Assignment 1

Andrew Abisha Hu, Elaine Su, Eric Xiong, Tianye Wang

9/5/2018

1. Read the data in R and summarize all variables. (provide R codes, 0.5%)

```
library(car)
```

```
## Loading required package: carData
```

```
setwd("/Users/andrewhu/Documents/GitHub/Machine-Learning/Class Practice/Assignment 1-Linear Regression I

prop <- read.csv("Property.csv")
summary(prop)
```

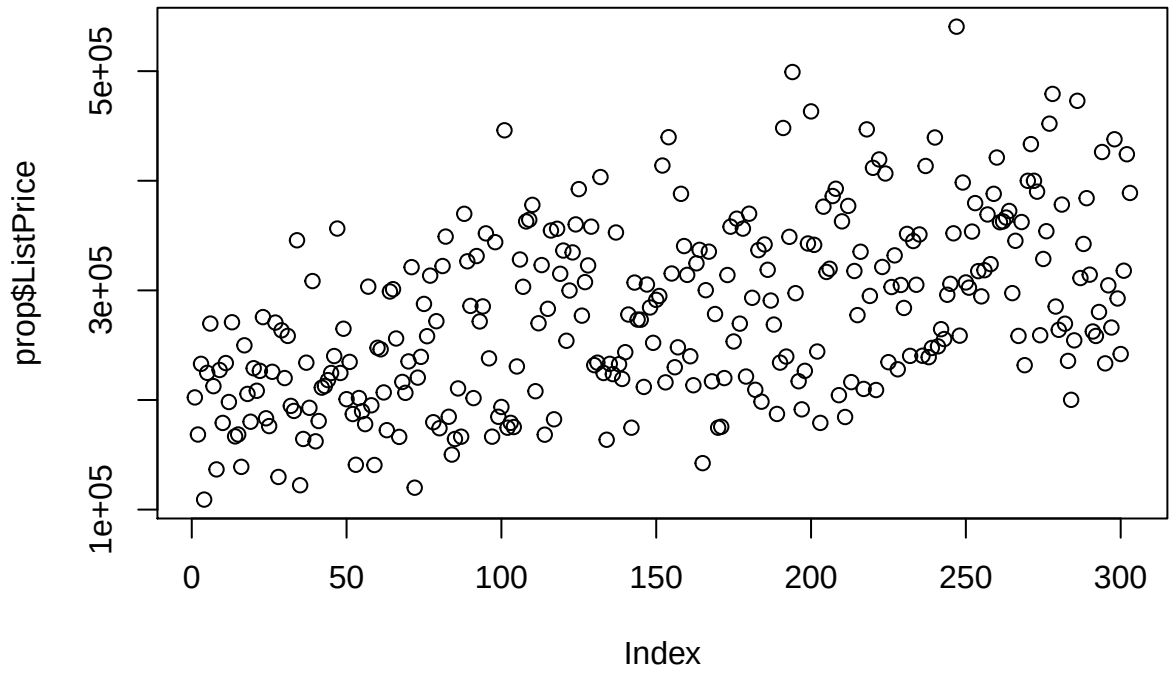
```
##     Listing        ListPrice         Sold              Age
##  Min.   :10001   Min.   :109140   Mode :logical   Min.   : 2.00
##  1st Qu.:10100   1st Qu.:217040   FALSE:213       1st Qu.: 4.00
##  Median :10932   Median :270912   TRUE :90        Median :11.00
##  Mean   :10756   Mean   :279189                   Mean   :13.21
##  3rd Qu.:11108   3rd Qu.:336834                   3rd Qu.:25.00
##  Max.   :11203   Max.   :540568                   Max.   :41.00
##
##    SquareFeet        Beds           Baths           County
##  Min.   :1132   Min.   :2.000   Min.   :1.000   Bergen :49
##  1st Qu.:2192   1st Qu.:3.000   1st Qu.:1.000   Essex  :87
##  Median :2684   Median :4.000   Median :2.000   Hudson :17
##  Mean   :2779   Mean   :3.818   Mean   :2.096   Mercer :40
##  3rd Qu.:3366   3rd Qu.:5.000   3rd Qu.:3.000   Morris :64
##  Max.   :5493   Max.   :6.000   Max.   :4.000   Passaic:26
##                                                 Sussex :20
##          Style     Construction            Garage            Roof
##  2 Story    : 72   Brick:100    1 Car Attached:48   Composition:63
##  Ranch      :123   Frame:106    1 Car Detached:87   Shaker     :63
##  Split Foyer: 53   Stone: 97    2 Car Attached:48   Shingle    :63
##  Split Level: 55                2 Car Detached:42   Slate      :52
##                                 3 Car Attached:39   Tile       :62
##                                 Carport       :39
##
```

2. Which of the following variables are qualitative variables: Listing, ListPrice, Sold, Beds, County? (0.5%)

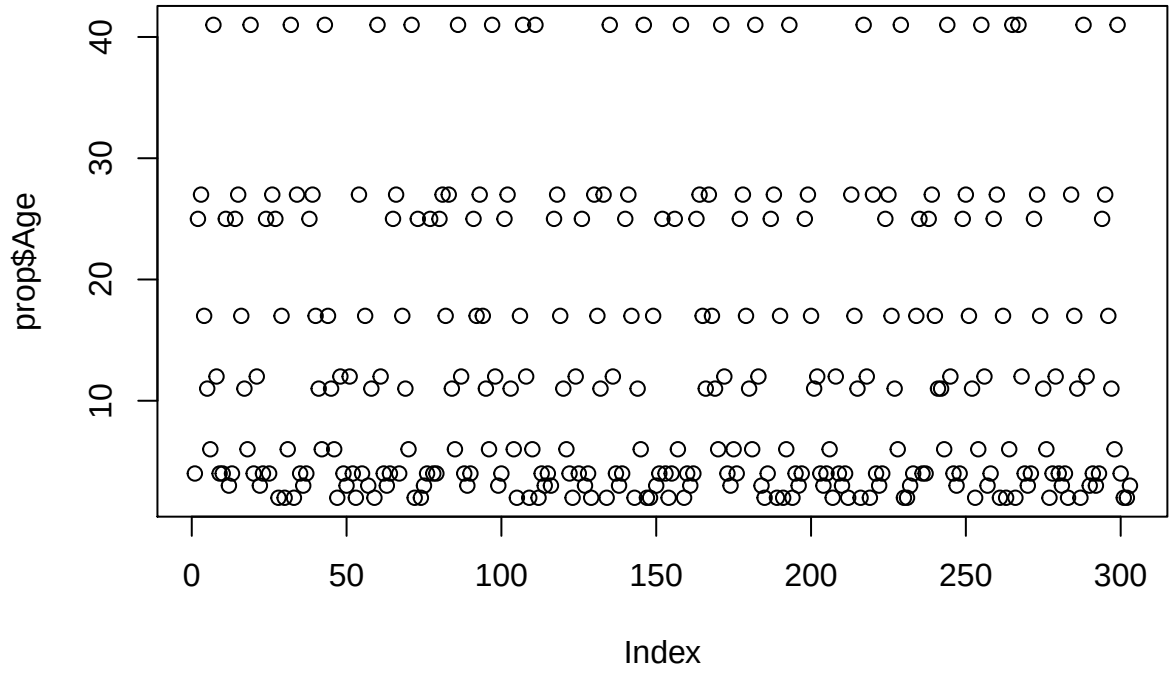
`Listing, Sold, County`

3. Create scatter plots for ListPrice, Age, SquareFeet, Beds, and Baths. (provide R codes, 0.5%)

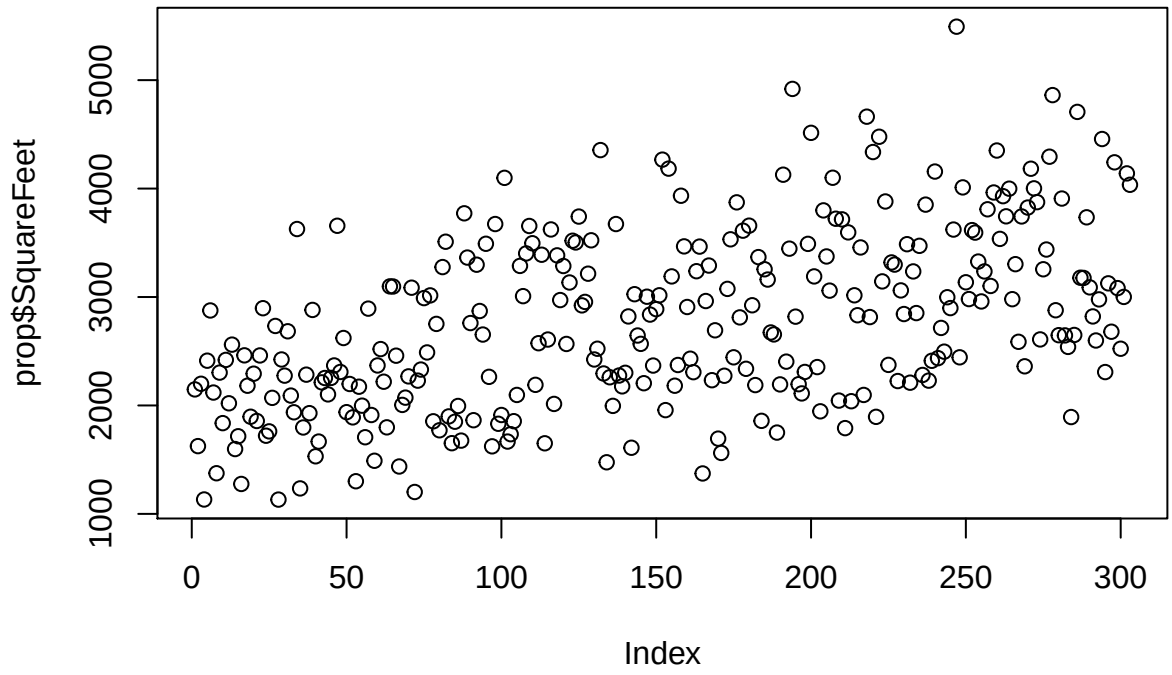
```
plot(prop$ListPrice)
```



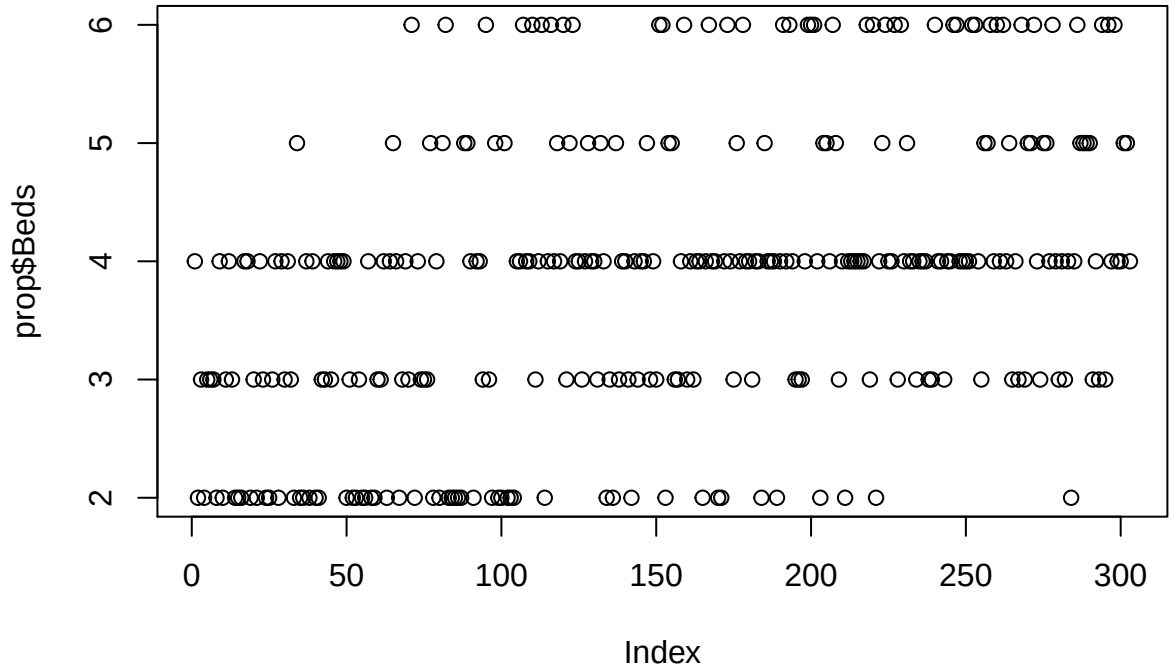

```
plot(prop$Age)
```



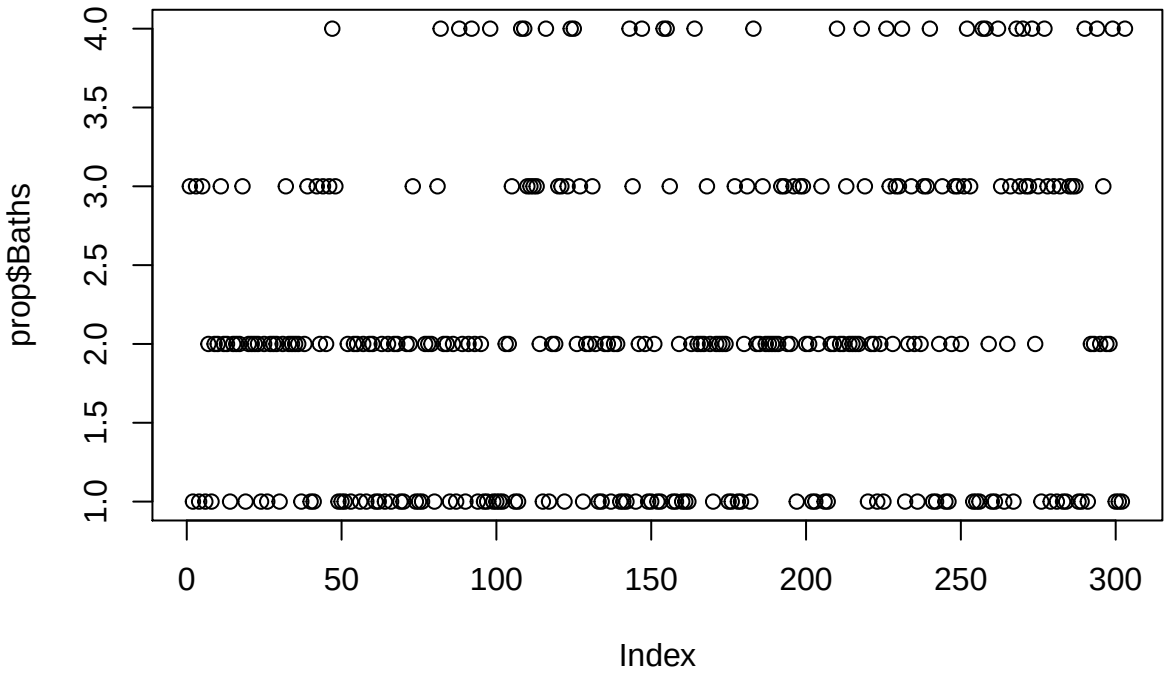

```
plot(prop$SquareFeet)
```



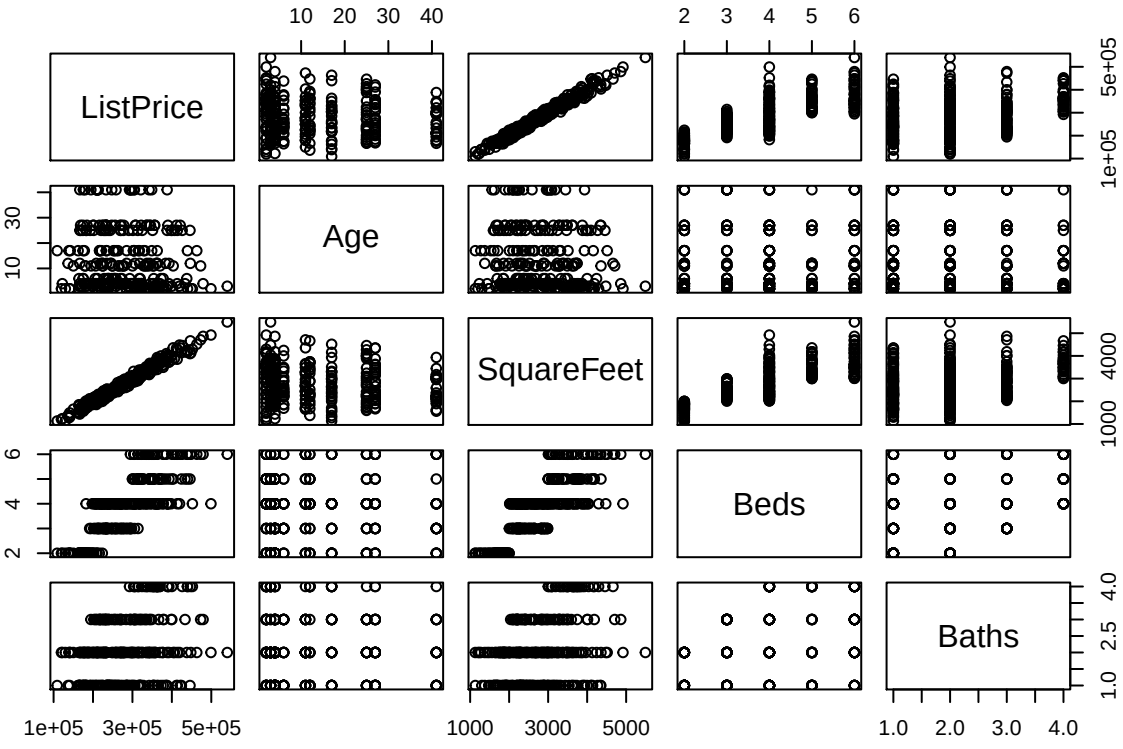

```
plot(prop$Beds)
```




```
plot(prop$Baths)
```




```
pairs(~ListPrice +Age + SquareFeet +Beds + Baths, prop)
```



4. According to the scatter plots in question (3), which variables may be linear correlated with ListPrice? (0.5%)

SquareFeet and ListPrice are most likely to be linear correlated. Beds may also have some kind of linear correlation with List Price but it isn't that strong.

5. Compute the correlation coefficients for the variables listed in question (3). Round the coefficients to the second decimal places. (provide R codes, 0.5%)

```
round(cor(prop$ListPrice,prop$Age),digits = 2)
```

```
## [1] -0.09
```

```
round(cor(prop$ListPrice,prop$SquareFeet),digits = 2)
```

```
## [1] 0.99
```

```
round(cor(prop$ListPrice,prop$Beds),digits = 2)
```

```
## [1] 0.79
```

```
round(cor(prop$ListPrice,prop$Baths),digits = 2)
```

```
## [1] 0.33
```

```
round(cor(prop$Age,prop$SquareFeet),digits = 2)
```

```
## [1] -0.08
```

```
round(cor(prop$Age,prop$Beds),digits = 2)
```

```
## [1] 0
```

```
round(cor(prop$SquareFeet,prop$Beds),digits = 2)
```

```
## [1] 0.8
```

```
round(cor(prop$SquareFeet,prop$Baths),digits = 2)
```

```
## [1] 0.34
```

```
round(cor(prop$Beds,prop$Baths),digits = 2)
```

```
## [1] 0.31
```

6. Comment on the correlation coefficients in question (5). Do the coefficients support your answer to question (4)? (0.5%)

```
Yes, SquareFeet has the highest correlation with ListPrice (0.99). Beds also has a
pretty high correlation with Listprice(0.79).
```

7. Do the scatter plots in question (4) and the correlation coefficients in question (6) make sense to you? Why? (0.5%)

```
Yes, it makes sense, because SquareFeet and ListPrice have the highest correlation coefficient and the
variables. Similarly, Beds and ListPrice have lower correlation coefficient and the plot shows that the
```

8. Regress ListPrice on each of Age, SquareFeet, Beds, and Baths. (provide R codes, 0.5%)

```
fit1 = lm(prop$ListPrice~prop$Age)
fit2 = lm(prop$ListPrice~prop$SquareFeet)
fit3 = lm(prop$ListPrice~prop$Beds)
fit4 = lm(prop$ListPrice~prop$Baths)
```

9. According to the regression results, which variables are significantly associated with ListPrice at the 95% confidence level?

```
SquareFeet, Beds, Baths
```

10. What is the R2 value of each model? (0.5%)

```
fit1 : 0.007461
fit2 : 0.9756
```

```
fit3 : 0.6226
fit4 : 0.1097
```

11. Compute the square of correlation coefficients among ListPrice, Age, SquareFeet, Beds, and Baths. (provide R codes, 0.3%)

```
cor(prop$ListPrice,prop$Age)^2
```

```
## [1] 0.007460781
```

```
cor(prop$ListPrice,prop$SquareFeet)^2
```

```
## [1] 0.9755554
```

```
cor(prop$ListPrice,prop$Beds)^2
```

```
## [1] 0.6225838
```

```
cor(prop$ListPrice,prop$Baths)^2
```

```
## [1] 0.1096655
```

12. Are the R2 values in question (10) different from the square of correlation coefficients between ListPrice and each independent variable in question (11)? (0.2%)

```
No. They are the same.
```

13. What is the estimated regression equation regarding ListPrice and SquareFeet? (0.5%)

```
y= 6657.4873+ 98.0794*SquareFeet
```

14. What does the slope of the equation in question (13) tell you? (0.5%)

```
A one-unit increase in SquareFeet leads to a 98.0794 unit increase in List Price.
```

15. Compute the 95% confidence interval for the slope in question (14). What is the upper limit of the confidence interval? (provide R codes, 0.5%)

```
confint(fit2, level=0.95)
```

```
##                       2.5 %      97.5 %
## (Intercept)     1559.13444 11755.84021
## prop$SquareFeet   96.31841    99.84039
```

16. Regress ListPrice on Age, SquareFeet, Beds, and Baths. (provide R codes, 0.5%)

```
fit5= lm(ListPrice~Age+SquareFeet+Beds+Baths, prop)
```

17. What is the p-value of the significance test for the overall model? (0.5%)

```
P-value < 2.2e-16
```

18. Given the 95% confidence interval, what is your conclusion on the overall model significance? (0.5%)

```
The p-value < 2.2e-16, and the F-statistic is 2978, which is very large. We can thus reject the null hyp
significantly to the overall model. Therefore, our model predicts better than simply using average valu
```

19. What is the estimated regression equation? (0.5%)

```
y= 7420.042-33.971*Age+ 97.859*SquareFeet+248.693*Beds-310.484*Baths
```

20. What fraction of the variation of ListPrice is explained by using the model? (0.5%)

```
R-Square=0.9756, indicating that the model can explain 97.56% of the variance of ListPrice.
```

21. According to the regression results, which variables are significantly associated with ListPrice at the 95% confidence level? (0.5%)

```
SquareFeet
```

22. Is your answer to question (21) different from your answer to question (9)? Why? (0.5%)

```
Yes, this may be due to multicollinearity. Independent variables SquareFeet and beds are highly correla
```

23. How do you interpret the coefficient corresponding to SquareFeet? (0.5%)

```
A one-unit increase in SquareFeet leads to a 97.859 unit increase in ListPrice.
```

24. What is the predicted ListPrice of a two-year-old, 1500 square-foot house with two bedrooms and 1 bathroom? (0.5%)

```
154327.3
```

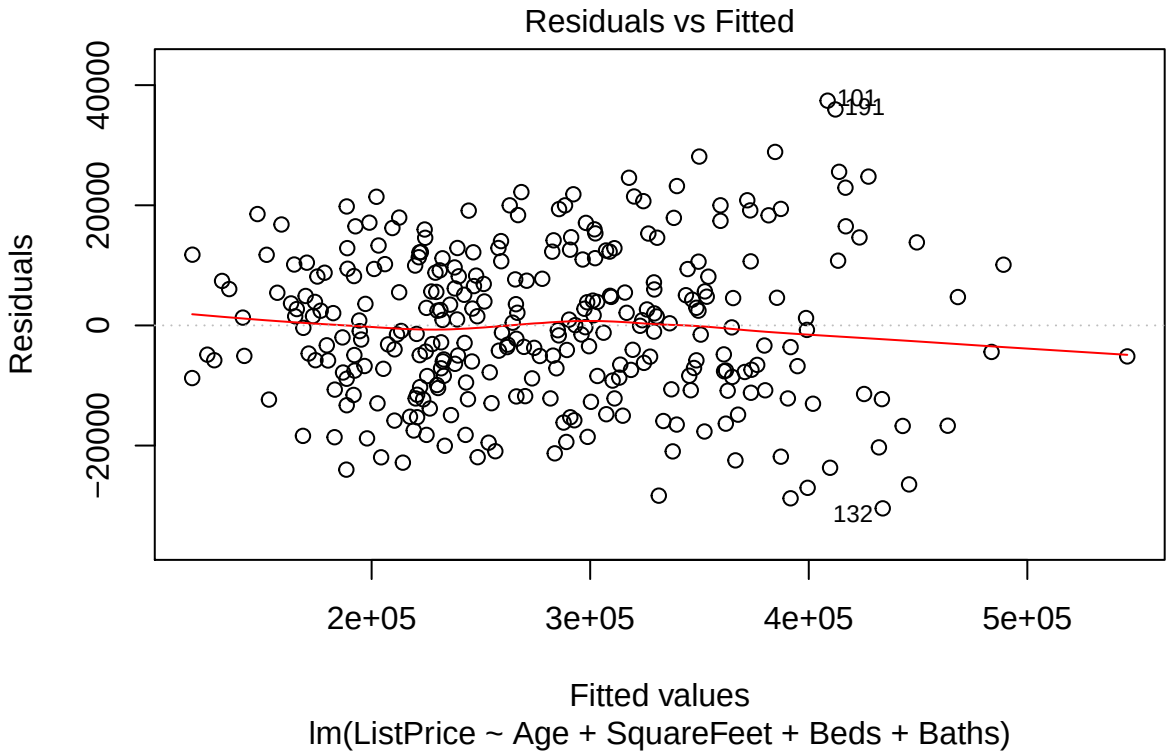
25. What is the 99% prediction interval associated with your prediction in question (24)? (Provide R codes, 0.5%)

```
predict(fit5, data.frame(Age=2, SquareFeet=1500, Beds=2,Baths=1), interval= "prediction", level=.99)
```

```
##        fit    lwr      upr
## 1 154327.3 121086 187568.7
```

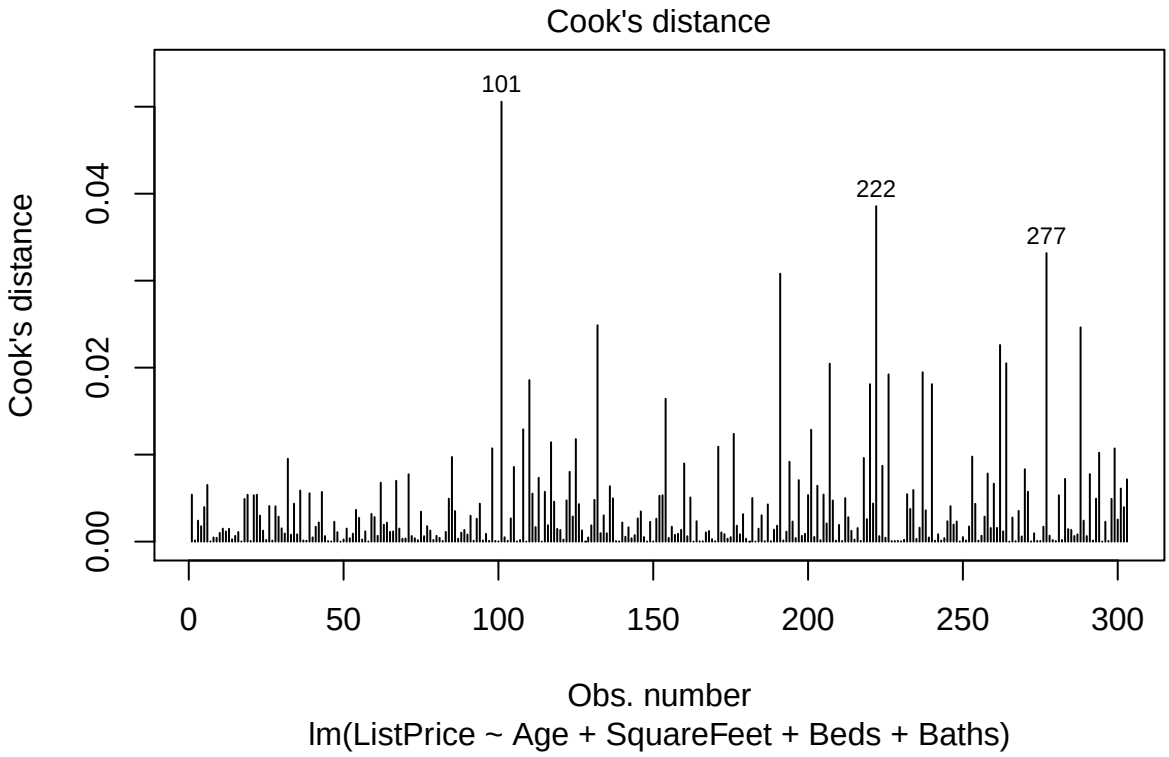
26. Create the scatter plot of the residuals and the fitted values. Do you notice any discernible pattern of the scatter plot? (provide R codes, 0.5%)

```
plot(fit5,1) #no
```



27. Create the plot that shows each observation's Cook's distance. Are there influential observations based on the Cook's distance? (provide codes, 0.5%)

```
plot(fit5,4)
```




```
The consensus is that when Cook's distance>1, then the observation is considered an influential observa
may still want to look more into observations 101, 222, and 277 as they stick out significantly from th
```

28. What is the VIF value for SquareFeet? (0.5%)

```
car::vif(fit5)
```

```
##        Age SquareFeet       Beds      Baths
##   1.018418   2.869123   2.798437   1.135860
```

```
VIF value for SquareFeet = 2.869123
```

29. Comment on the concern of multicollinearity based on the VIF values of all independent variables. (0.5%)

```
VIF > 10 signals serious multicollinearity requiring correction. None of these variables have high VIF
```

30. Larger houses tend to have more bedrooms and/or bathrooms. Given this fact, comment on using VIF to diagnose multicollinearity. (0.5%)

```
VIF can be used to detect serious multicollinearity problems (when VIF>10). However, when
VIF<10, it indicates that multicollinearity may exist but it isn't serious enough to cause severe proble
multicollinearity.
```