DSC520_Week8

Krista Knuckey

2024-02-02

Uploading dataset

```
housing_data <- read.csv("/Users/kristaknuckey/Desktop/DSC520/housing.csv")
head(housing_data)
```

```
##   Sale.Date Sale.Price sale_reason sale_instrument sale_warning sitetype
## 1    1/3/06     698000           1               3                    R1
## 2    1/3/06     649990           1               3                    R1
## 3    1/3/06     572500           1               3                    R1
## 4    1/3/06     420000           1               3                    R1
## 5    1/3/06     369900           1               3           15       R1
## 6    1/3/06     184667           1              15        18 51       R1
##            addr_full  zip5 ctyname postalctyn       lon      lat building_grade
## 1  17021 NE 113TH CT 98052 REDMOND    REDMOND -122.1124 47.70139              9
## 2  11927 178TH PL NE 98052 REDMOND    REDMOND -122.1022 47.70731              9
## 3 13315 174TH AVE NE 98052            REDMOND -122.1085 47.71986              8
## 4  3303 178TH AVE NE 98052 REDMOND    REDMOND -122.1037 47.63914              8
## 5  16126 NE 108TH CT 98052 REDMOND    REDMOND -122.1242 47.69748              7
## 6   8101 229TH DR NE 98053            REDMOND -122.0341 47.67545              7
##   square_feet_total_living bedrooms bath_full_count bath_half_count
## 1                     2810        4               2               1
## 2                     2880        4               2               0
## 3                     2770        4               1               1
## 4                     1620        3               1               0
## 5                     1440        3               1               0
## 6                     4160        4               2               1
##   bath_3qtr_count year_built year_renovated current_zoning sq_ft_lot prop_type
## 1               0       2003              0             R4      6635         R
## 2               1       2006              0             R4      5570         R
## 3               1       1987              0             R6      8444         R
## 4               1       1968              0             R4      9600         R
## 5               1       1980              0             R6      7526         R
## 6               1       2005              0          URPSO      7280         R
##   present_use
## 1           2
## 2           2
## 3           2
## 4           2
## 5           2
## 6           2
```

Transformations of dataset

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
housing_data <- select(housing_data, -lon)
housing_data <- select(housing_data, -lat)
housing_data <- select(housing_data, -building_grade)
housing_data <- select(housing_data, -current_zoning)
housing_data <- select(housing_data, -prop_type)
housing_data <- select(housing_data, -present_use)
housing_data <- select(housing_data, -sale_warning)
housing_data <- select(housing_data, -ctyname)
head(housing_data)
```

```
##   Sale.Date Sale.Price sale_reason sale_instrument sitetype          addr_full
## 1    1/3/06     698000           1               3       R1  17021 NE 113TH CT
## 2    1/3/06     649990           1               3       R1  11927 178TH PL NE
## 3    1/3/06     572500           1               3       R1 13315 174TH AVE NE
## 4    1/3/06     420000           1               3       R1  3303 178TH AVE NE
## 5    1/3/06     369900           1               3       R1  16126 NE 108TH CT
## 6    1/3/06     184667           1              15       R1   8101 229TH DR NE
##    zip5 postalctyn square_feet_total_living bedrooms bath_full_count
## 1 98052    REDMOND                     2810        4               2
## 2 98052    REDMOND                     2880        4               2
## 3 98052    REDMOND                     2770        4               1
## 4 98052    REDMOND                     1620        3               1
## 5 98052    REDMOND                     1440        3               1
## 6 98053    REDMOND                     4160        4               2
##   bath_half_count bath_3qtr_count year_built year_renovated sq_ft_lot
## 1               1               0       2003              0      6635
## 2               0               1       2006              0      5570
## 3               1               1       1987              0      8444
## 4               0               1       1968              0      9600
## 5               0               1       1980              0      7526
## 6               1               1       2005              0      7280
```

```
housing_data <- housing_data %>%
  rename(sale_price = Sale.Price)
housing_data <- housing_data %>%
  rename(sale_date = Sale.Date)
housing_data <- housing_data %>%
  rename(site_type = sitetype)
housing_data <- housing_data %>%
  rename(zip_5 = zip5)
housing_data <- housing_data %>%
```

```
  rename(postal_city = postalctyn)
head(housing_data)
```

```
##   sale_date sale_price sale_reason sale_instrument site_type          addr_full
## 1    1/3/06     698000           1               3        R1  17021 NE 113TH CT
## 2    1/3/06     649990           1               3        R1  11927 178TH PL NE
## 3    1/3/06     572500           1               3        R1 13315 174TH AVE NE
## 4    1/3/06     420000           1               3        R1  3303 178TH AVE NE
## 5    1/3/06     369900           1               3        R1  16126 NE 108TH CT
## 6    1/3/06     184667           1              15        R1   8101 229TH DR NE
##   zip_5 postal_city square_feet_total_living bedrooms bath_full_count
## 1 98052     REDMOND                     2810        4               2
## 2 98052     REDMOND                     2880        4               2
## 3 98052     REDMOND                     2770        4               1
## 4 98052     REDMOND                     1620        3               1
## 5 98052     REDMOND                     1440        3               1
## 6 98053     REDMOND                     4160        4               2
##   bath_half_count bath_3qtr_count year_built year_renovated sq_ft_lot
## 1               1               0       2003              0      6635
## 2               0               1       2006              0      5570
## 3               1               1       1987              0      8444
## 4               0               1       1968              0      9600
## 5               0               1       1980              0      7526
## 6               1               1       2005              0      7280
```

```
housing_data <- housing_data %>%
  arrange(sale_price)
head(housing_data)
```

```
##   sale_date sale_price sale_reason sale_instrument site_type
## 1    7/6/10        698           1              26        R1
## 2    7/6/10        698           1              26        R1
## 3  12/29/09        873           1              26        R1
## 4   1/28/10        873           1              26        R1
## 5  12/22/09        998           1              26        R1
## 6   3/20/07       1000           1              15        R1
##                  addr_full zip_5 postal_city square_feet_total_living bedrooms
## 1 19805 NE NOVELTY HILL RD 98053     REDMOND                     5830        4
## 2 19805 NE NOVELTY HILL RD 98053     REDMOND                     1040        3
## 3        8332 196TH AVE NE 98053     REDMOND                     2160        2
## 4        8340 196TH AVE NE 98053     REDMOND                     3430        3
## 5        8226 196TH AVE NE 98053     REDMOND                     1850        3
## 6         22340 NE 65TH PL 98053     REDMOND                     4610        4
##   bath_full_count bath_half_count bath_3qtr_count year_built year_renovated
## 1               4               0               1       1969              0
## 2               1               0               0       1900              0
## 3               1               0               1       1968              0
## 4               1               1               0       1955              0
## 5               1               1               0       1960           1989
## 6               2               0               2       2015              0
##   sq_ft_lot
## 1   1127205
## 2   1127205
## 3    102505
```

```
## 4    105660
## 5    209589
## 6     95989
```

Transformation

In order to make the data set easier to use I made a few transformations that I thought would be most helpful. First, I deleted multiple columns that will not be necessary to complete the project. I then renamed columns so they were all in the same format. Also, I arranged the sale_price from lowest to highest for readability purposes.

Model 1 plot

```
model <- lm(sale_price ~ sq_ft_lot, data = housing_data)
summary(model)
```

```
##
## Call:
## lm(formula = sale_price ~ sq_ft_lot, data = housing_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2016064  -194842   -63293    91565  3735109
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.418e+05  3.800e+03  168.90   <2e-16 ***
## sq_ft_lot   8.510e-01  6.217e-02   13.69   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 401500 on 12863 degrees of freedom
## Multiple R-squared:  0.01435,    Adjusted R-squared:  0.01428
## F-statistic: 187.3 on 1 and 12863 DF,  p-value: < 2.2e-16
```

```
ggplot(housing_data, aes(x = sq_ft_lot, y = sale_price)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Linear Regression: sq_ft_lot vs. sale_price",
       x = "Square Footage of Lot",
       y = "Sale Price")
```

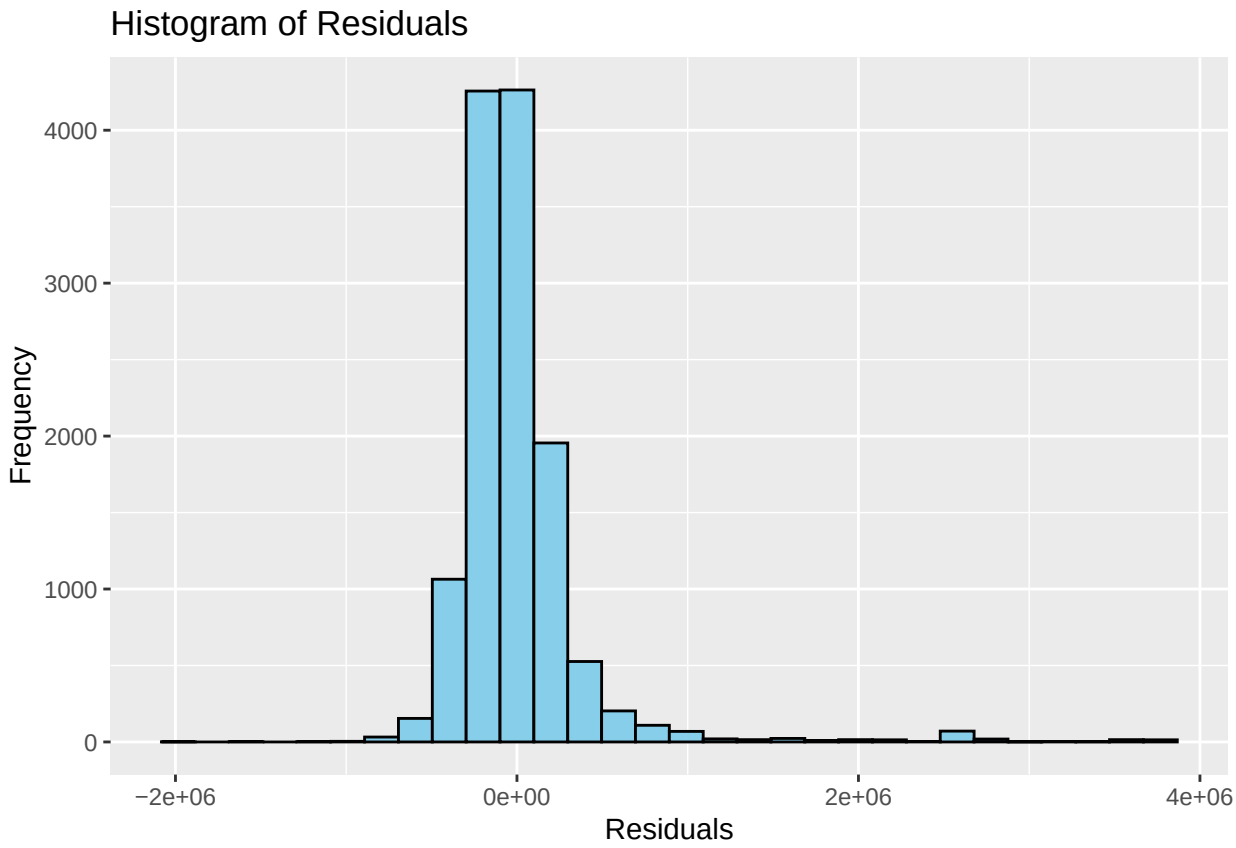
```
## `geom_smooth()` using formula = 'y ~ x'
```

Explanation of results

R2: .01435 This will measure the variance of "sale_price" against "sq_ft_lot" meaning that 1.43% of this variance is explained withing the "sq_ft_lot". Since this is a low score it indicates that this model may not be best for this data. Adj R2: .01428 This is slightly lower than R2, but takes into consideration the number of predictors, which is only sq_ft_lot in this model. This model also has a low score and indicates that the model may not be best for this data.

Plot of residuals

```
residuals <- resid(model)
ggplot() +
  geom_histogram(aes(x = residuals), bins = 30, fill = "skyblue", color = "black") +
  labs(title = "Histogram of Residuals",
       x = "Residuals",
       y = "Frequency")
```



Explanation

The residuals have a positive skewed distribution meaning that the model does not fit well with the dataset- in order to have the best accurary we would want to see a normal distrubition.

QQ Plot of residuals

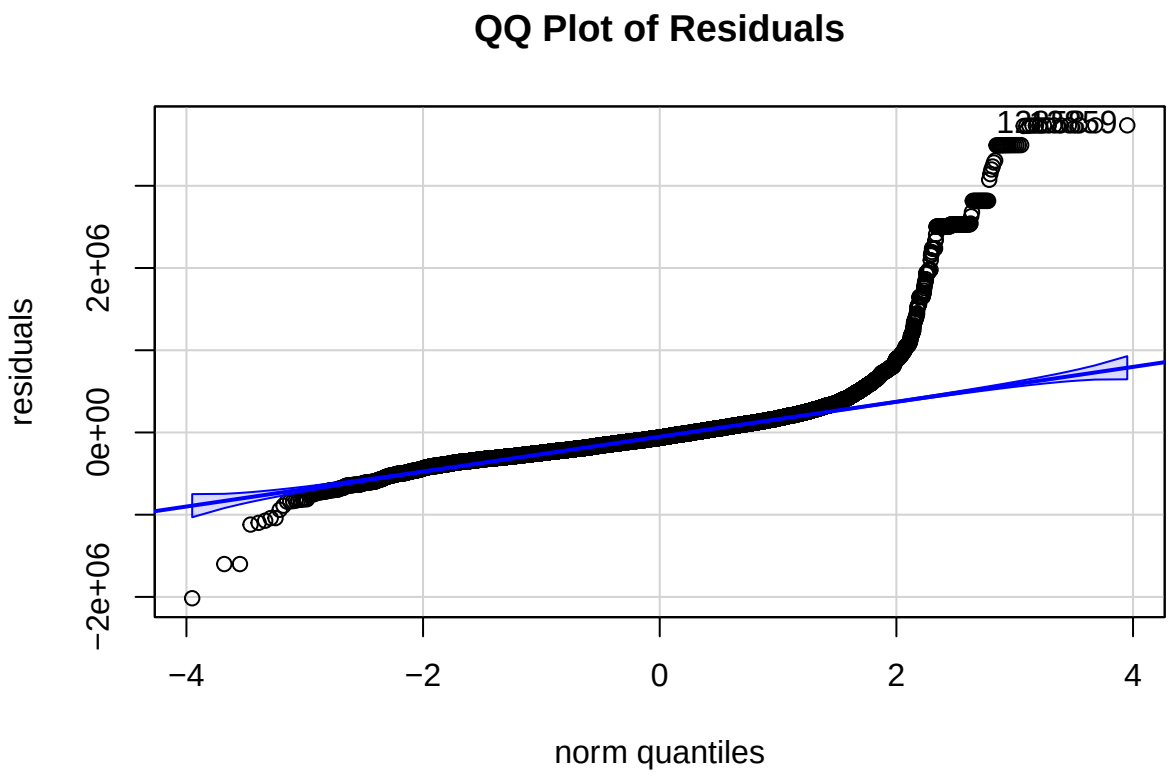
```
library(car)
```

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

```
##
## Attaching package: 'car'
```

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

```
model <- lm(sale_price ~ sq_ft_lot, data = housing_data)
residuals <- resid(model)
qqPlot(residuals, main = "QQ Plot of Residuals")
```




```
## [1] 12859 12858
```

Explanation

From the plot, we can see that the QQ plot does meet the normality assuption, but is skewed it terms of the distribution at the higher data points.

Model 2

```
model <- lm(sale_price ~ bedrooms + bath_full_count + square_feet_total_living, data = housing_data)
summary(model)
```

```
##
## Call:
## lm(formula = sale_price ~ bedrooms + bath_full_count + square_feet_total_living,
##     data = housing_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1760583  -117559   -41529    43918  3832099
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              204679.013  14013.468  14.606  < 2e-16 ***
## bedrooms                 -25206.328   4417.404  -5.706 1.18e-08 ***
## bath_full_count           42309.808   5685.497   7.442 1.06e-13 ***
## square_feet_total_living    184.150      4.353  42.302  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 359000 on 12861 degrees of freedom
## Multiple R-squared:  0.212,  Adjusted R-squared:  0.2118
## F-statistic:  1153 on 3 and 12861 DF,  p-value: < 2.2e-16
```

Explanation

When choosing a new model I thought adding the variables bedrooms, bath_full_count, and square_feet_total_living would be good indicators to predict sale_price. Upon review we can see that for each square foot increases the sales price by about $184.15. However, there is a decrease in sales prices for each additonal bedroom, which I find to be an odd finding when reviewing the results. One finding that I thought was very interesting was that for each full bath there was a sale increase of about 42,000.

Model comparison with ANOVA

```
model1 <- lm(sale_price ~ sq_ft_lot, data = housing_data)
model2 <- lm(sale_price ~ bedrooms + bath_full_count + square_feet_total_living, data = housing_data)
summary(model1)
```

```
##
## Call:
## lm(formula = sale_price ~ sq_ft_lot, data = housing_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2016064  -194842   -63293    91565  3735109
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.418e+05  3.800e+03  168.90   <2e-16 ***
## sq_ft_lot   8.510e-01  6.217e-02   13.69   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 401500 on 12863 degrees of freedom
## Multiple R-squared:  0.01435,    Adjusted R-squared:  0.01428
## F-statistic: 187.3 on 1 and 12863 DF,  p-value: < 2.2e-16
```

```
summary(model2)
```

```
##
## Call:
## lm(formula = sale_price ~ bedrooms + bath_full_count + square_feet_total_living,
##     data = housing_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1760583  -117559   -41529    43918  3832099
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              204679.013  14013.468  14.606  < 2e-16 ***
## bedrooms                 -25206.328   4417.404  -5.706 1.18e-08 ***
## bath_full_count           42309.808   5685.497   7.442 1.06e-13 ***
## square_feet_total_living    184.150      4.353  42.302  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 359000 on 12861 degrees of freedom
## Multiple R-squared:  0.212,  Adjusted R-squared:  0.2118
## F-statistic:  1153 on 3 and 12861 DF,  p-value: < 2.2e-16
```

```
anova_result <- anova(model1, model2)
print(anova_result)
```

```
## Analysis of Variance Table
##
## Model 1: sale_price ~ sq_ft_lot
## Model 2: sale_price ~ bedrooms + bath_full_count + square_feet_total_living
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1  12863 2.0734e+15
## 2  12861 1.6576e+15  2 4.1574e+14 1612.8 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Explanation

After reviewing the ANOVA we can see that model 2 has a better fit that model 1. The F- vlaue of 1612.8 and P- value of $<$ 2.2e-16 indicate that the variables have a large factor when predicting the sale price.

RMSE of models

```
install.packages("Metrics", repos = "http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/7s/jr1rn37s2wx1zv5ywl0kgr3m0000gn/T//Rtmp3LbwQ8/downloaded_packages
```

```
library(Metrics)
model1 <- lm(sale_price ~ sq_ft_lot, data = housing_data)
preds_model1 <- predict(object = model1, newdata = housing_data)
rmse_model1 <- rmse(housing_data$sale_price, preds_model1)
print(paste("RMSE for Model 1:", rmse_model1))
```

```
## [1] "RMSE for Model 1: 401452.546946962"
```

```
model2 <- lm(sale_price ~ bedrooms + bath_full_count + square_feet_total_living, data = housing_data)
preds_model2 <- predict(object = model2, newdata = housing_data)
rmse_model2 <- rmse(housing_data$sale_price, preds_model2)
print(paste("RMSE for Model 2:", rmse_model2))
```

```
## [1] "RMSE for Model 2: 358955.129516924"
```

Explanation

The RMSE score is used to measure the models ability to predict the target value. The RMSE for Model 1 is 401,452 and RMSE for Model 2 is 358,955. Typically the lower RMSE is more accurate when looking at the model. After reviewing the scores we can see that the second score is more accurate, but it does have a large score of 358,955 meaning that the model is off 358,955 compared to the dataset.