Critical Thinking Week 8: Portfolio Project

Justin Cull

MIS470-1: Data Science Foundations

Professor Alin Tomoiaga

3/12/2023

**R Studio Step 1:**

Screenshot setting the working directory using setwd () function. Screenshot of R code reading

in the MIS470HousingTraining(1000x25) and MIS470HoustingTesting(460x25) csv files from

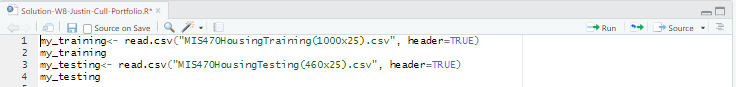
the MIS470 folder in the working directory for the Ames, IA housing data from 2006-2010. The

MIS470HousingTraining(1000x25) dataset has a total of 1000 observations, and 25 variables

in the data frame. The MIS470HousingTesting(460x25) has a total of 460 observations

and 25 variables in the dataset.





**R Studio Step 2:**

Screenshot using the summary () function and sd () function to display summary statistics and

standard deviation for the MIS470HoustingTesting(460x25) testing dataset. Below are the

screenshots displaying summary statistics for the 25 variables in the data frame and the standard

deviation for the variable – SalePrice. We are specifically interested in the SalesPrice column.

**Minimum = 52500 -> interpretation:** the lowest sale price for a residential dwelling in Ames,

IA between 2006-2010 was $52,500.

**1st Quartile = 129000-> interpretation:** 25% of the sale prices for a residential dwelling in

Ames, IA between 2006-2010 was less than or equal to $129,000.

**2nd Quartile (Median) = 1617500 -> interpretation:** half of the sale prices for a residential

dwelling in Ames, IA between 2006-2010 were greater than or equal to $161,750, and half of the

sale prices were less than or equal to $161,750.

**Mean = 177958 -> interpretation:** on average, a residential dwelling in Ames, IA between

2006-2010 sold for $177,958.

**3rd Quartile = 206925 -> interpretation:** 75% of the sale prices for a residential dwelling in

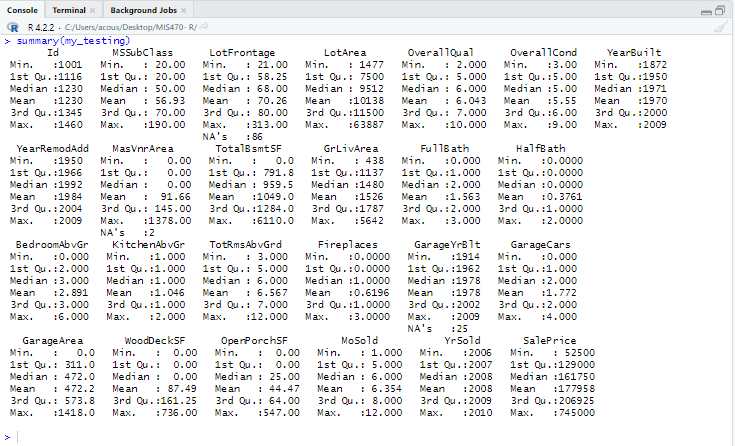
Ames, IA between 2006-2010 were less than or equal to $206,925.

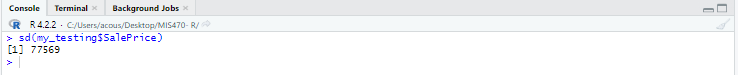
**Maximum = 745000 -> interpretation:** the highest sale price for a residential dwelling in

Ames, IA between 2006-2010 was $745,000.

**Standard Deviation = 77569 ->** interpretation: on average, a sale price for a residential

dwelling in Ames, IA deviates from the mean by $77,569





**R Studio Step 3:**

Screenshot using the hist () function to generate a visual histogram for the variable – SalePrice

for 460 observations in the Ames, IA dataset for house sale prices from 2006-2010. The first

histogram was generated with scientific notation on the (x) axis. I used the options () formula to

change this setting to numeric with the scipen option. The histogram is right skewed. Siegel &

Wagner (2022) state a skewed distribution is neither symmetric nor normal because the data

values trail off more sharply on one side than the other. The histogram also has somewhat of a

normal shape since most of the value’s center around the mean. We can see the highest

frequency of homes sold is between $100,000-$200,000 representing well over 50% of the full

dataset. The minimum sale price also falls within two standard deviations of the mean, which is

not uncommon. 75% of the full dataset sale price is $206,925 or below, so most homes were sold

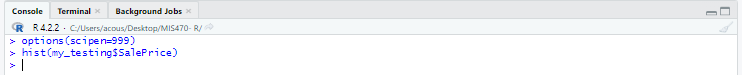
at this price or below. Factoring in the standard deviation (sd), 99.7% of all home prices fall

within three standard deviations of the mean, meaning $177,985 + $77,569 \* 3 = $410,692. We

can see 0.01, approximately 1.09% or 5 home sale prices fall within a $410,693 – $745,000 price

range and considered outliers which in this case creates the right skewness. We can conclude the

dataset has a large range, or large scatter of values around the mean.



Chart, histogram

Description automatically generated

**R Studio Step 4:**

First screenshot using R code to remove the ID column from both datasets and installing

the library(gdata), using the combine () function in the gdata library. RDocumentation (2023)

states the combine function takes a sequence of vector, matrix or data frames and combine into

rows of a common data frame with an additional column ‘source’ indicating the source object. In

this case, we are combining the MIS470HoustingTesting(460x25) testing and

MIS470HousingTraining(1000x25) dataset to create a combined 1460 observation data frame

and with 24 variables. Variable-ID was removed, although variable-source was added through

the combine function, combined into my\_data but does not account for a true independent

variable. If we run the summary statistics for my\_data, summary(my\_data), we can see the new

variable-source listed next to variable-SalePrice. The histogram continues to be right skewed.

The histogram also still has somewhat of a normal shape since most of the value’s center around

the mean. Compared to the testing data histogram, we can see the x-axis is now in price intervals

of $200,000 instead of $100,000. The y-axis is now frequency intervals of 100, instead of 50,

with a maximum frequency interval of 500, compared to 150 in the testing data set and 300 in

the training data set. The highest frequency distribution of homes sold are still between

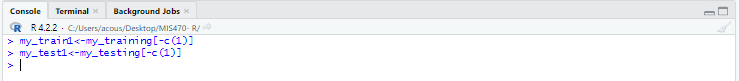
$100,000-$200,000 now represent well over 60% of the full dataset. 75% of the full dataset sale

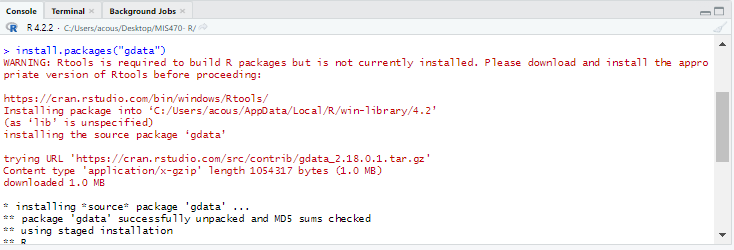
price is $214,000 or below, or 1095 homes were sold at this price or below. The median sale

price was $163,000. The minimum sale price was $34,900, and maximum sale price was

$755,000. We can conclude the dataset still has a large range, or large scatter of values around

the mean.



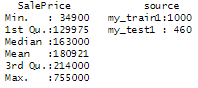


Graphical user interface, text, application

Description automatically generated

Graphical user interface, application

Description automatically generated





Chart, histogram

Description automatically generated

**R Studio Step 5:**

Screenshot using the lm () and summary () function creating a liner regression model and

summary statistics output data for the model. The summary () function displays the summary

statistics in the data frame, more importantly the y-intercept (b0), slope (b1), and p-value.  The

(b0) is the estimate for the y-intercept, and (b1) is the estimate for the slope. By setting the linear

regression model, we are programming R studio to predict variable (y). We can see the

coefficients for all 24 variables produce an intercept of approximately 215.77. We can also see a

mix of positive and negative coefficients between variables. To validate the model, we need to

attempt a hypothesis test on the significance of β1. In this case, a hypothesis is made around one

significant predictor variable-OverallQual. We can validate the model by performing a

hypothesis test on the null (H0) and the alternative hypothesis (H1) for listed below.

H0: The predictor *OverallQual* is not significant for our model.

H1: The predictor is significant for our model.

We are using multiple linear regression with 23 independent variables attempting to

analyze which variables are significant in the training dataset with p-value > 0.05. We need to

compare the significance level α=0.05, also known as the confidence level (CI) with the Pr(>|t|).

The **Pr(>|t|)** column represents the p-value associated with the value in the **t value** column. If the

p-value is less than the significance level α = .05, the independent variable (predictor variable)

has a statistically significant relationship to the dependent variable (response variable) in the

model. Each independent variable with Pr(>|t|) with at least one (\*) is a significant predictor.  We

can see 0 independent variables have (\*).  2 independent variables have a Pr(>|t|) > 0.01 or (\*\*).

11 independent variables have Pr(>|t|) > 0.001 or (\*\*\*). In this case, predictor variable-

*OverallQual* Pr(>|t|) =< 0.0000000000000002 which is far less than the significance level. The

model is validated, and we reject the null hypothesis in favor of the alternative hypothesis that

there is over a 95% probability that overall quality of a home affects sale price. Having 11

independent variables reach significance below 0.001 shows evidence of statistically highly

significant, also having 13 total independent variables with statistical significance with Pr(>|t|)

less than 0.01 shows a lot of strength in the dataset overall.

For the 13 significant predictors, understanding the estimated regression coefficients lets

us know the impact on the estimated sale price when the coefficient is positive or negative. A

positive coefficient lets us know when the value of the independent variable increases, the mean

of the dependent variable-SalePrice tends to increase. A negative coefficient lets us know when

the value of the independent variable increases, the dependent variable-SalePrice tends to

decrease. More importantly, the coefficient values let us know how much the mean of the

dependent variable changes when there is a 1-unit shift in the independent variable while all

other variables in the model are held constant. Listed below shows the significant coefficients in

the model. We can make assumptions on how these coefficients effect the response variable

based on the positive and negative numbers per coefficient.

|  |  |
| --- | --- |
| **Coefficient** | **Estimate** |
| MsSubClass | -100.4992 |
| LotArea | 0.7184 |
| OverallQual | 15279.9626 |
| OverallCond | 6579.6008 |
| YearBuilt | 428.9194 |
| MasVnrArea | 32.4241 |
| TotalBsmtSF | 25.9663 |
| GrLivArea | 47.3152 |
| BedroomAbvGr | -14906.0238 |
| KitchenAbvGr | -27864.4605 |
| TotRmsAbvGr | 9115.8338 |
| Fireplaces | 6198.0837 |
| GarageArea | 46.9568 |

The last step to interpet the estimated model is to obtain the multiple R-squared, or . The

is always between (0,1). This represents the proprtion of the variance in the response

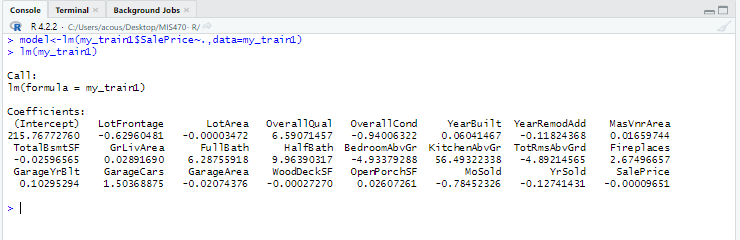
variable in the regression model that can be explained by the predictor variables. In this case, the

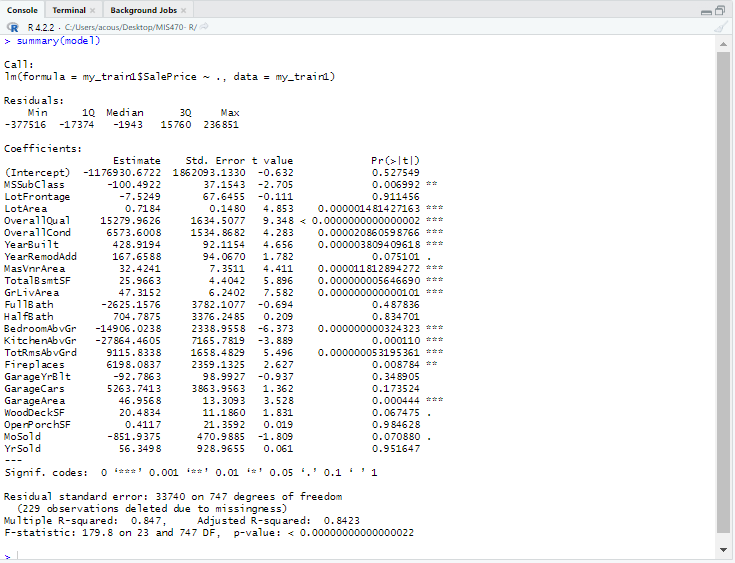
This explains, 84.7% of the variability of the variability of the dependent variable is

explained by the variability of the independent variables, or 84.7% of the variability of the

dependent variable is explained by the regression model. Typically, an shows a high

correlation between variables.





**R Studio Step 6:**

Screenshot using the head (), complete.cases () and predict () function. The head () function is

used to display the first (n) rows in the data frame. In this case, we are looking for the first 20

rows from the housing.testing.csv file. Using the R command-head(my\_test1,20) displays the

first 20 rows of the data frame, although some rows have missing values labeled as (NA). To

remove the missing values, a new dataset is created called my\_test2 using the complete.cases

() function to remove the rows with missing values, and display the first 20 rows with complete

rows from the housing.testing.csv file. Using R command- my\_test2<

my\_test1[complete.cases(my\_test1),] creates the new my\_test2 data frame. Using the head ()

function with R command - head(my\_test2,20) displays the 20 full rows. The next step is using

the predict () function. This function predicts future home sales prices based on the input data.

Using R command- predict (model, head(my\_test2,20)) displays 20 predicted sale prices.

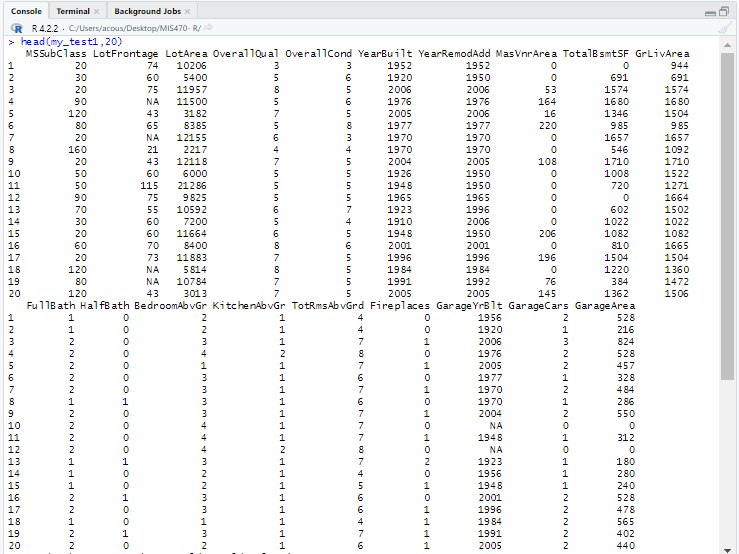
Comparing the testing dataset with the 20 new predicted sale prices, we can see all of the

predictions are within one standard deviation from the testing dataset. Row 5 has the largest

difference in prediction from $181,000 in the testing dataset, with predicted sale price of

$235,819.32. Row 1 has the second largest difference in prediction from $82,000 in the

testing dataset, with the predicted sale price of $39,816.42.



**Table

Description automatically generated with medium confidence**

Table

Description automatically generated

Graphical user interface, table

Description automatically generated

A picture containing Word

Description automatically generated

**R Studio Step 7:**

To measure the accuracy of the 20 predictions, we can use the mean absolute percentage error

(MAPE). Kim & Kim (2016) states the MAPE is one of the most widely used measures to

forecast accuracy due to its advantages of scale-independency and interpretability. MAPE helps

verify how accurate the predictions, or forecast are based on the final MAPE percentage, which

is the prediction error percentage. An easy way to show this equation is stated as MAPE = (1 /

sample size) x ∑ [(|actual - forecast|) / |actual| ] x 100. In this case, the sample size is 20

observations. A good rule of thumb when interpreting the MAPE percentage is obtaining a

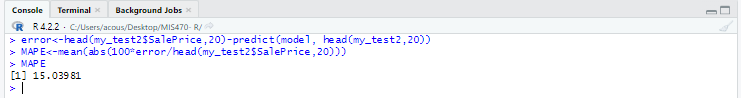
percentage of 20% or less which is considered good for predictions and indicates over a whole

time period, predictions were 20% away from the actual values. In this case, MAPE =

15.03981%, which verifies the accuracy for future predictions for variable-SalePrice in

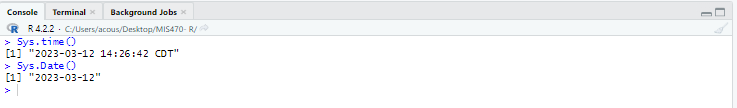
the my\_test2 dataset showing predictions are approximately 15.04% away from the actual sale

price values.



**R Studio Step 8:**

Screenshot using Sys.time () and Sys.Date () function to display completed assignment date and time.

****

References

Allright, S. (2022, August 15th). What is a good MAPE score? <https://stephenallwright.com/good-mape-score/>

CN, P. (2022, August 3rd). How to use predict function in R programming. <https://www.digitalocean.com/community/tutorials/predict-function-in-r>

Frost, J. (2023). How to interpret p-values and coefficients in regression analysis. <https://statisticsbyjim.com/regression/interpret-coefficients-p-values-regression/#:~:text=Interpreting%20Linear%20Regression%20Coefficients&text=A%20positive%20coefficient%20indicates%20that,dependent%20variable%20tends%20to%20decrease>

Kim, S. & Kim, H. (2016, September). A new metric of absolute percentage error for intermittent demand forecasts. International journal of Forecasting. *Volume32*(3). Pp. 669-670. <https://doi.org/10.1016/j.ijforecast.2015.12.003>

RDocumentation. (2023). Combine: combine R objects with a column labeling the source. <https://www.rdocumentation.org/packages/gdata/versions/2.18.0/topics/combine>

Siegel, A. F. & Wagner, M. R. (2022). Histograms: 3.4 skewed distribution and data transformation. *Practical Business Statistics (8th Edition).* <https://www.sciencedirect.com/topics/mathematics/skewed-distribution>

Statology. (2020, April 6th). How to calculate MAPE in R. <https://www.statology.org/mape-r/>

University of Colorado Boulder. (2023). Multiple linear regression. <https://www.colorado.edu/amath/sites/default/files/attached-files/lesson12_multregression.pdf>