Data Visualization Assignment 04

Jesse Conlon

2/16/2020

Data Sets

```
houses <- read_csv("data/KING COUNTY House Data.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   date = col_datetime(format = "")
## )
```

```
## See spec(...) for full column specifications.
```

```
countries <- st_as_sf(maps::map('world', plot = FALSE, fill = TRUE))

states <- st_as_sf(map("state", plot = FALSE, fill = TRUE))

counties <- st_as_sf(map("county", plot = FALSE, fill = TRUE))

counties_wa <-counties %>%
  filter(str_detect(ID, 'washington,'))

king_county_zips <- st_read('data/Zipcodes_for_King_County_and_Surrounding_Area_Shorelines__zipcode_sho
```

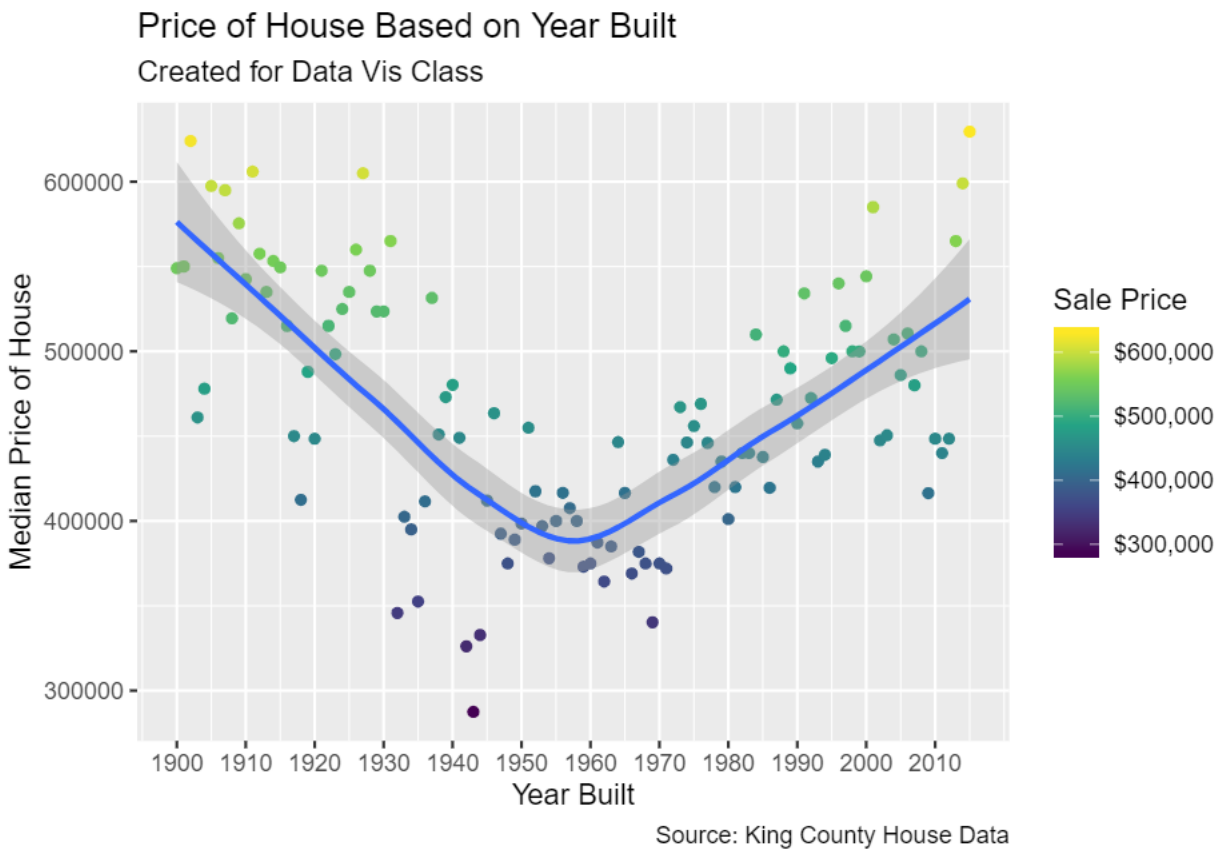
Question 1: What is happening to price over time (yr_built)?

For this one I didn't just want to make a plot of all the points, it was too much to look at initially and just not very interesting. I opted to manipulate my data set a bit to find the median price of houses per year. I plotted these values, added a bit of color, and then added a trend line to more clearly show that house prices took a dip for a time frame centered around 1960. I think this is a rather effective way of showing what is happening to house price over time. I chose to use the median value because there are instances of extremely high and low values that can really skew the data which is common with real estate (such as the Los Angeles house Jeff Bezos recently bought for $165 million!).

```
house_avg <- houses %>%
  group_by(yr_built) %>%
  summarize(price = median(price))

ggplot(data = house_avg, aes(x = yr_built, y = price, color = price)) +
  scale_colour_viridis_c("Sale Price", labels = dollar) +
  labs(x = "Year Built",
       y = "Median Price of House",
```

```
      title = "Price of House Based on Year Built",
      subtitle = "Created for Data Vis Class",
      caption = "Source: King County House Data") +
      scale_x_continuous(breaks = seq(1900,2015, by = 10)) +
      scale_y_continuous(labels = function(x) format(x, scientific = FALSE), ) +
      geom_point() +
      geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

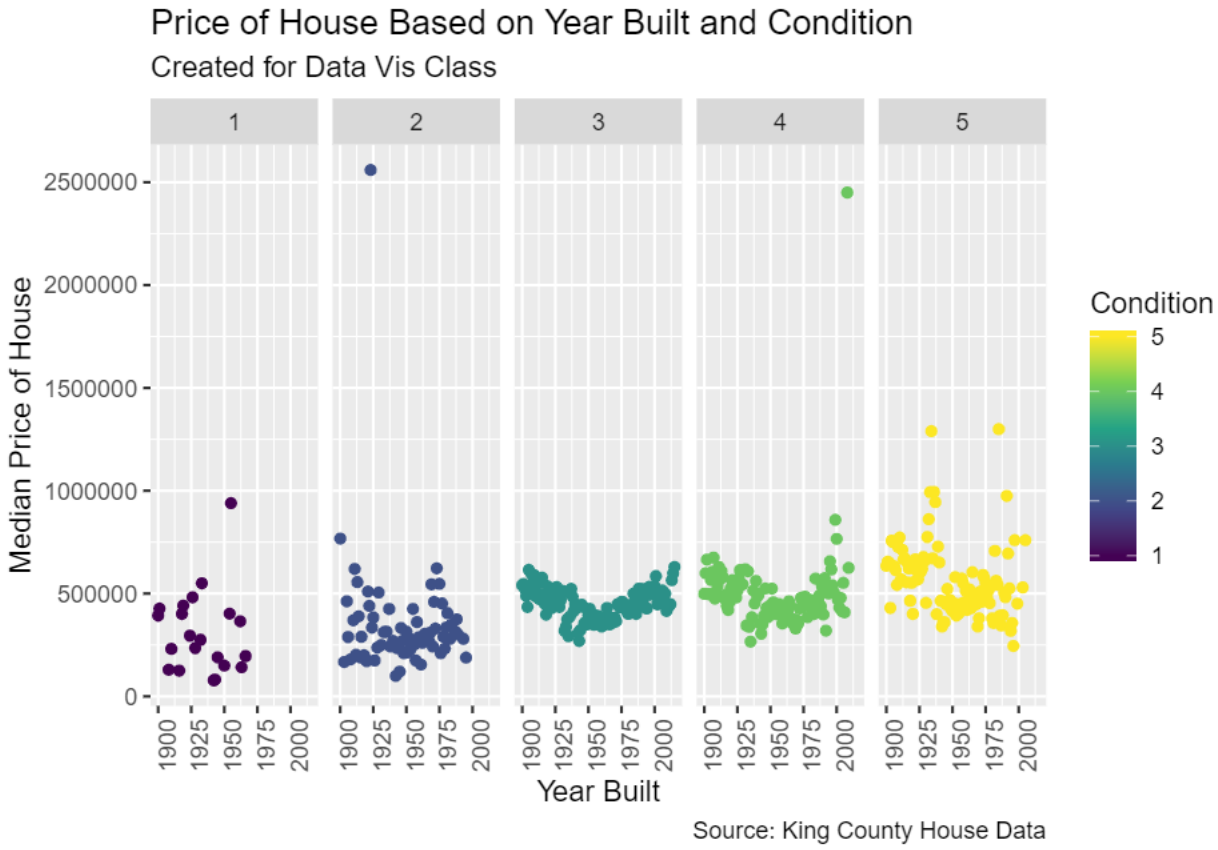


As an extension of the above question, I also broke down my data set to show the average value of the houses per year based on their condition. What I think is intriguing about this is that there is a clear difference in the average prices of the houses based on condition. There are a few outliers but for the most part the better the condition of the house the higher their value regardless of when they were built. Also, it is interesting to note that the occurence of houses rated at a condition of "1" are not even present after roughly 1975 and the same can be said for condition "2" houses after 2000. I feel like this could be valuable information when considering the upkeep and maintenance of a house over the course of time. I chose a facet grid to display this information as it shows how there is a progressive increase in the value of homes depending on their condition.

Extra Credit Price Based on Condition

```
house_avg_cond <- houses %>%
  group_by(yr_built, condition) %>%
  summarize(price = median(price))
```

```
ggplot(data = house_avg_cond, aes(x = yr_built, y = price, color = condition)) +
  scale_colour_viridis_c("Condition") +
  labs(x = "Year Built",
       y = "Median Price of House",
       title = "Price of House Based on Year Built and Condition",
       subtitle = "Created for Data Vis Class",
       caption = "Source: King County House Data") +
       scale_x_continuous(breaks = seq(1900,2015, by = 25)) +
       scale_y_continuous(labels = function(x) format(x, scientific = FALSE), ) +
       geom_point() +
       facet_grid(~condition) +
       theme(axis.text.x = element_text(angle=90))
```

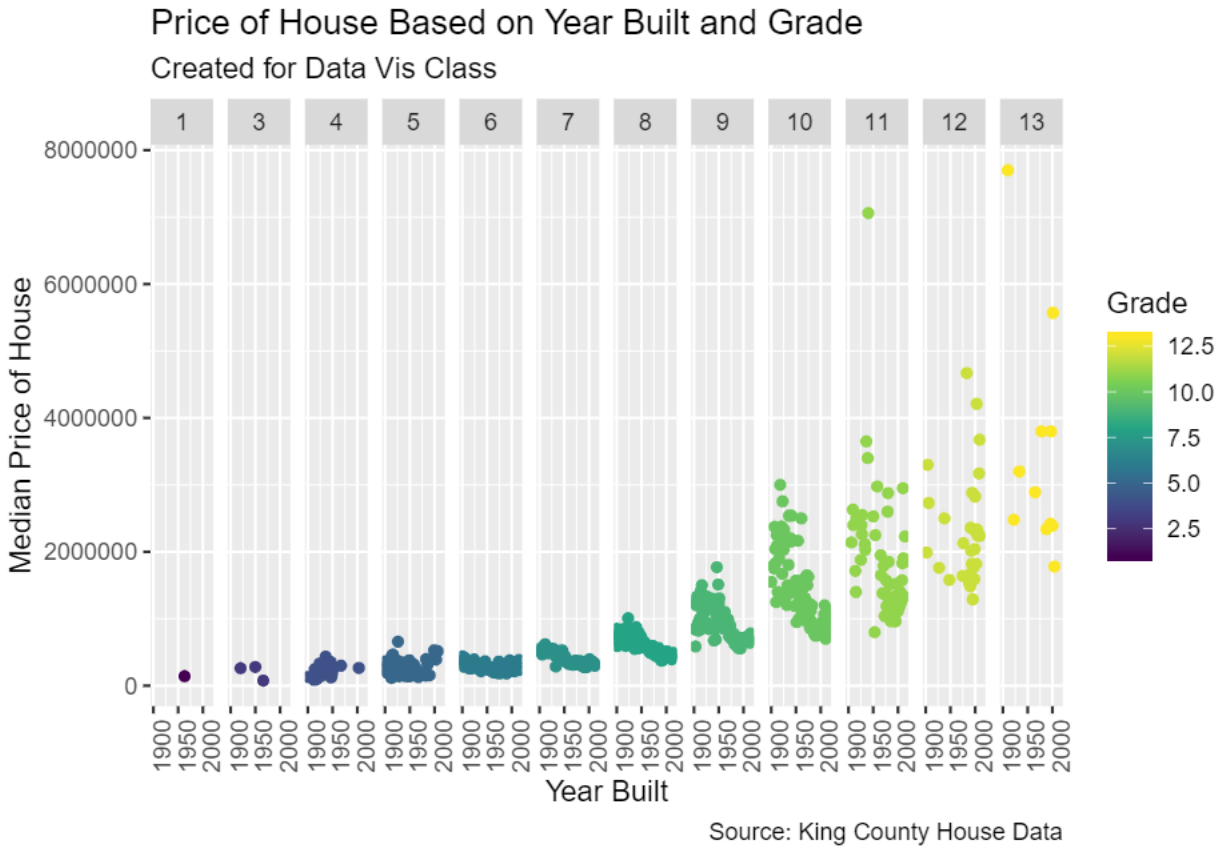


Again, another extension showing how grade also affects price, very similar to condition

Extra Credit Price Based on Grade

```
house_avg_grade <- houses %>%
  group_by(yr_built, grade) %>%
  summarize(price = median(price))

ggplot(data = house_avg_grade, aes(x = yr_built, y = price, color = grade)) +
  scale_colour_viridis_c("Grade") +
  labs(x = "Year Built",
```

```
      y = "Median Price of House",
      title = "Price of House Based on Year Built and Grade",
      subtitle = "Created for Data Vis Class",
      caption = "Source: King County House Data") +
      scale_x_continuous(breaks = seq(1900,2015, by = 50)) +
      scale_y_continuous(labels = function(x) format(x, scientific = FALSE), ) +
      geom_point() +
      facet_grid(~grade) +
      theme(axis.text.x = element_text(angle=90))
```



Question 2: What is happening to price over geographic space (Can be lat / long, zipcode, etc

For this question, I grouped by the zipcode, sumarized and took the average lat/long of the zipcodes to have a centralized location of all houses in the zipcode. I used both color and shape to emphasize the prices of the houses. What is a noticeable trend is that the further south, the lower the cost.

```
geo_price <- houses %>%
  group_by(zipcode) %>%
  summarize(price = median(price), lat = mean(lat), long = mean(long))

theme_set(theme_minimal())
counties_wa <-counties %>%
  filter(str_detect(ID, 'washington,'))
```

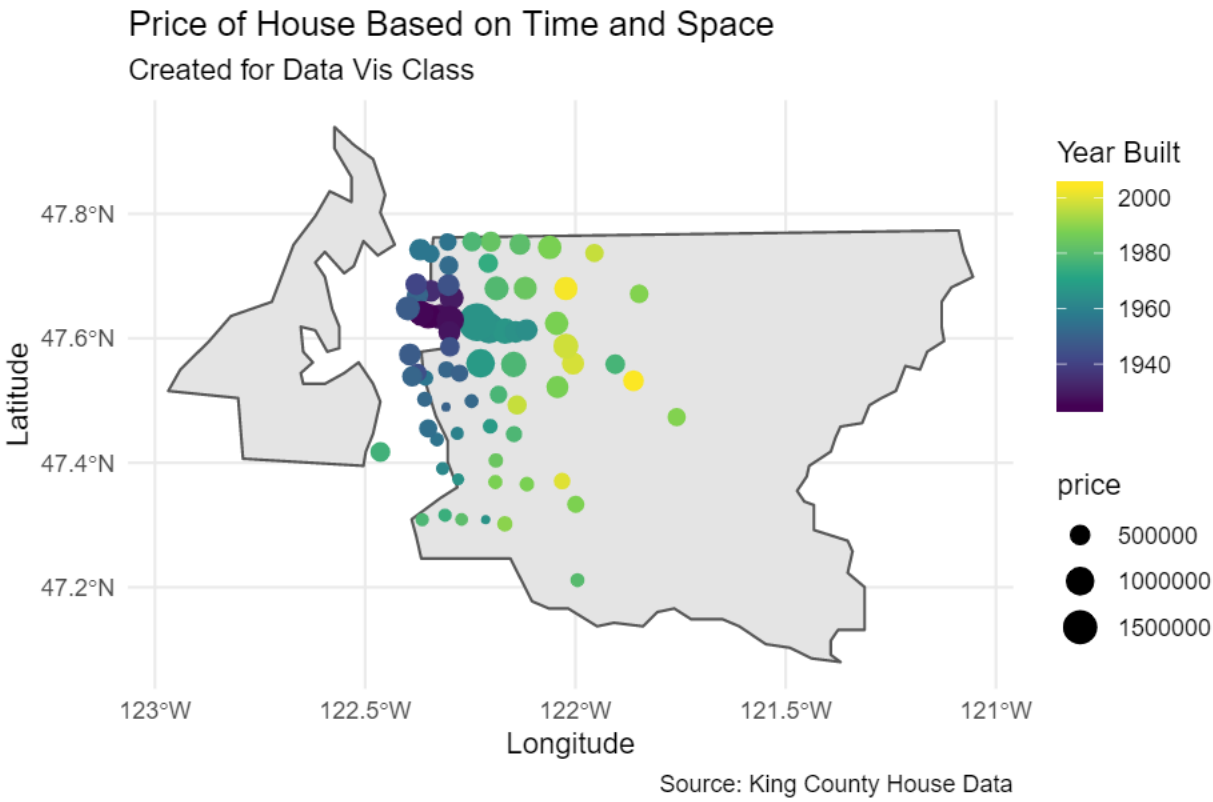
```
counties_wa %>%
  filter(str_detect(ID, "king|kitsap")) %>%
  ggplot() +
  geom_sf() +
  geom_point(data = geo_price, aes(x = long, y = lat, color = price, size = price)) +
  scale_colour_viridis_c("Sale Price", labels = dollar) +
  scale_size(name = "Sale Price", labels = dollar) +
  labs(x = "Longitude",
       y = "Latitude",
      title = "Price of House Based on Zipcode",
      subtitle = "Created for Data Vis Class",
      caption = "Source: King County House Data")
```

[](assignment_04_files/figure-latex/Question 2: What is happening to price over geographic space (Can be lat / long, zipcode, etc-1.pdf)

This following graphic I believe says alot about the area. For instance, properties close to the shore line tend to be older but not necessarily more expensive. The further away from the shore, the newer the properties appear to be. This suggests that the earlier generations built closer to the water and also that there might now be limited space to build new property. It also shows the areas where property value tends to be higher. Again, I opted for median values for price and year built. I believe this helps eliminate the extreme values to capture an accurate picture of an area. I think about my neighborhood, for example, and how most of the houses represent an accurate picture of where I live. There is a house a block away though that is falling apart, another that is immaculate and probably worth twice the value of the rest of the houses in the area. In my opinion they cancel out but their values might skew the representation of my neighborhood. I believe this is the true nature of real estate prices and why I feel using the median paints an accurate picture.

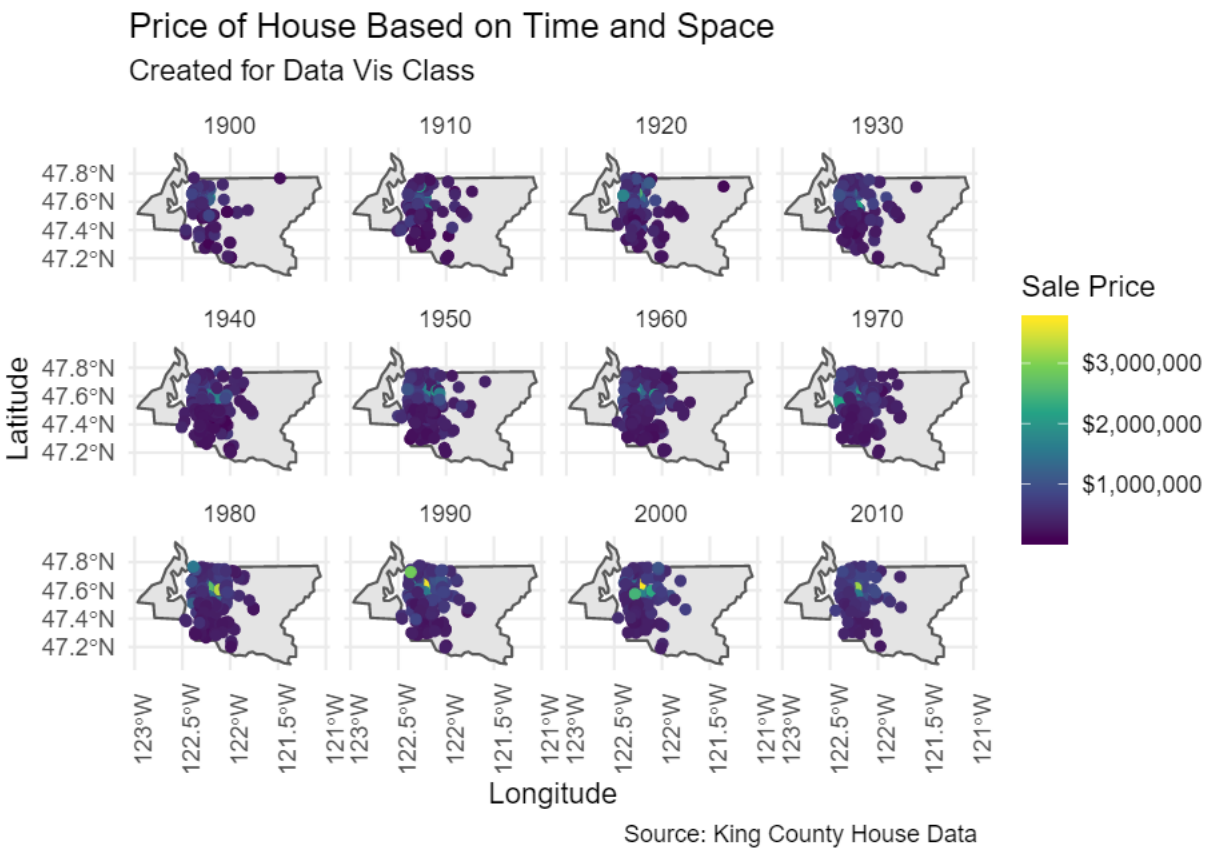
Question 3: What is happening to price over time and space?

```
time_space <- houses %>%
  group_by(zipcode) %>%
  summarize(price = median(price), yr_built = median(yr_built), lat = mean(lat), long = mean(long))

counties_wa %>%
  filter(str_detect(ID, "king|kitsap")) %>%
  ggplot() +
  geom_sf() +
  geom_point(data = time_space, aes(x = long, y = lat, color = yr_built, size = price)) +
  scale_colour_viridis_c("Year Built") +
  labs(x = "Longitude",
       y = "Latitude",
      title = "Price of House Based on Time and Space",
      subtitle = "Created for Data Vis Class",
      caption = "Source: King County House Data")
```



Here is another way of looking at the data which breaks it down into decades. If feel there is a lot of potential viewing the data this way and across other different variables as well. For instance, if we wanted to track the price, condition, grade, etc., of houses across the decades this could show the degradation of houses. I could also add in a count of houses to show how the area has increased in population and the number of houses.

```
time_space2 <- houses %>%
  group_by(zipcode, decade, condition) %>%
  summarize(price = median(price), lat = mean(lat), long = mean(long))

counties_wa %>%
  filter(str_detect(ID, "king|kitsap")) %>%
  ggplot() +
  geom_sf() +
  geom_point(data = time_space2, aes(x = long, y = lat, color = price)) +
  scale_colour_viridis_c("Sale Price", labels = dollar) +
  facet_wrap(~decade) +
  labs(x = "Longitude",
       y = "Latitude",
      title = "Price of House Based on Time and Space",
      subtitle = "Created for Data Vis Class",
      caption = "Source: King County House Data") +
  theme(axis.text.x = element_text(angle=90))
```

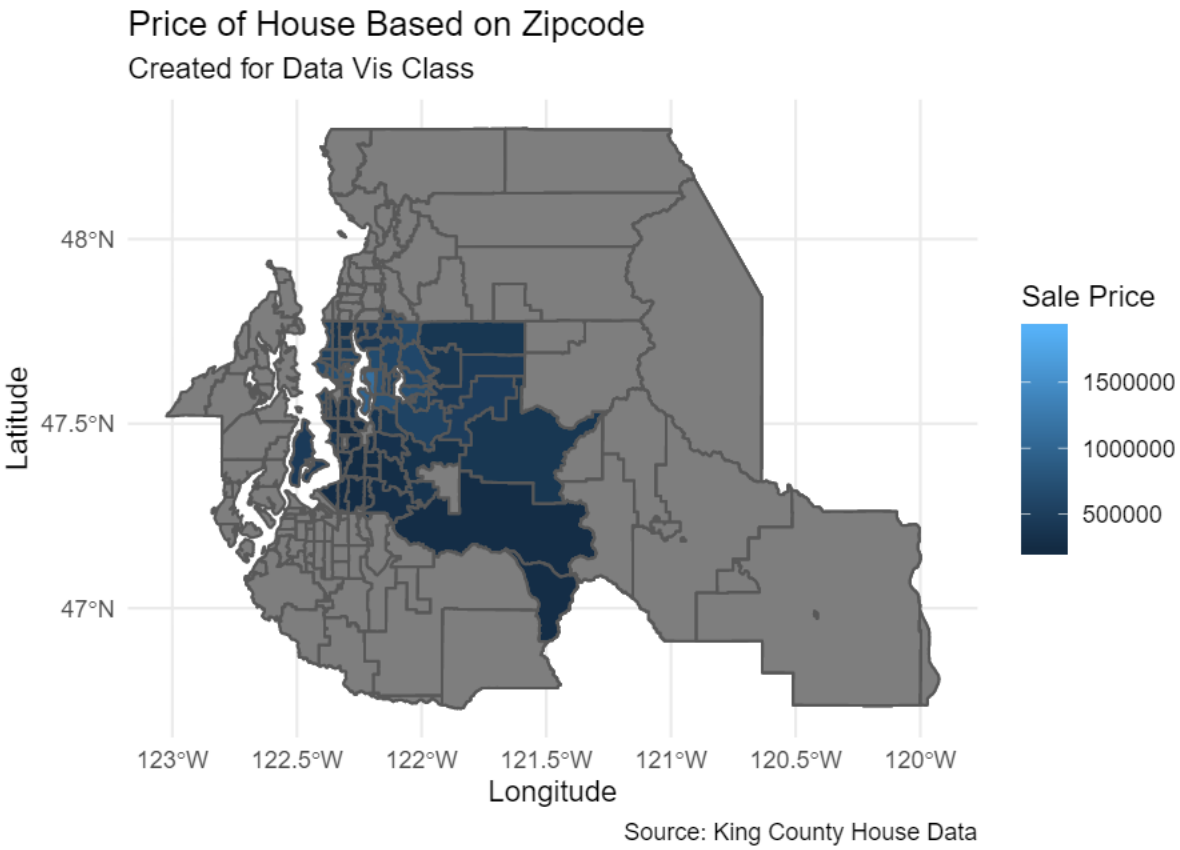


This one was a bit tricky and required some digging and research. To start, the shape file had to be downloaded and uploaded (in the data sets at the top of the assignment), then I had to join data sets so that I could generate a proper ggplot that allowed me to fill in the certain areas on the map. I was struggling with a few thing like eliminating some of the unnecessary zipcode areas as well as properly labeling. Nonetheless, it is another take on mapping prices to zipcodes in addition to the above examples.

```
zipcode_map <- king_county_zips %>% st_join(counties_wa, by = c("geometry" = "geom" ))
```

```
## although coordinates are longitude/latitude, st_intersects assumes that they are planar
```

```
zipcode_avgs <- zipcode_map %>%
  left_join(geo_price, by = c("ZIP" = "zipcode"))

ggplot(zipcode_avgs, aes(fill = price)) + geom_sf() +
    scale_colour_viridis_c("Sale Price", labels = dollar) +
    labs(x = "Longitude",
        y = "Latitude",
        title = "Price of House Based on Zipcode",
        subtitle = "Created for Data Vis Class",
        caption = "Source: King County House Data",
        fill = "Sale Price")
```


Price of House Based on Zipcode
Created for Data Vis Class
48°N
47.5°N
47°N
Latitude
123°W
122.5°W
122°W
121.5°W
121°W
120.5°W
120°W
Longitude
Sale Price
1500000
1000000
500000
Source: King County House Data