# Assignment 04

Amol Gote

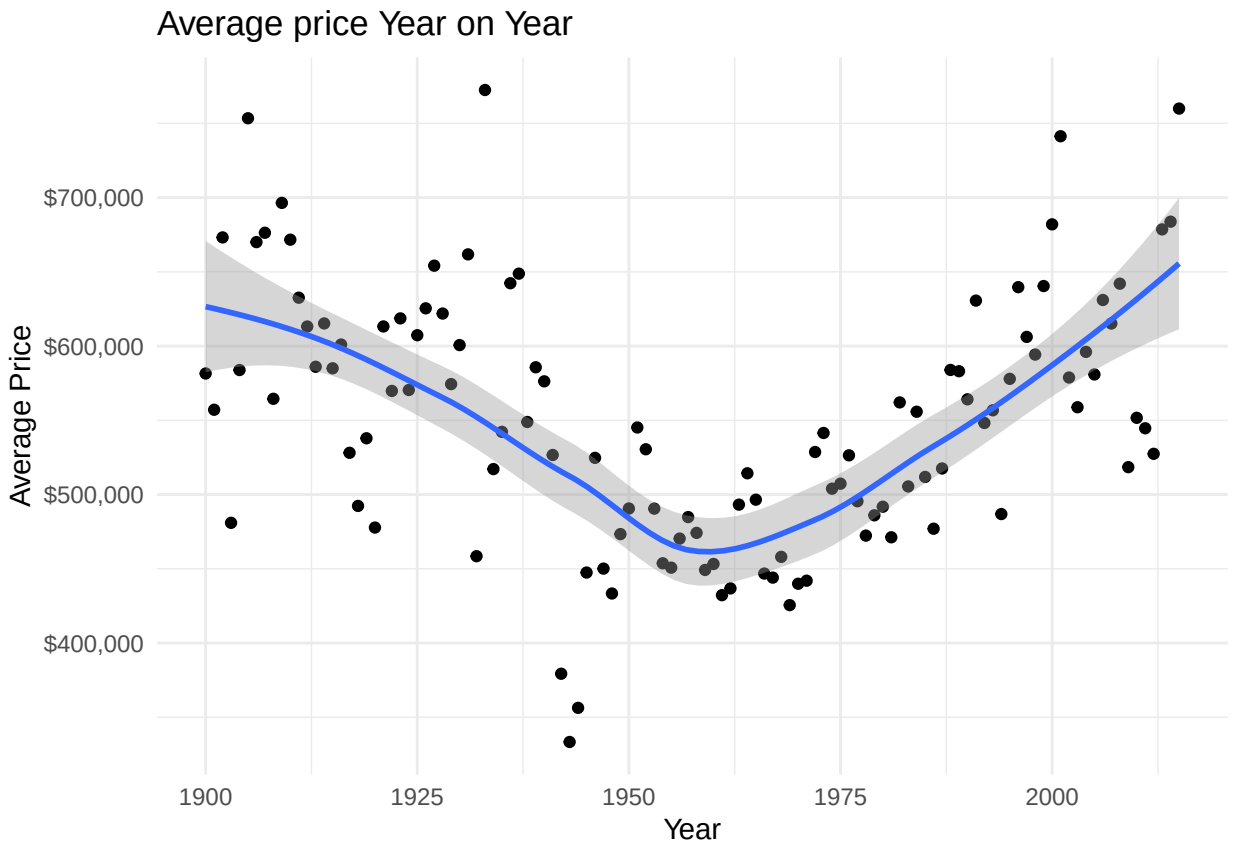
2/10/2020

## Question 1

What is happening to price over time (yr_built)

```
houses <- read_csv("data/KING COUNTY House Data.csv")
averagePriceEachYear <- houses %>%
  group_by(yr_built) %>%
  summarise(averagePrice = mean(price))

averagePriceEachYear %>%
  ggplot() +
  geom_point(aes(x = yr_built, y = averagePrice)) +
  geom_smooth(aes(yr_built, averagePrice)) +
  scale_y_continuous(labels = dollar) +
  labs(x = "Year", y = "Average Price",
  title = "Average price Year on Year") +
  theme_minimal()
```




1. For comparing price year on year, have taken mean house prices for each built year.

2. Average price had peaked in early 1900's, have been dropping then till late 1950's. Prior to 1950 there are points below the smooth line which is the lowest prices in the span of the 20th century, reason for the same could be World War 2.

3. Post 1960 it has started rising gradually till 2010.

4. There are points dropping below the smooth line after 2000, around 2009 this is due to economic depression in 2009.

5. Post 2009 average price has recovered and have hit peak, the peak numbers post 2009 are same that of peak number is early 1900's.

## Question 2

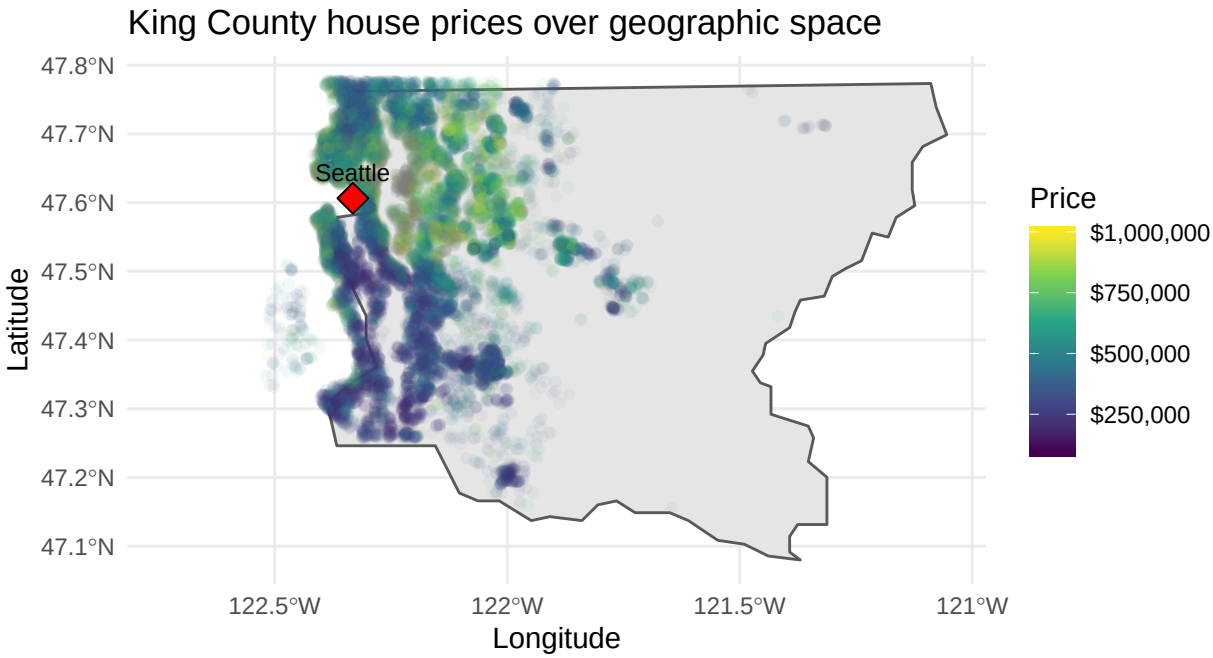
What is happening to price over geographic space (Can be lat / long, zipcode, etc)

```
counties <- st_as_sf(map("county", plot = FALSE, fill = TRUE))
counties_wa <-counties %>%
  filter(str_detect(ID, 'washington,'))

counties_wa_king <- counties_wa %>%
  filter(str_detect(ID, "king"))

sites <- data.frame(longitude = c(-122.3321), latitude = c(47.6062))
```

```
counties_wa_king %>%
  ggplot() +
  geom_sf() +
  geom_point(data = houses, aes(x = long, y = lat, color = price), alpha= .05) +
  geom_point(data = sites, aes(x = longitude, y = latitude), size = 4,
        shape = 23, fill = "red") +
  geom_text(data = sites, aes(x = longitude, y = latitude), label = 'Seattle', position =
              position_dodge(width = 0.8), size = 3, vjust = -1.0) +
  scale_colour_viridis_c("Price", limits = c(100000, 1000000), labels = dollar) +
  theme_minimal() +
  labs(x = "Longitude",
  y = "Latitude",
  title = "King County house prices over geographic space")
```




Note: Have added limit 100,000 to 1,000,000, so that all outliers will have same color.

1. North West side of the King county have lot of houses which are above 500K, that's reason there are lot of points with green to yellow color.

2. Reason for the higher price in the North West side is proximity to Seattle metro area.

3. Locations around Seattle have higher prices.

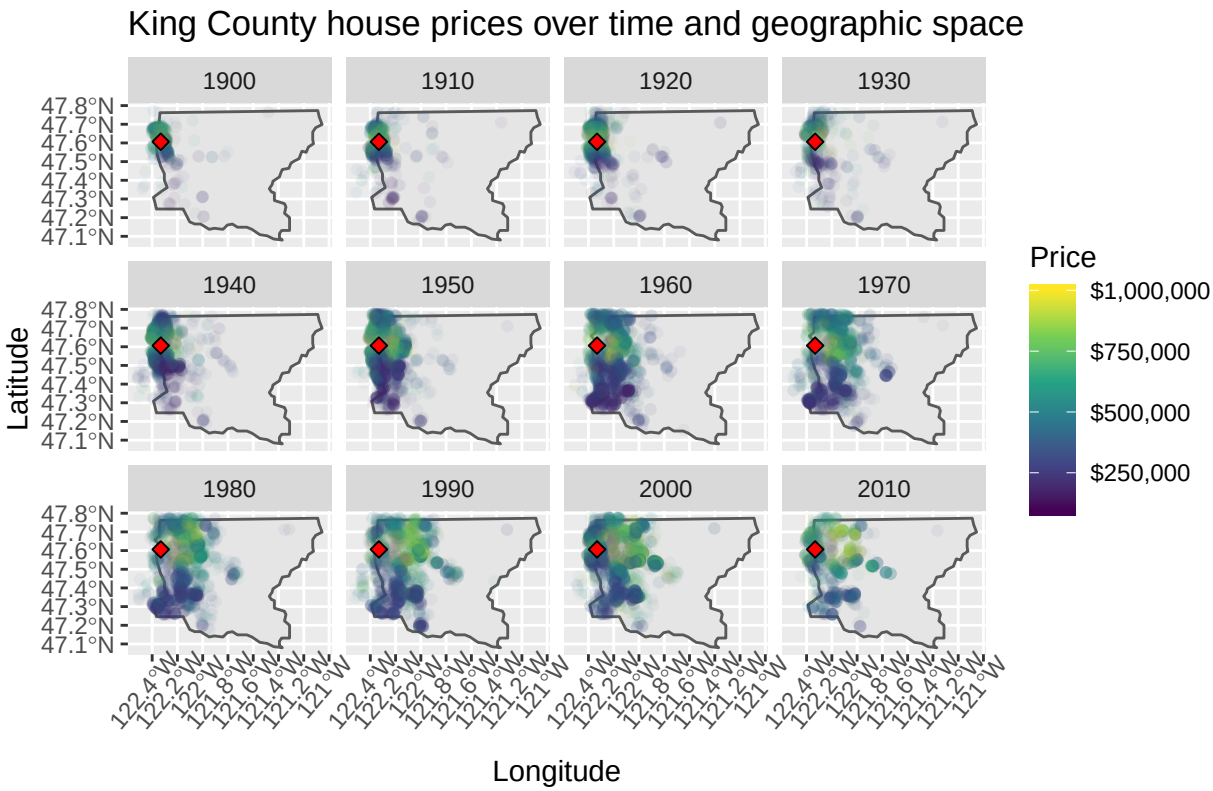
4. As we move away from Seattle towards east or towards south prices are dropping, especially in south, most of the houses are below 500k or less.

5. Outliers are getting highlighted by brown shade, there are shades of brown just beside east of Seatle (slight North East), those areas have highest average home prices and belong to richest people on earth, Jeff Bezos and Bill Gates.

## Question 3

What is happening to price over time and space?

```
counties_wa_king %>%
  ggplot() +
  geom_sf() +
  geom_point(data = houses, aes(x = long, y = lat, color = price ), alpha= .05) +
  geom_point(data = sites, aes(x = longitude, y = latitude), size = 2,
        shape = 23, fill = "red") +
  scale_colour_viridis_c("Price", limits = c(100000, 1000000), labels = dollar) +
  facet_wrap(~decade) +
  theme(axis.text.x = element_text(angle =50, hjust=0.75))+
  labs(x = "Longitude",
  y = "Latitude",
  title = "King County house prices over time and geographic space")
```




Note: Have added limit 100,000 to 1,000,000, so that all outliers will have same color.

1. Over the time of 20th Century (1901 - 2000), the density of points have grown, which indicates that more number of houses have been built.

2. Till 1950 density of houses around Seattle area is less, there are houses with price ranging from 500k to 1 Million, but number is less.

3. Post 1950 density of points has increased which indicates that more number of houses were built around Seattle metro area as well as towards south as well.

4. Post 1950 more and more number of points are getting from green to yellow which is clear indicator that pricing of the houses has increased.

5. From 1990 to 2010 there are more prominent green to yellow points which could be because technology giants like Amazon, Microsoft expanding.

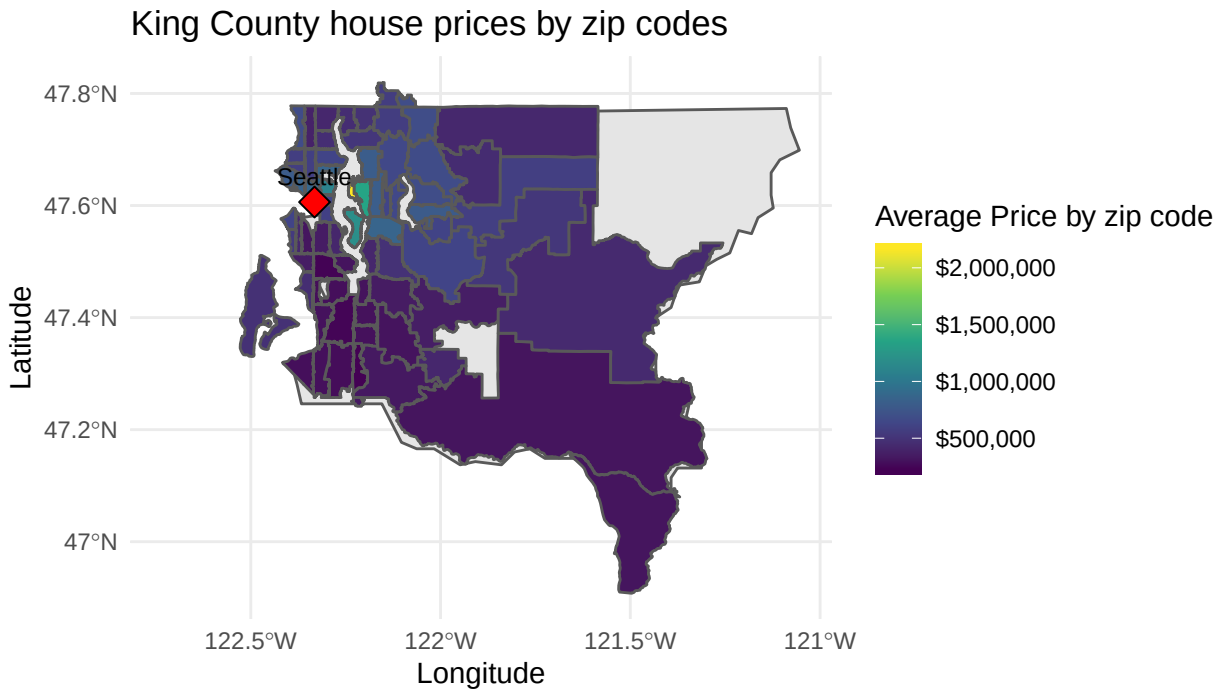
## Question 4

(Extra credit) Try to think about how you can use the zip code data for location information and map price to zipcodes

```
zipcodeShapeData <- st_read('data/Zipcodes_for_King_County/Zipcodes_for_King_County_and_Surrounding_Area

averagePriceByZip <- houses %>% group_by(zipcode) %>% summarize(averagePrice = mean(price), averageGrad
mergedShapeAndAvgZipCodeData <- merge(zipcodeShapeData,averagePriceByZip,by.x=c("ZIPCODE"),by.y=c("zipc
sites <- data.frame(longitude = c(-122.3321), latitude = c(47.6062))
```

```
mergedShapeAndAvgZipCodeData %>%
  ggplot() +
  geom_sf(data=counties_wa_king) +
  geom_sf(aes(fill=averagePrice)) +
  geom_point(data = sites, aes(x = longitude, y = latitude), size = 4,
        shape = 23, fill = "red") +
  geom_text(data = sites, aes(x = longitude, y = latitude), label = 'Seattle', position =
              position_dodge(width = 0.8), size = 3, vjust = -1.0) +
  scale_fill_viridis_c("Average Price by zip code", labels = dollar) +
  labs(x = "Longitude", y = "Latitude",
  title = "King County house prices by zip codes") +
  theme_minimal()
```




a. Mapped mean of price by zip code.

b. Around Seattle metro area, average house price hovers around 1M.

c. Interesting facts

1. Zipcode highlighted in yellow is 98039, it has an average house price of $2,161,300.0.
2. This zip code is where world's 2 richest people have their home Bill Gates and Jeff Bezos.
3. 98039 ZIP code ranks the top in Washington state in Forbes magazine's list of the most expensive ZIP codes in the country.

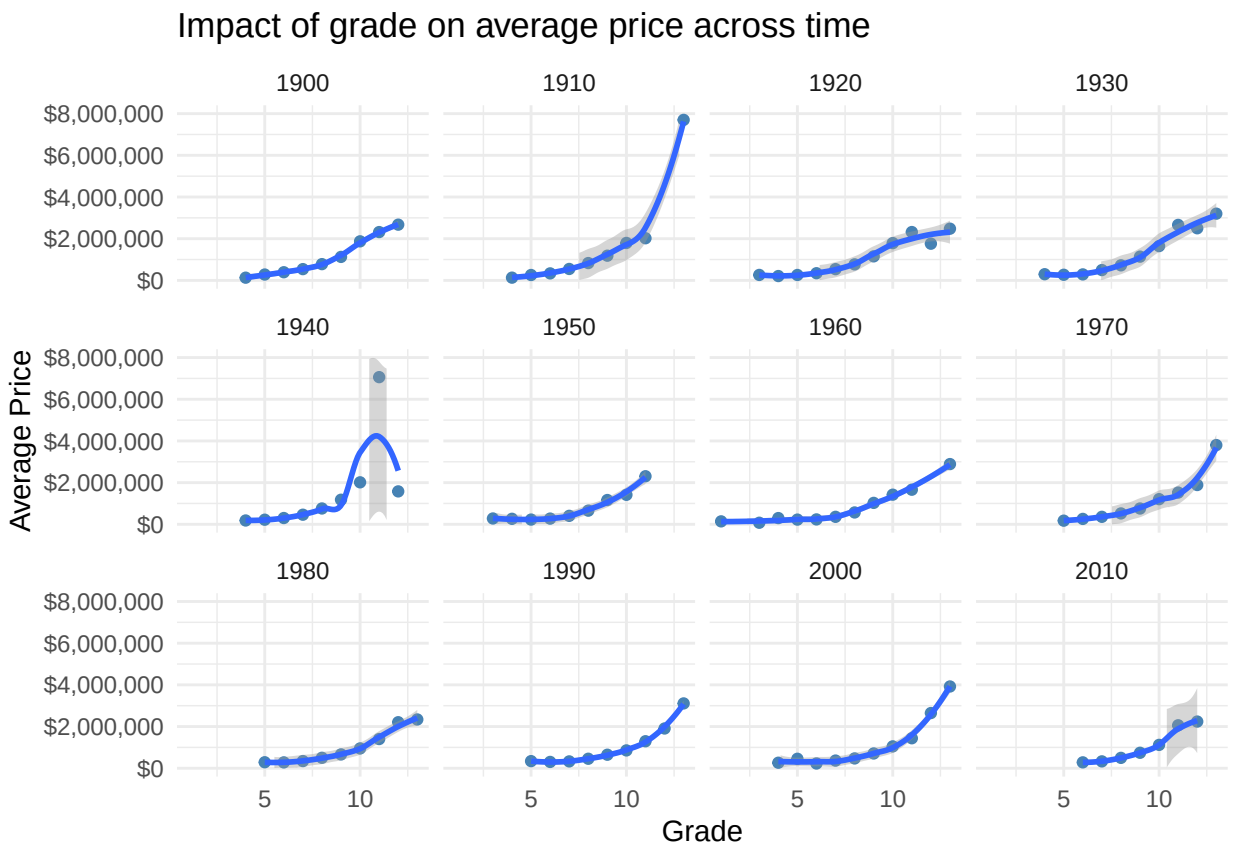
## Question 6:

Does grade impact prices across time?

Grade - An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high quality level of construction and design.

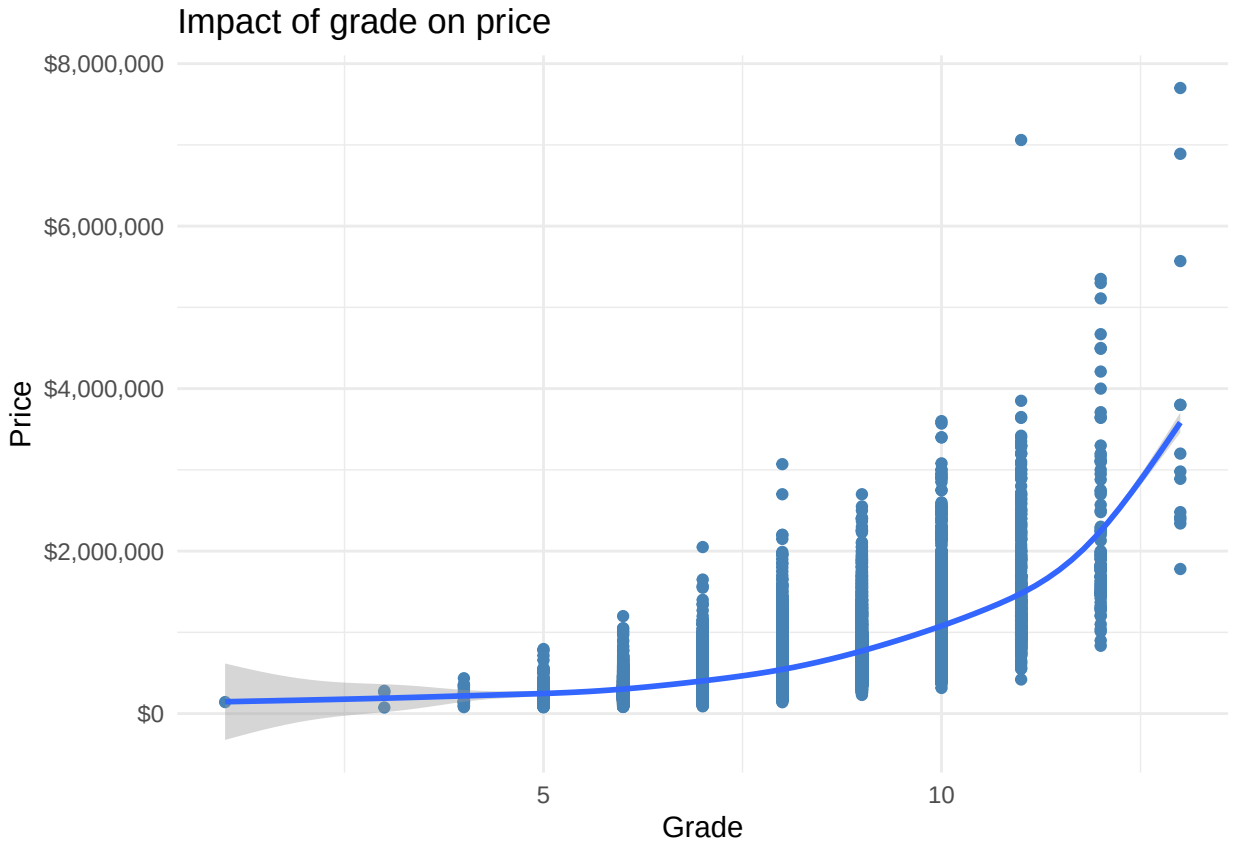
```
averagePriceDecadeGrade <- houses %>%
  group_by(decade, grade) %>%
  summarise(averagePrice = mean(price))

ggplot(averagePriceDecadeGrade, aes(x = grade, y = averagePrice)) +
    geom_point(color='steelblue') +
    geom_smooth(aes(grade, averagePrice)) +
    scale_y_continuous(labels = dollar, limits = c(0, 8000000)) +
    facet_wrap(~decade) +
    labs(x = "Grade", y = "Average Price",
    title = "Impact of grade on average price across time") +
    theme_minimal()
```



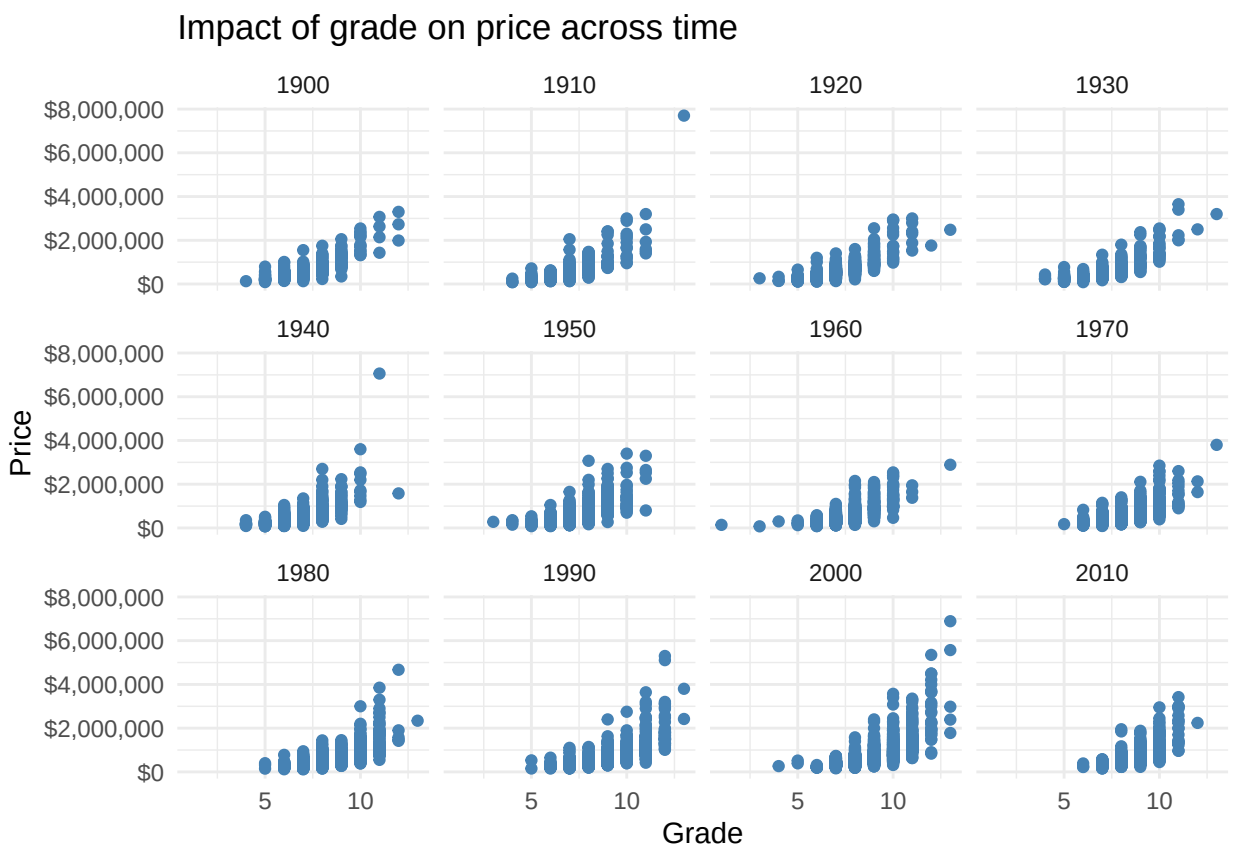

a. Based on the above visualization it is pretty evident that **grade does impact pricing** and overall general trend across time remains the **same for grade 1 to 10**.
b. **Above grade 10** over the period of time there has been **steep rise in the prices**, especially decade of **1990 and 2000**. Also houses built post 1970 were grade 5 and above.
c. In general if we consider all the data points instead of average it is evident that increase in grade does increase the price. For e.g. below visualization shows overall trend of price and grade.

```
ggplot(houses, aes(x = grade, y = price)) +
  geom_point(color='steelblue') +
  scale_y_continuous(labels = dollar) +
  geom_smooth(aes(grade, price)) +
  labs(x = "Grade", y = "Price",
  title = "Impact of grade on price") +
  theme_minimal()
```




d. Similar trend can be observed as time has passed by for 20th century

```
ggplot(houses, aes(x = grade, y = price)) +
  geom_point(color='steelblue') +
  scale_y_continuous(labels = dollar) +
  labs(x = "Grade", y = "Price",
  title = "Impact of grade on price across time") +
  facet_wrap(~decade) +
  theme_minimal()
```



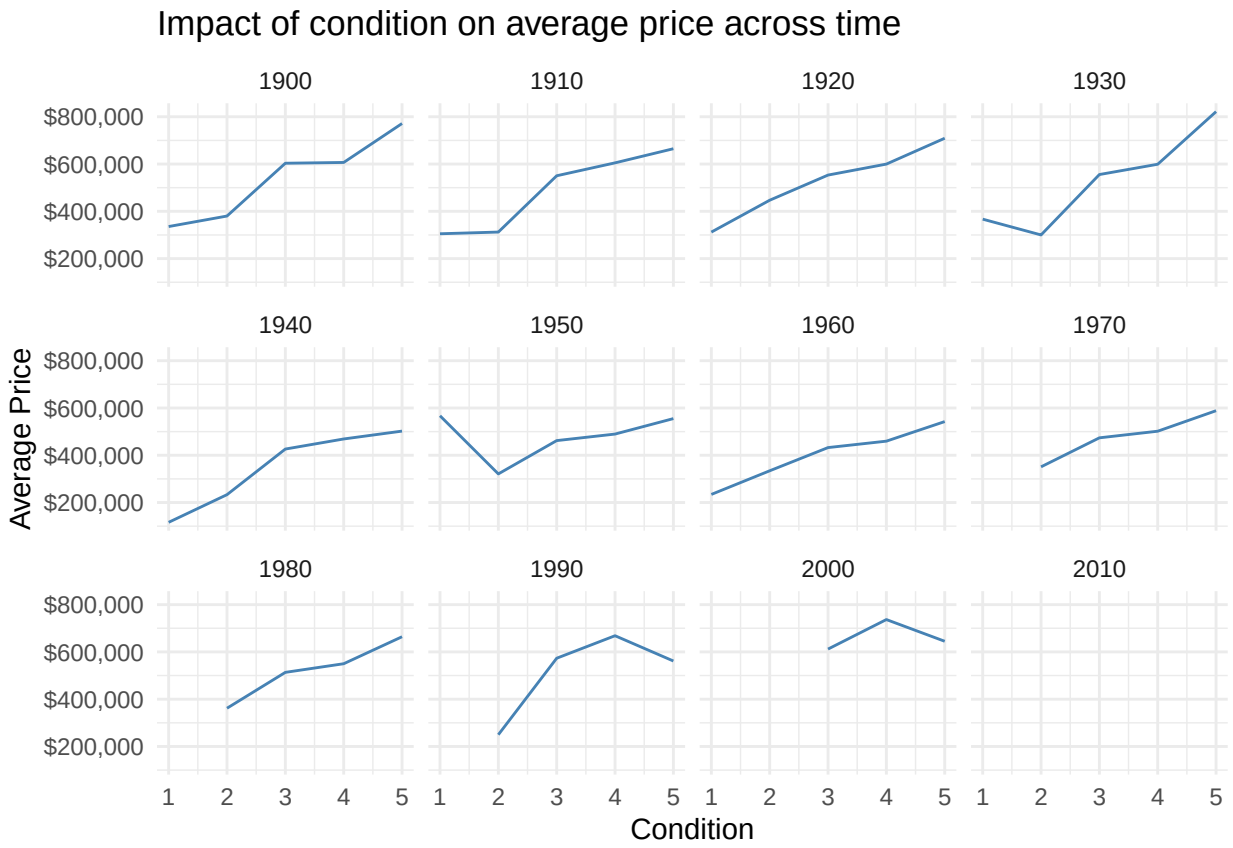

## Question 5:

Does condition impact prices across time?

Condition - An index 1 to 5 on the condition of the apartment.

```
averagePriceDecadeCondition <- houses %>%
  group_by(decade, condition) %>%
  summarise(averagePrice = mean(price))

ggplot(averagePriceDecadeCondition, aes(x = condition, y = averagePrice)) +
    geom_line(color='steelblue') +
    #geom_smooth(aes(condition, averagePrice)) +
    scale_y_continuous(labels = dollar) +
    facet_wrap(~decade) +
    labs(x = "Condition", y = "Average Price",
    title = "Impact of condition on average price across time") +
    theme_minimal()
```

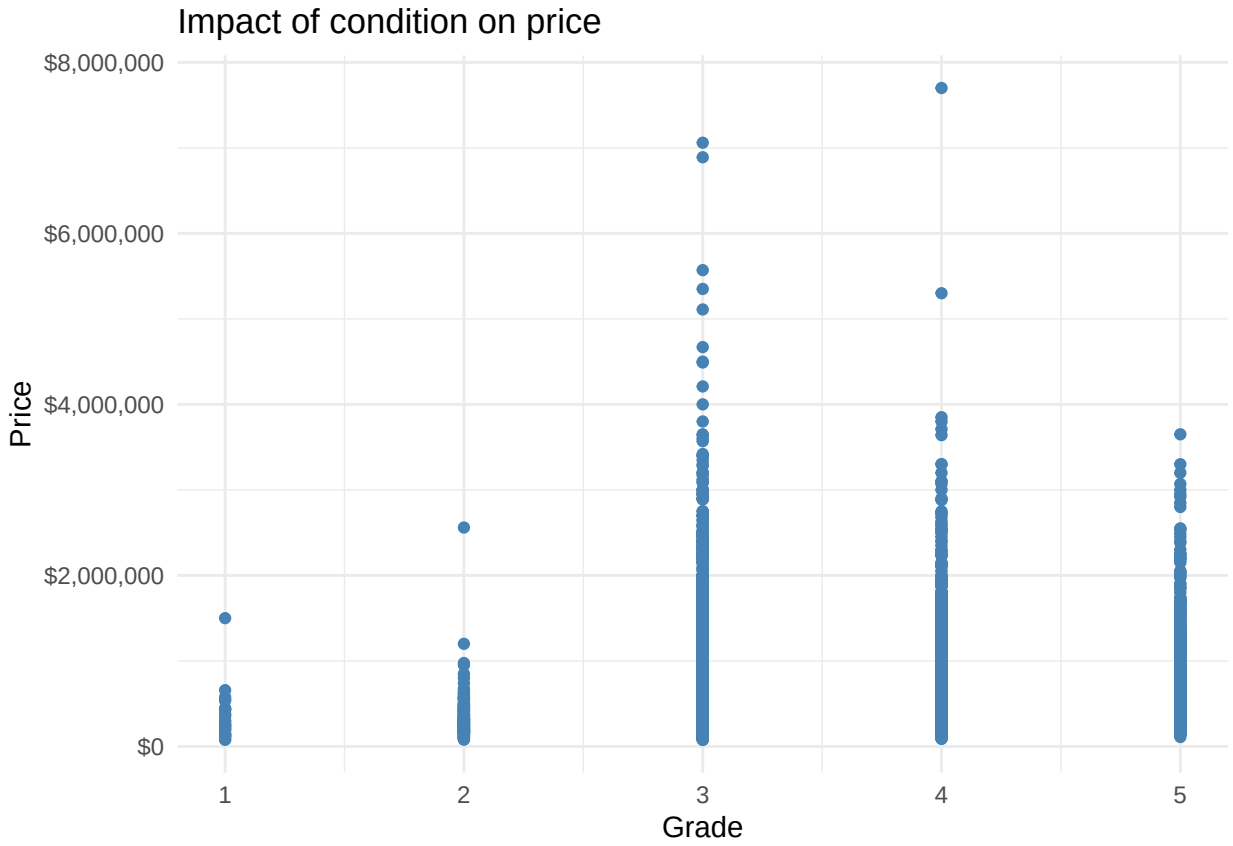



a. In general across time frame of 20th century **condition does impact** the house **price**, better the condition higher the price.

b. Post 1970 all houses that were built were with condition 2 and above.

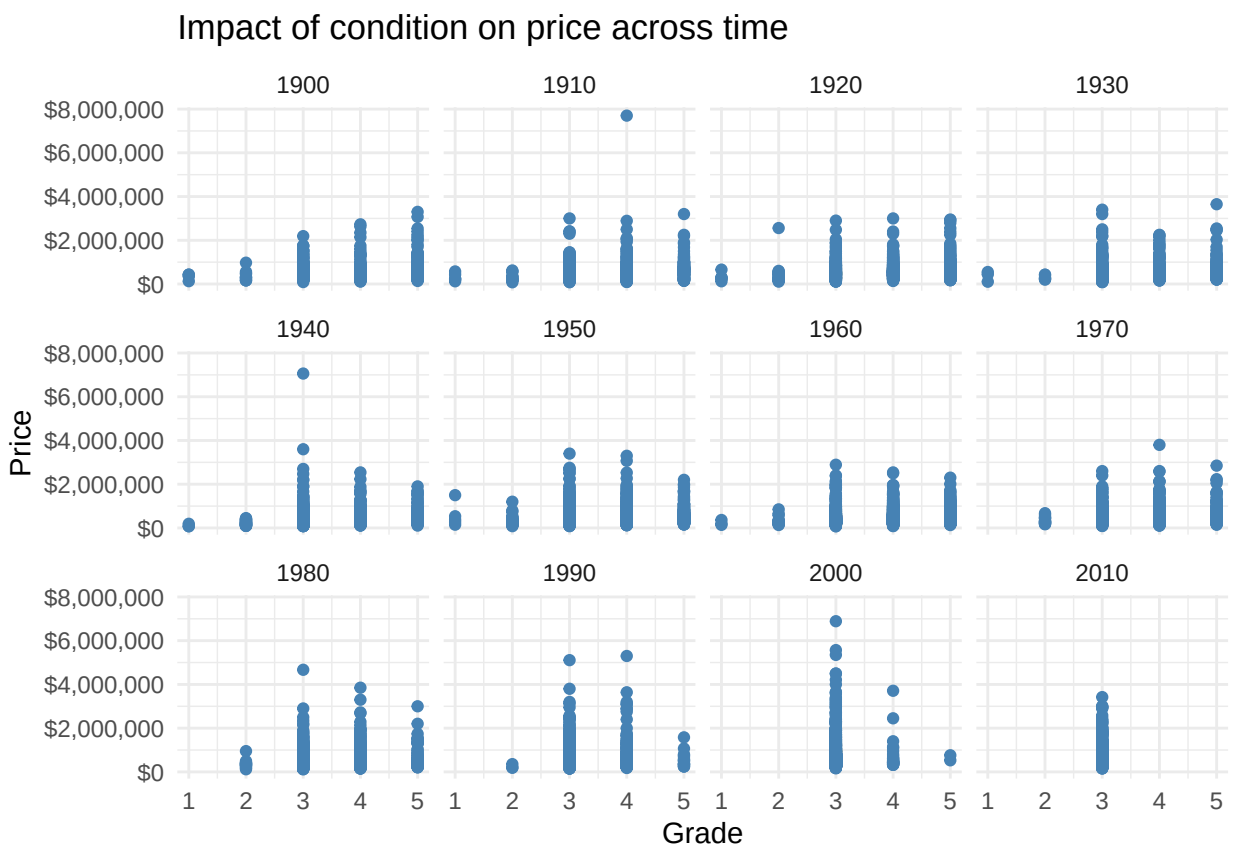
c. In general if consider all the data points instead of average it is evident condition 3 is where maximum number of houses are and some of the house prices of condition 3 are higher than condition 4 and 5. Below visualization shows overall trend of price and condition.

```
ggplot(houses, aes(x = condition, y = price)) +
  geom_point(color='steelblue') +
  scale_y_continuous(labels = dollar) +
  labs(x = "Grade", y = "Price",
  title = "Impact of condition on price") +
  theme_minimal()
```




d. Oveall there is increase in price from 1 to 3, but with condition 3, 4, 5 there is no significant impact on price. Infact in decade of 2000 there are more nouber of houses in condition 3 with higher prices than in 4 and 5.

```
ggplot(houses, aes(x = condition, y = price)) +
  geom_point(color='steelblue') +
  scale_y_continuous(labels = dollar) +
   facet_wrap(~decade) +
  labs(x = "Grade", y = "Price",
  title = "Impact of condition on price across time") +
  theme_minimal()
```



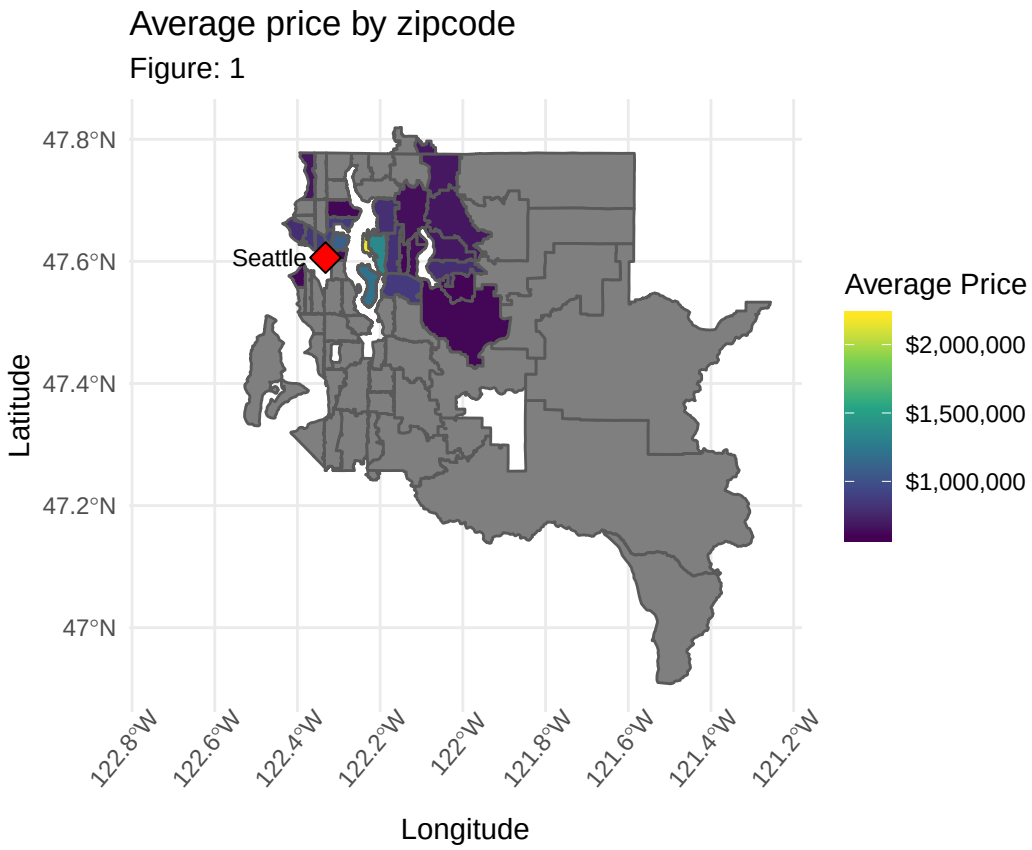

## Question 7

If you can figure out the maps then are location, grade, and prices concentrated in certain zipcodes

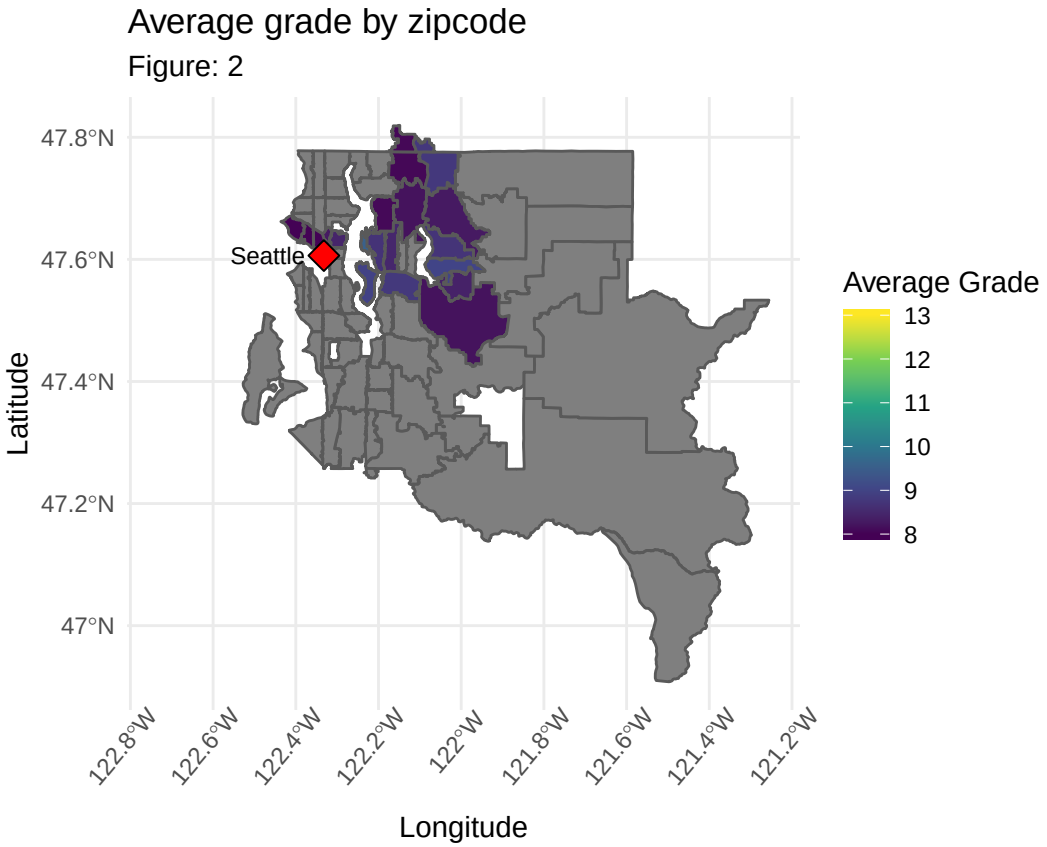
In order to find correlation between grade, prices and zip code, plotted 2 spatial visualizations
1. Figure: 1 - Average Price by zip codes and highlighted only those houses which have price higher than $600,000
2. Figure: 2 - Average Grade by zip codes and highlighted only those houses which have grade higher than 7

Below are the 2 visualizations (Figure: 1 and Figure: 2).

```
mergedShapeAndAvgZipCodeData %>%
  ggplot() +
  geom_sf(aes(fill=averagePrice)) +
  geom_point(data = sites, aes(x = longitude, y = latitude), size = 4,
        shape = 23, fill = "red") +
  geom_text(data = sites, aes(x = longitude, y = latitude), label = 'Seattle', position =
              position_dodge(width = 0.8), size = 3, hjust = 1.25) +
  scale_fill_viridis_c("Average Price", labels = dollar, limits=c(600000, 2200000)) +
  labs(x = "Longitude", y = "Latitude",
  title = "Average price by zipcode",subtitle = "Figure: 1") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle =50, hjust=0.75))
```




```
mergedShapeAndAvgZipCodeData %>%
  ggplot() +
  geom_sf(aes(fill=averageGrade)) +
  geom_point(data = sites, aes(x = longitude, y = latitude), size = 4,
        shape = 23, fill = "red") +
  geom_text(data = sites, aes(x = longitude, y = latitude), label = 'Seattle', position =
              position_dodge(width = 0.8), size = 3, hjust = 1.25) +
  scale_fill_viridis_c("Average Grade", limits=c(8, 13)) +
  labs(x = "Longitude", y = "Latitude",
  title = "Average grade by zipcode",subtitle = "Figure: 2") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle =50, hjust=0.75))
```




1. If you compare the visualizations with grade and price, then same set of zip codes are getting highlighted, so there is clear indication that houses with higher grades and higher prices are concentrated in certain set of zip codes.

2. These concentrated zip codes are around Seattle, towards slight North East and slight South West.