EDAProject

Team 3

2/20/2020

## House Prices Over Time

housePriceData = import("State\_and\_US\_SA.xls")

## New names:  
## \* `` -> ...1  
## \* `Economic and Housing Research` -> `Economic and Housing Research...2`  
## \* `` -> ...3  
## \* `` -> ...4  
## \* `` -> ...5  
## \* ... and 48 more problems

housePriceData = housePriceData[5:545, 1:52]  
#View(housePriceData)

# change column name  
colNames = as.list(housePriceData[1,])  
names(housePriceData) = colNames  
housePriceData = housePriceData[-c(1),] #deletes first data row of column names

period.list = strsplit(as.character(housePriceData$Month), "M")  
period.list = matrix(unlist(period.list), byrow = T, ncol = 2)  
Year = period.list[, 1]  
Month = period.list[, 2]  
housePriceData$Month = Month  
housePriceData$Year = Year  
housePriceData = data.frame(lapply(housePriceData, as.numeric), stringsAsFactors=FALSE)  
#View(housePriceData)

state\_abbrev = as.data.frame(read.delim(file = "state\_abbrevs.txt", header = TRUE, sep = " "))  
state\_abbrev = data.frame(lapply(state\_abbrev, as.character), stringsAsFactors=FALSE)  
stateCode = as.array(state\_abbrev$Code)  
  
housePriceData = housePriceData %>%  
 gather(key = "Code", value = "Price", stateCode)

## Adjusting for CPI

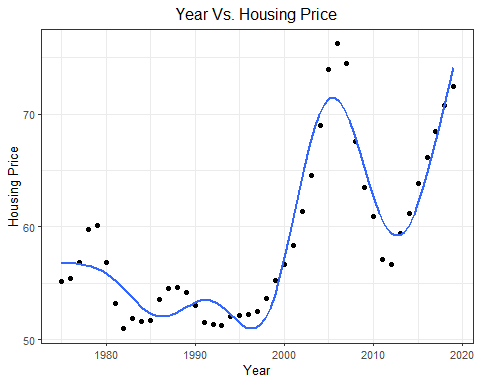
cpi.data = read.csv("cpi.csv")  
cpi.data$Period <- sapply(cpi.data$Period, function(x) gsub("M", "", x))  
#cpi.data$Period <- sapply(cpi.data$Period, function(x) gsub("0", "", x))  
#cpi.data  
cpi.data$Period = as.numeric(cpi.data$Period)  
names(cpi.data)[names(cpi.data) == "Period"] <- "Month"  
  
housePriceData = inner\_join(housePriceData, cpi.data, by = c("Year", "Month"))  
housePriceData <- transform(housePriceData, Price = (Price / Value)\* 100)

housePriceData = inner\_join(housePriceData, state\_abbrev, by = "Code")  
housePriceData.avgPriceUS = aggregate(Price ~ Year, FUN = mean, data = housePriceData)  
housePriceData.avgPriceRegion = aggregate(Price ~ Year + Region, FUN = mean, data = housePriceData)  
housePriceData.avgPriceState = aggregate(Price ~ Year + State, FUN = mean, data = housePriceData)

## 

## Change in prices US (CPI adjusted)

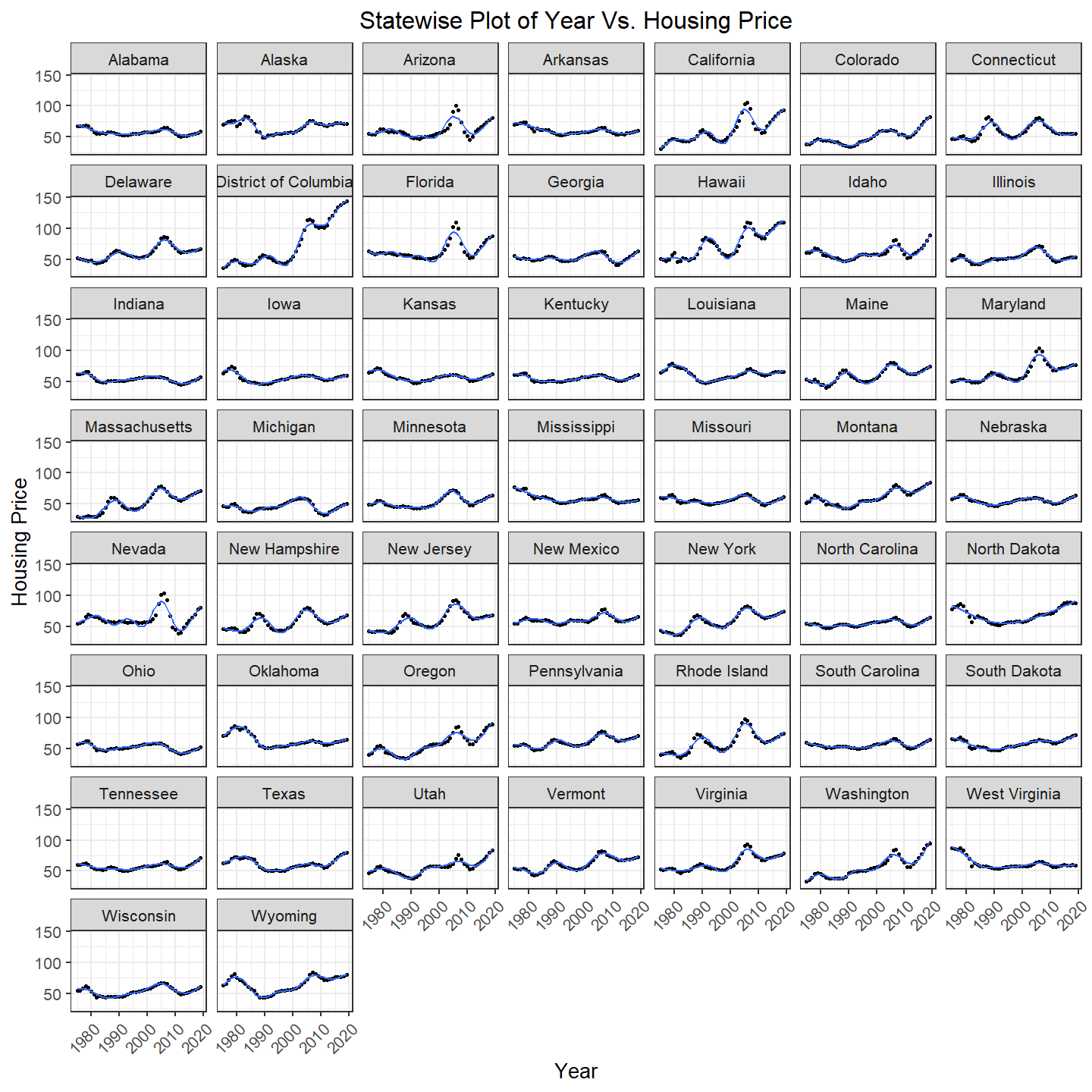
ggplot(housePriceData.avgPriceUS, aes(x = Year, y = Price)) +  
 geom\_point() +  
 geom\_smooth(method="gam", formula = y~s(x), se = FALSE) +  
 ggtitle("Year Vs. Housing Price") +  
 xlab("Year") +  
 ylab("Housing Price") +  
 theme\_bw(base\_size=10) +  
 theme(plot.title = element\_text(hjust = 0.5))



Year vs House Price: There are fluctuations in Housing Prices over the time period. The sharpest change (increase) in House prices can be seen be seen after 1995. These prices have been adjusted as per inflation hence the increase in value of currency has been taken into account. One more observation that can be derived is that the house price fluctuation is much more in the 2000s as compared to period 1975-2000. This can be interpolated to the fact of more economic dynamics in the recent years. There can be a correlated to increase in globalization and changes in market dynamics.

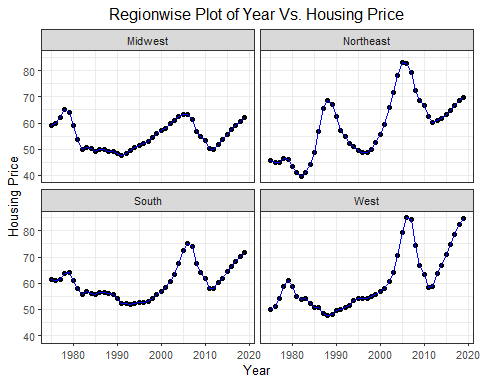
## Biggest increase and decrease

## changes in price varied by state



Different states showcase different relationships with house prices. Although each state obviously has a distinct curve that depicts relation between house prices and year, the underlying structure remains the same (showing an overall increase in price over the years). The district of Columbia has seen the biggest increase and Nevada has seen the maximum decrease in house prices. One state that can be equated as an outlier is West Virginia since its overall trend depicts a decrease in house price instead of increase, over the time span.

ggplot(housePriceData.avgPriceRegion, aes(x = Year, y = Price)) +  
 geom\_point() +  
 facet\_wrap(~ Region, ncol = 2) +   
 #geom\_smooth(method = "gam", formula = y ~ s(x), se = FALSE) +  
 geom\_line(color = 'blue') +  
 scale\_color\_manual(values = cb\_palette) +  
 ggtitle("Regionwise Plot of Year Vs. Housing Price") +  
 xlab("Year") +  
 ylab("Housing Price") +  
 theme\_bw(base\_size=10) +  
 theme(plot.title = element\_text(hjust = 0.5))



As seen region-wise, all the regions depict a similar trend of increase in house prices over the years especially in the later 2000s except Midwest which shows an overall flat trend as compared to all other regions. However, talking about absolute prices each region has its own numbers with West being the priciest of all, closely followed by Northeast. Midwest and South show quite identical trends though Midwest has an overall negligible change in house prices over the span of years.

## Population Density & Change in House Prices

data(state)  
stateAreaData = state.x77  
stateName = row.names(stateAreaData)  
stateAreaData = cbind(stateName,stateAreaData)  
stateAreaData = stateAreaData[, c("stateName", "Area")]  
rowname = seq(1, nrow(stateAreaData), 1)  
row.names(stateAreaData) <- rowname  
stateAreaData = cbind(stateAreaData[,1], as.numeric(stateAreaData[,2]))  
colnames(stateAreaData) = c('State', 'Area')  
stateAreaData = rbind(stateAreaData, c("District of Columbia", 61.05))  
#nrow(stateAreaData)

census\_api\_key("88a802fbbb65a6d48475e8b55569f2a836e11934")

## To install your API key for use in future sessions, run this function with `install = TRUE`.

census1990 = get\_decennial(geography = "state", variables = "P0010001", year = 1990)

## Getting data from the 1990 decennial Census

names(census1990)[names(census1990) == "NAME"] <- "State"  
census1990['Year'] = 1990  
census1990 = census1990[c('Year', 'State', 'value')]  
  
census2000 = get\_decennial(geography = "state", variables = "P001001", year = 2000)

## Getting data from the 2000 decennial Census

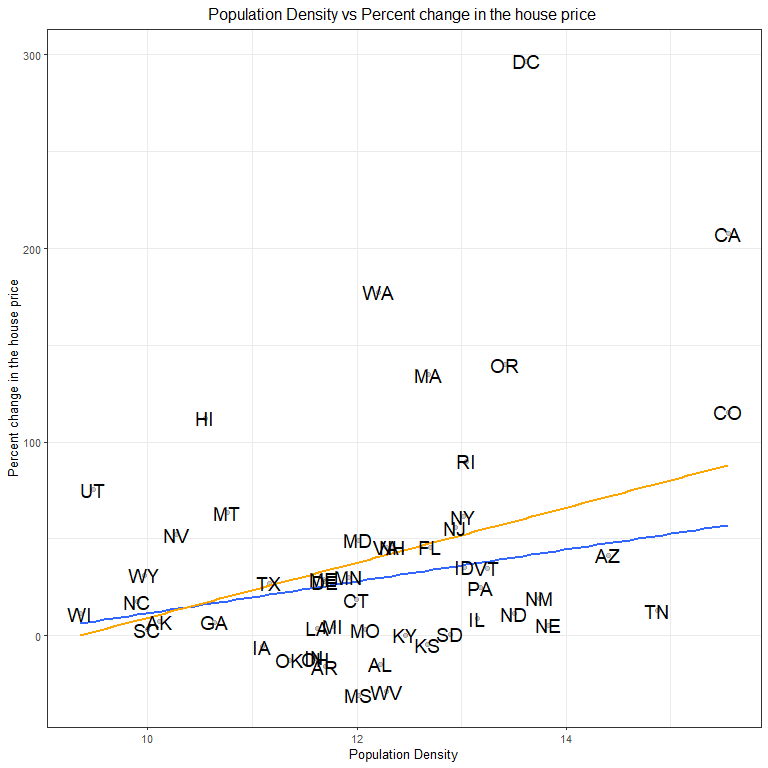
census2000['Year'] = 2000  
names(census2000)[names(census2000) == "NAME"] <- "State"  
census2000 = census2000[c('Year', 'State', 'value')]  
  
census2010 = get\_decennial(geography = "state", variables = "P001001", year = 2010)

## Getting data from the 2010 decennial Census

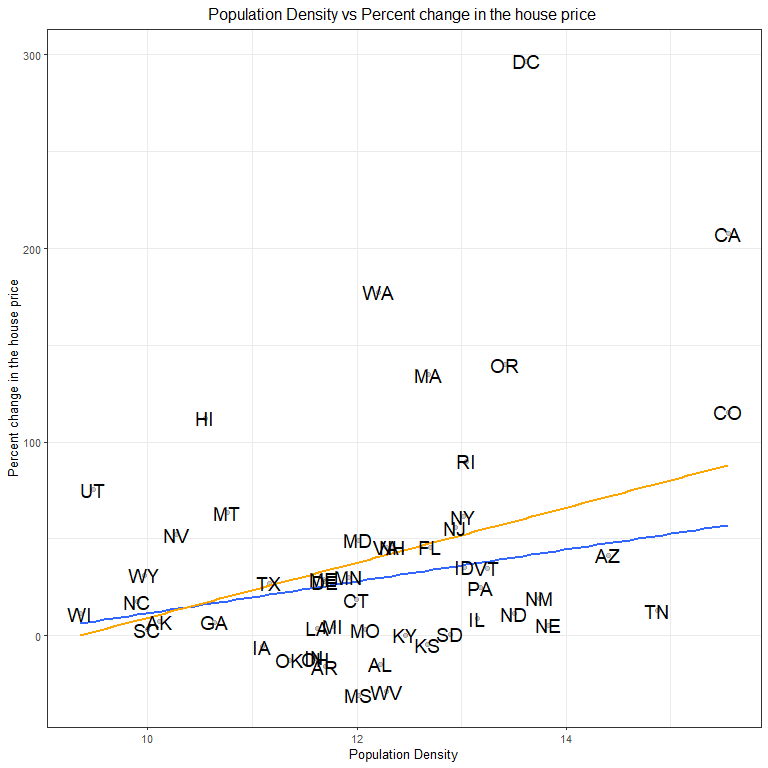
census2010['Year'] = 2010  
names(census2010)[names(census2010) == "NAME"] <- "State"  
census2010 = census2010[c('Year', 'State', 'value')]  
census2010 = census2010[0:51, ]  
  
census2018 <- get\_estimates(geography = "state",product = "population", )  
census2018 <- filter(census2018, variable == "POP")  
census2018 <- census2018[0:51,]  
stateAreaData = data.frame(stateAreaData)  
census2018 = data.frame(census2018)  
census2018$NAME = as.factor(census2018$NAME)  
names(census2018)[names(census2018) == "NAME"] <- "State"  
census2018 = subset(census2018, select = c('State', 'value'))  
year = rep("2018", 51)  
census2018$Year = year

joined\_data = inner\_join(stateAreaData, census2018, by = 'State')

joined\_data$density = as.numeric(joined\_data$value) / as.numeric(joined\_data$Area)  
dec18 = housePriceData[housePriceData$Month== 12 & housePriceData$Year== 2018,]  
jan75 = housePriceData[housePriceData$Month== 1 & housePriceData$Year== 1975,]  
#state = c(unique(housePriceData.avgPriceState$State))  
state = c(state\_abbrev$Code)  
change = c(((dec18$Price - jan75$Price)/ jan75$Price)\* 100)  
  
priceChange = data.frame(state,change, density = joined\_data$density )



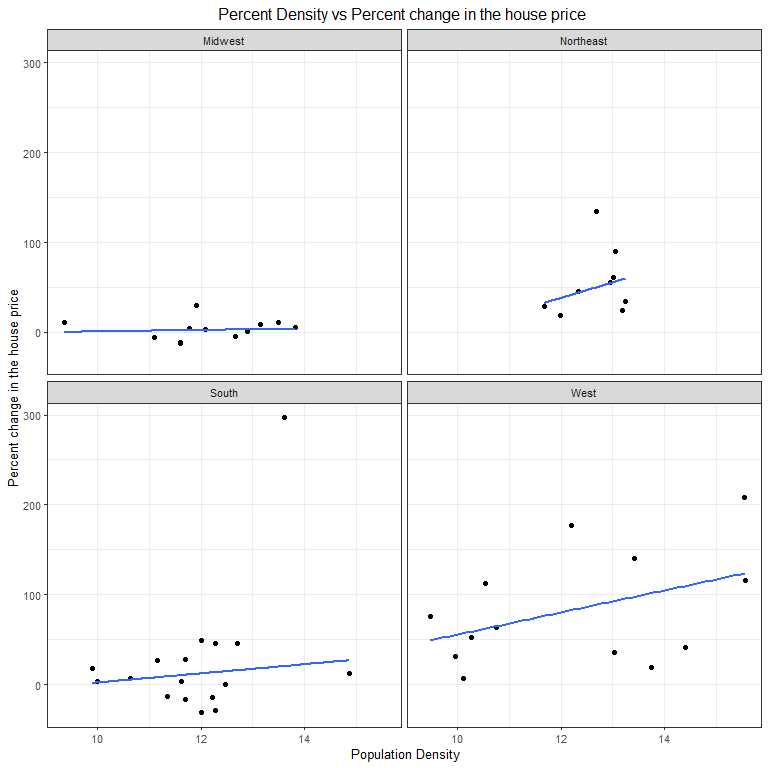
## Density vs Change in House prices after removing Outliers



With increase in density there is an increase in percent change in house prices. There are 2 outliers of DC and California which have been downweighed. This is done so because these 2 states do not follow the trend and change in house prices is extremely high. After this process the hypothesis is bolstered that with increase in density there is an increase in percent change in house prices. The removal of outliers considerably reduces the change in house prices.

## Change in House Prices vs Population Density (Region Wise)

data2018 <- filter(housePriceData, Year == 2018)  
data2018 <- filter(data2018, Month == 12)  
data1975 <- filter(housePriceData, Year == 1975)  
data1975 <- filter(data1975, Month == 1)  
dataChange <- inner\_join(data1975,data2018, by = c("State"))  
dataChange$regionPriceChange = c(((dataChange$Price.y - dataChange$Price.x)/ dataChange$Price.x) \* 100)  
density = joined\_data$density  
priceRegionChange.data = data.frame(density,dataChange )



Midwest shows no change in house price. Rest all three regions show increase in “percent changes in house price”. Though as previously observed South region shows a similar trend to Midwest which slight change in “change in house prices”.

## Changes in Population & Changes in House Prices

Population1990 <- get\_decennial(geography = "state", variables = "P0010001", year = 1990)

## Getting data from the 1990 decennial Census

Population1990 = as.data.frame((Population1990))  
Population1990 <- Population1990[c(2,4)]  
names(Population1990)[names(Population1990) == "value"] <- "Population1990"  
#head(Population1990)  
  
Population2000 <- get\_decennial(geography = "state", variables = "P001001", year = 2000)

## Getting data from the 2000 decennial Census

Population2000 = as.data.frame(Population2000)  
names(Population2000)[names(Population2000) == "value"] <- "Population2000"  
Population2000 = as.data.frame((Population2000))  
Population2000 <- Population2000[c(2,4)]  
  
#head(Population2000)  
  
Population2010 <- get\_decennial(geography = "state", variables = "P001001", year = 2010)

## Getting data from the 2010 decennial Census

Population2010 = as.data.frame(Population2010)  
names(Population2010)[names(Population2010) == "value"] <- "Population2010"  
Population2010 = as.data.frame((Population2010))  
Population2010 <- Population2010[c(2,4)]  
#head(Population2010)  
  
Population2018 <- get\_estimates(geography = "state", product = "population")  
Population2018 = as.data.frame(Population2018)  
names(Population2018)[names(Population2018) == "value"] <- "Population2018"  
Population2018 <- Population2018[c(1,4)]  
Population2018 = Population2018[1:52,]  
  
state\_abbrev = as.data.frame(read.delim(file = "state\_abbrevs.txt", header = TRUE, sep = " "))  
state\_abbrev = data.frame(lapply(state\_abbrev, as.character), stringsAsFactors=FALSE)  
names(state\_abbrev)[names(state\_abbrev)=="State"] <- "NAME"  
  
populationChanges = inner\_join(state\_abbrev, Population1990, by = "NAME")  
populationChanges = inner\_join(populationChanges, Population2000, by = "NAME")  
populationChanges = inner\_join(populationChanges, Population2010, by = "NAME")  
populationChanges = inner\_join(populationChanges, Population2018, by = "NAME")

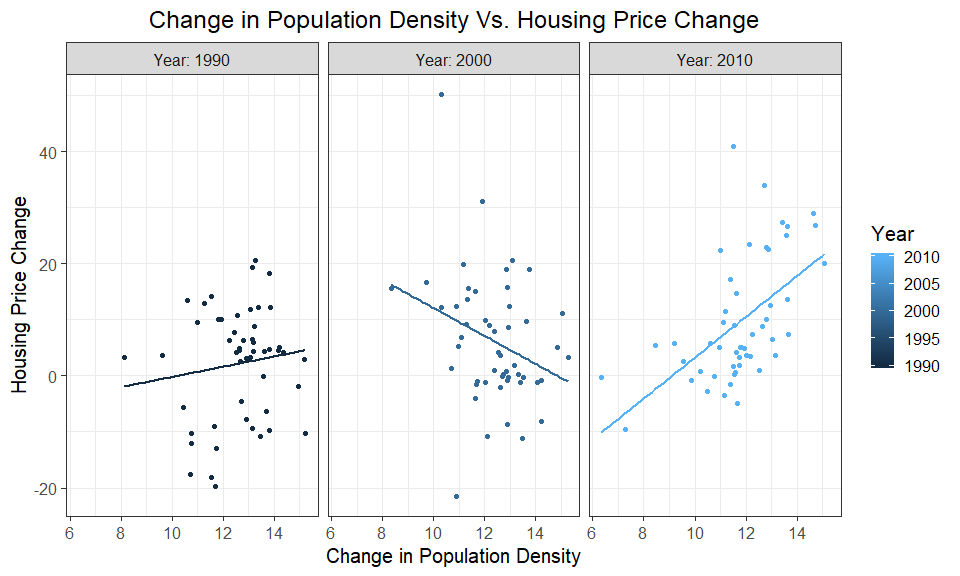
housingData = housePriceData[c(1:4,8,9)]  
  
jan1990 = housePriceData[housePriceData$Month == 1 & housePriceData$Year == 1990, ]  
dec1999 = housePriceData[housePriceData$Month == 12 & housePriceData$Year == 1999, ]  
dec1999$HousingPriceChange1999 = dec1999$Price - jan1990$Price  
dec1999 = dec1999[c(3,8:10)]  
  
jan2000 = housePriceData[housePriceData$Month == 1 & housePriceData$Year == 2000, ]  
dec2009 = housePriceData[housePriceData$Month == 12 & housePriceData$Year == 2009, ]  
dec2009$HousingPriceChange2009 = dec2009$Price - jan2000$Price  
dec2009 = dec2009[c(3,8:10)]  
  
jan2010 = housePriceData[housePriceData$Month == 1 & housePriceData$Year == 2010, ]  
dec2018 = housePriceData[housePriceData$Month == 12 & housePriceData$Year == 2018, ]  
dec2018$HousingPriceChange2018 = dec2018$Price - jan2010$Price  
dec2018 = dec2018[c(3,8:10)]  
  
housingPriceChange = inner\_join(dec1999, dec2009, by = "Code")  
housingPriceChange = inner\_join(housingPriceChange, dec2018, by = "Code")  
housingPriceChange = housingPriceChange[c(1,4,7:10)]  
  
populationChanges$PopulationChange1999 = populationChanges$Population2000 - populationChanges$Population1990  
populationChanges$PopulationChange2009 = populationChanges$Population2010 - populationChanges$Population2000  
populationChanges$PopulationChange2018 = populationChanges$Population2018 - populationChanges$Population2010  
populationChanges = populationChanges[-c(4:7)]  
  
populationHousingData = inner\_join(housingPriceChange, populationChanges, by = "Code")  
populationHousingData = populationHousingData[-c(7, 8)]  
populationHousingData = populationHousingData[c(1, 4, 5, 2, 3, 6:9)]  
  
populationHousingData1999 = populationHousingData[c(1:4, 7)]  
populationHousingData1999$Year = 1990  
names(populationHousingData1999)[names(populationHousingData1999)=="HousingPriceChange1999"] <- "HousingPriceChange"  
names(populationHousingData1999)[names(populationHousingData1999)=="PopulationChange1999"] <- "PopulationChange"  
  
populationHousingData2009 = populationHousingData[c(1:3, 5, 8)]  
populationHousingData2009$Year = 2000  
names(populationHousingData2009)[names(populationHousingData2009)=="HousingPriceChange2009"] <- "HousingPriceChange"  
names(populationHousingData2009)[names(populationHousingData2009)=="PopulationChange2009"] <- "PopulationChange"  
  
populationHousingData2018 = populationHousingData[c(1:3, 6, 9)]  
populationHousingData2018$Year = 2010  
names(populationHousingData2018)[names(populationHousingData2018)=="HousingPriceChange2018"] <- "HousingPriceChange"  
names(populationHousingData2018)[names(populationHousingData2018)=="PopulationChange2018"] <- "PopulationChange"  
  
populationHousingData = populationHousingData1999  
populationHousingData = rbind(populationHousingData, populationHousingData2009)  
populationHousingData = rbind(populationHousingData, populationHousingData2018)  
names(populationHousingData)[names(populationHousingData)=="Region.x"] <- "Region"  
  
  
library(viridis)

## Warning: package 'viridis' was built under R version 3.6.2

## Loading required package: viridisLite

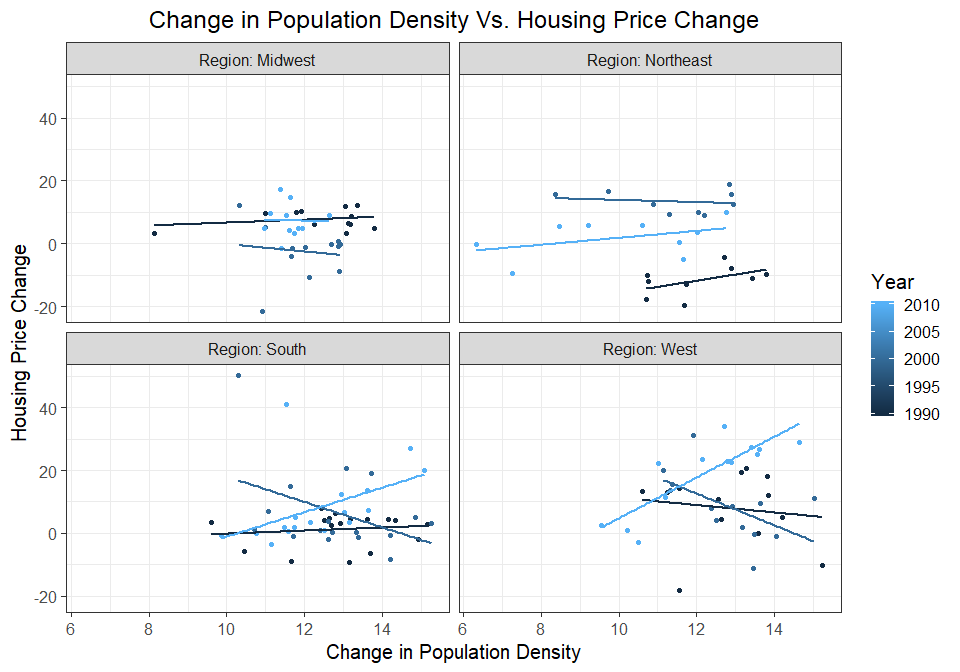
#cb\_palette = c("#999999", "#E69F00", "#56B4E9", "#009E73", "#F0E442", "#0072B2", "#D55E00", "#CC79A7")

ggplot(populationHousingData, aes(x = log(abs(PopulationChange)), y = HousingPriceChange, group = Year, color = Year)) +  
 geom\_point() +  
 geom\_smooth(method = "lm",se = FALSE) +  
 ggtitle("Change in Population Density Vs. Housing Price Change") +  
 xlab("Change in Population Density") +  
 ylab("Housing Price Change") +  
# scale\_color\_manual(values = cb\_palette) +  
 #scale\_color\_viridis(option = "cividis") +  
 facet\_wrap(Year ~ ., labeller = "label\_both") +  
 scale\_fill\_manual(values = cb\_palette) +  
 theme\_bw(base\_size=15) +  
 theme(plot.title = element\_text(hjust = 0.5))



A relationship can be observed between the two parameters of changes in population and changes in house prices. In the time period between 2010 and present the graph shows a positive trend which indicates a directly proportional relationship between the two aforementioned variables. In the time period of 2000-2010 there is a decrease in house price with respect to increase in change in population. There is a slight increase in change in house prices with respect to population change.

ggplot(populationHousingData, aes(x = log(abs(PopulationChange)), y = HousingPriceChange, group = Year, color = Year)) +  
 geom\_point() +  
 geom\_smooth(method = "lm",se = FALSE) +  
 ggtitle("Change in Population Density Vs. Housing Price Change") +  
 xlab("Change in Population Density") +  
 ylab("Housing Price Change") +  
# scale\_color\_manual(values = cb\_palette) +`  
 #scale\_color\_viridis(option = "cividis") +  
 facet\_wrap(Region ~ ., labeller = "label\_both") +  
 scale\_fill\_manual(values = cb\_palette) +  
 theme\_bw(base\_size=15) +  
 theme(plot.title = element\_text(hjust = 0.5))



The region wise graphs compel us to introspect on the various trends between changes in population and changes in house prices. Starting with the year 2018 the graph shows an upward trend for all the four regions i.e. with increase in population change there is an increase in change in house prices. In the year range 2010s the population change is in inverse proportion with change in house prices. In 1990s there are no interpretable trends.

Conclusion:

The graph for house prices shows that there is an increasing trend for the house prices over the year. It is also a known fact that the population of various states has grown over the year. Keeping the above constraints in mind it is expected to observe a positive linear relationship between the changes in the population over the year with respect to the changes in house prices. The house prices here are adjusted to the CPI because we need to take inflation into account while dealing with the time series analysis of any financial data. Hence, on observing the data and knowing the logical trends and the graphs observed from the data plotted above, the natural conclusion that Changes in house prices with respect to change in Population shows a positive relationship can be derived.