Statistics 133 Project Report:Determinants of Home Price and Its Fluctuation over Time in California

Group 14(A+): Hong Jabbari,Yangyang Li,Zhiyue Hu,and Renjun Zhu

May 6th, 2016

Source file ⇒ Project_Report.Rmd (data:text/x-markdown;base64,LS0tCnRpdGxlOiAiU3RhdGlzdGljcyAxMzMgUHJvamVjdCBSZXBvcnQ6RGV0ZXJtaW5hbnRzIG9mIEhvbWUgUHJpY2UgYW5klApJdHMgRmx1

1. Introduction

The idea of having a home dates back 12 to 15,000 years. Home, brings people closer; home, generates family; home, sweet home. As soon as ownership of land or property developed, houses quickly became commodities that had their own market value. Purchasing a home was, is, and will be the dream for many centuries in cultures across the globe. Home is not just a place of shelter, but more a place for you to experience life and leave traces of that life for future generations. Throughout modern history, the topic of affordable housing has remained a significant force in society. From individual concerns to the moving curve of global economy, home price is one of the most important indicators to measure the growing wealth in a society. Many issues are related to real estate, such as the housing bubble that triggered the financial crisis in 2008. Therefore, it is worthwhile to take a close look at the determinants of home prices and their fluctuation over time.

California is located on the West coast of the United States and crosses a wide latitude, spanning some of the most pleasant climates. It comes as no surprise that it attracts people from so many different cultures to settle down, and contains some of the most expensive residential areas for home buyers. Thus, California can be used as a suitable model from which to examine the determinants of home prices on a regional level over time.

Our project started with brainstorming, making educated guesses about what components might affect the prices of homes. We listed reasonable determinants such as population, property taxes, household income, school district, unemployment rate, crime rates, and geographical location away from the seashore. However, due to some limitation in obtaining the data, we began by analyzing three of these from above list: population, unemployment rate, and school district performance. To test our hypothesis, first we needed to configure our tidy table[see Appendix: table1] using a total of 320 cities in California, making columns that contain numerical values of Home Price, Population, unemployment rate, and high school API score. In that table, to choose a single home price within a year range, we take the mean value of that year across monthly values. Additionally, home prices are usually viewed as a large number with five to six digits when using dollar as the base unit, so for convenience, we take the integer part of the the price and remove the fractional part. We fix population by looking at city population. Naively, because higher population indicates more competition, so we hypothesize that cities with huge population will have higher home price. For unemployment rate, within a scope of cities, high unemployment rate means less people working, and this indicates fewer people will have consistent income to support purchasing a house, so we claimed that high unemployment rate will correspond to a relatively low housing price in that city. In terms of high school district, we focus on the Academic Performance Index (API), a measurement of academic achievement for all the public high schools quantified by the California Department of Education. In this case, we do not consider private schools because the rich 1% can send their children to anywhere they want, so it would not affect the decision-making by our general population that much.

Secondly, we also want to compare the home price geographically by subdividing California into regions both horizontally(latitude) and vertically(coastal vs. inland). For horizontal subdivision, we roughly define the North(1-4), Central(5-7), and South(8-11) regions by their latitude with relatively the same climate as the graph below, and make a new tidy table [see Appendix: table2] which groups all the cities into these three regions, and then takes home price as a variable. For vertical subdivision, we take compare the coastal regions with those more inland. In comparing home value between the coast and inland, we needed to do more data wrangling [see table 3] using figures from a span of 10 years beginning in 1996, and overlaying these data onto a California map to see the variation over the regions vertically.

Finally, we picked some representative cities in each region to see the change of their home prices on a timeline, and made a line graph from 1996 to present to see their home price moving curve. Namely, we picked San Francisco and Cupertino to represent cities from North, Bakersfield as a representative city of the central region, and Los Angeles and Compton for the South. We choose these cities, because some of them are big cities in California, and some of them have recently become well-known for their skyrocketing home prices, so it would be interesting to trace their home price history and compare with the present.

We raise the following questions: (1) Will our initial guess of choosing these three determinants really affect home price in reality? (2) If they do, how do these three determinants affect home prices, and to what degree of correlation? (3) Since California crosses a large latitude, if we divide it into North, Central, and South roughly, what will the home price be for each region in such a comparison? (4) Then we ask a similar question about the regions from the coast to inland? (5) Finally, within a 10 year span, how do some cities' home prices fluctuate?

In order to give different perspectives of our analysis and enhance the accuracy of calculation, we use R and specifically use the ggplot2 library to generate dot plots with an added layer of a linear regression model to visualize the correlation of three determinants with home price [plot: page 6 & R code: Appendix]. Additionally, we use violin and density plots to visualize the variation of home prices among North, Central, and South regions [plot: page 7-8 & R code: Appendix]. Furthermore, our individual houses over one million dollars will be represented by the animation through google earth. We will also make a California map to map the home price of 2576 cities into different alpha levels of the same color in order to see the geographic differences of home price between the coast and inland [plot: page 9 & R code: Appendix]. Finally, we will put those five selected cities into their time series, and see the price fluctuation over the last decade [plot: page 11 & R code: Appendix].

2.Data Resource:

Web searching allows us to collect huge data from 1996 to present, and our source came from different websites, namely zillow.com, a database website which provides home prices; labormarketinfo.edu.ca.gov, where gives us detailed information about unemployment rate; dof.ca.gov, a government website which updates the city population in California each year; and cde.ca.gov, an official website run by California government to report the API score of each public high schools. We are aiming to get data for these three determinants from 2010 to 2014 because the year before and around 2008 have been widely studied by other researcher and scholars due to the financial crisis happened in 2008. In order to avoid the redundancy, we are more interested in the analysis of recent year home price after the depression in 2008. To see the tendency of each determinant versus home price, we put all the 5 years together to get the graph for each determinant. We use the same set of data and do further cleaning for the region dividing graph. However, to make the map, the animation, and the graph of home price fluctuation time series, zillow.com again provides us data from 1996 to present.

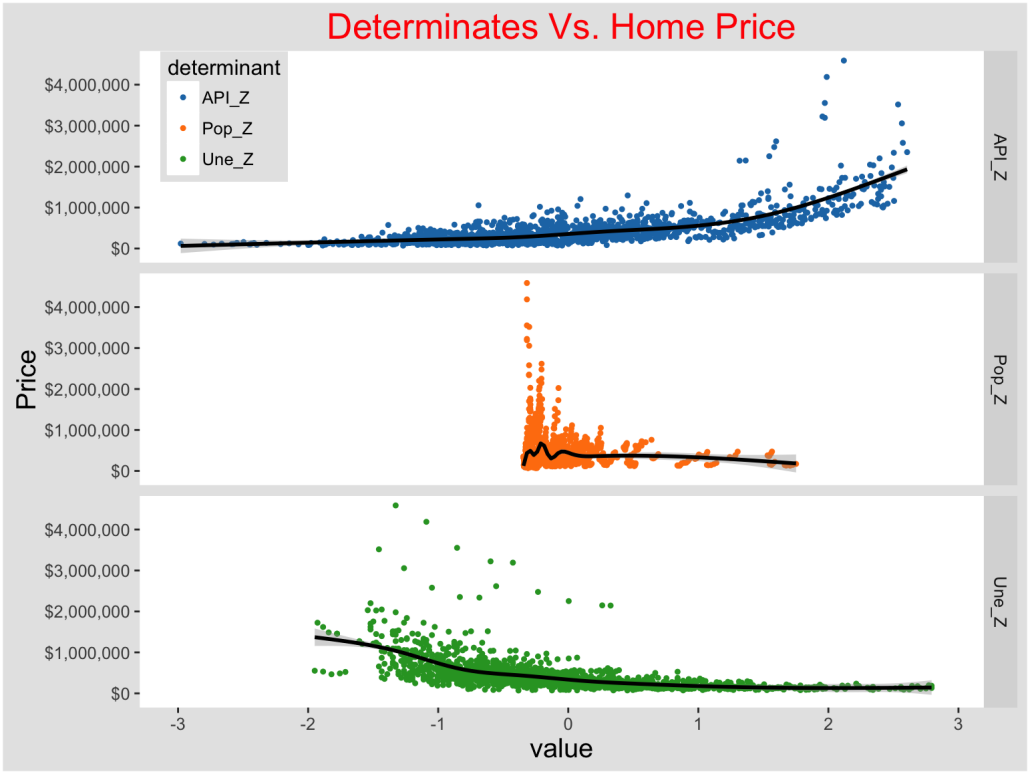
3. Analysis and Visualization:

3.1 Determinant of Home Price: Population, API and Unemployment Rate

In order to compare the effectiveness of these three determinants, we have to normalize each determinant by the idea of Z value: by applying our knowledge from statistics class, we use below formula to make such normalization; hence, we can compare them in the same graph:

$$z = \frac{x_{ij} - \bar{x_i}}{\sqrt{\frac{\sum_i (x_{ij} - \bar{x_i})^2}{(n-1)}}}$$

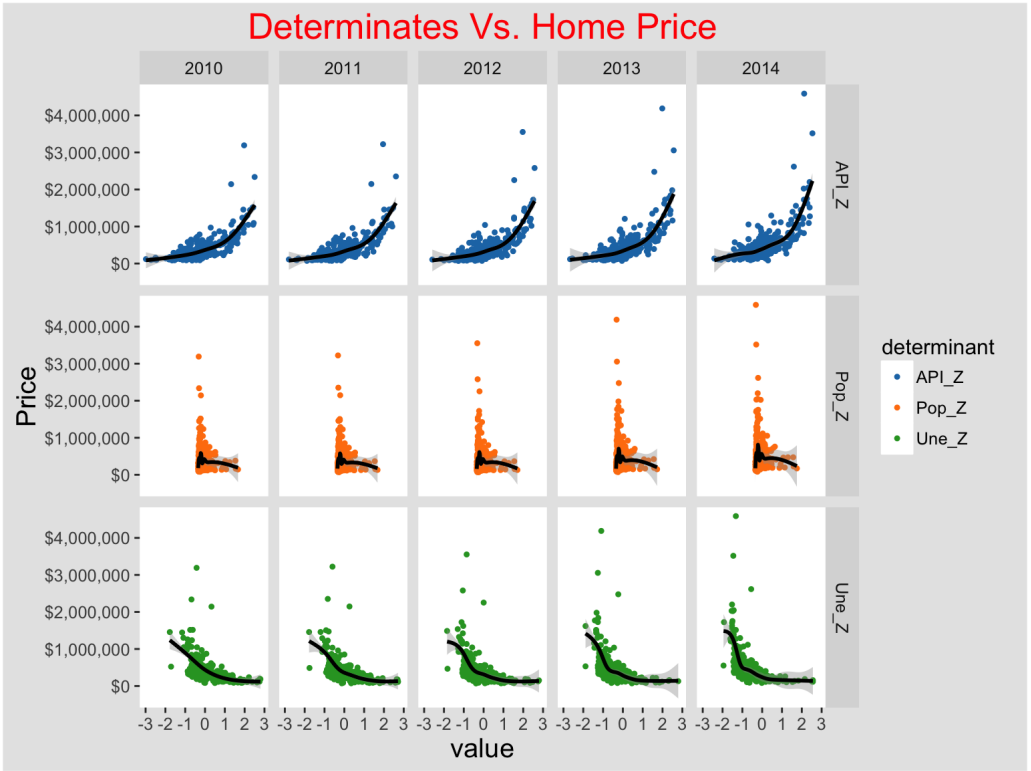
Note: i = each determinant (Population, API, and Unemployment Rate) j = each city



Recall our naive guess such that cities with large population tend to have higher home price. To test this, we plotted 320 cities in 5 years with their population and average home prices. Surprisingly, the correlation coefficient of the regression r = -0.01439502 is a negative value and very closed to zero, which shows that the home price is almost not related to Population. Setting up 95 percent confidence interval, we applying t-test with 1559 degree of freedom, and get the p-value = 0.5698. This large p-value (greater than 0.5) will reject our hypothesis that Population is a determinant, which affects home price significantly. In addition, even though the regression lines within these 5 years are very stable, by comparing 2010 through 2014 y-axis (Home Price), observe the fact of home price shifting up each year is quite obvious. This makes sense because our living standard is moving up, and the inflation rate for currency is still on-going and we didn't consider the inflation rate for home price yet when collecting data.

The first panel has showed a strong correlation with r = 0.6908296 between High School API Score and home price, which respond our initial hypothesis that high school with higher academic performance tend to have higher housing around them. Same t-test applied here with 1559 degree of freedom, and we also get the same p-value < 2.2e-16 as determinant unemployment rate. Comparing such small p-value with 95 percent of confident interval rate, which supports our initial claim. Despite the fact that not every homeowner around school district carries children in the house, having an house near school district can alternately become an investment for them to rent out. This argument can explain that within our 5 consecutive years, the slope of each line become steeper and steeper, so more and more people like to invest their money on school district, and the tendency of home price rise higher as time goes by.

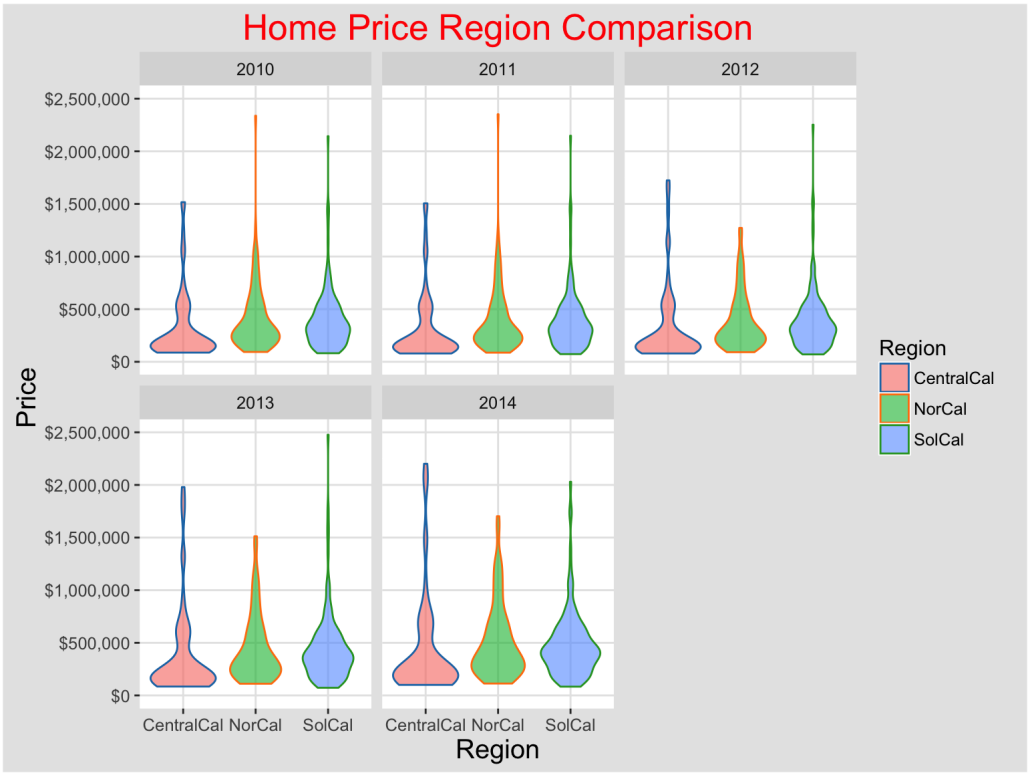
The third panel of unemployment rate versus home price below echo our claim that high unemployment rate will negatively affect home price. There is a -0.5147542 correlation between unemployment rate and home price. Looking at P-value, less than 2.2e-16 indicates a significant negative correlation between home price. History has shown that most homeowners purchase house with a stable job because they need the monthly salary to support the mortgage. However, in recent year, there is also a fair amount of the homeowners use all cash to pay for the total amount at once, and of course this group is not bounded by job security. It would be normal to have such an okay correlation value of -.5 rather than a strongly negative correlation such as values closer to -1. Moreover, within 2010 to 2014 these consecutive years, the lines are tilted up gradually. Simply by looking the intersection of each lines with zero value of home price. Notice that not just the unemployment rate is getting closer towards smaller value in order to reaching zero value of home price, all the dots in each plot have been shifted inner towards within 5 % of unemployment rate, and the tails are getting diminished. Such a shift shows that our economy environment is getting better each year after the recession of 2008.



3.2 Comparison of North, Central, and South

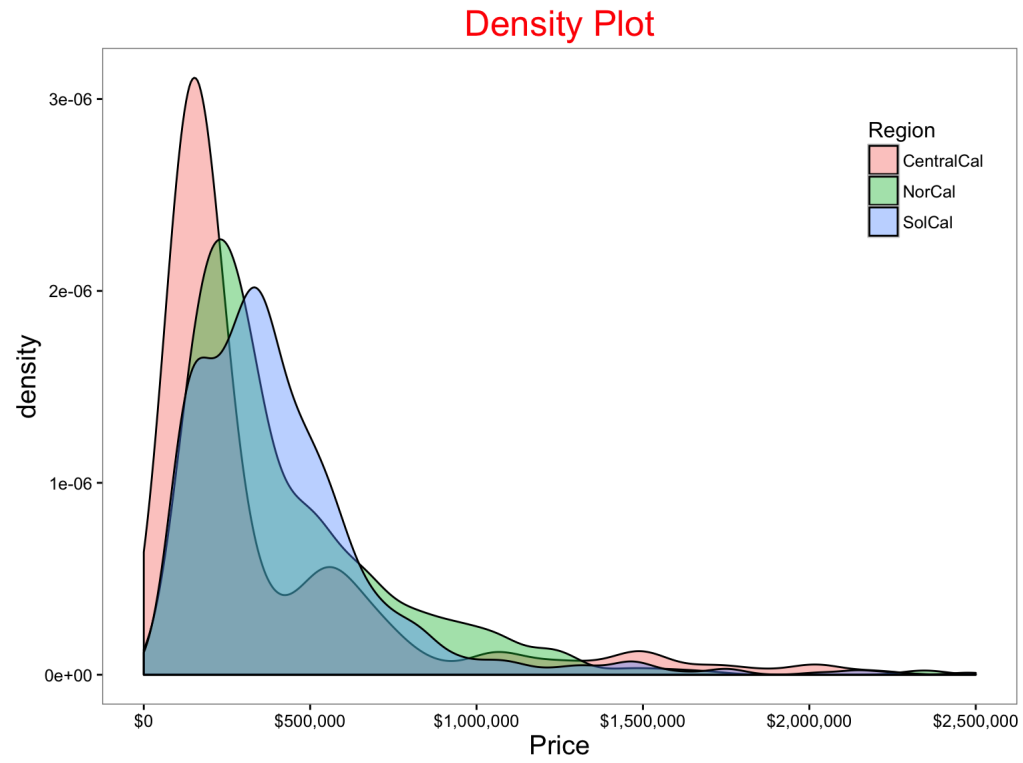
3.2.1 Graphs in Different Year

In terms of comparing three regions(North, Central, and South) within California. We plot a violin graph to show the variation of home prices by region. By looking at the shapes of each "violin" along time, the graph below has suggested that the "fat" shape of Central California in the bottom part of its "violin" stays pretty constant below 200,000 dollars, and also there is a small wave at around 500,000; however, after 2011, more and more over million dollars houses appear, and its "violin" shape started to grow bigger and shift up a little, especially the year in 2014. For Northern California, the "violin" shape shrank a little bit at 2012 and started to grow back the year after, but roughly the mean value of home prices is at around 250,000 and gradually narrow down as the price goes up. Also, besides the vertical growing, it also increases the size of its home price at any level horizontally, Finally, Southern California inherits about the same shape over time, it also has mean value higher than the other two regions with around 300,000 dollars and beyond. At year 2014, the quantity of home buyers corresponding to home price at each level below 2 million has increased significantly. Moreover, the "leaf" shape of Southern California indicates its little variation of home price compared to other regions.



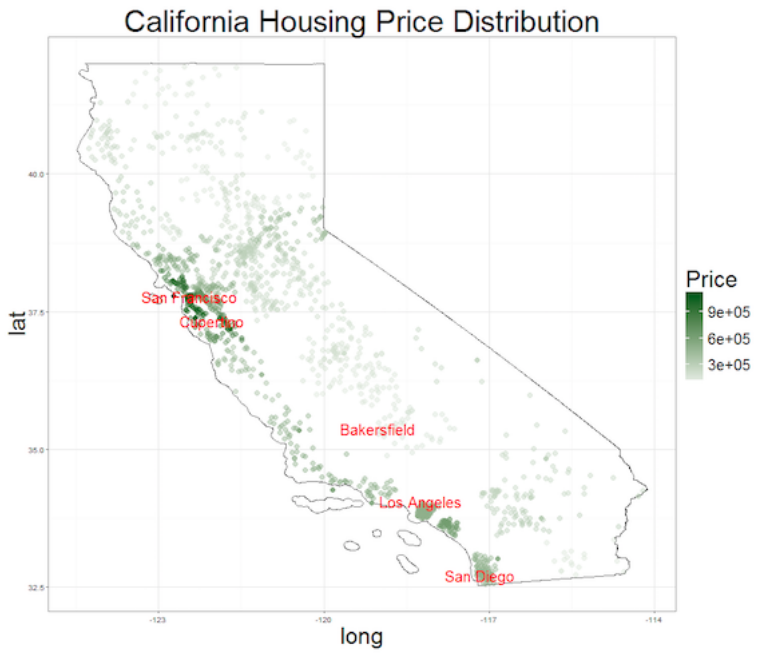
3.2.2 Graph as a Whole (disregard year)

If disregard the distinguishes between each year, and put these three regions into a graph by looking at their density, we can also conclude their prices variation. The red graph(Central) has a sharp peak suggests that the high density of buyers to purchase their home in the price below 250,000 dollars, and this matches our violin graph. Same argument apply to Northern and Southern California. If by looking at the width of each density plot, then we can conclude the variation of home prices within each region, which has a better visualization compared to the violin graph above. The graph has shown that Southern California has wider range in terms of variation, because it is the widest among all three regions why applying a horizontal line at any density(y-axis), then next is followed by Northern California. It is noticeable that Northern California significantly surpasses other two regions on approximately 0.75 million to 1.5 million range, while its peak is less expansive than that of Southern California's. We examine our data and conclude that this density makes sense as the price of the bay area and that of the Silicon Valley are booming, while in the far-northern of California, such as Modoc, housing price stays low.



3.3 From the Coast to Inland

Besides the glance of putting California into three regions horizontally, it is worthwhile to study the home price vertically from the coast to inland. The map below maps 2576 cities(or towns) into a dot, which represent in a different range of home price by the green color: the darker, the more expensive. It appears the darker color are mostly in the coast area, especially in bay area and the south bay area. Also, by looking at the overlapped dots, we notice that some areas such as longitude around 37.5 and 33.5 have more houses both along the coast and inland despite the variation of their home prices in comparison with other places with the same latitude.

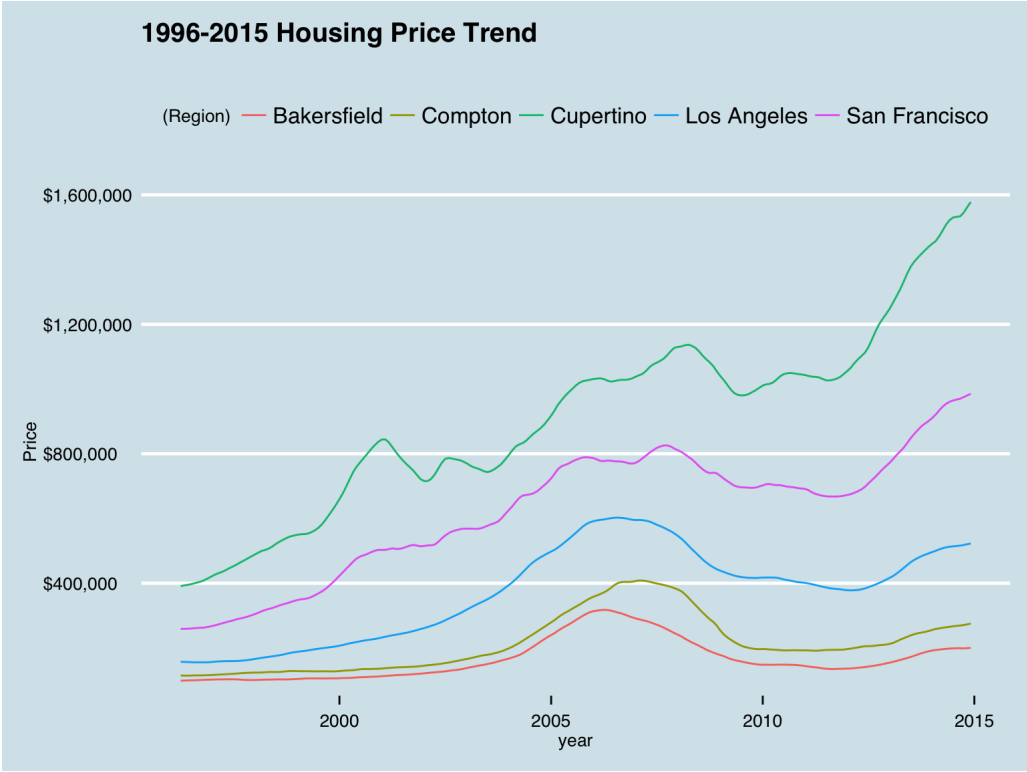


3.4 Over One Million Dollars Home:

In order to represent the dynamic movement of home price within regions, we collect data from 1996 to 2016, and filter out the the location(latitude and longitude) of houses with price over 1 million dollars, and organized in a kml.file which will use google earth software to present during our 6 minutes slides presentation during RRR week. (google earth: May, 4th & Code: Appendix) In 1996, most cities with an average price of more than 1 million are located in Southern California, for example, the infamous Rancho Santa Fe and Beverly Hills. The animation shows that within these 20 years, started from Southern California, moving up along the coast, more and more over million home was built. Especially in bay area, more precisely in the south bay along Silicon Valley, million dollars houses are piled. We analyzed that many big companies such as Apple, Google, ebay, and Yahoo set their headquarter in this region. Many of those are high technology companies with a great market value, and a well-known fact is that 21st century is an era of technology, so no surprising that tech-related jobs with high-paying salary are always in demand. With higher income and opportunities in this region, the home price collectively over million dollars will not be considered as a legend. Although generally the pins keep appearing, there is, however, a significant wave of disappearing of pins from 2008 to 2011, which indicate the effect of financial crisis on housing prices.(Partial Animation)

3.5 In a time frame

Finally, we extract some representative cities such as San Francisco, Cupertino, Bakersfield, Los Angeles, and Compton into a time frame from 1996 to present. Despite the effect of inflation, the tendency of home price is in an upslope growing trend. However, as stated before, home price dropped on the year 2008 due to the depression. It is amazing to see city like Cupertino, compare the price ten years ago, the home value quadruple. Because of the rise of Silicon Valley, the prices of regions around surge. Also, Cupertino attracts many wealthy new immigrants from China recently, who are both able and willing to spend more money in order for their children to receive better education. There is no surprise that San Francisco home price triple, and Los Angeles double. They are the two major and large cities in California, the history of becoming the first home for many settlers from the westward movement in the 19 century and the development of the city within two hundred years guarantee the rises of their home values. Now, let us move our attention to Bakersfield and Compton, they are pretty much stay at the same level and do not have much fluctuation. Bakersfield is located in the middle of a desert, and considered as Central California. Its name remembered by the location that would regard as a must pass-by city if driving east to Las Vegas. Surprisingly, Compton as seem to be one of the poorest cities in Southern California, and its home value is even higher than Bakersfield, the biggest and the most representative city in Central California. The "stock-chart-liked" graph below provides us a micro-level analysis of the home price fluctuation through the lens of cities.



4. Conclusion:

Through the collaboration of this group project, we found there is almost no correlation between home price and city population. While, there is a strong correlation with high school API score, and a moderate level of negative correlation with unemployment rate. Along the way of our analysis of these two determinants, we also discovered the economic growth of our nation within the 5 years after the recession began in 2008. In addition to demonstrating conclusions about our several hypotheses, we also visualized the variation of home price over a span of 5 years across the map of California, divided into North, Central, and South regions. The map of California home price highlights the coastline home value compared to inland, and also outlines the general region of financial centers inside of California. Moveover, the animation using google earth pinpoints the aggregation of wealth corresponding to technological intelligence. Finally, the home price fluctuation graph, with the five selected cities, depicts the significant impact on home prices due to some unfortunate stressors in the recent history.

Our study of the determinants of home price and its fluctuation in California only scratches the surface of this subject, no matter which period of time we selected. With respect to the fluctuation of home prices curve, we only made cursory descriptions of the change. Beyond the skills that we applied from our knowledge of statistics, there are also other cross-disciplinary perspectives, which when combined can provide a wider frame of reference to investigate this economic phenomenon.

Appendix:

Table 1:(5 years data) serve as our basis table, generate plot 1 and plot 2

```
table1 <- read.csv("/Users/renjunzhu/Downloads/R Study/POP+UN+PRI+API.CSV")
head(table1, 10)
```

```
##     X     Area Year      API Population Unemployment_Rate  Price
## 1   1   Albany 2010 886.0000      18539               7.2 558500
## 2   2   Albany 2011 896.8000      18343               6.6 540000
## 3   3   Albany 2012 901.2000      18481               5.7 556775
## 4   4   Albany 2013 906.0000      18483               4.7 666358
## 5   5   Albany 2014 901.8000      18457               3.8 721783
## 6   6 Alhambra 2010 800.9444      83089               8.3 456200
## 7   7 Alhambra 2011 809.6667      83362               8.1 443158
## 8   8 Alhambra 2012 810.8333      83841               7.2 429733
## 9   9 Alhambra 2013 826.0000      84390               6.4 474442
## 10 10 Alhambra 2014 826.7222      84736               5.4 532250
```

Table 2:(5 years data) add on top of table 1, generate plot 3 and plot 4

```
table2 <- read.csv("/Users/renjunzhu/Downloads/R Study/joinfinal.csv")
head(table2, 10)
```

```
##     X CountyName Region       Area Year      API Population
## 1   1    Alameda NorCal   Piedmont 2010 938.4000      10667
## 2   2    Alameda NorCal    Hayward 2012 727.0625     147090
## 3   3    Alameda NorCal   Berkeley 2013 839.2000     116118
## 4   4    Alameda NorCal   Berkeley 2012 831.0667     114807
## 5   5    Alameda NorCal   Piedmont 2011 942.8000      10708
## 6   6    Alameda NorCal Union City 2014 774.0000      72109
## 7   7    Alameda NorCal  Livermore 2014 842.4000      84815
## 8   8    Alameda NorCal   Berkeley 2011 814.0667     113918
## 9   9    Alameda NorCal    Hayward 2010 703.3750     144186
## 10 10    Alameda NorCal    Hayward 2011 718.8750     145090
##    Unemployment_Rate   Price
## 1                6.3 1258150
## 2               12.0  279092
## 3                5.8  769183
## 4                7.0  653592
## 5                5.8 1209692
## 6                5.2  583392
## 7                4.0  604542
## 8                8.2  618150
## 9               14.9  288500
## 10              13.8  272908
```

Table 3-4:(10 years data) generate plot 5(map) and google earth animation.

table 3(bbtable): table 3 is the raw data->bbtable is after first layer cleanning.

```
table3 <- read.csv("/Users/renjunzhu/Downloads/R Study/City_Zhvi_AllHomes.csv")
head(bbtable , 10)
```

```
##           Region       Year  Price       year
## 1    Los Angeles 1996.04.01 157300 1996-04-01
## 2  San Francisco 1996.04.01 258700 1996-04-01
## 3    Bakersfield 1996.04.01  98900 1996-04-01
## 4        Compton 1996.04.01 114900 1996-04-01
## 5      Cupertino 1996.04.01 390900 1996-04-01
## 6    Los Angeles 1996.05.01 156800 1996-05-01
## 7  San Francisco 1996.05.01 259100 1996-05-01
## 8    Bakersfield 1996.05.01  99300 1996-05-01
## 9        Compton 1996.05.01 114400 1996-05-01
## 10     Cupertino 1996.05.01 393300 1996-05-01
```

table 4:

```
table4<- read.csv("/Users/renjunzhu/Downloads/R Study/cityloc.csv")
head(table4, 10)
```

```
##    zip_code latitude longitude       city state    county
## 1       501 40.92233 -72.63708 Holtsville    NY   Suffolk
## 2       544 40.92233 -72.63708 Holtsville    NY   Suffolk
## 3       601 18.16527 -66.72258   Adjuntas    PR  Adjuntas
## 4       602 18.39310 -67.18095     Aguada    PR    Aguada
## 5       603 18.45591 -67.14578  Aguadilla    PR Aguadilla
## 6       604 18.49352 -67.13588  Aguadilla    PR Aguadilla
## 7       605 18.46516 -67.14149  Aguadilla    PR Aguadilla
## 8       606 18.17295 -66.94411    Maricao    PR   Maricao
## 9       610 18.28869 -67.13970     Anasco    PR    Anasco
## 10      611 18.27953 -66.80217    Angeles    PR    Utuado
```

R Code:

Table 1:

```
## After download data from zillow.com for house price, calculate the mean value of home price from 2010 to 2014

multifam <- read.csv("City_Zhvi_AllHomes.csv")
multifamca <- multifam %>%
  filter (State=="CA")  %>%
  select(RegionName, X2010.11:X2014.12) %>%
mutate(avg14=round(1/12*(X2014.01+X2014.02+X2014.03+X2014.04+X2014.05+X2014.06+X2014.07+X2014.08+X2014.09+X2014.1
0+X2014.11+X2014.12)),
avg13=round(1/12*(X2013.01+X2013.02+X2013.03+X2013.04+X2013.05+X2013.06+X2013.07+X2012.08+X2013.09+X2013.10+X201
3.11+X2013.12)),     avg12=round(1/12*(X2012.01+X2012.02+X2012.03+X2012.04+X2012.05+X2012.06+X2012.07+X2012.08+X2
012.09+X2012.10+X2012.11+X2012.12)),
avg11=round(1/12*(X2011.01+X2011.02+X2011.03+X2011.04+X2011.05+X2011.06+X2011.07+X2011.08+X2011.09+X2011.10+X201
1.11+X2011.12)),
avg10=round(1/2*(X2010.11+X2010.12)))

### make a narrow table to get ready for joining with other determinants

multifamcamean <- multifamca %>%
select ( RegionName,avg14:avg10) %>%
arrange(RegionName) %>% gather(key=Year,value=pricemean,avg14,avg13,avg12,avg11,avg10)

### change the structure of character to numeric

subs = function(x){
  for (i in 1:length(x)){
    if (grepl("avg14", x[i])){
      x[i] = 2014
    }
    else if (grepl("avg13", x[i])){
      x[i] = 2013
    }
    else if(grepl("avg12", x[i])){
      x[i] = 2012
    }
    else if(grepl("avg11", x[i])){
      x[i] = 2011
    }
    else if(grepl("avg10", x[i])){
      x[i] = 2010
    }
  }
  return(x)
}
multifamcamean$Year = subs(multifamcamean$Year)
str(multifamcamean)

## After download data from labormarketinfo.edu.ca.gov for unemployment rate

UR <-read.csv("/Users/renjunzhu/Downloads/R Study/1014UR.csv")
UR1 <- UR %>%
  filter(!grepl("CDP",Area))%>%
  mutate(Area=gsub(" city$| town$","",Area))%>%
  select(-Period,-Adjusted,-Preliminary)
UR1$Year <- as.character(UR1$Year)
str(UR1)

## Join Home Price with determinant Unemployment rate into one table

colnames(multifamcamean)[colnames(multifamcamean)=="RegionName"] <- "Area"
Right <- multifamcamean %>%
select(Area,Price,Year) %>%
group_by(Area) %>%
arrange(Area)
join<- merge(UR1,Right,all=FALSE)

## After download data from dof.ca.gov for population

pop <-read.csv("/Users/renjunzhu/Downloads/R Study/1015population.csv")
population <- pop %>%
select(COUNTY.CITY,X4.1.10:X1.1.14)
temp = read.csv("1015population.csv", header=TRUE, sep=",")
subs = function(x){
  for (i in 1:length(x)){
    if (grepl("avg10", x[i])){
      x[i] = 2010
    }
    else if (grepl("avg11", x[i])){
      x[i] = 2011
    }
    else if(grepl("avg12", x[i])){
      x[i] = 2012
```

```
    }
    else if(grepl("avg13", x[i])){
      x[i] = 2013
    }
    else if(grepl("avg14", x[i])){
      x[i] = 2014
    }
  }
  return(x)
}
colnames(temp) = c("City", "avg10", "avg11", "avg12", "avg13", "avg14")
temp = na.omit(temp)

### In the original csv file, the population data is a factor, needs to "numericalize" it.

temp$avg10 = as.numeric(gsub(",", "", temp$avg10))
temp$avg11 = as.numeric(gsub(",", "", temp$avg11))
temp$avg12 = as.numeric(gsub(",", "", temp$avg12))
temp$avg13 = as.numeric(gsub(",", "", temp$avg13))
temp$avg14 = as.numeric(gsub(",", "", temp$avg14))
temp2 = temp %>%
select(City, avg10, avg11, avg12, avg13, avg14)
Population = temp2 %>%
gather(Year, Population, avg10:avg14)
Population$Year = subs(Population$Year)
join2 <- join %>%
group_by(Area) %>%
arrange(Area)
population2 <- Population %>%
group_by(Area) %>%
arrange(Area)
join3 <- merge(population2,join2,all=FALSE)
write.csv(join3,file="POP+UN+PRI.csv")

## After download data from cde.ca.gov for high school API score.

### Cut special schools, such as deaf schools, continuation schools, regional occupational program as those are p
ublic high schools that can be accessed by all cities schools and thus are directly under the management of Count
y office of education, not under any specific city school district.

temp = read.dbf("14avgdb.dbf", as.is = FALSE)
temp2 = temp %>%
select(sname, cname, dname, api11, api12, api13)
delete_na = function(x){
  cutlist = c()
  for (i in 1:nrow(x)){
    for(k in 4:6){
    if (is.na(x[i,k])){
      cutlist = c(cutlist, i)}
    }
  }
  return(cutlist)
}
cutlist = delete_na(temp2)
temp3 = temp2[-cutlist,] #Cut School that does not have an API score
tempAPI = temp3 %>%
filter(grepl("City|Unified|Union|Elementary|District", dname)) %>%
na.omit()

### Some schools are missing data on 2010 and thus marked as "B".

temp = read.dbf("api10gdb.dbf", as.is = FALSE)
temp2 = temp %>%
select(SNAME, CNAME, DNAME, API09, API10)
temp3 = temp2 %>%
na.omit()
tempAPI2 = temp3 %>%
filter(grepl("City|Unified|Union|Elementary|District", DNAME))
colnames(tempAPI2) = c("sname", "cname", "dname", "api09", "api10")
tempAPI2 = tempAPI2[!grepl("B|C", tempAPI2$api09),]
extractcombine = function(x){
  foo = data.frame(dname=c(),
                   city=c(),
                   cname=c())
  for (i in 1:length(x)){
    tablelist = readHTMLTable(x[i])
    table = tablelist[[2]]
    colnames(table) = c("dname", "city", "cname")
    foo = rbind(foo, table)
  }
  return(foo)
```

```
}
page1 = "http://www.greatschools.org/schools/districts/California/CA"
page2 = "http://www.greatschools.org/schools/districts/California/CA/2"
page3 = "http://www.greatschools.org/schools/districts/California/CA/3"
page4 = "http://www.greatschools.org/schools/districts/California/CA/4"
pagelist = c(page1, page2, page3, page4)
tempDistrict = extractcombine(pagelist)

tempDistrict$dname=gsub("[[:blank:]]School[[:blank:]]District","",tempDistrict$dname)
APItabletemp = inner_join(tempAPI, tempDistrict, by = c("dname", "cname"))
APItable = APItabletemp %>%
inner_join(tempAPI2, by = c("dname", "cname", "sname")) %>%
na.omit()
APItable$api09 = as.numeric(as.character(APItable$api09))
APItable$api10 = as.numeric(as.character(APItable$api10))
APItable$api11 = as.numeric(as.character(APItable$api11))
APItable$api12 = as.numeric(as.character(APItable$api12))
APItable$api13 = as.numeric(as.character(APItable$api13))

### Some of the data are numeric while others are characteristic, thus need to be converted in this form.

API_almost = APItable %>%
group_by(city) %>%
mutate(avg14 = mean(api13)) %>%
mutate(avg13 = mean(api12)) %>%
mutate(avg12 = mean(api11)) %>%
mutate(avg11 = mean(api10)) %>%
mutate(avg10 = mean(api09))
API = API_almost[!duplicated(API_almost["city"]), ] %>%
select(city, avg10, avg11, avg12, avg13, avg14)

APIG = API %>%
gather(Year, API, avg10:avg14)
subs = function(x){
  for (i in 1:length(x)){
    if (grepl("avg10", x[i])){
      x[i] = 2010
    }
    else if (grepl("avg11", x[i])){
      x[i] = 2011
    }
    else if(grepl("avg12", x[i])){
      x[i] = 2012
    }
    else if(grepl("avg13", x[i])){
      x[i] = 2013
    }
    else if(grepl("avg14", x[i])){
      x[i] = 2014
    }
  }
  return(x)
}
APIG$Year = subs(APIG$Year)
colnames(APIG)[colnames(APIG)=="city"] <- "Area"

### Due to the limitation of source, we need to get 2015 data seperately

APIG2 <- APIG %>%
group_by(Area) %>%
arrange(Area)

## Join table for table 1:

join4 <- merge(APIG2,join3,all=FALSE)
write.csv(join4,file="POP+UN+PRI+API.CSV")
```

Table 2:

```
## Making three vectors represent the regions, which contain the counties corresponding to the map above.

NorCal <- c("Humboldt","Mendocino","Lake","Sonoma","Siskiyou","Modoc",
"Trinity","Shasta","Lassen","Tehama","Plumas","Butte","Glenn","Colusa","Yuba", "Sutter","Yolo","Sierra","Nevada",
"Placer","El Dorado","Alpine","Sacramento","Napa","Solano","Contra Costa","San Francisco",
"Alameda","San Mateo","Marin","Alaleda","San Benito","Monterey")

CentralCal <- c("Santa Cruz","Santa Clara","San Joaquin","Stanislaus",
"Amador","Calaveras","Tuolumne","Mono","Inyo","Mariposa","Merced","Madera","Fresno","Kings","Tulare")

SolCal <- c("San Luis Obispo","Kern","Santa Barbara","Ventura","Los Angeles",
"San Bernardino","Riverside","Orange","San Diego","Imperial")

## Putting vectors into data frame.

CountyName <- c(NorCal, CentralCal, SolCal)
Region <- c(rep("NorCal",33), rep("CentralCal", 15), rep("SolCal",10))
regiondf <- data.frame(CountyName,Region)

## Join with table 1

joinfinal <- merge(regiondf,join4,all=FALSE)
write.csv(joinfinal,file="joinfinal.csv")
```

Plot (page 4-11)

Plot 1-2

dot-linear plot:

```
### Data Cleaning: apply formula and use gather function

table1 <- read.csv("/Users/renjunzhu/Downloads/R Study/POP+UN+PRI+API.CSV")
A<-mean(table1$API)
sdA<-sd(table1$API)
B<-mean(table1$Population)
sdB<-sd(table1$Population)
C<-mean(table1$Unemployment_Rate)
sdC<-sd(table1$Unemployment_Rate)
table1<-table1%>%mutate(API_Z=(API-A)/sdA, Pop_Z=(Population-B)/sdB,Une_Z=(Unemployment_Rate-C)/sdC)
```

Plot1: Determinants Vs. Home Price (page 4)

```
new2<-table1 %>% select(Year, API_Z, Pop_Z, Une_Z, Price)
New2<-new2 %>% gather(key=determinant, value=value, API_Z, Pop_Z, Une_Z)
ggplot(New2, aes(x=value,y=Price))+geom_point(aes(color=determinant),size=0.9)+geom_smooth(col="black")+facet_gri
d(determinant~.)+labs(title="Determinates Vs. Home Price")+scale_y_continuous(labels = dollar)+scale_x_continuous
(limits = c(-3,2.9))+theme_igray()+ scale_colour_tableau()+theme(plot.title=element_text(size = 20, color = "Red"
),axis.title= element_text(size = 15),legend.position =c(0.1,0.9),panel.grid = element_blank())
```

Plot2: Determinants Vs. Home Price facet by Year (page 6)

```
new2<-table1 %>% select(Year, API_Z, Pop_Z, Une_Z, Price)
New2<-new2 %>% gather(key=determinant, value=value, API_Z, Pop_Z, Une_Z)
%>%ggplot(aes(x=value,y=Price))+geom_point(aes(color=determinant),size=0.9)+geom_smooth(col="black")+facet_grid(d
eterminant~Year)+labs(title="Determinates Vs. Home Price")+scale_y_continuous(labels = dollar)+scale_x_continuous
(limits = c(-3,2.9))+theme_igray()+ scale_colour_tableau()+theme(plot.title=element_text(size = 20, color = "Red"
),axis.title= element_text(size = 15),legend.position =c(0.1,0.9),panel.grid = element_blank())

## For Analysis:

### Calculate the correlation between Population and Home Price

table1 <- read.csv("/Users/renjunzhu/Downloads/R Study/POP+UN+PRI+API.CSV")
lm.out<-lm(formula = Price ~ Population,data =table1)
summary(lm.out)
cor.test(table1$Population,table1$Price)

### Calculate the correlation between Unemployment Rate and Home Price
lm.out<-lm(formula = Price ~ Unemployment_Rate,data =table1)
summary(lm.out)
cor.test(table1$Unemployment_Rate,table1$Price)

## Calculate the correlation between API and Home Price
lm.out<-lm(formula = Price ~ API,data =table1)
summary(lm.out)
cor.test(table1$API,table1$Price)
```

Plot 3-4: Region Graph (page 7-8)

Plot3: violin graph (page 7)

```
table1 <- read.csv("/Users/renjunzhu/Downloads/R Study/joinfinal.csv")%>%
group_by(Region)
p <- table1 %>%
ggplot(aes(x=Region,y=Price))+labs(title="Home Price Region Comparison")+scale_y_continuous(labels = dollar,limit
s = c(0,2500000))+theme_igray() +scale_colour_tableau()+geom_violin(aes(color=Region,fill=Region),alpha=0.6)+face
t_wrap(~Year)+theme(axis.title= element_text(size = 15),plot.title=element_text(size = 20, color = "Red"))
p
```

Plot 4: density graph (page 8)

```
table1 <- read.csv("/Users/yangyangli/Desktop/joinfinal.csv")%>%
group_by(Region)
p <- table1 %>%
ggplot(aes(x=Price))+geom_density(aes(fill=Region),alpha=0.4)+labs(title="Density Plot")+scale_x_continuous(label
s = dollar,limits = c(0,2500000))+theme_bw()+theme(plot.title=element_text(size = 20, color = "Red"),axis.title=
element_text(size = 15),legend.position =c(0.9,0.8), panel.grid = element_blank())
p
```

Plot 5: California Map (page 9)

```
### Download bycounty.csv file from zillow.com

county = read.csv("bycounty.csv")
countytidy = county %>%
filter(State == "CA") %>%
select(RegionName,X2015.12) %>%
na.omit()
loc = read.csv("cityloc.csv")
locca = loc %>%
filter(state == "CA") %>%
na.omit()

#locca = locca[!duplicated(locca["county"]),]

loctidy = locca %>%
select(county, latitude, longitude)
colnames(loctidy) = c("RegionName", "lat", "long")
location = merge(countytidy, loctidy, all=TRUE) %>%
na.omit()

### getting the region of California

library(geosphere)
geo.dist = function(df) {
  require(geosphere)
  d = function(i,z){         # z[1:2] contain long, lat
    dist = rep(0,nrow(z))
    dist[i:nrow(z)] = distHaversine(z[i:nrow(z),1:2],z[i,1:2])
    return(dist)
  }
  dm = do.call(cbind,lapply(1:nrow(df),d,df))
  return(as.dist(dm))
}
loc2 = read.csv("cityloc.csv")
set.seed(420)
CA     = loc2[loc2$state=="CA",]
CA     = CA[sample(1:nrow(CA),500),] %>% na.omit()
long=CA$longitude
lat=CA$latitude
city=CA$city
df     = data.frame(long, lat, city)
d      = earth.dist(df)   # distance matrix
hc     = hclust(d)      # hierarchical clustering
df$clust = cutree(hc,k=4)

### Install many packages and apply functions inside of these packages.

install.packages("rgdal")
install.packages("maptools")
install.packages("rgeos")
install.packages("C:\\Users\\Tom\\Documents\\gpclib.tar.gz", repos = NULL, type="source")
install.packages("ez", dependencies = TRUE)
library(rgeos)
library(maptools)
library(rgdal)
library(ez)
gpclibPermit()
map.US  = readOGR(dsn=".", layer="tl_2013_us_state")
map.CA  = map.US[map.US$NAME=="California",]
map.df  = fortify(map.CA)
ggplot(map.df)+
  geom_path(aes(x=long, y=lat, group=group))+
  geom_point(data=location, aes(x=long, y=lat, color=X2015.12), size=3, alpha=0.4)+ scale_colour_gradient2( "Pric
e", high=muted("green"))+
  coord_fixed()+
  theme_bw()+
  labs(title="California Housing Price Distribution")+
  theme(plot.title=element_text(size=40), axis.title=element_text(size=20),
axis.title.x=element_text(size=30),
axis.title.y=element_text(size=30),
legend.text=element_text(size=20),
legend.title=element_text(size=30))+
  annotate("text", y = 37.7739, x = -122.4313, label = "San Francisco", color="red", size=7)+
  annotate("text", y = 34.0522, x = -118.2437, label = "Los Angeles", color="red", size=7)+
  annotate("text", y = 32.7157, x = -117.1611, label = "San Diego", color="red", size=7)+
  annotate("text", y = 35.3744, x = -119.0187, label = "Bakersfield", color = "red", size = 7)+
  annotate("text", y = 37.3230, x = -122.0322, label = "Cupertino", color = "red", size = 7)
```

Plot 6: Price Fluctuation Plot (page 11)

```
### Making graph ready table and cleaning

library(lubridate)
library(ggplot2)
library(scales)
library(ggthemes)
table <- read.csv("/Users/renjunzhu/Downloads/R Study/City_Zhvi_AllHomes.csv")

btable <- table %>% filter (State=="CA")%>%
select(RegionName, X1996.04:X2014.12) %>%
filter( RegionName %in% c("San Francisco","Los Angeles","Cupertino","Compton","Bakersfield")) %>%
gather (key=RegionName, value=Year) %>% na.omit()
colnames(btable) = c("Region", "Year", "Price")
btable$Year = gsub("X","",btable$Year)
btable$Year = paste(btable$Year,"01",sep=".")

### Making fluctuation graph over time

bbtable <- btable %>% mutate(year=lubridate::ymd(Year))
ggplot(bbtable,aes(x= year, y = Price,col=(Region)))+scale_y_continuous(labels = dollar)+labs(title="1996-2015 Ho
using Price Trend")+theme_bw()+theme(plot.title=element_text(size = 20, color = "Red"),axis.title= element_text(s
ize = 15),legend.position =c(0.1,0.8), panel.grid = element_blank())+theme_economist()+geom_line(linetype=1)
```

Animation (present during RRR week through google earth)

```
## Downloads data from zillow.com, and filter out the cases of California over 1 million home price from 1996 to
 2016.

cityloc = read.csv("cityloc.csv", as.is = FALSE)
cityprice = read.csv("cityprice.csv", as.is = FALSE)
citypriceca = cityprice %>%
  filter(State == "CA")
citylocca = cityloc %>%
  filter(state == "CA")
citylocca = citylocca[!duplicated(citylocca["city"]), ] %>%
na.omit()
colnames(citylocca) = c("zip", "lat", "long", "RegionName", "State", "County")
Loc = merge(citylocca,citypriceca,all=FALSE) %>%
na.omit()
Locti = Loc %>%
gather(Time, Price, 11:250)
Loctidy = Locti %>%
select(RegionName, lat, long, Time, Price) %>%
filter(Price >= 1000000)

Latitude = Loctidy$lat
Longitude = Loctidy$long
subtimes = function(x){
  x1 = gsub("X", "", x)
  x2 = gsub("\\.", "-", x1)
  return (x2)
}

Loctidy$Time = subtimes(Loctidy$Time)
Time = Loctidy$Time
```

Making a kml file within R.

```
doc = newXMLDoc()
root = newXMLNode("kml", namespaceDefinitions = "http://www.opengis.net/kml/2.2", doc=doc)
documents = newXMLNode("Document", parent=root)
name = newXMLNode("name", "time", parent=documents)
description = newXMLNode("description", "Cities with average housing price bigger than 1 Million, 1996-present",
parent=documents)
for (i in 1:length(Latitude)){
  lat = Latitude[i]
  log = Longitude[i]
  time = Time[i]
  loc = paste(log, lat, "0", sep=",")
  placemark = newXMLNode("Placemark", parent=documents)
  timestamp = newXMLNode("TimeStamp", parent=placemark)
  newXMLNode("when", time, parent=timestamp)
  point = newXMLNode("Point", parent=placemark)
  newXMLNode("coordinates", loc, parent=point)
}
saveXML(doc, "/Users/Tom/Desktop/housing.kml")
```