Assignment #1

Jacob Shore & Derek Huang

February 5, 2018

Name of the GitHub repository: jbshore/Math158-Semester-Project Location of the GitHub repository: https://github.com/jbshore/Math158-Semester-Project

This dataset describes housing in King County, WA (Seattle area). It includes homes sold between May 2014 and May 2015, for a total of 21,000 observations. It was originally assembled by user "harlfoxem" on kaggle.

21,000 houses, while surely not the entirety of property sales in Seattle, is a very large sample. The plot of latitude/longitude below indicates that the sample is spread relatively uniformly across the county as well, so this should be a dataset that well-represents the actual population (of house sales).

Each row in the dataset is a particular house, described by 19 variables including numeric variables such number of bedrooms, number of bathrooms, and square footage of the lot and house itself, as well as categorical variables such as condition, grade, zipcode, and year, and a binary variable, waterfront.

The variables include price, the number of bedrooms/bathrooms, the size of the house/parking lot/basement, the number of floors, whether waterfront or not, overall condition/grade, built year, renovation year, zip code and latitude/longitude coordinate.

Appropriate summary statistics:

1. Mean price: $540088 Standard deviation: $367127.1 Waterfront mean price: $1661876 on 163 observations This is very indicative of the Seattle market, which is one of the priciest in the country. Also note the high standard deviation, suggesting a wide range of available houses.
2. Median price: $450000. The median below the mean suggests that the data is skewed to the right, and there are outliers in the higher price ranges (makes sense). The first plot below confirms this, and shows that housing prices are sort of bell-shaped with a very heavy positive tail.
3. Average size of the house/Living room area: The average square footage of the house is 2080. There is a positive relationship between the price and the living room area. The bigger the living space, the more expensive the house is, which makes sense.
4. Overall grade: grade 7 is the most common, followed by grade 6. Seems to be relatively bell-shaped (good job appraisers!). Grade seems to have a pretty strong effect on pricing, but a high grade does not guarantee a high price.
5. Overall condition: condition 3 is the most common, followed by condition 4. Again, relatively bell-shaped.
6. Latitude/longitude coordinate: The majority of the houses are located in the range of 122.40 - 122 (longitude), 47.3 -47.8(latitude). And it seems like the most expensive houses are located around longitude 122.25. The lat-long plot shows most houses are in the city, but Vashon Island and some other rural areas of the east-sound are also included.
7. Year built: Generally, the more recent that the house is built, the more expensive it is, which also makes sense.
8. Waterfront: Given two houses of the same size, the waterfront one is most likely more expensive than the one that is not waterfront.
9. View: Like waterfront, a better view translates to more expensive property most of the time. We wouldn't be surprised if view was correlated with waterfront, and possibly some other variables though. It is also unclear how this was measured.

Things of interest:

Overall, the data are what we expected. Price of a house is generally affected by variables such as waterfront, square footage, and location. Honestly, we were a bit surprised by how strong and straightforward some of these relationships are (especially square footage).

Some interesting caveats:

1. Pricing is generally not affected by number of bedrooms, suggesting maybe the more expensive houses are still geared towards smaller families/retirees, rather than enormous families.
2. Square footage of lot and living seem to be rather uncorrelated — a larger lot does not correspond to a larger house.

Note that many of the plots created below have been supressed, with the most interesting ones shown.

```
mean(housing$price)
```

```
## [1] 540088.1
```

```
sd(housing$price)
```

```
## [1] 367127.2
```

```
waterfront <- subset(housing, housing$waterfront == 1)
mean(waterfront$price)
```

```
## [1] 1661876
```

```
median(housing$price)
```

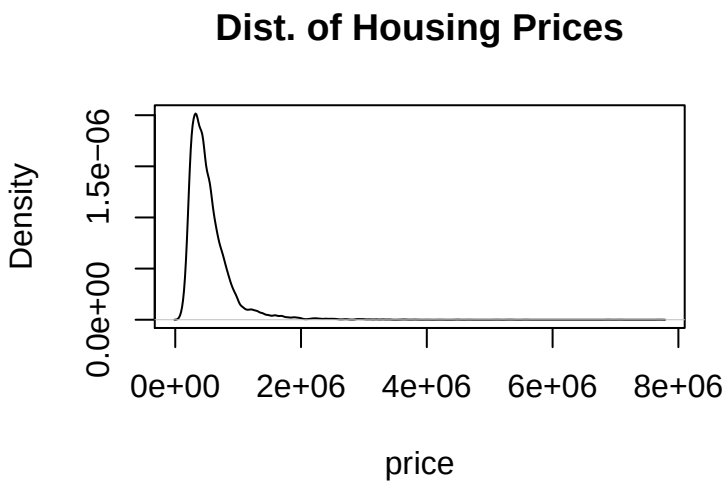
```
## [1] 450000
```

```
mean(housing$sqft_living)
```

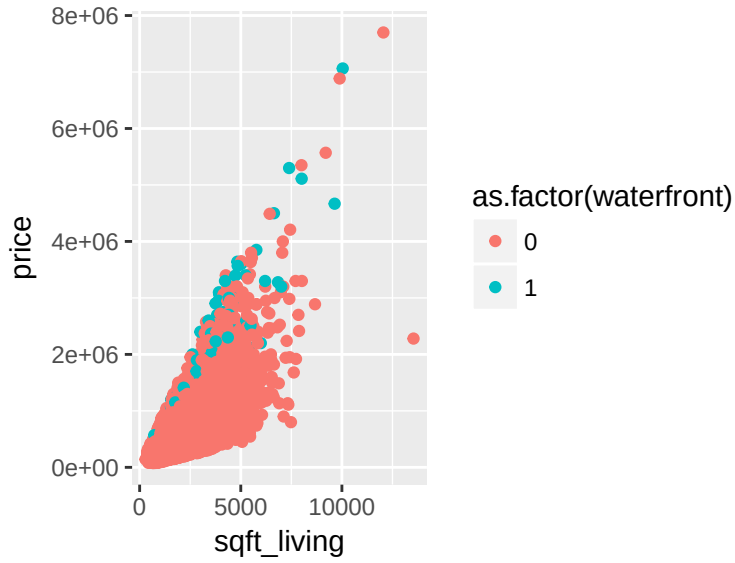
```
## [1] 2079.9
```

```
housing.1 <- subset(housing, housing$bedrooms < 15)
zips <- housing %>% group_by(zipcode) %>% arrange(desc(price)) %>% select(zipcode, price)

plot(density(housing$price), main = "Dist. of Housing Prices", xlab = "price")
```

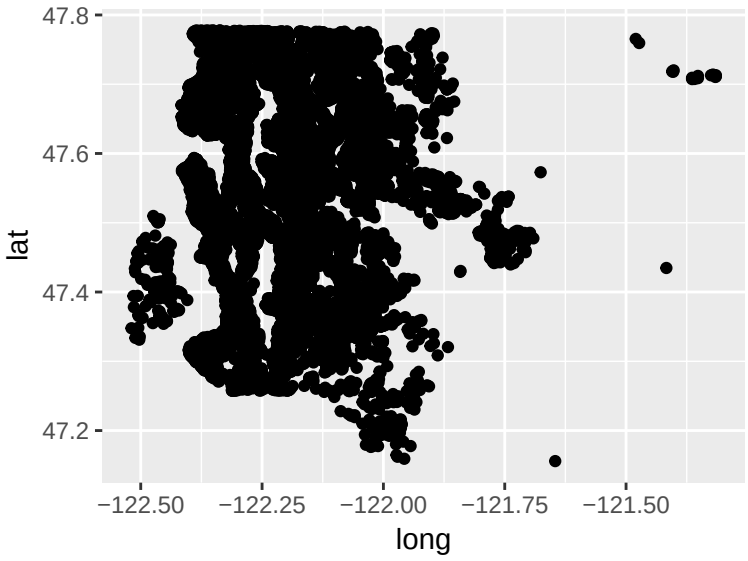



```
# housing price distribution
ggplot(housing)+geom_point(aes(x = sqft_living, y = price, col = as.factor(waterfront)))
```




```
# house size vs price, with waterfront highlighted
ggplot(housing)+geom_point(aes(x = sqft_living, y = price, col = as.factor(view)))
```


```
# house size vs price, with view highlighted
ggplot(housing)+geom_point(aes(x = long, y = lat))
```




```
# map of houses roughly corresponds to shape of King County. Seattle is visible between lat 47.4-47.6 a
#ggplot(housing)+geom_point(aes(x = long, y = price))
# longitude vs price
#ggplot(housing)+geom_point(aes(x = yr_built, y = price))
# year built vs. price
#ggplot(housing)+geom_bar(aes(grade))
# dist. of grades
#ggplot(housing)+geom_point(aes(grade,price))
# grade vs. price
#ggplot(housing)+geom_bar(aes(condition))
# dist. of conditions
#ggplot(housing.1)+geom_point(aes(bedrooms,price))
# number of bedrooms vs price
#ggplot(housing)+geom_point(aes(sqft_lot, sqft_living))
# house size vs lot size

zips[1:10,]
```

```
## # A tibble: 10 x 2
## # Groups:   zipcode [6]
##    zipcode   price
##      <int>   <dbl>
##  1   98102 7700000
##  2   98004 7062500
##  3   98039 6885000
##  4   98039 5570000
##  5   98004 5350000
##  6   98040 5300000
##  7   98033 5110800
##  8   98040 4668000
##  9   98155 4500000
## 10   98004 4489000
```

```
# most expensive zipcodes
```