Final Project Report Tips and Outline

Rebecca LI (1456424)

December, 2018

Data Preparation

```
# Feel free to insert R code chunks where needed
house_csv <- read.csv( "house.csv")
View(house_csv)
str(house_csv)
```

```
## 'data.frame':    150 obs. of  10 variables:
##  $ PID              : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ home.size        : int  600 1050 1800 922 1950 1783 1008 1840 3700 1092 ...
##  $ lot.size         : num  0.5 0.43 0.68 0.3 0.75 0.22 0.5 1.16 1.1 0.26 ...
##  $ rooms            : int  3 5 7 5 8 8 6 8 10 6 ...
##  $ bathrooms        : num  1 1.5 1.5 1 2.5 1.5 1 2 3 1 ...
##  $ utilities        : int  252 216 207 249 217 208 243 242 256 222 ...
##  $ year.built       : int  1960 1961 1958 1968 1998 1998 2001 1992 1996 1999 ...
##  $ overall.condition: int  3 6 6 3 5 6 5 5 5 5 ...
##  $ brick            : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 1 1 1 ...
##  $ price            : int  102000 146300 182000 110500 171900 154000 147000 195900 183500 156500 ...
```

```
##Randomly select subsets for training and testing
# Randomly select a subset of the 150 houses. This random sample should contain 130 houses.
set.seed(1)
training.houses <- sample_n(house_csv,130)
# Create another data frame of the rest of the houses (n = 20)
testing.houses <- subset(house_csv, !(PID %in% training.houses$PID))

write.csv(training.houses, file = "trainingset.csv")
write.csv(testing.houses, file = "testingset.csv")
```

Make sure the training and testing sets are exclusive to each other !

Main Task Overview and Data Preparation

Problem 1

1. Is there a significant difference in sales prices between brick and non-brick houses?

```
# inspect what the factor levels of this variable are.
class(training.houses$brick)
```

```
## [1] "factor"
```

```
# Compute descriptive statistics
( bricks <- group_by(training.houses, brick))
```

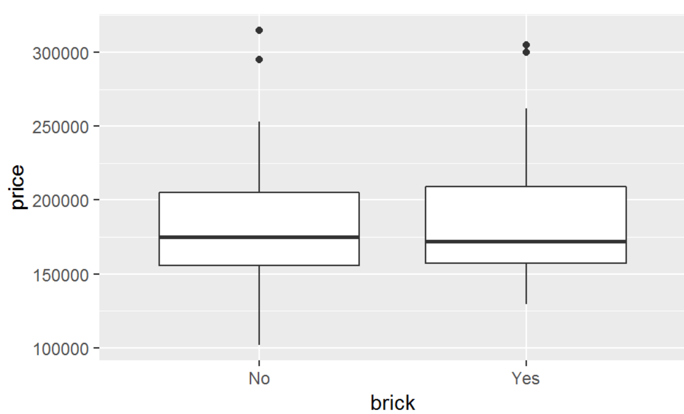
```
## # A tibble: 130 x 10
## # Groups:   brick [2]
##      PID home.size lot.size rooms bathrooms utilities year.built
##   * <int>     <int>    <dbl> <int>     <dbl>     <int>      <int>
##  1    40      1839     2.6      7       1.5       259       2004
##  2    56      1908     0.46     7       2         244       2007
##  3    85       864     0.32     4       1         265       2005
##  4   134      2473     1.25     9       2.5       223       2007
##  5    30      2000     0.5      8       2         264       1971
##  6   131      2000     0.65     7       1         263       1952
##  7   137      2300     0.91     8       2.5       252       2000
##  8    95      1980     0.7      8       2.5       259       1978
##  9    90      2400     2        7       2         202       1999
## 10     9      3700     1.1     10       3         256       1996
## # ... with 120 more rows, and 3 more variables: overall.condition <int>,
## #   brick <fct>, price <int>
```

Also use a side-by-side boxplot to present the prices of the two types of houses.

```
kable((scores_table <- summarize(bricks,
                       group.size   = length(price),
                       mean.bricks = round( mean(price), digits = 2),
                       sd.bricks   = round( sd(price), digits = 2)
                            )))
```

brick	group.size	mean.bricks	sd.bricks
No	85	179873.0	39293.21
Yes	45	185127.8	41500.94

```
# draw boxplot using geom_boxplot()
(bp <- ggplot(bricks, aes(x = brick, y = price))+geom_boxplot())
```



Test the equal variance assumption

```
( result_vartest <- var.test(price ~ brick, data = training.houses)  )
```

```
##
##  F test to compare two variances
##
## data:  price by brick
## F = 0.89644, num df = 84, denom df = 44, p-value = 0.6574
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.5209624 1.4778062
## sample estimates:
## ratio of variances
##          0.8964356
```

```
result_vartest$p.value # What is the p-value for the variance test?
```

```
## [1] 0.657426
```

Interpret the results and/or provide disucssions:

We obtained p-value (i.e., p = 0.657426) greater than 0.05; this result supports the equal-variance assumption. So we can assume that the two variances are equal/homogeneous.

On the other hand, we can tell from the *table and boxplot that the brick and non-brick houses shares similar mean and std

Perform t-test.

use the `t.test()` function. Also consider,need to include the `var.equal = TRUE` arugment based on the result of part (e).

```
# Edit me
(result_t_test <- t.test(price ~ brick, data = training.houses, var.equal=TRUE) )
```

```
##
##  Two Sample t-test
##
## data:  price by brick
## t = -0.71142, df = 128, p-value = 0.4781
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -19869.93   9360.40
## sample estimates:
##  mean in group No mean in group Yes
##          179873.0          185127.8
```

```
result_t_test$p.value
```

```
## [1] 0.4781214
```

```
result_t_test$estimate
```

```
##  mean in group No mean in group Yes
##          179873.0          185127.8
```

Our study found that on average t tatus is 179873 for houses without bricks, and 185127.8 for houses with bricks, thus non-brick houses have higher prices than brick houses

t-statistic = -0.71, p = 0.4781214, The 95 % confidence interval (CI) is (-19869.93, 9360.40)

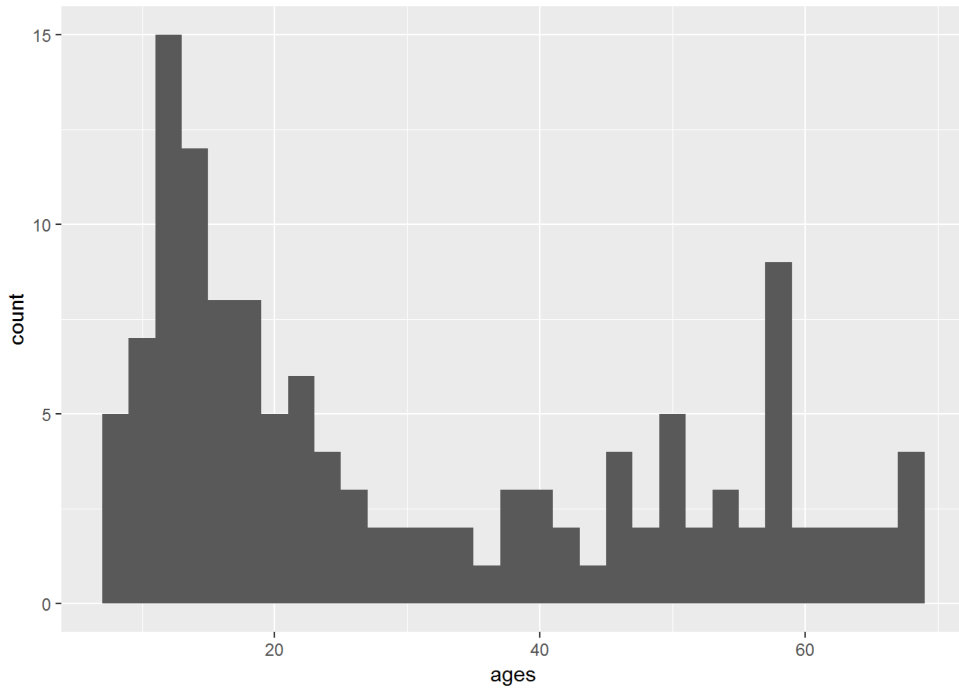
Problem 2

2. Calculate the age of the houses in 2018. In addition, present a histogram of the house age variable

```
( ages <- 2018 - training.houses$year.built)
```

```
##   [1] 14 11 13 11 47 66 18 40 19 22 47 48 66 20 59 34 16 63 17 12 59 47 14
##  [24]  8 40 14 57 14 51 16 22 68 10 48 13  8 15 20 13 12 10 37 52 13 20 14
##  [47] 60 18 51 33 59 13 12 50 17 19 58 59 13 13 15 19  8 11 14 64 18 16 64
##  [70] 54 59 39 25 47 13 26 24 13 25 15 41 52 16 54 25 28 22 26 58 14 33 50
##  [93] 55 26 15 29 12 56 51 30 43 30 68 13 38 68 20 18 12 22 68 23 38 14 23
## [116] 58 59 10 62 44  9 43 16 20 60 11 17 18  8 34
```

```
ggplot(training.houses, aes(x = ages)) +
  geom_histogram(binwidth = 2)
```



Regression Modeling and Interpretations

use multiple regression to explain house sales price

Problem 3

Fit a regression model to explain house sales price

```
##Randomly select subsets for training and testing
( house_lm <- lm(price ~ home.size + lot.size + rooms+ bathrooms+ utilities+ year.built + overall.condition + brick , data =
training.houses) )
```

```
##
## Call:
## lm(formula = price ~ home.size + lot.size + rooms + bathrooms +
##     utilities + year.built + overall.condition + brick, data = training.houses)
##
## Coefficients:
##       (Intercept)          home.size           lot.size
##        -297831.37              24.11            6719.78
##             rooms          bathrooms          utilities
##           -439.00           17728.63             -40.40
##        year.built  overall.condition           brickYes
##            183.49            8912.85            1787.28
```

Problem 4

Interpret the regression results. You should follow the 5 steps covered in class to analyze the results. The 5th step - prediction - is conducted in the next question

(Step 1) Interpret the overall model

Interpret the results of the F-test and its associated p-value for the overall significance of the regression model. Explain the results.

```
# Edit me
(lm.results<- summary(house_lm))
```

```
##
## Call:
## lm(formula = price ~ home.size + lot.size + rooms + bathrooms +
##     utilities + year.built + overall.condition + brick, data = training.houses)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -64521 -12715  -1654  11365  85950
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -2.978e+05  2.053e+05  -1.451 0.149432
## home.size          2.411e+01  7.285e+00   3.309 0.001233 **
## lot.size           6.720e+03  1.317e+03   5.104 1.25e-06 ***
## rooms             -4.390e+02  2.303e+03  -0.191 0.849170
## bathrooms          1.773e+04  4.650e+03   3.812 0.000218 ***
## utilities         -4.040e+01  8.258e+01  -0.489 0.625583
## year.built         1.835e+02  1.051e+02   1.746 0.083436 .
## overall.condition  8.913e+03  1.906e+03   4.677 7.63e-06 ***
## brickYes           1.787e+03  4.141e+03   0.432 0.666806
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21870 on 121 degrees of freedom
## Multiple R-squared:  0.7195, Adjusted R-squared:  0.7009
## F-statistic: 38.79 on 8 and 121 DF,  p-value: < 2.2e-16
```

It can be seen that p-value of the F-statistic is 2.2e-16 (< 0.05), which is highly significant. We say the regression model overall is significant. This means that, at least, one of the explanatory variables is significantly related to the response variable.

(Step 2) Interpret the beta coefficients

Test each of the individual regression coefficients (beta 1 and beta 2). To do this, first extract the coefficients table from the test ouput. Then, answer the questions afterward.

1. Extract the coefficients table from the test ouput

```
# Edit me
( lm.coef <- lm.results$coefficients )  # extract the coefficients table
```

```
##                        Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)       -297831.36836 205292.25697 -1.4507677 1.494321e-01
## home.size              24.10665      7.28457  3.3092753 1.232612e-03
## lot.size             6719.78191   1316.53318  5.1041493 1.249755e-06
## rooms                -438.99866   2303.41975 -0.1905856 8.491697e-01
## bathrooms           17728.62865   4650.29679  3.8123650 2.179868e-04
## utilities             -40.39722     82.57755 -0.4892034 6.255833e-01
## year.built            183.49063    105.12201  1.7455015 8.343605e-02
## overall.condition    8912.84841   1905.50212  4.6774277 7.628573e-06
## brickYes             1787.27679   4141.12361  0.4315922 6.668058e-01
```

```
# Display the coefficients table in a nice `kable` table. Meanwhile, round all the numbers to three decimal places in the ta
ble.
(lm.coef.table <-kable(lm.coef, digits = c(3, 3, 3, 3)))   # display the featuers
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-297831.368	205292.257	-1.451	0.149
home.size	24.107	7.285	3.309	0.001
lot.size	6719.782	1316.533	5.104	0.000
rooms	-438.999	2303.420	-0.191	0.849
bathrooms	17728.629	4650.297	3.812	0.000
utilities	-40.397	82.578	-0.489	0.626
year.built	183.491	105.122	1.746	0.083
overall.condition	8912.848	1905.502	4.677	0.000
brickYes	1787.277	4141.124	0.432	0.667

2. Extract the significant variables:

```
# find out those variables P values lower than 0.05
( sig_variables <- which ( lm.coef[,4] <0.05) )
```

```
##        home.size         lot.size        bathrooms overall.condition
##                2                3                5                 8
```

```
lm.coef[sig_variables,4]   # display the p values of  the significant variables
```

```
##        home.size         lot.size        bathrooms overall.condition
##     1.232612e-03     1.249755e-06     2.179868e-04      7.628573e-06
```

i. Estimated coefficents of the regression model

As can be seen the lm.coef.table, we have the estimate value and p-values of the F-statistic for all variables.

First, let's check out the **p-values**. The significant variables are: home.size, lot.size, bathrooms, overall.condition because their p values are less than 0.05 (1.2e-03, 1.2e-06, 2.2e-04, 7.6e-06 ,respectively). This tells us that only for those variables, the estimated beta coefficients are statistically significantly different from 0. Thus, only the variables ** home.size, lot.size, bathrooms, overall.condition ** are significant predictors of the response variable price.

Second, let's interpret the **beta coefficients** of the significant predictors. See the coefficients of home.size, lot.size, bathrooms, overall.condition in the model. They are all positive, and are 24.11, 6719.78, 17728.63, 8912.85 respectively.

Here is how we interpret their coefficients:

```
1) Assuming that we hold all else constant, as the home.size is increased by one square feet, the price increases by 24.11
2) Assuming that we hold all else constant, as the lot.size is increased by one acre, the price increases by 6719.78
3) Assuming that we hold all else constant, as the bathrooms is increased by one room, the price increases by 17729
4) Assuming that we hold all else constant, as the overall.condition is increased by one score, the price increases by 8912.
8484124
```

ii. Explanatory variables should be removed from the model

Extract the insignificant variables:

```
( insig_variables <-which ( lm.coef[,4] >0.05) [-1] )
```

```
##      rooms  utilities year.built   brickYes
##          4          6          7          9
```

```
lm.coef[insig_variables,4]   # display the p values of  the insignificant variables
```

```
##      rooms  utilities year.built   brickYes
## 0.84916966 0.62558333 0.08343605 0.66680583
```

The p values of the variables of **rooms, utilities, year.built, brick** are 0.849, 0.626, 0.083, 0.667 ,respectively. Since they are larger than 0.05, they are **insignificant coefficients, need to be removed**.

The **new linear regression** should be

```
( house_lm.new <- lm(price ~ home.size + lot.size + bathrooms + overall.condition , data = training.houses) )
```

```
##
## Call:
## lm(formula = price ~ home.size + lot.size + bathrooms + overall.condition,
##     data = training.houses)
##
## Coefficients:
##       (Intercept)          home.size           lot.size
##          55891.95              22.87            6967.24
##         bathrooms  overall.condition
##          18280.63            8828.12
```

```
(lm.results.new<- summary(house_lm.new))
```

```
##
## Call:
## lm(formula = price ~ home.size + lot.size + bathrooms + overall.condition,
##     data = training.houses)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -63673 -11527   -386  10888  89573
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       55891.95    8982.25   6.222 6.77e-09 ***
## home.size            22.87       5.28   4.332 3.00e-05 ***
## lot.size           6967.24    1286.29   5.417 3.00e-07 ***
## bathrooms         18280.63    4597.92   3.976 0.000118 ***
## overall.condition  8828.12    1883.19   4.688 7.11e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21800 on 125 degrees of freedom
## Multiple R-squared:  0.712,  Adjusted R-squared:  0.7028
## F-statistic: 77.27 on 4 and 125 DF,  p-value: < 2.2e-16
```

```
(lm.coef.new <- lm.results.new$coefficients )  # extract the coefficients table
```

```
##                      Estimate  Std. Error  t value     Pr(>|t|)
## (Intercept)       55891.95004 8982.246883 6.222491 6.774812e-09
## home.size            22.87432    5.279804 4.332419 2.998904e-05
## lot.size           6967.24216 1286.289612 5.416542 2.997520e-07
## bathrooms         18280.63447 4597.917670 3.975851 1.178945e-04
## overall.condition  8828.12222 1883.194868 4.687843 7.105519e-06
```

(Step 3) Report the regression equation

Determine the regression equation with the explanatory variable(s) identified in the previous step.

$$EstimatedPrice = 5.589195 \times 10^4 + 22.87 * home.size + 6967.24 * lot.size + 1.828063 \times 10^4 * bathrooms + 8828.12 * overall.cond$$

(Step 4) Assess the model

The value of R2 will always be positive and will range from zero to one. The R2 value close to 1 indicates that the model explains a large portion of the variance in the response variable.

To report the results of multiple regression, a more common practice is to report the “adjusted R2”, which is essentially the same as R2 but it also takes into account the number of predictors in the model.

```
# Edit me
(lm.results.new$r.squared)
```

```
## [1] 0.7120368
```

```
(lm.results.new$adj.r.squared)
```

```
## [1] 0.702822
```

The interpretation of R-square and adjusted R-square value:

```
The R2 value is 0.7120368 in our regression model. This means that our model can explain 71.2% of the variation of worker-ho
urs.

With four predictor variables, the adjusted R2 = 0.702822, meaning that 70.28% of the variance of overhead can be predicted
byhome.size, lot.size, bathrooms, overall.condition".
```

Prediction and Validation

Problem 5

5. Based on your regression model, predict which 10 properties in the testing set ranked highest in sales price. Clearly state your answer in the report.

To get the prediction for the prices of testing samples:

```
( predictions <-predict( house_lm.new, data.frame(testing.houses)) )
```

```
##       11       17       19       23       35       45       52       63
## 175542.1 219923.6 152879.3 276811.9 225753.4 150140.0 320626.5 179530.6
##       76       92       93      104      106      112      115      121
## 147504.2 193875.2 148061.6 200737.5 213107.1 264955.6 188670.0 204116.1
##      124      126      135      144
## 186996.8 195852.2 127003.1 161635.8
```

To get a 95% confidence interval for the prices:

```
( predictions_CI <-predict( house_lm.new, data.frame(testing.houses) , interval = "confidence") )
```

```
##          fit      lwr      upr
## 11  175542.1 169465.6 181618.6
## 17  219923.6 208879.8 230967.5
## 19  152879.3 143469.6 162289.0
## 23  276811.9 247160.3 306463.6
## 35  225753.4 217354.9 234151.9
## 45  150140.0 143385.3 156894.7
## 52  320626.5 275229.0 366024.0
## 63  179530.6 172345.1 186716.1
## 76  147504.2 140741.9 154266.5
## 92  193875.2 187071.3 200679.1
## 93  148061.6 141328.7 154794.4
## 104 200737.5 195362.0 206112.9
## 106 213107.1 200003.3 226210.8
## 112 264955.6 251902.9 278008.4
## 115 188670.0 180501.4 196838.6
## 121 204116.1 198366.8 209865.5
## 124 186996.8 179314.9 194678.6
## 126 195852.2 188403.8 203300.7
## 135 127003.1 119036.9 134969.3
## 144 161635.8 156108.4 167163.3
```

Sort the predicted prices values

```
(properties_top10<-sort(predictions,decreasing = TRUE)[1:10] )
```

```
##       52       23      112       35       17      106      121      104
## 320626.5 276811.9 264955.6 225753.4 219923.6 213107.1 204116.1 200737.5
##      126       92
## 195852.2 193875.2
```

```
names(properties_top10)
```

```
##  [1] "52"  "23"  "112" "35"  "17"  "106" "121" "104" "126" "92"
```

The 10 properties in the testing set ranked highest in sales price are (from highest prices to 10 highest price):

```
52, 23, 112, 35, 17, 106, 121, 104, 126, 92
```

Problem 6

6. Validate your prediction using the real price values in the testing set. Is your prediction correct (and/or to what extent)? Provide a discussion/conclusion of your analysis.

Ground Truth value:

```
(ground_truth <- testing.houses$price)
```

```
##  [1] 152000 199000 153500 332000 199900 166000 297000 185000 149000 165000
## [11] 159000 202000 171900 265000 210000 212000 207000 195000 137000 189000
```

Validation:

To evaluate the predicted and the ground truth values(i.e., observed), we used root of **mean square error and R square.**

```
# Root of Mean square error
RMSE  <- function(actual, preds)
{
  sqrt(mean((preds -actual)^2))
}

# R square
R_square <- function(actual, preds)
{
  rss <- sum((preds - actual) ^ 2)          ## residual sum of squares
  tss <- sum((actual - mean(actual)) ^ 2)   ## total sum of squares
  rsq <- 1 - rss/tss
}

(rmse_house <- RMSE(ground_truth,predictions))
```

```
## [1] 22325.1
```

```
(rsq_house <- R_square(ground_truth,predictions))
```

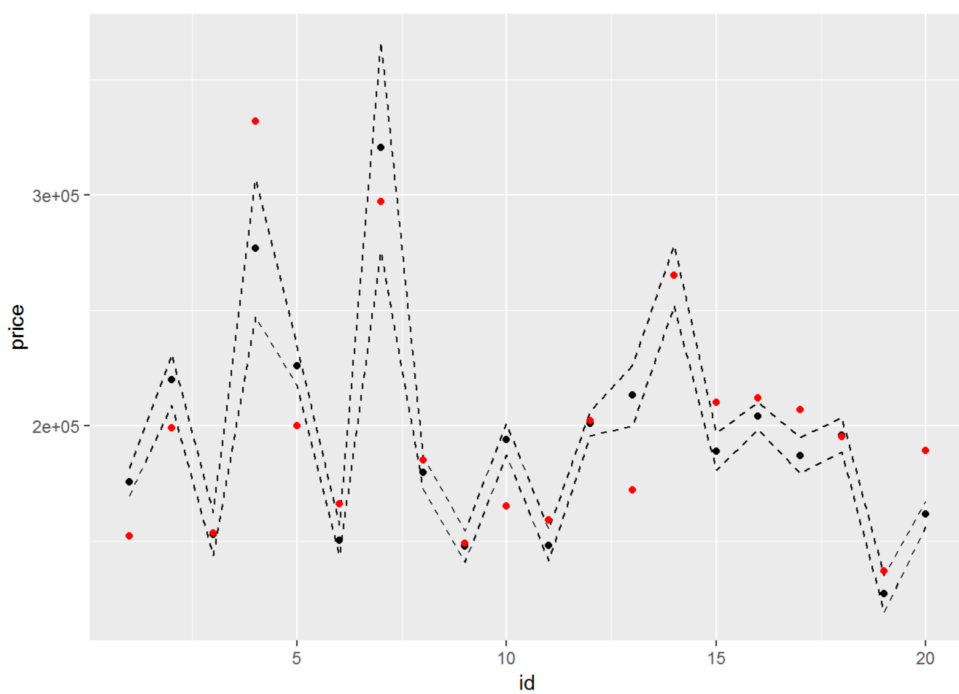
```
## [1] 0.7900136
```

```
RMSER = 22325.0952, and R square = 0.7900136.
```

This implies that 79% of the variability of the dependent variable has been accounted for.

```
fit.data <-data.frame( id = c( 1:20), price = predictions_CI[,1])
lwr.data <-data.frame( id = c( 1:20), price = predictions_CI[,2])
upr.data <-data.frame( id = c( 1:20), price = predictions_CI[,3])
gt.data <-data.frame( id = c( 1:20), price = ground_truth)

ggplot(lwr.data, aes(x = id, y = price))+  geom_line(linetype = "dashed") +
  geom_line(data = upr.data,linetype = "dashed") +
  geom_point(data = fit.data)+
  geom_point(data = gt.data, colour = "red")
```



This plot shows the comparing of the ground truth values(red points), the predicted values(black points) and the confidential interval (area between top and bottom dashed lines).

We can see most of the ground truth value are located in CI. So we had a fair-good prediction.