PORFOLIO PROJECT

Sonya E. Ghebrechristos

MIS470: DATA SCIENCE FOUNDATION

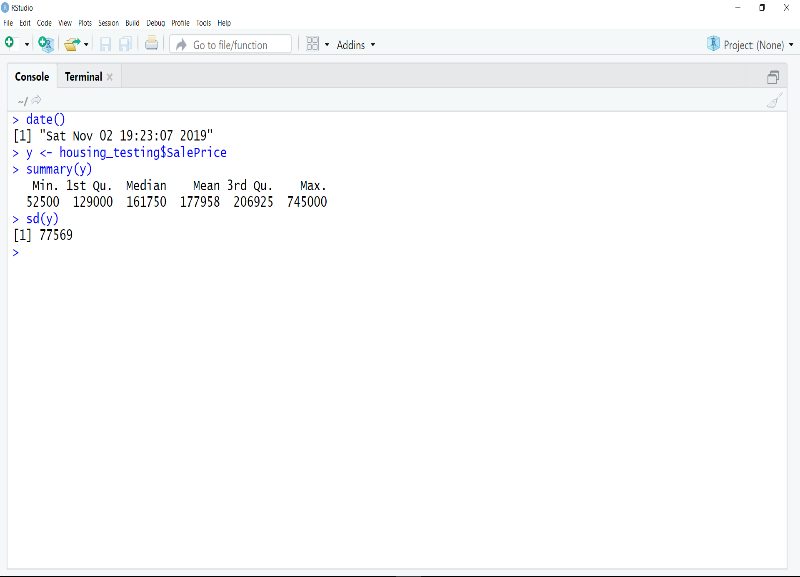
CSU-Global

Chung, Steve

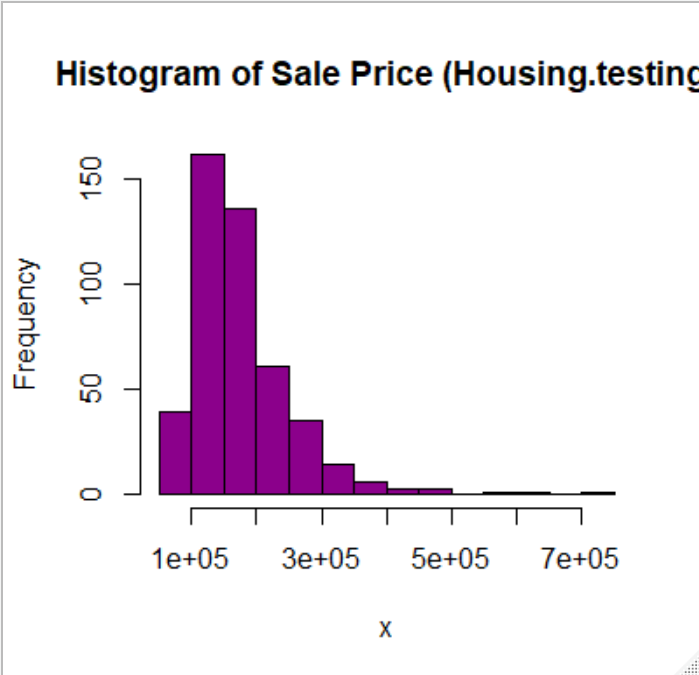
11/03 /2019

1. Using R, calculate the summary of statistics (min, max, mean, median, and sd) and create a histogram of sale price for housing\_training and housing\_testing dataset.

Figure1 is the output of summary statistics and standard deviation (sd) for housing.testing, and figure 2 is the histogram of housing.testing dataset.

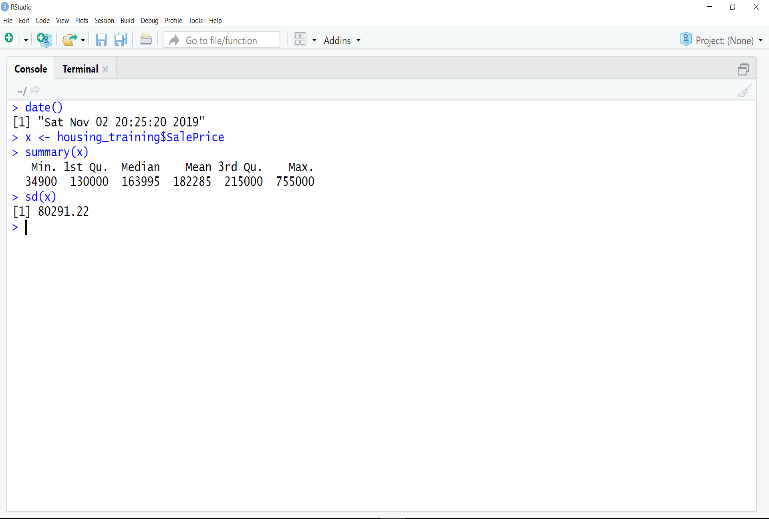
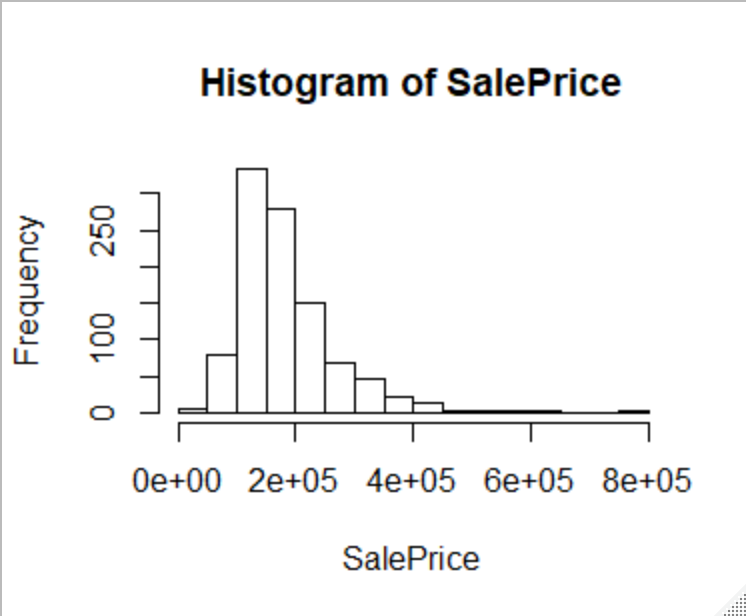


*(housing. testing)Figure1: Summary statistics (of SalePrice)*



*(housing.testing)Figure2:histogram*

Figure 3 is the output of summary statistics and standard deviation for housing.training dataset and figure 4 is the histogram of housing.training dataset.

*(**housing. training)Figure3: summary statistics and sd of saleprice Figure4: histogram of sale price*

The similarity between housing.training and housing.testing dataset

Summary statistics (min, max,mean, median) and standard deviation

The minimum value of sale price of housing.testing and housing.training is a bit difference. The maximum value is however lands within the similar numbers 745000/755000.

The mean of SalePrice, that is very helpful in determining the conclusion. Mean plays apart in avoiding error from the sample population used. the mean of sale price (for both dataset) was calculated using the sum of the data values in a saleprice dataset divided by the number of values (n) in a dataset. both dataset mean is greater than the median. looking at the figure2/figure4 the mean land on the second bar that shows that people are more likely to buy a house that cost 1e+04 to 2e+05. In housing.training and housing.testing dataset, both mean and median skew to minimum value of dataset values, can conclude that less evenly distributed. the majority buyers buy housing less than the median/mean. Different of factors might contribute such as the year of the house, model etc, because buyers buy a house costs less than the median/mean.

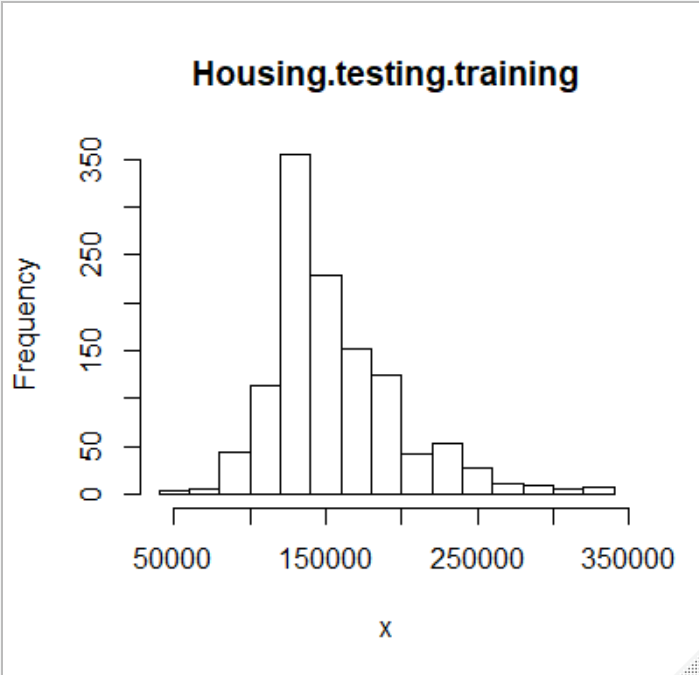
Difference between (housing.training and housing.testing datateset)

The standard deviation of the sale price tell how measurements for a group are spread out from the mean. The standard deviation of the sale price (housing training) is 80291.22, using the minimum and maximum value, the standard deviation is low. That shows that most of the dataset values are closer to the mean. The standard deviation of housing.testing is 77569, which is different from the housing training sd value. The minimum value of both dataset is difference, that might play apart on the difference of the sd value for both dataset. both have similar histogram, but if looking at figure4 the histogram of housing.training it skew more to left than the housing.testing.

Shape of the distribution(histogram)

A frequency distribution is used in sale price histogram graph, that displays how often and outcome (sale price) occurs for the sample. In housing.training and housing.testing,the highest sale occurs around 300 frequency. The highest sales occur between the minimum and the mean. The lowest sales occur between the median and the maximum. The curve of the graph smoothly used without space, that shows the dataset values are accurately given. The distribution of both dataset is skewed to the right. That means the picture is not symmetric around the mean, which is greater than the median. The most sale price occurs median and mean. Fewer data values are tail on the right side that shows majority of population from sample given buy a house cost less than 4e+05.

1. Combine(), two datasets(housing.training.csv and housing.testing.csv). create histogram of sale prices and compare it with the histograms from training and testing datasets.

 *Figure5: istogram(housing.tesitng.training)*

**The Similarities**

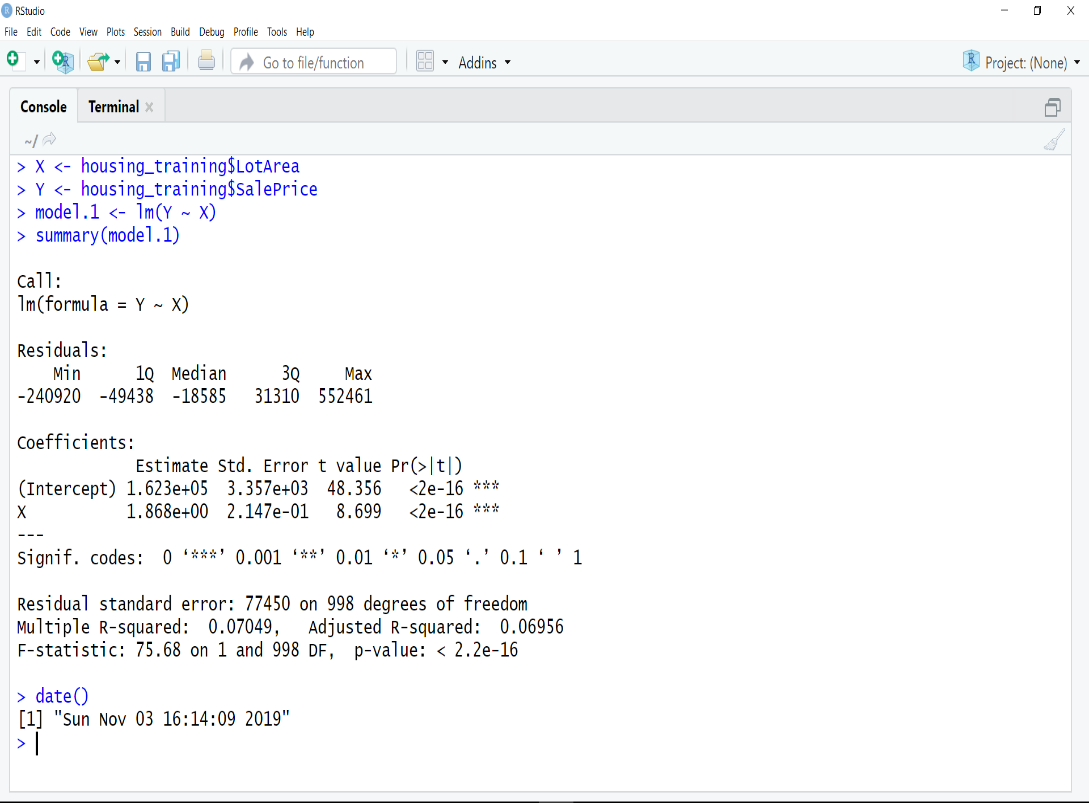
The similarities between the histogram of the combine datasets with the housing.training histogram are in housing.training and combine datasets the highest sale occurs around 300 frequency. The highest sales occur between the minimum and the mean. The lowest sales occur between the median and the maximum. The distribution of both datasets is skewed to the right. In both histogram the most sale price occurs median and mean. Fewer data values are tail on the right side that shows majority of population from sample given buy a house cost less than 350000. In both histograms, it is safe to say that the sale price is around the same area.

The combine datasets and the housing.testing histogram has similar relationship with the housing.training. The highest sale occurs around 300 frequency, which is between the minimum and the mean. The lowest sales occur between the median and maximum values. The distribution of both is skewed to the right. In the combine and testing histogram, sale occur in the minimum vale in the comparison of the training histogram.

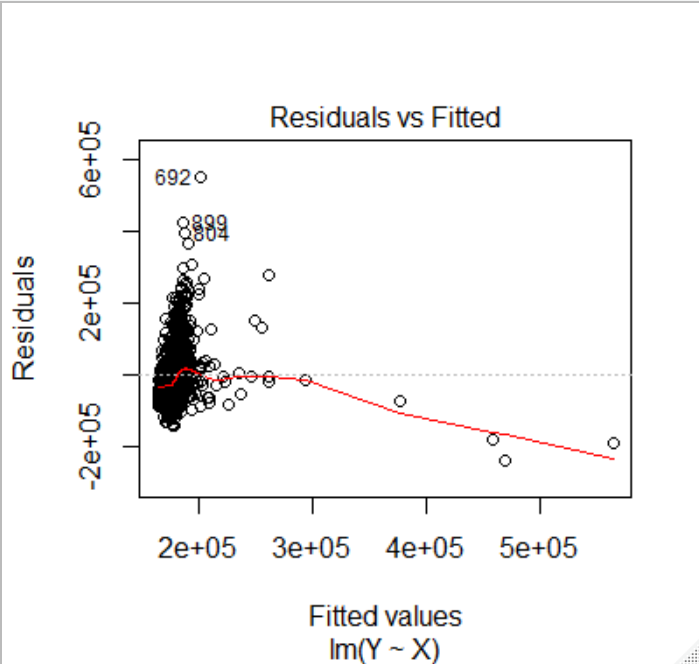
**The Differences**

From the histogram of the combine datasets with testing and training, there is less differences. training and testing histogram are skewed to the left just like the combine histogram. The only difference that I look is around the mean value, there are more sales in the combine datasets in comparison with training and testing histogram. Even though less sales occurs from mean to the maximum values but there are more sales in the combine histogram between the men and maximum values than there are in testing and training histogram.

1. Using only the dataset *housing.training.csv*, fit a linear regression model using all the explanatory variables and SalePrice as the response variable.



*fitted(LotArea vs saleprice)*

*Figure6: Scatterplot of LotArea vs SalePrice*

1. What are the significant factors? How do these variables relate to the sale price? Interpret your estimated model.

> X <- housing\_training$LotArea

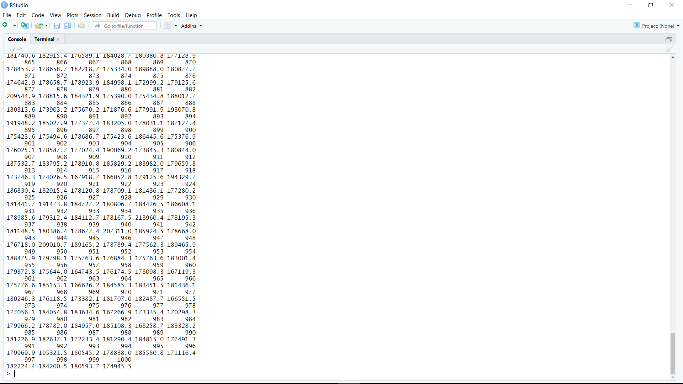
> Y <- housing\_training$SalePrice

> model.1 <- lm(Y ~ X, data= housing\_training)

> print(coefficients(model.1))

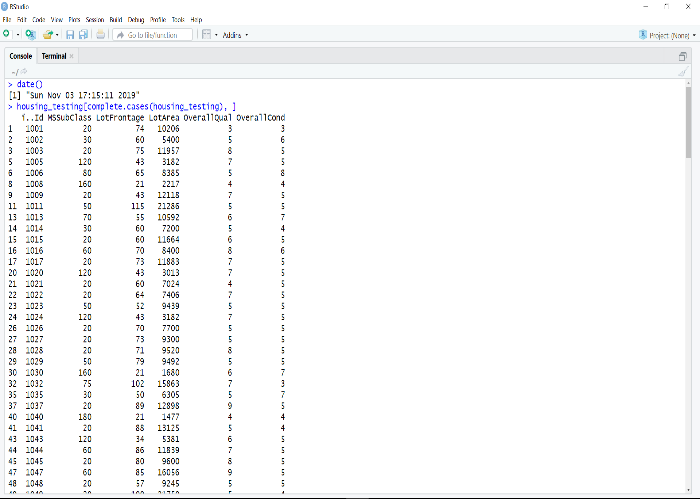
(Intercept) X

1.623153e+05 1.867817e+00

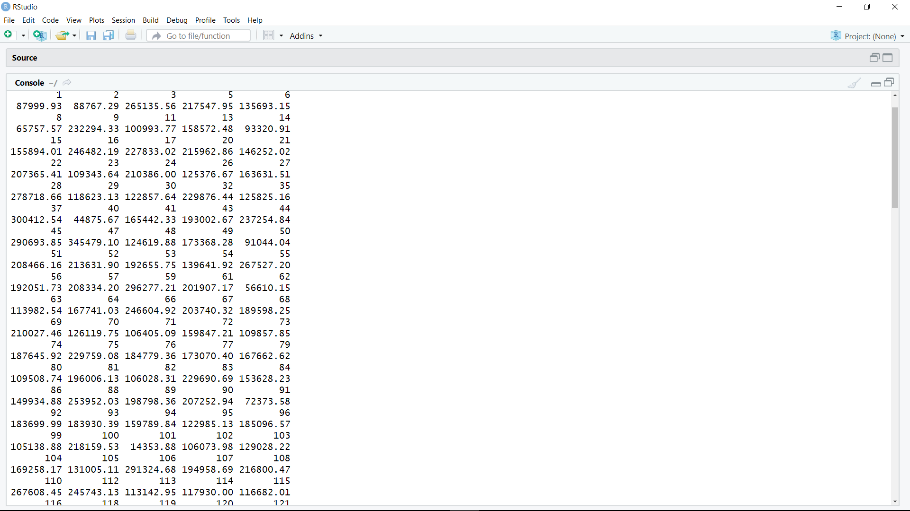
 *fitted(LotArea vs saleprice)*

A fitting linear regression helps in results to predicting values close to the observed data values. In linear regression, statistics are used in least square regression to evaluate model fit. Least square increase as predictors are added to the regression model. The figure above the fitted value and the residuals value that helps predict and plot the lotarea, saleprice of housing.training dataset. the residual standard error is a measure of the precision of a model prediction. The estimate standard error of the housing training with explanatory variables and SalePrice as the response variable is 1.623+05, with explanatory variable (lotarea) that is less than 0.5. the standard error result can help predict of lot area makes a difference in sale price of the housing. This the result is less than 0.5, it is safe to say that the lot area of the housing might make a difference in sale price of the housing. The plot of lot area, and sale price with rest of values given in the housing dataset, it shows that lot area of the house cost around 2e+05. Around the sale price above the mean, the relationship of lot area and sale price is hard to determine.

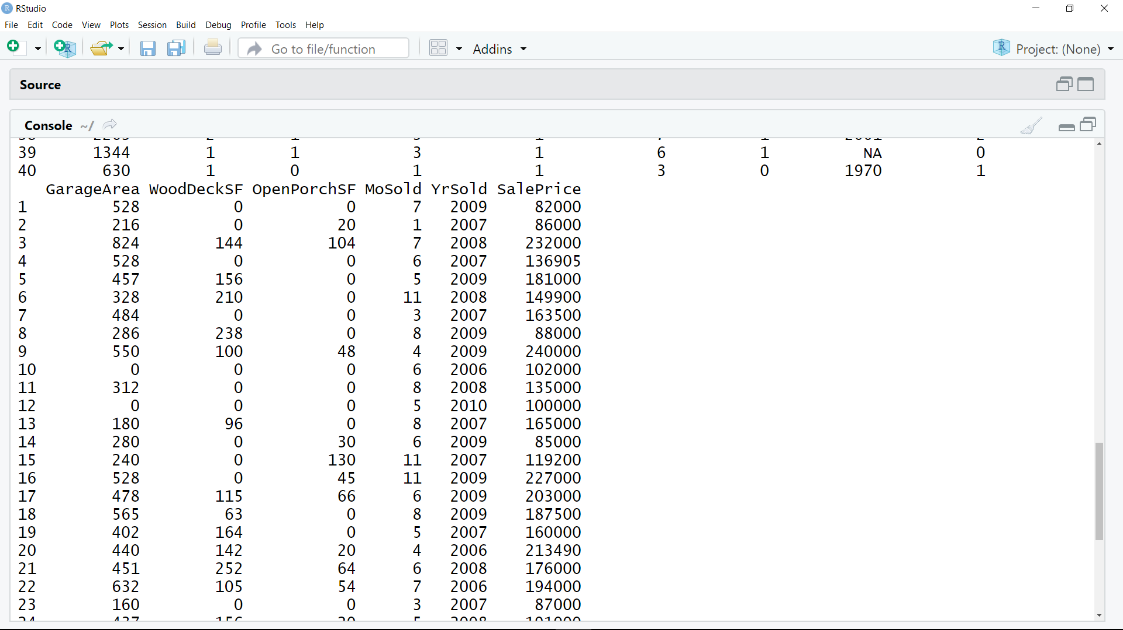
1. Remove all the rows with missing values (NA) from the dataset *housing.testing.csv*. The function *complete.cases()* can be used.

 *Remove all the rows with NA*

Using only the first 20 rows from *housing.testing.csv*, predict the sale price.  The R function *predict()* can perform this task. You should have 20 predicted sale prices.

*prediction*

1. Compare the predicted sale prices to the actual sale prices from the *housing.testing.csv* dataset (the first 20 rows). How good is your prediction?

 *actual saleprice*

The actual sale price of housing.testing.csv and the prediction are similar. The numbers land within few differences. It is safe to say that the well predicted. if we take the GarageArea with the relationship of sale price, the numbers are similar with the 20row of the dataset. with the given values of the actual dataset, we can predict the sale price of the housing with different factors such as lotarea, garagearea etc of the housing. with varies factors that makes a difference in the sale of the house, we can focus on the changes that needs to make in order to keep the housing sale going.

References:

Chen, Jiong, and Cheng, Guo. “Computer Aided Regression Analysis of Synergistic Effect of Energy Industry.” *Applied Mechanics and Materials* 543-547.Vehicle, Mechatronics and Information Technologies II (2014): 4667–4670. Web.

ZyBook – Module1 , module 2, Module 3, Module 4, and Module 5

Knapp, Herschel. *Logistic Regression.* Thousand Oaks, United States: SAGE Publications, Inc., 2017. Film.

Koch, R. (2015). [From business intelligence to predictive analytics](https://search-proquest-com.csuglobal.idm.oclc.org/docview/1645020434/F358ECAECB534FFCPQ/1?accountid=38569). *Strategic Finance, 96*(7), 56-58.

*Histograms and Distributions*. Place of publication not identified: Annenberg Learner Firm,, 2013. Film.

Bannister, Hugh et al. “Multiperiod Mean-Standard-Deviation Time Consistent Portfolio Selection.” *Automatica* 73 (2016): 15–26. Web.