**Module 8: Portfolio Project Option 1**

Joshua Humphreys

Colorado State University Global

Course Code:MIS470

Dr. Osama Morad

December 5, 2021

**Module 8: Portfolio Project Option 1**

For my portfolio project I have selected option 1. I began this assignment by calculating the summary statistics of both the housing training and housing testing datasets and comparing the two. I then combined the two datasets and created a histogram of the combined dataset. Then, using only the training dataset, I fit a multiple linear regression model using SalePrice as the response variable. I then determined the significant factors and interpreted my model. Finally, I used my model to predict the sales price of the first twenty houses in the testing dataset.

**Summary Statistics**

In week four of the class, I calculated the mean, median, max, min, and standard deviation of the testing dataset. This week I made the same calculations for the training dataset. The figures below contain the R input I used as well as the R output.

**Figure 1.**

*R input and output for testing dataset saleprice summary statistics and R input for histogram*

Graphical user interface, text, application, email

Description automatically generated

**Figure 2.**

*R input and output for training dataset saleprice summary statistics and R input for histogram*

Graphical user interface, text, application

Description automatically generated

**Figure 3.**

*R output of the testing sale price histogram*

Chart, histogram

Description automatically generated

**Figure 4.**

*R output of the training sale price histogram*

Chart, histogram

Description automatically generated

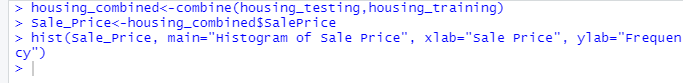
Comparing the two datasets we see that the range of the training dataset is larger than the testing. This makes sense given that the training dataset is a larger set. The medians of the two datasets are within 2245 of one another. Stated differently, the median of the testing dataset is within 1.3% of the training dataset. Similarly, the testing mean is within 2.4% of the training mean. Standard deviation is also close between the two datasets. Looking at the two histograms we see, once again, that the range is wider for the training dataset. Both histograms are asymmetric, unimodal, and skewed right. There is no significant difference in the mode between the two.

**Combining the Datasets**

The next step in this assignment was to combine the two datasets and produce a histogram of the resulting dataset. The first issue I ran into resulted from the way in which the *load.csv()* function was encoding the testing dataset. The *load.csv* added “Ï..” to the beginning of the first column label. This meant the column names were different between the two datasets. The *combine()* function requires the column names be the same between the two sets in order to function. The fix, courtesy of Roelpi (2019) was simply to reload the .csv files and include “fileEncoding = 'UTF-8-BOM'” in the arguments. The following figures contain the R input and output for combining the two datasets and producing the histogram.

**Figure 5.**

*R input to combine datasets and produce histogram*

**Figure 6.**

*R output of the combined sale price histogram*

Chart, histogram

Description automatically generated

Once again, we don’t see a significant difference between this distribution and the distribution of the other datasets. All the datasets are asymmetric, unimodal, and skewed right. There is no significant difference between mean, median, or mode between any of the sets

The important takeaway from the comparison between the testing, training, and combined datasets is that the three datasets are similar enough that it is reasonable to use these subsets as training and testing sets for our linear regression model. That is, we shouldn’t suspect any anomalies that could confuse our analysis in either set.

**Figure 7.**

*Screenshot with date demonstrating completion of assignment steps to this point*

Graphical user interface, application

Description automatically generated

**Linear Regression Model**

To create the linear regression model I used the *lm()* function. The figure below contains the R input and output for creating the linear regression model.

**Figure 8.**

*R input and output of creating the linear regression model*

Text

Description automatically generated with medium confidence

Table

Description automatically generated

The *summary()* output helps us answer several questions about our model. First, looking at the P-value allows us to determine what the significant factors are. LotArea, OverallQaul, OverallCond, Year Built, MasVnrArea, TotalBsmtSF, GrLivArea, BedroomAbvGr, KitchenAbvGr, TotRmsAbvGrd, and GarageArea all have a p-value less than .001. That is, each of those factors is unlikely to be associated with the sale price due to chance (Gareth et. al., 2017). Also, based on the estimate, we see that each of those factors has a positive correlation with the Sale Price. That is, as each of those increases, so does the sale price.

Lastly, the Residual Standard Error and the R-squared value are important components of determining model fit. We do see that, relative to the mean and median sale price, the Residual standard error is small. The R-squared value tells us, in this case, that the proportion of the variability in the sale price that can be explained using the model is .84. (Gareth et. al., 2017). Stated differently, 84% of the variance in sale price is explained by this model. I would consider this model a good predictor of sale-price.

**Predictions Using the Model**

For the next part of the assignment, I first removed all the rows in the testing data set with an “NA” in any column. I then used the model that I created to predict the sale price for each of the first 20 houses in the testing dataset. The figure below shows the R input I used. I started by using the *complete.cases()* function to remove the rows with columns containing “NA”. I then put the first 20 rows from the testing set into a new set. I then used the *predict()* function to predict the sales price for those cases.

**Figure 9.**

*R input and output of predicting sale price*

A picture containing text

Description automatically generated

In order to evaluate the model, because this is such a small set, I chose to keep things simple. I simply calculated the difference between the predicted value and the real value. If we were to use the entire testing dataset, more rigorous methods would be appropriate. The following figure contains my R input and output for evaluating the model’s predictions.

**Figure 9.**

*R input and output of predicting sale price*

Text

Description automatically generated with medium confidence

To me, the most notable part of this output are 6 rows with missing values. Considering that’s greater than 25% of this sample, I’m a little concerned. We do know that those missing values are not a result of missing values in the testing dataset as we eliminated those. Therefore, it’s reasonable to assume that for whatever reason, the model was unable to predict the sale price in those cases. I think a different and larger test set would be warranted to see if the problem persists.

Next, we see that the model was off significantly for extremely low sale prices. However, we also see that the model’s results were far mor accurate as the sale price approached mean, median, and mode prices. This isn’t surprising as the extremely low sale prices are outliers and likely difficult to predict regardless of the model. Based on this abbreviated analysis, I would say the model performed well at making predictions of house sale prices.

**Conclusion**

In this assignment, I started with two datasets: a testing and training dataset. I compared those two datasets and combined them. I then compared the combination to the original dataset. I then created a multiple linear regression model from the training dataset. I used the model to make predictions on the testing dataset and evaluated the prediction. Overall, I think the model performed well.

**References**

Berrier, J., Nestler, S., Pardoe, I., Sturdivant, R.X., Watts, K., McCown, F. (2018, July). *MIS 470: Data Science Foundations.* zyBooks.

Gareth, J., Witten, D., Hastie, T., Tibshirani, R. (2017). *An Introduction to Statistical Learning.* Springer. <https://doi.org/10.1007/978-1-4614-7138-7>

Roelpi (2019, July 9). *Removing Ï..,I and two dots or ulaut, when using read.csv in R.* Roel Peters. <https://www.roelpeters.be/removing-i-umlaut-two-dots-data-frame-column-read-csv/>

Zach (2021, June 3). *How to Use complete.cases in R (With Examples).* Statology*.* https://www.statology.org/complete-cases-in-r/