House Price Prediction

By Neeraj Walia

Problem to solve

- Provide research tool to homebuyer for searching houses and predict the home prices .
- Homeowners can use this to get an estimate of price their house before putting it on sale.

Introduction

- Home buyers need a tool that can easily narrow down their search selection for the house based on their **budget**, **requirements** like number of bedrooms etc. and **location**.

- Predict/forecast home prices so home buyers can have an estimate and make an informed decision at the time of home purchase.

- Home owners can use this tool to estimate the market value of their house.

Introduction

- I will be using advanced regression models to prediction:
  - Linear regression
  - Random forest regression.

- Website will be designed for easy to use and visual representation of data.

Tools used

- Tableau for website design
- Python for predictive analytics.

Data

- The dataset is available on kaggle website and its direct link is provided below.

https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data

- The dataset contain residential homes information from Ames, Iowa. It contains 80 features (variables) representing every aspect of a house.

- There are two datasets train and test. Both has 1461 records of data. Train data has house price information. This data will be used to build the model.

Data

```
df_train.head()
```

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	PoolQC	Fence
0	1	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN
1	2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN
2	3	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN
3	4	70	RL	60.0	9550	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN
4	5	60	RL	84.0	14260	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN

5 rows × 81 columns

```
df_test.head()
```

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	ScreenPorch	PoolArea
0	1461	20	RH	80.0	11622	Pave	NaN	Reg	Lvl	AllPub	...	120	0
1	1462	20	RL	81.0	14267	Pave	NaN	IR1	Lvl	AllPub	...	0	0
2	1463	60	RL	74.0	13830	Pave	NaN	IR1	Lvl	AllPub	...	0	0
3	1464	60	RL	78.0	9978	Pave	NaN	IR1	Lvl	AllPub	...	0	0
4	1465	120	RL	43.0	5005	Pave	NaN	IR1	HLS	AllPub	...	144	0

5 rows × 80 columns

Data: Features

- MSSubClass: type of dwelling (category)
  - 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
  - 1201-STORY PUD (Planned Unit Development) - 1946 & NEWER
  - 1501-1/2 STORY PUD - ALL AGES
  - 1602-STORY PUD - 1946 & NEWER
  - 180PUD - MULTILEVEL - INCL SPLIT LEV/FOYER
  - 1902 FAMILY CONVERSION - ALL STYLES AND AGES

- MSZoning: zoning classification (category)
  - 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

Data: Features

- LotFrontage: Linear feet of street connected to property (numerical)
- Street: Type of road access to property (category)
  - Grvl Gravel
  - Pave Paved
- Alley: Type of alley access to property (category)
  - Grvl Gravel
  - Pave Paved
  - NA No alley access
- LotShape: General shape of property
  - Reg Regular
  - IR1Slightly irregular
  - IR2Moderately Irregular
  - IR3Irregular
- 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

Data: Features

- 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

Data: Features

- Condition1: Proximity to various conditions
  - Artery — Adjacent to arterial street
  - Feedr — Adjacent to feeder street
  - Norm — Normal
  - RRNn — Within 200' of North-South Railroad
  - RRAn — Adjacent to North-South Railroad
  - PosN — Near positive off-site feature--park, greenbelt, etc.
  - PosA — Adjacent to postive off-site feature
  - RRNe — Within 200' of East-West Railroad
  - RRAe — Adjacent to East-West Railroad

- Condition2: Proximity to various conditions (if more than one is present)
  - Artery — Adjacent to arterial street
  - Feedr — Adjacent to feeder street
  - Norm — Normal
  - RRNn — Within 200' of North-South Railroad
  - RRAn — Adjacent to North-South Railroad
  - PosN — Near positive off-site feature--park, greenbelt, etc.
  - PosA — Adjacent to postive off-site feature
  - 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
  - TwnhsI — Townhouse Inside Unit

Data: Features

- HouseStyle: Style of dwelling
  - 1Story — One story
  - 1.5Fin — One and one-half story: 2nd level finished
  - 1.5Unf — One and one-half story: 2nd level unfinished
  - 2Story — Two story
  - 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
  - 6 Above 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

Data: Features

- YearBuilt: Original construction date
- YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)
- RoofStyle: Type of roof
  - FlatFlat
  - Gable Gable
  - Gambrel Gabrel (Barn)
  - Hip Hip
  - Mansard Mansard
  - Shed Shed

- RoofMatl: Roof material
  - ClyTile Clay or Tile
  - CompShg Standard (Composite)
  - Shingle
  - Membran Membrane
  - Metal Metal
  - RollRoll
  - Tar&Grv Gravel & Tar
  - WdShake Wood Shakes
  - WdShngl Wood Shingles
- 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

Data: Features

- 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
- MasVnrType: Masonry veneer type
  - BrkCmn Brick Common
  - BrkFace Brick Face
  - CBlock Cinder Block
  - None None
  - Stone Stone
- 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

Data: Features

- 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

Data: Features

- 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

Data: Features

- 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 gradeKitchenQual: 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

Data: Features

- 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

Data: Features

- 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
  - ElevElevator
  - Gar2 2nd Garage (if not described in garage section)
  - Othr Other
  - Shed Shed (over 100 SF)
  - TenC Tennis Court
  - NA None
- MiscVal: $Value of miscellaneous feature
- MoSold: Month Sold (MM)
- YrSold: Year Sold (YYYY)

Data: Features

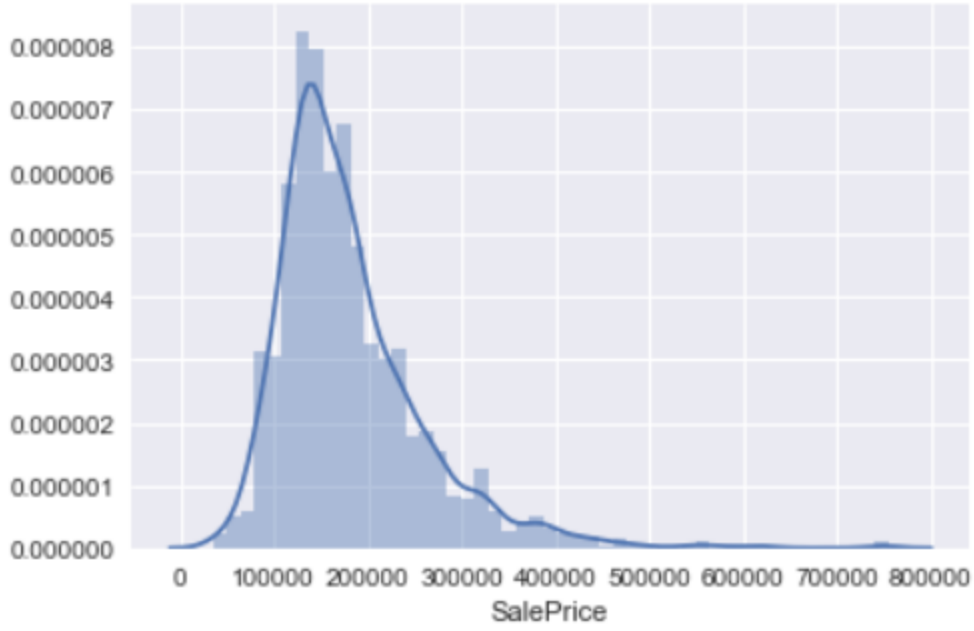
- SaleType: Type of sale
  - WD Warranty Deed - Conventional
  - CWD Warranty Deed - Cash
  - VWD Warranty Deed - VA Loan
  - New Home just constructed and sold
  - COD Court Officer Deed/Estate
  - ConContract 15% Down payment regular terms
  - ConLw Contract Low Down payment and low interest
  - ConLI Contract Low Interest
  - ConLD Contract Low Down
  - Oth Other

- SaleCondition: Condition of sale
  - Normal Normal Sale
  - Abnorml Abnormal Sale - trade, foreclosure, short sale
  - AdjLand Adjoining Land Purchase
  - Alloca Allocation - two linked properties with separate deeds, typically condo with a garage unit
  - Family Sale between family members
  - Partial Home was not completed when last assessed (associated with New Homes)

Data: Exploration

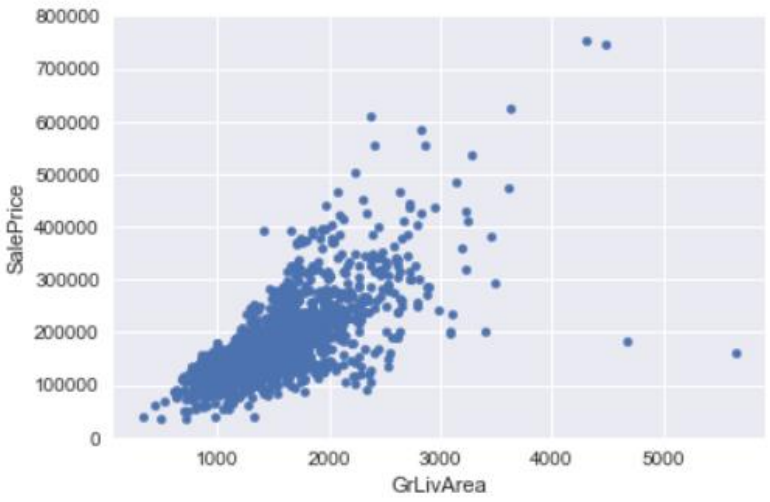
- **Sales Price**

```
count      1460.000000
mean     180921.195890
std       79442.502883
min       34900.000000
25%      129975.000000
50%      163000.000000
75%      214000.000000
max      755000.000000
Name: SalePrice, dtype: float64
```

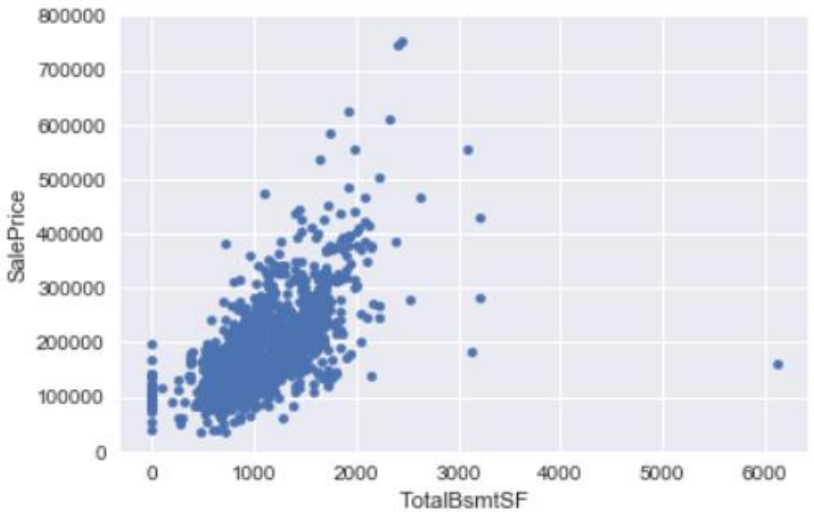


Data: Exploration

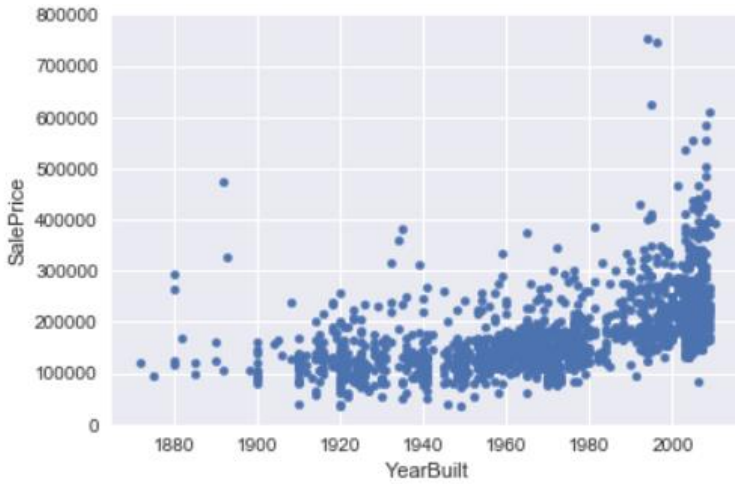
- **Sales Price vs GrLiv Area**



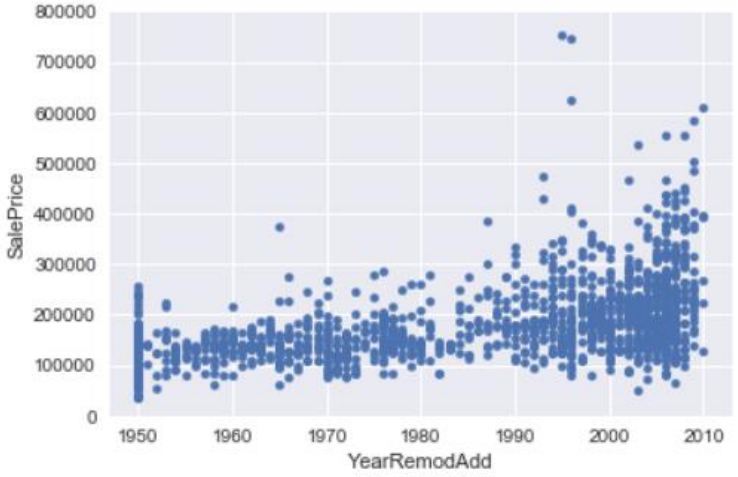
Sales Price vs Total Basement SF



Sales Price vs Year Built

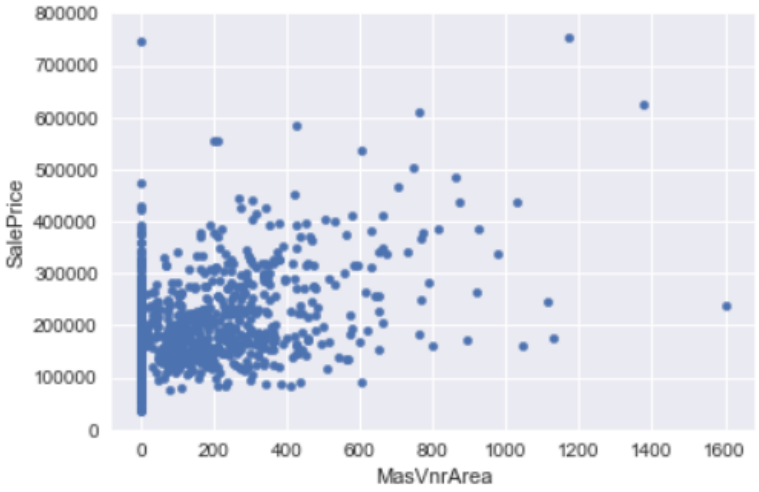


Sales Price vs Year Remod Add

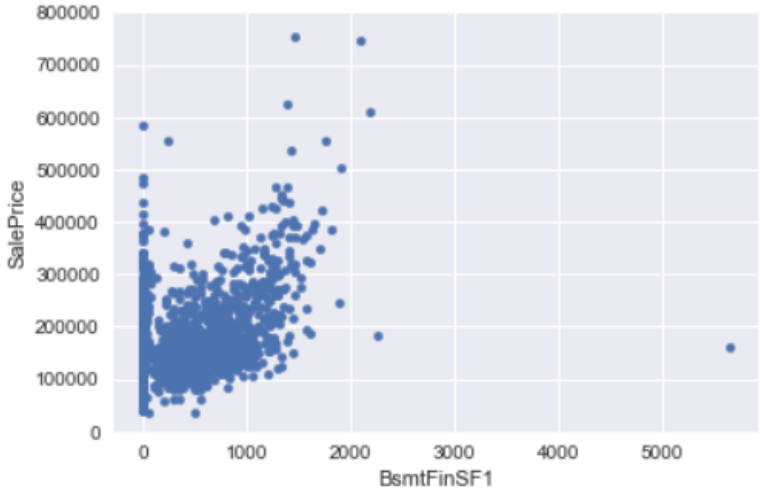


Data: Exploration

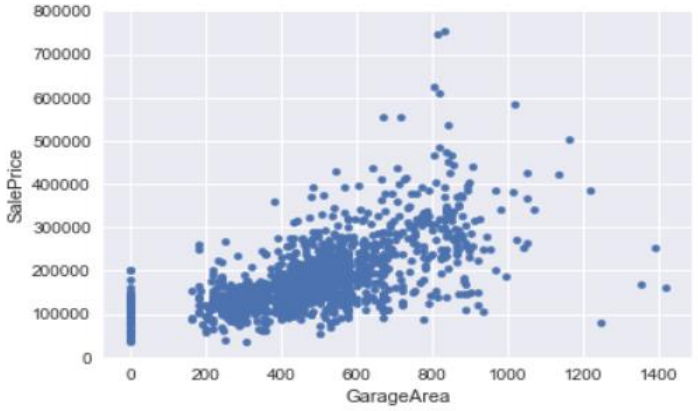
- **Sales Price vs MasVnrArea**



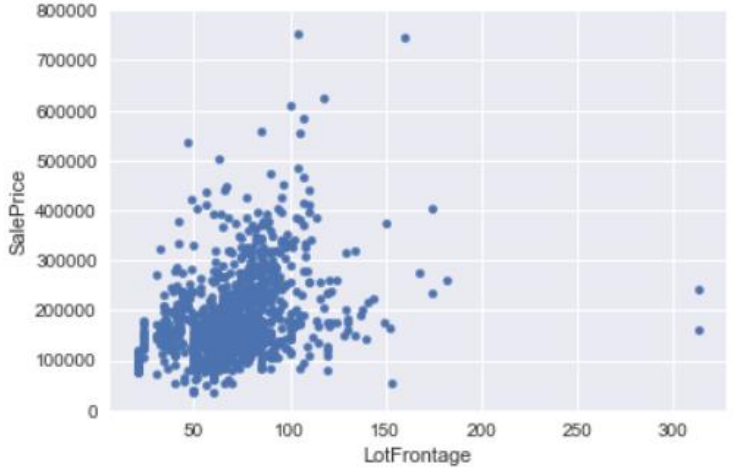
Sales Price vs Basement FinishedSF



Sales Price vs Garage Area

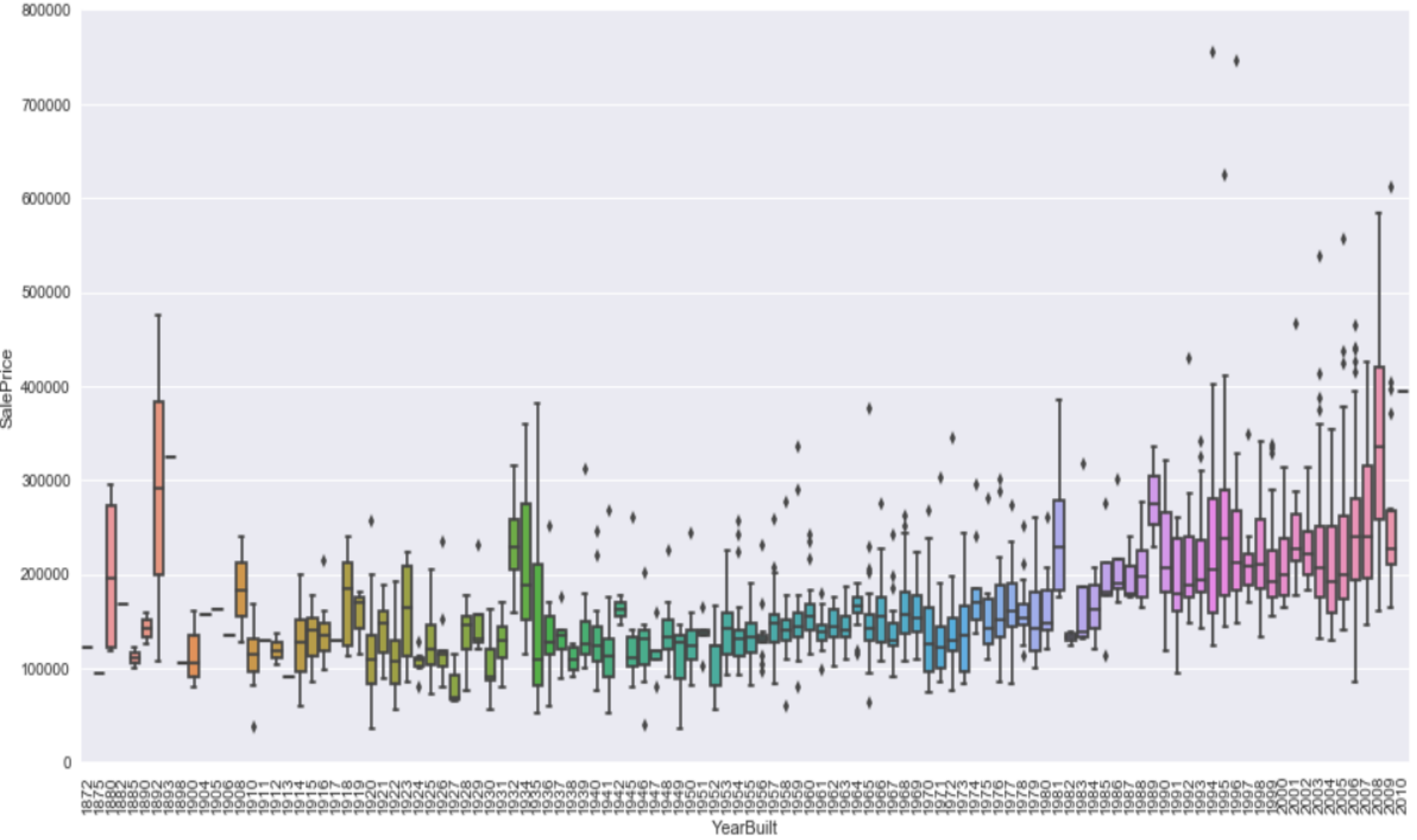


Sales Price vs Lot Frontage



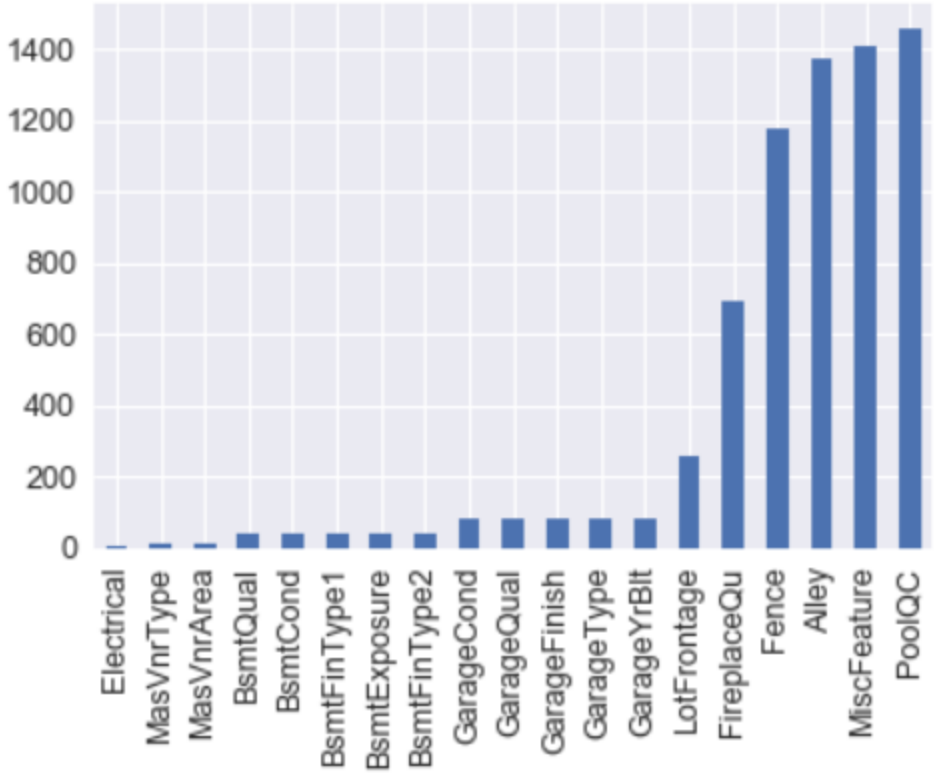
Data: Exploration

- Sale Price vs Year Built



Train Data: Missing Values

- Train Data will NA values

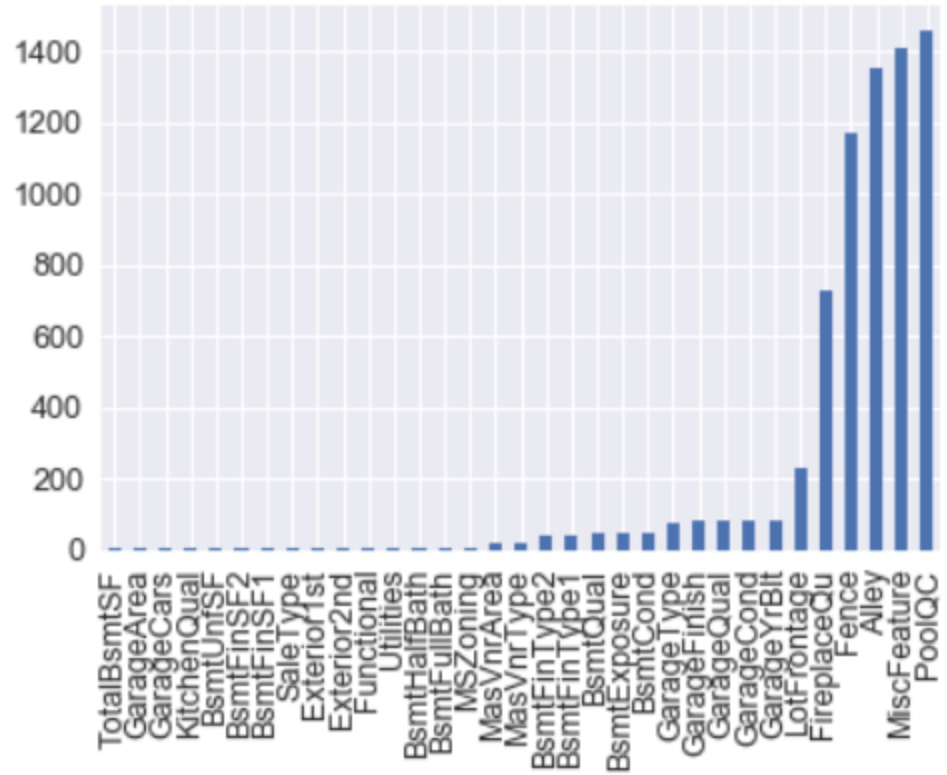


- In PoolQA (pool quality), NA is No pool.
- Alley: NA is No alley access.
- Fence: NA is No fence.
- FireplaceQu: NA is no fireplace

Test Data: Missing Values

- Test Data will NA values

	Total	Percent
PoolQC	1453	0.997940
MiscFeature	1405	0.964973
Alley	1349	0.926511
Fence	1168	0.802198
FireplaceQu	728	0.500000
GarageFinish	77	0.052885
GarageQual	77	0.052885
GarageCond	77	0.052885
GarageType	76	0.052198
BsmtCond	43	0.029533
BsmtExposure	42	0.028846
BsmtQual	42	0.028846
BsmtFinType1	40	0.027473
BsmtFinType2	40	0.027473
MSZoning	4	0.002747
Utilities	2	0.001374
Functional	2	0.001374
KitchenQual	1	0.000687



Data: Missing Values

- Following features are removed due to missing values:
  - LotFrontage: 17% of missing values.
  - MasVnrType
  - MasVnrArea
  - GarageYrBlt
  - Id

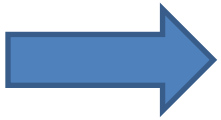
- Remove any row of NULL values that may be in dataset.

Categorical Data

- Convert text data to numerical by using One hot encoder from Sklearn.
- It converts categorical values into binary.
- Train Dataset

OLD Column

Alley
Grvl
NA
Pave



New Columns

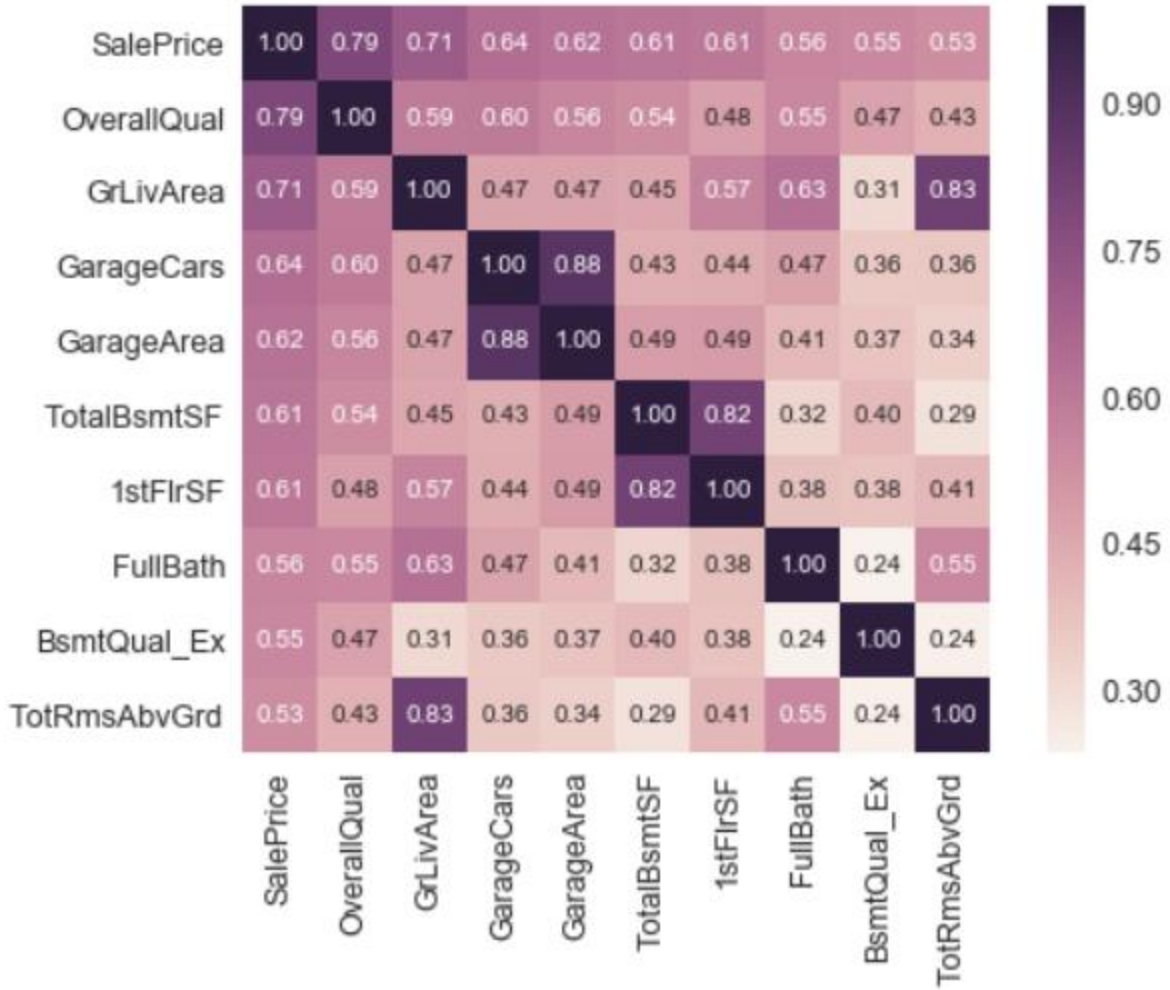
Alley_Grvl	Alley_Pave
1	0
0	0
0	1

Categorical Data

- Convert text data to numerical by using One hot encoder from Sklearn.
- Test Dataset may not have all the data values as train dataset so after One hot encoder transformation, test dataset will not have same number of columns as train dataset.
- To avoid this issue, we will have column names of train data set in test and if data is not available in test will be given zero value.

```
one_hot_encoded_training_predictors  = pd.get_dummies(df_train)
one_hot_encoded_test_predictors  = pd.get_dummies(df_test)
final_train, final_test = one_hot_encoded_training_predictors.align(one_hot_encoded_test_predictors,  join='left',  axis=1)
```

Correlation Matrix



Feature Selection

- Run data through sklearn cross validation *cross_val_score function*.
- Data is split into k folds and (k-1) folds are used to train the data and one fold is to test. Repeat this process number of times and take the average of the target values.
- This process works in avoiding **overfitting**.

Linear Regression

- Linear regression model is build on the processed train data.
- Cross validation cross_val_score function is used five times.

cross_validation.cross_val_score(model_name,X,y,cv=5, scoring='mean_squared_error')

- *Root mean square of Linear regression is* 35154.66

Random Forest Regression

- Random forest is an estimator that fits classifying decision tree on various samples of data and uses average to improve the predictive accuracy and control overfitting.
- Random forest regression model

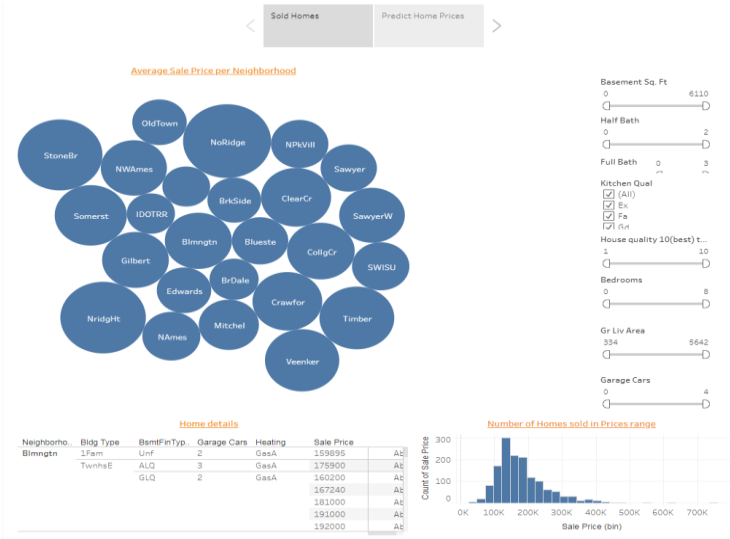
rf = RandomForestRegressor(n_estimators = 1000, random_state = 42)rf.fit(X_train, y_train)

- Root mean square error of Random Forest Regressor is 29921.69

- Prediction on test data is going to be done using this model as this has lower root mean square error value.

Website

- Research tool for home buyers and seller is the following website:
- https://public.tableau.com/profile/neeraj3209#!/vizhome/AmnesHousing/AmnesHousing



References

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