REGRESSION ANALYSIS TO PREDICT HOUSE PRICES IN SAN JOSE 2017

Final Report

Jay Patel

Group Mate: Van Anh Le

EE102

Professor Birsen Sirkeci

2. Abstract

The goal of this project is to predict the house prices of the San Jose city in summer 2017 based on the data collected from the summer 2016. The collected data is used to create several linear and non-linear models to predict the best house price in San Jose for summer 2017.

3. List of Figures and Tables

4. Glossary

Number of Bedroom (Bedrooms)
Number of Bathroom (Bathrooms)
Number of Stories (Stories)
number of Garage (garage space)
Postal Code (Zip code)
House Size [Sq Ft] (HouseSize in square foot)
Lot Size [Sq Ft] (in square foot)
Complex/House Age (Age)
Total Units in Complex (Units)
Sale Price: House prices in dollars
_train: training data, 90% of the whole data
_test: test data, 10% of the whole data
mdl: model

5. Introduction and Background

The project is going to predict the house prices for the San Jose City for summer 2017. As the house prices are on the rise, the prediction of house prices can help house buyers save money on buying house in San Jose. The project is important as it predicts the house prices in San Jose based on very reliable statics model, and it also helps consumer save money.

There are linear models that already exists in the market to predict the house prices in future. The linear model uses linear combination of the most relevant data variables to predict the house prices.

6. Discussion of Data

1. The data was collected from online sources such as Zillow.com, trulia.com, realtor.com, etc. Most of the data collected from the sources stated above was in the range of the months from May 2016 to September 2016.

2. The input variables that we observed to be the most important to decide the price of a house in San Jose are the following.
Number of Bedroom, Number of Bathroom, Number of Stories, number of Garage, Postal Code, House Size [Sq Ft], Lot Size [Sq Ft], Complex/House Age, Total Units in Complex.
The output variable was chosen to the Sale Price which accounts for the rice of the house based on the input variables.
In total, we collected the data for 110 house for nine input variables and one output variable. The data was collected for the house that were sold in summer 2016.

3. The data of the houses in San Jose regarding the nine-input variable and one output variable was collected one house by one house through the online retail sources. The data collected from the sources was saved directly in an excel file Part A.

7. Analysis

Part B:

1. For each variable collected in the part A, the summary of statics such as mean, median, mode, minimum, maximum, variance, standard deviation was calculated and put in the table shown below.

Table 1: Summary of Statistics for The Collected Data

	Houses	Mean	Median	Mode	Minimum	Maximum	Variance	Standard Div
Input Variables	# of Bedroom	3.449541284	3.5	4	1	6	1.032775855	1.016255802
	# of Bathroom	2.353211009	2	2	1	7	0.73983564	0.860136989
	# of Stories	1.348623853	1	1	1	5	0.351403306	0.592792802
	# of Garage Space	1.403669725	1	2	0	5	0.832083814	0.912186282
	Postal Code	95125.62385	95125	95124	95110	95148	57.88321799	7.608102128
	House Size [Sq Ft]	1801.724771	1626	2124	824	6880	811953.9697	901.0848849
	Lot Size [Sq Ft]	8300.458824	6200	6000	436	78000	103159389	10156.74106
	Complex Age	47.79411765	46.5	45	0	110	447.8301615	21.16199805

	Total Units in Complex	1	1	1	1	1	0	0
Output Variable	Sale Price	1150529.559	902500	1350000	410000	11990000	1.38386E+12	1176377.374

As shown in the table above, the average bedroom size of the houses is about 3.5 and the average bathroom size of the houses in San Jose is 2.35. Most of the houses are 1 stories buildings and contain about three garage space per two houses. The average postal code is 95125, and the average house size is 1800 square feet. The average lost size is 8300 square feet which seems to be too high. The average complex age is 48 years, and most of the houses have only one unit in the complex. The average price of the houses in San Jose in summer 2016 based on the collected data is 1150529 dollars which also seems very high.

The median and mode of the data describes the middle value of the data and the most repeated values of the data respectively. Most of the house had four bedrooms and two bathrooms, two garage space and one stories. The common house sizes are between 1600 to 2100 square feet and the lot sizes are common at around 6200 square feet. The Lot sizes here seems better and more intuitive here than that of the average. All the houses have only one house per unit.

The minimum and the maximum values for each input variables were also calculated. The minimum values for the input variables was about 1 or zero for bedrooms, bathrooms, stories, garage and units. The minimum house price based on the minimum specs of the input variable was observed to be 410000 dollars. Similarly, the maximum values of the input variables were observed to be 5,6, or 7 for the bedrooms, bathrooms, stories, garage and units. The maximum house price was observed as 11990000 dollars which is very high compare to the stander of the San Jose City.

The variance means the deviation of a random variable form the mean. The variance for the garage, stories, bathrooms and bed rooms is around 0.5 to 1 which is expected and conceivable. The postal code varies by 58 from the mean which seems little too high. House sizes in squared feet along with lot sizes varies a lot from the mean with 811954 and 103159189 numbers respectively. The house prices also varies a lot with the margin of E+12 which is very high.

The standard deviation which represents the variance of a random variable for the whole collected data seems to be very high for the house size, lot size, and house prices. The postal code deviation seems fine in the stander deviation. The bedrooms, stories, bathrooms and garage data also seems fine and conceivable based on the overall look of the data.

2. The data was cleaned based on the observation made in the section 1 of the part B. As the lot size and the house prices of the house seemed very high, some of the most expensive house with a huge lot size were removed from the data. Also, the house with very low house prices and very minimum specifications were also removed to make the data more consistent. Over all, out of 110 rows of data, 8 rows of outlier data were removed.

Table 2: Summary of Statistics for The Collected Data After Removing Outliers

	Houses	Mean	Median	Mode	Minimum	Maximum	Variance
Input Variables	# of Bedroom	$3	3	4	1	6	0.895520953
	# of Bathroom	$2	2	2	1	4	0.445045175
	# of Stories	$1	1	1	1	5	0.339869281

	# of Garage Space	$1	1	2	0	5	0.769800077
	Postal Code	$95,126	95125	95125	95110	95148	56.85313341
	House Size [Sq Ft]	$1,748	1628	2124	824	3940	413684.4517
	Lot Size [Sq Ft]	$7,621	6192.5	6000	436	78000	80322212.53
	Complex Age	$49	48	45	0	110	429.9001346
	Total Units in Complex	$1	1	1	1	1	0
Output Variable	Sale Price	$1,027,809	920250	1350000	503000	1877777	1.18364E+11

3. The scatter plot Metrix of the nine input variables and one output variable was created using the METLAB command plotmetrix. The plotmetrix command took all the variables as array and created the graph as shown below.

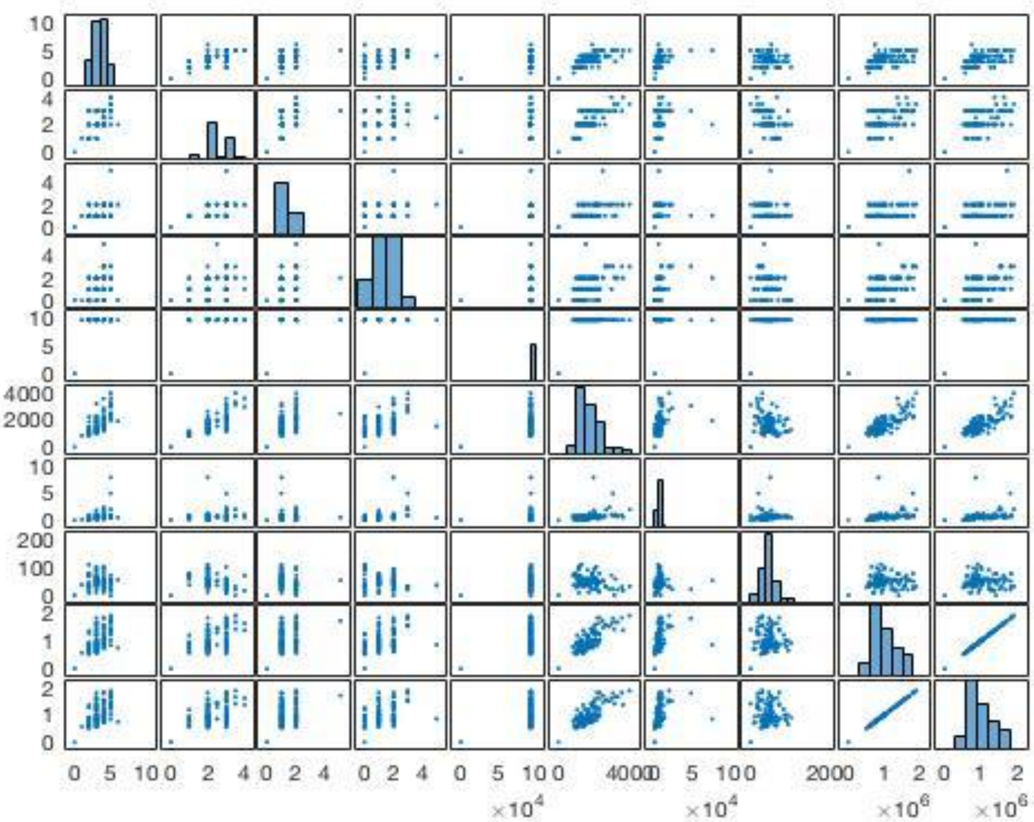

Figure 1: Scatter Plot Matric for All Input and output variables

The scattered plot matric shown above is for the variables Number of Bedroom, Number of Bathroom, Number of Stories, number of Garage, Postal Code, House Size [Sq Ft], Lot Size [Sq Ft], Complex/House Age, Total Units in Complex and Sale Price from top to bottom respectively and left to right respectively.

The scatter plot is used to visually see the correlation between two different variables. Therefore, the scatter plot matric will help observe the correlation between each variable. Most of the graphs for the variables bedrooms, bathrooms, stories, and garages looks like horizontal or vertical scattered bars. As the number of bedrooms, bathrooms, garage space, stories, postal code are discrete numbers, the graph looks like a scattered bar graph. The more the scattered the bars are, the less correlated the variables are. Therefore, the relation between variables bedrooms, bathrooms, stories, garage and postal code look very weak.

However, the relation between the variable house size, lot size, age and sale price looks strong as the graph of these variable against all the other variables looks denser and less scattered.

4. The correlation coefficient was calculated as shown in the table 3 below.

Table 3: Correlation Coefficient Respect to Sale Price

	Houses	Cov (sale price)	Variance	(Variance)^(1/2)	Corr. CoEffi.
Input Variables	# of Bedroom	$142,307	0.895520953	0.946319689	0.437096694
	# of Bathroom	$102,433	0.445045175	0.667117062	0.446301931
	# of Stories	$52,229	0.339869281	0.582983088	0.260403813
	# of Garage Space	$117,324	0.769800077	0.877382515	0.388676274
	Postal Code	-$83,073	56.85313341	7.540101684	-0.032023814
	House Size [Sq Ft]	$172,508,632	413684.4517	643.1830624	0.779588697
	Lot Size [Sq Ft]	$832,002,104	80322212.53	8962.266038	0.269833518
	Complex Age	-$580,996	429.9001346	20.73403324	-0.081447653
	Total Units in Complex	$0	0	0	0
Output Variable	Sale Price	$118,364,479,973	1.18364E+11	344041.3928	1

In table 3, the second column shoes all the input and output variable names. The third column shoes the values of the covariance of each variable in column 2. The variance of the variables is also shown in the table 3. The square root of the variance is also shown in the table since it is necessary to calculate the value of correlation coefficient.

Correlation coefficient varies between the value -1 and 1. If the correlation coefficient is closer to 1, the variables are highly correlated and the increase in one variable implies increase in another variable. Similarly, If the correlation coefficient is closer to - 1, the variables are highly correlated and the increase in one variable implies decrease in another variable. The value of 0 for correlation coefficient suggest there is no correlation between variables.

The correlation between the sale price and the house size seems to be very high as the correlation coefficient is 0.779 which is very close to the value 1. Also, the correlation of sale price with bedrooms and bathrooms also seems to be high as the values 0.437 and 0.446 seems to be closer to the 1. The variables complex age and postal code have negative correlation values which means as the age and postal code increases, the house price decreases. The other variable is not strongly correlated as the correlation coefficient is close to zero.

We chose the following three variables as most relevant variables for predicting house prices for 2017 using linear model.

Table 4: Selected Variables

Variable	Correlation Coefficient respect to Sale Price
House Size	0.779588697
Bathrooms	0.446301931
Bedrooms	0.437096694

The house size, bedrooms and the bathrooms are the variable selected as the most relevant variables as they are highly correlated to the house prices. Their correlation coefficient respect to the sale price is shown in the table 4.

Table 5: Correlation Between the Variable

Corr. CoEffi.	# of Bedroom	# of Bathroom	House Size [Sq Ft]
# of Bedroom	1	0.520621789	0.635271385
# of Bathroom	0.520621789	1	0.726386736
House Size [Sq Ft]	0.635271385	0.726386736	1

Part C:
1. Our data set had, 102 data rows of the nine input and one output variable. To choose the 90% of the data for the training, the first 92 data rows were chosen as the training data rows and the last 10 rows were chosen as the test data rows as stated in the excel file part C.2.

All the calculations for the linear model was done in the MATLAB using fitlm command which generates a linear model based on the input variables' data and output variable's data.

2. The linear models for the different combination of the input variable is shown below.

Linear Regression Model 1:

Input Variable: Bedrooms (90% of the data)
Output Variable: House Price (90% of the data)

```
mdl =


Linear regression model:
  SalePrice ~ 1 + Bedrooms

Estimated Coefficients:
              Estimate       SE       tStat      pValue
             __________  __________  ________  __________


  (Intercept)    4.508e+05    1.2427e+05    3.6274    0.00047441
  x1                34863    4.7601    7.3513e-06
```

```
Number of observations: 92, Error degrees of freedom: 90
Root Mean Squared Error: 3.05e+05
R-squared: 0.201,  Adjusted R-Squared 0.192
F-statistic vs. constant model: 22.7, p-value = 7.35e-06
```

Model equation: House Price = (4.508e+05) + (1.6595e+05) * Bedrooms

Linear Regression Model 2

Input Variable: Bathrooms (90% of the data)
Output Variable: House Price (90% of the data)

```
mdl2 =


Linear regression model:
  SalePrice ~ 1 + Bathrooms

Estimated Coefficients:
                Estimate        SE        tStat       pValue
               __________    __________    ______    __________

  (Intercept)    5.1068e+05    1.1533e+05    4.4279    2.6676e-05
  x1            2.1805e+05        47191    4.6205    1.2713e-05


Number of observations: 92, Error degrees of freedom: 90
Root Mean Squared Error: 3.07e+05
R-squared: 0.192,  Adjusted R-Squared 0.183
F-statistic vs. constant model: 21.3, p-value = 1.27e-05
```

Model equation: House Price = (5.1068e+05) + (2.1805e+05) * Bathrooms

Linear Regression Model 3

Input Variable: House Size (90% of the data)
Output Variable: House Price (90% of the data)

```
mdl3 =


Linear regression model:
  SalePrice ~ 1 + HousePrice

Estimated Coefficients:
                Estimate      SE      tStat       pValue
               __________    ______    ______    __________

  (Intercept)    3.1868e+05     65730    4.8483    5.1776e-06
  x1                403.06    35.305    11.416    3.4435e-19
```

```
Number of observations: 92, Error degrees of freedom: 90
Root Mean Squared Error: 2.18e+05
R-squared: 0.592,  Adjusted R-Squared 0.587
F-statistic vs. constant model: 130, p-value = 3.44e-19
```

Model equation: House Price = (3.1868e+05) + (403.06) * House Size (in Sq Ft)

Linear Regression Model 4

Input Variable: Bedrooms, Bathrooms (90% of the data)
Output Variable: House Price (90% of the data)
mdl4 = fitlm(tb1)

```
mdl4 = 


Linear regression model:
  SalePrice ~ 1 + Bathrooms + Bedrooms

Estimated Coefficients:
                  Estimate        SE        tStat      pValue
                 __________    __________    _______    _________

  (Intercept)       3.1763e+05    1.3151e+05    2.4152     0.017773
  Bathrooms_train    1.3766e+05         54083    2.5453      0.01264
  Bedrooms_train      1.108e+05         40189     2.757    0.0070763


Number of observations: 92, Error degrees of freedom: 89
Root Mean Squared Error: 2.96e+05
R-squared: 0.255,  Adjusted R-Squared 0.239
F-statistic vs. constant model: 15.3, p-value = 2.01e-06
```

Model equation: House Price = (3.1763e+05) + (1.3766e+05) * Bathrooms + (1.108e+05) * Bedrooms

Linear Regression Model 5

Input Variable: Bedrooms, House Size (90% of the data)
Output Variable: House Price (90% of the data)

```
mdl5 = 


Linear regression model:
  SalePrice ~ 1 + Bedrooms + HouseSize
```

```
Estimated Coefficients:
                 Estimate      SE      tStat       pValue
                ___________    ______    ________    __________

  (Intercept)       3.6188e+05     89623      4.0378    0.00011409
  Bedrooms_train        -22936     32242    -0.71138       0.47871
  HouseSize_train       423.57    45.661      9.2763    1.0058e-14


Number of observations: 92, Error degrees of freedom: 89
Root Mean Squared Error: 2.19e+05
R-squared: 0.594,  Adjusted R-Squared 0.585
F-statistic vs. constant model: 65.1, p-value = 3.87e-18
```

Model equation: House Price = (3.6188e+05) + (423.57) * House Size + (-22936) * Bedrooms

Linear Regression Model 6

Input Variable: Bathrooms, House Size (90% of the data)
Output Variable: House Price (90% of the data)

```
mdl6 =

Linear regression model:
  SalePrice ~ 1 + Bathrooms + HouseSize

Estimated Coefficients:
                 Estimate      SE      tStat       pValue
                ___________    ______    ________    __________

  (Intercept)       4.5935e+05     78788     5.8301    8.7532e-08
  Bathrooms_train   -1.419e+05     47679    -2.9761     0.0037587
  HouseSize_train       513.27    50.177     10.229    1.0787e-16


Number of observations: 92, Error degrees of freedom: 89
Root Mean Squared Error: 2.09e+05
R-squared: 0.628, Adjusted R-Squared 0.62
F-statistic vs. constant model: 75.3, p-value = 7.3e-20
```

Model equation: House Price = (4.5935e+05) + (513.27) * House Size + (-1.419e+05) * Bathrooms

Linear Regression Model 7

Input Variable: Bathrooms, House Size, Bedrooms (90% of the data)
Output Variable: House Price (90% of the data)

```
mdl7 =

Linear regression model:
  SalePrice ~ 1 + Bathrooms + Bedrooms + HouseSize

Estimated Coefficients:
                          Estimate       SE      tStat       pValue
                         ___________    ______   ________   __________

  (Intercept)             4.7659e+05    94873      5.0234   2.6367e-06
  Bathrooms_train        -1.3967e+05     48394    -2.8861    0.0049053
  Bedrooms_train            -10329      31297    -0.33002      0.74217
  HouseSize_train           520.77      55.323     9.4133   5.7811e-15

Number of observations: 92, Error degrees of freedom: 88
Root Mean Squared Error: 2.1e+05
R-squared: 0.629, Adjusted R-Squared 0.616
F-statistic vs. constant model: 49.7, p-value = 6.84e-19
```

Model equation: House Price = (4.7659e+05) + (520.77) * House Size + (-1.3967e+05) * Bathrooms + (-10329) * Bedrooms

3. Test models using 10% of the data:

Table 6: MSE error and test data predicted prices

Model Equation 1	Model Equation 2	Model Equation 3	Model Equation 4	Model Equation 5	Model Equation 6	Model Equation 7
1.61201E+11	1.62649E+11	72082487173	1.80554E+11	65111924610	43207171751	41009860483
2.62097E+11	3.88486E+11	49919838598	2.51459E+11	54397141763	32515548150	34547479219
12943612900	2371963603	5104772200	7083059476	2230526632	5441173124	3839371621
96411253156	95297424278	15287886937	82416947591	14612711854	11089317955	10878617833
1.90294E+11	3.62762E+11	27506217014	2.62833E+11	25953707224	947255849.7	957438158.5
38916727578	973615288.8	23889970565	14599394626	20615426084	45094608223	42421678479
4.93406E+11	43186994835	1.20321E+11	2.70326E+11	83659912271	1.66903E+11	1.43965E+11
82434890710	1.18116E+11	95241607133	86186241526	1.01461E+11	98647798221	1.01632E+11
4830474541	4583670345	8602583619	2123690837	8887757926	11028124080	11136768595
34310152900	16561973080	8900448804	5605346489	12514465380	38470287437	41122154052

MSE Error:

1.37844E+12	**1.19658E+12**	**4.54231E+11**	**1.16625E+12**	**1.16625E+12**	**4.71788E+11**	**4.53037E+11**

The table 6, shows the predicted house prices based on the model equations derived in model 1 through 7. The MSE error, was calculated based on the sale price obtain through the test data and the sale price obtain through the original data. The MSE error for each model is shown in the last row of the table 6 above.

Based on the observation, the model 7 which include linear combination of all three selected variables performs the best with minimum error in the house price.

Model 7 with all the three variables, house size, bathrooms, bedrooms performs the best as the MSE error is the lowest among the seven tables.

Perdition of the house prices in 2017 based on the best model 7 is shown below in table 7.

Table 7: House Price Prediction for Summer 2017 Based on Model 7

Bedrooms	Bathrooms	House Size [Sq Ft]	Predicted Price for 2017 [$]
1	1	800	727857.6818
2	1	950	796949.4899
2	2	1050	716199.8364
3	1	1100	866041.2979
3	2	1200	785291.6444
3	3	1300	704541.9909
3	3.5	1600	798514.1378
4	2	1800	1096208.005
4	3	2000	1069197.141
4	4	2500	1203402.645
5	3	2900	1541329.87
5	4	3500	1729274.164

Part D:
To derive heuristic model to predict the prices better than the liner model, we decided to incorporate an additional variable which is the multiplication of the variables house size and complex age. Therefore, there were four input variables were used to derive the heuristic model. The four variables are house size, bedrooms, bathrooms and multiplication of house size and complex age.

The following table 8 shows the summary of the for all the data set including the new variable which is multiplication of house size and complex age.

Table 8: Summary of Statics for Heuristic model

Houses	**# of Bedroom**	**# of Bathroom**	**House Size [Sq Ft]**	**Complex Age**	**Age*HouseSize**	**Sale Price**
Mean	3.460784314	2.318627451	1748.137255	49.23529412	81713.44118	1027808.971
Median	3	2	1628	48	71629	920250
Mode	4	2	2124	45	-	1350000
Minimum	1	1	824	1	3094	503000
Maximum	6	4	3940	110	210888	1877777
Variance	0.895520953	0.445045175	413684.4517	428.9446367	1534727954	1.18364E+11
Standard Div	0.946319689	0.667117062	643.1830624	20.71097865	39175.60407	344041.3928
Cov (sale price)	142306.9449	102433.3819	172508631.5	-577346.9637	6099277170	1.18364E+11
Variance	0.895520953	0.445045175	413684.4517	428.9446367	1534727954	1.18364E+11
(Variance)^(1/2)	0.946319689	0.667117062	643.1830624	20.71097865	39175.60407	344041.3928

Table 9: Correlation Coefficient respect to Sale Price for Heuristic model

Houses	**# of Bedroom**	**# of Bathroom**	**House Size [Sq Ft]**	**Complex Age**	**Age*HouseSize**	**Sale Price**
Cov (sale price)	142306.9449	102433.3819	172508631.5	-577346.9637	6099277170	1.18364E+11
Variance	0.895520953	0.445045175	413684.4517	428.9446367	1534727954	1.18364E+11
(Variance)^(1/2)	0.946319689	0.667117062	643.1830624	20.71097865	39175.60407	344041.3928
Corr. CoEffi.	0.437096694	0.446301931	0.779588697	-0.081026218	0.452534788	1

Based on the correlation coefficient table sown above, the variable HouseSize * Age is highly correlated with the house price as its coefficient is 0.45.

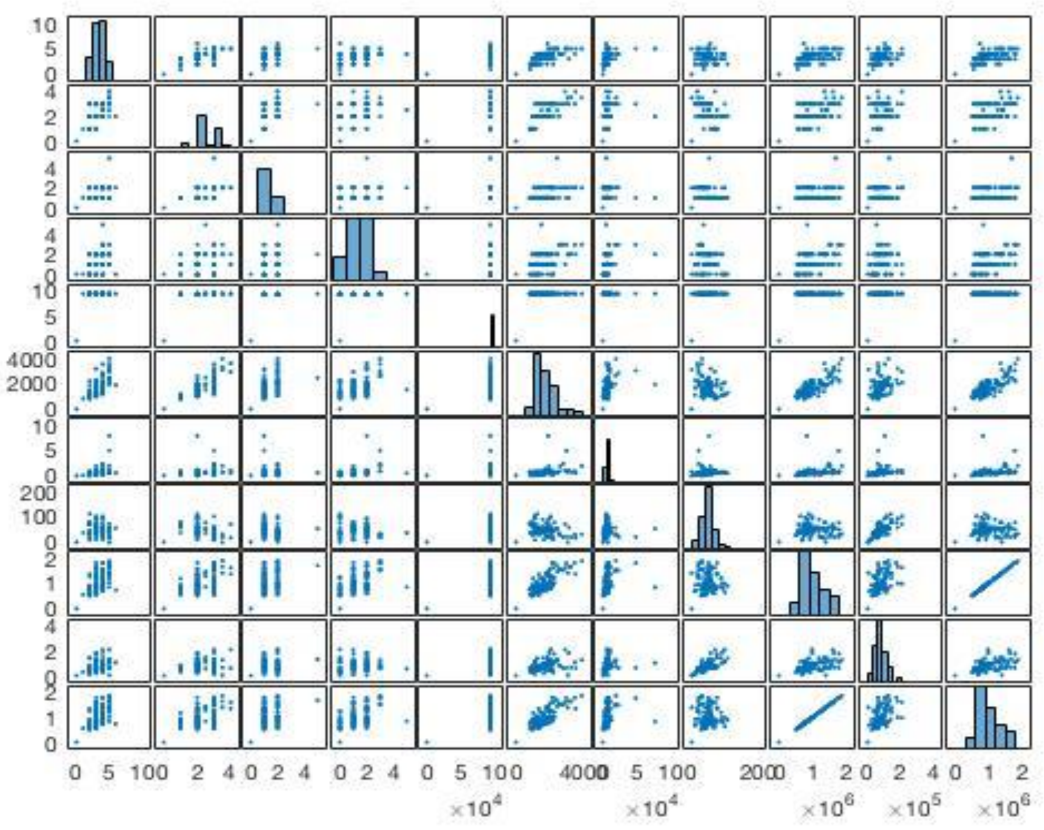

Figure 2: Scatter Plot Matric for Heuristic Model

The scatter plot matric shown in the figure 2 displays the similar scatter plot as in figure 1 except the variable House Size * Complex Age is added to it right before Sale price variable.

Table 10: Correlation between Selected variables

# of Bedroom	1	0.520621789	0.635271385	0.2877766
# of Bathroom	0.520621789	1	0.726386736	0.161204217

House Size [Sq Ft]	0.635271385	0.726386736	1	0.382540688
Age*HouseSize	0.2877766	0.161204217	0.382540688	1

From table 10, the variable Age*HouseSize is not highly correlated with the other three variables since the correlation coefficient is less than 0.3 in most cases. However, there is a weak correlation between the house size and the new variable as the correlation coefficient is close to 0.38.

Linear Regression Model 8

```
mdl8 =


  SalePrice  ~ 1 + Age * HouseSize

Estimated Coefficients:
                   Estimate      SE      tStat      pValue
                 ___________  ________  _______  ___________

  (Intercept)       6.8077e+05     68901    9.8803     5.064e-16
  HouseSize 4.2829    0.77248    5.5443    2.9208e-07


Number of observations: 92, Error degrees of freedom: 90
Root Mean Squared Error: 2.95e+05
R-squared: 0.255,  Adjusted R-Squared 0.246
F-statistic vs. constant model: 30.7, p-value = 2.92e-07


Model equation: House Price = (6.8077e+05) + (4.2829) * Age * HouseSize
```

Linear regression model 9

```
mdl9 =
  SalePrice ~ 1 + HouseSize * Age + HouseSize + Bedrooms+ Bathrooms

Estimated Coefficients:
                   Estimate       SE       tStat       pValue
                 ____________  ________  _________  ___________

  (Intercept)         4.09e+05     94306       4.337    3.8729e-05
  Age_Hsize_train        1.6986    0.59702      2.8452      0.005533
  HouseSize_train        465.73     56.626      8.2247    1.727e-12
  Bedrooms_train         -21179      30348    -0.69789      0.48711
  Bathrooms_train    -1.1176e+05      47575     -2.3492      0.021081
```

Number of observations: 92, Error degrees of freedom: 87
Root Mean Squared Error: 2.02e+05
R-squared: 0.661, Adjusted R-Squared 0.645
F-statistic vs. constant model: 42.3, p-value = 1.15e-19

Model equation: House Price = (4.09e+05) + (465.73) * House Size + (-1.1176e+05 * Bathrooms + (-21179) * Bedrooms + (1.6986) *Age * HouseSize

Table 11: MSE Error Calculation Data For Model 1-9

Sale Price	**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**	**Model 7**	**Model 8**	**Model 9**
1350000	948650	946780	1074014.44	925350	1086842.18	1137417.98	1142185.98	1114182.348	1166612.546
1788999	1280550	1164830	1533905.9	1284610	1524263.55	1581159.05	1576056.55	951941.8135	1479547.809
898000	782700	946780	831775.38	814550	855212.61	828942.71	839532.21	1280504.487	973851.248
638000	948650	946780	770107.2	925350	767470.4	750412.4	749525.4	1170048.496	837608.664
1549000	1114600	946780	1360590.1	1036150	1365064.45	1502352.95	1502124.45	1046122.785	1449575.123
915500	1114600	946780	1063131.82	1036150	1052469.79	1123559.69	1117796.19	1092116.848	1124107.648
738888	1446500	946780	1078045.04	1257750	1022269.88	1142550.68	1116406.68	1132633.082	1115050.414
1399888	1114600	1055805	1083284.82	1104980	1073648.29	1078273.19	1073999.69	1306368.92	1176486.613
879000	948650	946780	793484.68	925350	792037.46	780182.06	779730.06	988530.6282	792630.9388
599000	782700	728730	705214.54	676890	722211.63	809675.93	815680.43	972388.3781	817173.0454
MSE Error	**1.37844E+12**	**1.19658E+12**	**4.54231E+11**	**1.16625E+12**	**4.26625E+12**	**4.71788E+11**	**4.53037E+11**	**1.78E+12**	**4.74856E+11**

Based on the MSE error calculation shown in the table 11 above, the heuristic model 8 and 9 has grater error than the liner model 6 and 7; therefore, the heuristic model does not predict the price of the house in San Jose city better than the liner model.

Linear model 7 which includes the linear combination of three variables, house size, bedrooms and bathrooms, performs the best as it has the least MSE error value.

Table 12: Predicted House Prices for Summer 2017 Based on Heuristic Model

Bedrooms	Bathrooms	House Size [Sq Ft]	Age	Predicted Price for 2017 [$]
1	1	800	45	709794.6
2	1	950	40	761872.3
2	2	1050	30	685644.4
3	1	1100	50	839429

3	2	1200	25	731777
3	3	1300	40	703959.2
3	3.5	1600	45	821770.2
4	2	1800	40	1061377.2
4	3	2000	45	1073338
4	4	2500	40	1211429
5	3	2900	45	1540109.3
5	4	3500	40	1723924

8. Results

Based on the analysis of the data, the linear model derived based on the most related three variables performs the best. The heuristic model derived based on non-linear combination comes very close to the best performed linear model 7. However, the heuristic model fails to outperform the linear model 7. The house prices predicted based on the best performed linear model 7 is shown in table 7. The house price prediction in the table 7 is the most accurate prediction of the prices of houses in San Jose based on the analysis done in this project.

9. Conclusion

The project was done successfully with the help of the tools like excel and MATLAB. The project was very useful in terms of understanding the concepts of the statistics. Doing linear regression gives us a good knowledge of doing data analysis in future. The challenges that we faced in the project was the use of MATLAB to generate the linear models and matric scatter plot. Overall, the project experience was excellent and we learned new concepts of the statistics.

10. Team Member’s Performance

My team mate was Van Ahn le. She was very helpful and cooperative. We collected all the data with 50% effort from each member. However, the part B, C and D was done by me completely. My partner did make sure that all the work that I have done is correct by re calculation. Overall, it was a great experience working with Van Ahn Le.

11. References

Zillow, Inc. "Zillow: Real Estate, Apartments, Mortgages & Home Values". *Zillow*. N.p., 2017. Web. 17 May 2017.