Data Visualization: Final Project

James Dale

8/18/2020

Prep house dataset

```r
keycol  <- "years"
valuecol <- "med_sales_price"
gathercols <- c()

raw_names <- names(house_data)
for (i in 1:length(raw_names)){
  if (raw_names[i] %in% c("RegionID","RegionName", "State", "Metro",
                 "CountyName", "SizeRank")){
  } else {
    gathercols[length(gathercols)+1] <- raw_names[i]
  }
}

house_data_reshape <- gather_(house_data, key_col = keycol,
                              value_col = valuecol,
                              gather_cols = gathercols)

house_data_reshape <- house_data_reshape %>%
  separate(years, sep = "-", into=c("year", "month"))
```

Prep GDP data

```r
yearly_gdp_data <- yearly_gdp_data %>%
  select(-c("GDP in billions of current dollars")) %>%
  rename("Yearly GDP in billions" = "GDP in billions of chained 2009 dollars")

quarterly_gdp_data <- quarterly_gdp_data %>%
  separate("Year", sep = "q", into = c("year", "quarter")) %>%
  select(-c("GDP in billions of current dollars")) %>%
  spread(key = "quarter", value="GDP in billions of chained 2009 dollars") %>%
  rename("Q1 GDP" = "1", "Q2 GDP" = "2", "Q3 GDP" = "3", "Q4 GDP" = "4")
```

Prep univeristy dataset

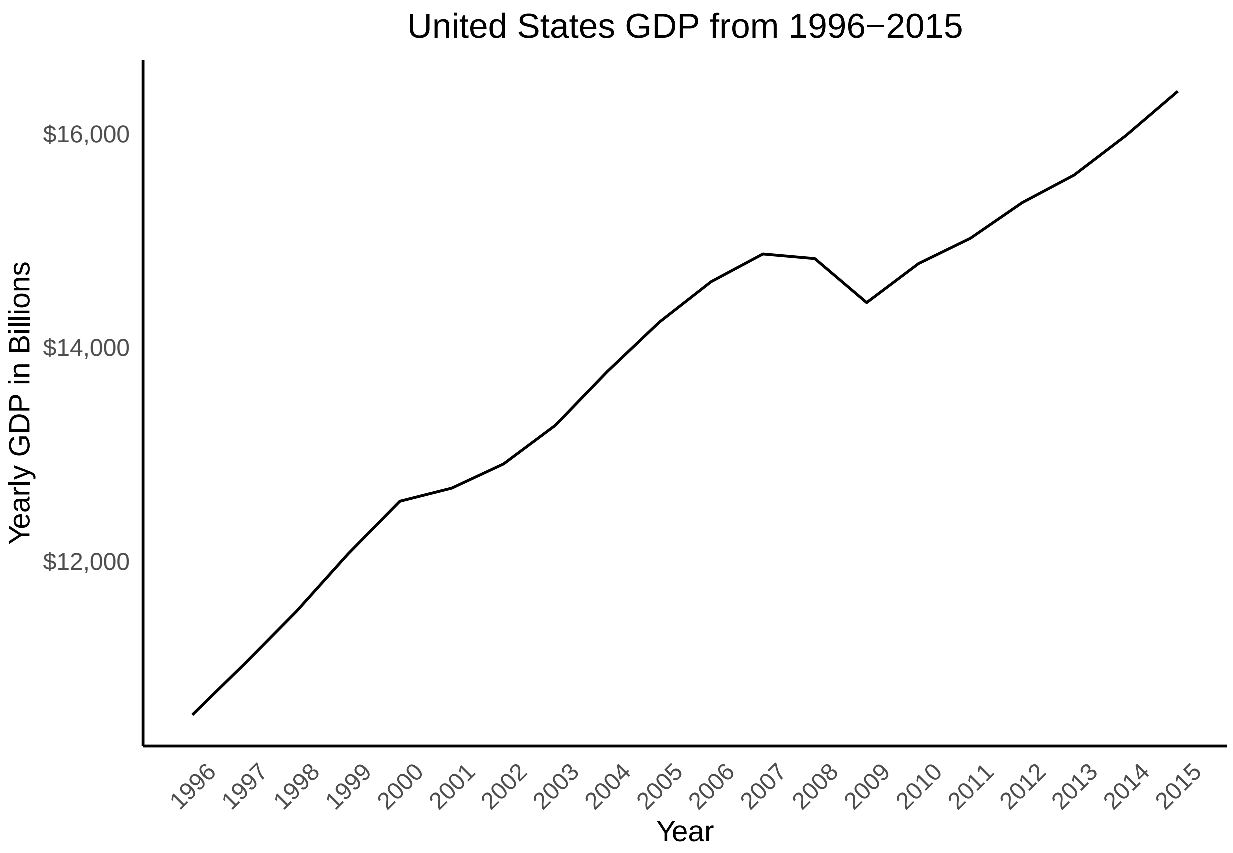
```
university_clean <- university_raw %>%
  mutate(is_state = if_else(str_detect(raw_txt, "\\[edit\\]"), TRUE,FALSE),
         state = if_else(is_state, str_remove(raw_txt, "\\[edit\\]"), NA_character_),
         college_and_city = if_else(is_state, NA_character_, str_remove_all(raw_txt, "\\[[:digit:]+\\]"
         city = str_remove_all(college_and_city, "\\s\\(.+\\)"),
         college = str_remove_all(str_extract(college_and_city, "\\(.+\\)"), "\\(|\\)"))  %>%
  tidyr::fill(state, .direction = ’down’) %>%
  filter(!is.na(college)) %>%
  select(state, city, college)
```

Prep Geographic data

```
shapes <- st_crop(shapefile,  xmin = -200, xmax = 0,
                  ymin = 20, ymax = 150)
```

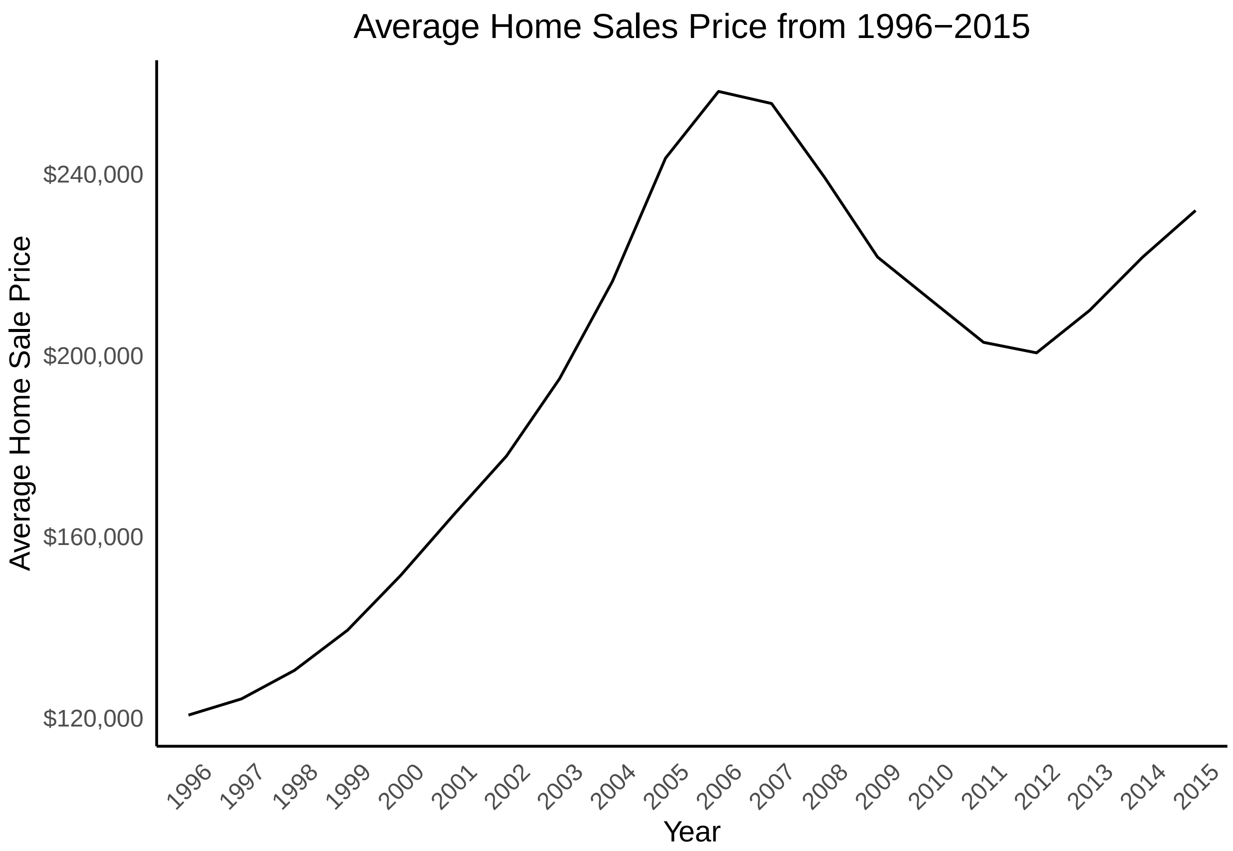
Plot 1

```
plot1_data <- yearly_gdp_data %>%
  filter(Year >= 1996) %>%
  group_by(Year) %>%
  summarise(GDP = na.omit(first(‘Yearly GDP in billions‘)))

ggplot(plot1_data, aes(x=Year, y=GDP, group=1)) +
  geom_line() +
  xlab("Year") +
  ylab("Yearly GDP in Billions") +
  scale_y_continuous(labels = dollar) +
  scale_x_continuous(breaks = pretty(plot1_data$Year, n = 15)) +
  ggtitle("United States GDP from 1996-2015") +
  theme_minimal() +
  theme(panel.border = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(size = 0.5, colour = "black"),
        axis.text.x = element_text(angle = 45, vjust = 0.5, hjust=0.5),
        legend.position = "none",
        plot.title = element_text(hjust = 0.5))
```



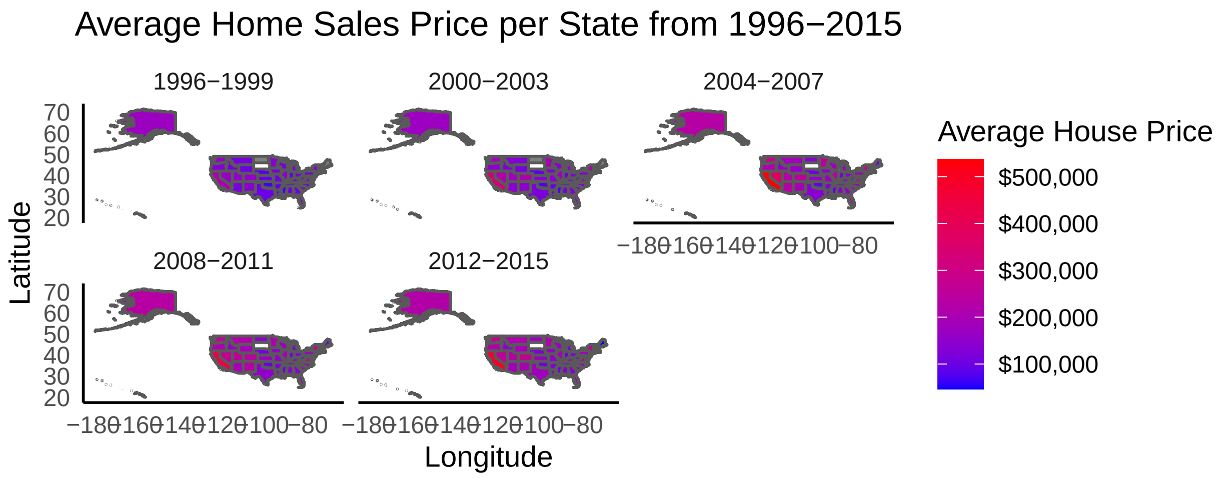
Plot 2

```
plot2_data <- house_data_reshape %>%
  filter(year <= 2015) %>%
  group_by(year) %>%
  summarise(average_price = mean(na.omit(med_sales_price)))

ggplot(plot2_data, aes(x=year, y=average_price, group=1)) +
  geom_line() +
  xlab("Year") +
  ylab("Average Home Sale Price") +
  scale_y_continuous(labels = dollar) +
  ggtitle("Average Home Sales Price from 1996-2015") +
  theme_minimal() +
  theme(panel.border = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(size = 0.5, colour = "black"),
        axis.text.x = element_text(angle = 45, vjust = 0.5, hjust=0.5),
        legend.position = "none",
        plot.title = element_text(hjust = 0.5))
```



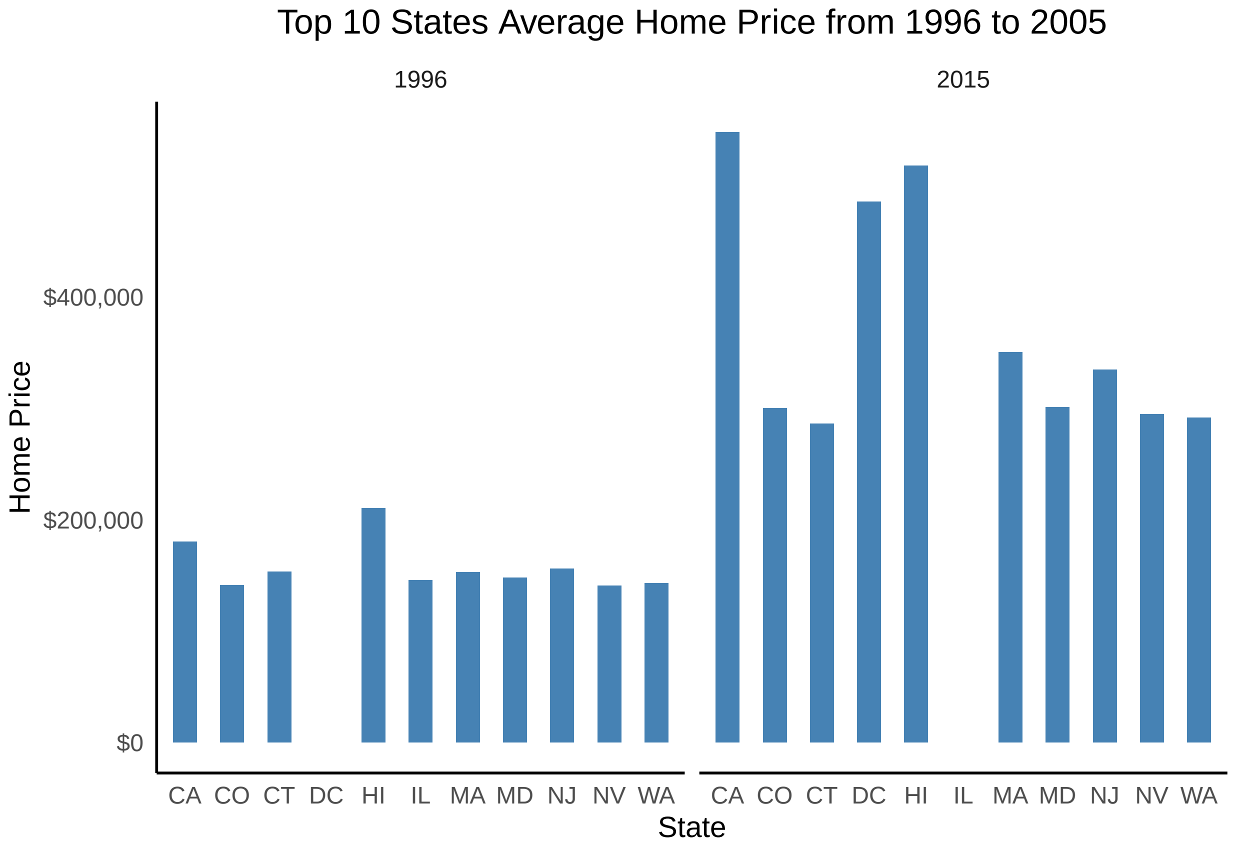
Plot 3

```
plot2_data <- house_data_reshape %>%
  mutate(year_bin = cut(as.numeric(year),
                        breaks = c(0, 1999, 2003, 2007, 2011, 2016),
                        labels = c("1996-1999", "2000-2003", "2004-2007",
                                   "2008-2011", "2012-2015"))) %>%
  group_by(year_bin, State) %>%
  summarise(mean_sales = mean(na.omit(med_sales_price))) %>%
  inner_join(shapes, by=c("State" = "STUSPS"))


ggplot(plot2_data) +
  geom_sf(data=plot2_data, aes(fill=mean_sales, geometry=geometry)) +
  scale_fill_gradient(low = "blue", high = "red",
                      name = "Average House Price",
                      labels = dollar) +
  ggtitle("Average Home Sales Price per State from 1996-2015") +
  xlab("Longitude") +
  ylab("Latitude") +
  facet_wrap(~ year_bin) +
  theme_minimal() +
  theme(panel.border = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(size = 0.5, colour = "black"),
        plot.title = element_text(hjust = 0.5))
```



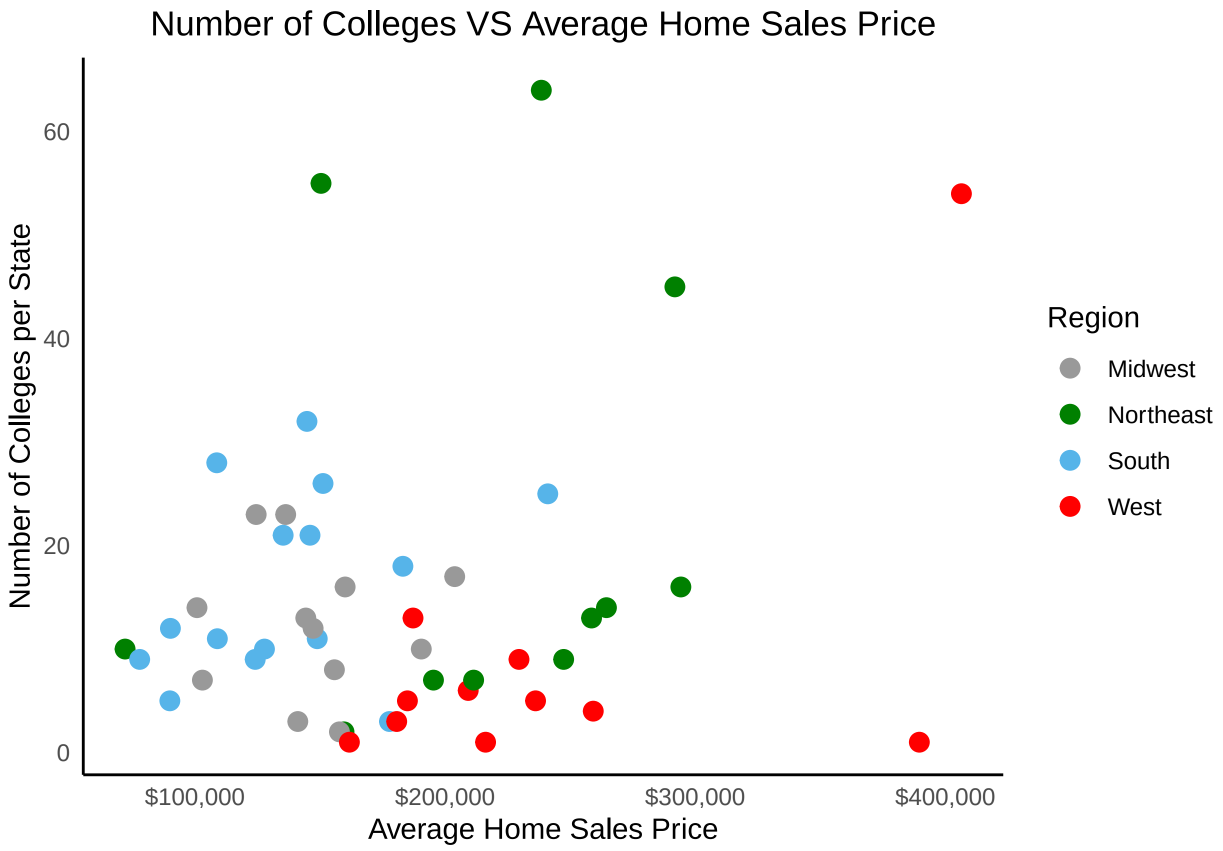
Plot 4

```
house_data_state_groups <- house_data_reshape %>%
  group_by(year, State) %>%
  summarise(med_sales_price = mean(na.omit(med_sales_price)))

top10_states_2015 <- house_data_state_groups %>%
  filter(year == "2015") %>%
  arrange(-med_sales_price) %>%
  distinct(State, .keep_all = TRUE) %>%
  slice(1:10) %>%
  select(c(State, med_sales_price, year))

top10_states_1996 <- house_data_state_groups %>%
  filter(year == "1996") %>%
  arrange(-med_sales_price) %>%
  distinct(State, .keep_all = TRUE) %>%
  slice(1:10) %>%
  select(c(State, med_sales_price, year))

plot4_data <- bind_rows(top10_states_1996, top10_states_2015)

ggplot(data=plot4_data, aes(x=State, y=med_sales_price)) +
  geom_bar(stat = "identity", width = 0.5, fill="steelblue") +
  facet_wrap(~ year) +
  ggtitle("Top 10 States Average Home Price from 1996 to 2005") +
  scale_y_continuous(labels = dollar) +
  xlab("State") +
  ylab("Home Price") +
  theme_minimal() +
  theme(panel.border = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(size = 0.5, colour = "black"),
        plot.title = element_text(hjust = 0.5))
```



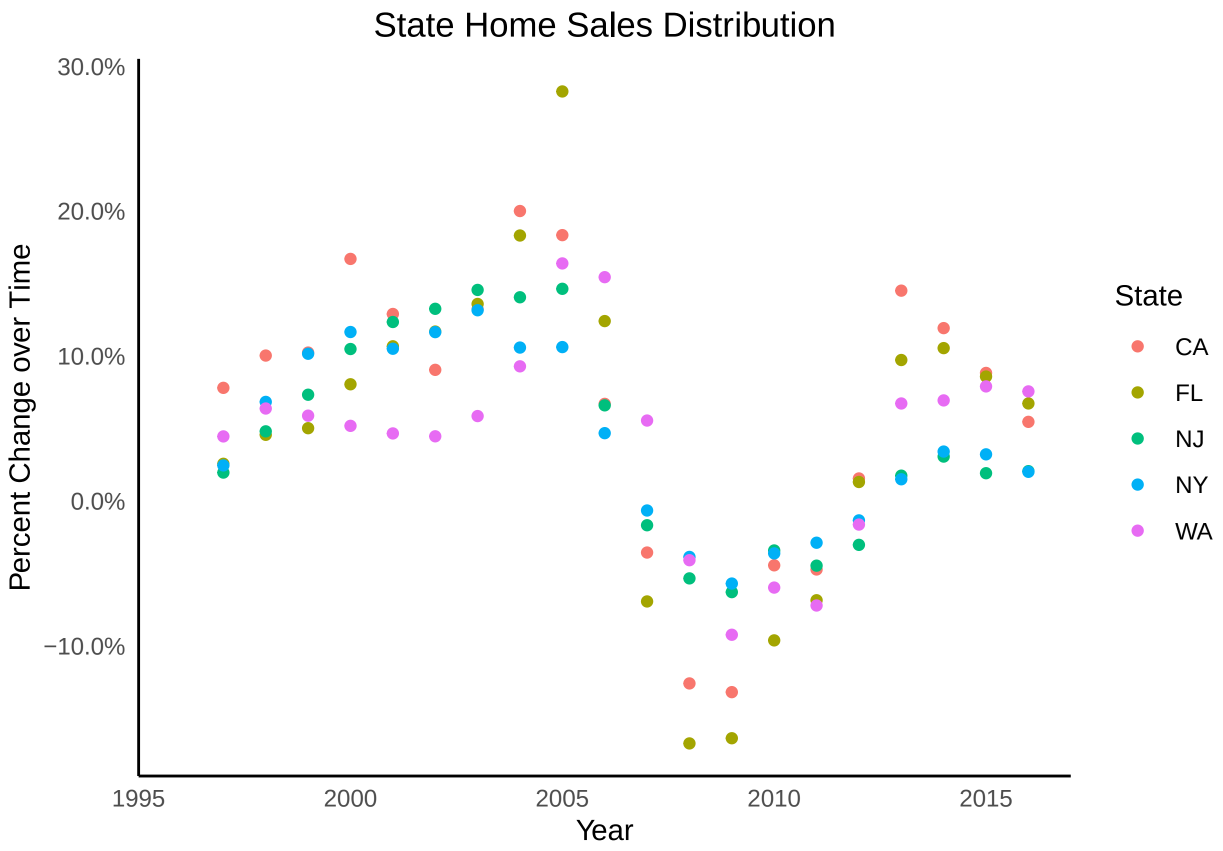
Plot 5

```
plot5_data <- university_clean %>%
  separate_rows(college, sep=",") %>%
  group_by(state) %>%
  summarise(n_colleges = n()) %>%
  inner_join(state_data, by = c("state" = "State")) %>%
  right_join(house_data_reshape, by = c("ABV" = "State")) %>%
  drop_na(med_sales_price, state) %>%
  group_by(state) %>%
  summarise(avg_sales_price = mean(med_sales_price),
            n_colleges = mean(n_colleges),
            region = first(Region))

ggplot(plot5_data, aes(x=avg_sales_price, y=n_colleges)) +
  geom_point(aes(color = region), size=3) +
  scale_color_manual(values=c("#999999", "#008000", "#56B4E9", "#FF0000")) +
  scale_x_continuous(label = dollar) +
  scale_y_continuous() +
  ggtitle("Number of Colleges VS Average Home Sales Price") +
  xlab("Average Home Sales Price") +
  ylab("Number of Colleges per State") +
  labs(color = "Region") +
  theme_minimal() +
  theme(panel.border = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(size = 0.5, colour = "black"),
        plot.title = element_text(hjust = 0.5))
```



Plot 6

```r
plot6_data <- house_data_reshape %>%
  drop_na(med_sales_price) %>%
  filter(State %in% c("FL","CA", "NY", "NJ", "WA")) %>%
  group_by(State, year) %>%
  summarise(med_sales_price = mean(med_sales_price)) %>%
  mutate(pct_change = (med_sales_price/lag(med_sales_price, k = 1) - 1))

ggplot(plot6_data, aes(x=as.numeric(year), y=pct_change, group=1)) +
  geom_point(aes(color = State)) +
  scale_x_continuous() +
  ggtitle("State Home Sales Distribution") +
  scale_y_continuous(labels = percent) +
  xlab("Year") +
  ylab("Percent Change over Time") +
  theme_minimal() +
  theme(panel.border = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(size = 0.5, colour = "black"),
        plot.title = element_text(hjust = 0.5))
```



Executive Summary

This report will summarize some general insights derived from several data sets from Kaggle.com titled Zillow All Homes Data. The data mainly consisted of the median sales price for a home in a specific region at a point in time. The data given allowed me to analyze the median sales price for the United States from 1996 to 2015. The data set also contained United States GDP value from 1995 to 2015 as well as each city in the United States that had a college. With this data and a spatial file from the United States census website I was able to derive several interesting graphics.

The first graphic I chose was a simple line graph showing the United States GDP from 1996 to 2015. I chose this because GDP is a descent representation of the economy and generally the better the economy is doing the higher home prices will be. We can see from the graph the GDP was rising at a steady rate right up to about 2006, which was the just about the start of the subprime mortgage crisis. The crisis latest approximately from 2006 to 2010 which is reflective of the graph.

For the second plot I also chose to use a line graph to represent the average home sales price from 1996 – 2015. We can see after steady growth there is an apex and decline in 2006 and lasting until about 2011. With is we can definitely see there was a disturbance in GDP and median home sales across the United State from 2006 – 2011.

The next graph I made was a choropleth map depicting average house price from 1996 – 2015 split into groups of 4 years. The first thing we can see is the increase in average home price over time. There is also a clear distinction the average home price in coastal states has been and remains higher than states in the middle of the country.

Next, I wanted to answer the following questions: what are the top 10 states with the highest average home sales price, and did they change from 1996 to 2015? To do this a made a bar graph showing the average sales price of each state in 1996 and 2015. We can see from the graph all the states stayed the same from 1996 to 2015 expect for Illinois which changed to Washington DC. It is slightly deceiving because instead of the actual value showing for Washington DC in 1996 and for Illinois in 2015 is slows as 0 but allows you to more easily visualize the states missing in each year. We can also see the average home price at least doubled in each of the common states.

For the 5th plot I decided to look at the effect the number of colleges a state has and the region the state is in on the average sales price from 1996 to 2015. I used a scatter plot to visualize this. Based on the graph we can see states in the Midwest and South have lower average home sales prices than that of states in the Northeast and West. However, it is not clear that the number of colleges a state has correlates to the average home sales price. It is also interesting that the only true outliers in terms of both number of colleges and average home sales price are in the Northeast and Western regions.

For my last plot I wanted to explore the percent change in average sales price over time between the top 5 states with highest average home sales prices. I used a scatter plot with percent change on the y-axis and year on the x-axis with each state being represented by color. We can see a common trend among all 5 states. There was a steady increase in price until about 2006 with a steep decline.

Overall, there are several findings to highlight in this report. The average home sales across the United States increased steadily until 2006 signifying the start of the subprime mortgage crisis and since 2011 or 2012 the average home sales price has recovered. Costal states are the most expensive states to live in and have experienced the highest growth in sales price. Illinois was once the most expensive places to live on average in 1996 and was overtaken by Washington DC in 2015. California, Florida, New Jersey, New York, and Washington are among the most expensive places to live on average based on data from 1996-2015 and show similar trends in price changes.

Dashboard Link

Click here to view the dashboard on Tableau Public