# Homework 6

*Alexander Van Roijen*

*February 14, 2019*

```
require('ggpubr')
```

```
## Loading required package: ggpubr

## Loading required package: ggplot2

## Loading required package: magrittr
```

```
cellData = read.csv("cells.csv")
salesData = read.csv("Sales.csv")
```

## Problem Set 1

### Problem 1.1

```
fTest = aov(count1~factor(dose),data=cellData)
print(summary(fTest))
```

```
##              Df  Sum Sq Mean Sq F value   Pr(>F)    
## factor(dose)  2 2701378 1350689   14.62 2.09e-05 ***
## Residuals    37 3417120   92355                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that there is a significant difference between the means of their post treatment cell count. Further, due to the non-observational nature of this study, we can conclude that it is in fact the level of dosage that is impacting this difference

### Problem 1.2

```
flmTest = lm(count1~dose,data=cellData)
print(summary(flmTest))
```

```
## 
## Call:
## lm(formula = count1 ~ dose, data = cellData)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -524.57  -70.44  -14.13   40.88 1401.43 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   30.369     62.254   0.488    0.628    
## dose           5.732      1.047   5.474 2.99e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 300 on 38 degrees of freedom
## Multiple R-squared:  0.4409, Adjusted R-squared:  0.4262
## F-statistic: 29.97 on 1 and 38 DF,  p-value: 2.993e-06
```

Again, we can see that the dose is a significant parameter in our regression, and in further detail, tells us there is a positive linear relationship between post treatment cell count and dosage. This follows in accordance with our anova test. In more concrete terms, our linear regression model states that for every unit change in dose, we expect to see a 5.732 mean increase in our cell count.

### Problem 1.3

```
#maybe use indicator random variables here?
fwsTest = aov(count1~factor(dose)+factor(sex),data=cellData)
print(summary(fwsTest))
```

```
##              Df  Sum Sq Mean Sq F value   Pr(>F)    
## factor(dose)  2 2701378 1350689  14.274 2.73e-05 ***
## factor(sex)   1   10610   10610   0.112     0.74    
## Residuals    36 3406510   94625                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
flmwsTest = lm(count1~dose+sex,data=cellData)
print(summary(flmwsTest))
```

```
##
## Call:
## lm(formula = count1 ~ dose + sex, data = cellData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -508.42  -71.45   -0.80   42.48 1388.58 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)    12.90      84.06   0.153    0.879    
## dose            5.73       1.06   5.407 3.98e-06 ***
## sex            30.49      97.12   0.314    0.755    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 303.7 on 37 degrees of freedom
## Multiple R-squared:  0.4424, Adjusted R-squared:  0.4123
## F-statistic: 14.68 on 2 and 37 DF,  p-value: 2.027e-05
```

According to both results, sex has no significance when presented at the same time as our dosage level, despite the lack of interaction included in the model. However, we do see our p-value decrease slightly in both cases, meaning that some variation was explained by the sex regardless of its lack of significance. I will note this change is only slight.

**Problem 1.4**

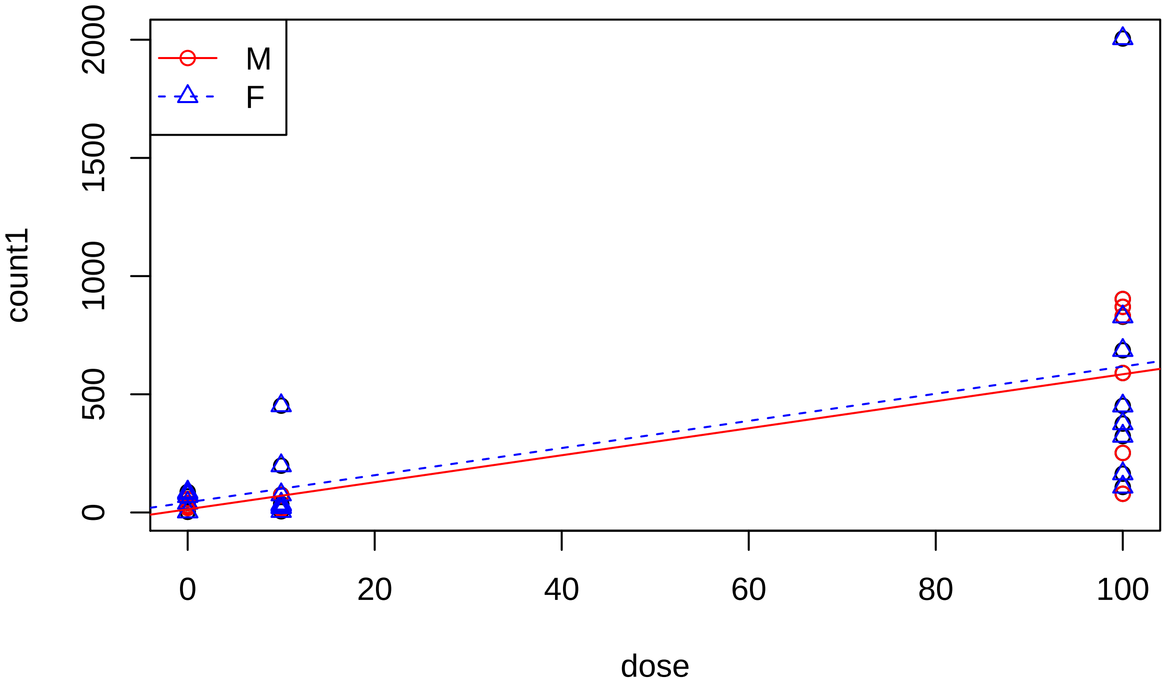
```
#also maybe use indicators here?
fwsITest = aov(count1~factor(dose)*factor(sex),data=cellData)
print(summary(fwsITest))
```

```
##                           Df  Sum Sq Mean Sq F value   Pr(>F)    
## factor(dose)               2 2701378 1350689  13.496 4.85e-05 ***
## factor(sex)                1   10610   10610   0.106    0.747    
## factor(dose):factor(sex)   2    3875    1938   0.019    0.981    
## Residuals                 34 3402635  100077                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
flmwsITest = lm(count1~dose+sex+dose:sex,data=cellData)
print(summary(flmwsITest))
```

```
## 
## Call:
## lm(formula = count1 ~ dose + sex + dose:sex, data = cellData)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -509.28  -71.04   -0.78   41.90 1387.72 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 13.59571   97.26802   0.140  0.88962   
## dose         5.71211    1.63049   3.503  0.00125 **
## sex         29.25634  128.97889   0.227  0.82184   
## dose:sex     0.03216    2.16753   0.015  0.98824   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 307.8 on 36 degrees of freedom
## Multiple R-squared:  0.4424, Adjusted R-squared:  0.3959 
## F-statistic: 9.521 on 3 and 36 DF,  p-value: 9.106e-05
```

```
plot(count1 ~ dose, data=cellData)
points(cellData$dose[cellData$sex==0],cellData$count1[cellData$sex==0],pch=1,col=2)
points(cellData$dose[cellData$sex==1],cellData$count1[cellData$sex==1],pch=2,col=4)
abline(lm(count1 ~ dose, data = cellData, subset=(sex==0)),lty=1,col=2)
abline(lm(count1 ~ dose, data = cellData, subset=(sex==1)),lty=2,col=4)
legend("topleft",c("M","F"),pch=1:2,lty=1:2,col=c(2,4))
```

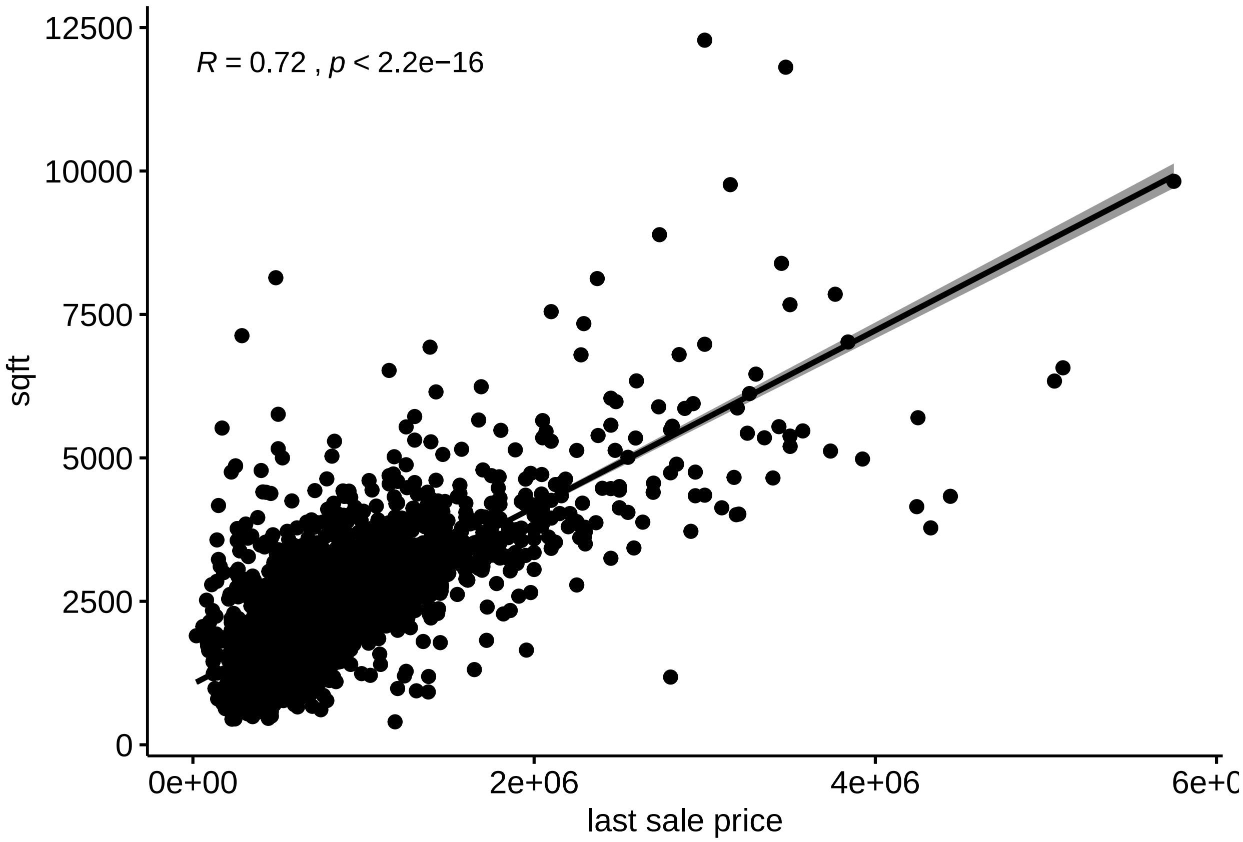


According to both the Anova and regression model, there is no clear interaction between the dosage between males and females. We can further see this through the parallel lines produced by plotting their regressions.It goes to show that our coefficient of dose for both males and females is roughly the same.

## Problem Set 2

### Problem 2.1a

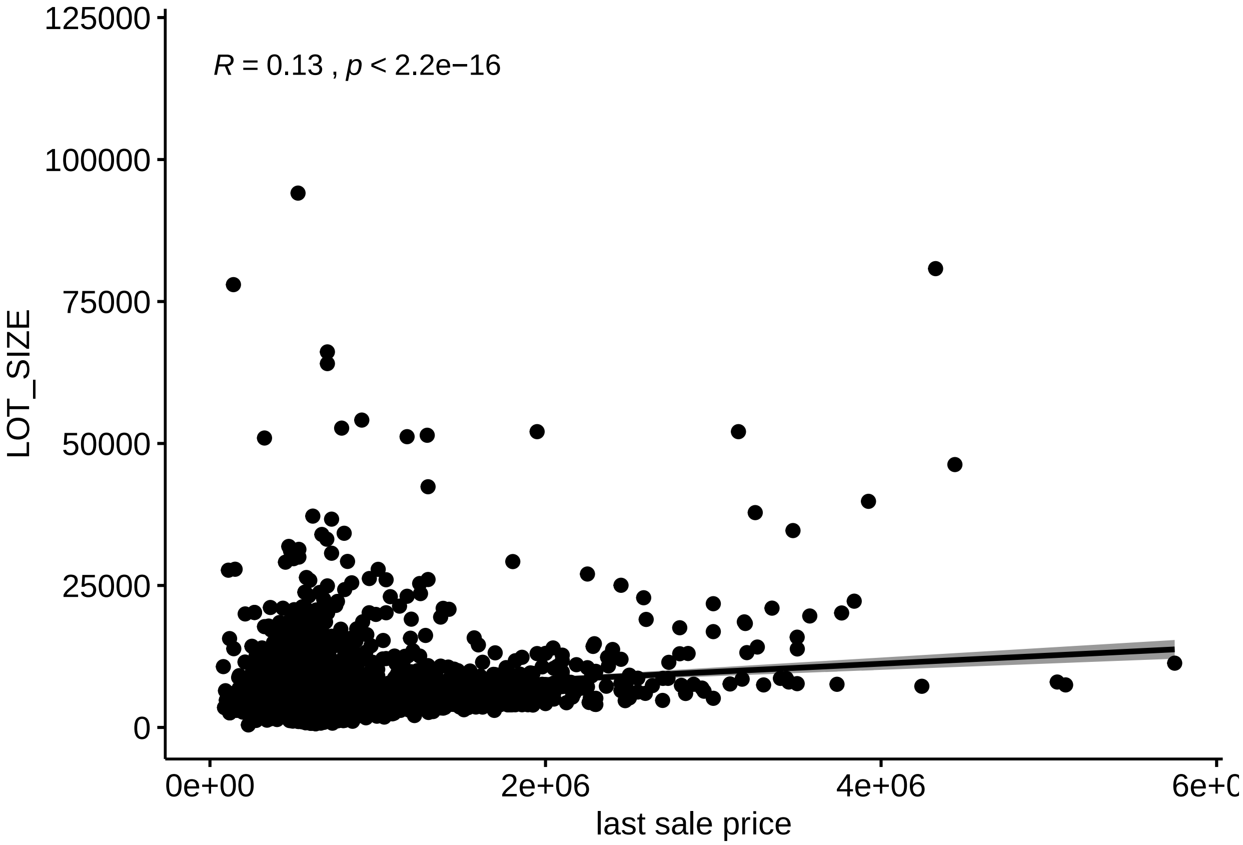
```
par(mfrow=c(1,2))
ggscatter(salesData, x = "LAST_SALE_PRICE", y = "SQFT",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "last sale price", ylab = "sqft")
```



We can see a strong positive linear trend, shown by the very low p-value and positive person correlation coefficient expressed by R

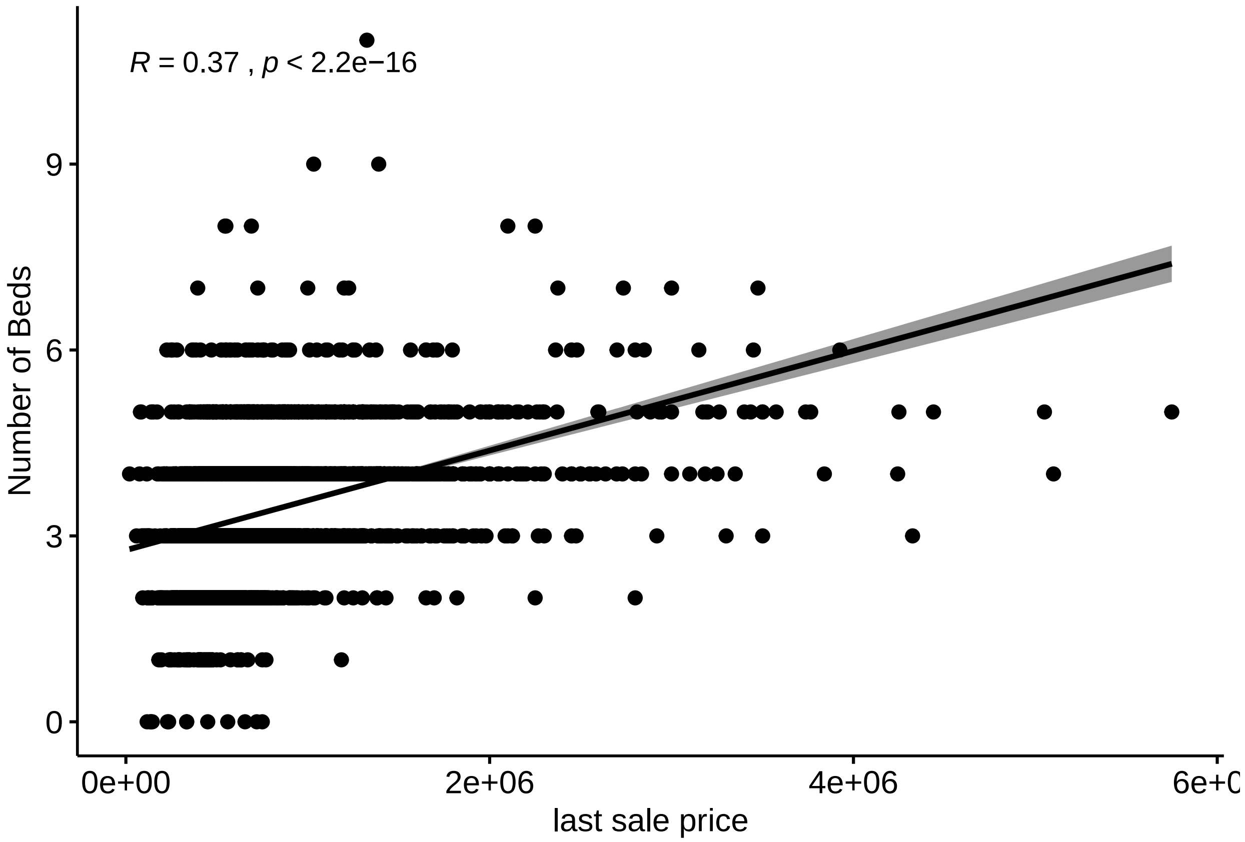
**Problem 2.1b**

```
ggscatter(salesData, x = "LAST_SALE_PRICE", y = "LOT_SIZE",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "last sale price", ylab = "LOT_SIZE")
```



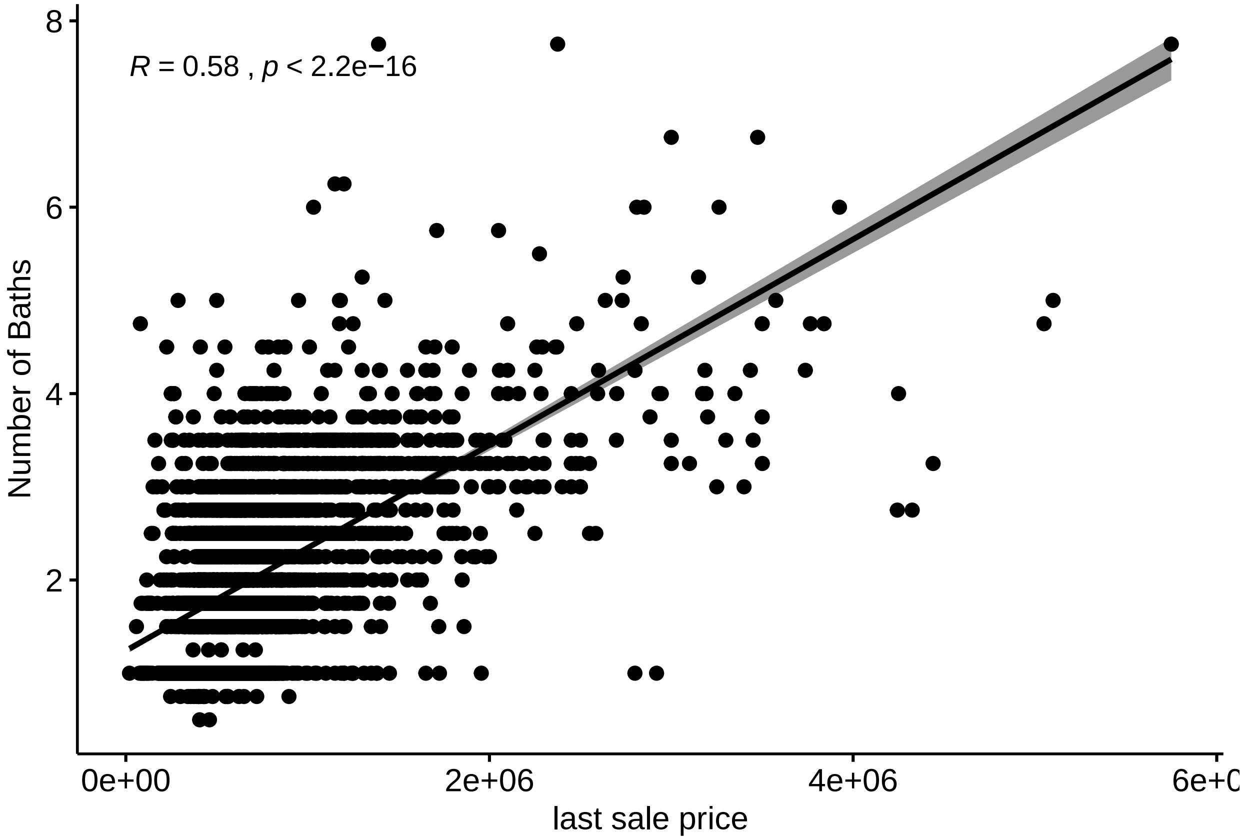
Again, we see a positive trend between last sale price and lot size, indicated by the low p-value and the positive correlation, but only slight, general increase in last sale price with increasing lot size. We do see some outliers that we may want to keep our eye on ##Problem 2.1c

```
ggscatter(salesData, x = "LAST_SALE_PRICE", y = "BEDS",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "last sale price", ylab = "Number of Beds")
```



According to our R value and p-value, we can see this factor of number of beds is in fact significant again with a positive correlation. However, I am slightly concerned as the relationship doesn't seem clearly linear pictorially. It may be that most beds are 5 or less and the higher end, if filled out more, would show no true relationship. Just a thought. ###Problem 2.1d

```
ggscatter(salesData, x = "LAST_SALE_PRICE", y = "BATHS",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "last sale price", ylab = "Number of Baths")
```



Similar to the previous plots, Baths seems to have a positive strong correlation with last sale price. Unlike the previous scatter between number of beds and sale price, I am more confident in this relationship pictorially.

**Problem2.2**

```
sqftvsprice = lm(LAST_SALE_PRICE~SQFT,salesData)
print(summary(sqftvsprice))
```

```
##
## Call:
## lm(formula = LAST_SALE_PRICE ~ SQFT, data = salesData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2271121  -149940   -14536   121363  3051927
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -13574.815  11452.860  -1.185    0.236
## SQFT           340.383      4.794  71.008   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 316000 on 4574 degrees of freedom
##   (119 observations deleted due to missingness)
## Multiple R-squared:  0.5243, Adjusted R-squared:  0.5242
## F-statistic:  5042 on 1 and 4574 DF,  p-value: < 2.2e-16
```

```r
spacevsprice = lm(LAST_SALE_PRICE~LOT_SIZE,salesData)
print(summary(spacevsprice))
```

```
##
## Call:
## lm(formula = LAST_SALE_PRICE ~ LOT_SIZE, data = salesData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1482805  -272636   -92585   116583  4949448
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.609e+05  1.177e+04  56.171   <2e-16 ***
## LOT_SIZE    1.234e+01  1.433e+00   8.611   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 457600 on 4093 degrees of freedom
##   (600 observations deleted due to missingness)
## Multiple R-squared:  0.01779,    Adjusted R-squared:  0.01755
## F-statistic: 74.15 on 1 and 4093 DF,  p-value: < 2.2e-16
```

```r
bedsvsprice = lm(LAST_SALE_PRICE~BEDS,salesData)
print(summary(bedsvsprice))
```

```
##
## Call:
## lm(formula = LAST_SALE_PRICE ~ BEDS, data = salesData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -983894 -236151  -69903  108014 4738101
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   150240      22194    6.769 1.46e-11 ***
## BEDS          172332       6345   27.161  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 424900 on 4588 degrees of freedom
##   (105 observations deleted due to missingness)
## Multiple R-squared:  0.1385, Adjusted R-squared:  0.1383
## F-statistic: 737.7 on 1 and 4588 DF,  p-value: < 2.2e-16
```

```r
bathsvsprice = lm(LAST_SALE_PRICE~BATHS,salesData)
print(summary(bathsvsprice))
```

```
##
## Call:
## lm(formula = LAST_SALE_PRICE ~ BATHS, data = salesData)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -1475615  -194857   -36477   130541  3494435
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   103831      14073   7.378  1.9e-13 ***
## BATHS         305628       6334  48.256  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 371900 on 4574 degrees of freedom
##   (119 observations deleted due to missingness)
## Multiple R-squared:  0.3374, Adjusted R-squared:  0.3372
## F-statistic:  2329 on 1 and 4574 DF,  p-value: < 2.2e-16
```

Clearly, separately, each factor is clearly significant with sales price, even though our R-squared values are quite varied, showing some clearly describe much more of the variance within our data than others. As a sanity check, we can easily see that the $R^2$ values are exactly the correlation values found before, squared.

- Accordingly,
  - for each increase by 1 square foot in our property, we see an average increase of last sales price by 340$
  - for each increase by 1 square foot in our lot size, we see an average increase of last sales price by 12$
  - for each increase by 1 bath in our property, we see an average increase of last sales price by 172332$
  - for each increase by 1 bathroom in our property, we see an average increase of last sales price by 305,628$

### Problem 2.3

```
togetherlm = lm(LAST_SALE_PRICE~BEDS+BATHS+SQFT+LOT_SIZE,data=salesData)
print(summary(togetherlm))
```

```
##
## Call:
## lm(formula = LAST_SALE_PRICE ~ BEDS + BATHS + SQFT + LOT_SIZE,
##     data = salesData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2270641  -136780    -6410   113262  3156003
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.364e+04  1.789e+04   5.235 1.74e-07 ***
## BEDS        -9.310e+04  6.665e+03 -13.969  < 2e-16 ***
## BATHS        8.694e+04  8.612e+03  10.095  < 2e-16 ***
## SQFT         3.554e+02  7.841e+00  45.319  < 2e-16 ***
## LOT_SIZE    -2.818e+00  9.736e-01  -2.895  0.00381 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 300500 on 4060 degrees of freedom
##   (630 observations deleted due to missingness)
## Multiple R-squared:  0.5754, Adjusted R-squared:  0.575
## F-statistic:  1376 on 4 and 4060 DF,  p-value: < 2.2e-16
```

```
#create bar plot of R values
```

Accordingly, we can see that our R-squared has gone up, as to be expected for adding more parameters to our linear model. What is encouraging is to see the adjusted R-squared is also the highest of all our other models. This is explained by the fact that despite taking into account all four factors, they each are still significant. We do note that the LOT_SIZE predictor has lost some significance, but not enough to not be considered significant under that 0.01 significance level.

SQFT and BATHS remain both positive parameter values, in particular, SQFT has remained pretty close to its initial value from its independent model, while BATHS has somewhat decreased, but is still positive. However, quite surprisingly, we see that BEDS and LOT_SIZE have become negative parameters rather than the positive parameters they were on their own. It is interesting to see such a relationship between our parameters when considered all together.