Big Data Analytics Project: AIT-580
Navya Mote

House Prices Data Set from Kaggle

This data set was put together from a larger real time data set known as The Ames Housing data. The original data set had 3970 observations with 113 variables. Some of the observations belonged to stand-alone garages, condos, and storage areas which was not relevant to the analysis of house prices. Therefore, these observations were not included. The house prices data set also contains recent sales data on a property when compared to the original data which had approximately 100 houses with changed ownership multiple times in a period of 4 years.

The data set is already split into a training and a test set. The training set is 449 KB and the test set is 440 KB in size. Training set has 1460 observations whereas the test set has 1459 observations. The training set has 81 variables. And the test set has 80(excluding Sale Price). The training set consists of 23 nominal, 23 ordinal, 15 discrete/interval and 20 ratio/continuous variables. The test set consists of 23 nominal, 23 ordinal, 15 discrete/interval and 19 ratio/continuous variables.

Some of the variables are:

Order (Discrete/Interval): Observation number
MS SubClass (Nominal): Identifies the type of dwelling involved in the sale.
MS Zoning (Nominal): Identifies the general zoning classification of the sale.
Lot Frontage (Continuous/Ratio): Linear feet of street connected to property
Lot Area (Continuous/Ratio): Lot size in square feet
Street (Nominal): Type of road access to property
Alley (Nominal): Type of alley access to property
Lot Shape (Ordinal): General shape of property
Year Built (Discrete/Interval): Original construction date
Exter Qual (Ordinal): Evaluates the quality of the material on the exterior
Fireplaces (Discrete/Interval): Number of fireplaces
SalePrice (Continuous/Ratio): Sale price $$[1]

Who (company, agency, organization) collected the data?

The data was collected by Ames City Assessor's Office for sales that occurred in Ames, Iowa between 2006 and 2010[2]

- **Who they are, what do they do?**

  The City Assessor's office exists by city ordinance and is governed by the City Conference Board. This office is responsible for assigning values to all taxable properties within Ames city limits
- **What is their role/purpose?**

[1] Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project, Dean De Cock, Truman State University, Journal of Statistics Education Volume 19, Number 3(2011), www.amstat.org/publications/jse/v19n3/decock.pdf

[2] Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project, Dean De Cock, Truman State University, Journal of Statistics Education Volume 19, Number 3(2011), www.amstat.org/publications/jse/v19n3/decock.pdf

The Assessor's office has an on-going process of gathering and reviewing information related to real estate which is used to provide accurate and current values of a property. New values are assigned in odd-numbered years. The City Assessor also processes applications for homestead credits, veterans' exemptions, industrial property tax abatements, urban revitalization exemptions, and various other programs.[3]This data set was acquired by Professor Dean De Cock of Truman State University for the regression project that would allow his students to display the skills they had learned in class.

Need

- ***Why* did they collect this data?**
  It is the duty of The City Assessor's office to provide significant amount of information regarding property values, selling prices, ownership, and physical characteristics to the public and other city and county departments. The Assessor's office is required to provide the County Auditor with a list of current values. These values are used by the auditor to calculate property taxes for the city, county and school governments. Property valued by the Ames City Assessor has a market value of approximately $4.82 billion for 2017 (which does not include ISU, IDOT, City, County, School, or Federal property) [4]

- What potential *questions* could be answered by studying this data?
  - List some *specific questions*, and be sure to *answer them* in your analysis

  The questions that can be answered by studying this data are:
  1. What information affects the Sale Price of a house?
  2. What is the relationship between the predictor and the response? Is it a positive or a negative correlation?
  3. Are there any outliers in the data set?
  4. Are there any null or missing values in the data set? What needs to be done about them?
  5. How can we try to get the best prediction of the Sale Price? Is there just one or multiple models that can get us the best price?
  6. What parameters can be used to measure the model performance?
  7. Which is the house that costs the most in the dataset? Which the least?
  8. How much does a house with largest square feet cost?
  9. On an average houses belonging to which sale condition and Sale type cost the most?
  10. Which house style on an average costs the most?
  11. Houses in which neighborhood cost the most/least?
- **Are there any privacy, quality, or other issues with this data?**
  Since the data belongs to the government, the author of the paper I referred to had to meet with the Assessor and Deputy Assessor of the city of Ames to seek access and permission.

---

[3] CITY ASSESSOR, 2018 City of Ames, IA, https://www.cityofames.org/government/departments-divisions-a-h/city-assessor

[4] CITY ASSESSOR, 2018 City of Ames, IA, https://www.cityofames.org/government/departments-divisions-a-h/city-assessor

Then the data was shared with the author in the form of a dump from their record system. The initial Excel file contained 113 variables describing 3970 property sales that had occurred in Ames, Iowa between 2006 and 2010. The variables were a mix of nominal, ordinal, continuous, and discrete variables. In order to make the data set easily understood by users of all levels, the author began by removing any variables that required prior knowledge of their use. Most of these deleted variables were related to weighting and adjustment factors used in the city's current modeling system.

Requirements, resources needed

- What software and hardware resources do you need to study this data?
  I would need R, SQL and Tableau as the software resources. As the hardware resource I would need my laptop and mouse.

Present the Results/Findings

- *Explore the dataset* using relevant tools discussed in the course (R, SQL, Python, Tableau, etc)
  - Prepare and describe relevant metadata definitions (define types of data in the dataset)

    The types of data in this file are:

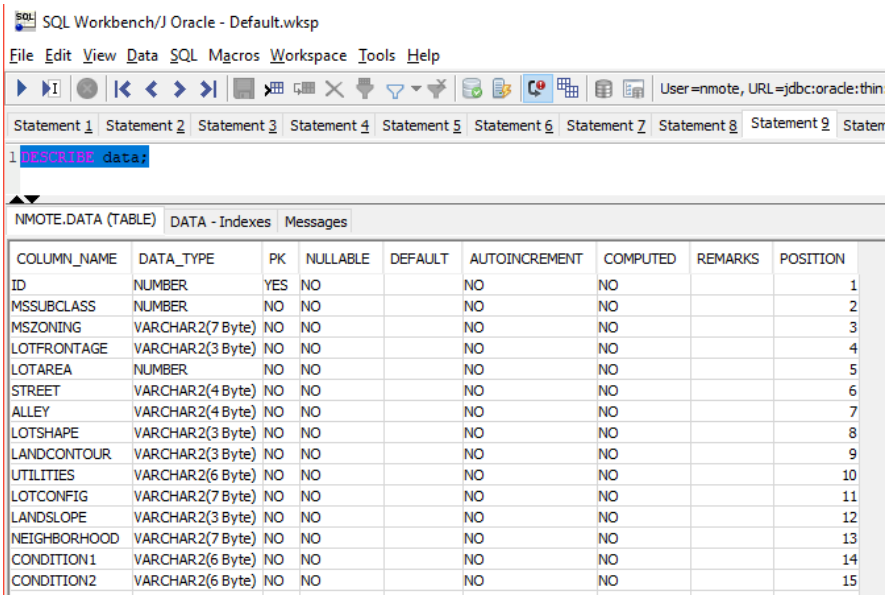
    Nominal : It has two or more categories, but there is no rank order to the categories. For example, Street(Type of road access to property) is a nominal variable having two categories (Grvl-Gravel and Pave-Paved) and there is no rank order to the categories

    Ordinal : An ordinal data is similar to a nominal data. The difference between the two is that there is a clear rank based ordering of the variables. For example, we consider Exter Qual (Evaluates the quality of the material on the exterior). Even though we can order these from lowest to highest(Ex-Excellent, Gd-Good, TA - Average/Typical, Fa-Fair, Po-Poor), the spacing between the values may not be the same across the levels of the variables.

    Interval : An interval data is similar to an ordinal data, except that the intervals between the values of the interval data are equally spaced. For example : Year Built (Original construction date)

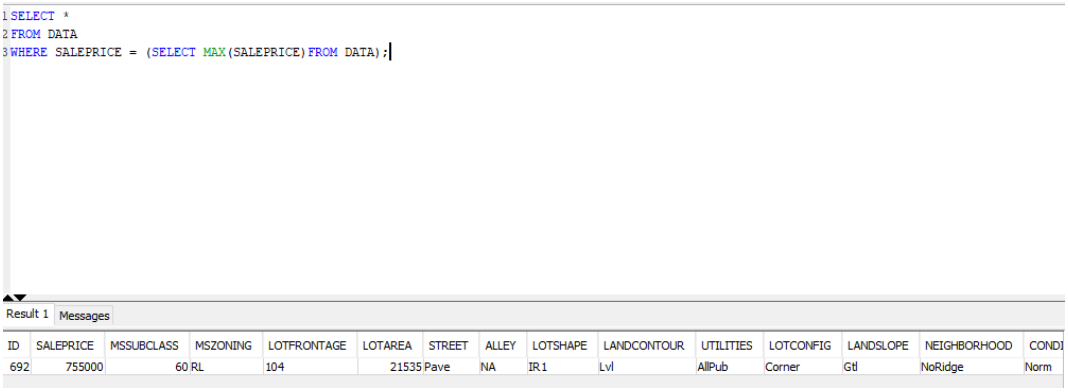
    Ratio : Ratio data is similar to interval data, but with the added condition that 0 (zero) of the measurement indicates that there is none of that variable. For example : Lot Area (Lot size in square feet)

  - include schema for SQL and do some SQL-based data exploration
- I used the training set to do the following SQL analysis:
  My table name is DATA and it has variables belonging to character and numeric data types.



COLUMN_NAME	DATA_TYPE	PK	NULLABLE	DEFAULT	AUTOINCREMENT	COMPUTED	REMARKS	POSITION
ID	NUMBER	YES	NO		NO	NO		1
MSSUBCLASS	NUMBER	NO	NO		NO	NO		2
MSZONING	VARCHAR2(7 Byte)	NO	NO		NO	NO		3
LOTFRONTAGE	VARCHAR2(3 Byte)	NO	NO		NO	NO		4
LOTAREA	NUMBER	NO	NO		NO	NO		5
STREET	VARCHAR2(4 Byte)	NO	NO		NO	NO		6
ALLEY	VARCHAR2(4 Byte)	NO	NO		NO	NO		7
LOTSHAPE	VARCHAR2(3 Byte)	NO	NO		NO	NO		8
LANDCONTOUR	VARCHAR2(3 Byte)	NO	NO		NO	NO		9
UTILITIES	VARCHAR2(6 Byte)	NO	NO		NO	NO		10
LOTCONFIG	VARCHAR2(7 Byte)	NO	NO		NO	NO		11
LANDSLOPE	VARCHAR2(3 Byte)	NO	NO		NO	NO		12
NEIGHBORHOOD	VARCHAR2(7 Byte)	NO	NO		NO	NO		13
CONDITION1	VARCHAR2(6 Byte)	NO	NO		NO	NO		14
CONDITION2	VARCHAR2(6 Byte)	NO	NO		NO	NO		15

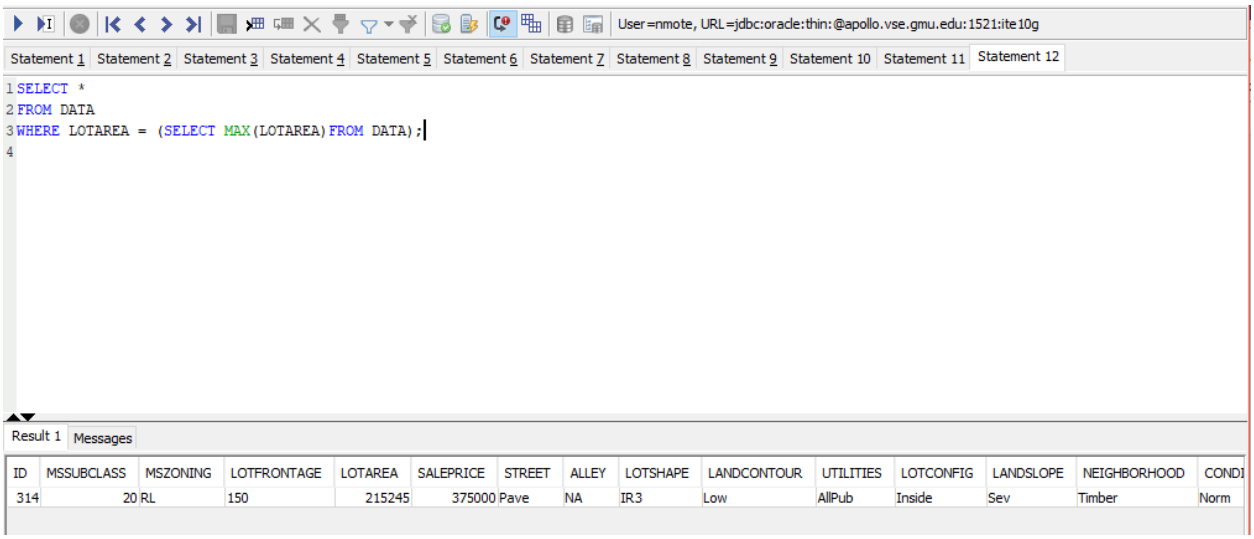
- **Which is the house that costs the most? Which the least?**



ID	SALEPRICE	MSSUBCLASS	MSZONING	LOTFRONTAGE	LOTAREA	STREET	ALLEY	LOTSHAPE	LANDCONTOUR	UTILITIES	LOTCONFIG	LANDSLOPE	NEIGHBORHOOD	CONDI
692	755000	60	RL	104	21535	Pave	NA	IR1	Lvl	AllPub	Corner	Gtl	NoRidge	Norm

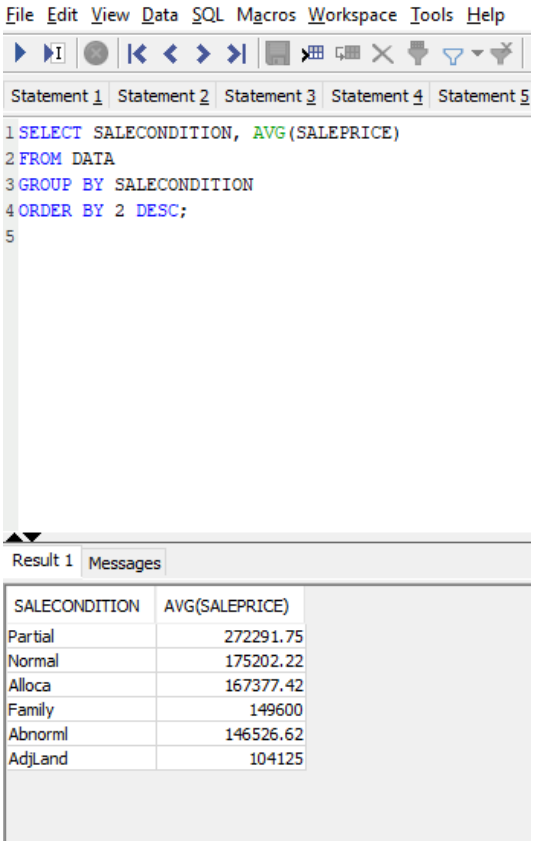
Looking at the result we can say that the costliest house belongs to NoRidge (Northridge)neighborhood, it is a 2 story house, built in 1994, Excellent exterior quality, warranty dead Sale type and Normal Sale Condition with a Sale Price of $755000.

- **How much does a house with largest Lot area cost?**



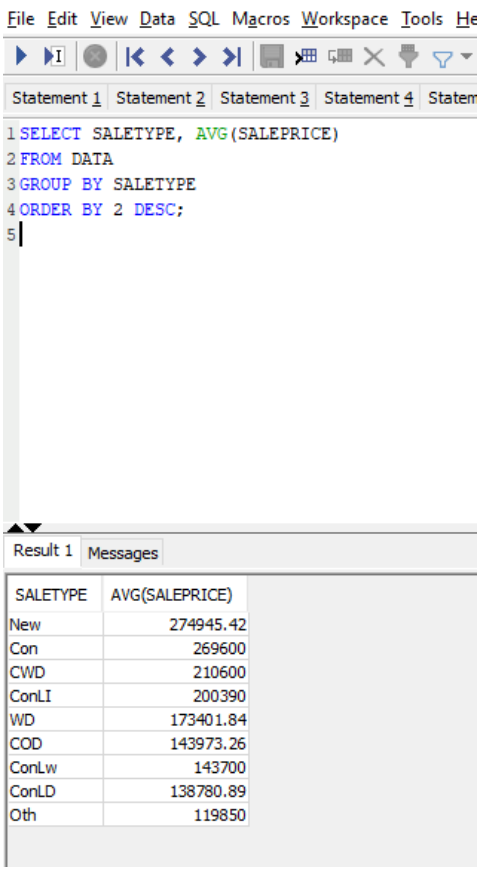
It costs $375000. Which is not the costliest house in the dataset. Which means bigger houses in general are not the costliest. By this we understand that there are also other parameters that help in determining the price of the house.

- **On an average houses belonging to which sale condition cost the most?**



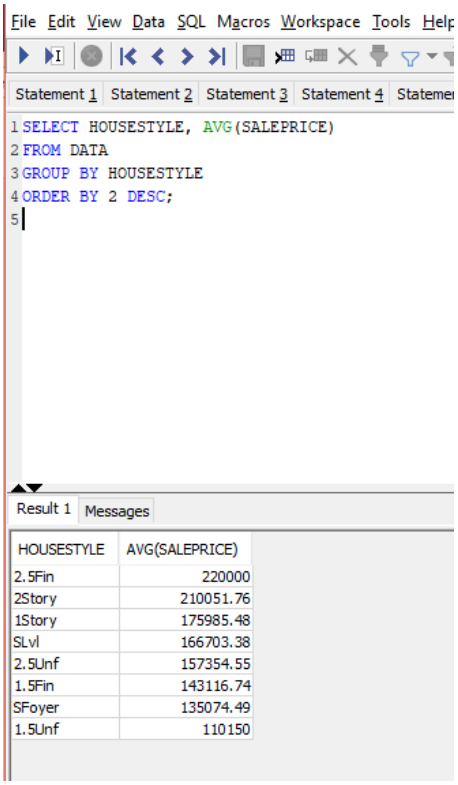
On an average houses belonging to the partial sale condition cost more. Partial stands for ‘Home was not completed when last assessed (associated with New Homes)’. Therefore the new houses cost more.

- **On an average house belonging to which Sale type cost the most?**



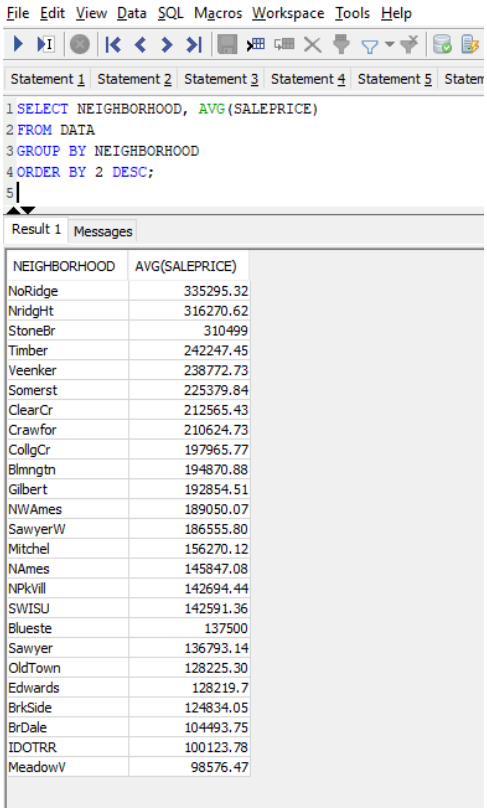
On an average the houses belonging to New sale type cost the most. New stands for 'Home just constructed and sold' which confirms that the new houses cost more. This tells us that there is a strong positive correlation between Sale type and Sale condition

- **Which house style on an average costs the most?**



Houses belonging to 2.5Fin house type cost more in an average. 2.5Fin stands for 'Two and one-half story: 2nd level finished'

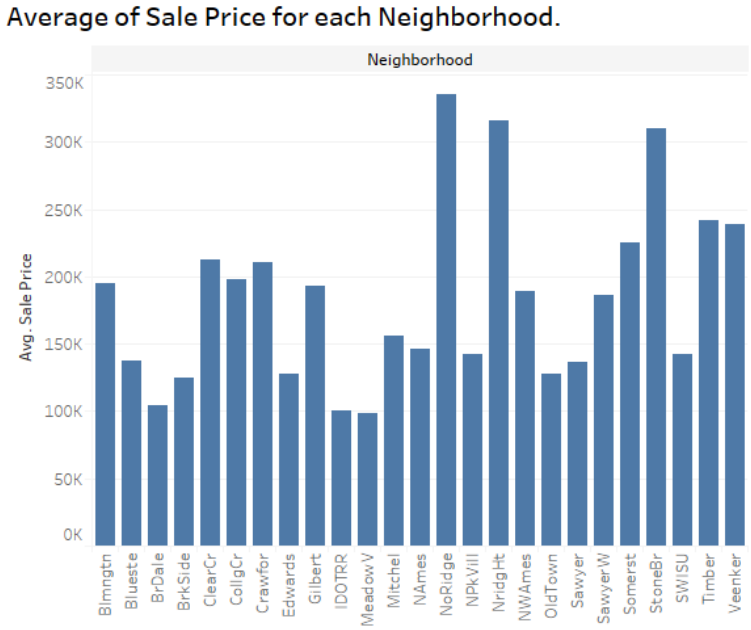
- **Houses in which neighborhood cost the most/least?**



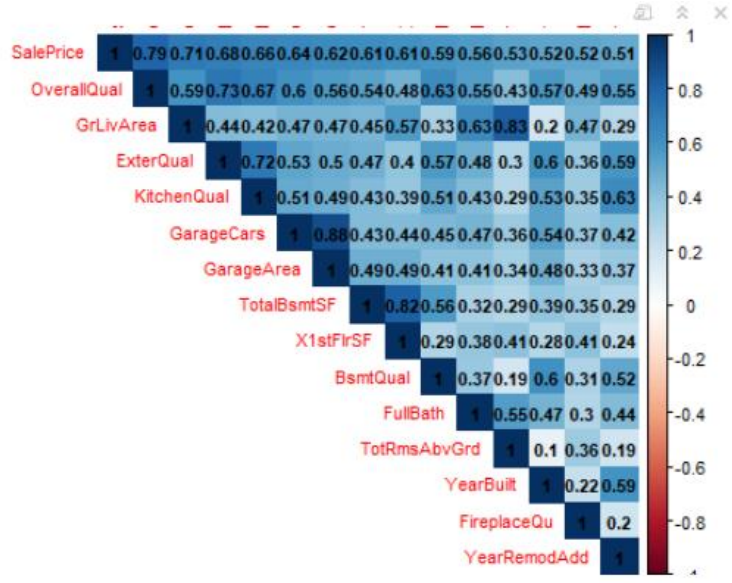
On an average houses in NoRidge(Northridge) cost the most and houses in MeadowV(Meadow Village) cost the least.

- Prepare relevant *descriptive statistics* and *visualizations* for selected data
  - (you *don't need* to analyze *all* the items in the dataset)
  - Graphics must follow good visualization practices discussed in course lectures

The below visualization was created using Tableau and shows us the Average sale price of each neighborhood in the dataset

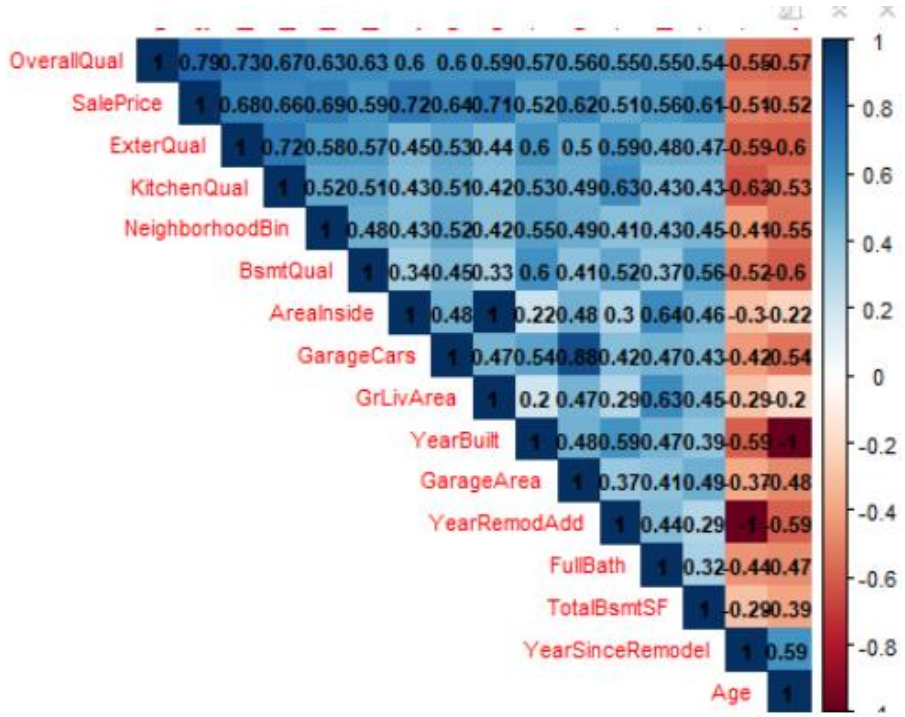


- **What is the relationship between the variable and the response? Is it a positive or a negative correlation?**



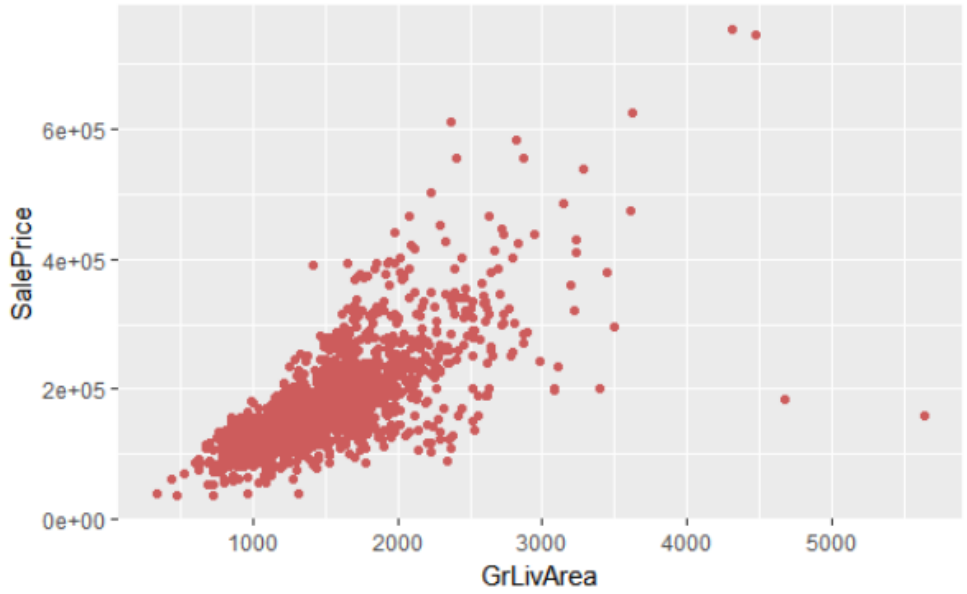
We can observe high positive correlation between the Sale price and the overall quality, living area, external built quality and kitchen quality. We can also make note of the high

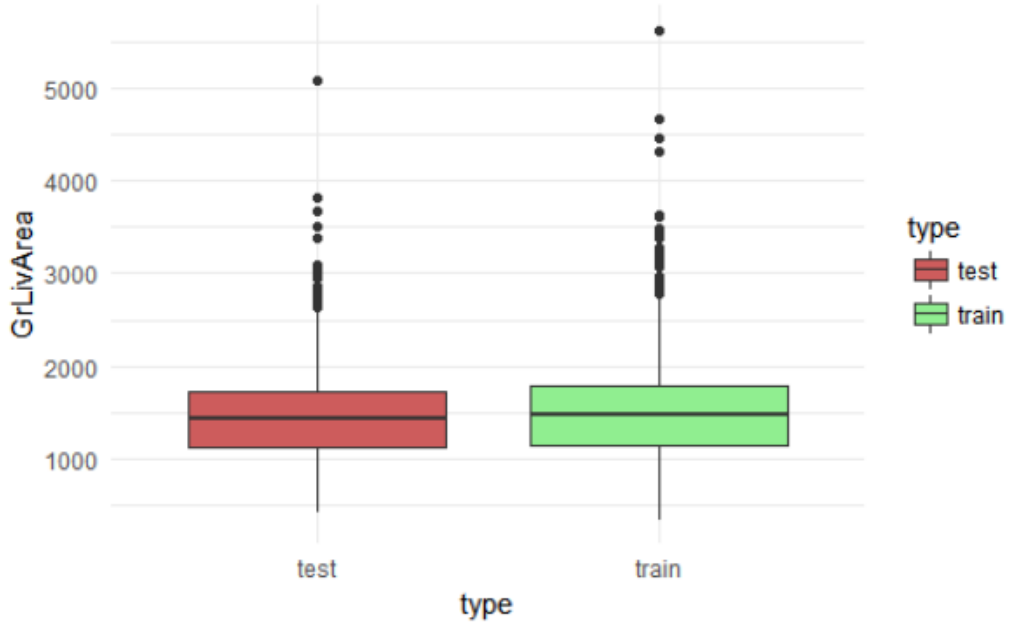
inter variable correlation. The living area and total rooms above ground have a high correlation of 0.83. Which means that they both are trying to say the same thing and are related to each other. Garage cars and garage area have a high correlation of 0.88 which says the no. of cars that fit in a garage depends on the area of the garage. Total basement area and 1st floor square feet are highly correlated as they both are trying to say the same thing.



This matrix consists of all predictors after cleaning and adding new predictors. Here we can observe the negative correlation between year remodeled and year since remodeled. Because as the year increases, the year since remodeled decreases. We can observe a similar relationship between year built and age of the house.

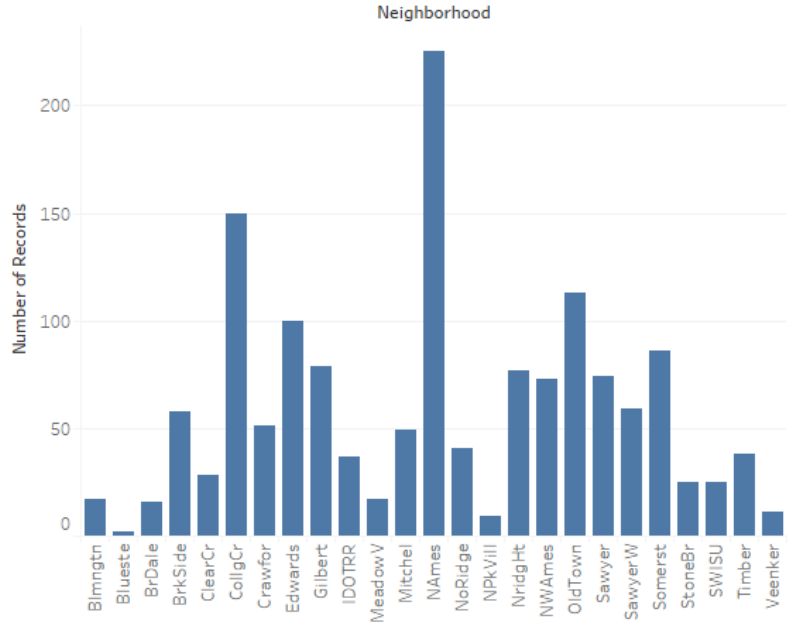
- **Are there any outliers in the data set?**



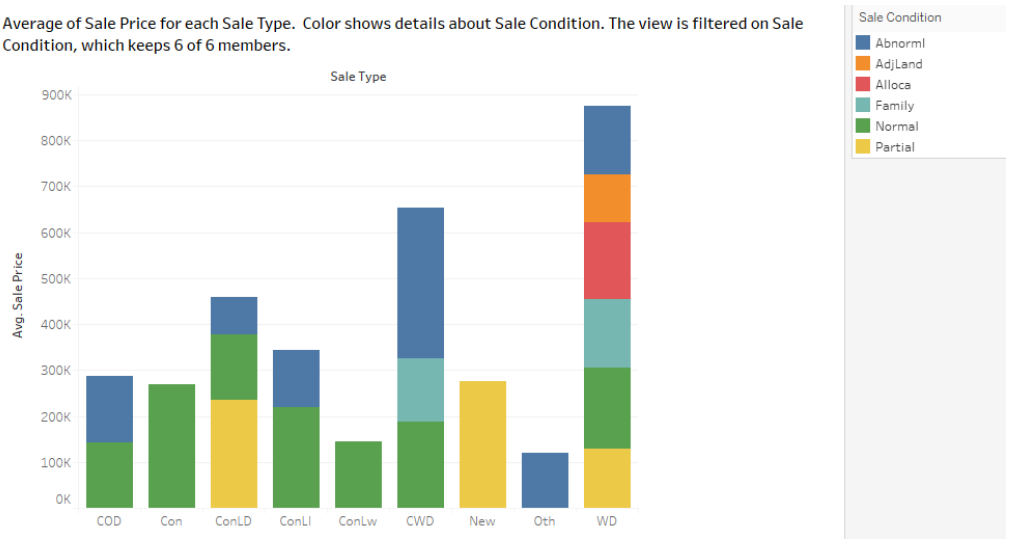


There are 4 houses in the training set having area greater than 4000sqft. This would cause heavy right skewness in Sale Price and the Living Area. Especially the 2 values that have above a 4000 GrLivArea(Above grade (ground) living area square feet) but low SalePrice are putting a constraint on the correlation between the 2 variables. Therefore, I decided to remove these records from my training data.

Sum of Number of Records for each Neighborhood.



In the training dataset, most of the houses are from the Northwest Ames neighborhood. The above graph was generated using Tableau



Warranty deed(a deed that guarantees a clear title to the buyer of real property) is the sale type for which the sale price is more. It consists of all the sale conditions. This graph was generated using Tableau

- *Interpret* the results; what *conclusions* can be supported?
    - This should reflect answers to the *specific questions* specified above
- **Are there any null or missing values in the data set? What needs to be done about them?**
  There are 34 columns with NA's in the dataset.

```
      PoolQC  MiscFeature        Alley        Fence
        2909         2814         2721         2348
 FireplaceQu  LotFrontage  GarageYrBlt GarageFinish
        1420          486          159          159
  GarageQual   GarageCond   GarageType     BsmtCond
         159          159          157           82
BsmtExposure     BsmtQual BsmtFinType2 BsmtFinType1
          82           81           80           79
  MasVnrType   MasVnrArea     MSZoning    Utilities
          24           23            4            2
BsmtFullBath BsmtHalfBath   Functional  Exterior1st
           2            2            2            1
 Exterior2nd   BsmtFinSF1   BsmtFinSF2    BsmtUnfSF
           1            1            1            1
 TotalBsmtSF   Electrical  KitchenQual   GarageCars
           1            1            1            1
  GarageArea     SaleType
           1            1
[1] "There are 34 columns with missing values"
```

Since this a combined dataset of training and test sets, we can see all the missing values here. The missing values could be because the house does not have that feature or it was not made a note of. It would not be a great idea to delete columns with large number of missing values right away as they might be helpful for our prediction. While building our models if we feel that removing one of these columns would lead to a better prediction we can then remove them. For now I imputed the missing values based on predictors that are related to them. For example: PoolQC is the quality rating of the house. If the rating is missing it could be because the house does not have a pool. So we can check against the pool area. If the pool area is greater than 0 we know that the house has a pool but it was not rated. In such cases, I have computed the mean pool area for each rating and depending on the nearest value the area belongs to, assigned a rating. For GarageYrBuilt(Year garage was built) I have imputed the values based on the year the house was built because it is most likely that the garage is built the same year as the house. Similarly I have computed all the other missing values.

- **What information affects the Sale Price of a house?**

  The Partial Least Squares predictive model with k=10 fold cross validation and 20 components gave me the best prediction of the house prices. I did a varImp(plsTune ,scale = FALSE) to get a list of predictors that have a strong effect on the Sale Price. As we can see, the top 5 predictors are : Overall Quality of the house which rates the overall material and finish of the house, The Above grade (ground) living area square feet, The neighborhood(Physical locations within Ames city limits), Size of garage in car capacity and Total square feet of basement area.

	Overall
OverallQual	100.00
GrLivArea	94.75
NeighborhoodBin	82.94
GarageCars	78.70
TotalBsmtSF	78.47
X1stFlrSF	77.11
ExterQual	76.53
GarageArea	76.49
KitchenQual	76.05
Age	71.49
TotRmsAbvGrd	68.70
FullBath	68.62
GarageYrBlt	68.51
YearRemodAdd	64.75
FoundationPConc	62.95
KitchenQualTA	61.49
Fireplaces	61.20
ExterQualGd	60.60
BsmtQualTA	56.56
GarageFinish	55.72

For Linear regression model with 10 folds CV:

lm variable importance

only 20 most important variables shown (out of 115)

	Overall
OverallCond	11.721
OverallQual	10.772
LotArea	7.806
GrLivArea	7.556
FunctionalTyp	6.007
TotalBsmtSF	5.699
NbrhRich	4.547
NeighborhoodSomerst	4.362
SaleConditionAbnorml	4.242
Condition1Norm	4.177
BsmtFullBath	4.106
GarageTypeDetchd	3.766
CentralAirY	3.745
LotConfigCulDSac	3.696
Fireplaces	3.606
HasScreenPorch	3.521
SaleTypeWD	3.438
Exterior1stWd.Sdng	3.314
Age	3.166
FullBath	3.156

For Robust linear regression with 10 folds CV:

rlm variable importance

only 20 most important variables shown (out of 74)

	Overall
PC1	130.358
PC3	52.966
PC4	20.627
PC5	15.305
PC44	11.873
PC6	11.237
PC9	9.861
PC32	9.298
PC30	6.817
PC15	6.631
PC54	6.515
PC46	6.384
PC22	6.057
PC72	5.972
PC16	5.900
PC8	5.636
PC10	5.142
PC43	4.769
PC19	4.686
PC7	4.678

Principal Component Regression (PCR) with 10 folds CV:

```
  only 20 most important variables shown (out of 122)

               Overall
OverallQual     0.6712
GrLivArea       0.5395
NeighborhoodBin 0.4930
ExterQual       0.4637
GarageCars      0.4630
GarageArea      0.4484
KitchenQual     0.4451
TotalBsmtSF     0.4215
X1stFlrSF       0.3873
FullBath        0.3492
Age             0.3469
GarageYrBlt     0.3263
YearRemodAdd    0.3237
KitchenQualTA   0.2905
TotRmsAbvGrd    0.2846
FoundationPConc 0.2820
ExterQualGd     0.2638
BsmtQual        0.2483
GarageFinish    0.2426
Fireplaces      0.2373
```

Ridge regression:

```
  only 20 most important variables shown (out of 122)

               Overall
OverallQual     0.6712
GrLivArea       0.5395
NeighborhoodBin 0.4930
ExterQual       0.4637
GarageCars      0.4630
GarageArea      0.4484
KitchenQual     0.4451
TotalBsmtSF     0.4215
X1stFlrSF       0.3873
FullBath        0.3492
Age             0.3469
GarageYrBlt     0.3263
YearRemodAdd    0.3237
KitchenQualTA   0.2905
TotRmsAbvGrd    0.2846
FoundationPConc 0.2820
ExterQualGd     0.2638
BsmtQual        0.2483
GarageFinish    0.2426
Fireplaces      0.2373
```

Elastic Net:

```
  only 20 most important variables shown (out of 122)

                Overall
OverallQual      0.6712
GrLivArea        0.5395
NeighborhoodBin  0.4930
ExterQual        0.4637
GarageCars       0.4630
GarageArea       0.4484
KitchenQual      0.4451
TotalBsmtSF      0.4215
X1stFlrSF        0.3873
FullBath         0.3492
Age              0.3469
GarageYrBlt      0.3263
YearRemodAdd     0.3237
KitchenQualTA    0.2905
TotRmsAbvGrd     0.2846
FoundationPConc  0.2820
ExterQualGd      0.2638
BsmtQual         0.2483
GarageFinish     0.2426
Fireplaces       0.2373
```

- **How can we try to get the best prediction of the Sale Price? Is there just one or multiple models that can get us the best price?**
  In my analysis I tried implementing the linear regression models. As there is multicollinearity in the variables, models like Ridge, Lasso and Elastic Net would help in a better prediction. The Ridge regression shrinks the coefficients of correlated predictors towards each other while Lasso tends to pick one of them and discards the others. I think there are multiple models that can get us a good prediction but some of them perform better on the training set and some on the test set. Ultimately the one that performs best on the test set is the one that we need to choose as it has the ability to properly predict the Sale price for unseen data. The modelling techniques that I implemented on the dataset are :

  1. Simple Linear Regression using few predictors
  2. Multiple Linear Regression using all predictors
  3. Multiple Linear Regression using filtered predictors and 10 fold CV
  4. Robust Linear Regression
  5. Partial Least Squares
  6. Principal Component Regression
  7. Ridge Regression
  8. Elastic net

- **What parameters can be used to measure the model performance?**
  RMSE, R squared, RMSE SD, R squared SD, AIC, BIC, Anova, k-fold CV

Table 1:

No.	Model	RMSE	R Squared	Anova	AIC	BIC

1	Using predictors : OverallQual+GarageCars+GarageArea	0.2046	0.7336	60.805	-481.9807	-455.5635
2	Using predictors : OverallQual+TotalBsmtSF+ExterQual	0.2024	0.7394	59.474	-514.2159	-487.7987
3	Using predictors : OverallQual+X1stFlrSF+KitchenQual	0.1917	0.7662	53.376	-671.7100	-645.2927

Looking at the results from Table 1, I would select model 3 as it has the lowest RMSE, Anova, AIC, BIC and the largest R squared.

Table 2:

No.	**Model**	**Train RMSE**	**R^2**	**RMSE SD**	**R^2 SD**	**Test RMSE**
1	Multiple LR with all predictors	0.1051	Multiple R-squared: 0.9361 Adjusted R-squared: 0.9296	-	-	0.12194
2	Linear regression model with 10 folds CV	0.1068	Multiple R-squared: 0.933, Adjusted R-squared: 0.9273	0.01434768	0.0181251 8	0.12213
3	Robust linear regression	0.1209279	0.9076801	0.01556433	0.0206334 1	0.13200
4	PLS	0.1132457	0.9184889	0.01414828	0.0178556 0	0.11948
5	PCR	0.1214349	0.9067512	0.01359602	0.0182284 9	0.13939
6	Ridge regression	0.1129433	0.9192039	0.01403201	0.0178930 3	0.12104
7	Enet	0.1170720	0.9145279	0.01412760	0.0171410 2	0.12006

8	**Ensemble (PLS Tune and Enet)**	-	-	-	-	**0.11849**

Looking at Table 2, model 4(PLS) has the lowest RMSE and highest R squared but an ensemble of PLS and Elastic Net has a more reduced RMSE on the test set. These test set RMSE values are the results got when the predictions were uploaded on Kaggle.

Removing predictors 'LotFrontage' and 'FireplaceQu' as they had many NA's led to a better test set prediction.

Conclusion:

After analyzing the 20 most important predictors given by all the linear models tried in my project, we can say that the Overall Quality of the house is of utmost importance, the size of the house(living, basement, garage areas), the neighborhood, number of cars that can fit in the garage, external quality of the house, total number of rooms, total number of bathrooms and age of the house are also important. Most of these top predictors can be said are important by us and do not need a modelling technique to help us decide. But predictors like the size of the basement, garage finish, total rooms above ground, year remodeled, kitchen quality, the number of cars that can fit in the garage and the importance of having a fireplace can be known with the help of these models.

Instructions to run the R code:
To open the R code and to change the path of the training and test sets at line numbers 33 and 34. Rest of the code can be run chunk by chuck to see the results. The predictions need to be in same format as the sample_submission.csv to be uploaded to the Kaggle competition.

Explain/define terms

- **Training set** : The training data is used to build the model. It already contains the response that we need to predict
- **Test set** : Once we try different techniques to get the best results, we evaluate the model using test set. The test set does not contain the response and is unseen data
- **Regression** : It is a process of depicting the relationship between the predictors and the response
- **Correlation** : It is a measure which gives us information on how interdependent the variables are on each other and on the response
- **Outliers** : They are data points that are very different from most of the other data points in the dataset
- **Predictors** : They are a set of variables that help in predicting the Sale Price
- **Partial Least Squares** : It is a supervised process where the prediction is made based on the response. It combines the predictors to form components which best summarize the predictor while maximizing the correlation with the response.

- **K fold Cross Validation** : It is process where the dataset is split into k equal portions. Out of which one subset is used as the test set while the remaining are used as training data. This process is repeated K times.
- **Linear Regression** : It is a modelling technique used to show us the relationship between the response and one or more variables.
- **Multicollinearity** : It is a phenomenon where one predictor can be accurately predicted with the help of another predictor
- **Robust Linear Regression** : It is a technique which is helpful in addressing influential observations
- **Ridge Regression** : It is a technique which shrinks the coefficients of correlated predictors towards each other, allowing them to borrow strength from each other.
- **Lasso Regression** : When there are highly correlated predictors, Lasso tends to pick one and ignores the rest.
- **Principal Component Regression** : It uses Principal component analysis as a preprocessing method and performs regression after that. It is an unsupervised process that does not consider the response.
- **Elastic Net** : Elastic net regularization is a combination of Ridge and Lasso regression techniques.
- **RMSE** : Root Mean Square Error is a measure of calculating the difference between the predicted and observed values. Lower the RMSE, better the model performance
- **R-Squared** : It is a measure that helps us determine the goodness of fit of our regression model. It ranges between 0-1. Higher value the better
- **RMSE SD** : It is the standard deviation of the RMSE when resampled. Lower value the better
- **R-Squared SD** : It is the standard deviation of the R-Squared when resampled. Lower value the better
- **AIC** : Akaike information criterion estimates the quality of a model based on the quality of the other models that it is compared with. It is a model selection criteria. The model with the lowest AIC is selected
- **BIC** : Bayesian information criterion is also a model selection criteria which is similar to AIC and the model with the lowest BIC is selected
- **Anova** : It stands for analysis of variance and is one of the model selection criteria which analyzes the models based on the difference in group means. Smaller the RSS(sum of the squared residuals) the better.
- **Ensemble** : It is technique where we combine two or more modelling techniques to get a better prediction[5]

• References

- Provide *appropriate citations and references…*
- Third party tool to convert csv into sql to create table, Convert CSV to SQL : http://www.convertcsv.com/csv-to-sql.htm

[5] Learn to use Forward Selection Techniques for Ensemble Modeling, TAVISH SRIVASTAVA , SEPTEMBER 3, 2015, https://www.analyticsvidhya.com/blog/2015/09/selection-techniques-ensemble-modelling/

- Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project, Dean De Cock, Truman State University, Journal of Statistics Education Volume 19, Number 3(2011), www.amstat.org/publications/jse/v19n3/decock.pdf
- Learn to use Forward Selection Techniques for Ensemble Modeling, TAVISH SRIVASTAVA , SEPTEMBER 3, 2015, https://www.analyticsvidhya.com/blog/2015/09/selection-techniques-ensemble-modelling/
- CITY ASSESSOR, 2018 City of Ames, IA, https://www.cityofames.org/government/departments-divisions-a-h/city-assessor
- OR/SYST568: Applied Predictive Analytics, Prof. Jie Xu, Dept. of SEOR, George Mason University, Linear Regression slides and R code
- I took reference for the cleaning part from : Detailed Data Analysis & Ensemble Modeling, Tanner Carbonati, April 5 2017, https://www.kaggle.com/tannercarbonati/detailed-data-analysis-ensemble-modeling
- The in detail model building was done as a part of OR568 – Applied Predictive Analytics project, Damuluri_Molugu_Mote_Pathuri_Shen_Zhang_House Price. The Linear Regression techniques were implemented by me and therefore I have used them for further analysis in this project.
    - …including *citation for the dataset*
        - http://infoguides.gmu.edu/citingdata
        - Dean De Cock, Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project, 2011, https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data