Assignment2

Ismail Olasege

4/11/2020

Rental Vacancy Data

Rental Data Cleaning

-Using the gather function from the tidyverse library to gather/clean the data

-Dropping of column 1 and 3 due to being irrelevant

-Renaming region name

```
rental_vacancy <- read_excel("Rental_Vacancy_Rate_by_State.xls")
names(rental_vacancy)[2] = "Region_Name"
library(tidyr)
rental_vacancy <- gather(rental_vacancy, year, rate, '2000':'2010')
rental_vacancy <- rental_vacancy[,c(-1,-3)]
head(rental_vacancy)
```

```
## # A tibble: 6 x 3
##   Region_Name year   rate
##   <chr>       <chr> <dbl>
## 1 Alabama     2000   14.4
## 2 Alaska      2000    6.9
## 3 Arizona     2000   10.7
## 4 Arkansas    2000   11.4
## 5 California  2000    4.5
## 6 Colorado    2000    5.4
```

```
str(rental_vacancy)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    561 obs. of  3 variables:
##  $ Region_Name: chr  "Alabama" "Alaska" "Arizona" "Arkansas" ...
##  $ year       : chr  "2000" "2000" "2000" "2000" ...
##  $ rate       : num  14.4 6.9 10.7 11.4 4.5 5.4 8.6 10.6 11.7 10.8 ...
```

```
names(rental_vacancy)
```

```
## [1] "Region_Name" "year"        "rate"
```

Describing Rental Vacancy data

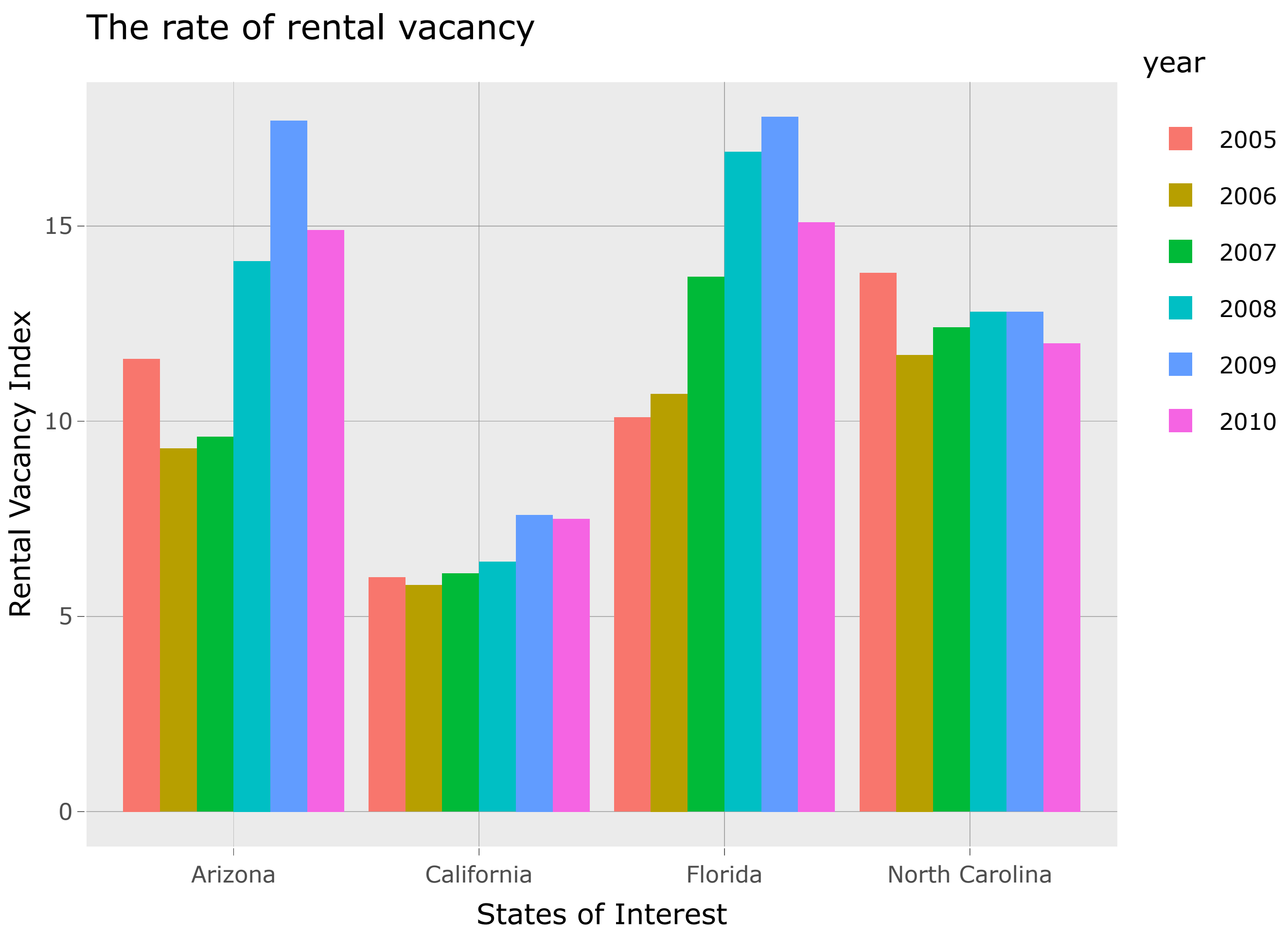
The following varianbles are attributes of the the rental vacancy data: `Region_Name, year, rate`

There are 3 number of columns and 561 number of rows

```
# The numeric variable data has a minimum value of -274 and a maximum value of 936
rental_vacancy %>%select(rate)%>%
  summary()
```

```
##       rate       
##  Min.   : 3.200  
##  1st Qu.: 7.400  
##  Median : 9.500  
##  Mean   : 9.578  
##  3rd Qu.:11.600  
##  Max.   :18.100  
```

```
a <- rental_vacancy %>% filter(Region_Name %in% c("North Carolina", "Arizona", "California", "Fl
orida") & year %in% c("2005", "2006", "2007", "2008", "2009", "2010")) %>%
  group_by(year, Region_Name) %>%
  ggplot() +
  geom_col(aes(Region_Name, rate, fill = year), position = "dodge")+
    labs(title = "The rate of rental vacancy",
         fill = "year",
         x="States of Interest",
         y = "Rental Vacancy Index")

ggplotly(a)
```



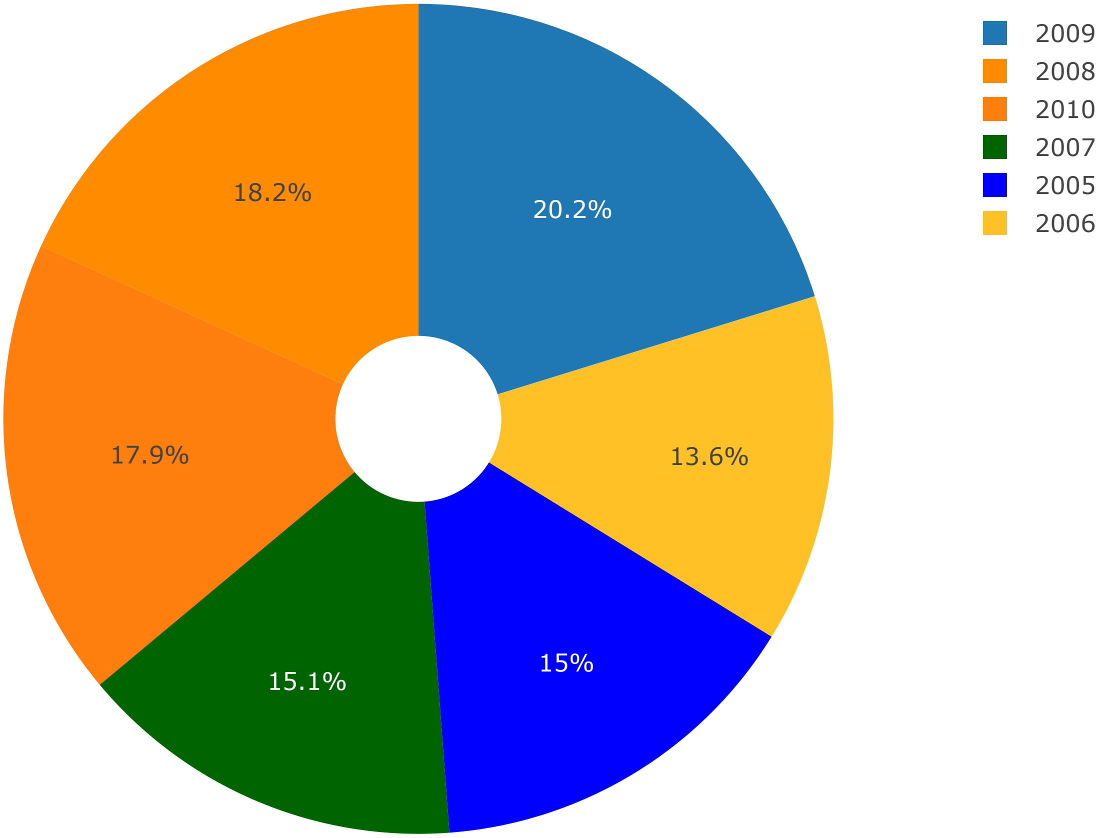
Rate of Rental Vacancy by State each year

It is generally seen that the of rental vacancy is very low in California while Arizona and Florida have a number of rental vacancy index. Aside that, it is shown that the number of rental vacancy increased between 2007 and 2010 and at the peak of rental vacancy in years 2009.

This shows the effect of the great recession, this means that there alot of people who could not afford to pay house rent eiather due to job loss or failed business.

There is an exception in North Carolina, there was no change in the rental vacancy index in 2008 and 2009.

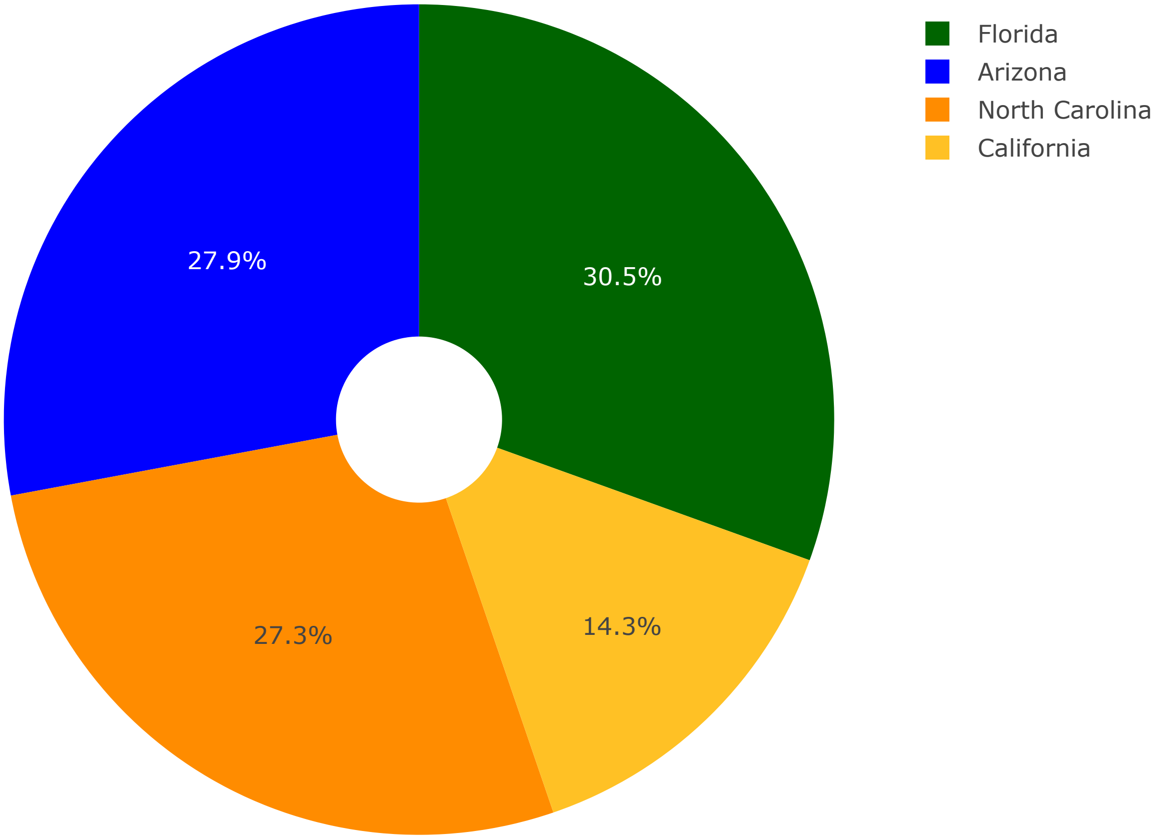
```
b <- rental_vacancy %>% filter(Region_Name %in% c("North Carolina", "Arizona", "California", "Fl
orida") & year %in% c("2005","2006", "2007", "2008", "2009", "2010")) %>%
  group_by(year) %>% summarise(vac_mean =mean(rate)) %>%
        plot_ly(labels = ~year,
                values = ~vac_mean,
                marker = list(colors = colors_used)) %>%
        add_pie(hole = 0.2) %>%
        layout(xaxis = list(zeroline = F,
                            showline = F,
                            showticklabels = F,
                            showgrid = F),
               yaxis = list(zeroline = F,
                            showline = F,
                            showticklabels=F,
                            showgrid=F))

b
```



The average rate of Rental Vacancy for six years

With a look at the Pie Chart, it is clear that the effect of the great recession took a significant effect in the year 2008 and increased till 2010. This means the was a continual job loss across the years

```
c <- rental_vacancy %>% filter(Region_Name %in% c("North Carolina", "Arizona", "California", "Florida") & year %in% c("2005", "2006", "2007", "2008", "2009", "2010")) %>%
  group_by(Region_Name) %>% summarise(vac_mean =mean(rate)) %>%
         plot_ly(labels = ~Region_Name,
                 values = ~vac_mean,
                 marker = list(colors = colors_used)) %>%
         add_pie(hole = 0.2) %>%
         layout(xaxis = list(zeroline = F,
                             showline = F,
                             showticklabels = F,
                             showgrid = F),
                yaxis = list(zeroline = F,
                             showline = F,
                             showticklabels=F,
                             showgrid=F))
c
```



The average rate of Rental Vacancy for each region in six years

Florida had the highest number of people that could not afford to apy for their rental apartment, this leeds to high rental vacancy across the state

Population Index Data

Population Data Cleaning

-Using the gather function from the tidyverse library to gather/clean the data

-Dropping of column 1 and 3 due to being irrelevant

-Renaming the region name

```
population <- read_excel("Resident_Population_by_State.xls")
names(population)[2] = "Region_Name"

population <- gather(population, year, rate, '2000':'2010')
population <- population[,c(-1,-3)]



#head(population)
```

Describing Population data

The following varianbles are attributes of the the rental vacancy data: `Region_Name, year, rate`

There are `3` number of columns and `561` number of rows

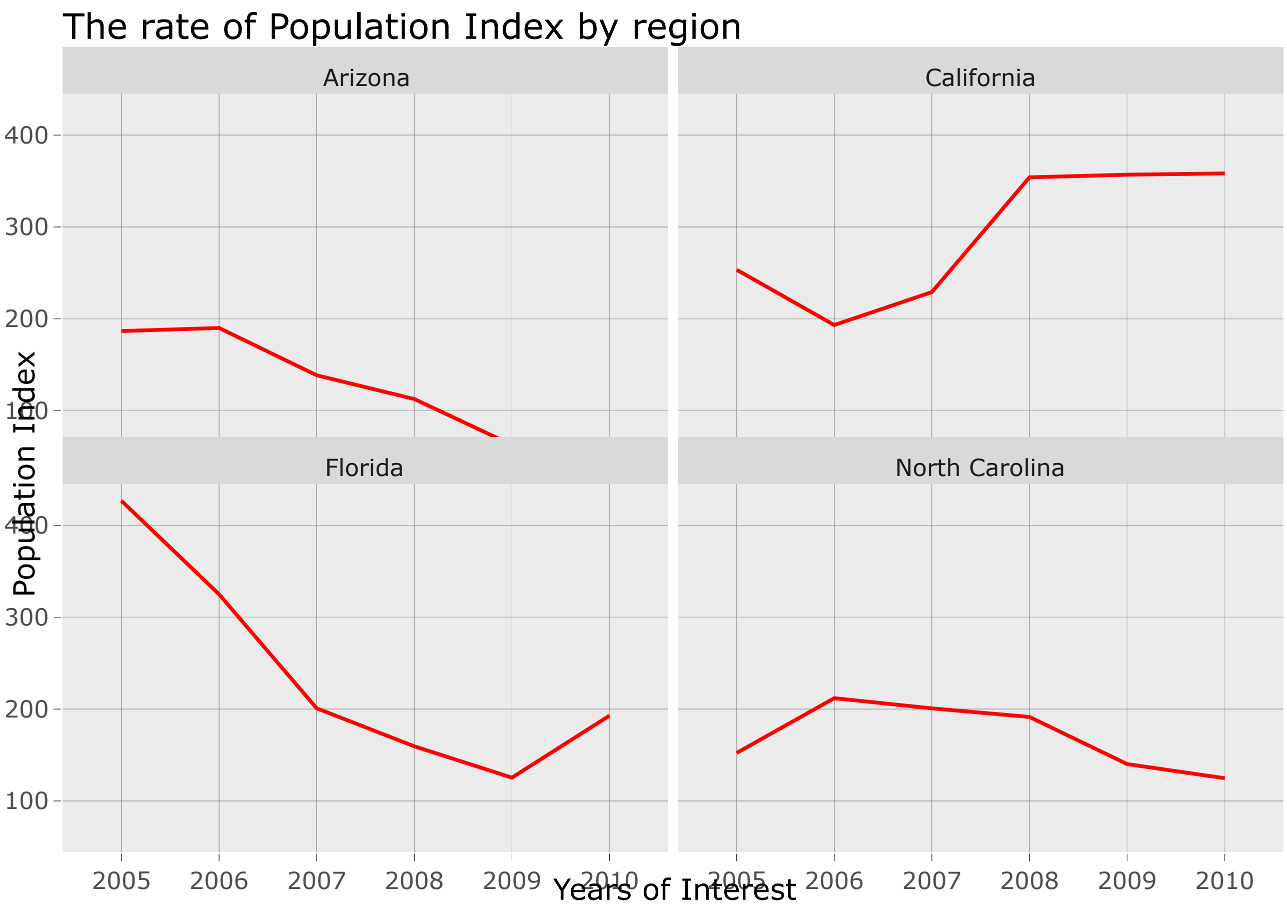
```
names(population)[3] = "Index"
head(population)
```

```
## # A tibble: 6 x 3
##   Region_Name year   Index
##   <chr>       <chr>  <dbl>
## 1 Alabama     2000   82.3
## 2 Alaska      2000    8.46
## 3 Arizona     2000  382.
## 4 Arkansas    2000  127.
## 5 California  2000  843.
## 6 Colorado    2000  271.
```

```
# The numeric variable data has a minimum value of -274 and a maximum value of 936
population %>%select(Index)%>%
  summary()
```

```
##      Index
##  Min.   :-273.96
##  1st Qu.:  10.86
##  Median :  31.11
##  Mean   :  65.30
##  3rd Qu.:  64.72
##  Max.   : 936.27
```

```
a <- population %>% filter(Region_Name %in% c("North Carolina", "Arizona", "California", "Florid
a"), year %in% c("2005","2006", "2007", "2008", "2009", "2010")) %>%
  ggplot(mapping = aes(year, Index)) +
  geom_line(aes(group = Region_Name),  color = "red") +
  labs(title = "The rate of Population Index by region",
         fill = "year",
         x="Years of Interest",
         y = "Population Index")+
  facet_wrap(~Region_Name)

ggplotly(a)
```



Arizona, Florida and North Carolina States show that there was a steady decrease in Population between 2006 and 2009. From the graph we can see that California had in increase in Population and a steady population between 2008 and 2010. Does this means California was cheaper to leave in compared to the three other states? or Was more jobs for people to do?

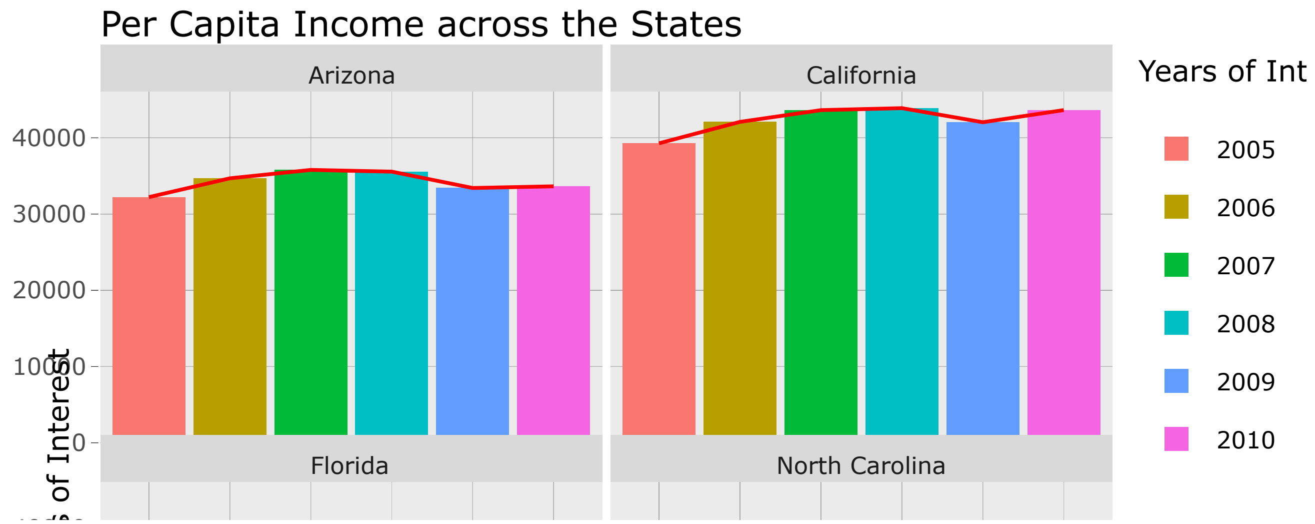
Effect of GDP and Per Capita Income

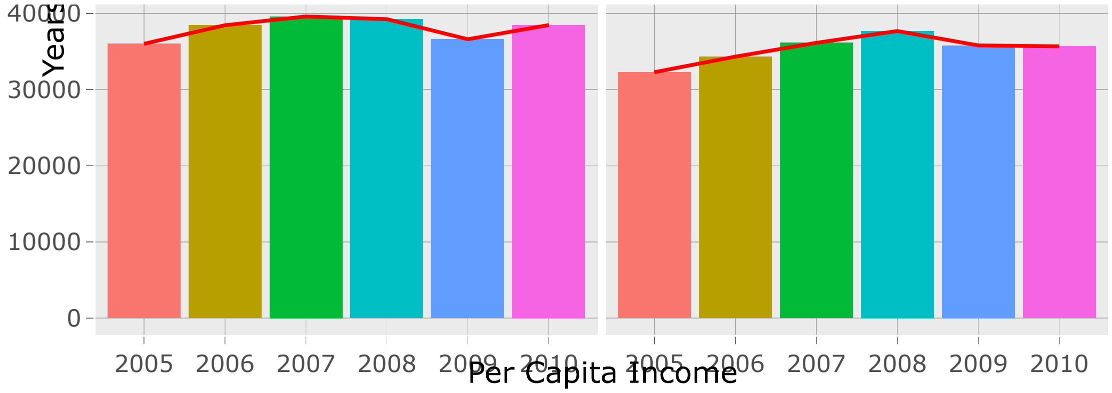
```
pci_state <- read_excel("Per_Capita_Personal_Income_by_State.xls")
pci_state <- pci_state[, c(-1,-3)]
names(pci_state)[1] <- "Region_Name"

pci_state <- gather(pci_state, year, income, "2000":"2010")
head(pci_state)
```

```
## # A tibble: 6 x 3
##   Region_Name year  income
##   <chr>       <chr>  <dbl>
## 1 Alabama     2000   24338
## 2 Alaska      2000   31974
## 3 Arizona     2000   26235
## 4 Arkansas    2000   22762
## 5 California  2000   33364
## 6 Colorado    2000   34187
```

Per Capita Income in four states over six years

```
a <- pci_state %>% filter(Region_Name %in% c("North Carolina", "Arizona", "California", "Florid
a") & year %in% c("2005","2006", "2007", "2008", "2009", "2010")) %>%
  group_by(year) %>%
  ggplot(aes(year, income)) +
  geom_col(aes(fill = year), position = "dodge") +
  geom_line(aes(group = Region_Name),  color = "red")+
    labs(title = "Per Capita Income across the States",
         fill = "Years of Interest",
         x="Per Capita Income",
         y = "Years of Interest")+

  facet_wrap(~Region_Name)

plotly::ggplotly(a)
```



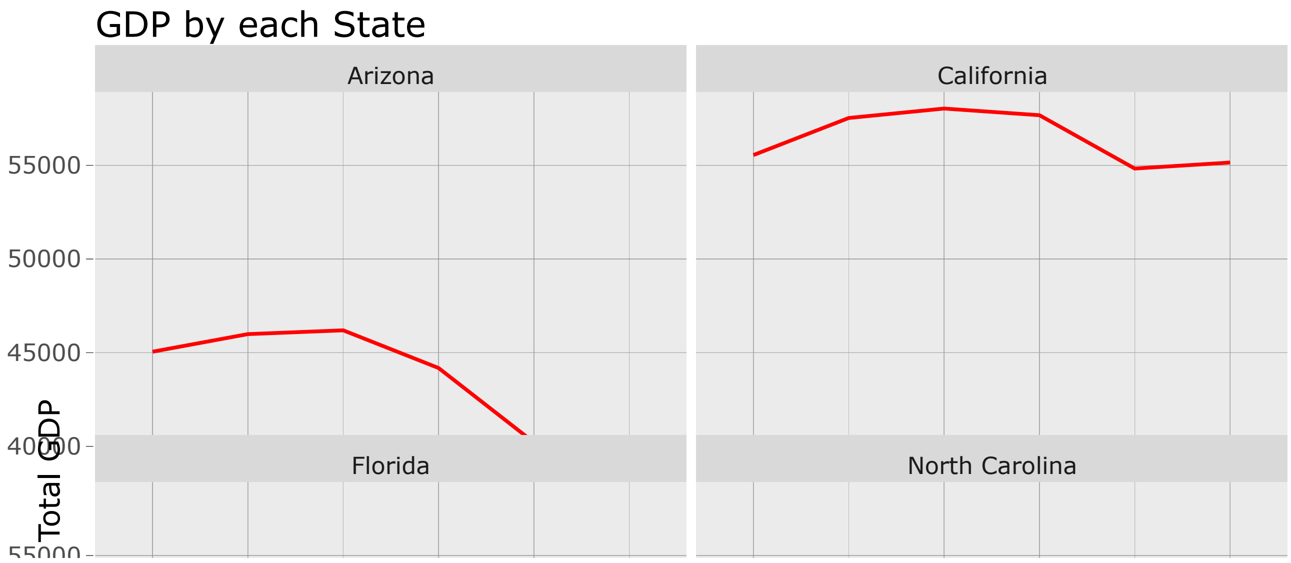


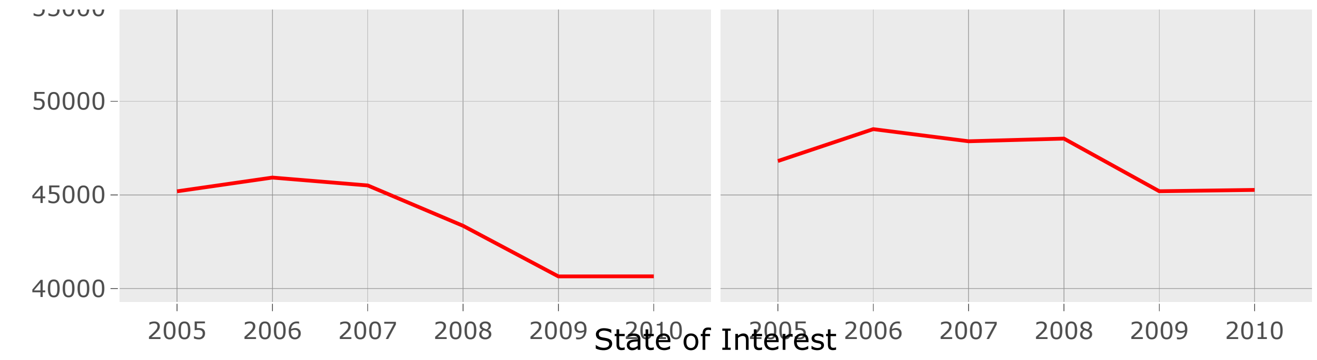
There was a slight changes in the Per Capita Income across the six years in each state but generall, California has more people that are averagely earning more than other state.

GDP Data

GDP by State

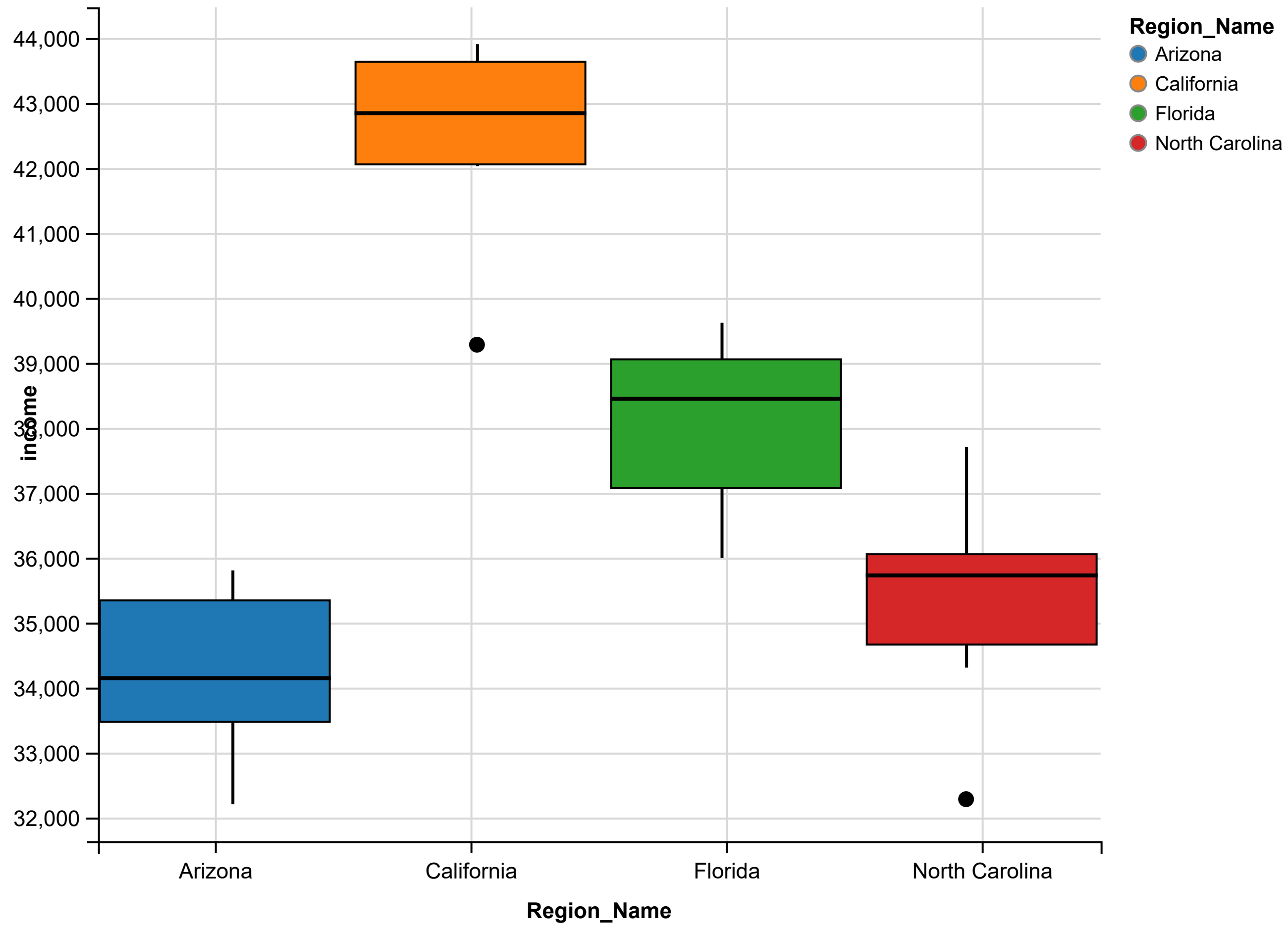
```
state_gdp <- read_excel("Real_Total_Gross_Domestic_Product_by_State.xls")
state_gdp <- state_gdp[,c(-1,-3)]
state_gdp <- gather(state_gdp, year, gdp, "2000": "2010")
names(state_gdp)[1] <- "Region_Name"
#head(state_gdp)
a <- state_gdp %>% filter(Region_Name %in% c("North Carolina", "Arizona", "California", "Florid
a"), year %in% c("2005","2006", "2007", "2008", "2009", "2010")) %>%
  ggplot(mapping = aes(year, gdp)) +
  geom_line(aes(group = Region_Name),  color = "red") +
  facet_wrap(~Region_Name)+
  labs(title = "GDP by each State",
         fill = "State of Interest",
         x="State of Interest",
         y = "Total GDP")

ggplotly(a)
```





Distribution of Income by Region

```
a <- pci_state %>% filter(Region_Name %in% c("North Carolina", "Arizona", "California", "Florid
a") & year %in% c("2005","2006", "2007", "2008", "2009", "2010")) %>%
        group_by(Region_Name) %>%
        ggvis(~Region_Name, ~income, fill = ~Region_Name) %>%
        layer_boxplots()
a
```



The boxplot shows that Arizona has the highest number of people with high income. Averagely the jobs is Arizona pay more than other states.

Homeownership Data

Homeownership Data Cleaning

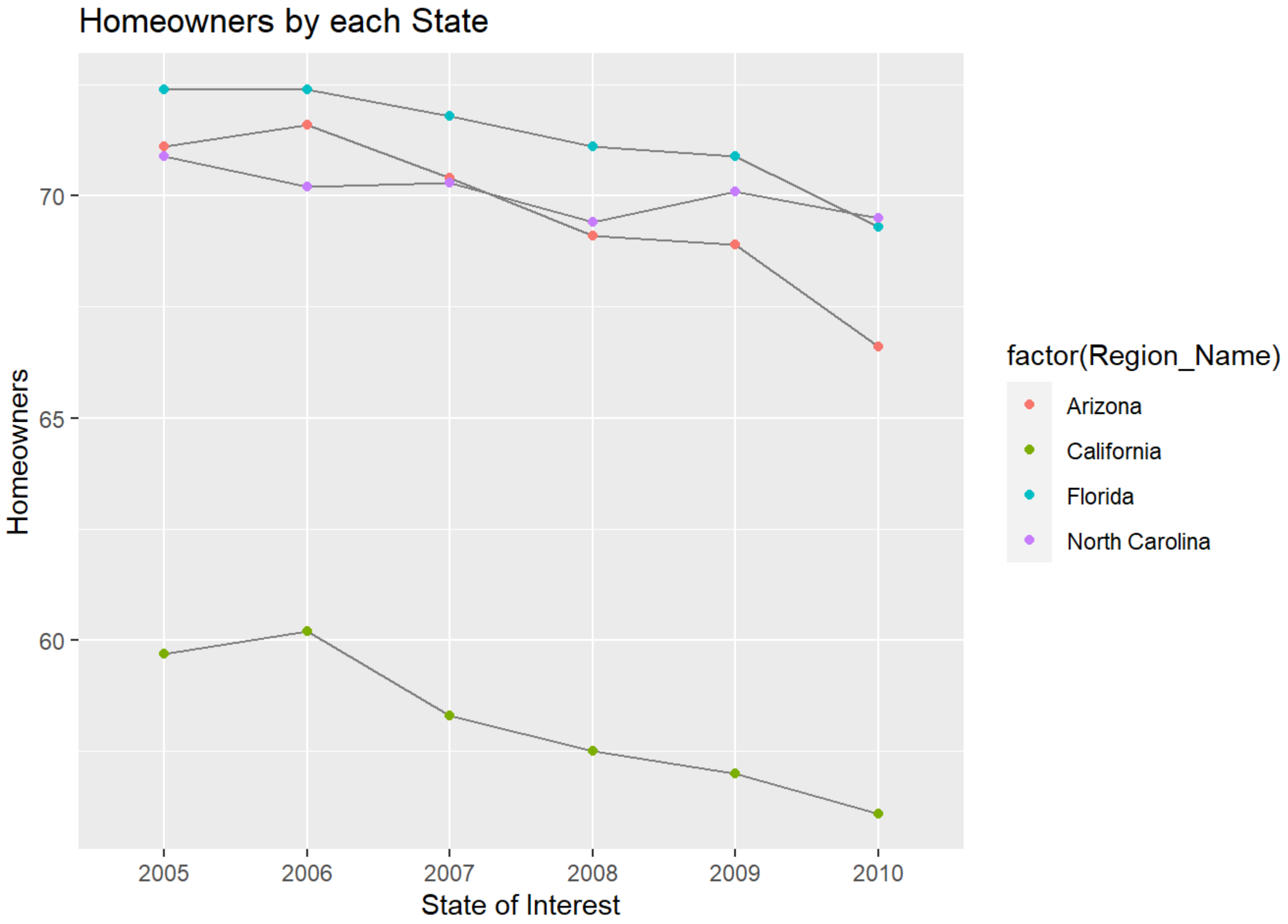
```
Home_owner <- read_excel("Homeownership_Rate_by_State.xls")
Home_owner <- Home_owner[,c(-1,-3)]
Home_owner <- gather(Home_owner, year, owner_index, "2000": "2010")
names(Home_owner)[1] <- "Region_Name"
head(Home_owner)
```

```
## # A tibble: 6 x 3
##   Region_Name year  owner_index
##   <chr>       <chr>       <dbl>
## 1 Alabama     2000         73.2
## 2 Alaska      2000         66.4
## 3 Arizona     2000         68
## 4 Arkansas    2000         68.9
## 5 California  2000         57.1
## 6 Colorado    2000         68.3
```

Time trend of homeowner in four states over six years

```
Home_owner %>% filter(Region_Name %in% c("North Carolina", "Arizona", "California", "Florida") &
year %in% c("2005","2006", "2007", "2008", "2009", "2010")) %>%
    group_by(Region_Name, year) %>%
    ggplot(., aes(year, owner_index) ) +
    geom_line(aes(group=Region_Name), color="grey50")+
    geom_point(aes(color=factor(Region_Name)))+
    geom_smooth(method = "lm")+
    scale_fill_brewer(palette="Set1")+
  labs(title = "Homeowners by each State",
         fill = "State of Interest",
         x="State of Interest",
         y = "Homeowners")
```

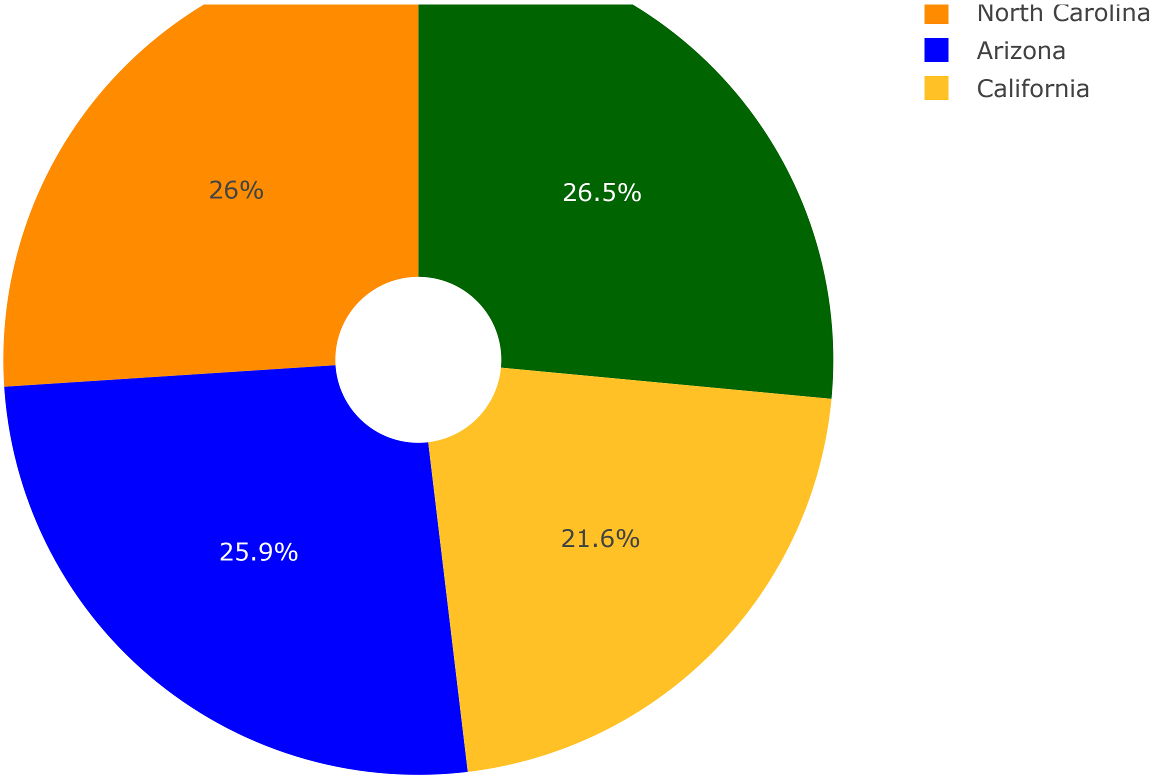
```
## `geom_smooth()` using formula 'y ~ x'
```



The Line graph shows a decline in the rate of homeowner over the years with Florida having th highest number of persons with home. The decline in homeowner started in 2007 and still going down untill 2010. This is as a result of the recession

Average homeowner by State

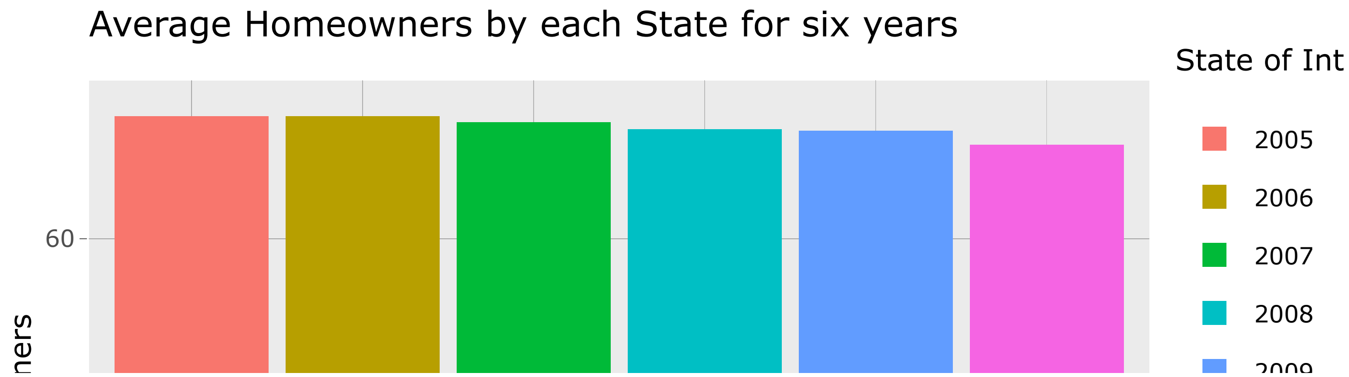
```
c <- Home_owner %>% filter(Region_Name %in% c("North Carolina", "Arizona", "California", "Florid
a") & year %in% c("2005", "2006", "2007", "2008", "2009", "2010")) %>%
  group_by(Region_Name) %>% summarise(avg_owner =mean(owner_index)) %>%
         plot_ly(labels = ~Region_Name,
                 values = ~avg_owner,
                 marker = list(colors = colors_used)) %>%
         add_pie(hole = 0.2) %>%
         layout(xaxis = list(zeroline = F,
                             showline = F,
                             showticklabels = F,
                             showgrid = F),
                yaxis = list(zeroline = F,
                             showline = F,
                             showticklabels=F,
                             showgrid=F))
c
```

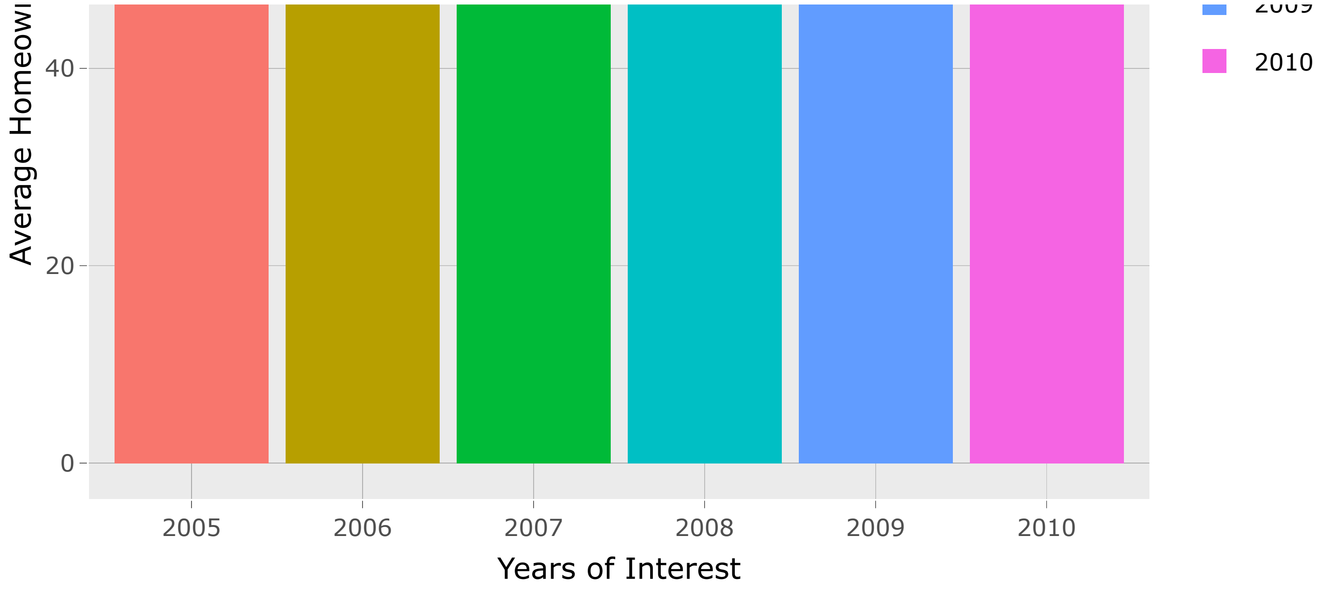


It is expected that Florida would have the average number of homeowner among the four states due to the high number of homeowner the line graph above shows and California having the lowest average number of homeowner.

Average homeowner over six years

```
a <- Home_owner %>% filter(Region_Name %in% c("North Carolina", "Arizona", "California", "Florida") & year %in% c("2005":"2010")) %>%
  group_by(Region_Name,year) %>% summarise(avg_owner =mean(owner_index))%>%
  ggplot(mapping = aes(x = year, y = avg_owner)) +
  geom_col(aes(year, avg_owner, fill = year), position = "dodge") +
  labs(title = "Average Homeowners by each State for six years",
         fill = "State of Interest",
         x="Years of Interest",
         y = "Average Homeowners")

ggplotly(a)
```

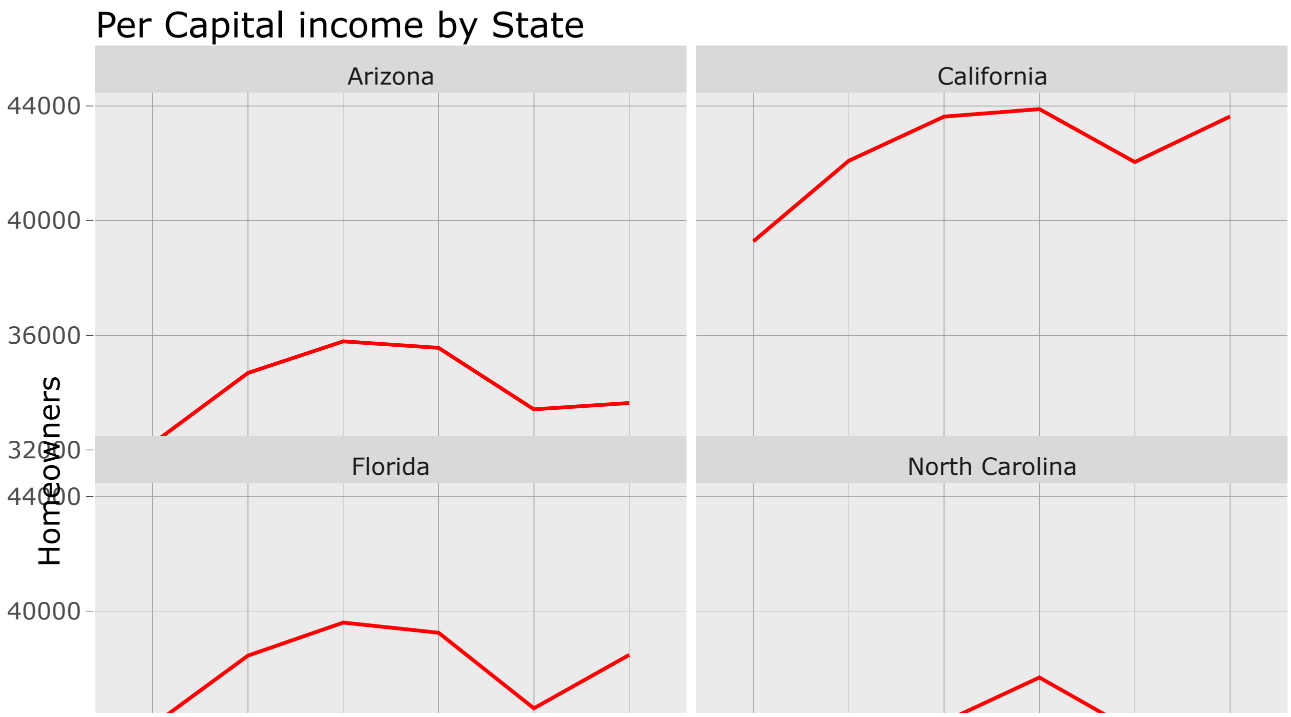


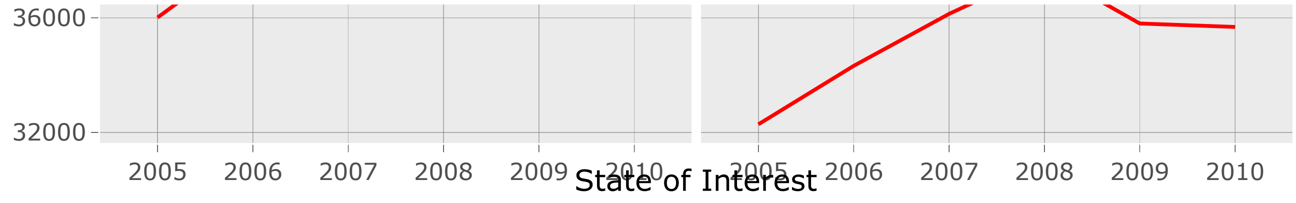


The four states had a slow decline in the average number of homeowner

Income by State

```
a <- pci_state %>% filter(Region_Name %in% c("North Carolina", "Arizona", "California", "Florid
a"), year %in% c("2005","2006", "2007", "2008", "2009", "2010")) %>%
  ggplot(mapping = aes(year, income)) +
  geom_line(aes(group = Region_Name),  color = "red") +
  facet_wrap(~Region_Name)+
  labs(title = "Per Capital income by State",
         fill = "State of Interest",
         x="State of Interest",
         y = "Homeowners")

ggplotly(a)
```





The Per Capita Income by each State experienced a decline from 2008 to 2010 except for California with an increase in the number of income

```
extra <- read.csv("extracredit.csv")
extra <- gather(extra, year, value, "X2004":"X2012")
#head(extra)
```

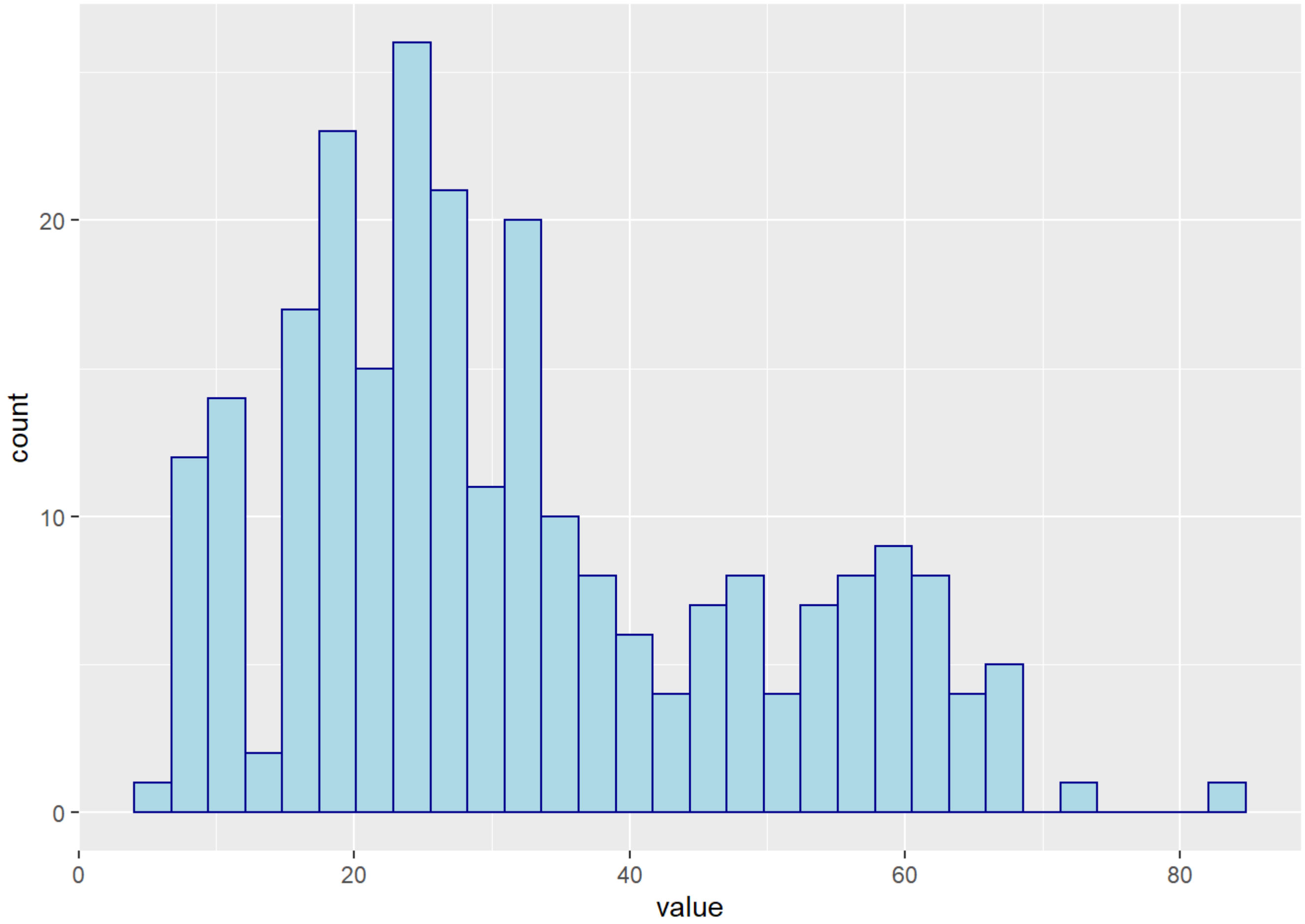
```
library(tidyr)
extra<- extra %>% mutate(year = ifelse(year == 'X2004', 2004,
                                       ifelse(year == 'X2005', 2005,
                                              ifelse(year == 'X2006', 2006,
                                                     ifelse(year == 'X2007', 2007,
                                                            ifelse(year == 'X2008', 2008,
                                                                   ifelse(year == 'X2009',2009,
                                                                          ifelse(year == 'X2010'
, 2010,
                                                                                 ifelse(year ==
'X2011', 2011, 2012)))))))))
#head(extra)
```

```
extra <- extra[, -1]
#head(extra)
extra_credit <- filter(extra, State %in% c("NC", "AZ", "CA", "FL"))
#head(extra_credit)
```

Histogram showing the distribution of depression in North Carolina, Florida, California and Arizona

```
#library(ggplot2)
ggplot(extra_credit, aes(x = value))+
  geom_histogram(color="darkblue", fill="lightblue")
```

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

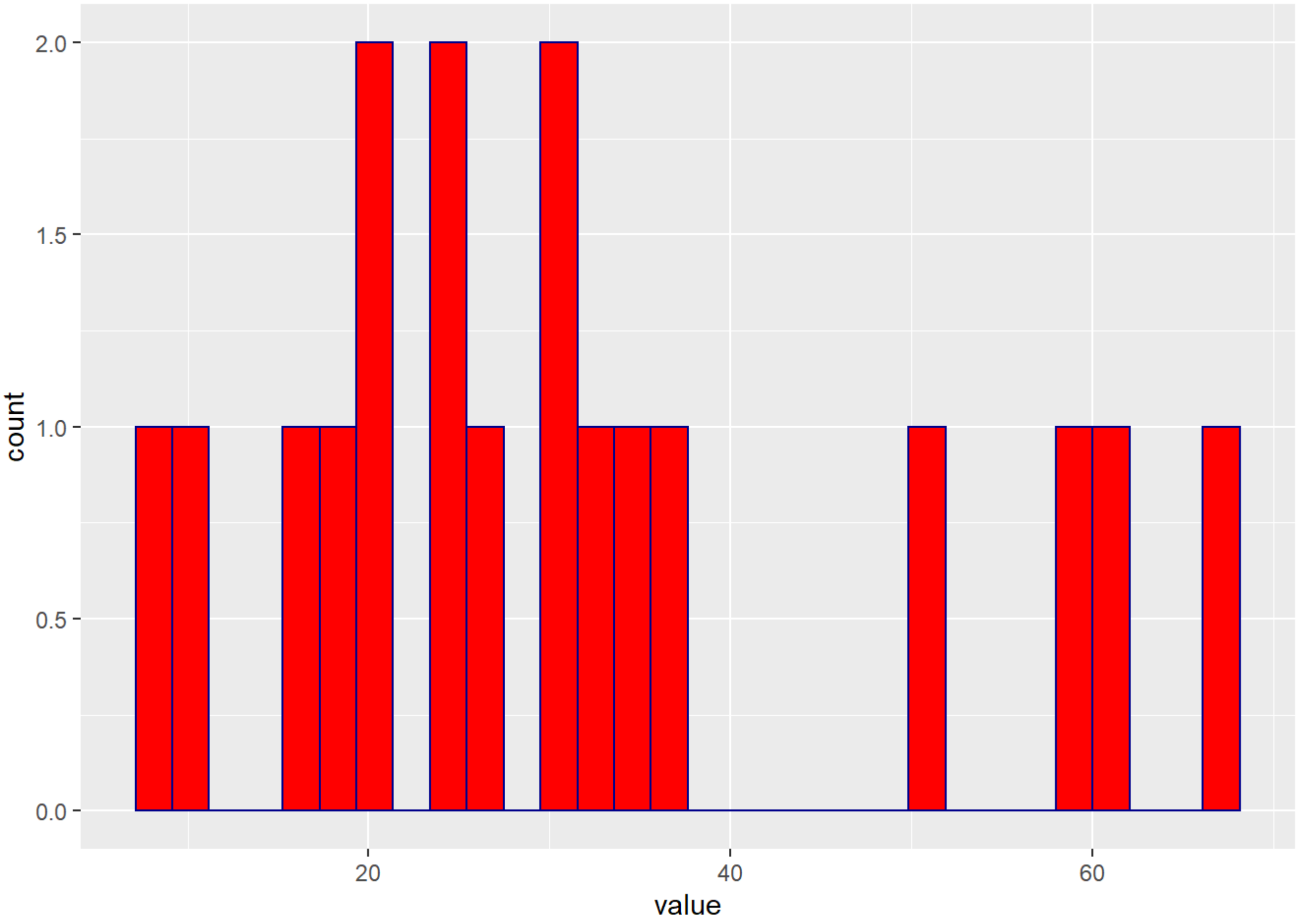


The histogram shows the distribution of the whole depression in four States between 2004 and 2012. There distribution shows there are some outliers and the distribution is not normal.

```
extra_AZ <- filter(extra_credit, State == "AZ")
#extra_AZ
```

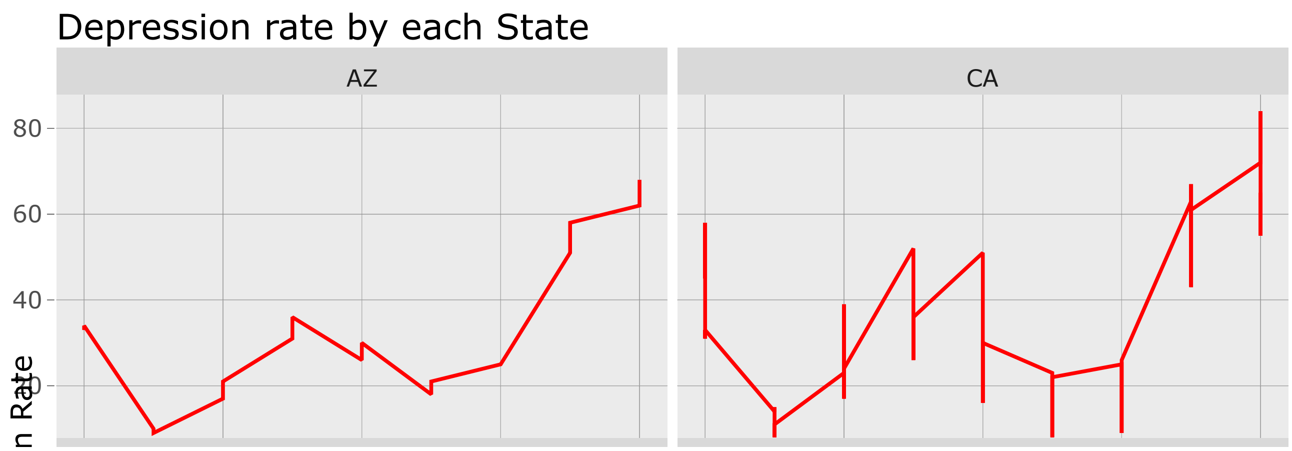
```
# Histogram showing the distribution of depression in Arizona
ggplot(extra_AZ, aes(x=value))+
  geom_histogram(color="darkblue", fill="red")
```

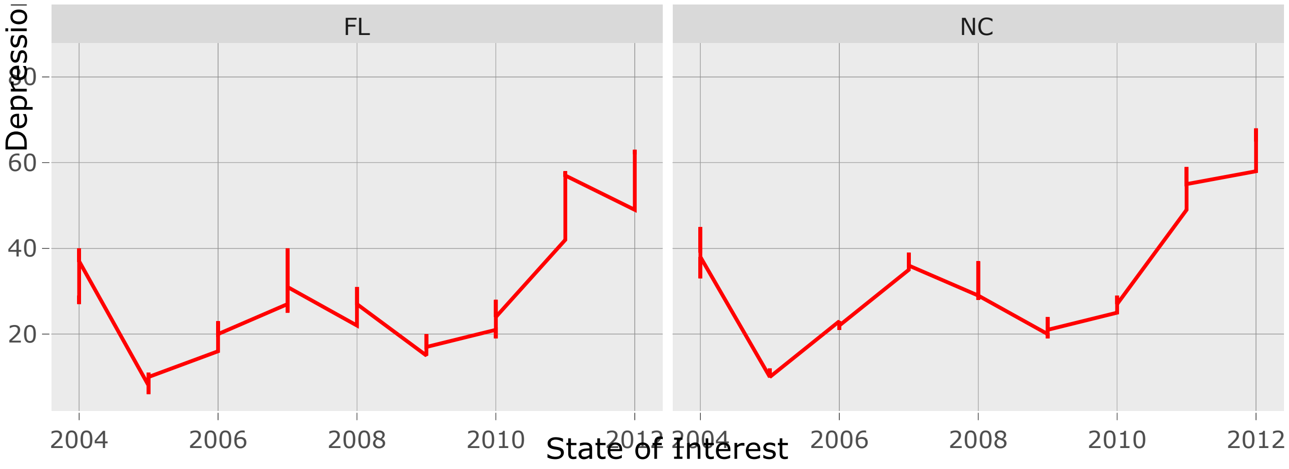
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The histogram showing irregularities in the distribution of depression rate in Arizona.

```
a <- extra_credit %>% filter(State %in% c("NC", "AZ", "CA", "FL"), year %in% c("2004":"2012")) %
>%
  ggplot(mapping = aes(year, value)) +
  geom_line(aes(group = State),  color = "red") +
  facet_wrap(~State)+
  labs(title = "Depression rate by each State",
         fill = "State of Interest",
         x="State of Interest",
         y = "Depression Rate")

ggplotly(a)
```



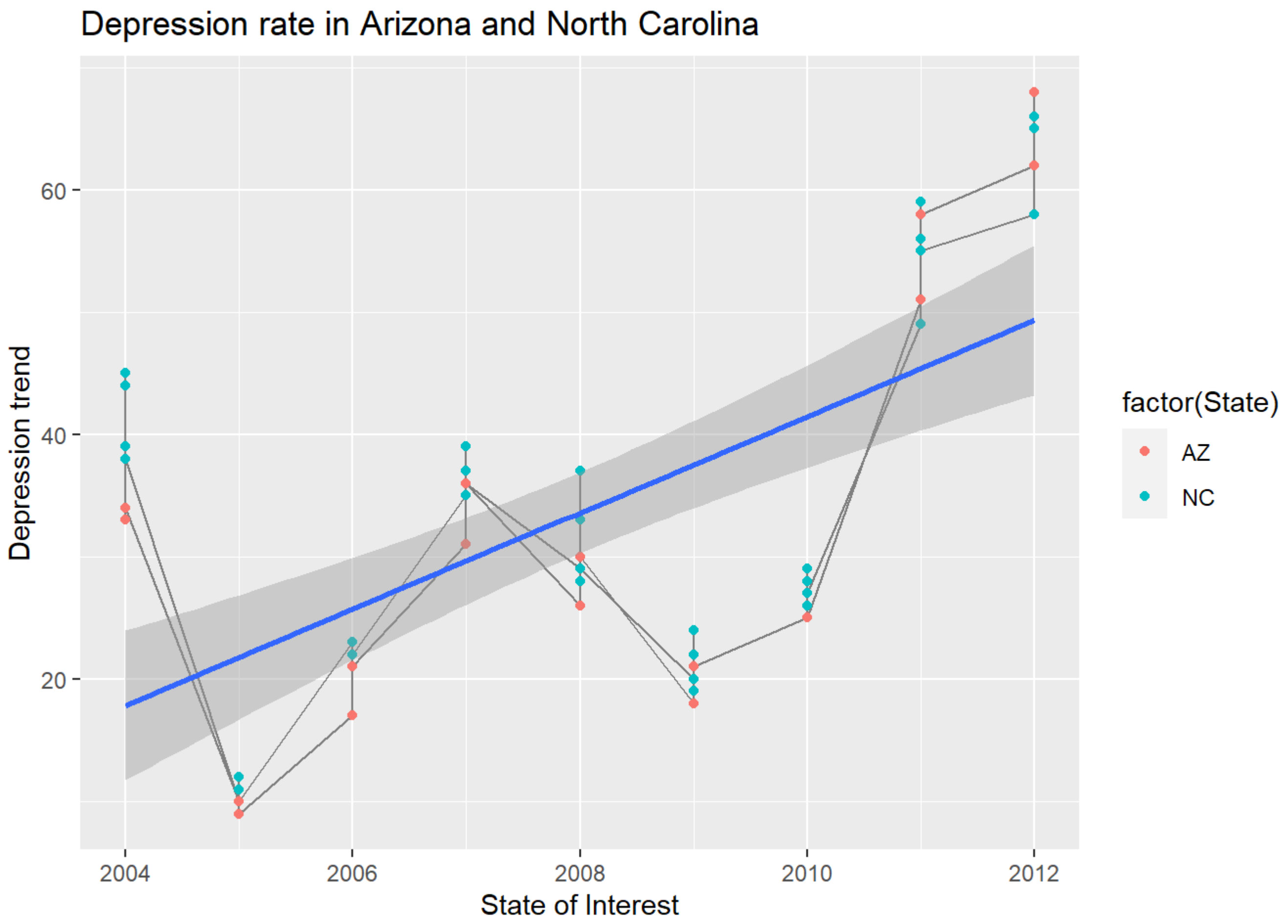


The line graph shows the the depression trend showing an increase in depression rate from 2010 and above. This could be a post recession effect.

Line trend showing the depression effect in Arizona and North Carolina between 2004 and 2012

```
extra_credit %>% filter(State %in% c("NC", "AZ") & year %in% c("2004":"2012")) %>%
    group_by(State, year) %>%
    ggplot(., aes(year, value) ) +
    geom_line(aes(group=State), color="grey50")+
    geom_point(aes(color=factor(State)))+
    geom_smooth(method = "lm")+
    scale_fill_brewer(palette="Set1")+
  labs(title = "Depression rate in Arizona and North Carolina",
         fill = "State of Interest",
         x="State of Interest",
         y = "Depression trend")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

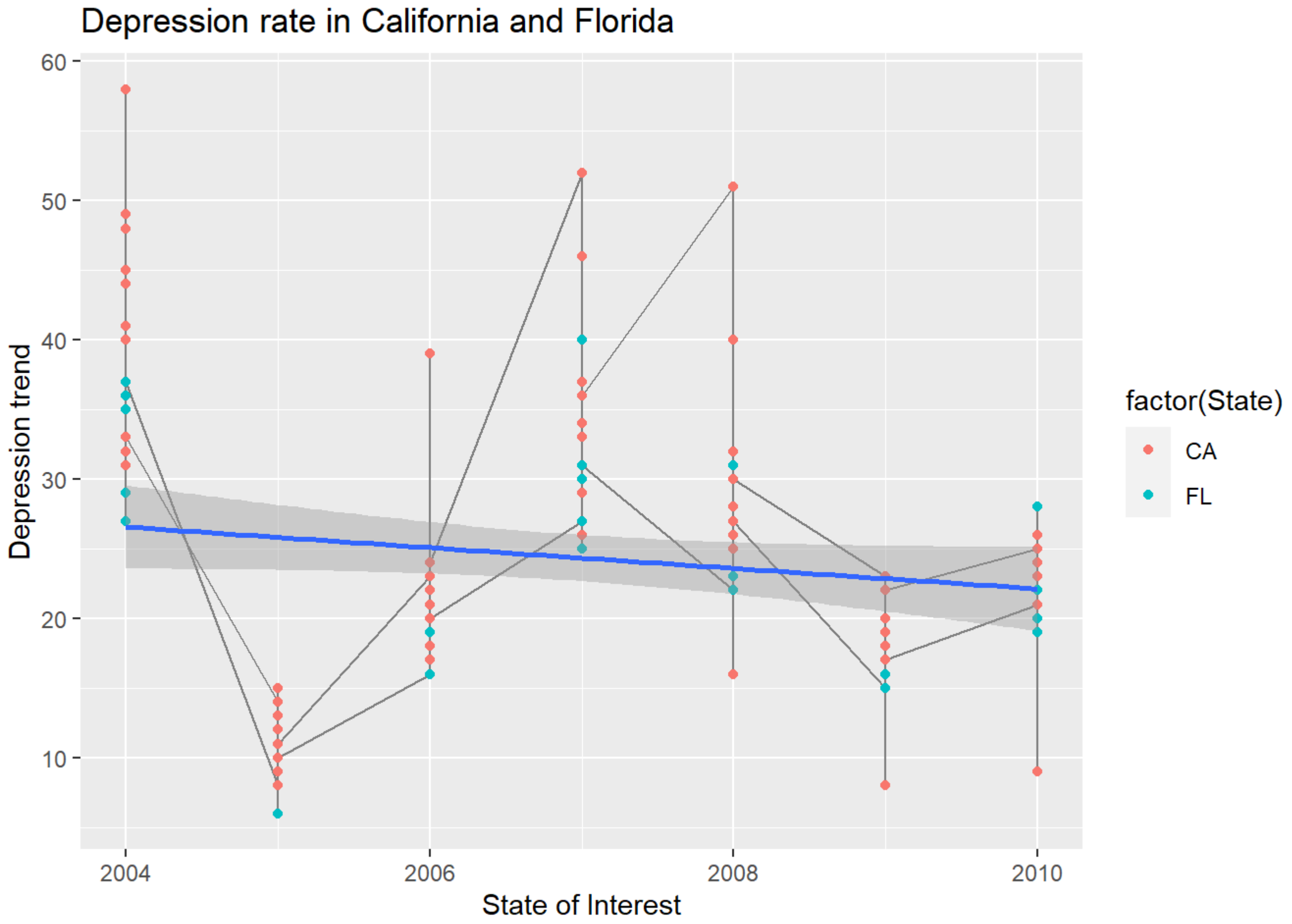


The trend shows an increase in the rate of depression in Arizona and North Carolina. The effect of the great recession has more effect on depression during the recession and after the recession.

Line trend showing the depression effect in California and Florida

```
extra_credit %>% filter(State %in% c("CA", "FL") & year %in% c("2004", "2005","2006", "2007", "2
008", "2009", "2010")) %>%
    group_by(State, year) %>%
    ggplot(., aes(year, value) ) +
    geom_line(aes(group=State), color="grey50")+
    geom_point(aes(color=factor(State)))+
    geom_smooth(method = "lm")+
    scale_fill_brewer(palette="Set1")+
  labs(title = "Depression rate in California and Florida",
        fill = "State of Interest",
        x="State of Interest",
        y = "Depression trend")
```

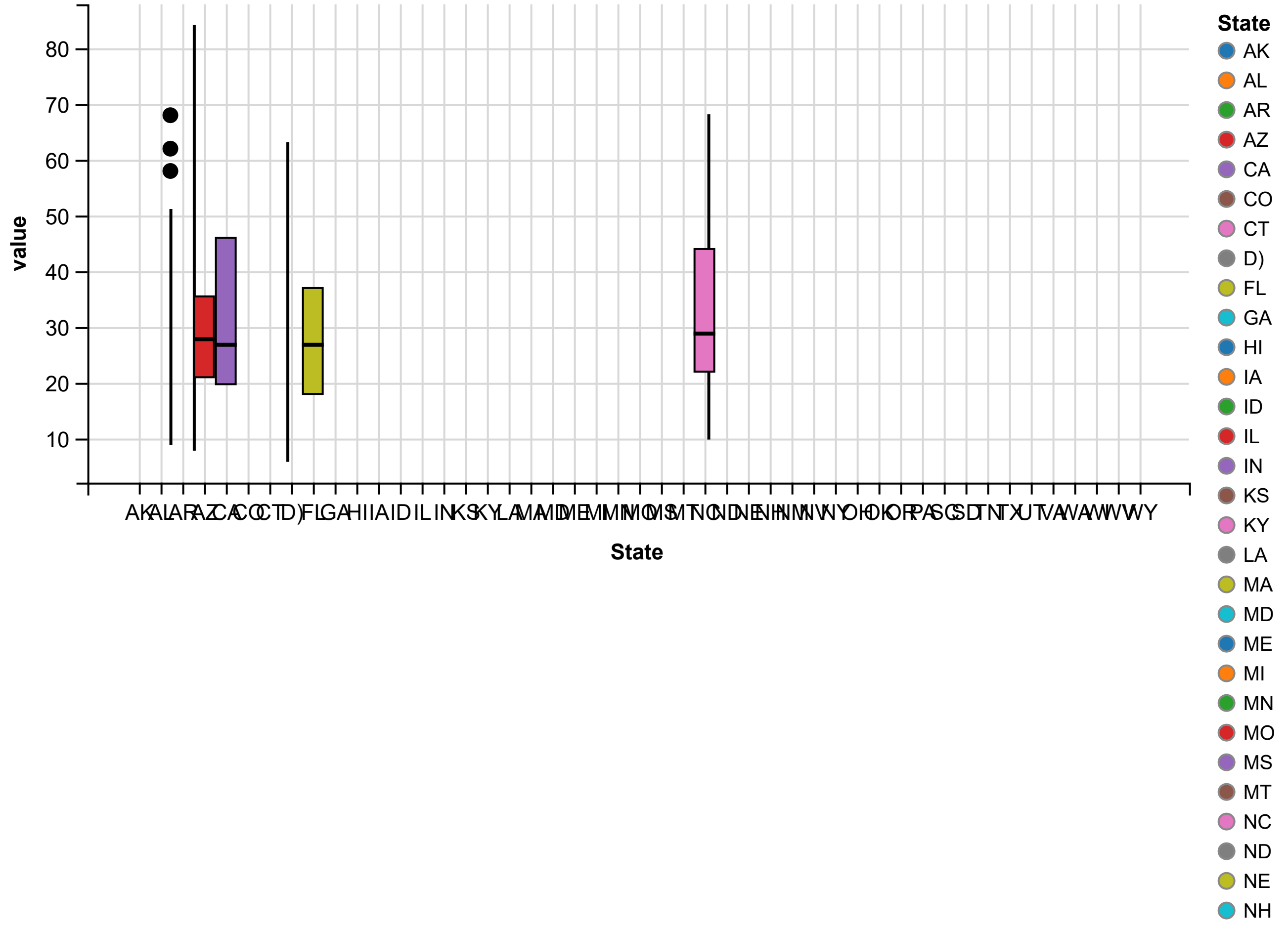
```
## `geom_smooth()` using formula 'y ~ x'
```



The depression trend was high during the recession in California and Florida between 2006 and 2009

Boxplot showing the depression distribution across four State between 2004 and 2012

```
library(dplyr)
a <- extra_credit %>% filter(State %in% c("NC", "AZ", "CA", "FL") & year %in% c("2004", "2005",
"2006", "2007", "2008", "2009", "2010", "2011", "2012")) %>%
         group_by(State) %>%
         ggvis(~State, ~value, fill = ~State) %>%
         layer_boxplots()
a
```



The distribution from Arizona has some outliers and also there is no equal variance.