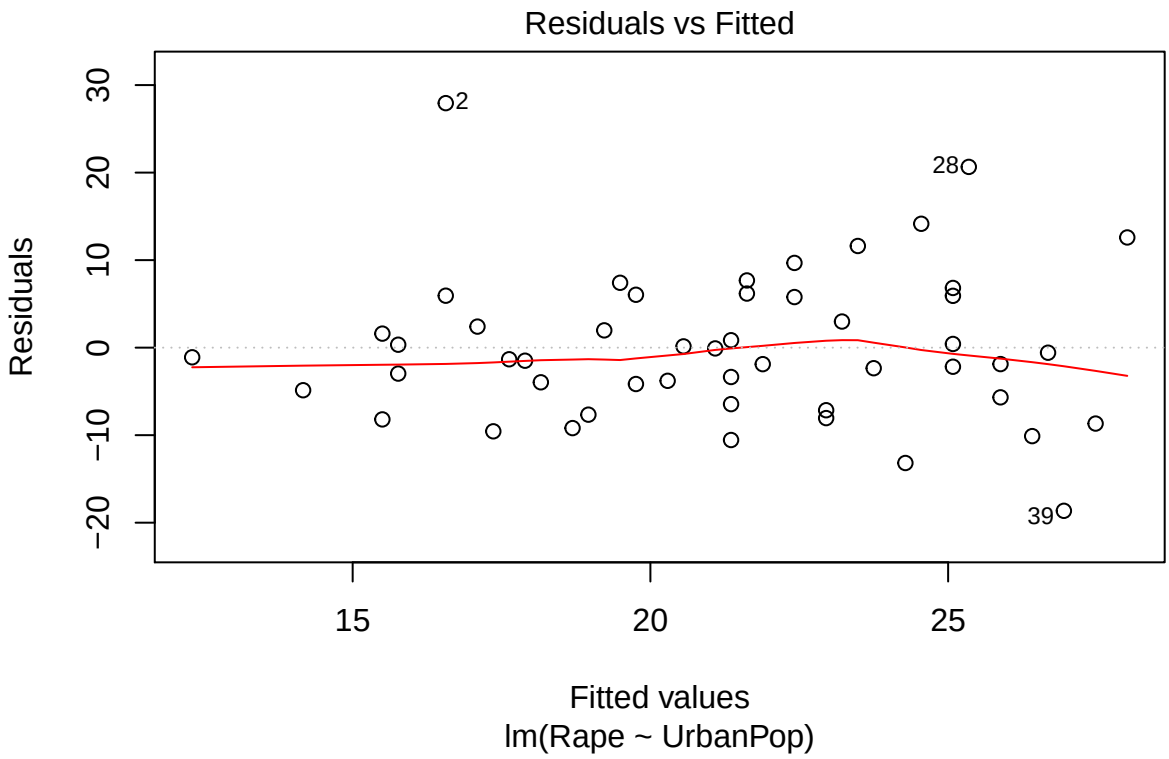



```
library(maps)
library(mapproj)
states <- map_data("state")
arsts <- arrests[c('region', 'Murder')]
arsts$region <- tolower(arsts$region)
map.df <- merge(states,arsts, by="region", all.x=T)
map.df <- map.df[order(map.df$order),]
ggplot(map.df, aes(x=long,y=lat,group=group))+
  geom_polygon(aes(fill=Murder))+
  geom_path()+
  scale_fill_gradientn(colours=rev(heat.colors(10)),na.value="grey90")+
  coord_map()
```



Regression

```
arrests_f <- arrests %>% dplyr::select(-Murder, -Assault)
linear_fit <-lm(Rape~UrbanPop, data = arrests_f)
plot(linear_fit, 1)
```




```
arrests %>% ggplot(aes(Rape, UrbanPop, label=region)) +geom_point(color = case_when(arrests$Rape>40 & a
 geom_smooth(method='lm', se = FALSE, aes(color = "Full data")) +
```

```
geom_smooth(data=arrests[arrests$region!='Alaska',],method='lm', se = FALSE, aes(color = "Alaska exclu
 ggtitle('Rape vs Urban Population regression plot')+
 xlab('Rape')+
 ylab('Urban Population')+
    scale_color_manual(name = "Linear fits",
                       breaks = c("Full data", "Alaska excluded"),
                       values = c("Full data" = "blue", "Alaska excluded" = "red") )
```