FLIP ROBO

HOUSING: PRICE PREDICTION

Submitted by:

SANDES P

ACKNOWLEDGMENT

I used different online sources and research papers to understand the domain. Some of them are given below. I used informative tutorials to follow some steps in the task from websites like geeksforgeeks, stackoverflow, etc.

http://www.econ.mq.edu.au/

https://www.livingin-australia.com/australian-house-prices/

INTRODUCTION

- Business Problem Framing

Houses are one of the necessary need of each and every person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors in the world's economy. It is a very large market and there are various companies working in the domain. Data science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases. Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for housing companies. Our problem is related to one such housing company.

A US-based housing company named Surprise Housing has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price. For the same purpose, the company has collected a data set from the sale of houses in Australia. The data is provided in the CSV file below.

The company is looking at prospective properties to buy houses to enter the market. You are required to build a model using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest in them or not. For this company wants to know:

• Which variables are important to predict the price of variable?

• How do these variables describe the price of the house?

Business Goal:

You are required to model the price of houses with the available independent variables. This model will then be used by the management to understand how exactly the prices vary with the variables. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns. Further, the model will be a good way for the management to understand the pricing dynamics of a new market.

- ## Conceptual Background of the Domain Problem

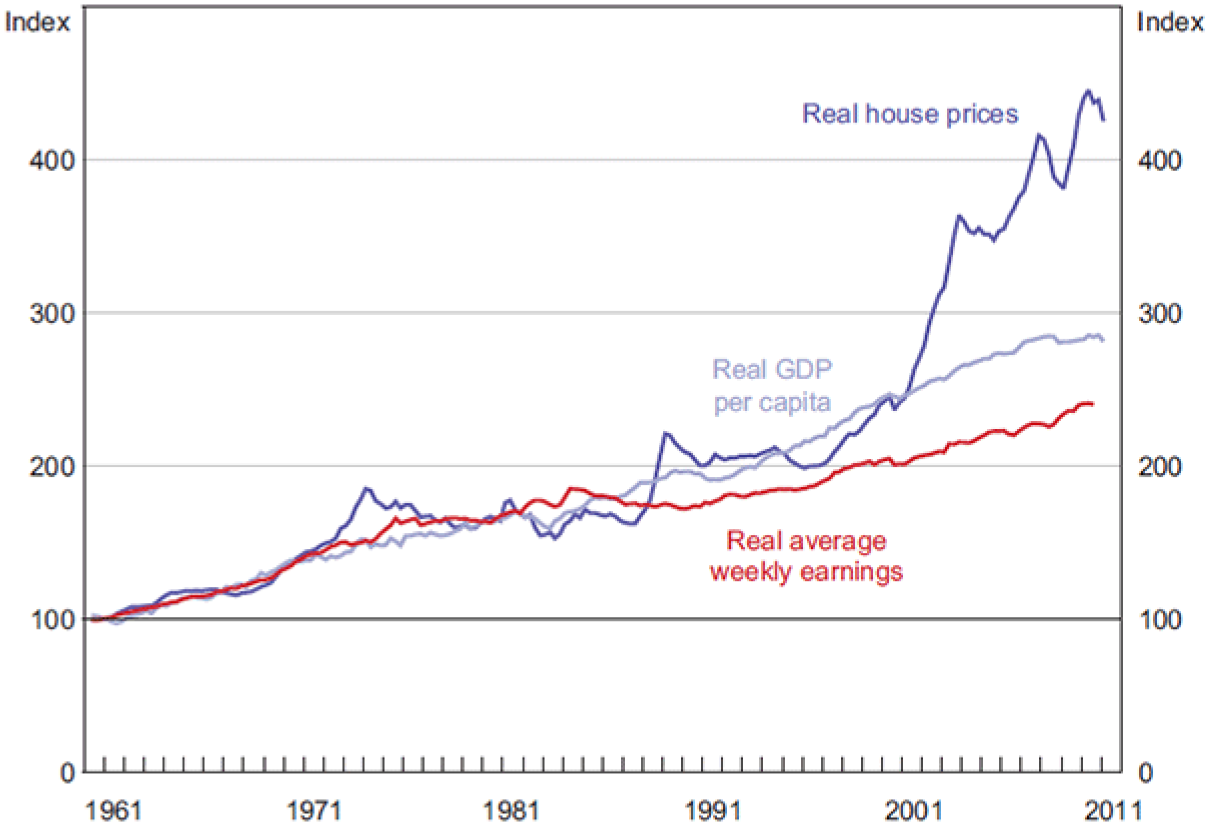
This domain is related to the real estate field in Australia. The aim of this project is to create a machine learning model that predicts the prices of houses using independent feature values. The price of a house not only depends on how big it is but also a lot of other factors like where it stands, how old it is, how far the street etc.

- ## Review of Literature

We have done some researches on the domain to gain more knowledge. The main feature that determines the price of a house is its location. In different cities price varies drastically. At present, the average price of a house lies between $150,000 and $280,000. But in our dataset, we are working on the sales from 2006-2011. The price of houses started increasing rapidly in the first decade of the 21st. The percentage of increase of price was 6%/year during 1995-2005 with an annual increase of 15% during 2003-2005.

Figure 1: Real House Prices, GDP per Capita and Earnings

1960/61 = 100



Sources: Abelson and Chung (2005); ABS; Australian Treasury; REIA

Due to the increase in household income, the demand for houses has increased in this time span. According to researches, demography, economic growth, and wealth effect are the three major long-term factors which determine the house price.

Tax benefits, underlying socio-demographic factors, institutional reforms, increasing income, broader development within the economy, and capital markets contributed to increased investment demand for housing. A sharp downward correction was observed in the equity market around the world in mid-2000; this caused a capital switch from shares to other assets like bonds and real estate. A sharp increase in the flow of debt-financed private investment into residential property occurred with the perception that investment in property is preferable in terms of risk and return.

Demographic factors significantly affect housing prices in the long run. Increasing population growth increases the number of households in a country which in turn increases the housing demand and thus housing price

A positive relationship exists between household income and housing demand. In a country like Australia, the income elasticity of demand for housing is likely to be one or greater. So the demand for housing increases at least proportionately with income. The long-term trend in real income is important in explaining house prices.

In every society, housing is generally considered as the major store of wealth. In Australia gross housing assets account for more than half of total personal wealth. Owner-occupiers and landlord investors feel wealthier with the rising prices of existing houses. This is called the ‘wealth effect’ which leads to an increase in consumption expenditure. As a result, aggregate demand, and thus economic growth, occurs which in turn support rising housing prices through a self-reinforcing cycle. This factor also contributed to the recent price hike of Australian housing.

- ## Motivation for the Problem Undertaken

  We have the data of different house prices with independent factors. Our objective is to find the important features that affect the house price and to build a model that predicts the house prices given the independent features are provided. This model will be helpful for people who are looking for new houses in Australia to estimate their expenditure.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  Here we tried to use different regression modeling techniques such as SVR, KNN, linear regression, etc to find the best accuracy.

- Data Sources and their formats

  The data source is a study done about the house prices in Australia during 2006-2010. There is one train dataset and test data set provided. Both data are provided in CSV format. The dataset has 81 columns and 1168 rows. The snapshot of the data is given below

Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope
127	120	RL	NaN	4928	Pave	NaN	IR1	Lvl	AllPub	Inside	Gtl
889	20	RL	95.0	15865	Pave	NaN	IR1	Lvl	AllPub	Inside	Mod
793	60	RL	92.0	9920	Pave	NaN	IR1	Lvl	AllPub	CulDSac	Gtl
110	20	RL	105.0	11751	Pave	NaN	IR1	Lvl	AllPub	Inside	Gtl
422	20	RL	NaN	16635	Pave	NaN	IR1	Lvl	AllPub	FR2	Gtl

The description of the columns

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
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

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

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
CompShg	Standard (Composite) Shingle
Membran	Membrane
Metal	Metal
Roll	Roll
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

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

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

NA None

MiscVal: $Value of miscellaneous feature

MoSold: Month Sold (MM)

YrSold: Year Sold (YYYY)

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

Con Contract 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 Preprocessing Done
    - The first task to treat the missing values.
    - Checked the null values and dropped the columns which have null values of more than 80%. Alley, MiscFeature, Fence, PoolQC are the columns that are dropped.
    - Plotted boxplot for the numerical values to find out which columns have outliers and found out the columns
    - The outlier values are replaced by nan values
    - The nan values are replaced by the median in numerical columns and by mode in categorical columns
    - Found out the categorical columns which have only one unique value and drop those columns. We get one unique value for the utility column here.
    - EDA
    - Plotted different scatterplots between sale price and other variables to get the relationship of columns with sale price
    - Dropped the single values columns from the numerical columns also
    - Found out the skewness of all the columns
    - Plotted distplots of all variables
    - Plotted count plots of all categorical variable
    - Found out and plotted the correlation values using heatmap
    - Compared the correlation values of skewed columns and removed the skewness of less correlated values
    - Created three new columns using the year related features

        - train["builtTosale"] = train["YrSold"] - train["YearBuilt"]
        - train["modiTosale"] = train["YrSold"] - train["YearRemodAdd"]
        - train["garageTosale"] = train["YrSold"] - train["GarageYrBlt"]
    - Labelencoded the categorical columns
    - Scaled the columns
- Data Inputs- Logic- Output Relationships

  * In sub class, Price decrease with increase in value within sub division of sub class(in the base of stories and styles)

  * Lot Fontage has no direct relationship with price. It might have dependent with other features

  * Price increases with overall quality

  * In over all condition, Price is higher in the middle range

  * Year built is approximately +ve correlated with price, but we have to check with other parameters too

  * BsmtFinSF2 column has only 1 value which is zero. so we can drop that

  * Other basement area(sq ft) related features is almost directly propotional to saleprice

  * Floor squre feet is + correlated with sale price

  * LowQualFinSF column has only 1 value which is zero. so we can drop that

  * BsmtHalfBath,kitchenabvgr,EnclosedPorch,3ssnporch, screenporch,poolarea,miscval columns have only 1 value which is zero. so we can drop that

  * GrLivArea is directly correlated with price

  * FullBath, BedroomAbvgr,fireplaces, GarageYrBlt,GarageCars is directly correlated with price

* columns MSSubClass, OverallQual, OverallCond, BsmtFullBath, Fullbath, HalfBath, BedroomAbvGr, TotRmsAbvGrd, FirePlaces, GarageCars, MoSols, YrSold can be treated as categorical values

* basment full bath and basementunitsf are -vely correlated

* First floor sqft and basement sqft are highly +ve correlated

* Overall quality and price are highly correlated

* garage car and garage area are +vely correlated

* Price decreases from MZZoning RL to C

* HOuses in Paved street is more costly

* Houses in flat level have more price and banked houses cost less

* House price decreases with slope of the place

* Typical basement houses have more price than excellent basement . So there is less over all consideration for basement

* Houses with gas forced heating have more price

* Central air conditioned houses have more price

* Mixed wired houses have lesser price

* Houses with attached garage have more price

* Houses with paved driveway have more price

- Hardware and Software Requirements and Tools Used

  The model building is done on a computer with specifications as follows

  - 8GB RAM
  - i5 7th gen processor

  Softer requirements

  - Python
  - Jupyter notebook

Libraries

- Numpy
- Pandas
- Matplotlib
- Seaborn
- SKLearn

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)
    - Treating outliers, missing values
    - EDA for finding the relationship
    - Feature Engineering
    - Encoding
    - Scaling

- Testing of Identified Approaches (Algorithms)
    - SVR
    - KNN
    - RandomForest
    - Linear Regression
    - Ridge
- Run and Evaluate selected models

```python
models = {"SVR":SVR(),"KNN":KNeighborsRegressor(), "RandomForest":RandomForestRegressor(),
          "LinearRegression":LinearRegression(), "Ridge":Ridge() }
acc = {}
mod_list = []
for i in models:
    mod = i
    mod = models[i]
    mod.fit(x_train, y_train)
    pred = mod.predict(x_test)
    r2_sc = r2_score(y_test,pred)
    acc[i] = r2_sc
    mod_list.append(mod)
print(acc)
```

We have used SVR, KNN, RandomForestRegressor, LinearRegression, and ridge. The result of these algorithm as follows

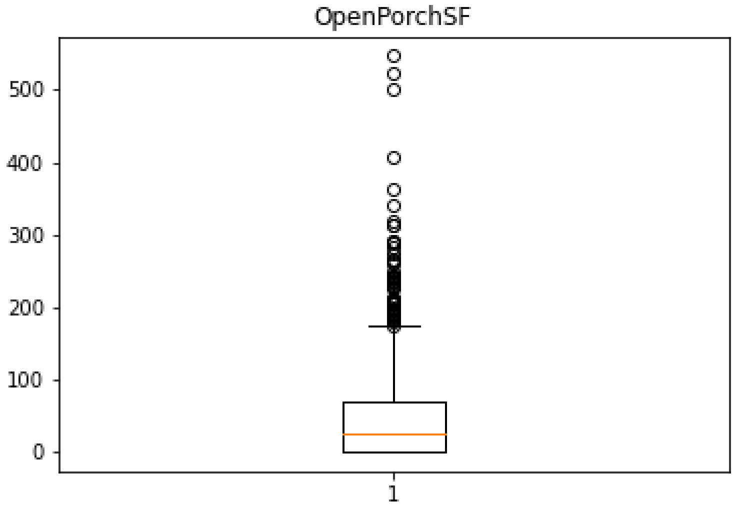
```
    mod.fit(x_train, y_train)

{'SVR': -0.08816674306526151, 'KNN': 0.8113179937166635, 'RandomForest': 0.8917710201079561, 'LinearRegression': 0.8676123464894193, 'Ridg
e': 0.867670566713543}
```

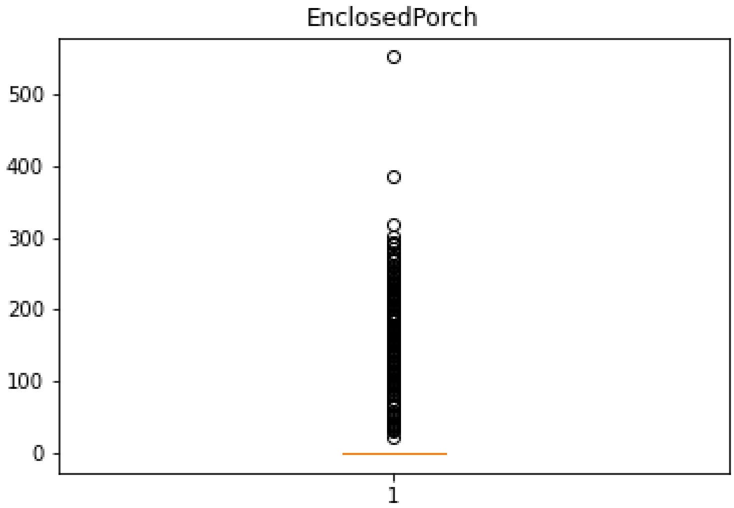
- Key Metrics for success in solving problem under consideration

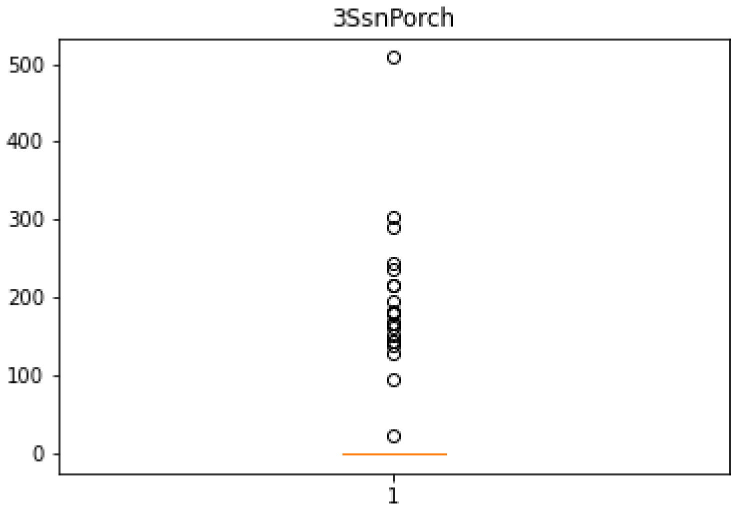
  Here we used r2score since this is a regression problem

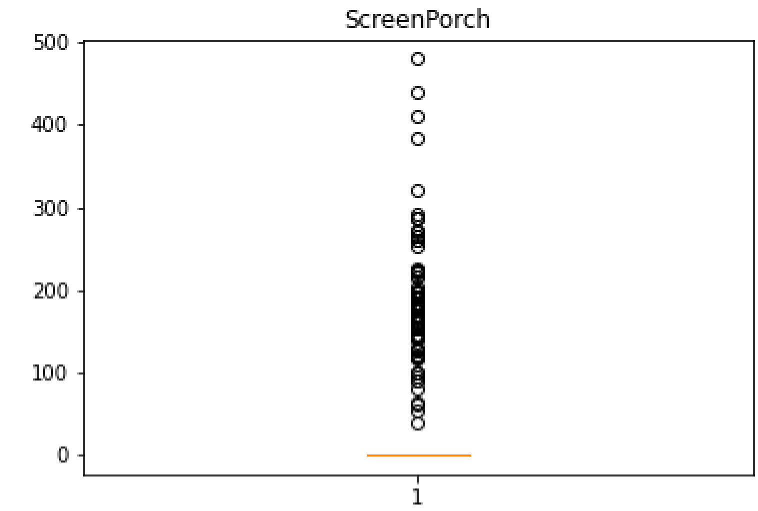
- Visualizations

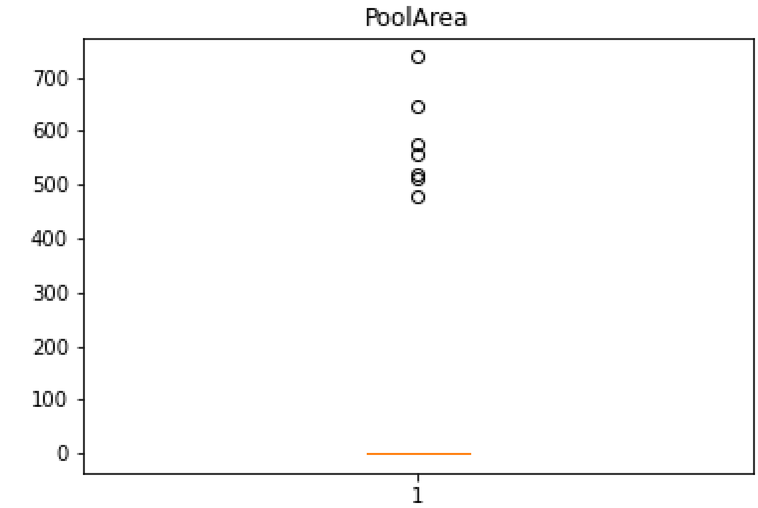
  Plotted outliers using boxplot

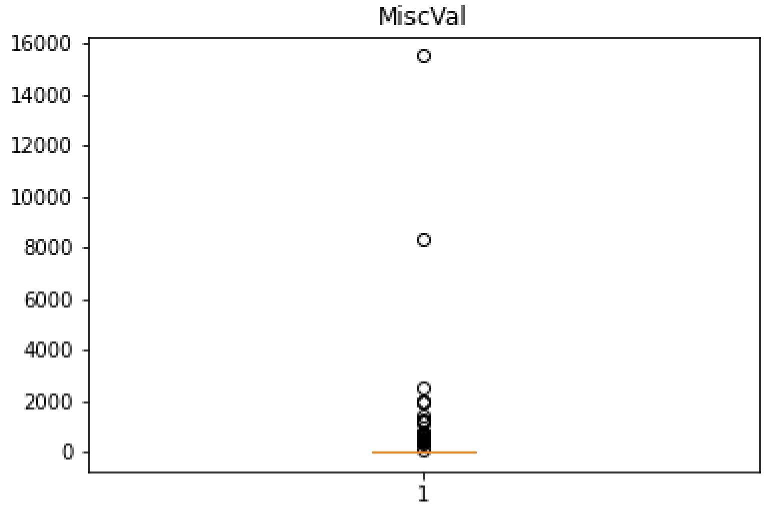

OpenPorchSF
500
400
300
200
100
0
1

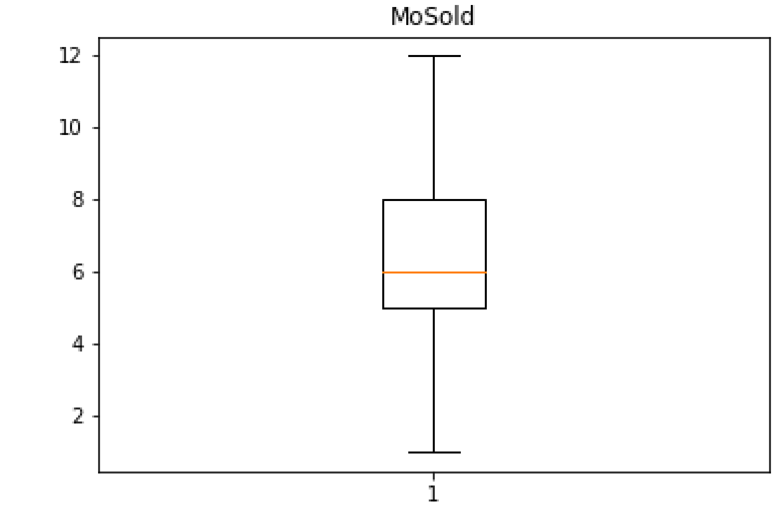

EnclosedPorch
500
400
300
200
100
0
1

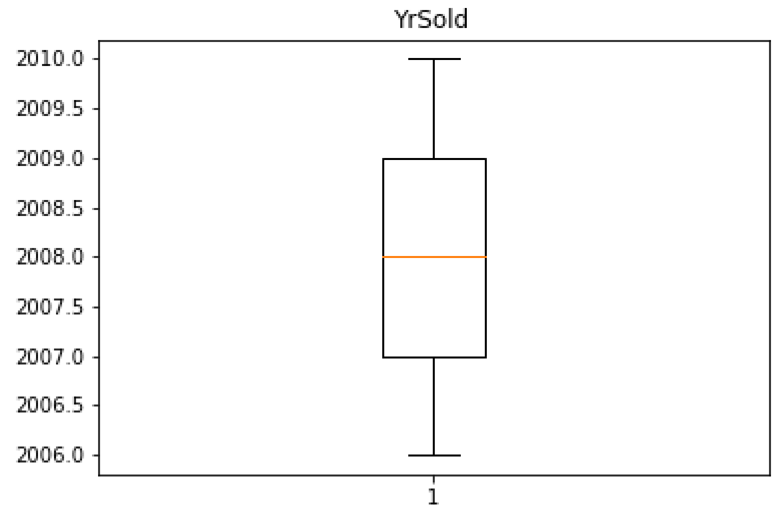

3SsnPorch
500
400
300
200
100
0
1


ScreenPorch
500
400
300
200
100
0
1

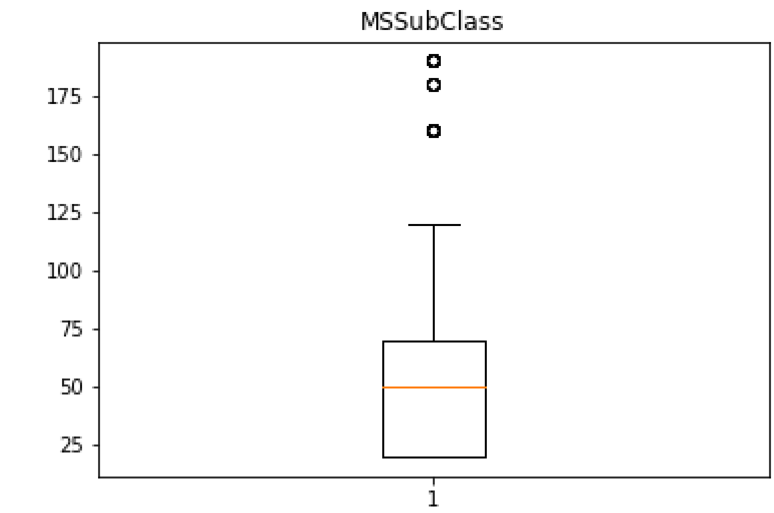

PoolArea
700
600
500
400
300
200
100
0
1

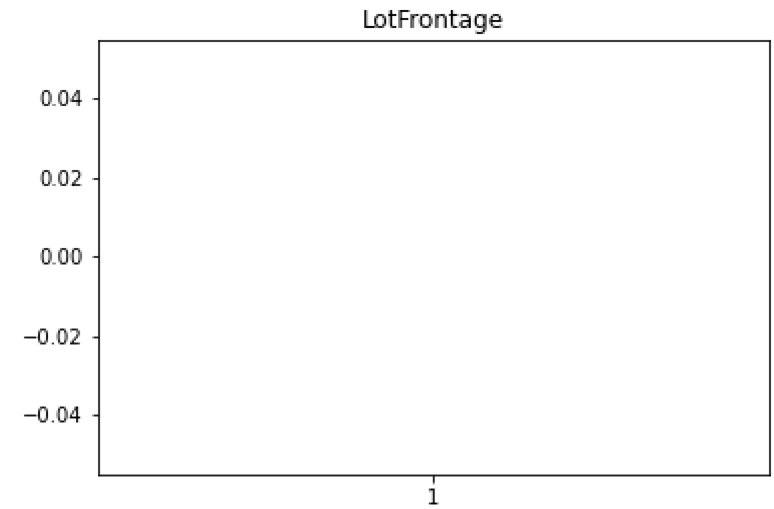

MiscVal
16000
14000
12000
10000
8000
6000
4000
2000
0
1

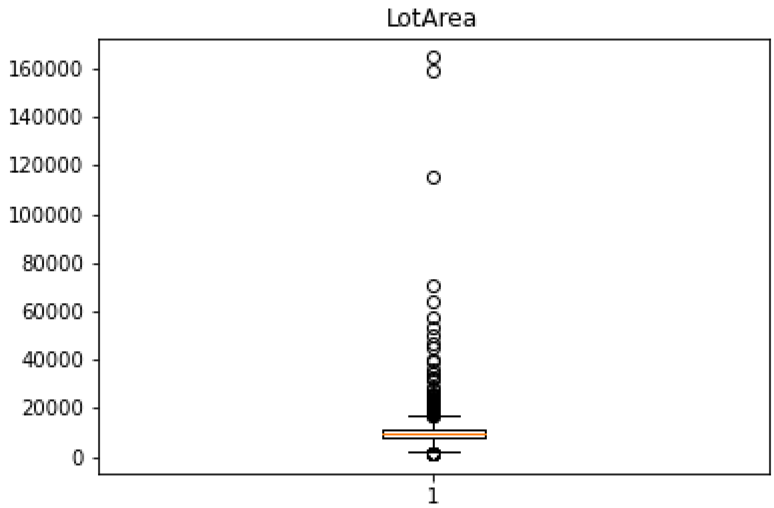

MoSold
12
10
8
6
4
2
1

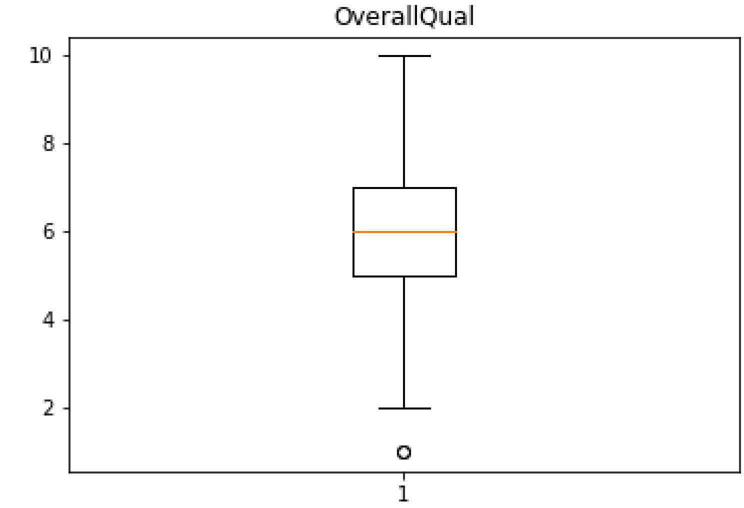

YrSold
2010.0
2009.5
2009.0
2008.5
2008.0
2007.5
2007.0
2006.5
2006.0
1

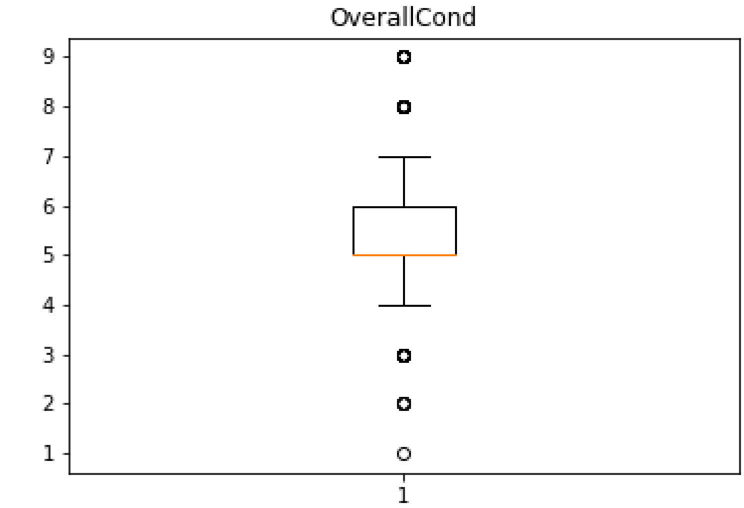

SalePrice
700000
600000
500000
400000
300000
200000
100000
0
1

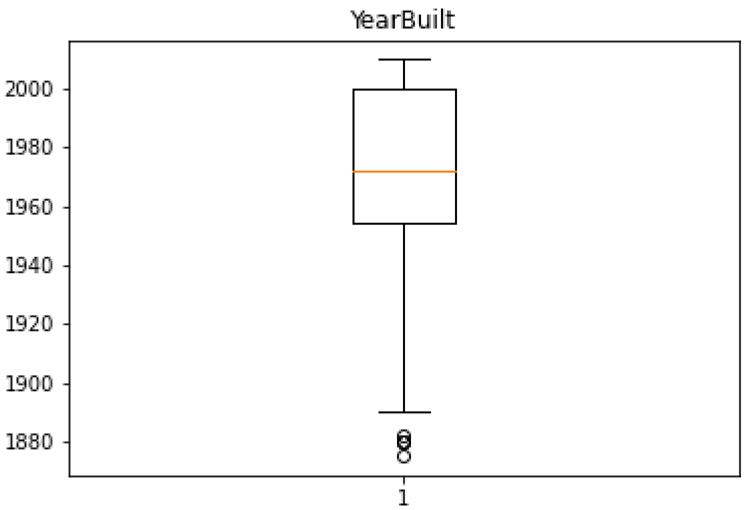

MSSubClass
175
150
125
100
75
50
25
1

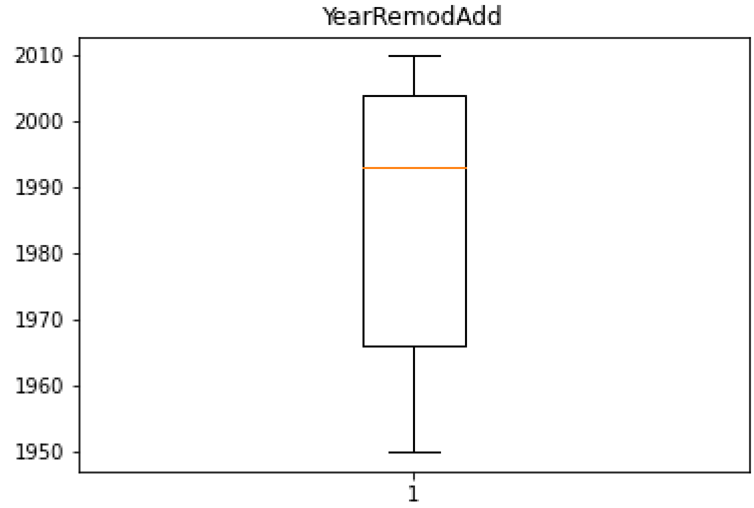

LotFrontage
0.04
0.02
0.00
−0.02
−0.04
1

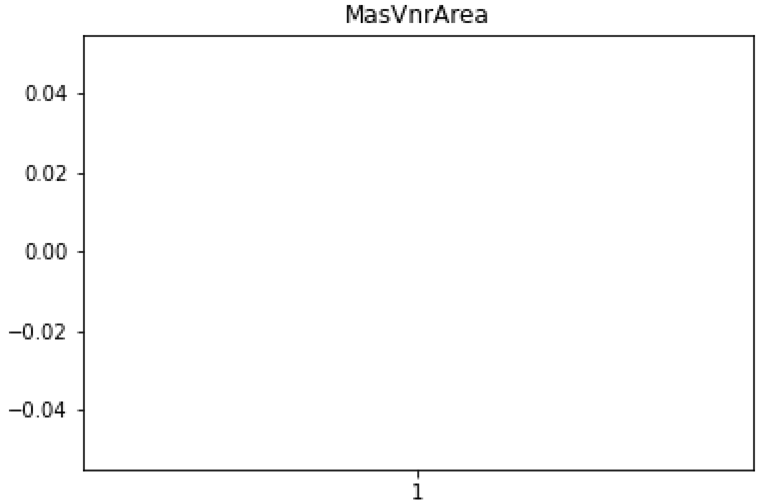

LotArea
160000
140000
120000
100000
80000
60000
40000
20000
0
1

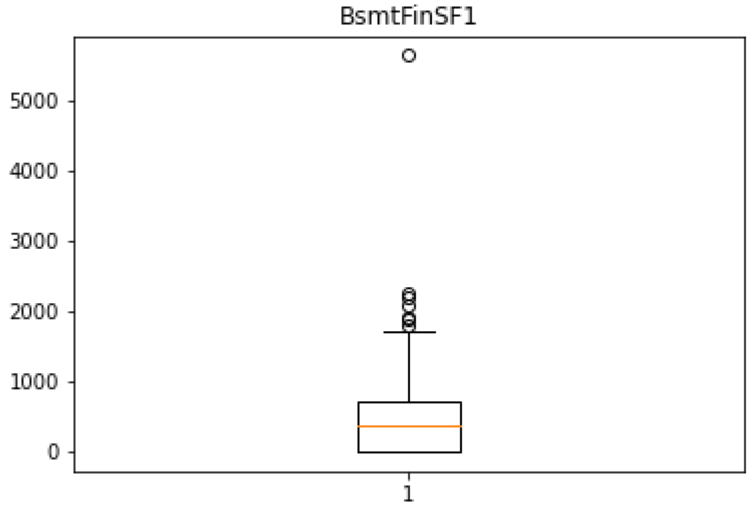

OverallQual
10
8
6
4
2
1

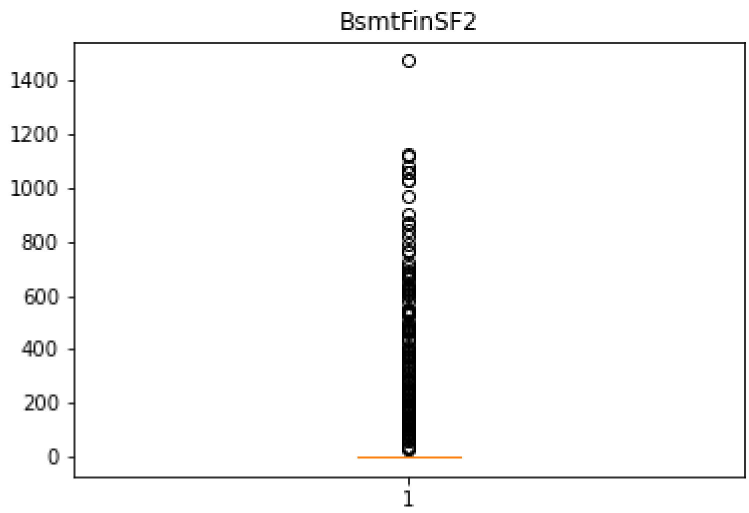

OverallCond
9
8
7
6
5
4
3
2
1
1

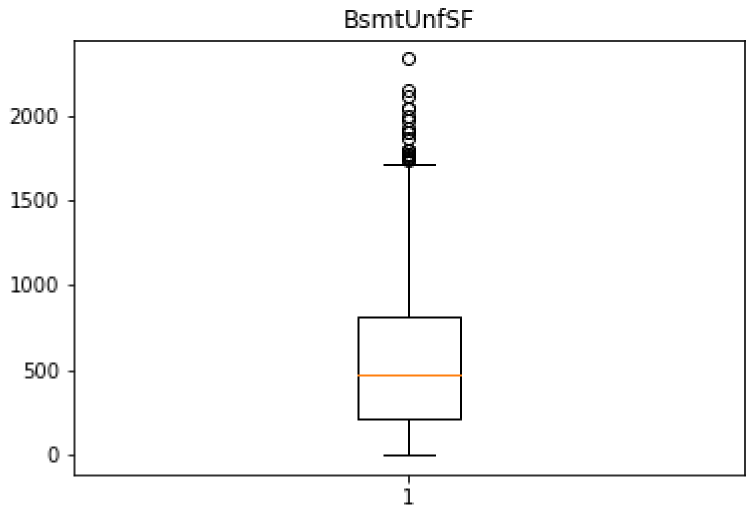

YearBuilt
2000
1980
1960
1940
1920
1900
1880
1

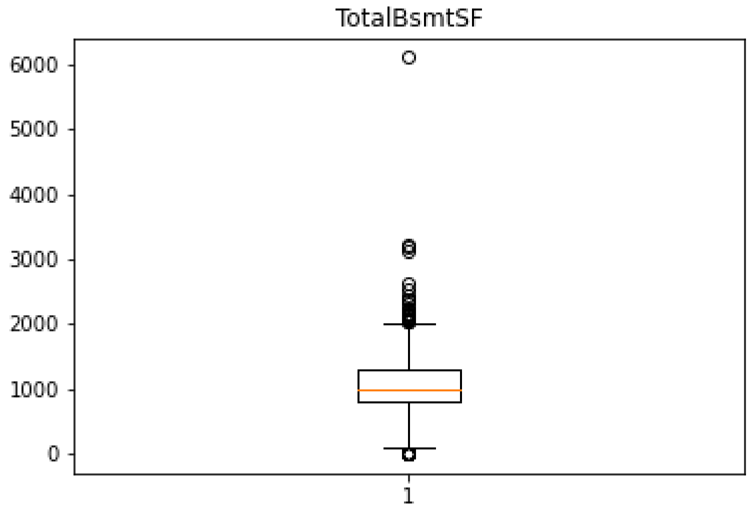

YearRemodAdd
2010
2000
1990
1980
1970
1960
1950
1

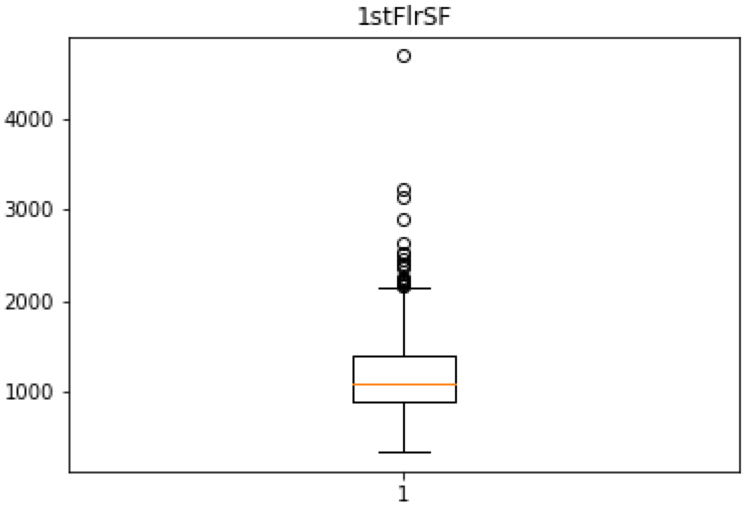

MasVnrArea
0.04
0.02
0.00
−0.02
−0.04
1

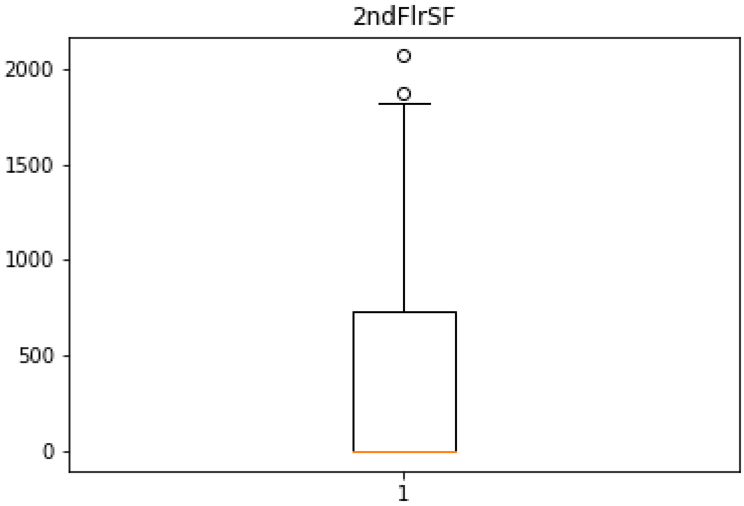

BsmtFinSF1
5000
4000
3000
2000
1000
0
1

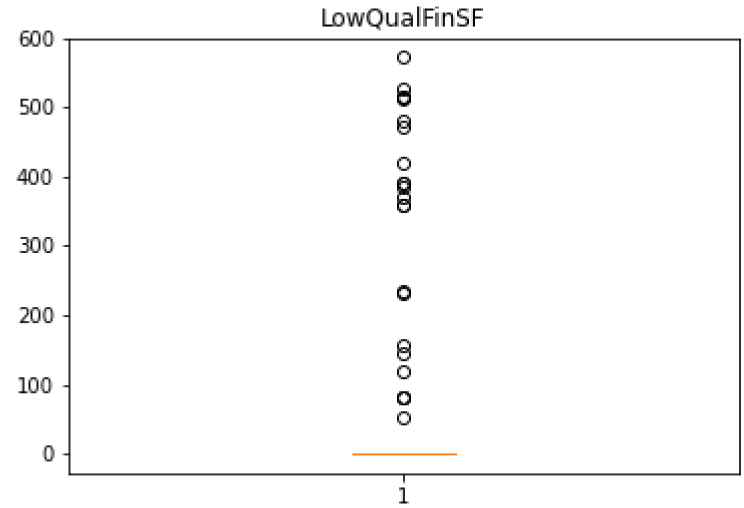

BsmtFinSF2
1400
1200
1000
800
600
400
200
0
1

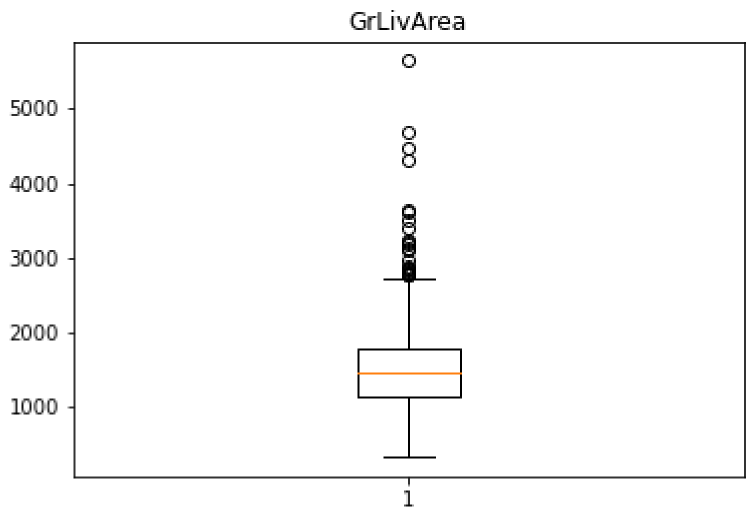

BsmtUnfSF
2000
1500
1000
500
0
1

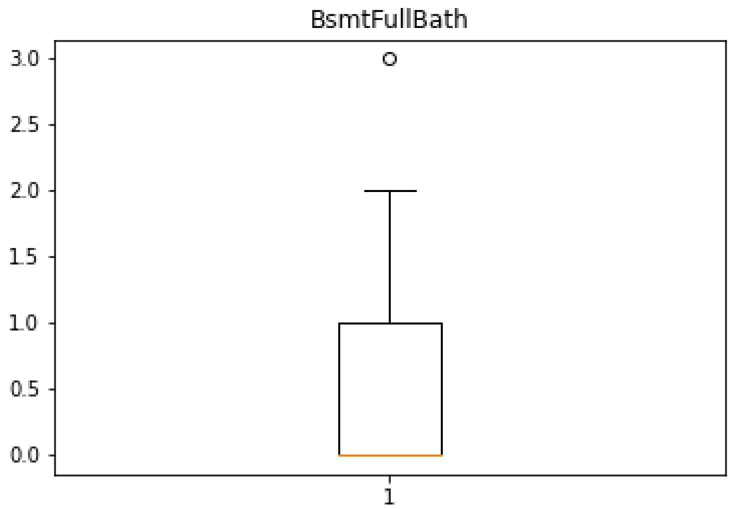

TotalBsmtSF
6000
5000
4000
3000
2000
1000
0
1

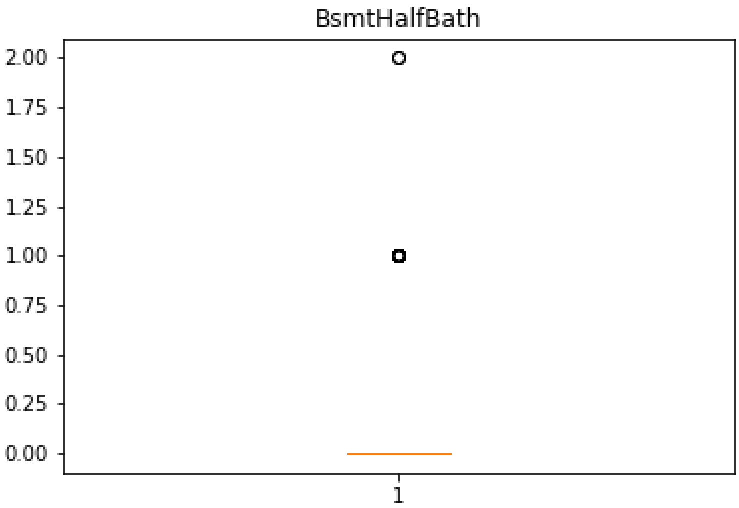

1stFlrSF
4000
3000
2000
1000
1

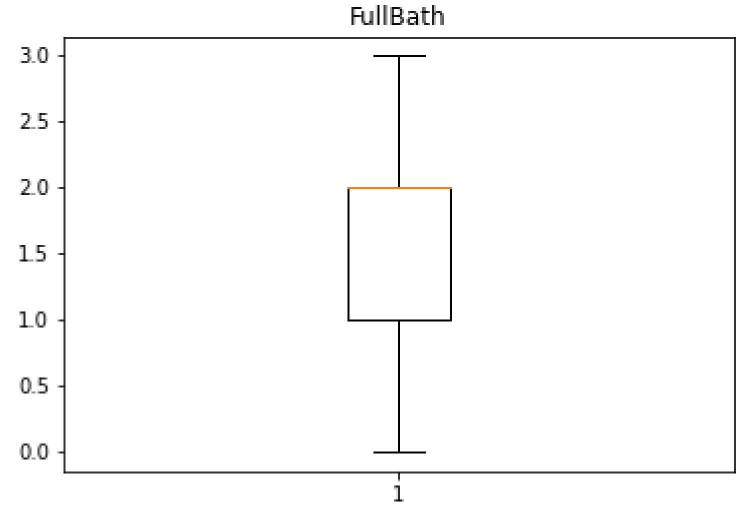

2ndFlrSF
2000
1500
1000
500
0
1

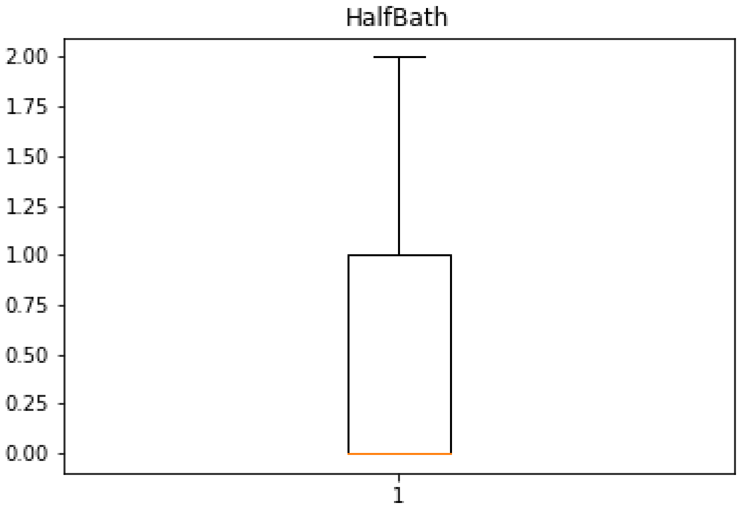

LowQualFinSF
600
500
400
300
200
100
0
1

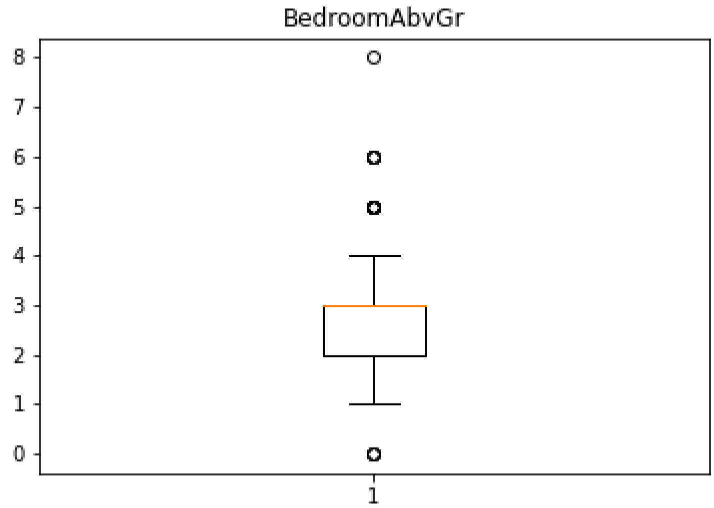

GrLivArea
5000
4000
3000
2000
1000
1

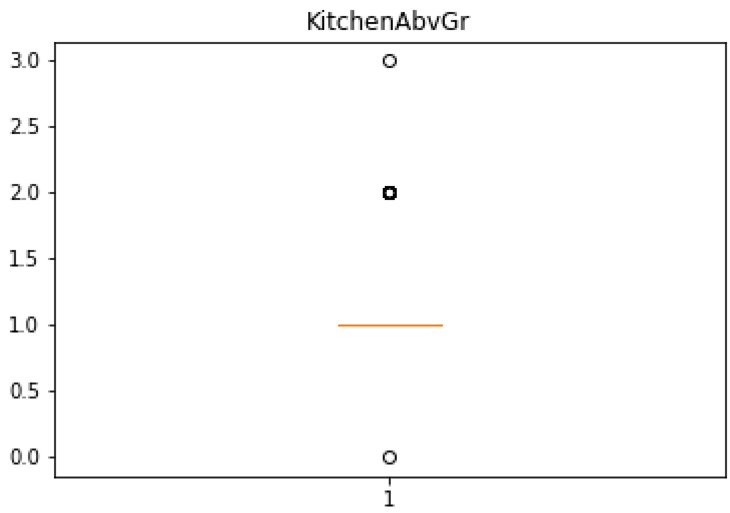

BsmtFullBath
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1

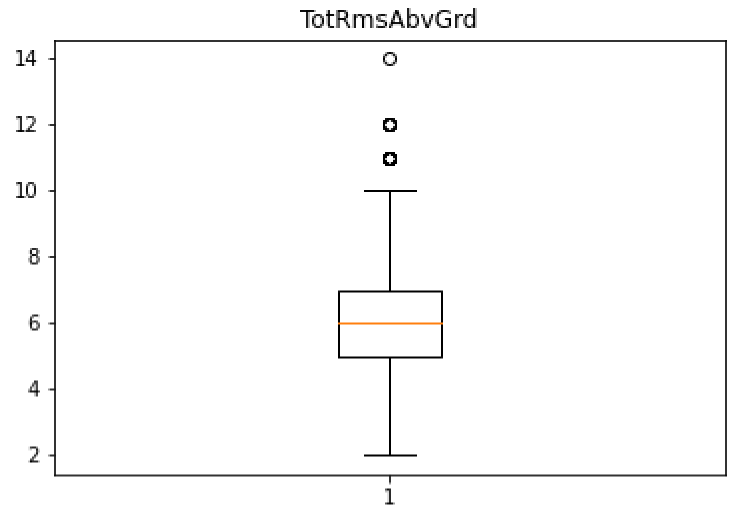

BsmtHalfBath
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
1

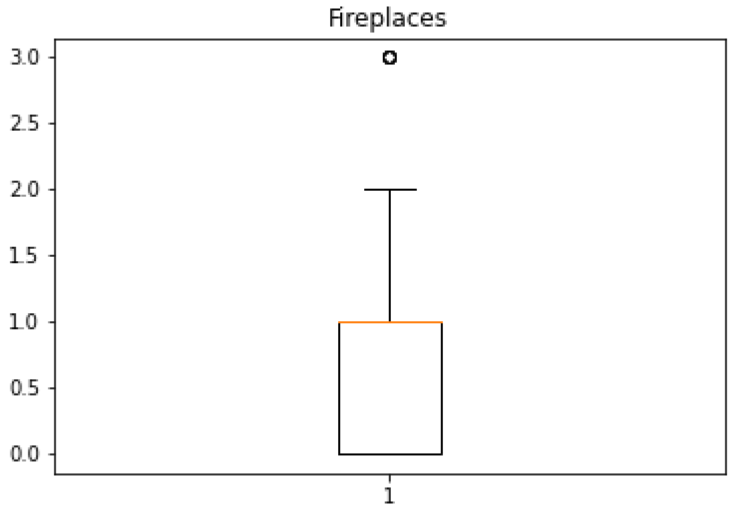

FullBath
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1

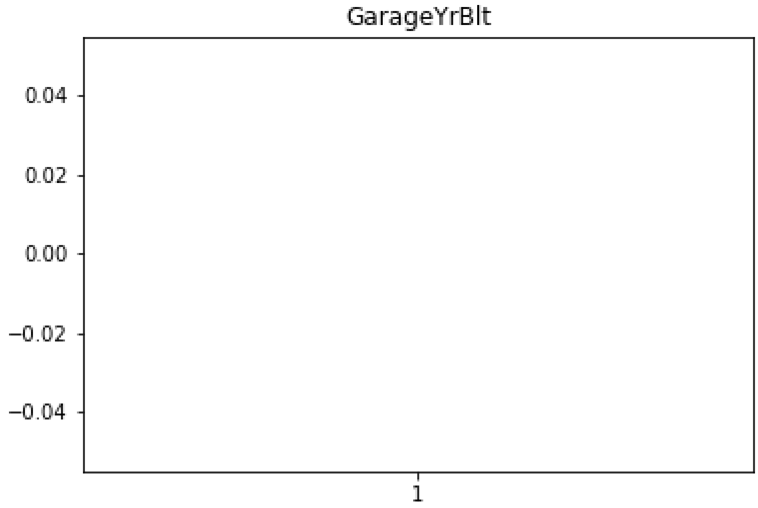

HalfBath
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
1

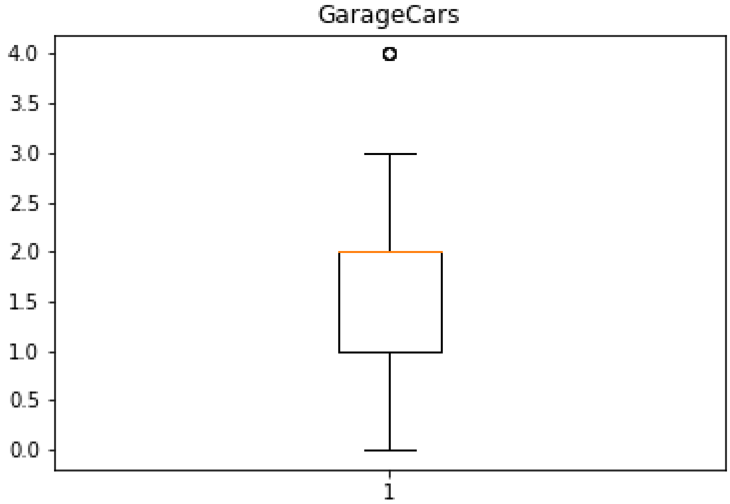

BedroomAbvGr
8
7
6
5
4
3
2
1
0
1

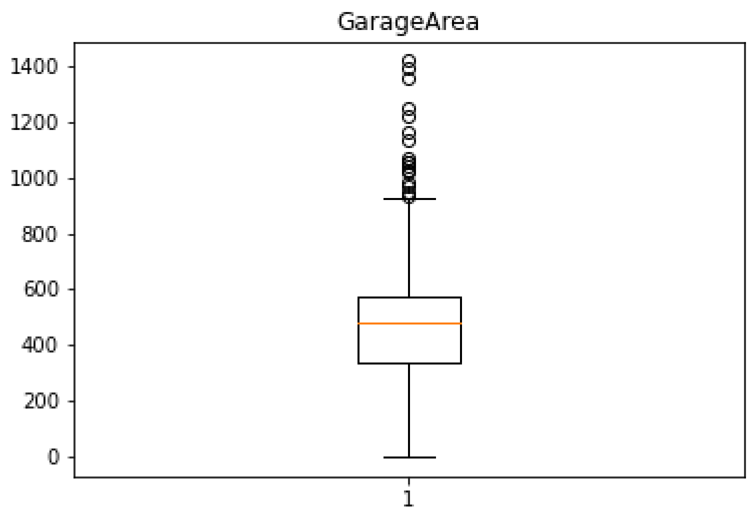

KitchenAbvGr
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1

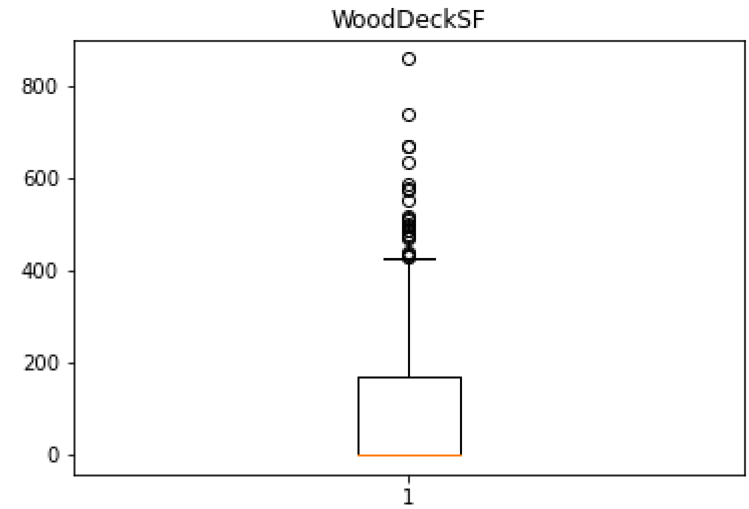

TotRmsAbvGrd
14
12
10
8
6
4
2
1


Fireplaces
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1

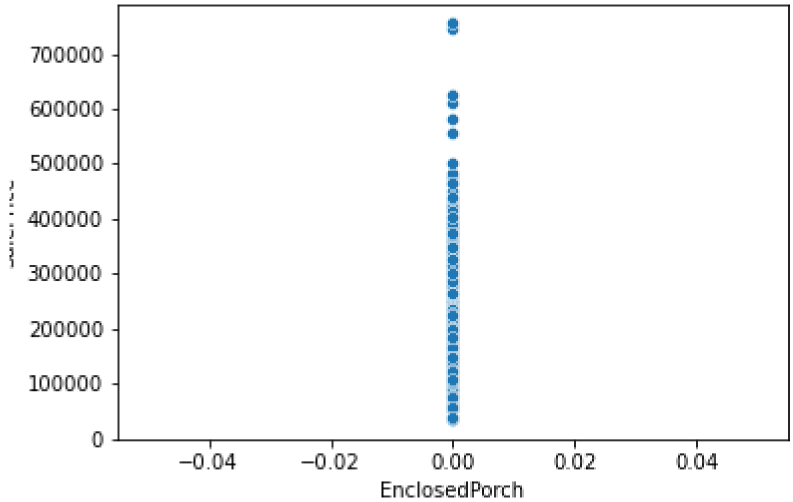

GarageYrBlt
0.04
0.02
0.00
−0.02
−0.04
1


GarageCars
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1

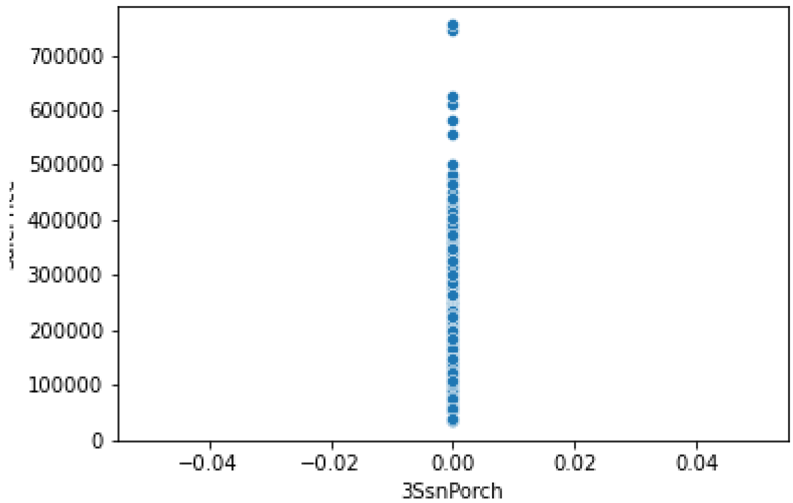




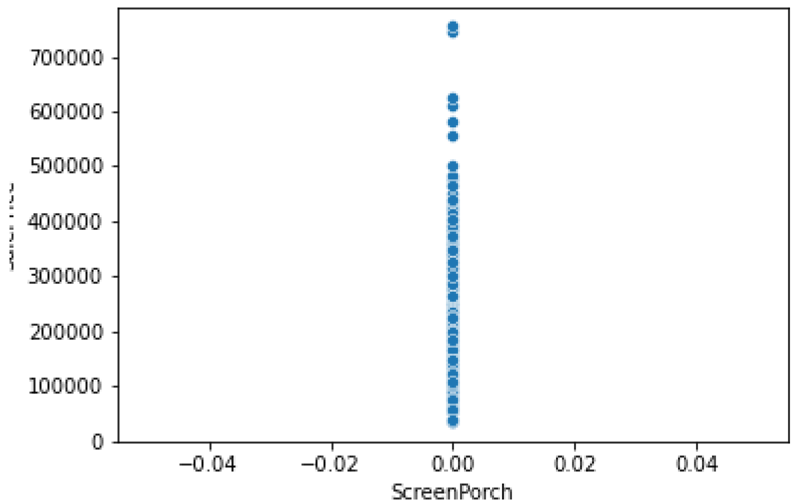
For EDA, Scatterplot, countplot, distplot and heatmap is used



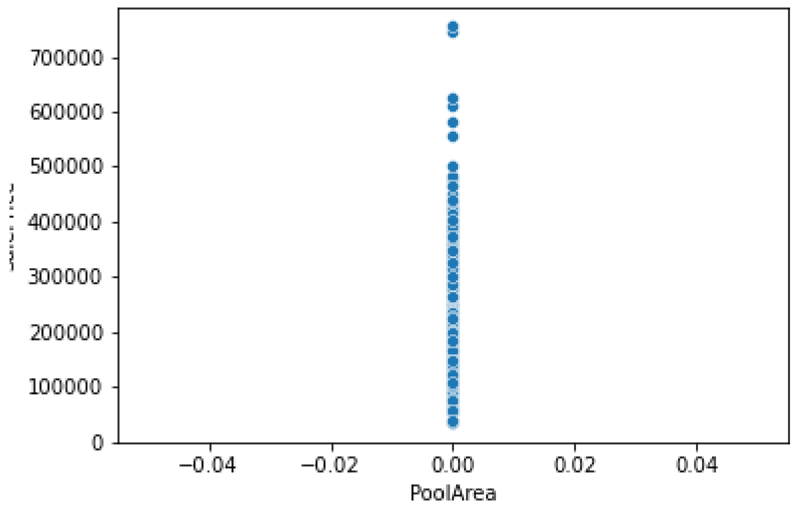
700000
600000
500000
400000
300000
200000
100000
0
−0.04
−0.02
0.00
0.02
0.04
EnclosedPorch



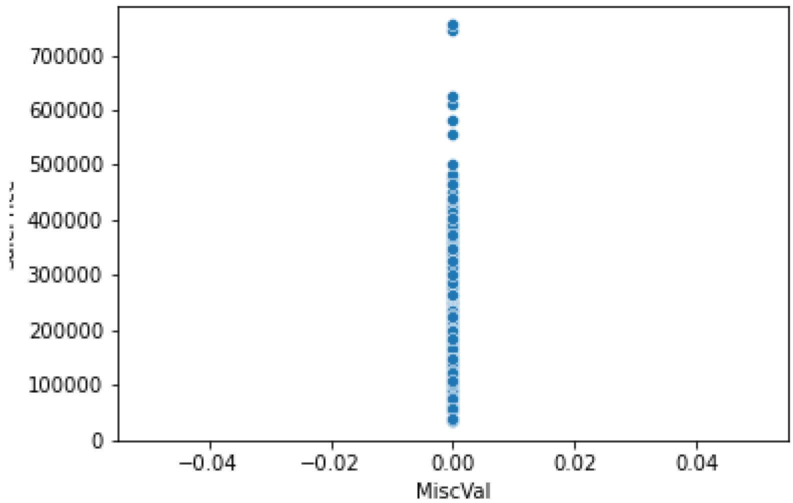
700000
600000
500000
400000
300000
200000
100000
0
−0.04
−0.02
0.00
0.02
0.04
3SsnPorch



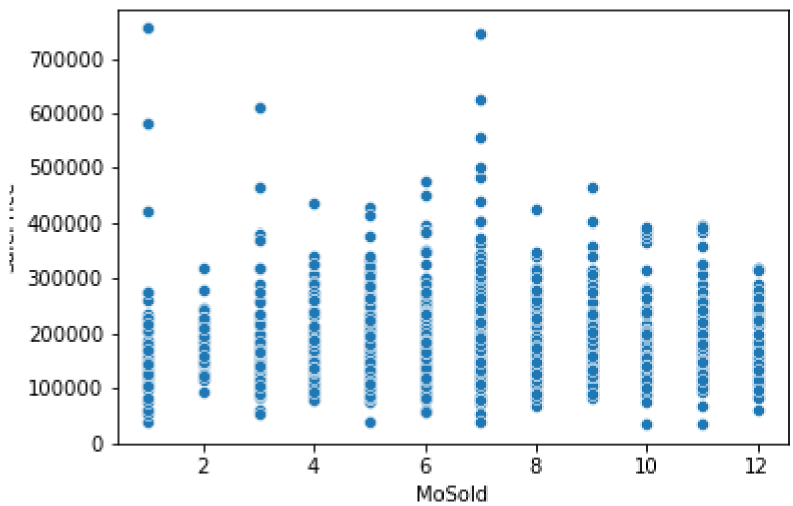
700000
600000
500000
400000
300000
200000
100000
0
−0.04
−0.02
0.00
0.02
0.04
ScreenPorch



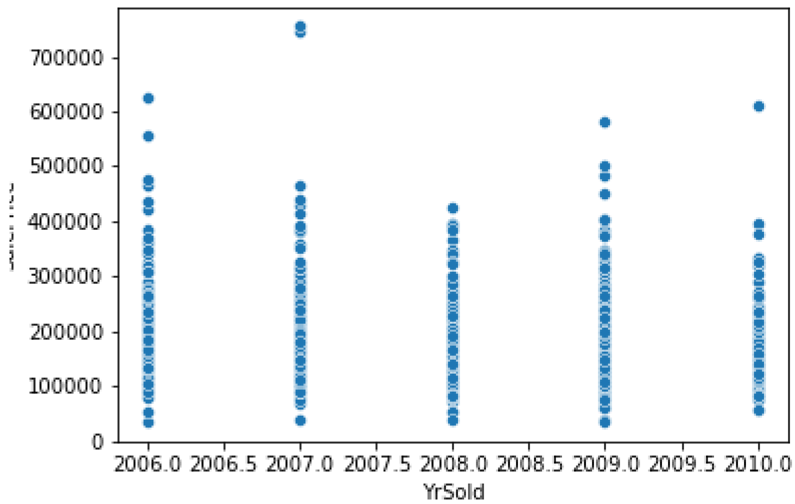
700000
600000
500000
400000
300000
200000
100000
0
−0.04
−0.02
0.00
0.02
0.04
PoolArea



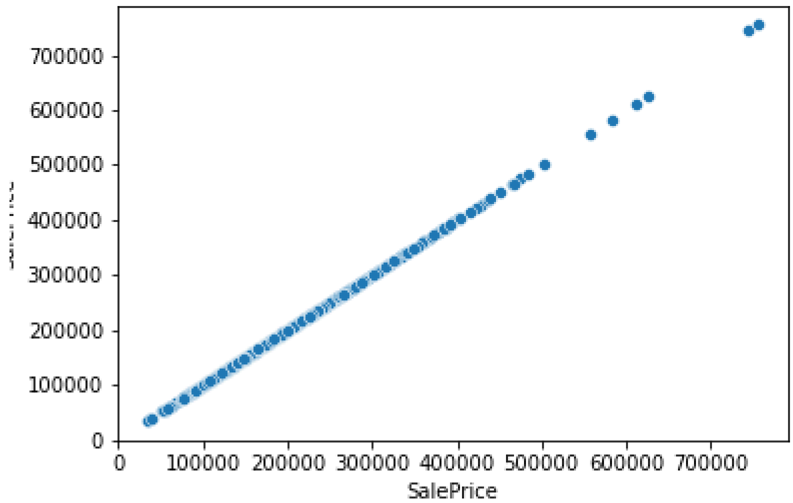
700000
600000
500000
400000
300000
200000
100000
0
−0.04
−0.02
0.00
0.02
0.04
MiscVal



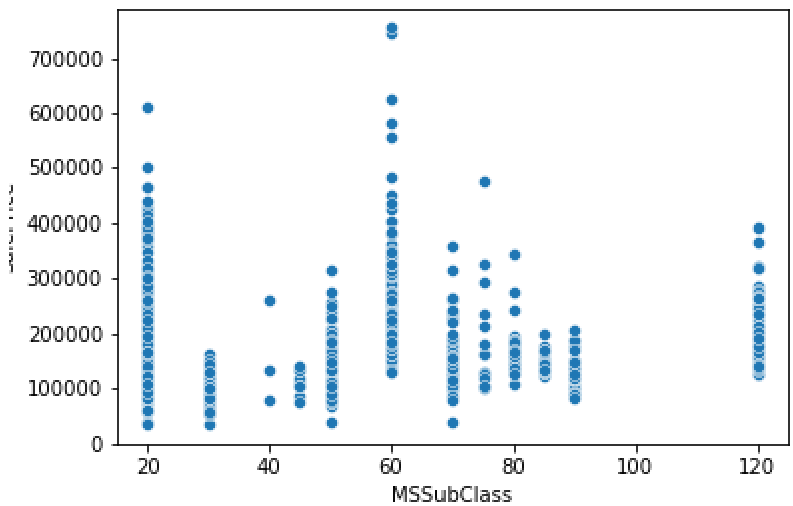
700000
600000
500000
400000
300000
200000
100000
0
2
4
6
8
10
12
MoSold



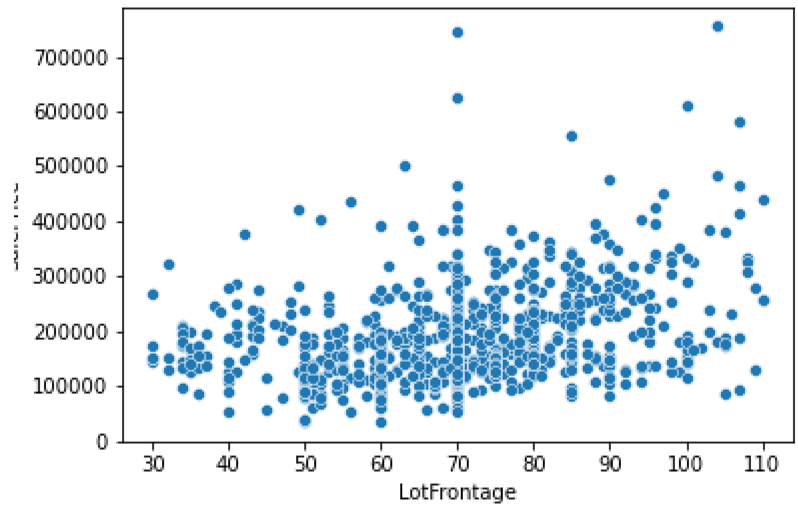
700000
600000
500000
400000
300000
200000
100000
0
2006.0 2006.5 2007.0 2007.5 2008.0 2008.5 2009.0 2009.5 2010.0
YrSold

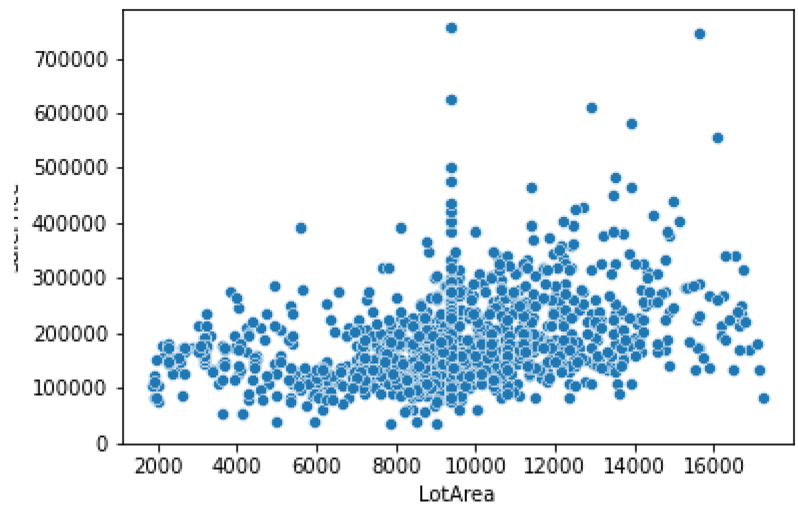


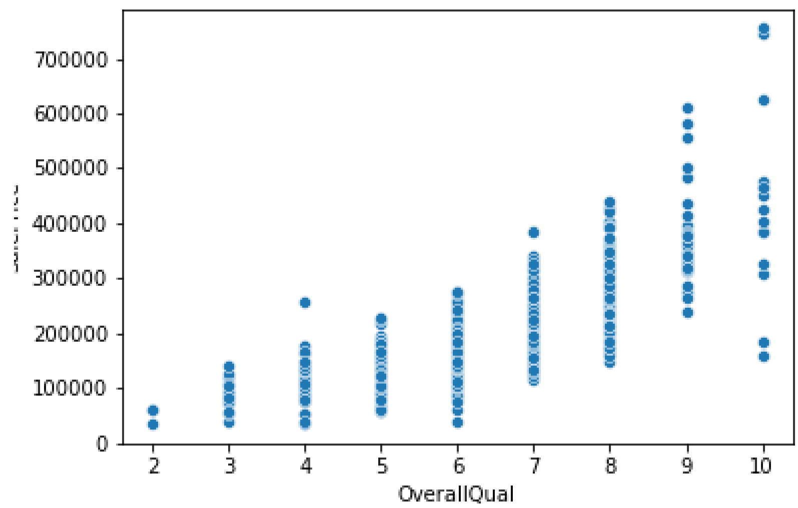
700000
600000
500000
400000
300000
200000
100000
0
0 100000 200000 300000 400000 500000 600000 700000
SalePrice

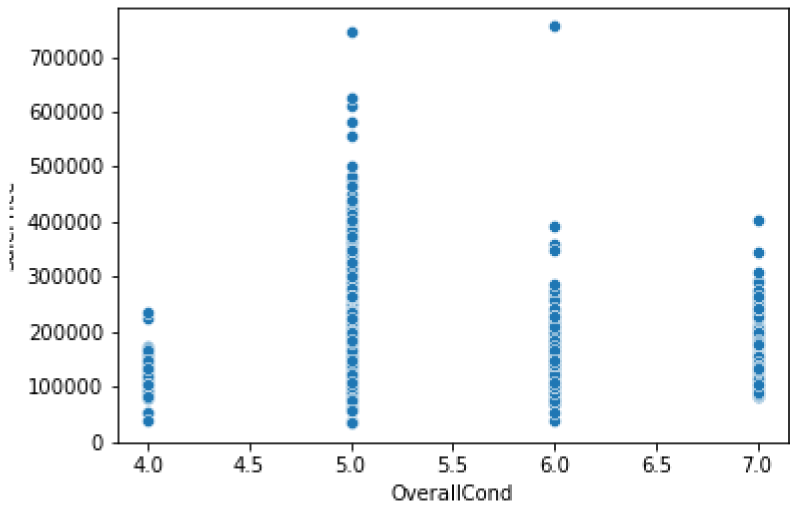


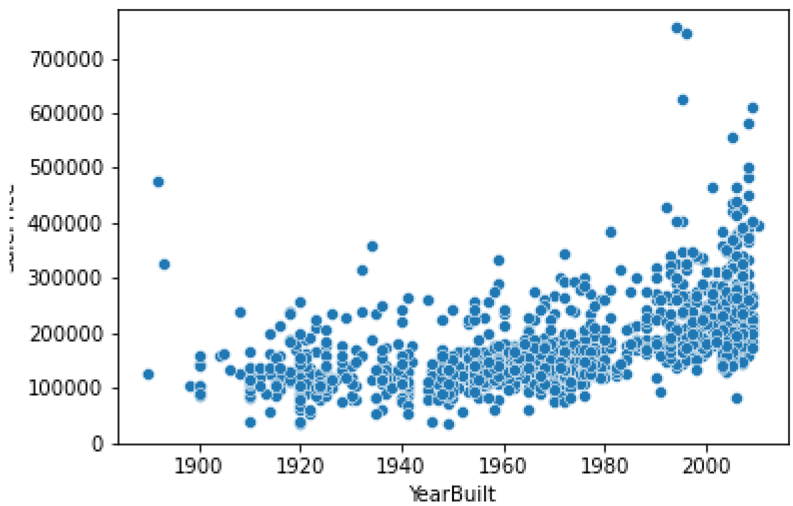
700000
600000
500000
400000
300000
200000
100000
0
20 40 60 80 100 120
MSSubClass

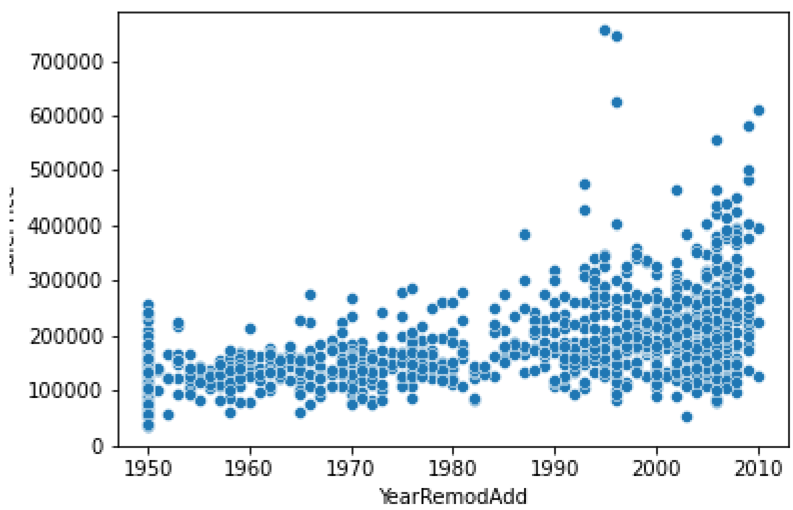

700000
600000
500000
400000
300000
200000
100000
0
30
40
50
60
70
80
90
100
110
LotFrontage

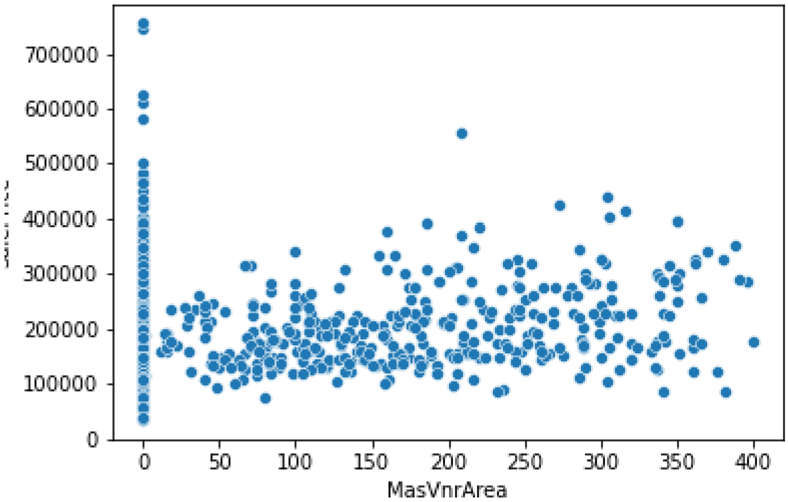

700000
600000
500000
400000
300000
200000
100000
0
2000
4000
6000
8000
10000
12000
14000
16000
LotArea


700000
600000
500000
400000
300000
200000
100000
0
2
3
4
5
6
7
8
9
10
OverallQual

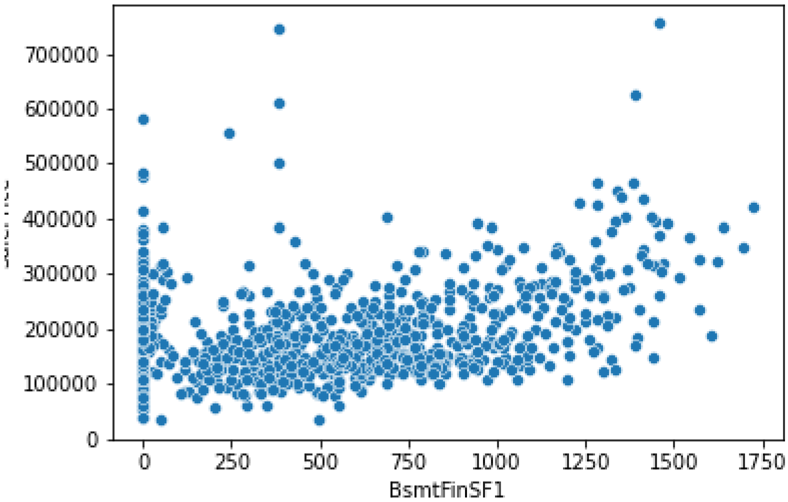

700000
600000
500000
400000
300000
200000
100000
0
4.0
4.5
5.0
5.5
6.0
6.5
7.0
OverallCond


700000
600000
500000
400000
300000
200000
100000
0
1900
1920
1940
1960
1980
2000
YearBuilt

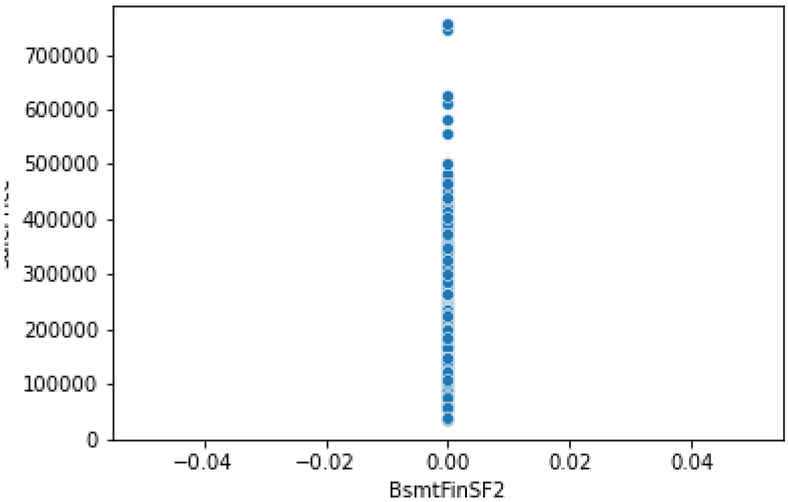

700000
600000
500000
400000
300000
200000
100000
0
1950
1960
1970
1980
1990
2000
2010
YearRemodAdd



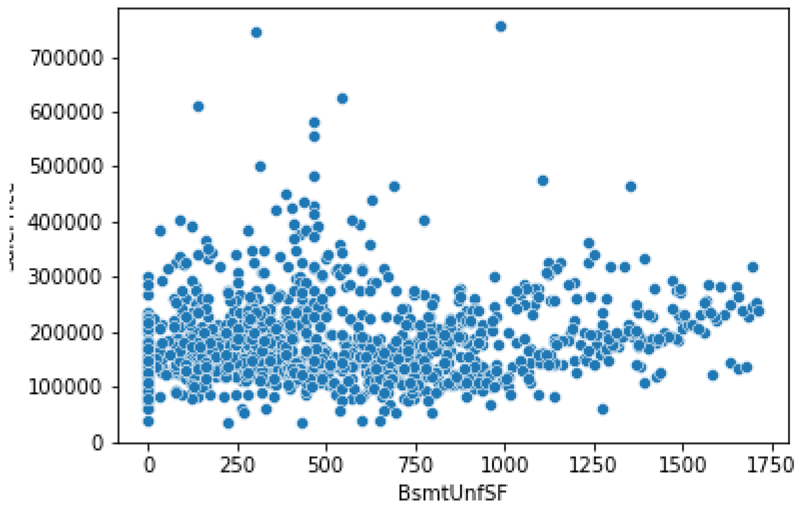
SalePrice
0
100000
200000
300000
400000
500000
600000
700000
0
50
100
150
200
250
300
350
400
MasVnrArea

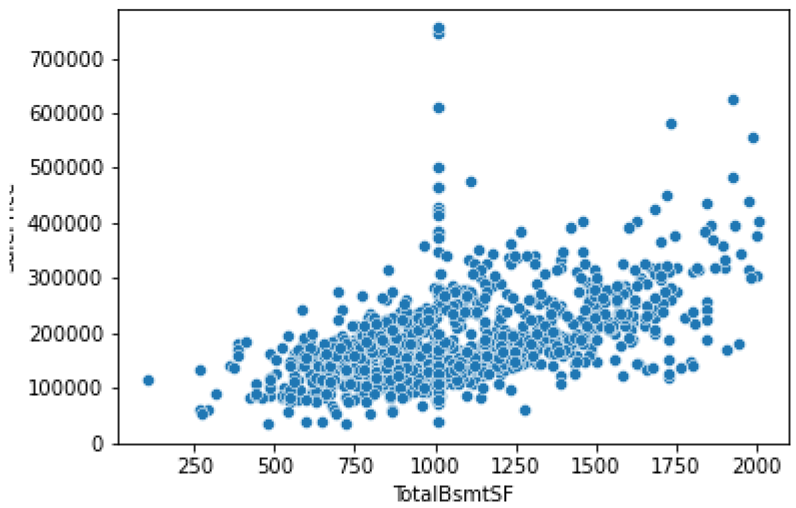


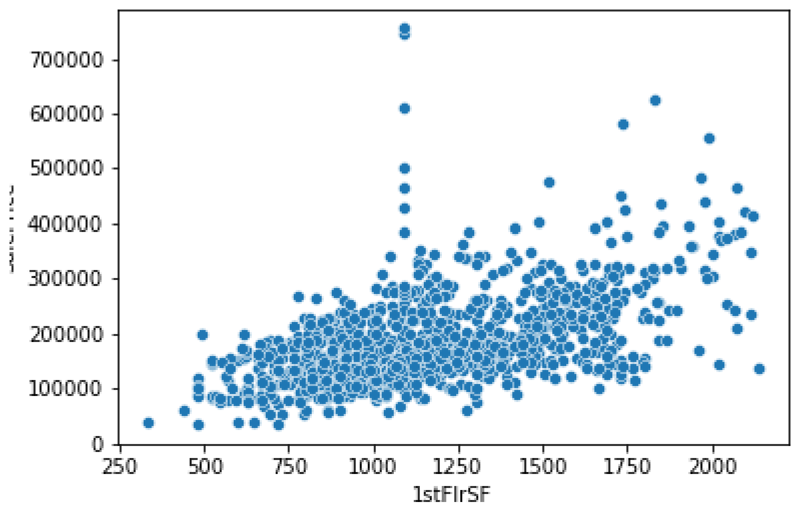
SalePrice
0
100000
200000
300000
400000
500000
600000
700000
0
250
500
750
1000
1250
1500
1750
BsmtFinSF1

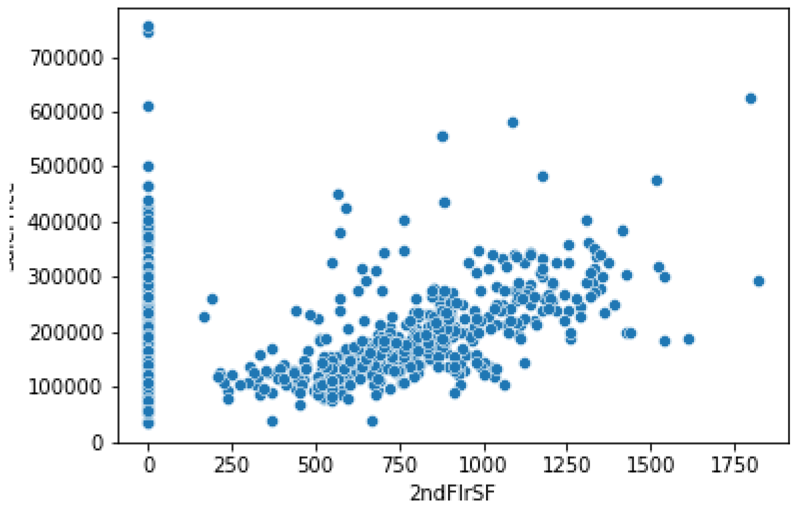


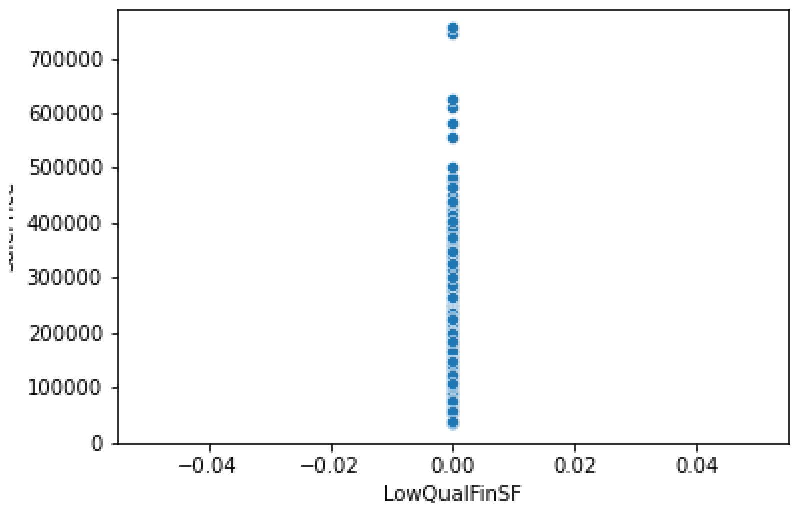
SalePrice
0
100000
200000
300000
400000
500000
600000
700000
−0.04
−0.02
0.00
0.02
0.04
BsmtFinSF2

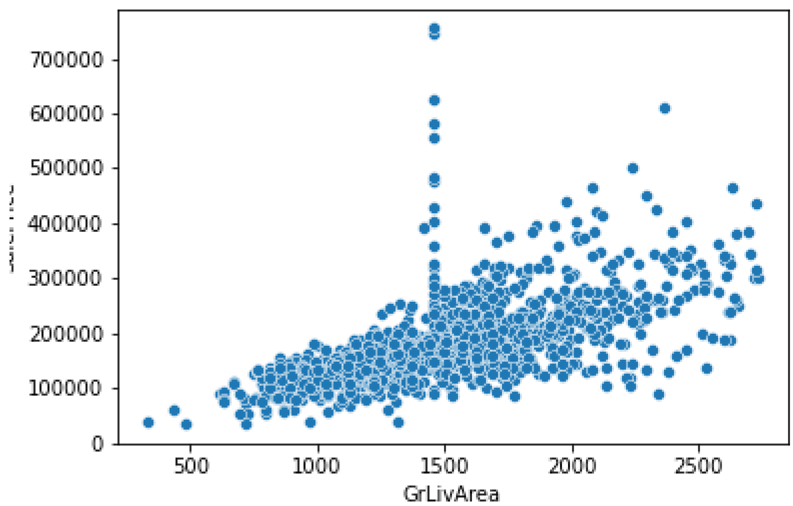

BsmtUnfSF

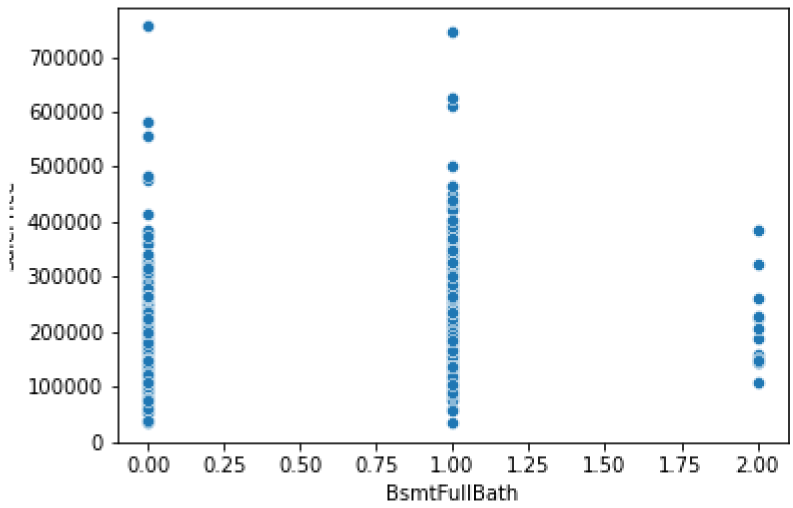

TotalBsmtSF


1stFlrSF

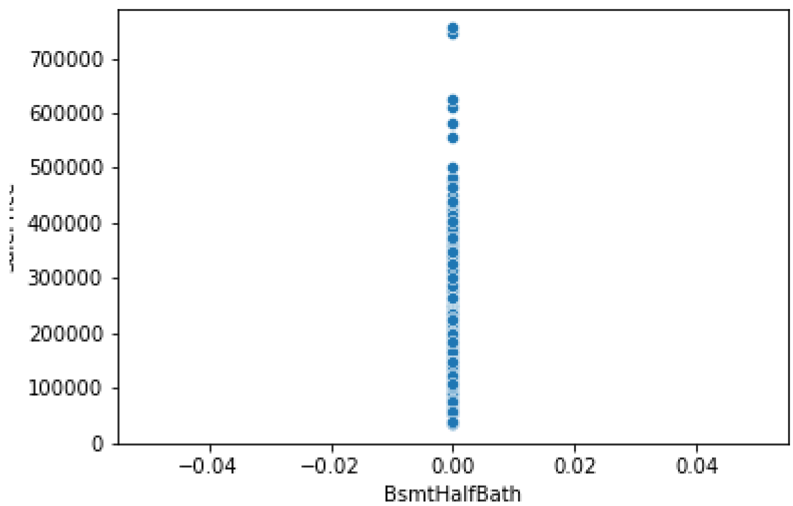

700000
600000
500000
400000
300000
200000
100000
0
0
250
500
750
1000
1250
1500
1750
2ndFlrSF


700000
600000
500000
400000
300000
200000
100000
0
−0.04
−0.02
0.00
0.02
0.04
LowQualFinSF

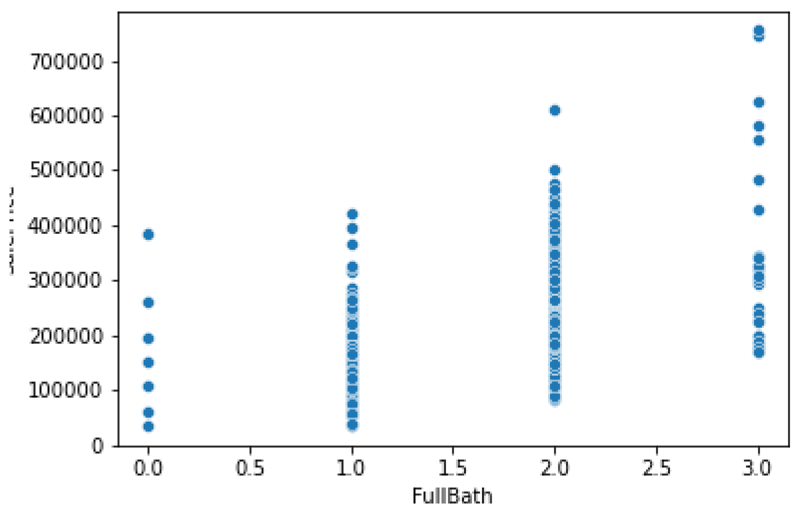

700000
600000
500000
400000
300000
200000
100000
0
500
1000
1500
2000
2500
GrLivArea



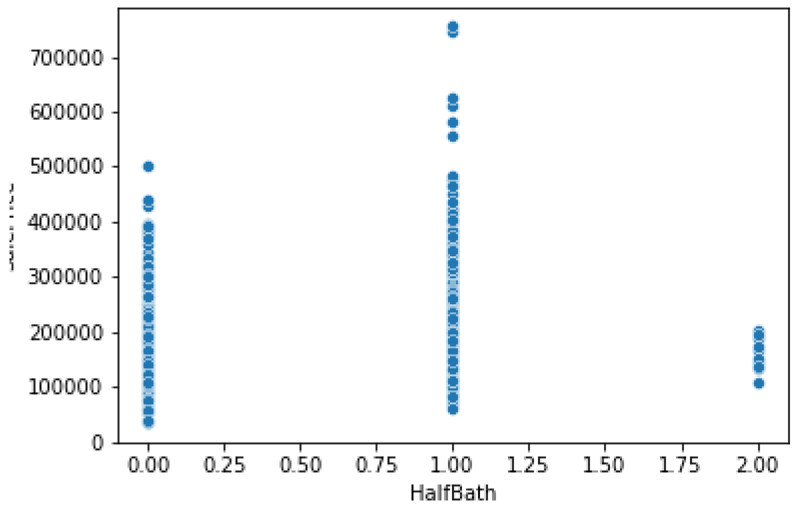
SalePrice
0
100000
200000
300000
400000
500000
600000
700000
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
BsmtFullBath

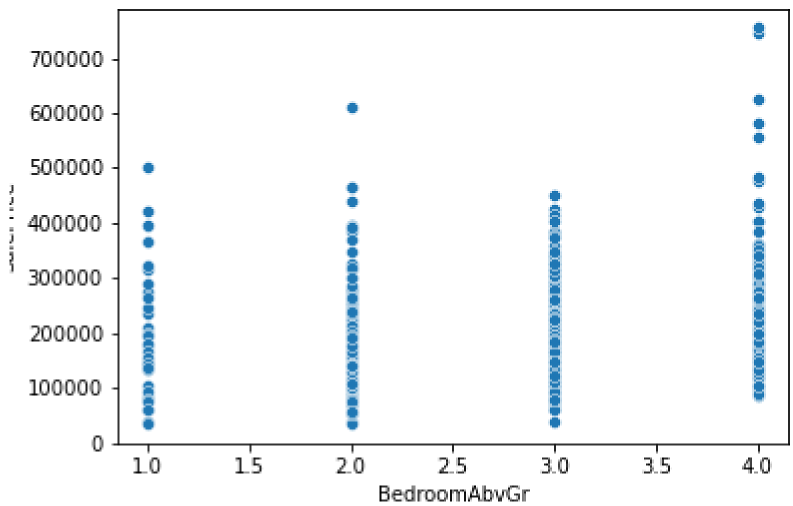


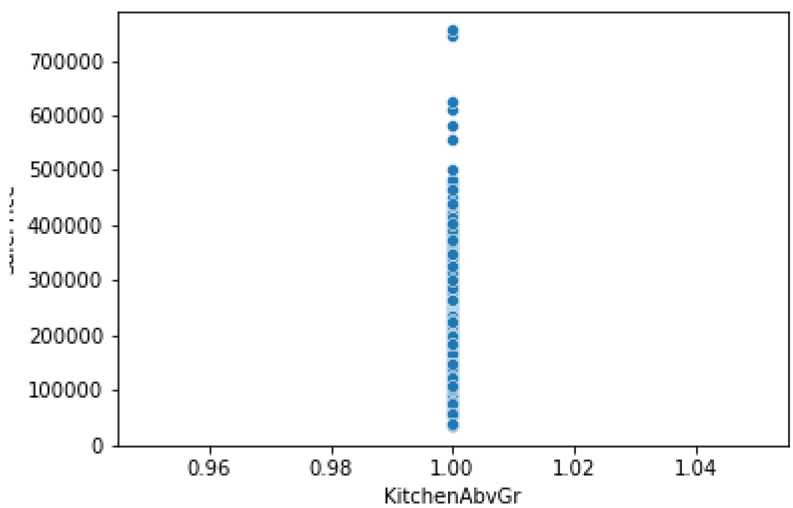
SalePrice
0
100000
200000
300000
400000
500000
600000
700000
−0.04
−0.02
0.00
0.02
0.04
BsmtHalfBath

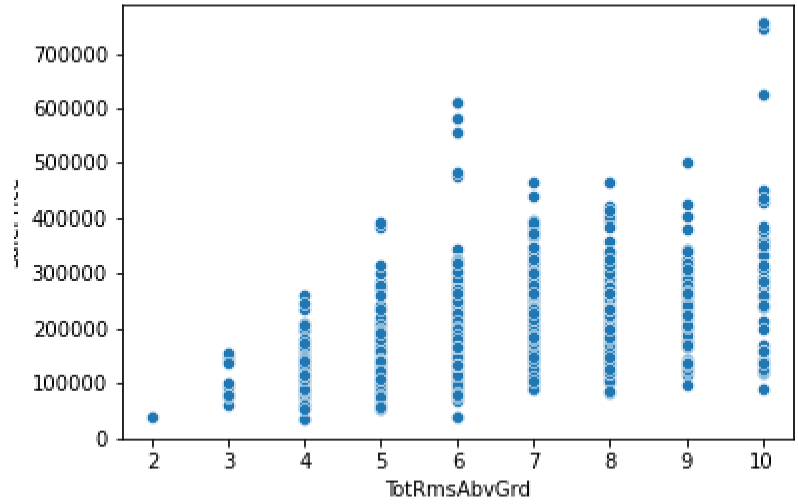


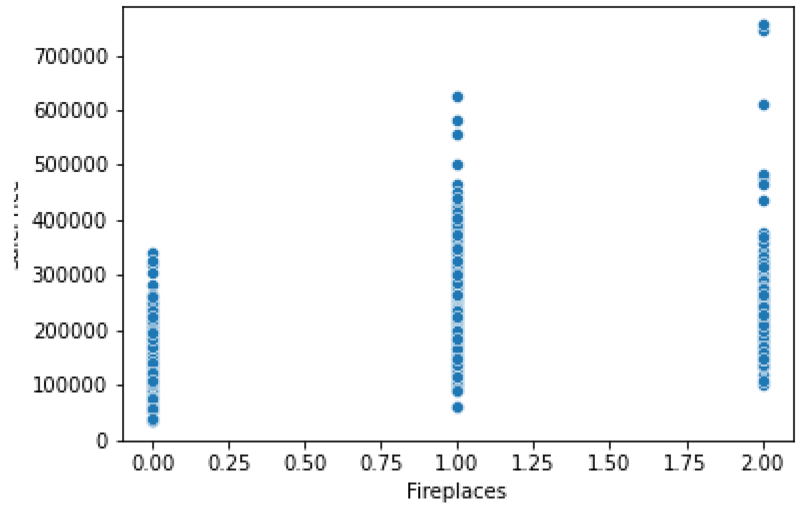
SalePrice
0
100000
200000
300000
400000
500000
600000
700000
0.0
0.5
1.0
1.5
2.0
2.5
3.0
FullBath

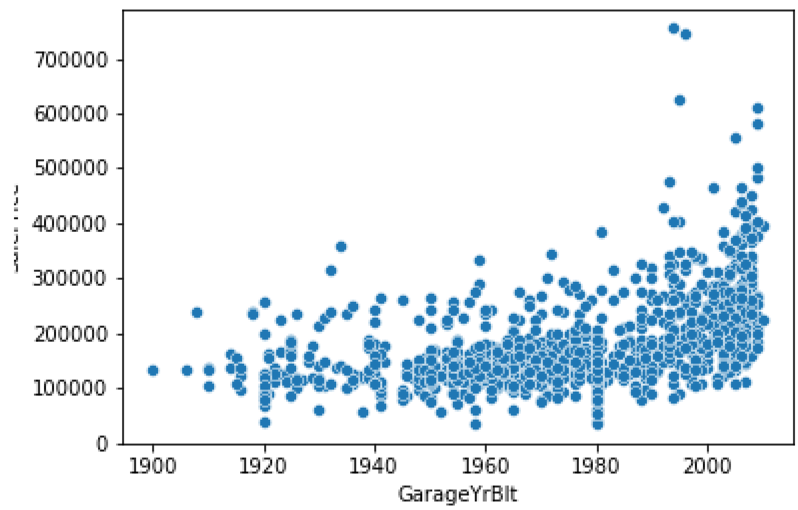

700000
600000
500000
400000
300000
200000
100000
0
0.00 0.25 0.50 0.75 1.00 1.25 1.50 1.75 2.00
HalfBath

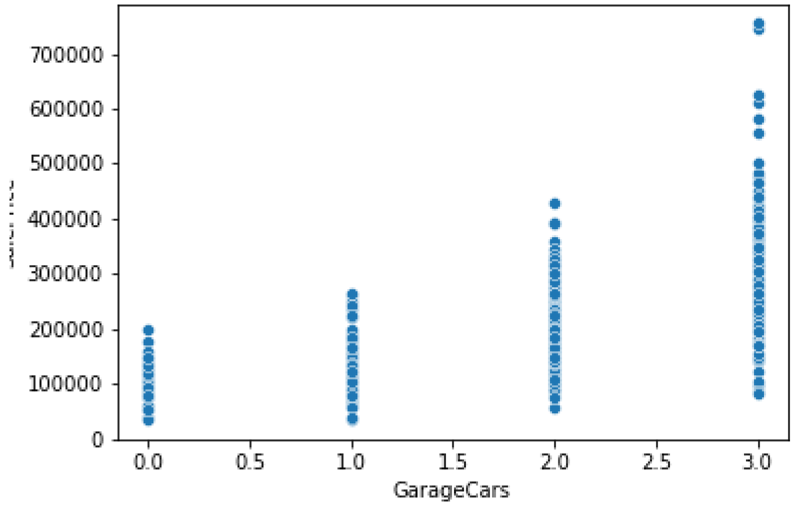

700000
600000
500000
400000
300000
200000
100000
0
1.0 1.5 2.0 2.5 3.0 3.5 4.0
BedroomAbvGr

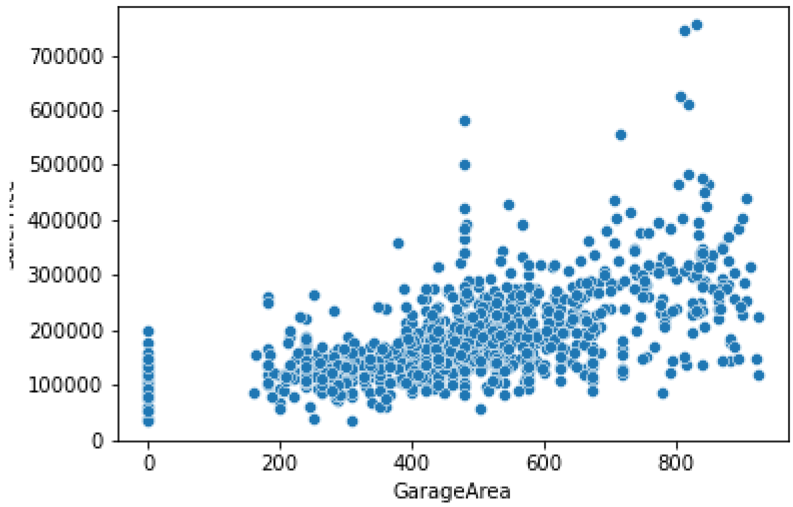

700000
600000
500000
400000
300000
200000
100000
0
0.96 0.98 1.00 1.02 1.04
KitchenAbvGr

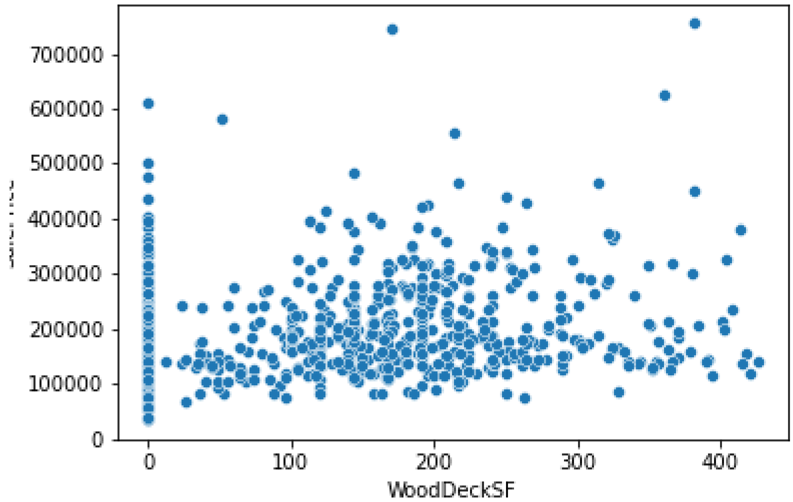

SalePrice
TotRmsAbvGrd


SalePrice
Fireplaces

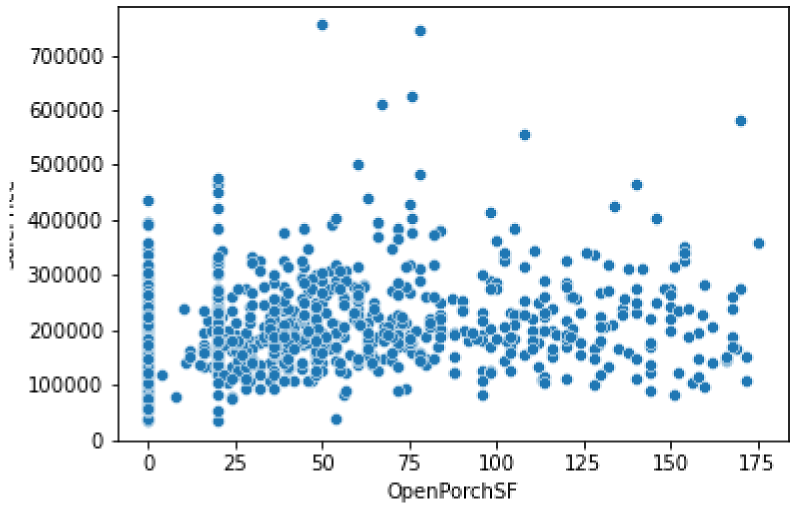

SalePrice
GarageYrBlt


700000
600000
500000
400000
300000
200000
100000
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
GarageCars

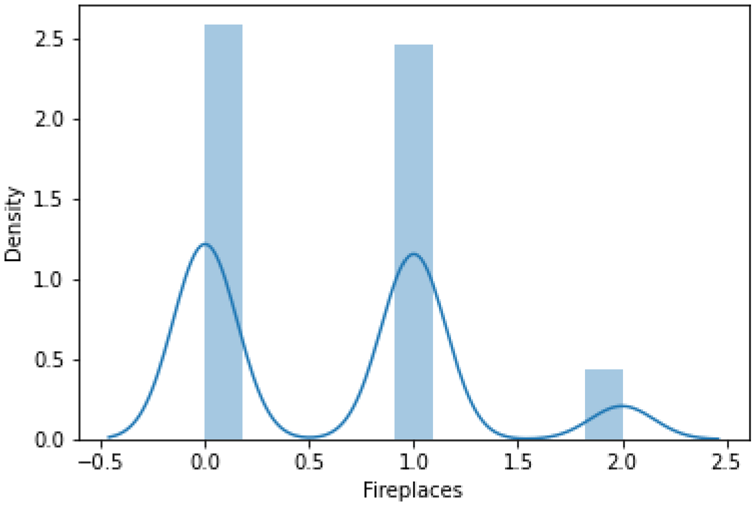

700000
600000
500000
400000
300000
200000
100000
0
0
200
400
600
800
GarageArea

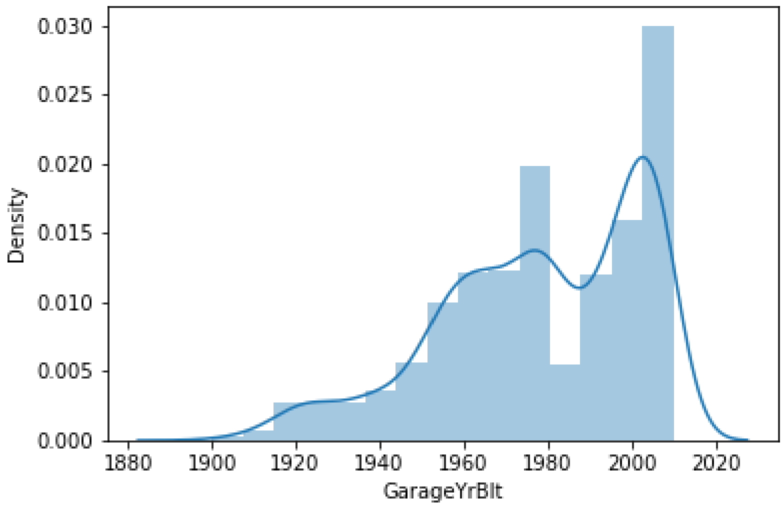


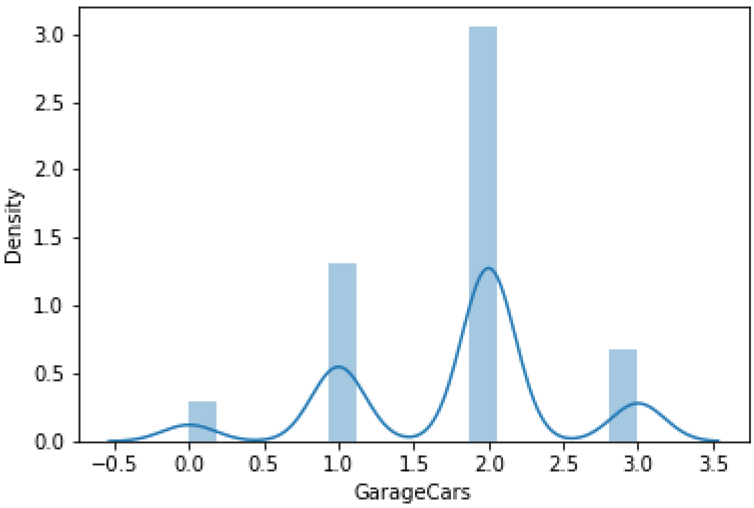
0
100000
200000
300000
400000
500000
600000
700000
0
100
200
300
400
WoodDeckSF

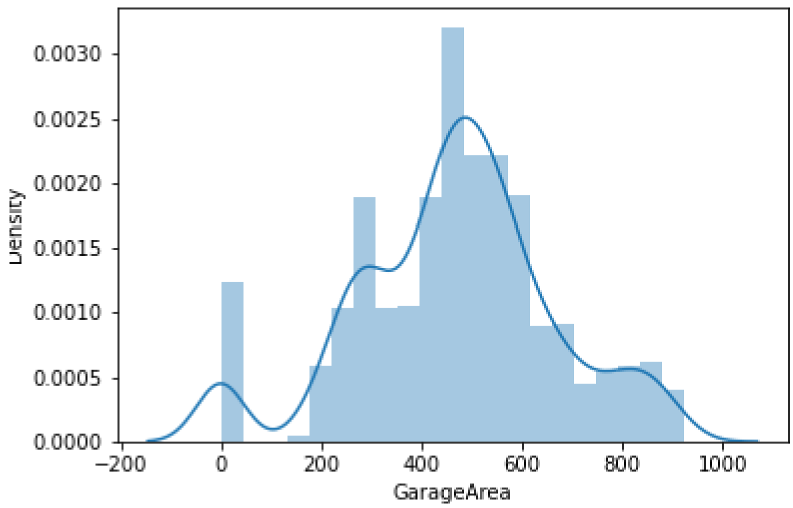


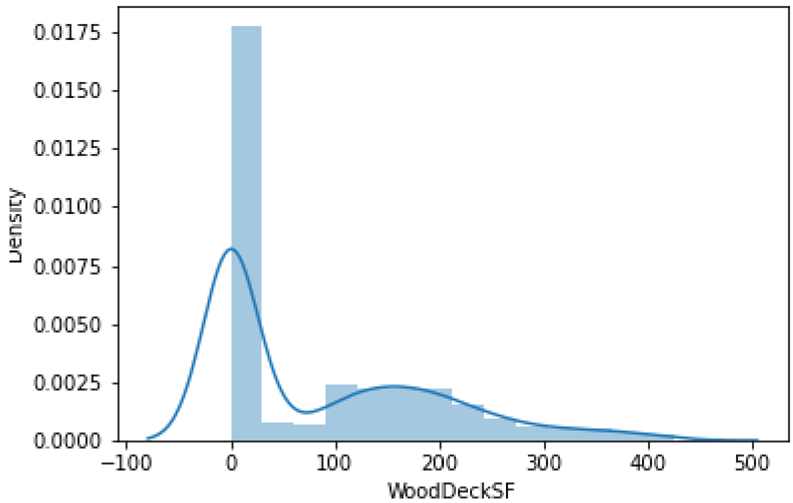
0
100000
200000
300000
400000
500000
600000
700000
0
25
50
75
100
125
150
175
OpenPorchSF

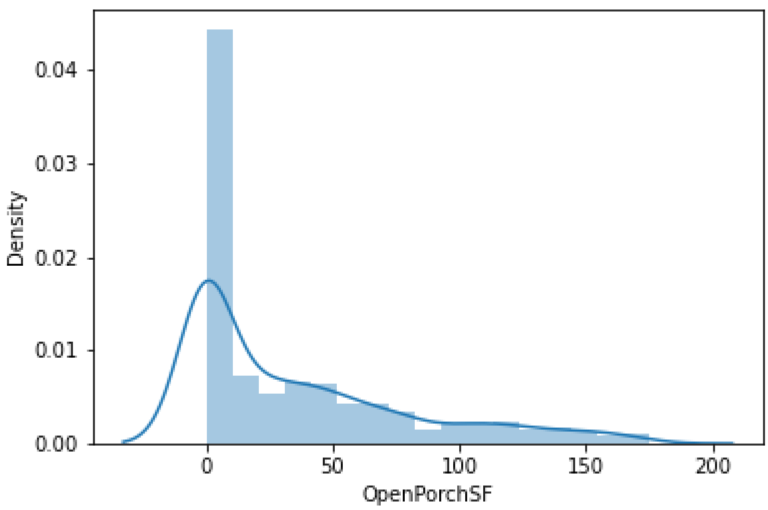

Density
Fireplaces
0.0
0.5
1.0
1.5
2.0
2.5
−0.5

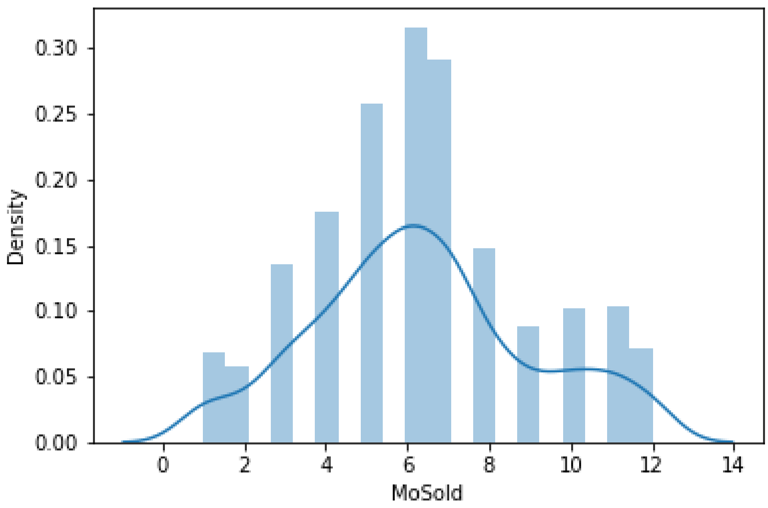

Density
GarageYrBlt
0.000
0.005
0.010
0.015
0.020
0.025
0.030
1880
1900
1920
1940
1960
1980
2000
2020

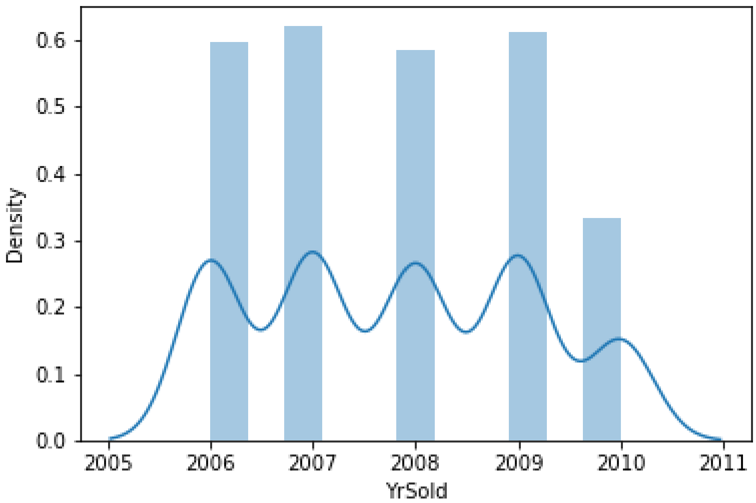

Density
GarageCars
0.0
0.5
1.0
1.5
2.0
2.5
3.0
−0.5
3.5

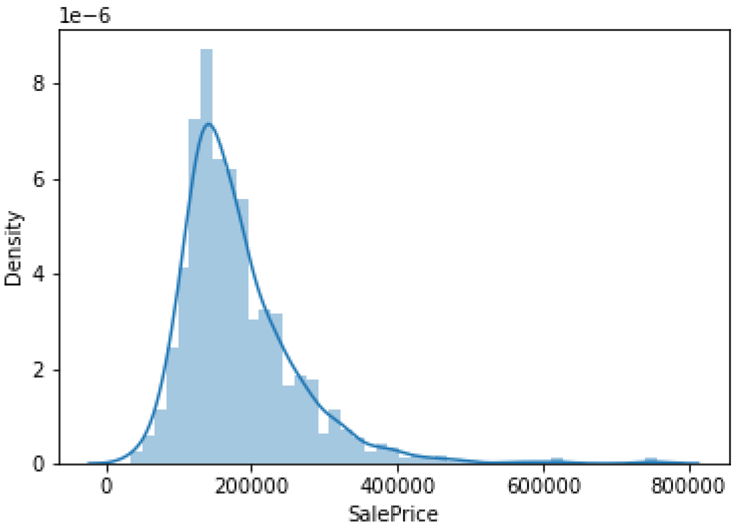

0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
Density
−200
0
200
400
600
800
1000
GarageArea

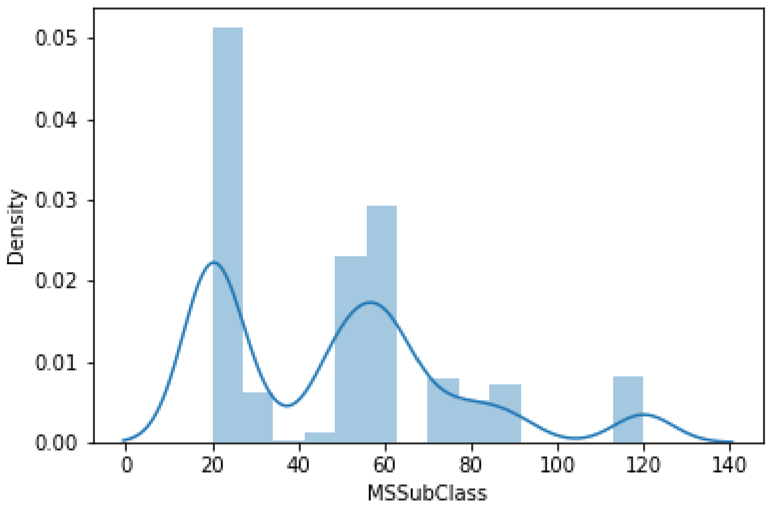

0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
Density
−100
0
100
200
300
400
500
WoodDeckSF


0.04
0.03
0.02
0.01
0.00
Density
0
50
100
150
200
OpenPorchSF

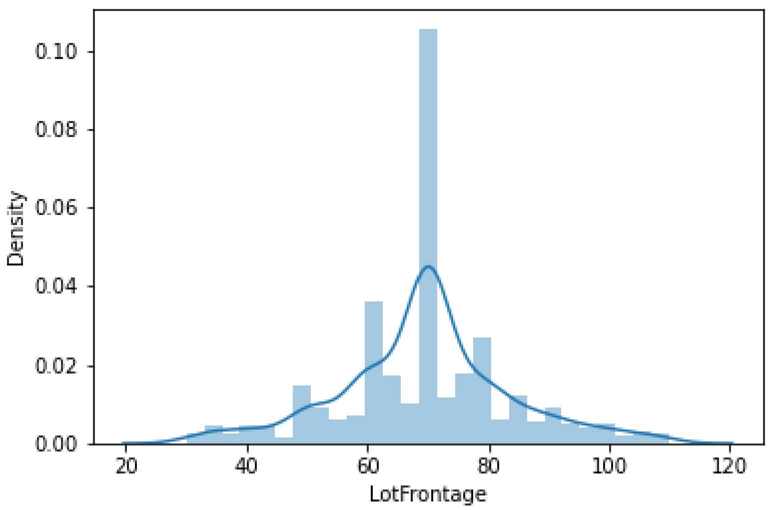

0.30
0.25
0.20
0.15
0.10
0.05
0.00
Density
0
2
4
6
8
10
12
14
MoSold


0.6
0.5
0.4
0.3
0.2
0.1
0.0
Density
2005
2006
2007
2008
2009
2010
2011
YrSold

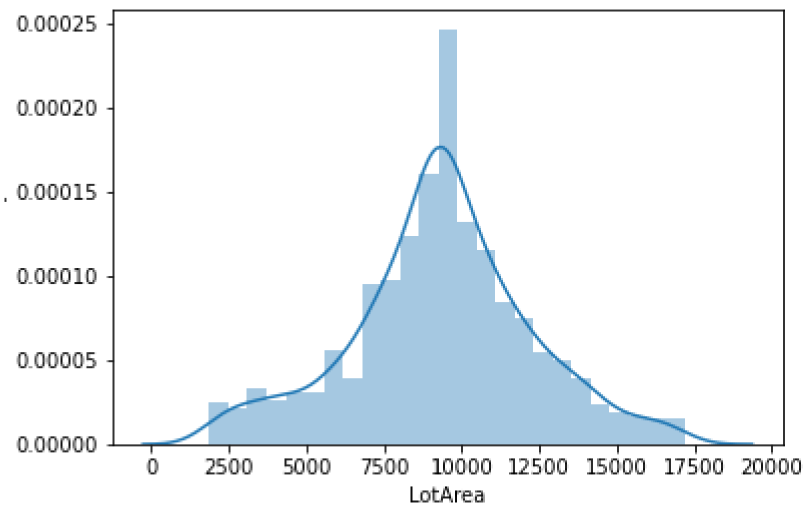

1e−6
8
6
4
2
0
Density
0
200000
400000
600000
800000
SalePrice



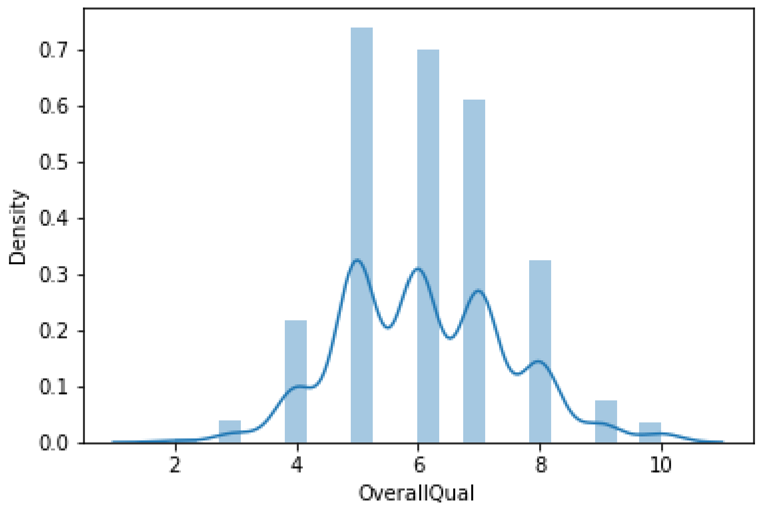
0.05
0.04
0.03
0.02
0.01
0.00
Density
0
20
40
60
80
100
120
140
MSSubClass

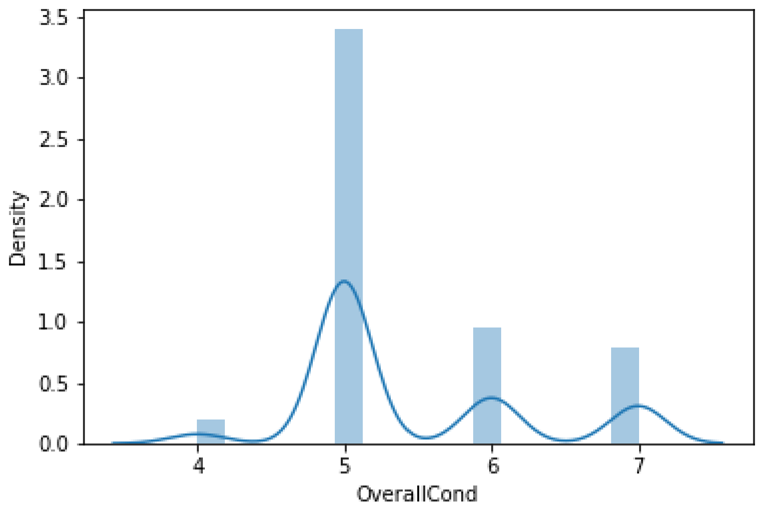


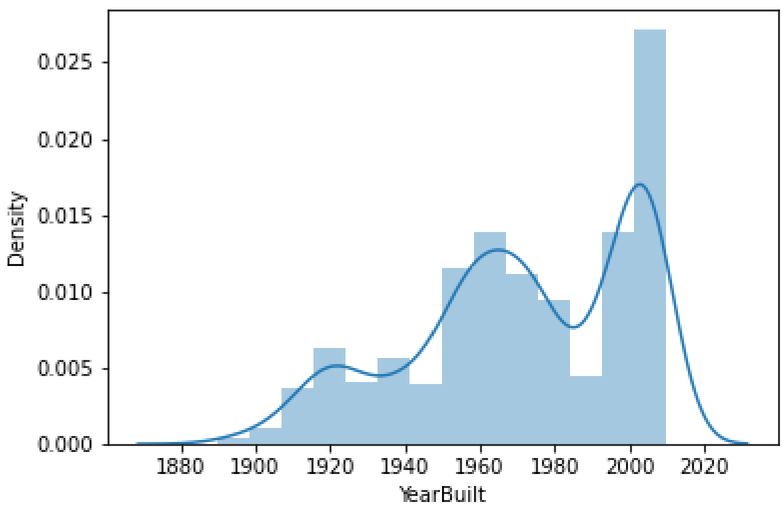
0.10
0.08
0.06
0.04
0.02
0.00
Density
20
40
60
80
100
120
LotFrontage

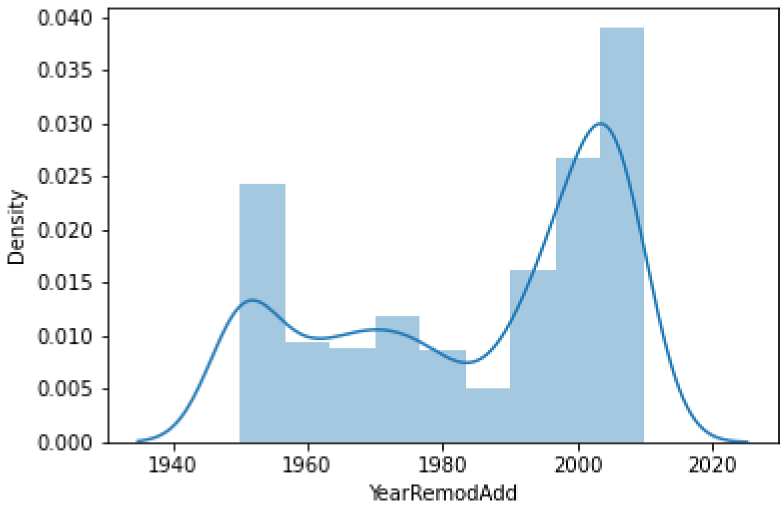


0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
0
2500
5000
7500
10000
12500
15000
17500
20000
LotArea

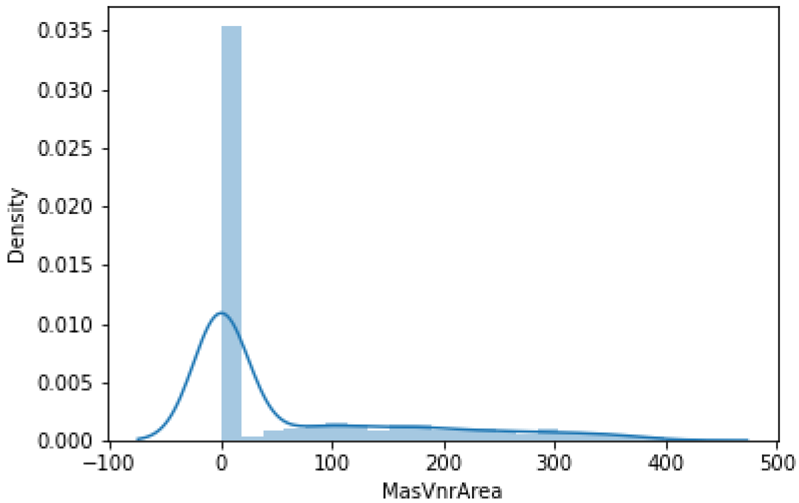

0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Density
2
4
6
8
10
OverallQual


3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Density
4
5
6
7
OverallCond

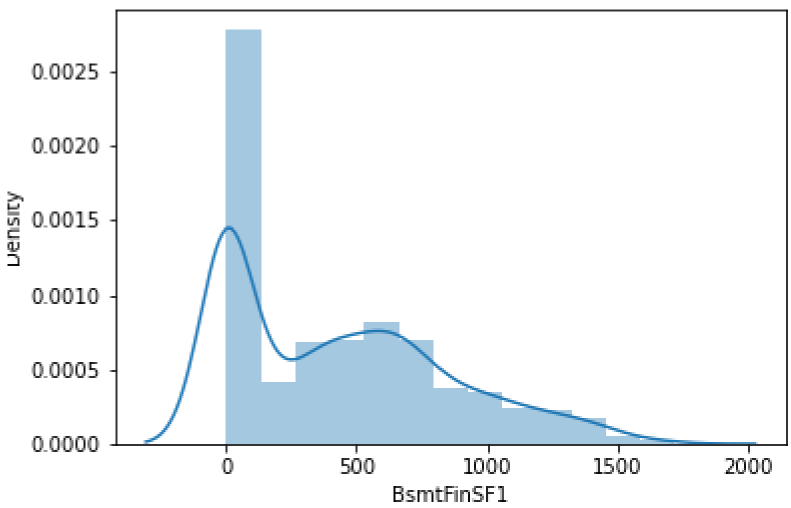

0.025
0.020
0.015
0.010
0.005
0.000
Density
1880
1900
1920
1940
1960
1980
2000
2020
YearBuilt



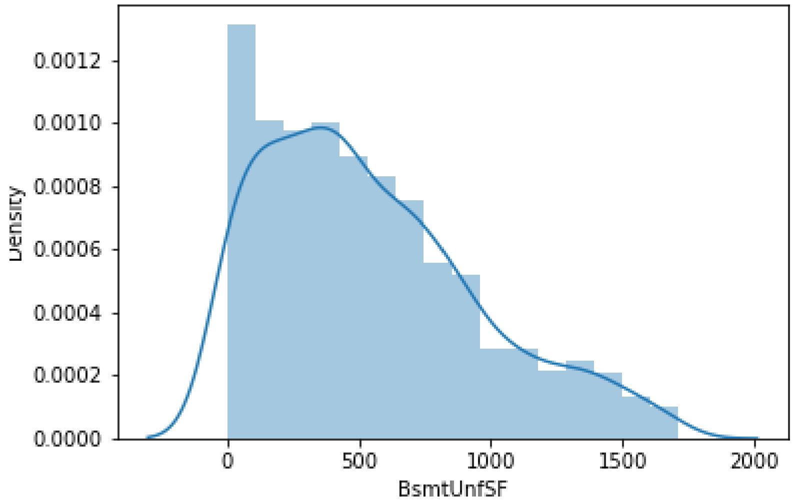
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
Density
1940
1960
1980
2000
2020
YearRemodAdd

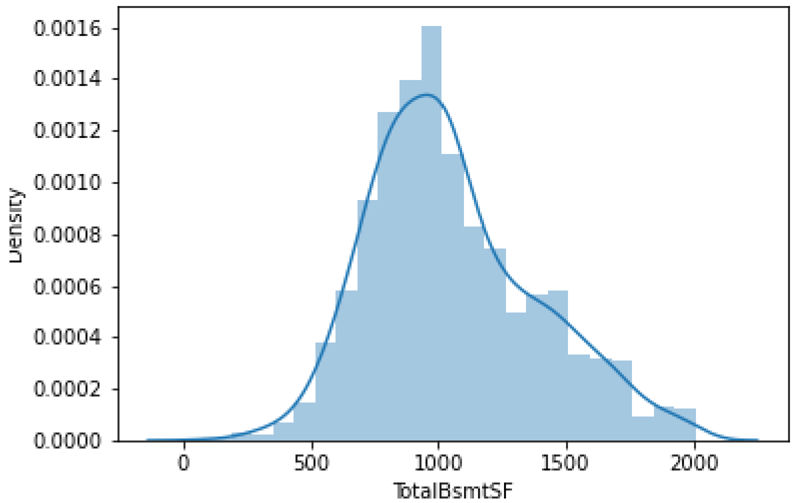


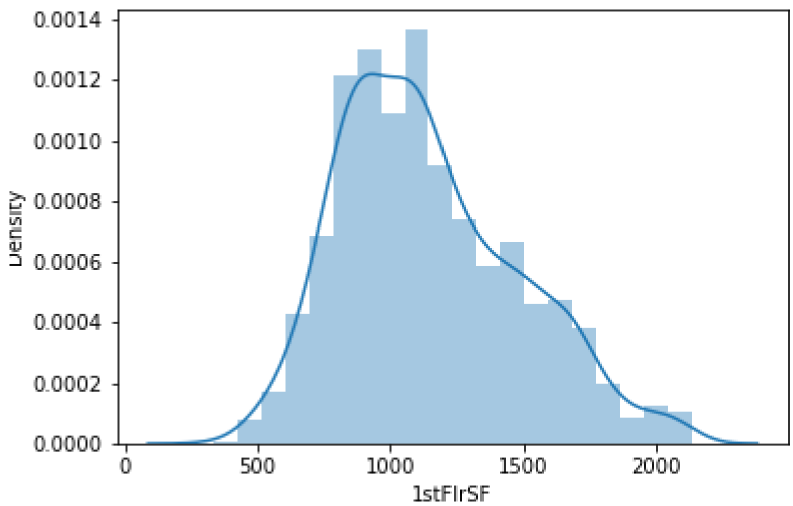
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
Density
−100
0
100
200
300
400
500
MasVnrArea

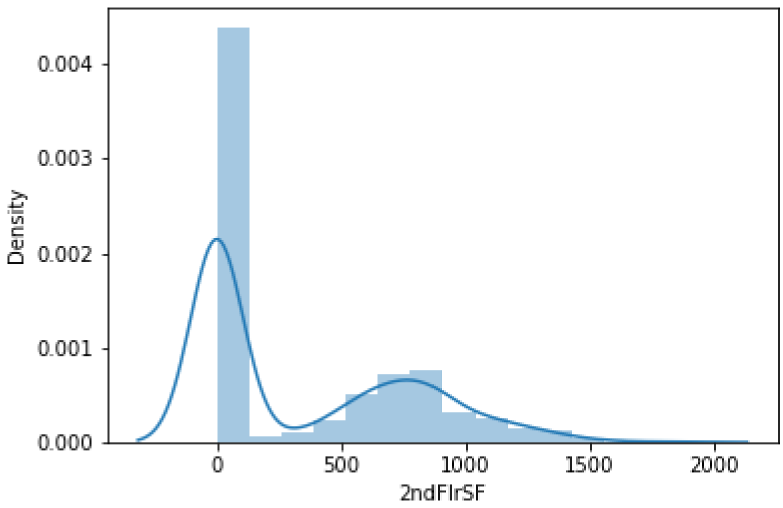


0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
Density
0
500
1000
1500
2000
BsmtFinSF1

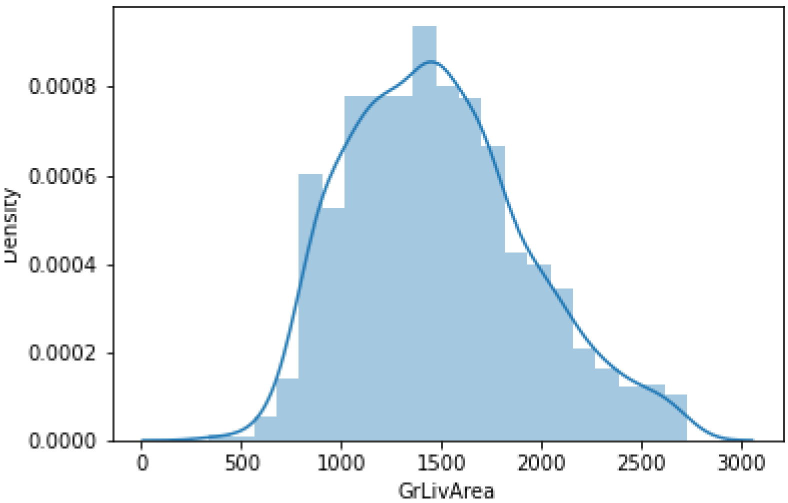

0.0012
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
Density
0
500
1000
1500
2000
BsmtUnfSF


0.0016
0.0014
0.0012
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
Density
0
500
1000
1500
2000
TotalBsmtSF

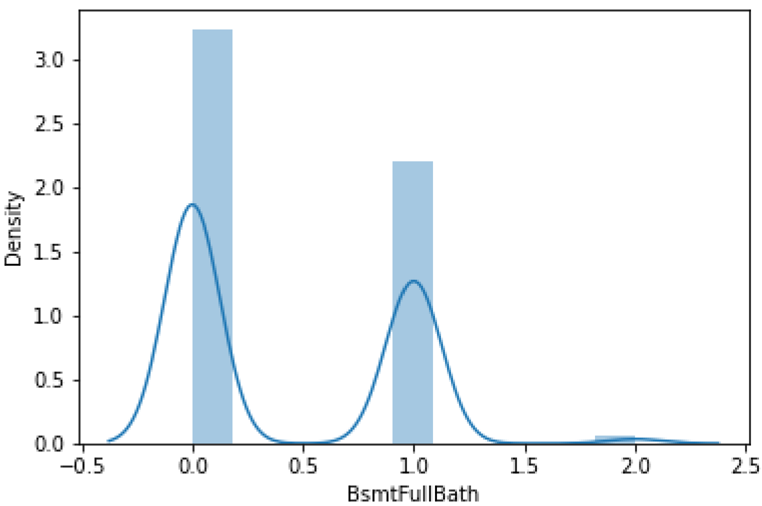

0.0014
0.0012
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
Density
0
500
1000
1500
2000
1stFlrSF



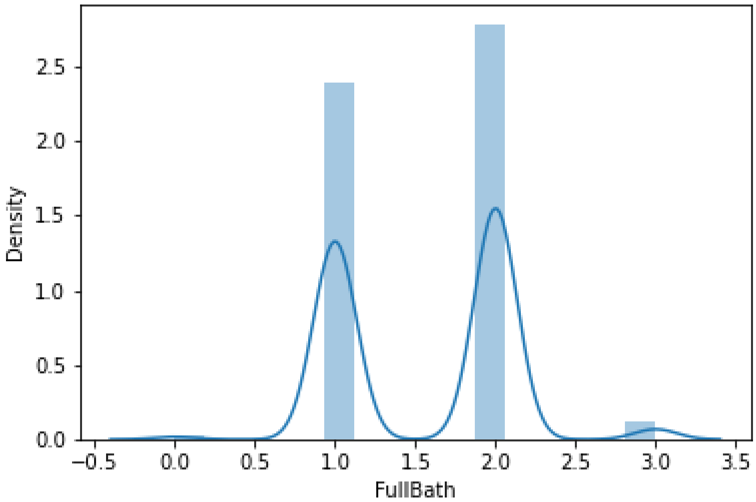
Density
0.000
0.001
0.002
0.003
0.004
0
500
1000
1500
2000
2ndFlrSF

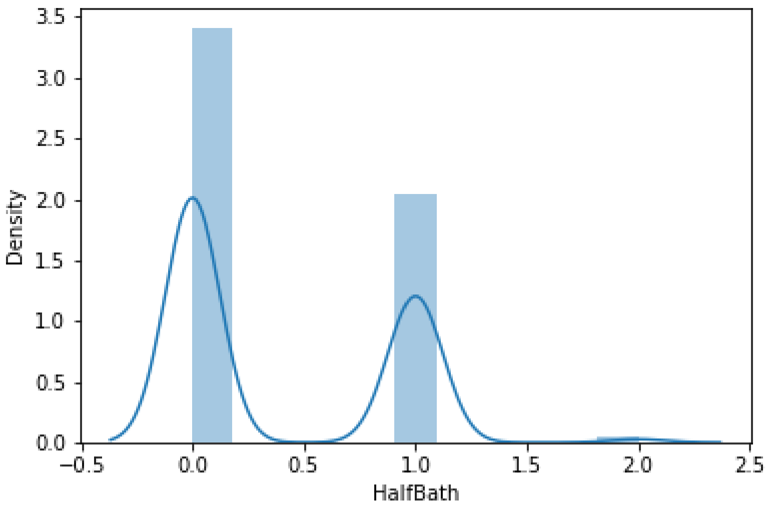


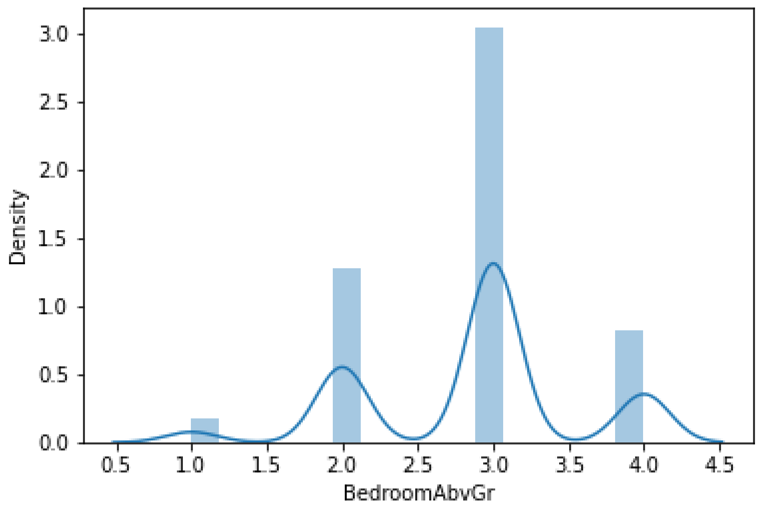
Density
0.0000
0.0002
0.0004
0.0006
0.0008
0
500
1000
1500
2000
2500
3000
GrLivArea



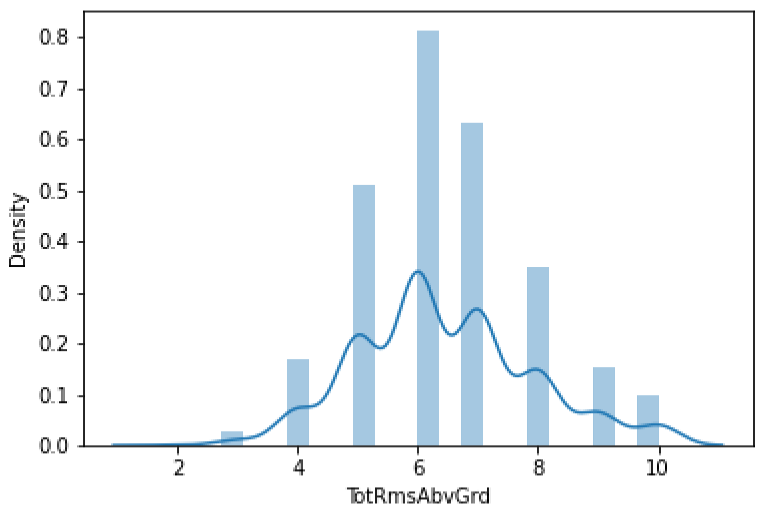
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
BsmtFullBath


Density
FullBath

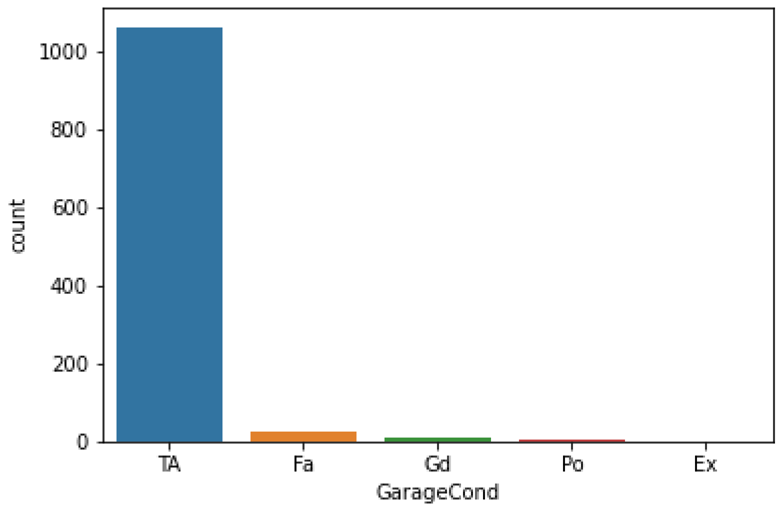

Density
HalfBath

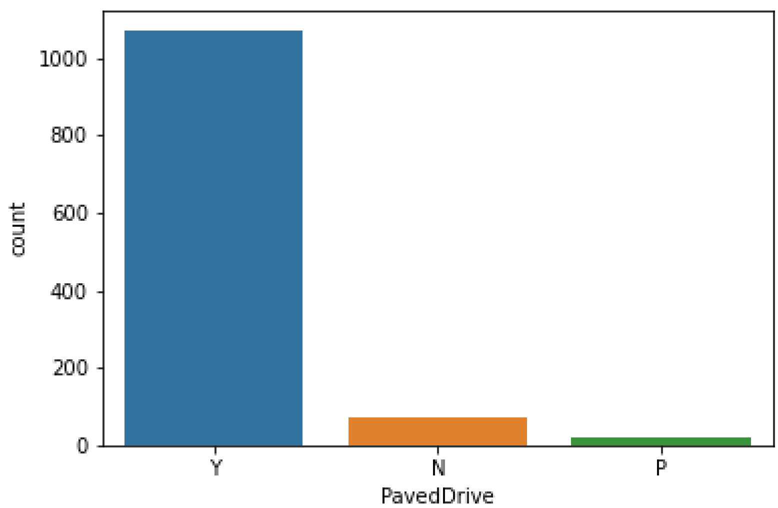


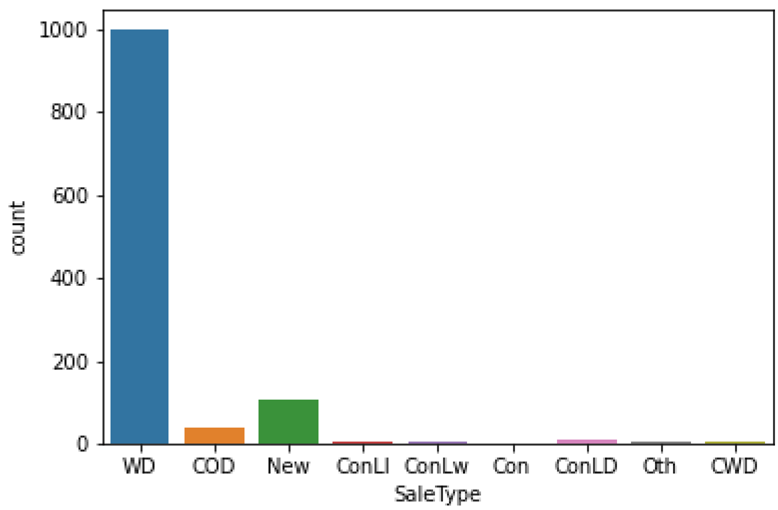
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
BedroomAbvGr

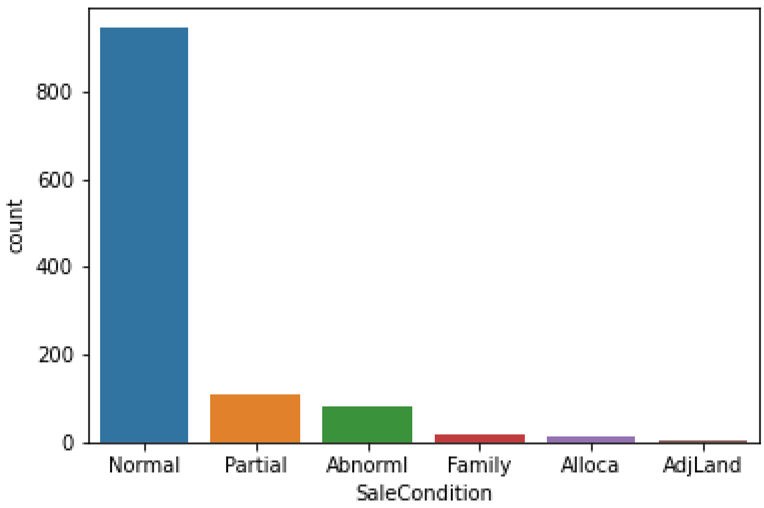


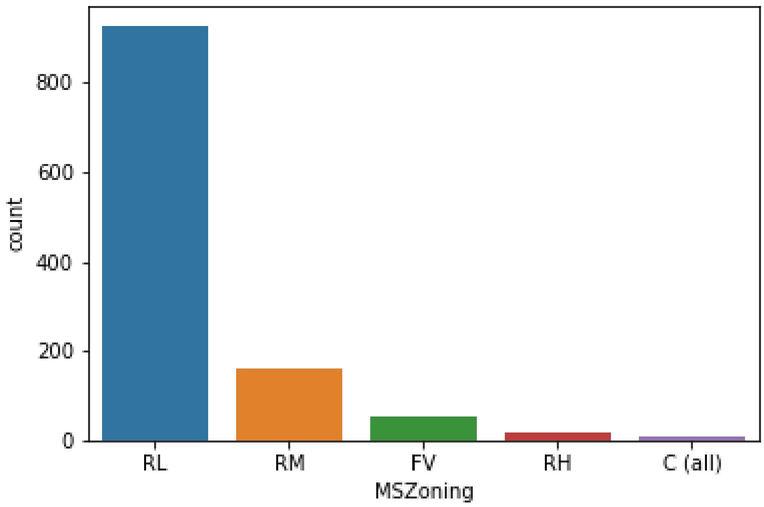
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
2
4
6
8
10
TotRmsAbvGrd

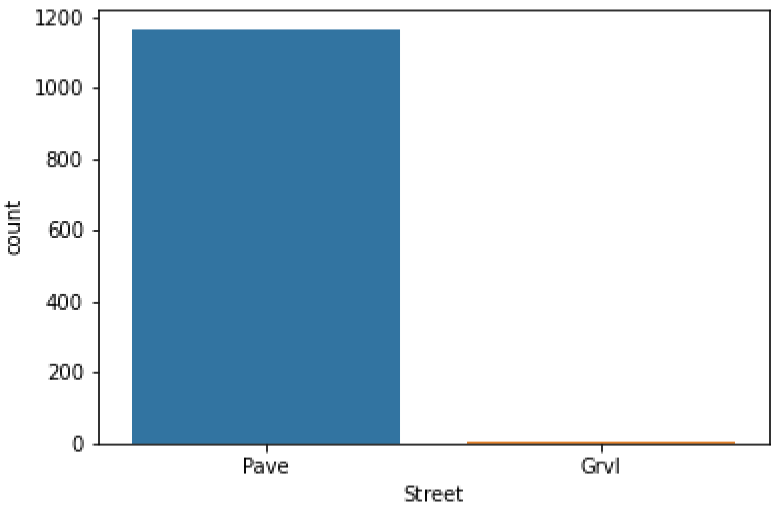

count
1000
800
600
400
200
0
TA
Fa
Gd
Po
Ex
GarageCond

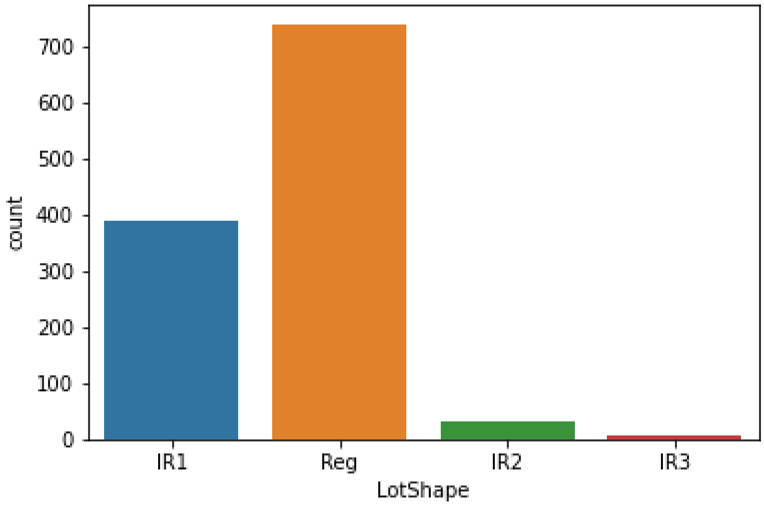

count
1000
800
600
400
200
0
Y
N
P
PavedDrive

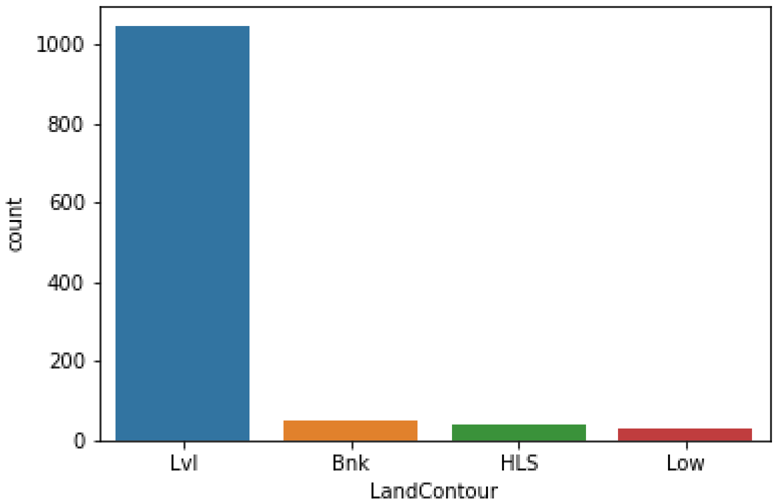

count
1000
800
600
400
200
0
WD
COD
New
ConLI
ConLw
Con
ConLD
Oth
CWD
SaleType

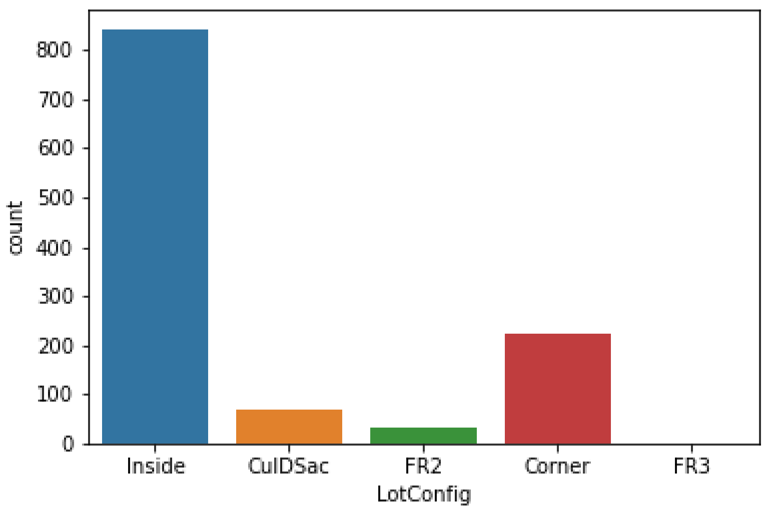

800
600
400
200
0
count
Normal
Partial
Abnorml
Family
Alloca
AdjLand
SaleCondition

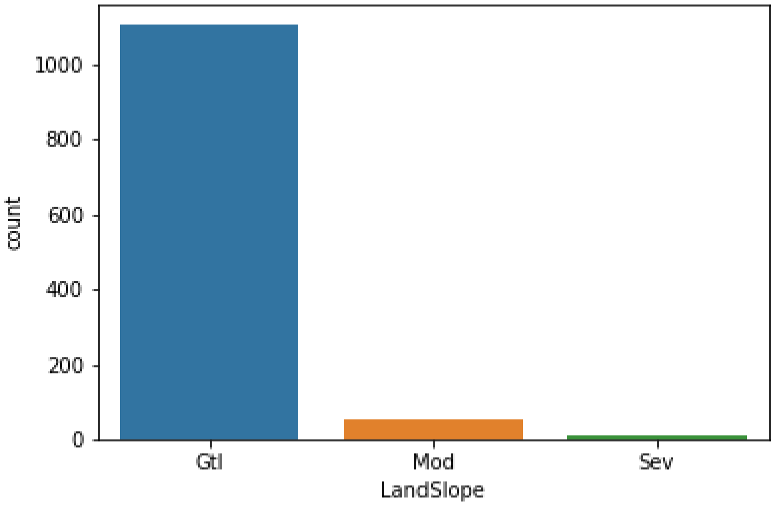

800
600
400
200
0
count
RL
RM
FV
RH
C (all)
MSZoning

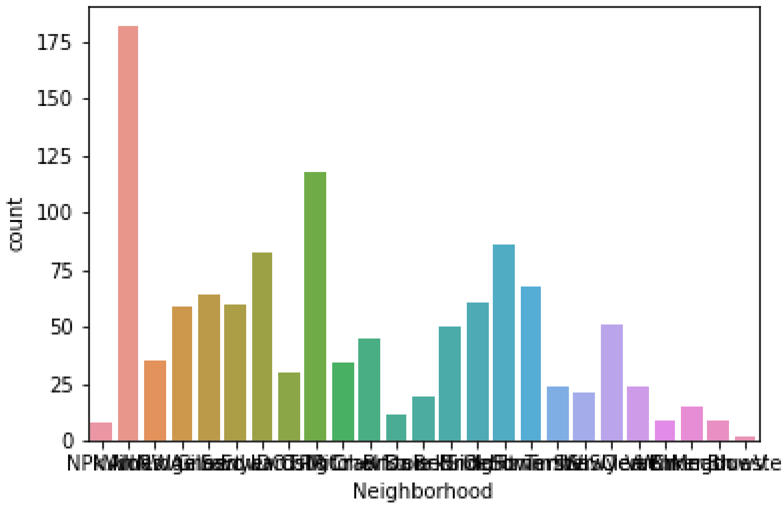

1200
1000
800
600
400
200
0
count
Pave
Grvl
Street

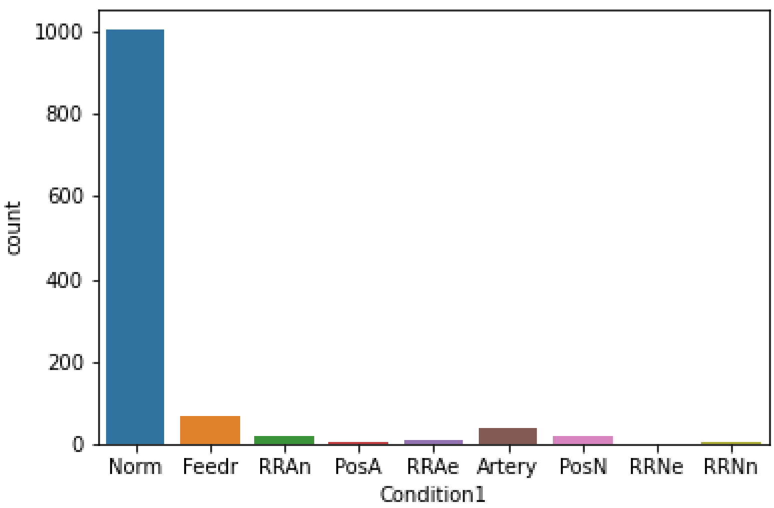

count
700
600
500
400
300
200
100
0
IR1
Reg
IR2
IR3
LotShape

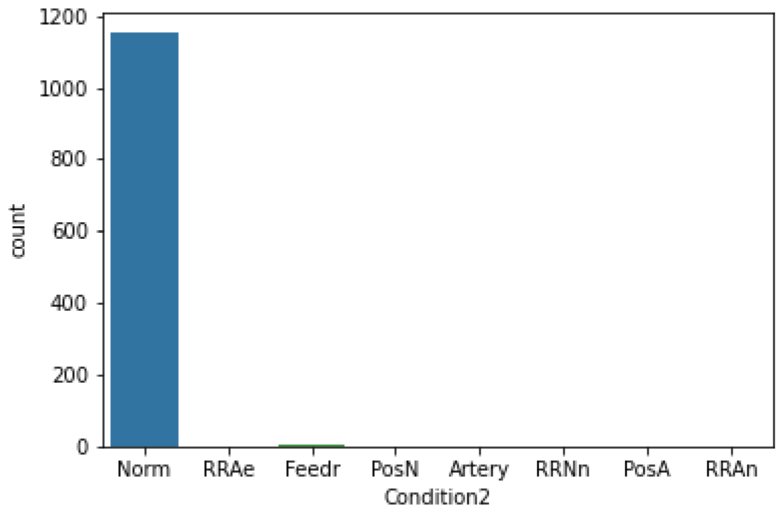

count
1000
800
600
400
200
0
Lvl
Bnk
HLS
Low
LandContour

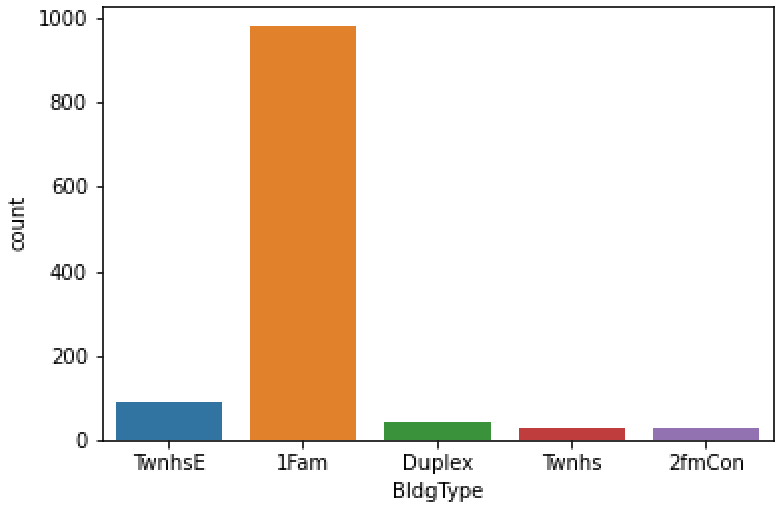

count
800
700
600
500
400
300
200
100
0
Inside
CulDSac
FR2
Corner
FR3
LotConfig

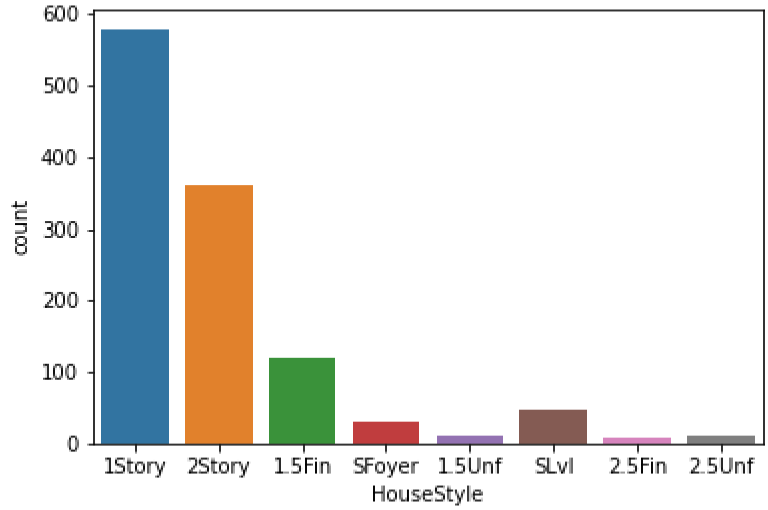

count
0
200
400
600
800
1000
Gtl
Mod
Sev
LandSlope

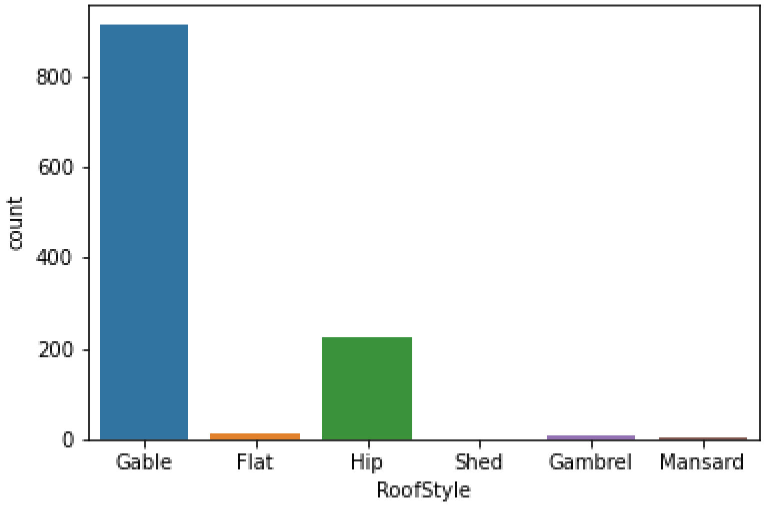

count
0
25
50
75
100
125
150
175
Neighborhood

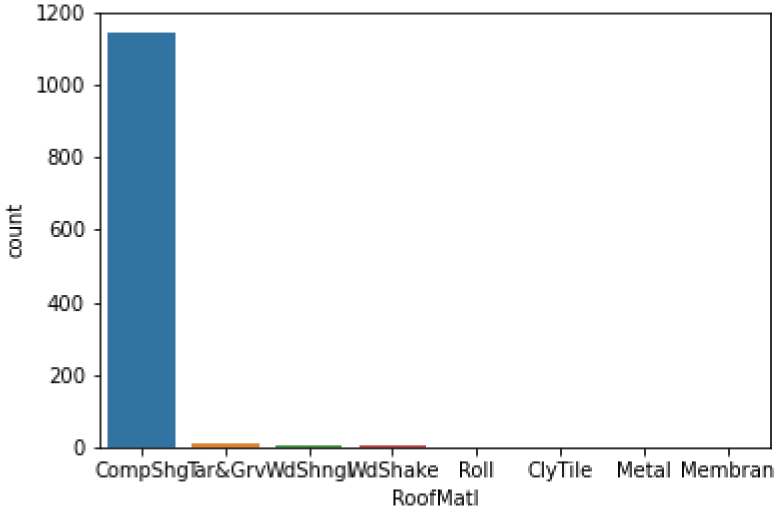

count
0
200
400
600
800
1000
Norm
Feedr
RRAn
PosA
RRAe
Artery
PosN
RRNe
RRNn
Condition1

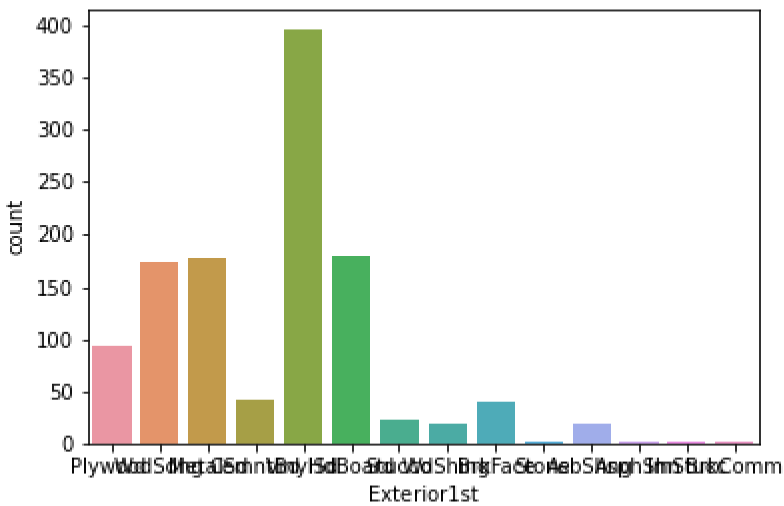

1200
1000
800
600
400
200
0
count
Norm
RRAe
Feedr
PosN
Artery
RRNn
PosA
RRAn
Condition2

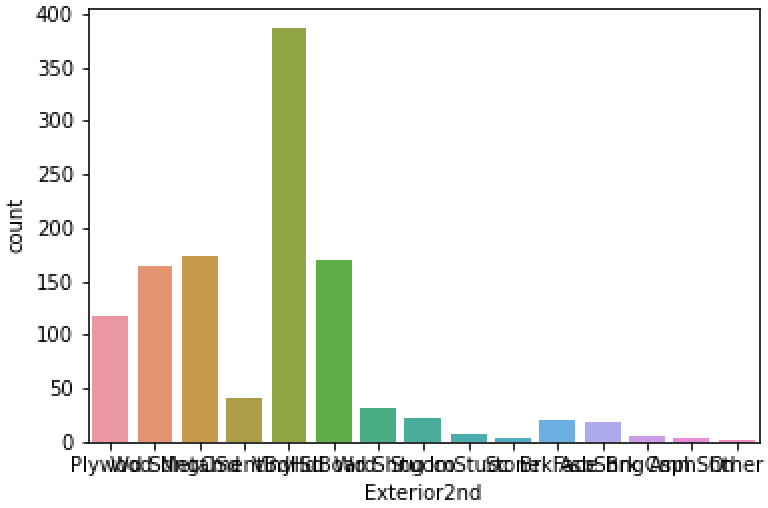

1000
800
600
400
200
0
count
TwnhsE
1Fam
Duplex
Twnhs
2fmCon
BldgType

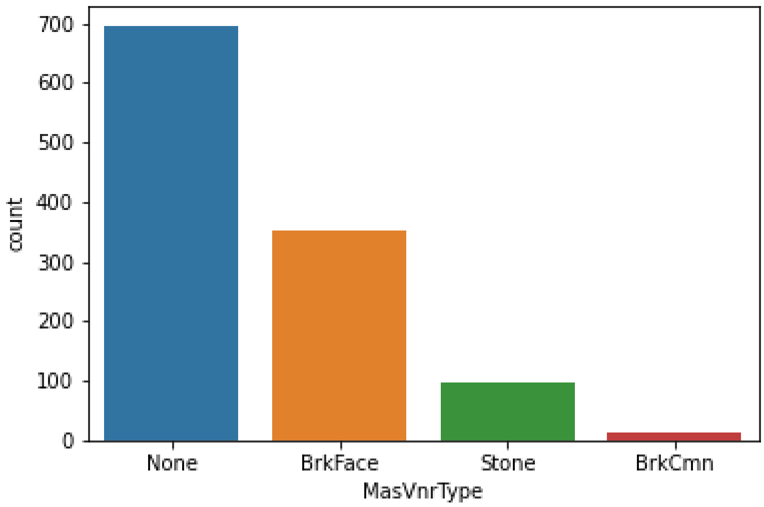

600
500
400
300
200
100
0
count
1Story
2Story
1.5Fin
SFoyer
1.5Unf
SLvl
2.5Fin
2.5Unf
HouseStyle

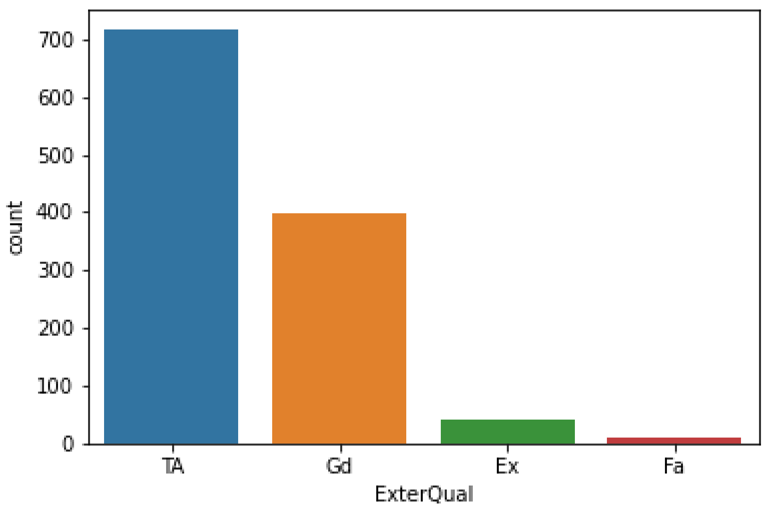

count
800
600
400
200
0
Gable
Flat
Hip
Shed
Gambrel
Mansard
RoofStyle

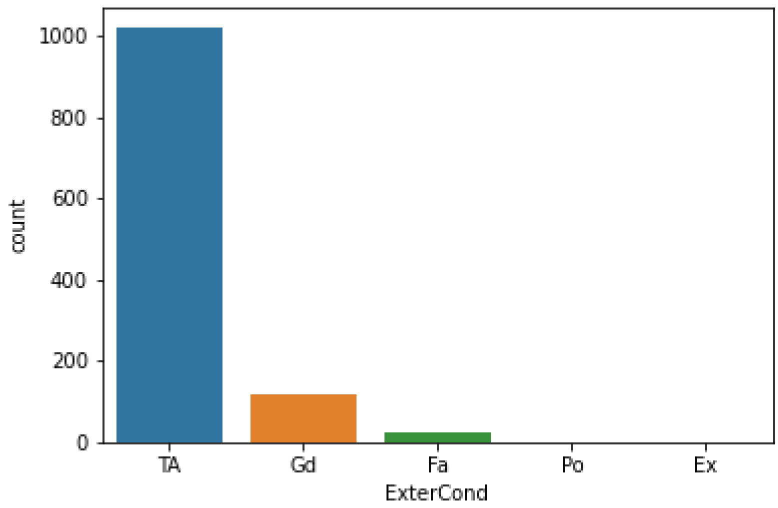

count
1200
1000
800
600
400
200
0
CompShg
Tar&Grv
WdShngl
WdShake
Roll
ClyTile
Metal
Membran
RoofMatl


count
400
350
300
250
200
150
100
50
0
Exterior1st

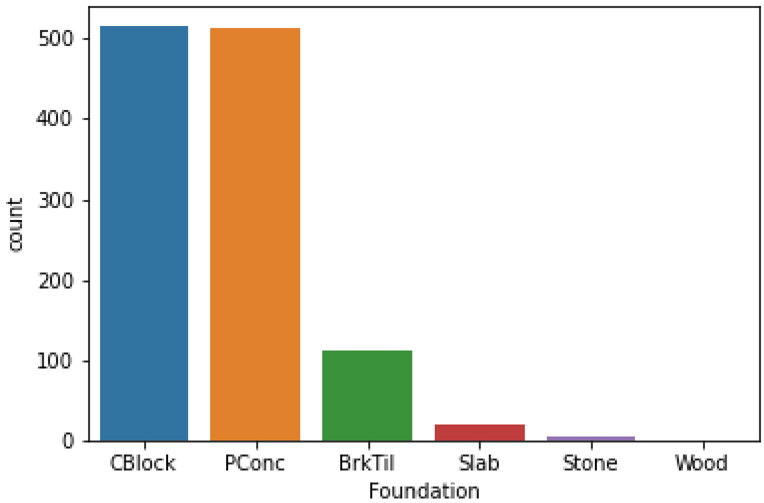

count
Exterior2nd
0
50
100
150
200
250
300
350
400


count
MasVnrType
None
BrkFace
Stone
BrkCmn
0
100
200
300
400
500
600
700

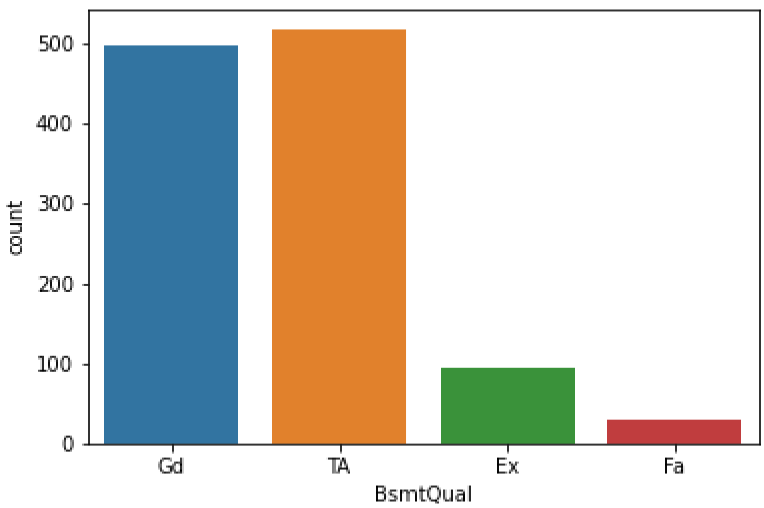

count
ExterQual
TA
Gd
Ex
Fa
0
100
200
300
400
500
600
700



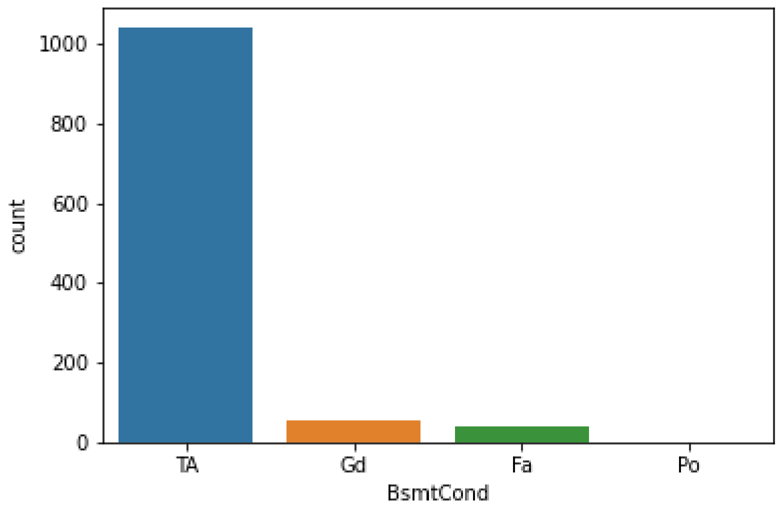
count
1000
800
600
400
200
0
TA
Gd
Fa
Po
Ex
ExterCond



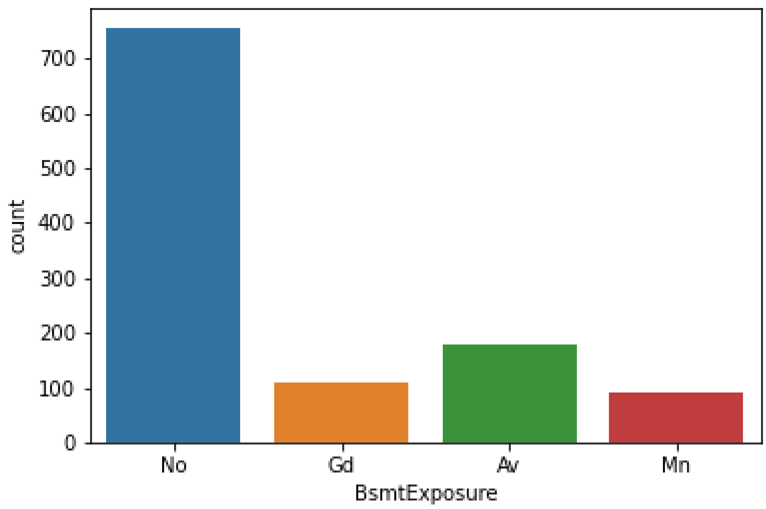
count
500
400
300
200
100
0
CBlock
PConc
BrkTil
Slab
Stone
Wood
Foundation



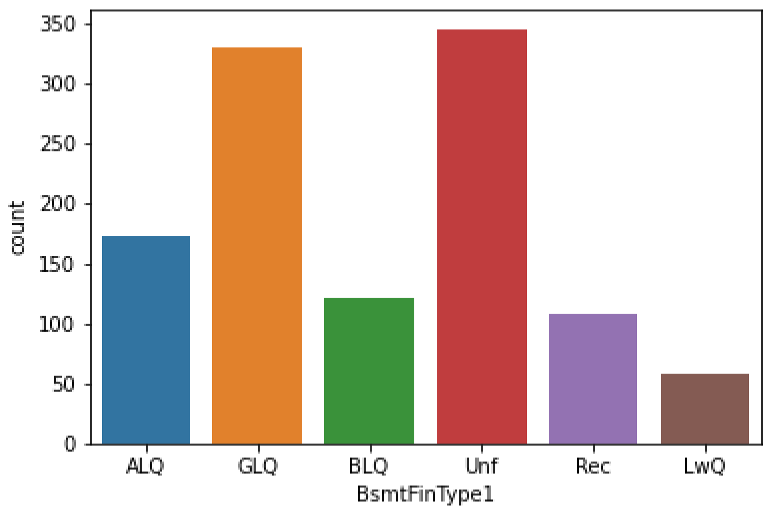
count
500
400
300
200
100
0
Gd
TA
Ex
Fa
BsmtQual



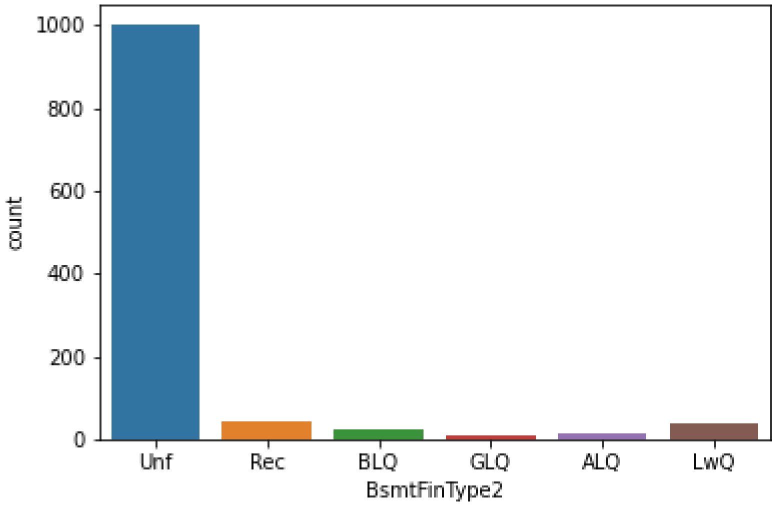
1000
800
600
400
200
0
count
TA
Gd
Fa
Po
BsmtCond



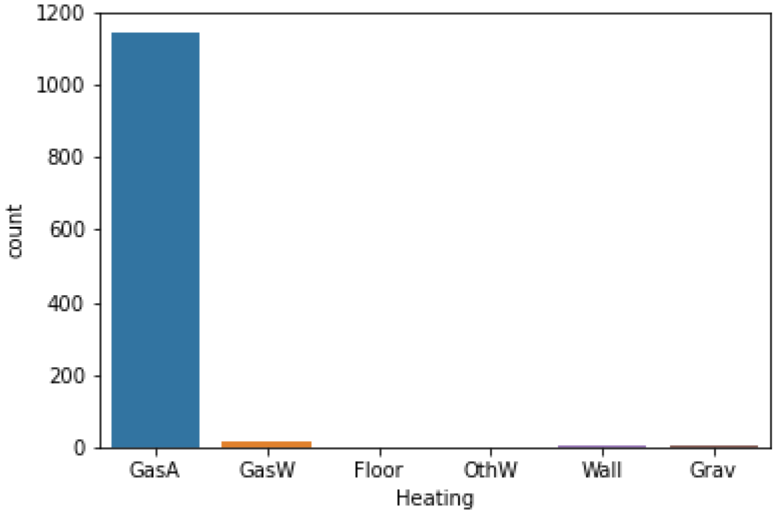
700
600
500
400
300
200
100
0
count
No
Gd
Av
Mn
BsmtExposure



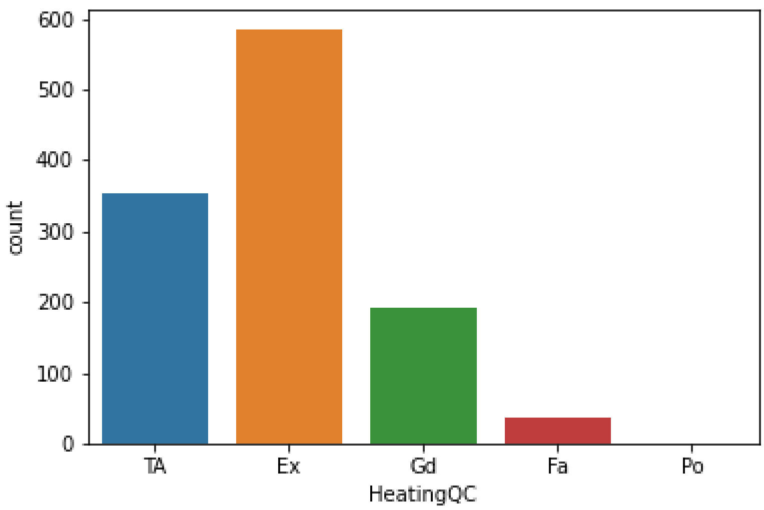
350
300
250
200
150
100
50
0
count
ALQ
GLQ
BLQ
Unf
Rec
LwQ
BsmtFinType1



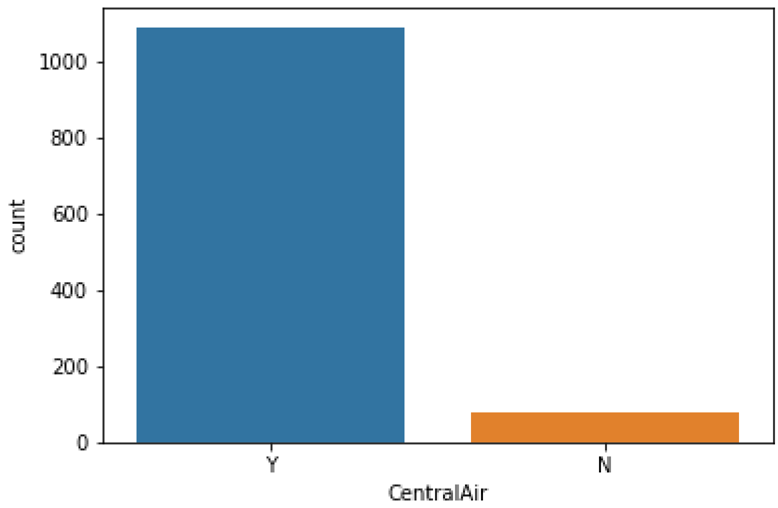
1000
800
600
400
200
0
count
Unf
Rec
BLQ
GLQ
ALQ
LwQ
BsmtFinType2

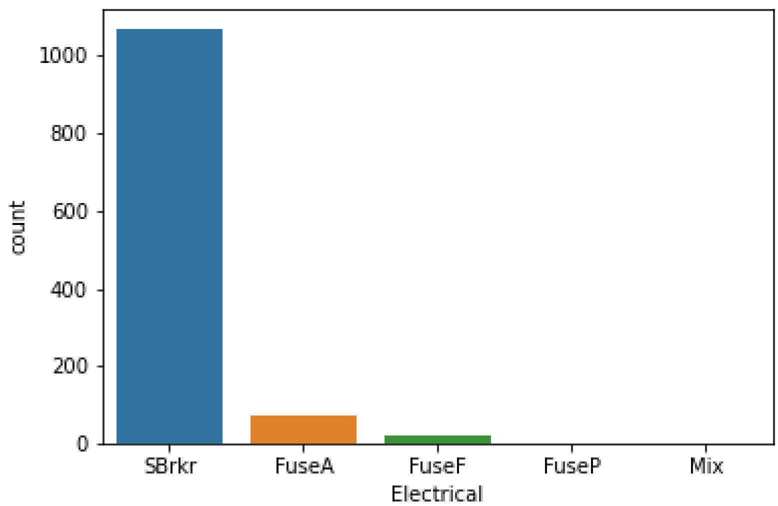


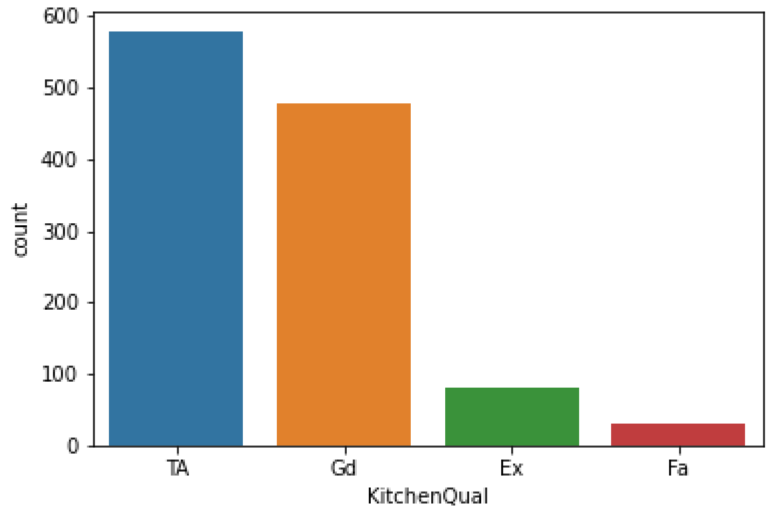
1200
1000
800
600
400
200
0
count
GasA
GasW
Floor
OthW
Wall
Grav
Heating

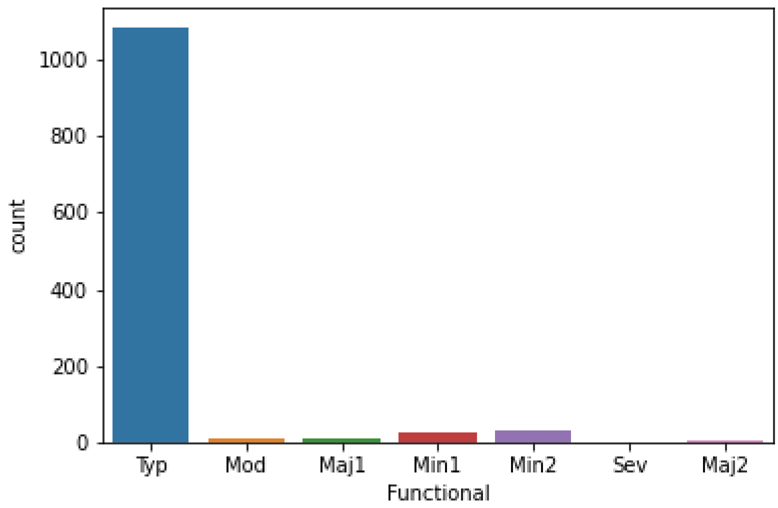


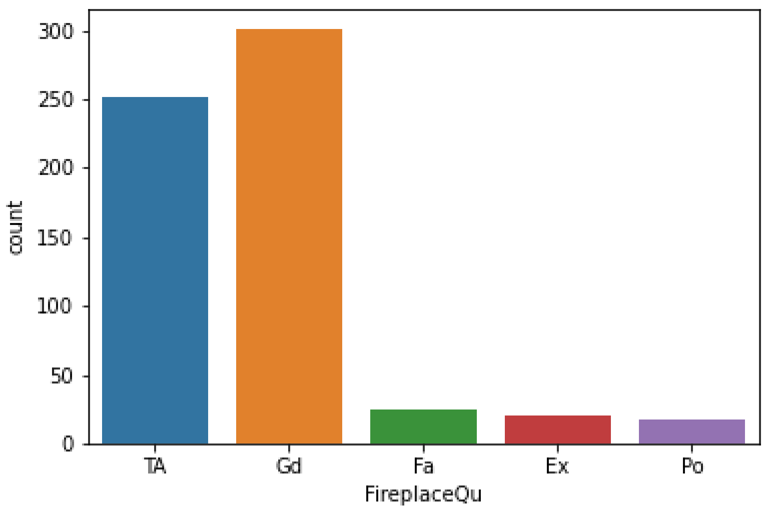
600
500
400
300
200
100
0
count
TA
Ex
Gd
Fa
Po
HeatingQC

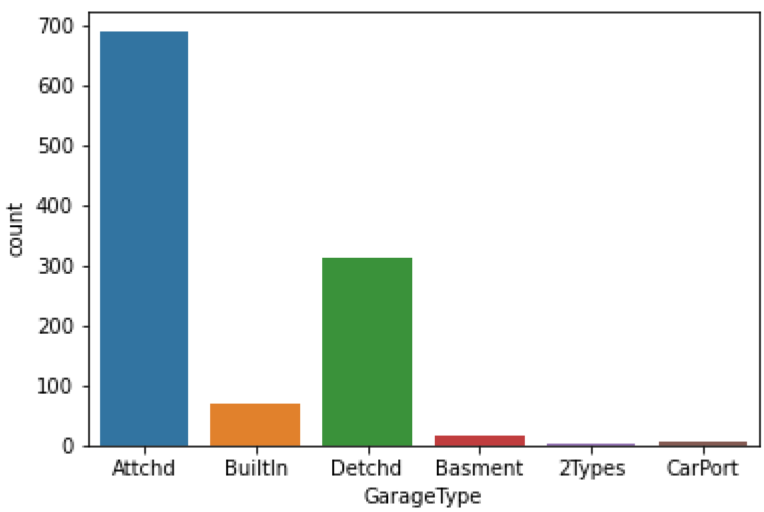

1000
800
600
400
200
0
count
Y
N
CentralAir

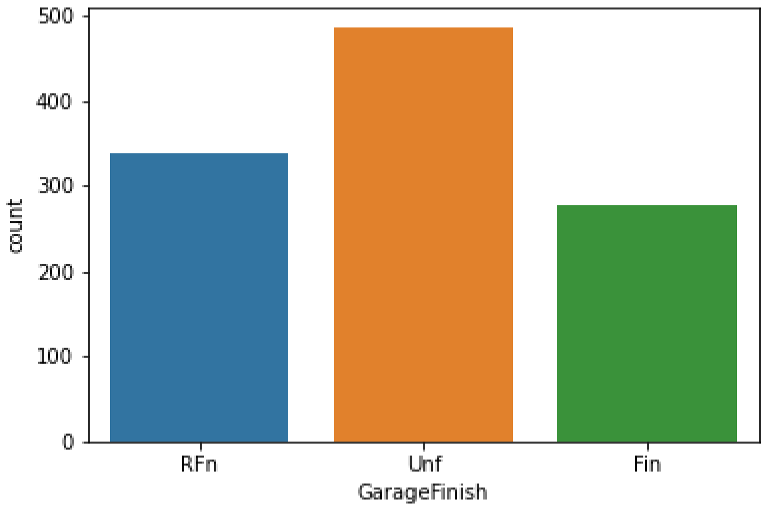

1000
800
600
400
200
0
count
SBrkr
FuseA
FuseF
FuseP
Mix
Electrical

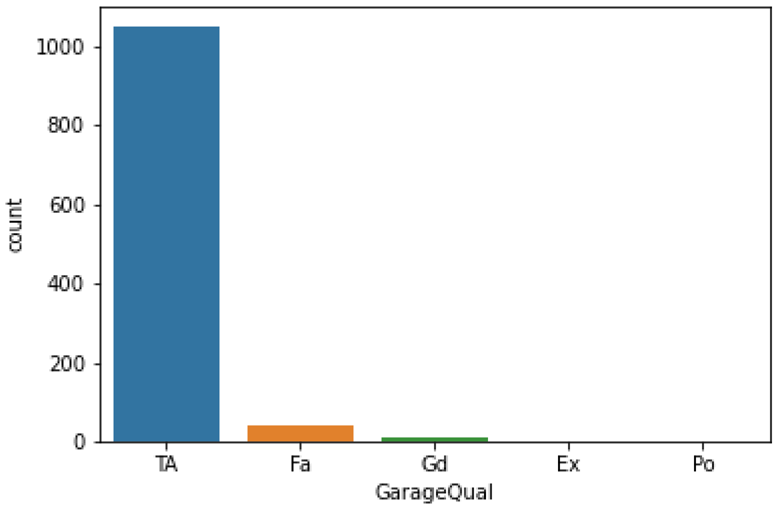

600
500
400
300
200
100
0
count
TA
Gd
Ex
Fa
KitchenQual

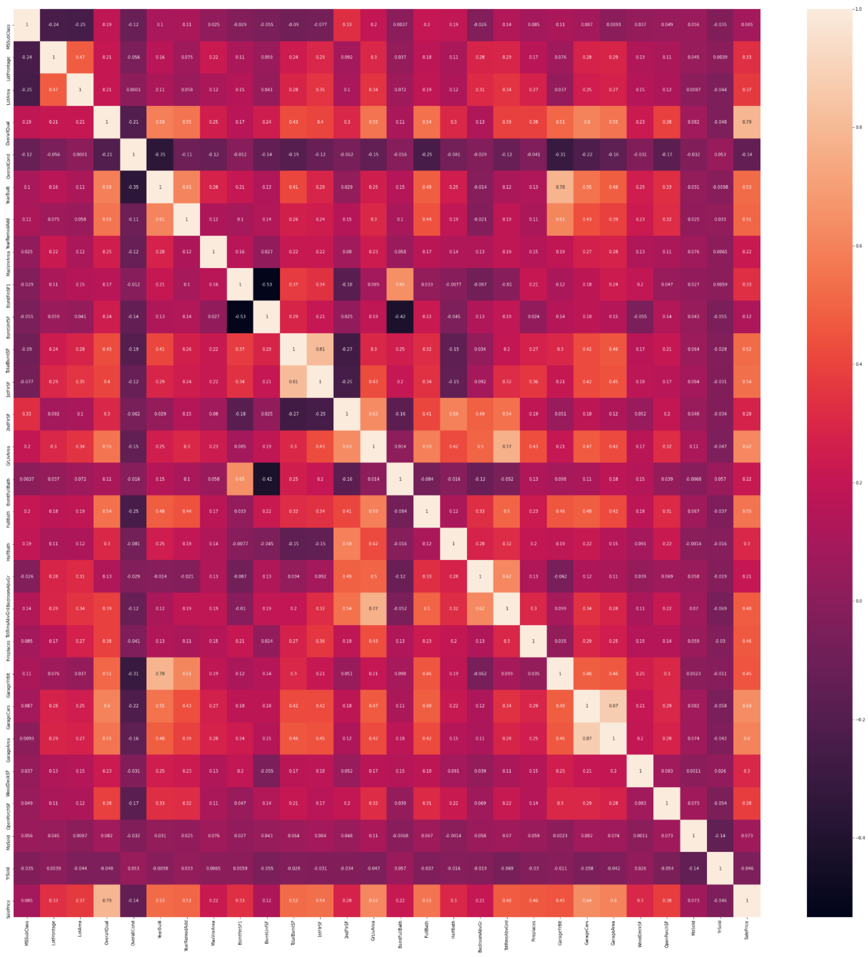

1000
800
600
400
200
0
count
Typ
Mod
Maj1
Min1
Min2
Sev
Maj2
Functional


300
250
200
150
100
50
0
count
TA
Gd
Fa
Ex
Po
FireplaceQu


700
600
500
400
300
200
100
0
count
Attchd
BuiltIn
Detchd
Basment
2Types
CarPort
GarageType


500
400
300
200
100
0
count
RFn
Unf
Fin
GarageFinish


1000
800
600
400
200
0
count
TA
Fa
Gd
Ex
Po
GarageQual



- Interpretation of the Results

  * In sub class, Price decrease with increase in value within sub division of sub class(in the base of stories and styles)

  * Lot Fontage has no direct relationship with price. It might have dependent with other features

  * Price increases with overall quality

  * In over all condition, Price is higher in the middle range

  * Year built is approximately +ve correlated with price, but we have to check with other parameters too

  * BsmtFinSF2 column has only 1 value which is zero. so we can drop that

  * Other basement area(sq ft) related features is almost directly propotional to saleprice

  * Floor squre feet is + correlated with sale price

  * LowQualFinSF column has only 1 value which is zero. so we can drop that

  * BsmtHalfBath,kitchenabvgr,EnclosedPorch,3ssnporch, screenporch,poolarea,miscval columns have only 1 value which is zero. so we can drop that

  * GrLivArea is directly correlated with price

  * FullBath, BedroomAbvgr,fireplaces, GarageYrBlt,GarageCars is directly correlated with price

  * columns MSSubClass, OverallQual, OverallCond, BsmtFullBath, Fullbath, HalfBath, BedroomAbvGr, TotRmsAbvGrd, FirePlaces, GarageCars, MoSols, YrSold can be treated as categorical values

  * basment full bath and basementunitsf are -vely correlated

  * First floor sqft and basement sqft are highly +ve correlated

  * Overall quality and price are highly correlated

* garage car and garage area are +vely correlated

* Price decreases from MZZoning RL to C

* Houses in Paved street is more costly

* Houses in flat level have more price and banked houses cost less

* House price decreases with slope of the place

* Typical basement houses have more price than excellent basement . So there is less over all consideration for basement

* Houses with gas forced heating have more price

* Central air conditioned houses have more price

* Mixed wired houses have lesser price

* Houses with attached garage have more price

* Houses with paved driveway have more price

CONCLUSION

- Key Findings and Conclusions of the Study

  Found out the key features(like living area, squarefeet, overall quality, etc) that related to the price of the house and abled to make a machine learning model that predicts the house price.

- Learning Outcomes of the Study in respect of Data Science

  Got the opportunity to work with different Ml algorithms like SVR, KNN, etc

- Limitations of this work and Scope for Future Work

  The city in which the house is located is not provided in the dataset. I think city plays an important role in the price as we got information from different research papers