DATA EXPLORATION TECHNIQUES TO DETERMINE HOUSE PRICES

BY

SUSHAANTH SRIRANGAPATHI

SUDIPTI DEY

NIKITA SHROTE

LAKSHMI SRINIVASAN

KAUSHIK KOMPELLA

JAIDEEP ADUSUMELLI

ABHINAV SINGH

TABLE OF CONTENTS

1. Introduction

Problem Statement: Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.
With 80 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, the dataset challenges to predict the final price of each home.

1.1. Dataset description:

- The given haggle problem had 2 datasets in csv file - training and set.
- Both the files contain 1460 rows.
- The training dataset contains 80 independent variables along with 1 dependent variable
- The dependent/target variable is called Sales Price which is a continuous variable giving the price value of a house/property
- The test dataset contains only the 80 independent variables as we are required to predict the Sale price of the houses in the test dataset

Following is the count of continuous vs categorical variables in the dataset

Variable type	**Count**
Continuous	35
Categorical	46
Total	**81**

1.1.1. Source of data

The following business problem has been taken from Kaggle. Following the link for the problem- https://www.kaggle.com/c/house-prices-advanced-regression-techniques

Following is the code used to solve the business problem -

https://github.com/sushaanths/kaggle_houseprice_predict/blob/master/model%20fitting.ipynb

1.1.2. Codebook

variable names and their units of measurement
levels of categorical variables (*e.g.,* red, green, blue; M, F)

MSSubClass: Identifies the type of dwelling involved in the sale.

20 1-STORY 1946 & NEWER ALL STYLES
30 1-STORY 1945 & OLDER
40 1-STORY W/FINISHED ATTIC ALL AGES
45 1-1/2 STORY - UNFINISHED ALL AGES
50 1-1/2 STORY FINISHED ALL AGES
60 2-STORY 1946 & NEWER
70 2-STORY 1945 & OLDER
75 2-1/2 STORY ALL AGES
80 SPLIT OR MULTI-LEVEL
85 SPLIT FOYER
90 DUPLEX - ALL STYLES AND AGES
120 1-STORY PUD (Planned Unit Development) - 1946 & NEWER
150 1-1/2 STORY PUD - ALL AGES
160 2-STORY PUD - 1946 & NEWER
180 PUD - MULTILEVEL - INCL SPLIT LEV/FOYER
190 2 FAMILY CONVERSION - ALL STYLES AND AGES
MSZoning: Identifies the general zoning classification of the sale. A Agriculture, C Commercial, FV Floating Village, Residential, I Industrial, RH Residential High Density, RL Residential Low Density, RP Residential Low Density Park RM Residential Medium Density

LotFrontage: Linear feet of street connected to property
LotArea: Lot size in square feet
Street: Type of road access to property: Grvl Gravel, Pave Paved
Alley: Type of alley access to property: Grvl Gravel, Pave Paved, NA No alley access
LotShape: General shape of property: Reg Regular, IR1 Slightly irregular, IR2 Moderately Irregular, IR3 Irregular
LandContour: Flatness of the property
Lvl Near Flat/Level
Bnk Banked - Quick and significant rise from street grade to building
HLS Hillside - Significant slope from side to side
Low Depression
Utilities: Type of utilities available
AllPub All public Utilities (E,G,W,& S)
NoSewr Electricity, Gas, and Water (Septic Tank)
NoSeWa Electricity and Gas Only

ELO Electricity only
LotConfig: Lot configuration
Inside Inside lot
Corner Corner lot
CulDSac Cul-de-sac
FR2 Frontage on 2 sides of property
FR3 Frontage on 3 sides of property
LandSlope: Slope of property
Gtl Gentle slope
Mod Moderate Slope
Sev Severe Slope
Neighborhood: Physical locations within Ames city limits
Blmngtn Bloomington Heights
Blueste Bluestem
BrDale Briardale

BrkSide Brookside
ClearCr Clear Creek
CollgCr College Creek
Crawfor Crawford
Edwards Edwards
Gilbert Gilbert
IDOTRR Iowa DOT and Rail Road
MeadowV Meadow Village
Mitchel Mitchell
Names North Ames
NoRidge Northridge
NPkVill Northpark Villa
NridgHt Northridge Heights
NWAmes Northwest Ames
OldTown Old Town
SWISU South & West of Iowa State University
Sawyer Sawyer
SawyerW Sawyer West
Somerst Somerset
StoneBr Stone Brook
Timber Timberland
Veenker Veenker
Condition1: Proximity to various conditions
Artery Adjacent to arterial street
Feedr Adjacent to feeder street
Norm Normal
RRNn Within 200' of North-South Railroad

RRNe Within 200' of East-West Railroad
RRAe Adjacent to East-West Railroad
BldgType: Type of dwelling
1Fam Single-family Detached
2FmCon Two-family Conversion; originally built as one-family dwelling
Duplx Duplex
TwnhsE Townhouse End Unit
2.5Fin Two and one-half story: 2nd level finished
2.5Unf Two and one-half story: 2nd level unfinished
SFoyer Split Foyer
SLvl Split Level
OverallQual: Rates the overall material and finish of the house: 10 Very Excellent, 9 Excellent, 8 Very Good,7 Good, 6Above Average,5 Average,4 Below Average,3 Fair, 2 Poor, 1 Very Poor
OverallCond: Rates the overall condition of the house: 10 Very Excellent, 9 Excellent, 8 Very Good, 7 Good, 6 Above Average, 5 Average, 4 Below Average, 3 Fair, 2 Poor, 1 Very Poor
YearBuilt: Original construction date
YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)
RoofStyle: Type of roof
Flat Flat
Gable Gable
Gambrel Gabrel (Barn)
Hip Hip
Mansard Mansard
Shed Shed
RoofMatl: Roof material
ClyTile Clay or Tile

Exterior1st: Exterior covering on house
AsbShng Asbestos Shingles
AsphShn Asphalt Shingles
BrkComm Brick Common
BrkFace Brick Face
CBlock Cinder Block
CemntBd Cement Board
HdBoard Hard Board
ImStucc Imitation Stucco
MetalSd Metal Siding
Other Other
Plywood Plywood
PreCast PreCast
Stone Stone
Stucco Stucco
VinylSd Vinyl Siding
Wd Sdng Wood Siding
WdShing Wood Shingles
Exterior2nd: Exterior covering on house (if more than one material)
AsbShng Asbestos Shingles
AsphShn Asphalt Shingles
BrkComm Brick Common
BrkFace Brick Face
CBlock Cinder Block
CemntBd Cement Board
HdBoard Hard Board
ImStucc Imitation Stucco
MetalSd Metal Siding
Other Other
Plywood Plywood
PreCast PreCast
Stone Stone
Stucco Stucco
VinylSd Vinyl Siding
Wd Sdng Wood Siding
WdShing Wood Shingles
MasVnrArea: Masonry veneer area in square feet
ExterQual: Evaluates the quality of the material on the exterior
Ex Excellent
Gd Good
TA Average/Typical
Fa Fair
Po Poor

ExterCond: Evaluates the present condition of the material on the exterior
Ex Excellent
Gd Good
TA Average/Typical
Fa Fair
Po Poor
Foundation: Type of foundation
BrkTil Brick & Tile
CBlock Cinder Block
PConc Poured Contrete

Slab Slab
Stone Stone
Wood Wood
BsmtQual: Evaluates the height of the basement
Ex Excellent (100+ inches)
Gd Good (90-99 inches)
TA Typical (80-89 inches)
Fa Fair (70-79 inches)
Po Poor (<70 inches
NA No Basement
BsmtCond: Evaluates the general condition of the basement
Ex Excellent
Gd Good
TA Typical - slight dampness allowed
Fa Fair - dampness or some cracking or settling
Po Poor - Severe cracking, settling, or wetness
NA No Basement
BsmtExposure: Refers to walkout or garden level walls
Gd Good Exposure
Av Average Exposure (split levels or foyers typically score average or above)
Mn Mimimum Exposure
No No Exposure
NA No Basement
BsmtFinType1: Rating of basement finished area
GLQ Good Living Quarters
ALQ Average Living Quarters
BLQ Below Average Living Quarters
Rec Average Rec Room
LwQ Low Quality
Unf Unfinshed
NA No Basement
BsmtFinSF1: Type 1 finished square feet
BsmtFinType2: Rating of basement finished area (if multiple types)
GLQ Good Living Quarters
ALQ Average Living Quarters
BLQ Below Average Living Quarters
Rec Average Rec Room
LwQ Low Quality
Unf Unfinshed
NA No Basement

BsmtFinSF2: Type 2 finished square feet
BsmtUnfSF: Unfinished square feet of basement area
TotalBsmtSF: Total square feet of basement area
Heating: Type of heating
Floor Floor Furnace
GasA Gas forced warm air furnace
GasW Gas hot water or steam heat
Grav Gravity furnace
OthW Hot water or steam heat other than gas
Wall Wall furnace
HeatingQC: Heating quality and condition
Ex Excellent
Gd Good

TA Average/Typical
Fa Fair
Po Poor
CentralAir: Central air conditioning
N No
Y Yes
Electrical: Electrical system
SBrkr Standard Circuit Breakers & Romex
FuseA Fuse Box over 60 AMP and all Romex wiring (Average)
FuseF 60 AMP Fuse Box and mostly Romex wiring (Fair)
FuseP 60 AMP Fuse Box and mostly knob & tube wiring (poor)
Mix Mixed
1stFlrSF: First Floor square feet
2ndFlrSF: Second floor square feet
LowQualFinSF: Low quality finished square feet (all floors)
GrLivArea: Above grade (ground) living area square feet
BsmtFullBath: Basement full bathrooms
BsmtHalfBath: Basement half bathrooms
FullBath: Full bathrooms above grade
HalfBath: Half baths above grade
Bedroom: Bedrooms above grade (does NOT include basement bedrooms)
Kitchen: Kitchens above grade
KitchenQual: Kitchen quality
Ex Excellent
Gd Good
TA Typical/Average
Fa Fair
Po Poor
TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)

Functional: Home functionality (Assume typical unless deductions are warranted)
Typ Typical Functionality
Min1 Minor Deductions 1
Min2 Minor Deductions 2
Mod Moderate Deductions
Maj1 Major Deductions 1
Maj2 Major Deductions 2
Sev Severely Damaged
Sal Salvage only
Fireplaces: Number of fireplaces
FireplaceQu: Fireplace quality
Ex Excellent - Exceptional Masonry Fireplace
Gd Good - Masonry Fireplace in main level
TA Average - Prefabricated Fireplace in main living area or Masonry Fireplace in basement
Fa Fair - Prefabricated Fireplace in basement
Po Poor - Ben Franklin Stove
NA No Fireplace
GarageType: Garage location
2Types More than one type of garage
Attchd Attached to home
Basment Basement Garage
BuiltIn Built-In (Garage part of house - typically has room above garage)
CarPort Car Port

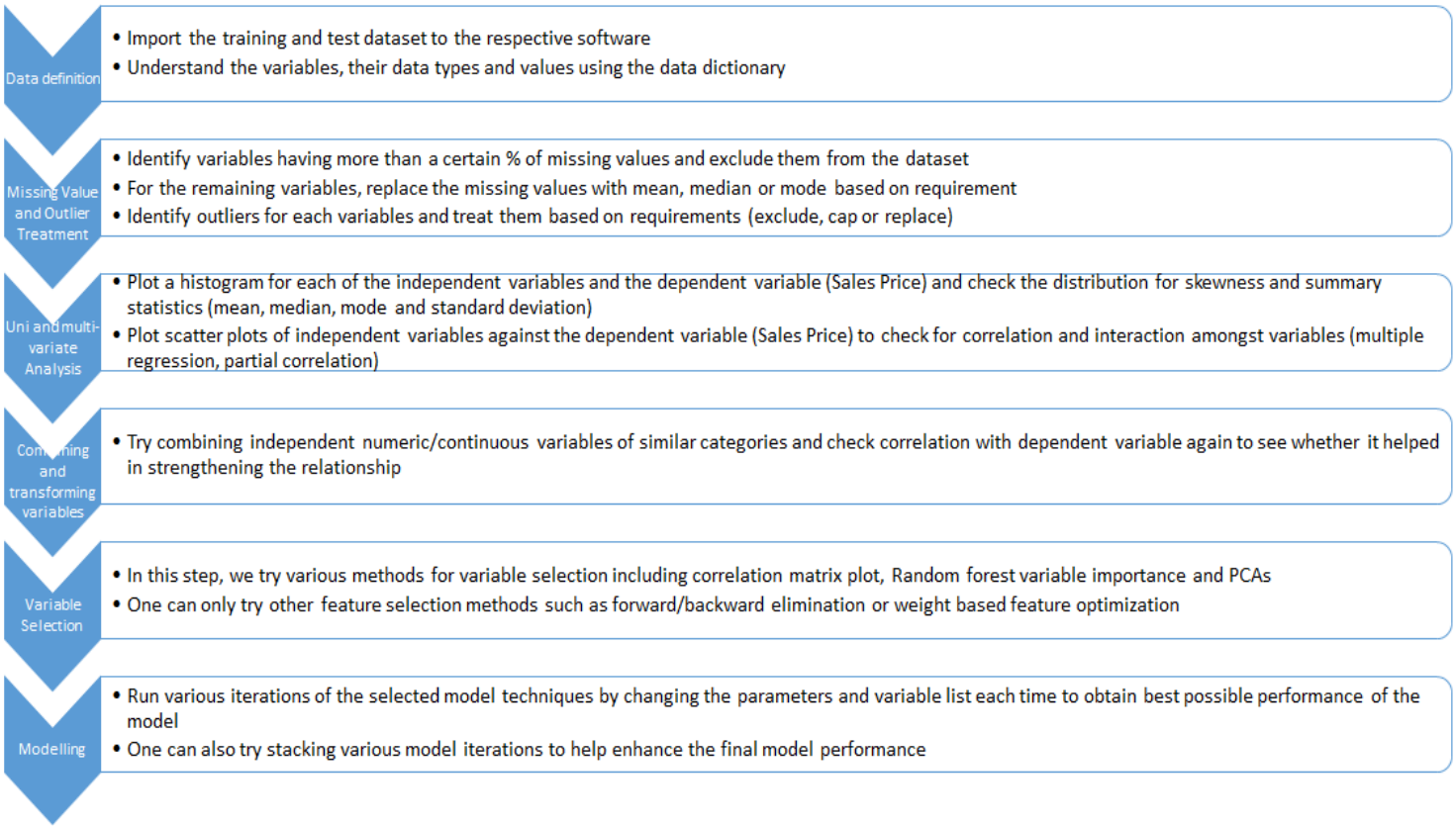
Detchd Detached from home
NA No Garage
GarageYrBlt: Year garage was built
GarageFinish: Interior finish of the garage
Fin Finished
RFn Rough Finished
Unf Unfinished
NA No Garage
GarageCars: Size of garage in car capacity
GarageArea: Size of garage in square feet
GarageQual: Garage quality
Ex Excellent
Gd Good
TA Typical/Average
Fa Fair
Po Poor
NA No Garage
GarageCond: Garage condition
Ex Excellent
Gd Good
TA Typical/Average
Fa Fair
Po Poor
NA No Garage

PavedDrive: Paved driveway
Y Paved
P Partial Pavement
N Dirt/Gravel
WoodDeckSF: Wood deck area in square feet
OpenPorchSF: Open porch area in square feet
EnclosedPorch: Enclosed porch area in square feet
3SsnPorch: Three season porch area in square feet
ScreenPorch: Screen porch area in square feet
PoolArea: Pool area in square feet
PoolQC: Pool quality
Ex Excellent
Gd Good
TA Average/Typical
Fa Fair
NA No Pool
Fence: Fence quality
GdPrv Good Privacy
MnPrv Minimum Privacy
GdWo Good Wood
MnWw Minimum Wood/Wire
NA No Fence
MiscFeature: Miscellaneous feature not covered in other categories
Elev Elevator
Gar2 2nd Garage (if not described in garage section)
Othr Other
Shed Shed (over 100 SF)
TenC Tennis Court

2. Statistical Analysis

2.1. Business problem approach

We followed a structured approach to solve the business problem -



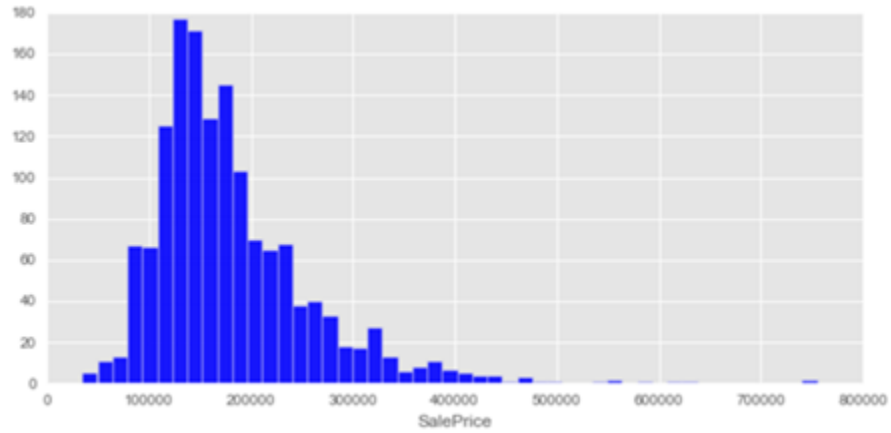
2.2.Pre-processing, if any (e.g., differencing; transformation)

After understanding the data with the help of the data dictionary we move towards null value and outlier treatment.
For null values - we exclude any variables having more than 85% of missing values.
4 variables were excluded
For the remaining variables -
Continuous variables: replace missing values with median value
Categorical variables: replace missing values with 'MIA' which creates a new category of missing values

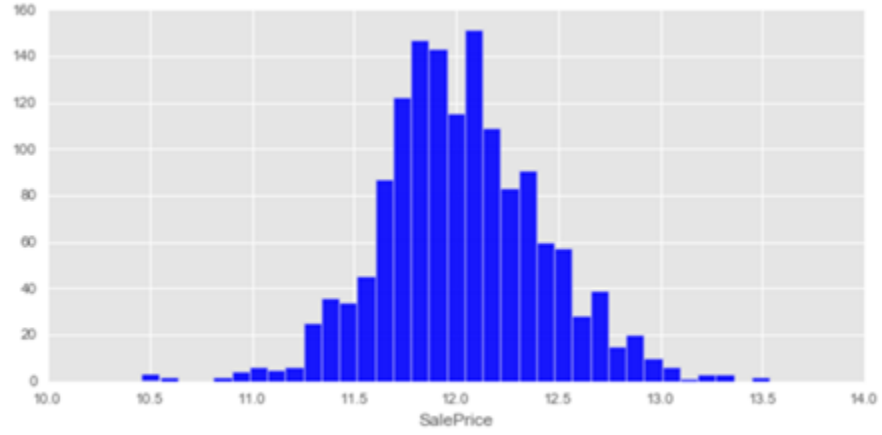
Outlier treatment: calculate the absolute difference of data points with the median value of the variable and exclude those which have a significantly high value

Univariate and bivariate analysis: To understand which variables to transform and which to combine etc. we performed univariate analysis and correlation analysis.

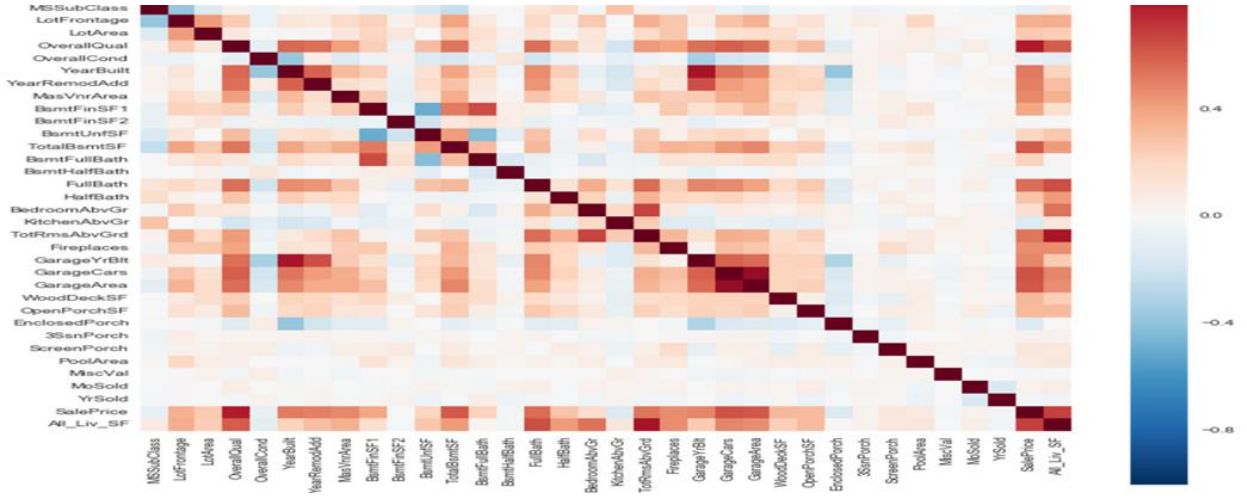
Univariate: As the term suggests, this process includes understanding one variable at a time, plotting its distribution across data points and understand the summary statistics (mean, median and mode)
Here's an example of histogram of the target variable (Sales Price)



We can see from plot that Sales Price follows a left skewed distribution. To overcome this, we perform a variable transformation, namely logarithmic. Below is the distribution of this variable post transformation -

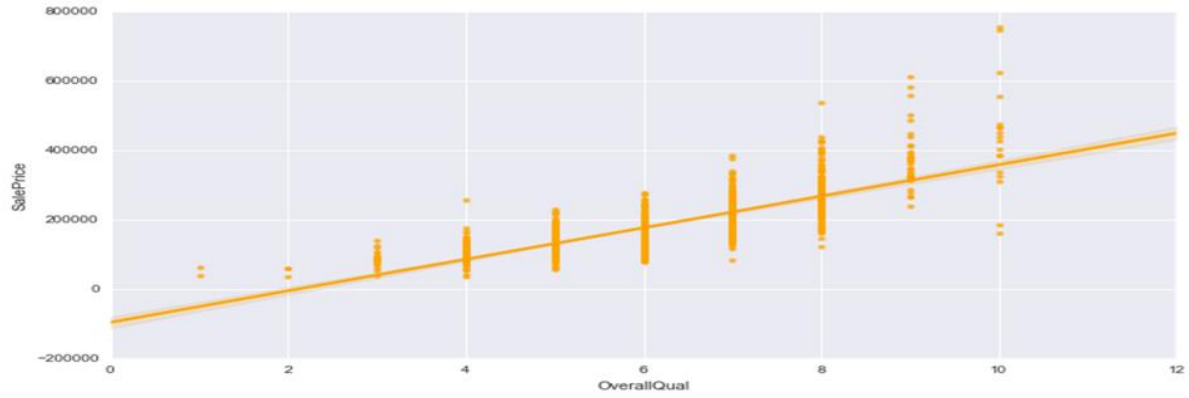


Bivariate: As the term suggest, this involves studying two variables at a time to understand the relationship between them. Through regression plots we look at the correlation value and sign to see whether two variables have a positive or negative relationship and the absolute value of correlation to know the strength of the relationship

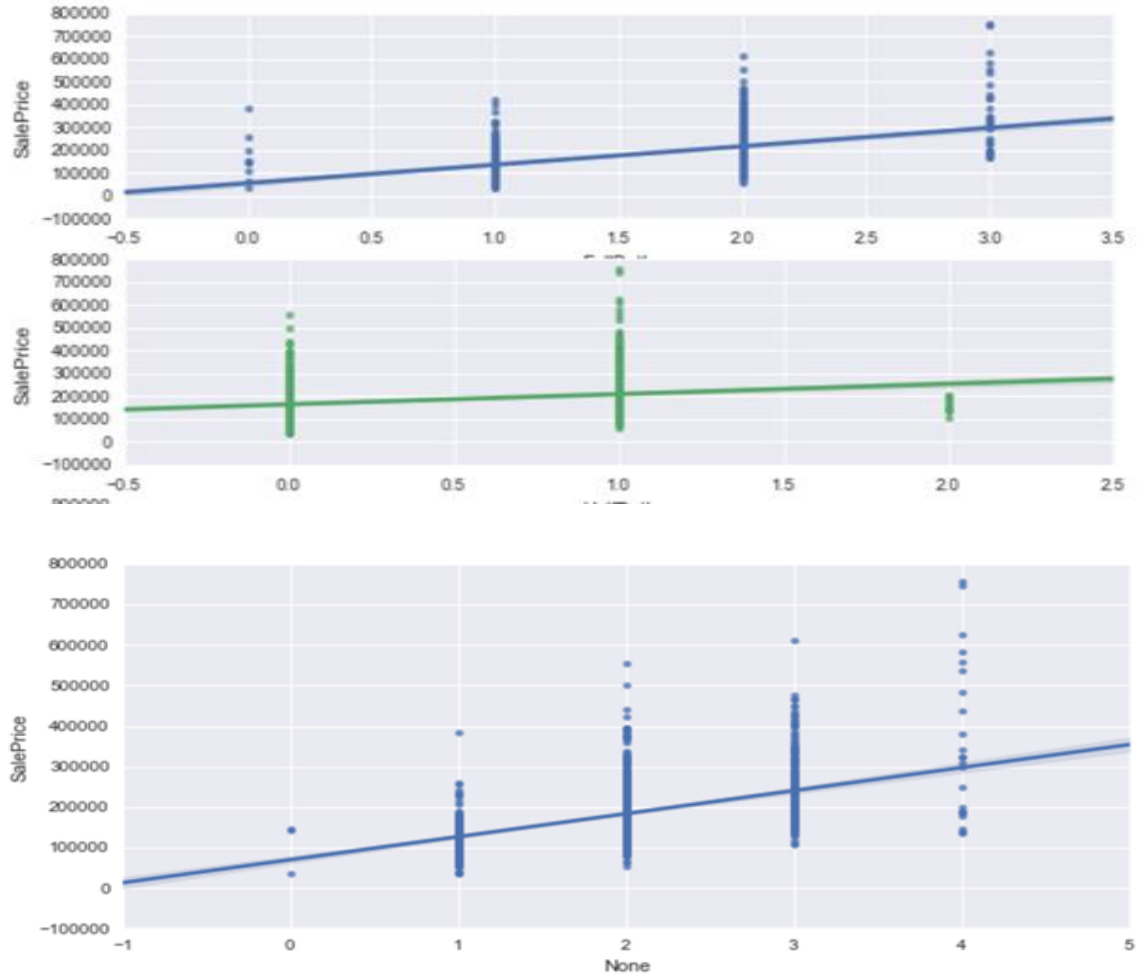


This is a correlation matrix of some of the variables from the training dataset. Darker shades of red show a positively strong relationship and darker shades of blue show a negatively strong relationship.
We can also see which variables influence the target variable (Sales Price) by seeing its correlation with the other independent variables.

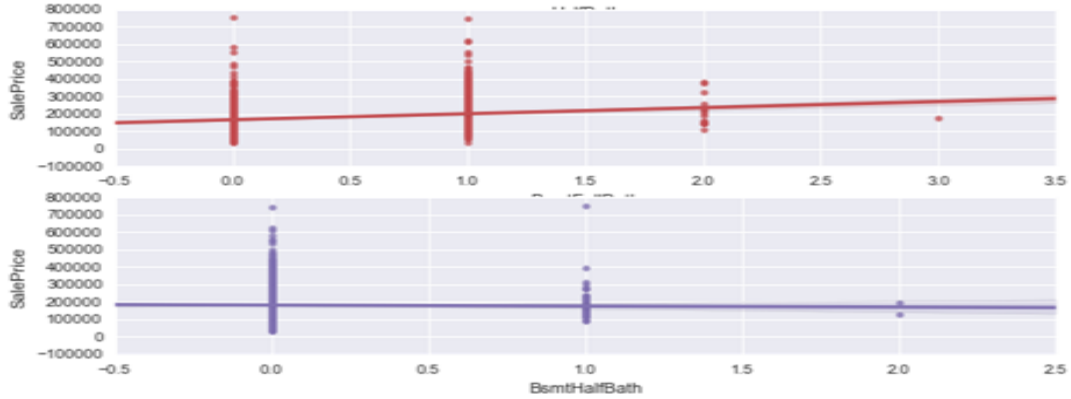
Variable transformations: after understanding the relationships between the variables, we can see which variables can be combined to strengthen its influence on the target variable

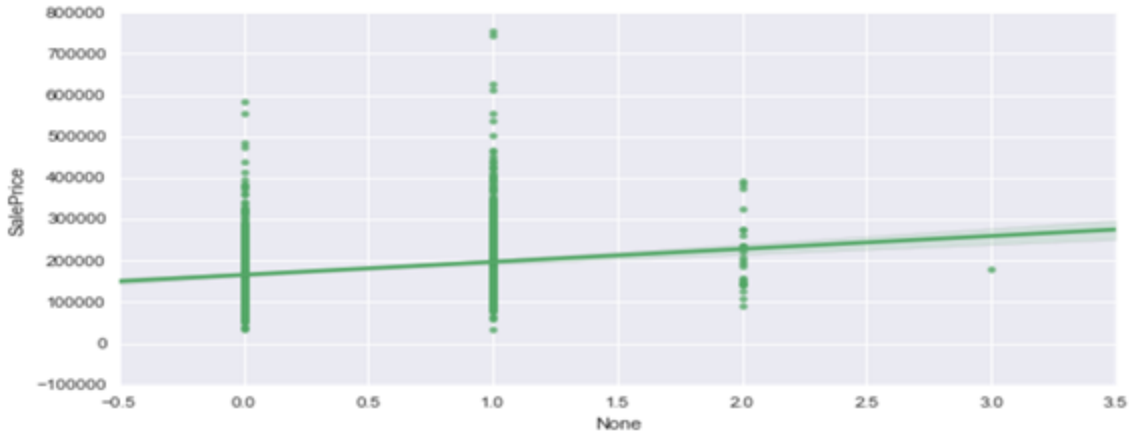


•Here is an example of a regression plot of a categorical variable (Overall quality) against the dependent variable (Sale Price)
•The fitted line shows some relationship between Overall quality and Sales price (as the line is on an upward slope and not flat)
•However, we can see that there are very few data points lying in the category of overall quality between 8-12
•As the Overall quality of the house increases, the sale price of the house increase (correlation not causation)



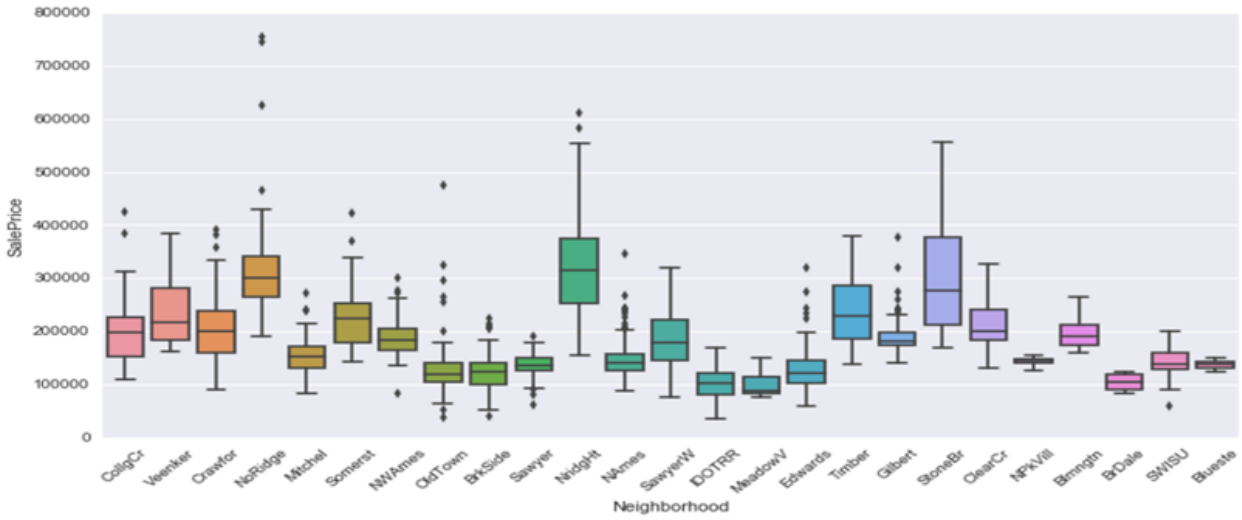
•We looked at two variables – number of full and half baths in the house individually against the dependent variable (sales price) and noticed that though there is some correlation between sales price and full bath or half bath, it's not such a strong relationship (images on the left)
•After combining the two variables as one (adding the values) and plotting it against Sales Price, we saw a stronger relationship between the new transformed variable and Sales Price





•We performed the same variable transformation on another set of two variables – Basement half bathrooms and basement full bathrooms

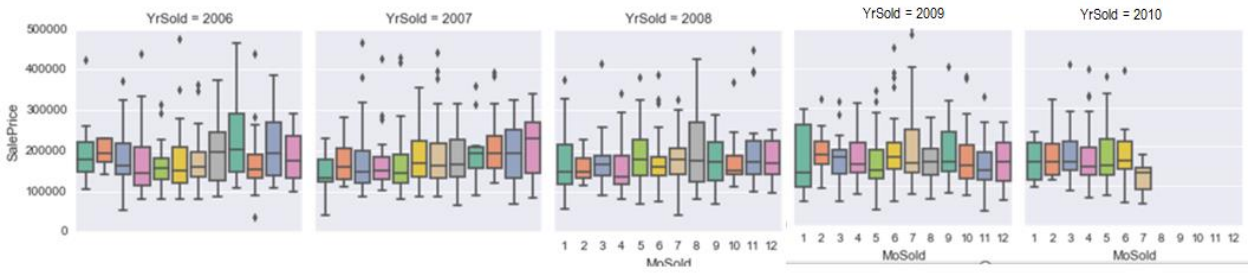
•We observed an increase in the correlation between the two independent variables and sales price after they were combined as one



•One of the independent categorical variables named ‘Neighborhood’ had 25 categories

•We plotted a box plot for each category of Neighborhood against the sales price to check which of the categories have a higher median sales price - NridgHt and StoneBr (from the above image)

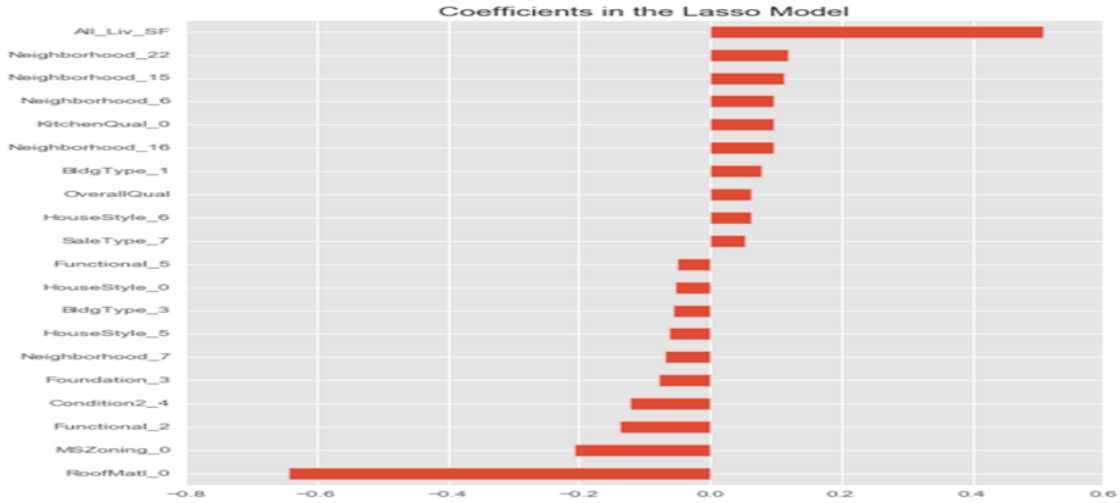
•We transformed these 2 categories into binary variables (Eg- if data point (House ID) belongs to NridgHt then 1 else 0) as they have a relationship with the dependent variable



•Plotting box plots for each month of the year (2006 – 2010) against the Sales Price to see if there is a significant increase or decrease in Sale price during any months of the year
•We can notice that there is no increase/decrease in Sales Price during any months across the years, hence there is no effect of seasonality on house sale prices

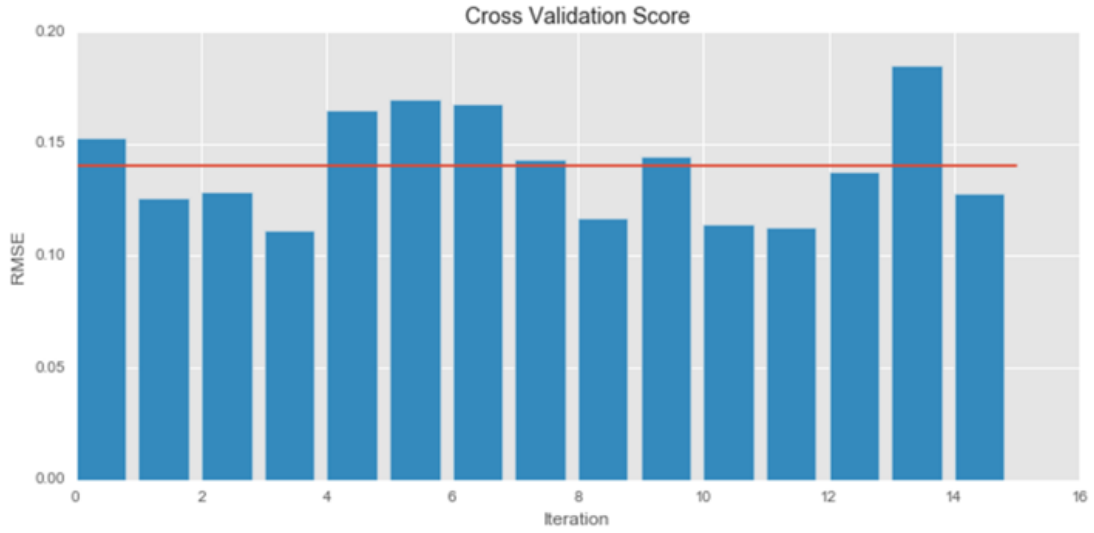
After all the variable transformations, we move towards the modelling aspect. We used two types of models:
Lasso regression and random forest

We checked whether the variables we selected as important were also selected by the lasso regression model



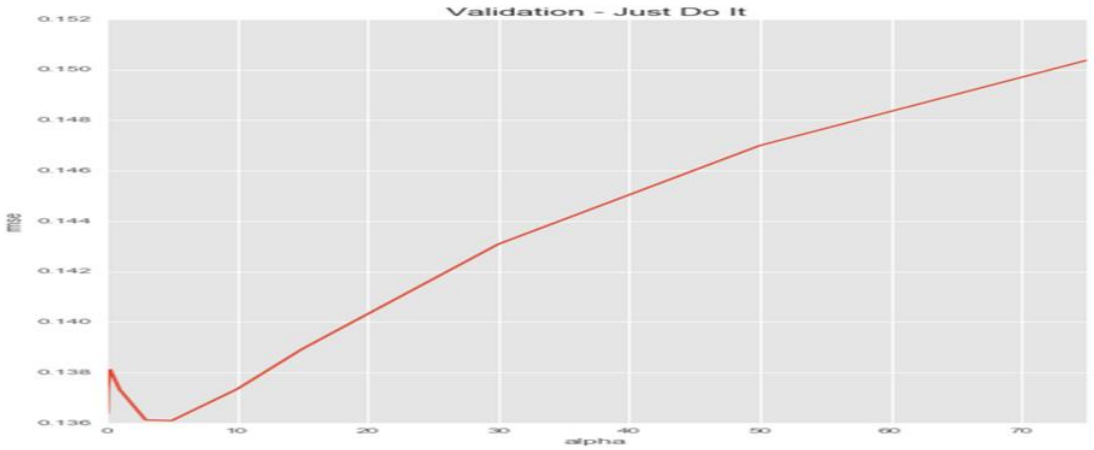
•The important exploratory variables identified through data exploration were considered important by the Lasso Regression model
•There were few attributes which were identified as important by the model but it was not identified during exploration process
•This emphasizes that using partial correlation will help us identify the more detailed insights

Our baseline random forest model (without any variable transformation or variable selection)-



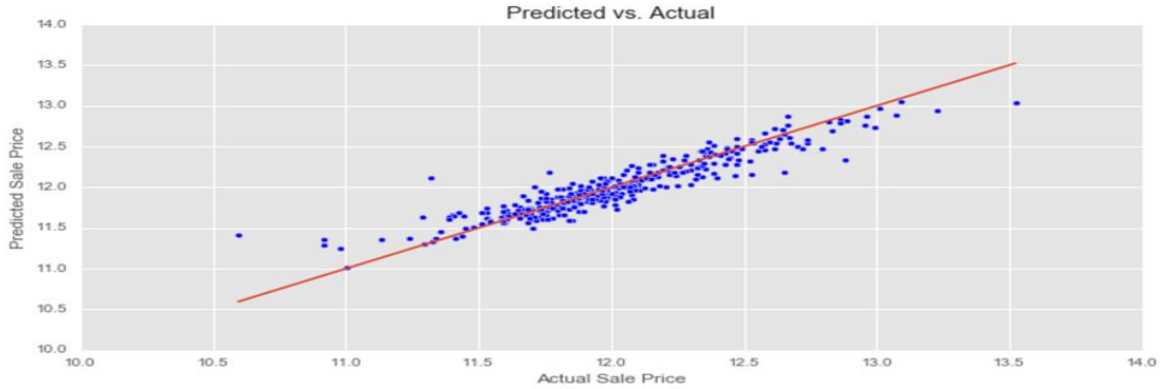
•Without any transformation and specific variable selection, we could obtain an RMSE score of 0.140 (Rank 1120 in haggle)

Lasso regression model -



•After the transformation process and proper feature selection we could obtain MSE of 0.136 (Rank 712 in kaggle)

Predicted v/s actual sales for Lasso model



•The above plot shows the goodness of fit line for the Lasso model.
•We can observe that there are many data points away from the line (high residuals) which shows that this model may not be the best
•Requires more feature engineering to be performed to get a better goodness of fit

2.3. Software used:

We used Python for our analysis. Some of the libraries used were

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib
import matplotlib.pyplot as plt
from sklearn.preprocessing import scale
from scipy.stats import skew, skewtest
%config InlineBackend.figure_format = 'png'
%matplotlib inline
```

2.4. Procedures used

- Outlier treatment
  - Median absolute deviation
  - Bivariate analysis
  - Correlation matrix and regression plots
- Feature Engineering:
  - Random forest for variable importance plot
  - Correlation matrix
  - Forward/backward feature selection
  - Optimize weight feature selection
- Modelling:
  - Random forest: it is a method for classification and regression that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.
  - Lasso regression: Least absolute shrinkage and selection operator (also Lasso or LASSO) is a regression analysis method that performs both variable selection and regularization to enhance the prediction accuracy and interpretability of the statistical model it produces.
- Performance evaluation:
  - RMSE: **root-mean-square error (RMSE)** is a frequently used measure of the differences between values (sample and population values) predicted by a model or an estimator and the values actually observed.

3. Conclusions

Data exploration Summary

- ❖ **Numeric Variables**

•1stFlrSF,2ndFlrSF,LowQualFinSF,GrLivArea are all skewed but influences house prices together. Hence we could group them to reduce dimensionality (All_Liv_SF)

•OverallQual is a key indicator for the house prices

•GarageCars and GarageArea have a positive relation with sales prices

•YearBuilt and YearRemodAdd can provide some effect on sales price (not much)

- ❖ **Categorical Variables**

•neighborhood is indicative or related to sales prices. eg. StoneBR and NridgHT have higher median price. We can group regions with similar prices into buckets to reduce dimension of this attribute

•sales condition can be partial indicative of sales prices. eg. partial sales have higher house prices. Can be used

•Sale's timing does not seem to hugely affect the house.

•HouseStyle shows some indication of sales prices. But it doesnt provide enough info to be used all by itself like OveralQual

•FireplaceQu actually determines the price. "Fireplaces" doesnt contribute much to the price compared to quality of the fireplace

•CentralAir is a key indicator of price along with fireplace

•KitchenQual is related to sales price of house

•price is related to the MSZoning to some extent

Model Summary:

- Lasso regression performed better than random forest as it gave a lower value for RMSE

In conclusion we can say the following-

- Data exploration is a very important step of this business problem - it helps us understand which variables influence the target variable , which variables have interaction amongst each other (partial correlation)
- After the lasso regression model, we now know which variables influence the price of a house and can consider these factors while determining the price of a house
- This model can be very useful in the real estate field, agents can use information from this model while determining house prices

References:

•Git link https://github.com/sushaanths/kaggle_houseprice_predict/blob/master/model%20fitting.ipynb

•Kaggle Dataset https://www.kaggle.com/c/house-prices-advanced-regression-techniques

•Exploratory data analysis https://en.wikipedia.org/wiki/Exploratory_data_analysis

•Partial correlation https://en.wikipedia.org/wiki/Partial_correlation

•Correlation matrix in R http://www.sthda.com/english/wiki/correlation-matrix-a-quick-start-guide-to-analyze-format-and-visualize-a-correlation-matrix-using-r-software

All definitions of terms were taken from Wikipedia