LukeLadasLab5

The data we are looking at for this lab is focused on apartment sales in Sao Palo, Brazil. These are apartments that were either for sale or for rent and were advertised in April 2019. There are many attributes for apartments such as condo expenses, size, amount of rooms suites etc. We want to use multiple linear regression to determine if there is a way to predict an apartment's price based on features. This dataset was found on kaggle and the link is given here. https://www.kaggle.com/datasets/argonalyst/sao-paulo-real-estate-sale-rent-april-2019 . There doesn't seem to be many flaws in the way sampling was done for this dataset, there are no missing values and the data is taken from reliable real estate property sources. There is also many features and rows to use for prediction.

Before I imported this dataset, some columns were either removed or edited slightly. I removed the columns property type and district, first column being removed was that every row was the same value (apartment), and district being removed to see if longitude and latitude are more specific location features. I changed negotiation type so that instead of it being the values rent and sale they became 0 and 1 for ease of use in predicting with linear regression. I changed some column titles to have an underscore instead of space as well in order for them to be able to be accessed in R.

```
library(readr)
sao_paulo_properties <- read_csv("C:/Users/luker/Downloads/sao-paulo-properties-april-2019_-_sao-paulo-
```

```
## Rows: 13640 Columns: 14
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (14): Price, Condo, Size, Rooms, Toilets, Suites, Parking, Elevator, Fur...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Let variables x1, x2, x3, . . . , x13 be the respective weighted predictors condo expenses, apartment size, number of rooms, number of toilets, number of suites, has parking, has elevator, is furnished, has swimming pool, is new, is rent or sale, and longitude/latitude. The initial linear model chosen is y hat = B0 + B1x1 + B2x2 + B3x3 + . . . + B13x13, where betas are beta hats. We now put this model into R and look at the results produced below.

```
summary(lm(Price~Condo+Size+Rooms+Toilets+Suites
           +Parking+Elevator+Furnished+Swimming_Pool+New+Negotiation_Type+Latitude
           +Longitude, data=sao_paulo_properties))
```

```
##
## Call:
## lm(formula = Price ~ Condo + Size + Rooms + Toilets + Suites +
##     Parking + Elevator + Furnished + Swimming_Pool + New + Negotiation_Type +
##     Latitude + Longitude, data = sao_paulo_properties)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3049327  -201859    -1264   147237  7818574
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -4.213e+05  2.012e+04 -20.938  < 2e-16 ***
## Condo            -6.744e+01  7.887e+00  -8.550  < 2e-16 ***
## Size              3.982e+03  1.327e+02  30.016  < 2e-16 ***
## Rooms            -6.108e+04  6.686e+03  -9.135  < 2e-16 ***
## Toilets           8.766e+04  9.790e+03   8.954  < 2e-16 ***
## Suites           -4.082e+04  1.115e+04  -3.663 0.000251 ***
## Parking           8.858e+04  7.689e+03  11.521  < 2e-16 ***
## Elevator          1.496e+04  8.302e+03   1.802 0.071546 .
## Furnished         1.090e+04  1.078e+04   1.012 0.311717
## Swimming_Pool     2.003e+04  8.127e+03   2.465 0.013710 *
## New              -1.055e+05  3.067e+04  -3.440 0.000583 ***
## Negotiation_Type  6.450e+05  7.733e+03  83.404  < 2e-16 ***
## Latitude          4.107e+03  2.963e+03   1.386 0.165789
## Longitude        -1.551e+03  1.512e+03  -1.025 0.305185
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 432900 on 13626 degrees of freedom
## Multiple R-squared:  0.4636, Adjusted R-squared:  0.4631
## F-statistic: 905.9 on 13 and 13626 DF,  p-value: < 2.2e-16
```

Our predicted model initially is y hat = -4.213e+05 -6.744e+01x +3.982e+03x2 - 6.108e+04x3 + … + -1.551e+03x13. At confidence level (alpha) being 0.05 we find out that not all of our features are significant. For our first model, we found out that having an elevator, being furnished, and an apartments longitude and latitude are not significant because their respective p-values are greater than 0.05. The remaining features would be considered important/significant predictors since their p-values are below 0.05. The R-squared for this model is 0.4636, which means that around 46.36% of observed variation in apartment price can be explained by our features. Our p-value relating to the F-statistic being < 2.2e-16 means that it would reject the null hypothesis of the model utility test (all B's are equal to zero), meaning there is significant evidence that not all (at least one) B is not equal to zero. We should remove the non-significant features in an updated model to see if we can improve significance as well as increase our R-squared for our model. We should first drop the highest p-value which would be if an apartment is furnished, then longitude, then latitude, then elevator. Our new predicted model should look like this… y hat = B0 + B1x1 + B2x2 + B3x3 + … + B11x11

```r
summary(lm(Price~Condo+Size+Rooms+Toilets+
             Suites+Parking+Elevator+Swimming_Pool+New+Negotiation_Type
           +Latitude+Longitude, data=sao_paulo_properties))
```

```
##
## Call:
## lm(formula = Price ~ Condo + Size + Rooms + Toilets + Suites +
##     Parking + Elevator + Swimming_Pool + New + Negotiation_Type +
##     Latitude + Longitude, data = sao_paulo_properties)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3050094  -201763      -11   147730  7817186
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -4.194e+05  2.004e+04 -20.935  < 2e-16 ***
## Condo            -6.687e+01  7.867e+00  -8.500  < 2e-16 ***
## Size              3.979e+03  1.326e+02  30.000  < 2e-16 ***
## Rooms            -6.186e+04  6.642e+03  -9.313  < 2e-16 ***
## Toilets           8.801e+04  9.784e+03   8.996  < 2e-16 ***
## Suites           -4.100e+04  1.114e+04  -3.679 0.000235 ***
## Parking           8.887e+04  7.683e+03  11.567  < 2e-16 ***
## Elevator          1.534e+04  8.293e+03   1.850 0.064294 .
## Swimming_Pool     2.099e+04  8.072e+03   2.600 0.009319 **
## New              -1.067e+05  3.064e+04  -3.483 0.000498 ***
## Negotiation_Type  6.445e+05  7.719e+03  83.499  < 2e-16 ***
## Latitude          4.095e+03  2.963e+03   1.382 0.167017
## Longitude        -1.546e+03  1.512e+03  -1.022 0.306706
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 432900 on 13627 degrees of freedom
## Multiple R-squared:  0.4636, Adjusted R-squared:  0.4631
## F-statistic: 981.3 on 12 and 13627 DF,  p-value: < 2.2e-16
```

```
summary(lm(Price~Condo+Size+Rooms+Toilets
           +Suites+Parking+Elevator+Swimming_Pool+New+
             Negotiation_Type+Latitude, data=sao_paulo_properties))
```

```
##
## Call:
## lm(formula = Price ~ Condo + Size + Rooms + Toilets + Suites +
##     Parking + Elevator + Swimming_Pool + New + Negotiation_Type +
##     Latitude, data = sao_paulo_properties)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3056194  -202138       63   147776  7817272
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -4.174e+05  1.994e+04 -20.937  < 2e-16 ***
## Condo            -6.682e+01  7.867e+00  -8.494  < 2e-16 ***
## Size              3.979e+03  1.326e+02  29.997  < 2e-16 ***
## Rooms            -6.188e+04  6.642e+03  -9.317  < 2e-16 ***
## Toilets           8.809e+04  9.784e+03   9.003  < 2e-16 ***
## Suites           -4.095e+04  1.114e+04  -3.674 0.000239 ***
## Parking           8.889e+04  7.683e+03  11.570  < 2e-16 ***
## Elevator          1.519e+04  8.292e+03   1.832 0.066905 .
## Swimming_Pool     2.087e+04  8.071e+03   2.586 0.009715 **
## New              -1.065e+05  3.064e+04  -3.474 0.000513 ***
## Negotiation_Type  6.446e+05  7.718e+03  83.512  < 2e-16 ***
## Latitude          1.138e+03  6.439e+02   1.768 0.077136 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 432900 on 13628 degrees of freedom
## Multiple R-squared:  0.4635, Adjusted R-squared:  0.4631
## F-statistic:  1070 on 11 and 13628 DF,  p-value: < 2.2e-16
```

```
summary(lm(Price~Condo+Size+Rooms+Toilets+Suites+
             Parking+Elevator+Swimming_Pool+
             New+Negotiation_Type, data=sao_paulo_properties))
```

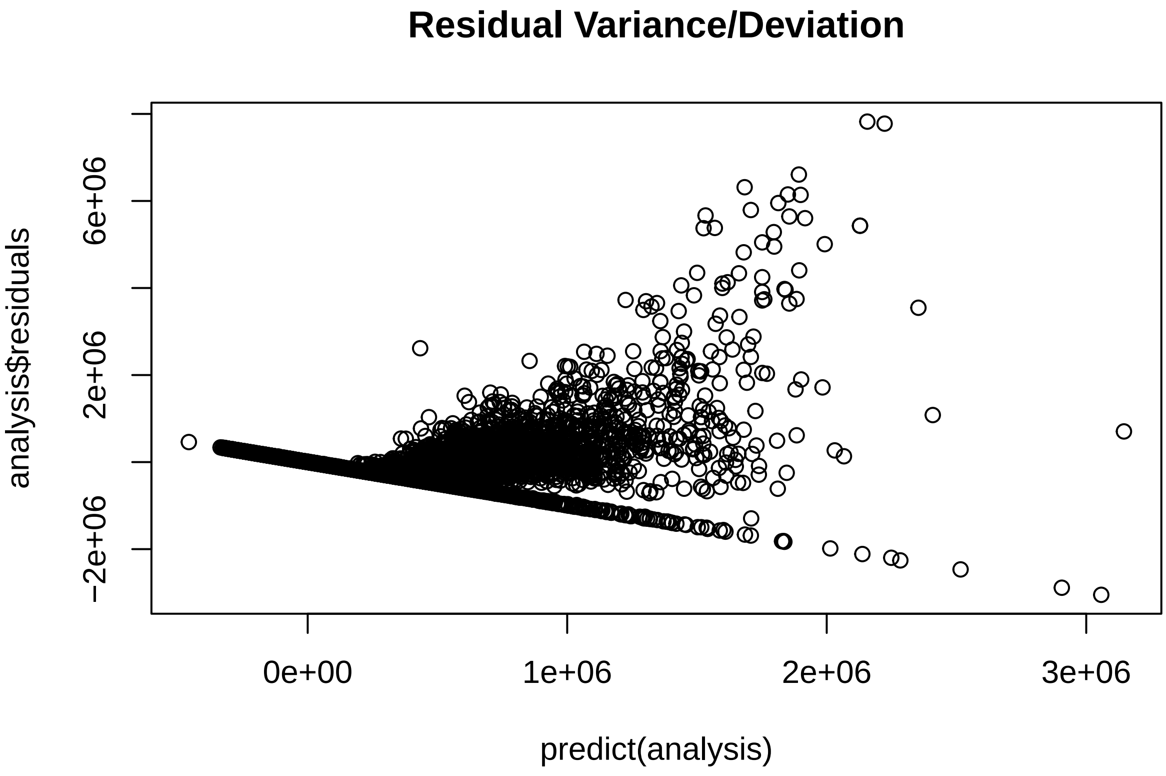
```
##
## Call:
## lm(formula = Price ~ Condo + Size + Rooms + Toilets + Suites +
##     Parking + Elevator + Swimming_Pool + New + Negotiation_Type,
##     data = sao_paulo_properties)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3063634  -201098     -995   147460  7818160
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -4.410e+05  1.482e+04 -29.747  < 2e-16 ***
## Condo            -6.685e+01  7.868e+00  -8.497  < 2e-16 ***
## Size              3.991e+03  1.325e+02  30.125  < 2e-16 ***
## Rooms            -6.203e+04  6.642e+03  -9.339  < 2e-16 ***
## Toilets           8.760e+04  9.781e+03   8.956  < 2e-16 ***
## Suites           -4.206e+04  1.113e+04  -3.780 0.000157 ***
## Parking           8.906e+04  7.683e+03  11.591  < 2e-16 ***
## Elevator          1.374e+04  8.252e+03   1.665 0.095845 .
## Swimming_Pool     2.117e+04  8.070e+03   2.623 0.008717 **
## New              -1.057e+05  3.064e+04  -3.451 0.000560 ***
## Negotiation_Type  6.446e+05  7.719e+03  83.510  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 433000 on 13629 degrees of freedom
## Multiple R-squared:  0.4634, Adjusted R-squared:  0.463
## F-statistic:  1177 on 10 and 13629 DF,  p-value: < 2.2e-16
```

```
summary(lm(Price~Condo+Size+Rooms+Toilets
           +Suites+Parking+Swimming_Pool+New+
             Negotiation_Type, data=sao_paulo_properties))
```

```
##
## Call:
## lm(formula = Price ~ Condo + Size + Rooms + Toilets + Suites +
##     Parking + Swimming_Pool + New + Negotiation_Type, data = sao_paulo_properties)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3049221  -201207     -690   148008  7823459
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -4.398e+05  1.481e+04 -29.699  < 2e-16 ***
## Condo            -6.691e+01  7.868e+00  -8.504  < 2e-16 ***
## Size              3.979e+03  1.323e+02  30.077  < 2e-16 ***
## Rooms            -6.227e+04  6.641e+03  -9.377  < 2e-16 ***
## Toilets           8.993e+04  9.681e+03   9.290  < 2e-16 ***
## Suites           -4.249e+04  1.113e+04  -3.819 0.000134 ***
## Parking           8.837e+04  7.673e+03  11.517  < 2e-16 ***
## Swimming_Pool     2.338e+04  7.960e+03   2.938 0.003313 **
## New              -9.736e+04  3.023e+04  -3.221 0.001281 **
## Negotiation_Type  6.457e+05  7.691e+03  83.958  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 433000 on 13630 degrees of freedom
## Multiple R-squared:  0.4633, Adjusted R-squared:  0.4629
## F-statistic:  1307 on 9 and 13630 DF,  p-value: < 2.2e-16
```

The new model looks like this: y hat = -4.213e+05 -6.744e+01x +3.982e+03x2 - 6.108e+04x3 + … + 6.450e+05x11. Removing these features had little impact on the predictor it seems. The features that are remaining still have p-values less than 0.05 meaning they are good predictors but our R-squared did not change too much. It actually ended up decreasing to 46.33%, this isn't much but it still shows that removing those four insignificant features had an effect on the model somewhat. This linear model also rejected the model utility test null hypothesis and was also less than 2.2e-16. With the remaining models being highly significant as well as the R-squared not really changing much, we could use this as our finalized model. However we still need to check aassumptions for linear regression. The two being Shapiro test and to check if variance/standard deviation is constant. Because of restrictions in R, we are only allowed to include the first 5000 samples for the residuals in testing.

```
analysis <- lm(Price~Condo+Size+Rooms+Toilets+
                 Suites+Parking+Swimming_Pool+
                 New+Negotiation_Type, data=sao_paulo_properties)
plot(analysis$residuals~predict(analysis), main="Residual Variance/Deviation")
```



```
shapiro.test(analysis$residuals[0:5000])
```

```
##
##  Shapiro-Wilk normality test
##
```

```
## data:  analysis$residuals[0:5000]
## W = 0.73902, p-value < 2.2e-16
```

The shapiro test produced a p-value less than 2.2e-16 which is much smaller than 0.05 our significance level. This means we reject the null hypothesis of the shapiro test, which means that our model has very significant evidence that it is NOT normally distributed. Looking at the graph, we can also see that the standard deviation/variation is NOT constant, the plot somewhat resembles a cone shape. Because both of our assumptions were unsuccessful, we can deduce that our model and the predictions that come out of it may very likely be inaccurate.

In conclusion, we were able to get a somewhat accurate prediction model for predicting apartment sale prices in Sao Palo, Brazil. With a reduced model, with 46.33% of observed variation in apartment sales being explained by the features, condo price, size of apartment, number of rooms, num of toilets, num of suites, has parking, has swimming pool, is new, and if its rented or purchased.