Example Exploration

Clark

This is *a* way to answer the explorations, it is certainly not the *only* way.

Here I will go through Exploration P.B.: Housing Prices in Michigan

Problem 1

Prior to looking at the data a few things would be, waterfront or not, size of house, number of bedrooms, neighborhood the house is in, size of garage, etc.

Problem 2

The observational units are houses for sale north of Lake Macatawa, the response variable is price, which is quantitative. As it is quantitative we should look at sample mean, median, variance, perhaps skewness. We could use a histogram or boxplot to visualize this.

Problem 3 (Explore the data)

First we pull in the data from the website and run `glimpse()` to see the structure of the data. I like to do this to ensure that R is indeed reading in factors as factors and numerics as numerics.:

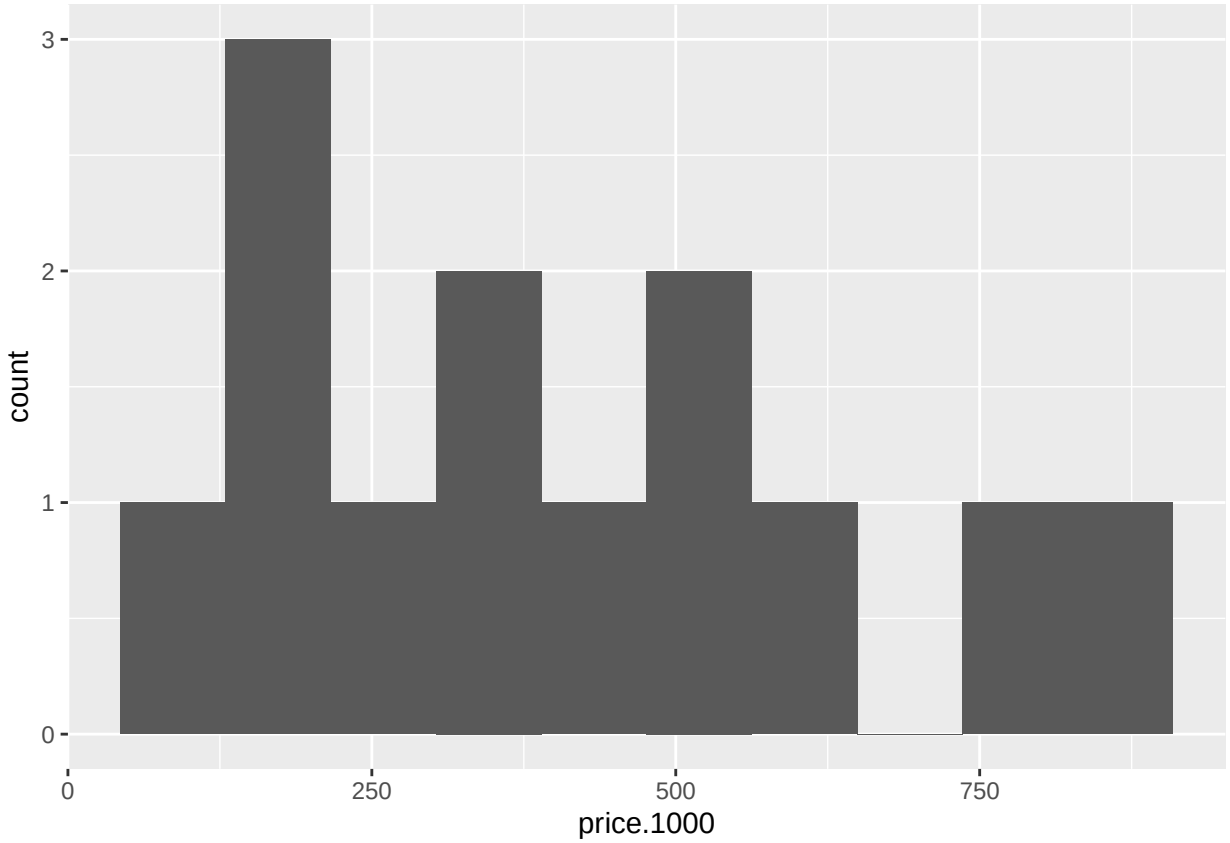
```
library(tidyverse)
housing.data<-read.table("http://www.isi-stats.com/isi2/data/homeprices.txt" ,header=T)
glimpse(housing.data)
```

```
## Observations: 13
## Variables: 3
## $ sqft       <int> 2700, 3353, 1600, 1740, 1907, 3262, 1336, 2410, 903, 316...
## $ price.1000 <dbl> 639.9, 898.0, 410.0, 529.9, 495.0, 749.9, 199.9, 324.9, ...
## $ lake       <fct> lakefront, lakefront, lakefront, lakefront, lakefront, l...
```

As an aside, say we are confused as to what R is doing here. We have both something called `dbl` and something called `int`. I would **highly** encourage you in this course to explore things like this. A simple google search would reveal something like https://stackoverflow.com/questions/23660094/whats-the-difference-between-integer-class-and-numeric-class-in-r and we can see what R is doing.

Back to the question at hand. To describe the shape we run

```
housing.data %>% ggplot(aes(x=price.1000))+geom_histogram(bins=10)
```



Here's where a bit of the art of statistics. I chose 10 bins, but I could easily have picked a different number. We want to ensure that the number of bins are large enough to group efficiently but small enough to show the structure of the data. Here it would appear that the data are skewed positively (also called right skewed).

Problem 4

The mean will over predict the price of many houses, as the data are skewed it may be better to use the median.

Problem 5

The mean should be higher than the median, which we can check:

```
xbar<-mean(housing.data$price.1000)
xtilde<-median(housing.data$price.1000)
xbar
```

```
## [1] 408.0615
```

```
xtilde
```

```
## [1] 324.9
```

Which matches our intuition

Problem 6

We can calculate by hand:

```
samp.var <- sum((housing.data$price.1000-xbar)^2)/(nrow(housing.data)-1)
samp.sd <- sqrt(samp.var)
samp.sd
```

```
## [1] 240.9228
```

Or we can rely on the built in function `sd()`

```
sd(housing.data$price.1000)
```

```
## [1] 240.9228
```

The interpretation is on average our houses are 241,0000 from the mean

Problem 7

Straight forward calculation

```
639-xbar
```

```
## [1] 230.9385
```

So we see we would be $ 230,938 from the correct price

Problem 8

The residual for the first house is:

```
all.resids<-housing.data$price.1000-xbar
all.resids[1]
```

```
## [1] 231.8385
```

So we under predicted the price of the first house

Problem 9

Here we have a vector called `all.resids` so let's take advantage of that to answer our question. To do this we will buid a logical. This returns true if our residual is positive, false if it is negative.

```
pos.resid <- (all.resids>0)
```

Now we can just sum this up

```
sum(pos.resid)
```

```
## [1] 6
```

To find the falses we could do:

```
length(pos.resid)-sum(pos.resid)
```

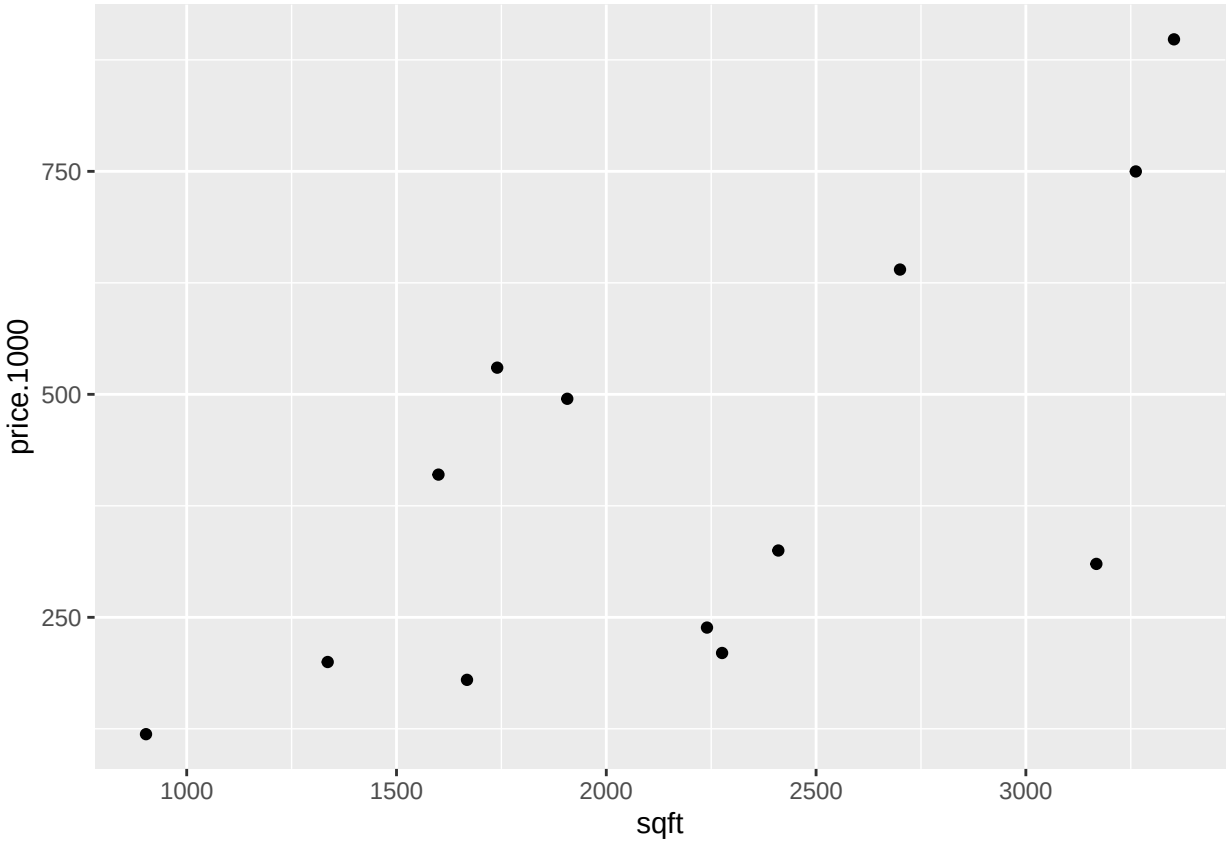
```
## [1] 7
```

So 6 over predictions, 7 under predictions

Problem 10

To see the association between square footage and price we can create a scatter plot

```
housing.data %>% ggplot(aes(x=sqft,y=price.1000))+geom_point()
```



From the plot it would appear that as square footage increases, price also increases, but not always.

Problem 11

From MA206, we can build a statistical model

$$
\begin{aligned}
i &= \text{House} \\
y_i &= \text{Price of house } i \\
x_{1,i} &= \text{Square footage of house } i \\
y_i &= \beta_0 + \beta_1 x_{1,i} + \epsilon_i
\end{aligned}
$$

Recalling our previous course, we can fit this using the `lm()` function in R

```
house.lm<-lm(price.1000~sqft,data=housing.data)
summary(house.lm)
```

```
##
## Call:
## lm(formula = price.1000 ~ sqft, data = housing.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -304.70 -128.44  -13.74  128.98  244.04
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -59.36870  161.36807  -0.368   0.7199
## sqft          0.21274    0.06963   3.055   0.0109 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 185.1 on 11 degrees of freedom
## Multiple R-squared:  0.4591, Adjusted R-squared:  0.4099
## F-statistic: 9.335 on 1 and 11 DF,  p-value: 0.01094
```

So our estimates are: $\hat{\beta}_0 = -59$, $\hat{\beta}_1 = 0.21$, and $\hat{\sigma} = 185$. Note that $\hat{\sigma}$ is the standard error of the residuals.

The intercept here is meaningless as it gives the price of a house with 0 sqft. The slope means that a change in one sqft corresponds to an increase in price of 212 dollars. (Recall our response is price/1000)

The standard error of the residuals for the first house can be found using:

```
house.lm$residuals[1]
```

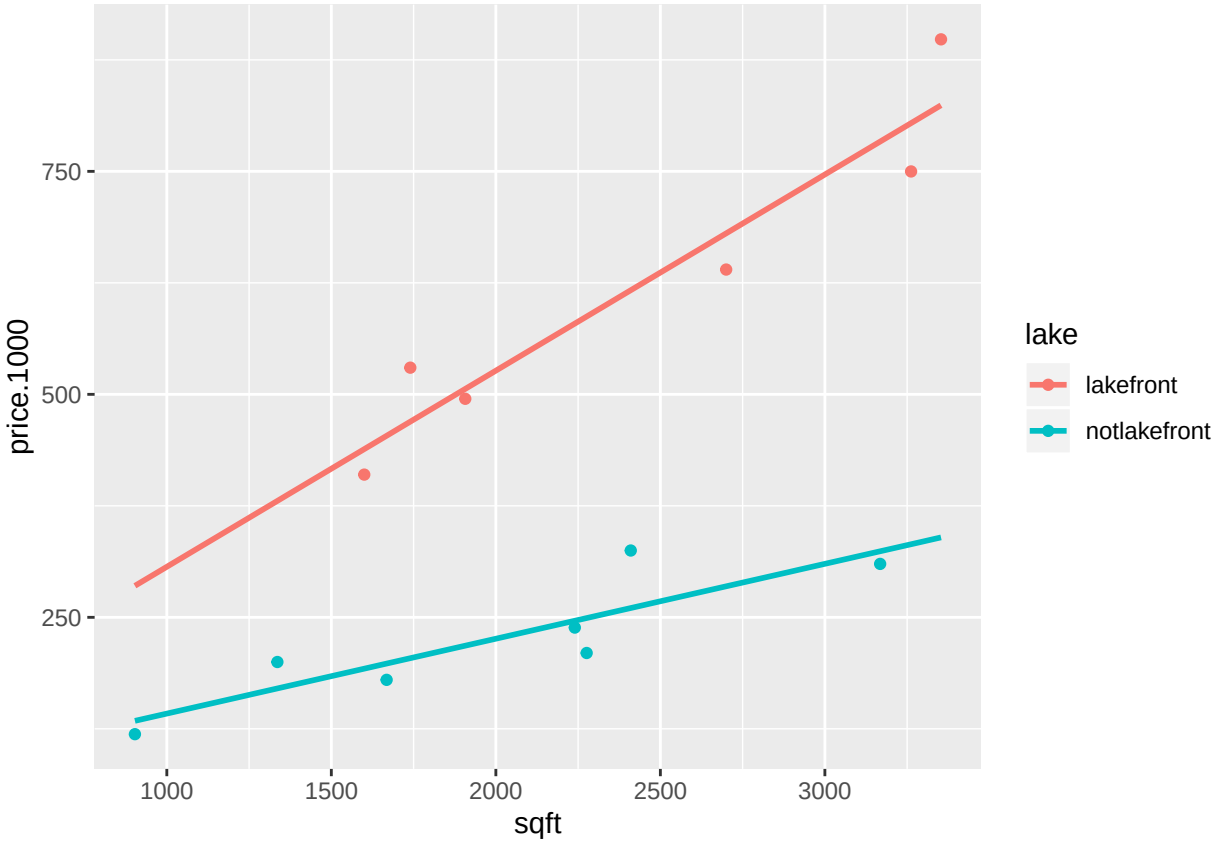
```
##        1
## 124.8612
```

And the standard error of the residuals is smaller than before.

Problem 12

This question is asking us to fit a seperate regression line for lake front and not lake front. Essentially we are treating them as two different datasets.

```
housing.data %>% ggplot(aes(x=sqft,y=price.1000,color=lake))+geom_point()+
  geom_smooth(method='lm',se=FALSE,fullrange = TRUE)
```



Problem 13

For a tricky reason, this problem is actually not straight forward. I'm going to answer it in a way that is slightly different then perhaps the approved solution (though I make the approved solution so...) We need differing slopes and differing intercepts for our two groups, which suggests an interaction term in our statistical model.

The complete model would be:

$i =$ House
$y_i =$ Price of house i
$x_{1,i} =$ Square footage of house i
$x_{2,i} = 1$ If not lakefront, 0 otherwise$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{1,i} x_{2,i} + \epsilon_i$

Which we fit via:

```
inter.lm<-lm(price.1000~sqft*lake,data=housing.data)
summary(inter.lm)
```

```
##
## Call:
## lm(formula = price.1000 ~ sqft * lake, data = housing.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -54.16 -28.60 -14.15  29.64  73.93
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              86.76438   71.60612   1.212  0.25648
## sqft                      0.21990    0.02831   7.769  2.8e-05 ***
## lakenotlakefront        -28.65098   91.32560  -0.314  0.76088
## sqft:lakenotlakefront    -0.13595    0.03895  -3.491  0.00682 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 49.48 on 9 degrees of freedom
## Multiple R-squared:  0.9684, Adjusted R-squared:  0.9578
## F-statistic: 91.84 on 3 and 9 DF,  p-value: 4.547e-07
```

Meaning, for lakefront houses our fitted model is $\hat{y}_i = 86.8 + 0.22x_{1,i}$ and for not lakefront our fitted model is $\hat{y}_i = 58.11 + 0.085x_{1,i}$

Problem 14

Our first house is 2700 sq ft and is on the lakefront so the predicted price is $86.8 + .22(2700) = 680.8$ the actual value is 639

Problem 15

The standard error of the residuals is now 49.5 so we have reduced it by quite a bit.

Problem 16

To see if it is coonfounded we want to know whether increasing square footage also increases the probability a house is on the lakefront or not. What we see here is out of our four biggest houses, three of them are on the lakefront and out of our four smallest houses three of them are not on the lakefront. Therefore we certainly have the potential for their to be confounding. Furthermore, from the output in problem 13 we see that both lakefront and square footage seem to impact price.

Our diagram is:

```
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
library(knitr)
sv.diagram<-data.frame(Observed_variation_in=c("Housing Price"),explained_variation=c("Square Footage, a

kable(sv.diagram, "latex",booktabs = T)%>%
  kable_styling(full_width = F)%>%
  column_spec(2:3, width = "10em")
```

Observed_variation_in	explained_variation	unexplained_variation
Housing Price	Square Footage, and Lakefront	Neighborhood, Number of Bedrooms, garage size

Problem 17

I would summarize by using the statistical model that includes location and home size, as this model explained more variance than the other models considered in the study. My conclusion is that location and home size has a positive impact on the home price - lake front properties are generally going to be more expensive and the larger the home on the lakefront will increase the property value. I would not be willing to generalize this study to a larger population of homes, as there are many other factors that could explainvariability in home prices, such as school districts, location to major cities, etc. In this study, home size and location are extremely likely to cause variation in home prices.

Problem 18

Information about homes in other areas would be particularly useful to this study, as this study only accounts for two possible explanations of variations and is very specific to the Lake Macatawa area.