R-Script Talking Points
King County Housing Prices

Libraries:

Ggmap – This library is a package built on top of ggplot that allows you to do statistical analysis in combination with geographical information. Since my dataset contains longitude and latitude, I wanted to visualize where some of the most expensive houses are located visually

Dplyr – This package makes it easy to work with tabular data such as mine. I like using the glimpse() compared to the built in str() to see a quick over view of my data

Script Walkthrough:

Misc.
I downloaded the .csv file for my dataset from a Kaggle dataset which I then read in via the read.csv() and stored as house.data

Since this is a larger dataset, I only wanted to use the first 5000 records. I made sure that the order wasn't important in this dataset

After I imported the data I attached the data I attached the variables names to the script so I could access them using the attach()

1. **Data Exploration**

Before doing any statistical analysis I wanted to view the data in my dataset and what type of data it was. To do this I used the glimpse() from the DPLYR package

Next I wanted to see some basic summary statistics of my data so I used the summary() built into R

The prices in this dataset are quite large, so to make them easier to work with I created a new variable to store all the house prices in terms of hundreds of thousands

I wanted to see how some of my data was distributed, so I ran a histogram on my prices

Next I wanted to see the distribution of bedrooms so I again created a bedrooms histogram

Lastly I wanted to see the distribution of conditions to get a sense for the quality of houses in that area.

Next I wanted to see the relationship between price and square feet so I created a scatter plot

I also wanted to look at prices by number of bedrooms.

2. **MLR Model**

Being that my data contains price, I was interested in seeing if I could accurately predict price based on a number of characteristics

I created a model called house.model that stores a linear regression model

Then I wanted to see how my model performed, so I ran summary on that model I just created to see how it did and what variables were the best predictors

3. **ANOVA**

I was interested in finding out whether or not house prices varied over time, every 25 years to be exact. So I ran an ANOVA

First I created a new data frame containing only the price and year built columns. I have to convert the years into groups, or factors so I can do an ANOVA.

First I found the oldest house

Then I found the newest house

Then I ran a histogram to see the distribution of years

Next I created a for loop that grouped each price into a specific factor based on the year it was built, there are 5 groups spanning from 1900 – present day

Before running an anova I made sure that the centuries were in factor by using the as.factor() on the decade column of the priceByDecade df I jest created

Then I ran a summary on the anova to see if it is significant

After the summary I also ran Tukey's test

After the tests were done, I plotted the relationship using a box plot

4. GGMAP

This was my favorite part of the project, as it provided me with tools to visualize my data on a real map of where the data was from. It is a lot more interesting than just a bar chart or a simple chart.

To make the map I first needed to get the data I wanted to plot. I was interested in seeing where the most expensive houses are. To determine what is 'expensive' I ran a summary on the house.data$price to see summary statistics. I decided to use the quartile 3 as a cutoff point. This happened to be 650,000+ is considered expensive

Next I created a new data frame containing only the rows that met this criteria and stored those records as mostExpensiveHouses. I also only stored the longitude, latitude, and price in this new table

Now that I have the data I have to get a map of King County.

I create a new variable called map and store a map in it using the get_map() from GGmap of King County, Washington where my data is from

Then I create a new variable called m that stores the actual map I will use to plot the data on

Next I plot green points on the map of the most expensive houses