# Housing Price Forecasting

Carlos Figueroa

Colorado State University Global

MIS470: Data Science Foundation

Kelly Wibbenmeyer

Due date

# 

# Housing Price Forecasting

## Introduction.

The importance of Housing Price Forecasting is undeniable, as Real State is one of the most critical sectors in the economy. The 2009 subprime financial crisis revealed that as an asset class, Real State is interconnected to the rest of the economic system through the financial system due to leverage, collaterals, and the securitization of loans. Therefore, forecasting models should be widely studied.

Housing prices impact the formation and burst of bubbles, macroeconomic processes, such as business cycles, unemployment, aggregate consumption, etc. While other assets such as bonds, commodities, and currencies have different price dynamics, Real State has specific characteristics because of its heterogeneity due to property location and physical attributes. (Ghysels, Plazzi , Tourus, & Valkanov, 2013)

Housing price predictability faces many challenges. For a start as an asset class, Housing Prices are infrequently traded. Therefore, Real State data is relatively short, unlike other assets such as bonds, commodities, and currencies that generate yearly, monthly, daily data, and even by-minute data. Also, house prices face high transaction costs, and they are inherently illiquid. (Ghysels, Plazzi , Tourus, & Valkanov, 2013)

Acknowledging the possible limitations, we wanted to have a first approach to analyze and forecast Housing Prices based on the sale prices in Ames, Iowa. We wanted to know which factors are significant to our model and how good is our prediction. We use a data set of 1,000 observations to fit a linear regression model, and then with a different data set, we tested how well the predicted prices fit the observed prices.

## 1.Examining housing.testing.csv.

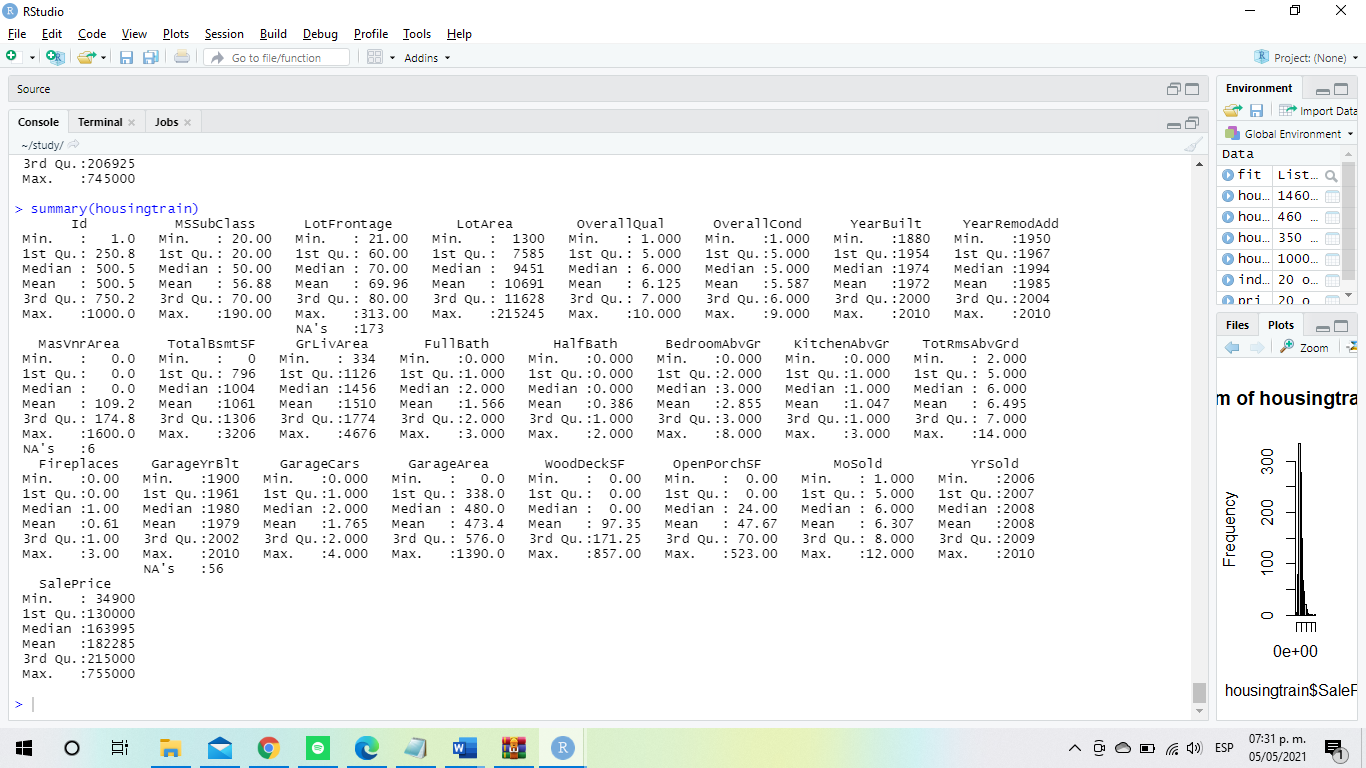


Figure .Training Data Summary

We can observe that the minimum value is very different from the two data sets, so we should expect that the forecasted price differs for small house prices.

Figure . Testing Data Summary

The 1st, 2nd(or median), and 3rd quantiles are very close to each other, so we should expect data to be similarly distributed across the price ranges. The mean and the maximum of the training data are higher than the mean from the testing data, but they are not significantly different.

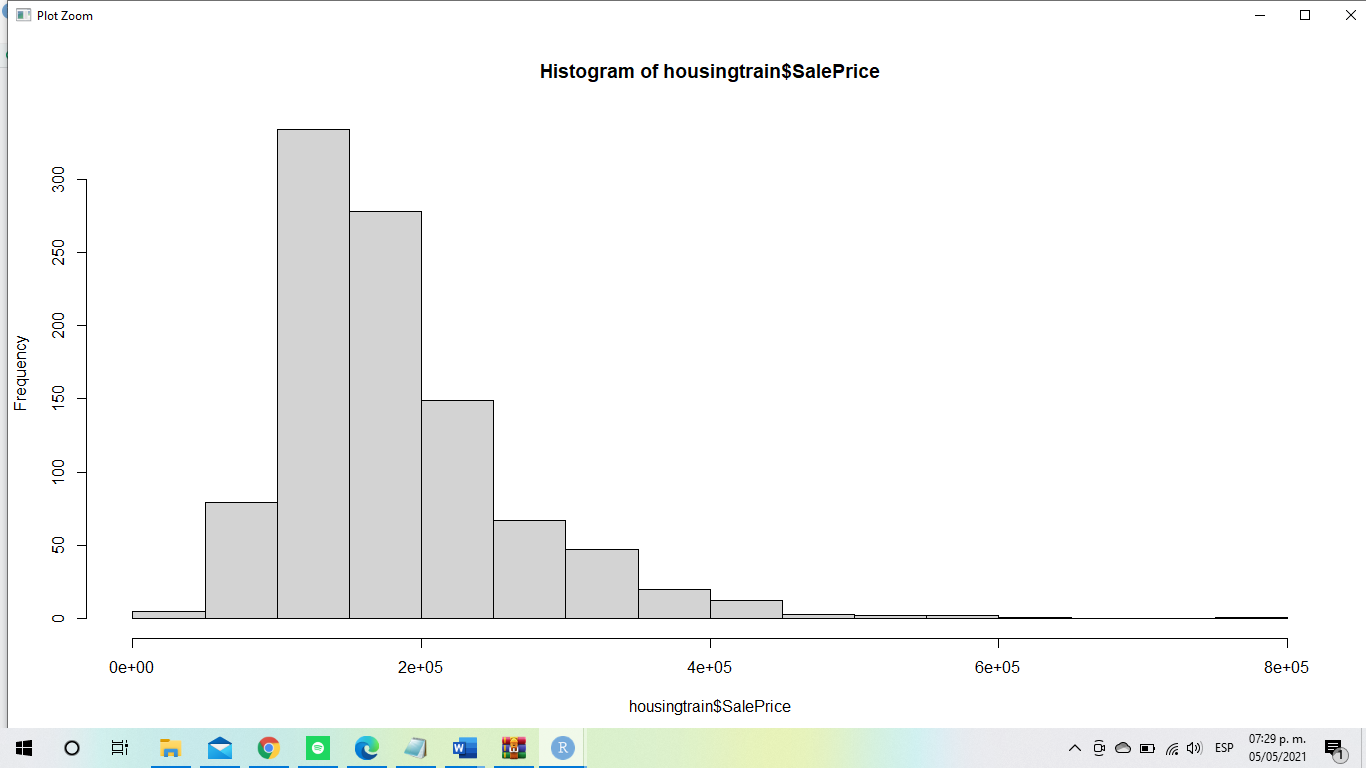


Figure . Sale Price Histogram for Training Data

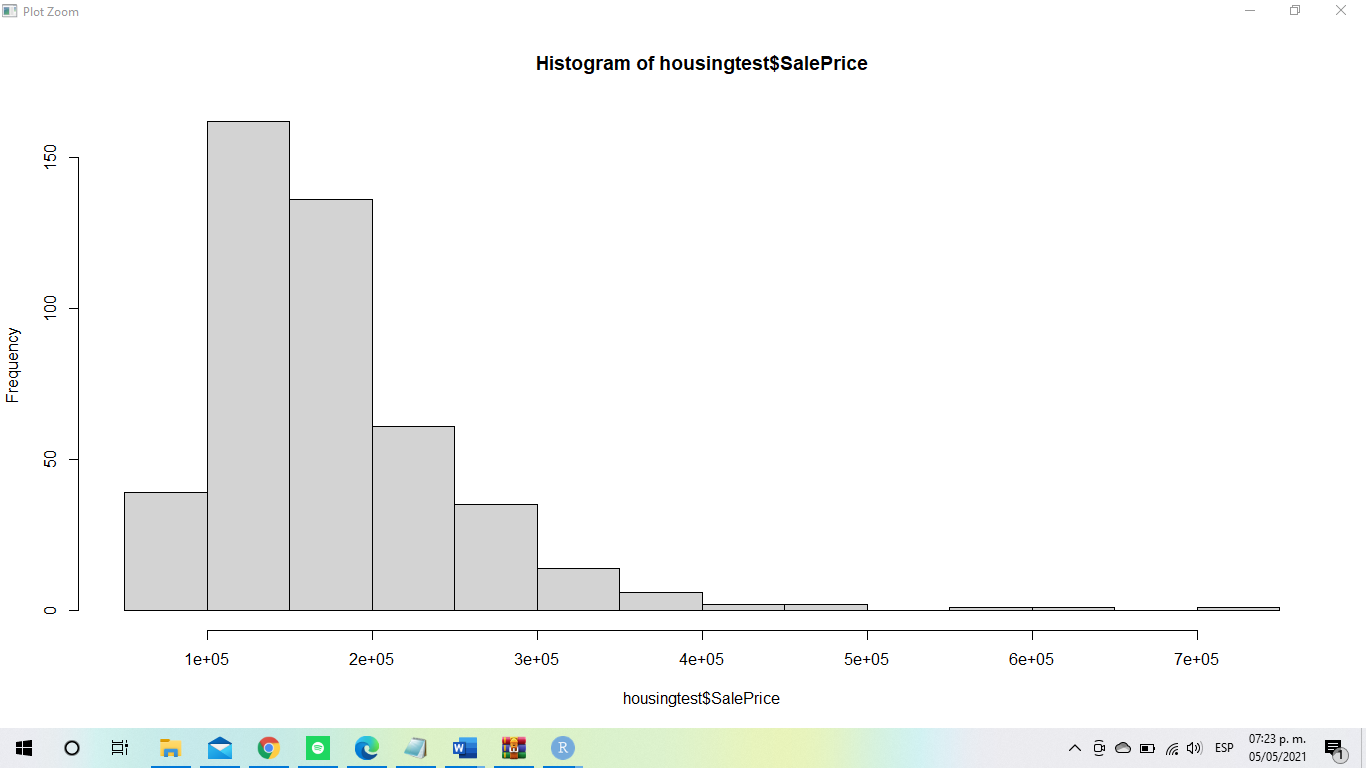


Figure . Sale Price Histogram for Testing Data

The histograms show that both data sets are left-skewed and similarly distributed, except for small values of SalePrice.

## 2. Combining the two data sets.

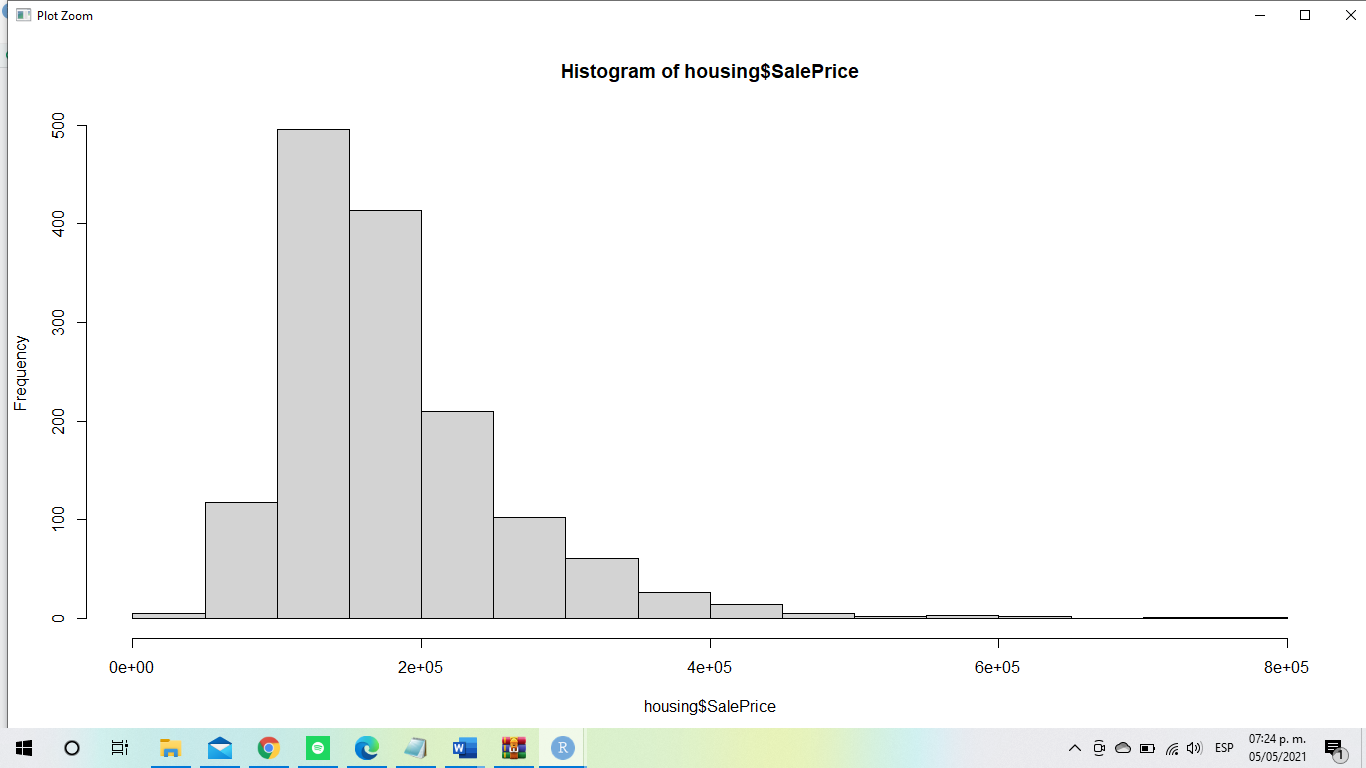


Figure . Sale Price Histogram for Combined Data

The histogram from the combined data is similar to the training data histogram. The similarity is because the training data is a larger (1,000 observations) than the testing data set (460 observations).

## 3. Linear regression.

We calculated a linear regression model using the training data set. The SalePrice as a linear function of the rest of the variables:

As we said before, a linear regression model is the first approximation into predicting housing prices based on a linear model that uses the characteristics of each sale. More advanced models are very complex and use data-rich approaches, "which summarize a large amount of information in a relatively small number of estimated factors and use these to forecast house price fluctuations." (Bork & Stig, 2018)

The previous does not mean that we dismiss the reach of our model, but that we understand that it can have limitations. Still, as we will see in part 5, the simple linear regression model has high predictive power.

The regression output was:

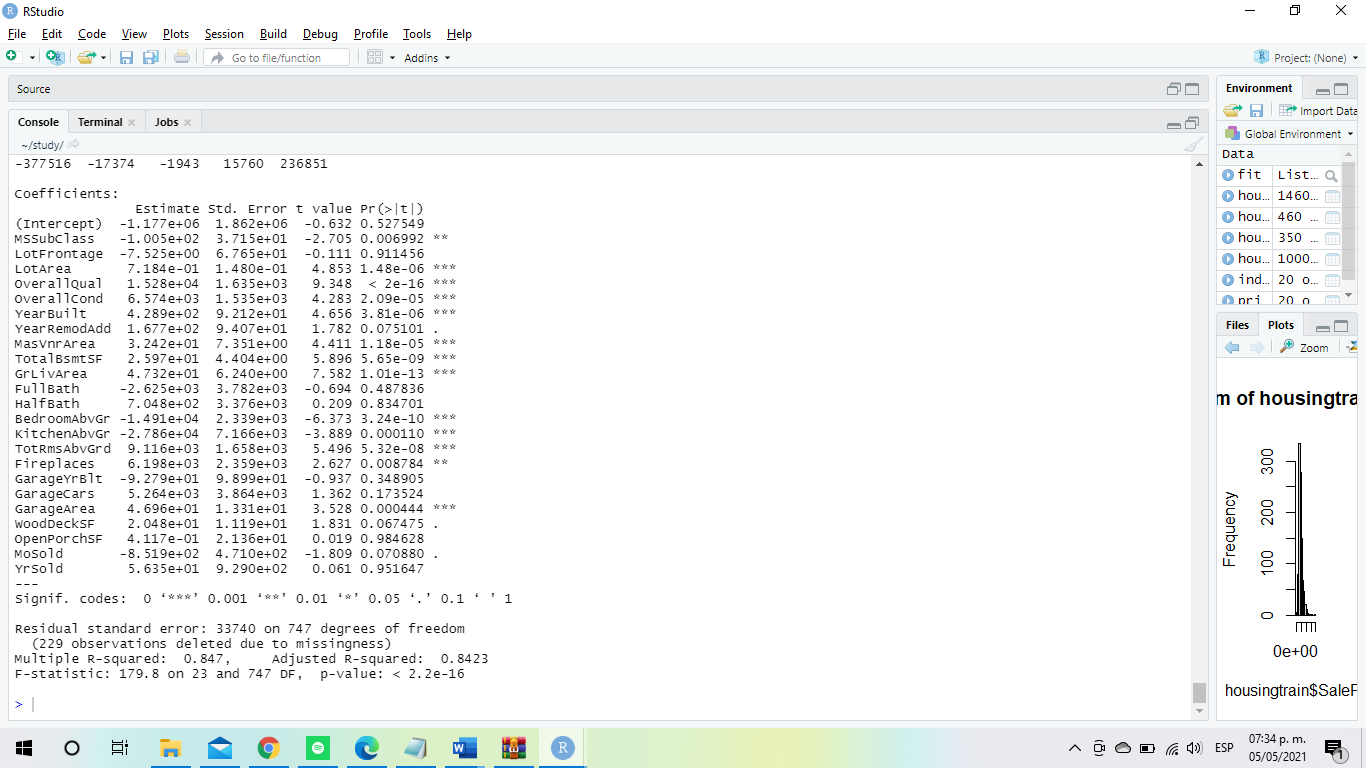


Figure . Regression Output

Before we interpret our model, we want to make some considerations: The model specification was made to analyze whether the independent variables impact housing prices. SalePrice is a linear function on 23 variables. The possible difficulty that could arise from this model is the problem of overfitting our data, which means that model is highly predictive for the training data, but it fails to replicate in future samples. Overfitting a model means that the results are overly optimistic, and the "findings" will not replicate on the population. (Babyak, 2004).

In this concern, it will be helpful to test our model's predictability with a testing data in part 5, which is different from the data from which we estimate our model (training data).

Additionally, we can observe that the and the adjusted-. The R-square is a measure of the goodness of fit of the model but always increases when we add a new variable. On the other hand, adjusted-R- squared does not always goes up when a variable is added (Hill, Griffiths, & Lim, 2011, pág. 237). The previous means that even if many variables can penalize the adjusted-R-squared since the two indicators are very close, we have not overfitted the model significantly.

### 4. Model interpretation

We performed hypothesis testing on each of the regression coefficients to test whether they are statistically significant. The hypothesis testing was given by:

The coefficient is statistically non-significant. The variable has no impact on the determination of Housing Prices.

The coefficient is statistically significant. The variable has an impact on Housing Prices.

To perform the hypothesis test, we used the p-value criteria. If the p-val<, then the variable is statistically significant. We condensed the results in a table, according to their p-values, for different significance levels:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Significant at α=.001 | Significant at α=.01 | Significant at α=.05 | Significant at α=.10 | Non-significant at any level |
| LotArea |  |  |  |  |  |
| OverallQual |  |  |  |  |  |
| OverallCond |  |  |  |  |  |
| YearBuilt |  |  |  |  |  |
| MasVnrArea |  |  |  |  |  |
| TotalBsmtSF |  |  |  |  |  |
| GrLivArea |  |  |  |  |  |
| BedroomAbvGr |  |  |  |  |  |
| KitchenAbvGr |  |  |  |  |  |
| TotRmsAbvGrd |  |  |  |  |  |
| GarageArea |  |  |  |  |  |
| MSSubClass |  |  |  |  |  |
| Fireplaces |  |  |  |  |  |
| YearRemodAdd |  |  |  |  |  |
| WoodDeckSF |  |  |  |  |  |
| MoSold |  |  |  |  |  |
| LotFrontage |  |  |  |  |  |
| FullBath |  |  |  |  |  |
| HalfBath |  |  |  |  |  |
| GarageYrBlt |  |  |  |  |  |
| GarageCars |  |  |  |  |  |
| OpenPorchSF |  |  |  |  |  |
| YrSold |  |  |  |  |  |

Therefore, the variables that have a more significant impact in our model are: LotArea, OveralQuall, OverallCond, YearBuilt, MasVnrArea, TotalBsmtSF, GrLivArea, BedroomAbvGr, KitchenAbvGr, TotRmsAbvGrd, GarageArea.

The variables, MSSubClass, Fireplaces, YearRemodAdd WoodDeckSF, MoSold, have a statistical significance depending on the significance level.

The variables LotFrontage, FullBath, HalfBath, GarageYrBlt, GarageCars, OpenPorchSF, and YrSold have no impact on the SalePrice.

Above the most significant variables, the impact SalePrice is:

|  |  |
| --- | --- |
| Variable | Impact on SalePrice |
| LotArea | Positive |
| OverallQual | Positive |
| OverallCond | Positive |
| YearBuilt | Positive |
| MasVnrArea | Positive |
| TotalBsmtSF | Positive |
| GrLivArea | Positive |
| BedroomAbvGr | Negative |
| KitchenAbvGr | Negative |
| TotRmsAbvGrd | Positive |
| GarageArea | Positive |

Figure .Signs of the regressors.

Therefore, an increase BedroomAbvGr, and in KitvchenAbvGrd have a negative impact the House Price: Having a kitchen or a bathroom above the grade floor (above the ground) diminishes the price of the house. The rest of the significant variables have a positive impact on the price.

## 5. Predicted sales prices.

Figure . Predicted / Oserved Sale Prices

We calculated the predicted values for the first 20 observations of our testing data and compared them with the observed values. In figure.9 'predict(fit, indep)' are the predicted prices while the 'housingtest1$SalePrice[1:20]' are the observed prices once we have eliminated the missing data.

To visualize the data, we graphed the predicted and the observed values in R. The red dots are the observed prices, and the blue diamonds are the predicted values:

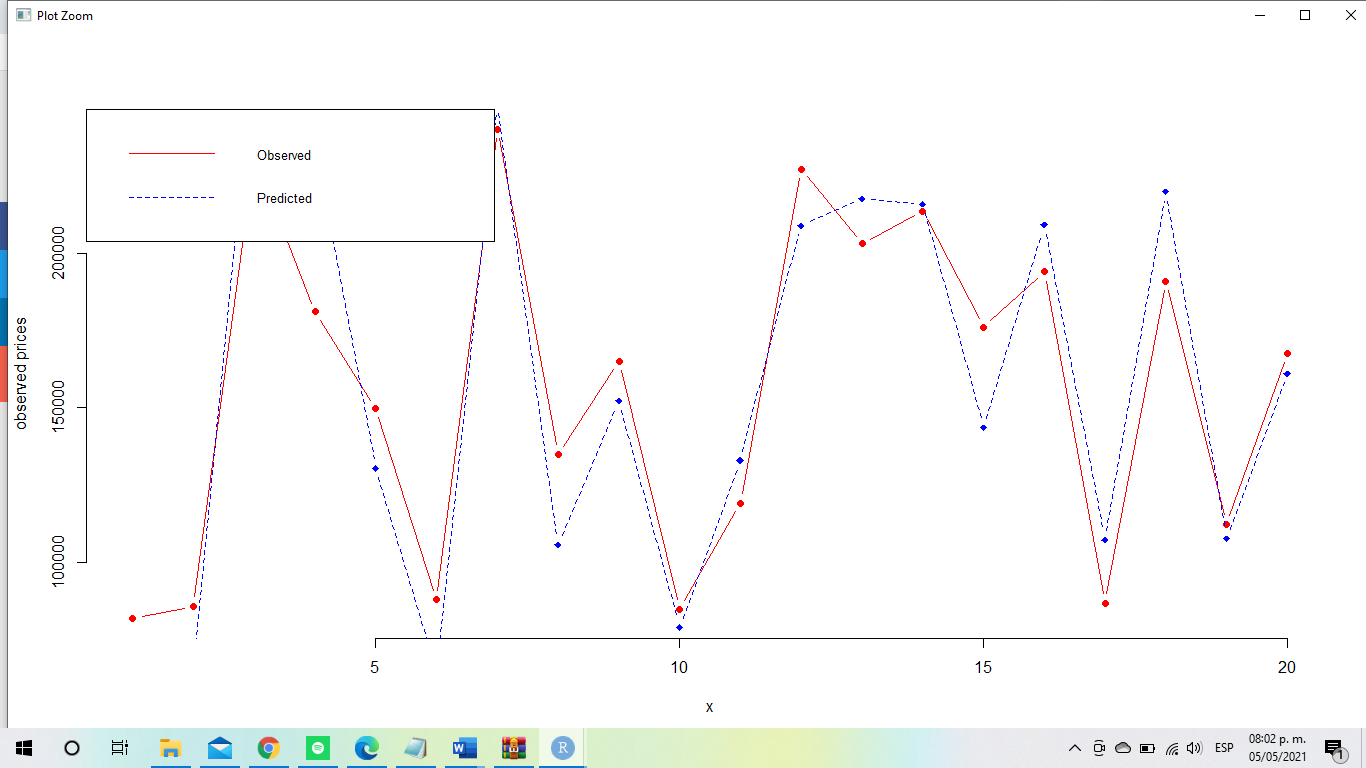


Figure . Predicted / Observed Sale Prices

We can notice that the observed prices and the predicted prices are very similar visually.

## Conclusions.

The relative importance of Real State within the economic dynamics makes it crucial to explore and analyze forecasting models to predict house pricing. We were initially worried that an overfitting problem could arise from a linear specification by introducing too many variables. To verify that overfitting does not create a model that only describes the data under which it was constructed, we use different data sets:

-The training data – to estimate our linear regression.

-The testing data – to test whether the predicted values are accurate.

By using different data sets, we can verify that our linear model is accurate. We compared the predicted and observed prices *figure.10* and observed that it is a good predictive model.

11 variables were statistically significant, 7 variables were statistically insignificant, and 5 variables were statistically significant depending on the significance level.

# References

Babyak, M. A. (2004). What You See May Not Be What You Get: A Brief, Nontechnical Introduction to Overfitting in Regression-Type Models. *Psychosomatic Medicine*, 411–421.

Bork , L., & Stig, M. V. (2018). Housing Price Forecastability: A Factor Analysis. *Real State Economics* , 582–611.

Ghysels, E., Plazzi , A., Tourus, W., & Valkanov, R. (2013). Forecasting Real Estate Prices. *Handbook of Economic Forecasting*, 509–580.

Hill, C. R., Griffiths, W. E., & Lim, G. C. (2011). *Principles of Econometrics* (Fourth ed.). United States : John Wiley & Sons, Inc.