Housing Project

Submitted By-
Akanksha Amarnani

Acknowledgement

I would like to thank Almighty for giving me the confidence to pursue this project. Further, the concepts from DataTrained Academy guided me to complete the project.
In addition I would like to thank my mentor from Flip Robo Technology, Ms Khushboo Garg for clarifying my doubts and queries.

The references used for the completion of this project are-

- House Price Prediction; Kristianstad University Sweden, Ahmad Abdulal & Nawar Aghi
- House Price Prediction using regression techniques; Shri Mata Vaishno Devi University, Udit Deo and Uday Deo
- Housing Price Project Report; Math Industry Workshop, Daniel Di Benedetto, Leimin Gao, Yiwei Huang, Neha Sharma and Dongying Wang

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. 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.

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. The company is looking at prospective properties to buy houses to enter the market.

The model build using Machine Learning should predict the actual value of the prospective properties and decide whether to invest in them or not

Conceptual Background of the Domain Problem-

The domain related concepts which help us in a better understanding are-

- Exploratory Data Analysis (EDA)- By conducting explanatory data analysis, we obtain a better understanding of our data. This yields insights that can be helpful later when building a model, as well as insights that are independently interesting.

- Feature Selection- In order to avoid overfitting issues, we select top 80% features using SelectPercentile and Chi2

- Modeling- We apply Linear Regression, Random Forest Regressor and Ada Boost Regressor models for prediction of the prices for the house

- Regularization- Models are regularized and the parameters are hypertuned to enhance the efficiency of the models

Review of Literature-

Machine learning is a subfield of Artificial Intelligence (AI) that works with algorithms and technologies to extract useful information from data. Machine learning methods are appropriate in big data since attempting to manually process vast volumes of data would be impossible without the support of machines. Machine learning in computer science attempts to solve problems algorithmically rather than purely mathematically. Therefore, it is based on creating algorithms that permit the machine to learn. However, there are two general groups in machine learning which are supervised and unsupervised. Supervised is where the program gets trained on pre-determined set to be able to predict when a new data is given. Unsupervised is where the program tries to find the relationship and the hidden pattern between the data

The performance of the model build will be measured upon predicting house prices since the prediction in many regression algorithms relies not only on a specific feature but on an unknown number of attributes that result in the value to be predicted. House prices depend on an individual house specification. Houses have a variant number of features that may not have the same cost due to its location. For instance, a big house may have a higher price if it is located in desirable rich area than being placed in a poor neighbourhood. The data used in the experiment will be handled by using a combination of pre-processing methods to improve the prediction accuracy

Motivation for the Problem Undertaken-

The purchase and sale of properties is considered an unpredictable, cumbersome process. Being a data analyst, it seems to be my responsibility to add the element of prediction in every unpredictable scenario, to solve the cumbersome unsure process into a more reliable, dependent matter. Therefore, the project motivated me to go further and predict the unpredictable. Further, every project has a lot to offer as well. The project and its attributes imparted a lot of knowledge about the real-estate sector, its dependable and the various criteria which varies the prices of various properties.

ANALYTICAL PROBLEM FRAMING

EDA Steps and Visualization

- The train datasheet is extracted and saved in a dataframe
- The shape of the dataframe is checked-

  There are 1168 rows and 81 columns
- **The columns areas follows-**

- Id
- MSSubClass
- MSZoning
- LotFrontage
- LotArea
- Street
- Alley
- LotShape
- LandContour
- Utilities
- LotConfig
- LandSlope
- Neighborhood
- Condition1
- Condition2
- BldgType
- HouseStyle
- OverallQual
- OverallCond
- YearBuilt
- YearRemodAdd
- RoofStyle
- RoofMatl
- Exterior1st
- Exterior2nd
- MasVnrType
- MasVnrArea
- ExterQual
- ExterCond
- Foundation
- BsmtQual
- BsmtCond
- BsmtExposure
- BsmtFinType1
- BsmtFinSF1
- BsmtFinType2
- BsmtFinSF2
- BsmtUnfSF
- TotalBsmtSF
- Heating
- HeatingQC
- CentralAir
- Electrical
- 1stFlrSF
- 2ndFlrSF
- LowQualFinSF
- GrLivArea
- BsmtFullBath
- BsmtHalfBath
- FullBath
- HalfBath
- BedroomAbvGr
- KitchenAbvGr
- KitchenQual
- TotRmsAbvGrd
- Functional
- Fireplaces
- FireplaceQu
- GarageType
- GarageYrBlt
- GarageFinish
- GarageCars
- GarageArea
- GarageQual
- GarageCond
- PavedDrive
- WoodDeckSF
- OpenPorchSF
- EnclosedPorch
- 3SsnPorch
- ScreenPorch
- PoolArea
- PoolQC
- Fence
- MiscFeature
- MiscVal
- MoSold
- YrSold
- SaleType
- SaleCondition
- SalePrice

The data type of each column is-

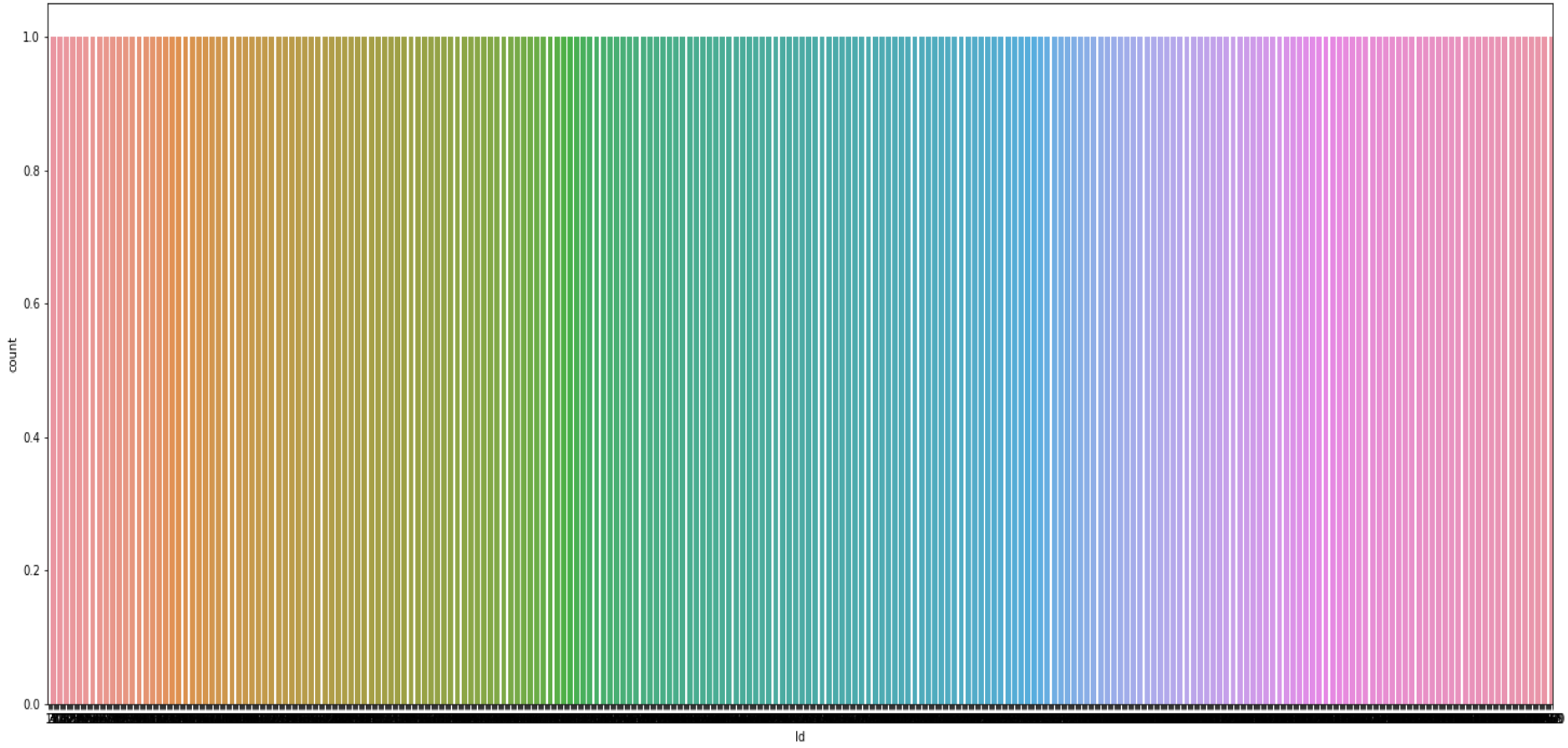
- Id ---------- int64
- MSSubClass ---------- int64
- MSZoning ---------- object
- LotFrontage ---------- float64
- LotArea ---------- int64
- Street ---------- object
- Alley ---------- object
- LotShape ---------- object
- LandContour ---------- object
- Utilities ---------- object
- LotConfig ---------- object
- LandSlope ---------- object
- Neighborhood ---------- object
- Condition1 ---------- object
- Condition2 ---------- object
- BldgType ---------- object
- HouseStyle ---------- object
- OverallQual ---------- int64
- OverallCond ---------- int64 Y
- earBuilt ---------- int64
- YearRemodAdd ---------- int64
- RoofStyle ---------- object
- RoofMatl ---------- object
- Exterior1st ---------- object
- Exterior2nd ---------- object
- MasVnrType ---------- object
- MasVnrArea ---------- float64
- ExterQual ---------- object
- ExterCond ---------- object
- Foundation ---------- object
- BsmtQual ---------- object
- BsmtCond ---------- object
- BsmtExposure ---------- object
- BsmtFinType1 ---------- object
- BsmtFinSF1 ---------- int64
- BsmtFinType2 ---------- object
- BsmtFinSF2 ---------- int64
- BsmtUnfSF ---------- int64
- TotalBsmtSF ---------- int64
- Heating ---------- object
- HeatingQC ---------- object
- CentralAir ---------- object
- Electrical ---------- object
- 1stFlrSF ---------- int64
- 2ndFlrSF ---------- int64
- LowQualFinSF ---------- int64
- GrLivArea ---------- int64
- BsmtFullBath ---------- int64
- BsmtHalfBath ---------- int64
- FullBath ---------- int64
- HalfBath ---------- int64
- BedroomAbvGr ---------- int64
- KitchenAbvGr ---------- int64
- KitchenQual ---------- object
- TotRmsAbvGrd ---------- int64
- Functional ---------- object
- Fireplaces ---------- int64
- FireplaceQu ---------- object
- GarageType ---------- object
- GarageYrBlt ---------- float64
- GarageFinish ---------- object
- GarageCars ---------- int64
- GarageArea ---------- int64
- GarageQual ---------- object
- GarageCond ---------- object
- PavedDrive ---------- object
- WoodDeckSF ---------- int64
- OpenPorchSF ---------- int64
- EnclosedPorch ---------- int64
- 3SsnPorch ---------- int64
- ScreenPorch ---------- int64
- PoolArea ---------- int64
- PoolQC ---------- object
- Fence ---------- object
- MiscFeature ---------- object
- MiscVal ---------- int64
- MoSold ---------- int64
- YrSold ---------- int64
- SaleType ---------- object
- SaleCondition ---------- object
- SalePrice ---------- int64

- The null values are present in the following columns-

Column	No of null values
LotFrontage	214
Alley	1091
MasVnrType	7
MasVnrArea	7
BsmtQual	30
BsmtCond	30
BsmtExposure	31
BsmtFinType1	30
BsmtFinType2	31
FireplaceQu	551
GarageType	64
GarageYrBlt	64
GarageFinish	64
GarageQual	64
GarageCond	64
PoolQC	1161
Fence	931
MiscFeature	1124

The data visualization, value counts encoding and imputation of null values for each column

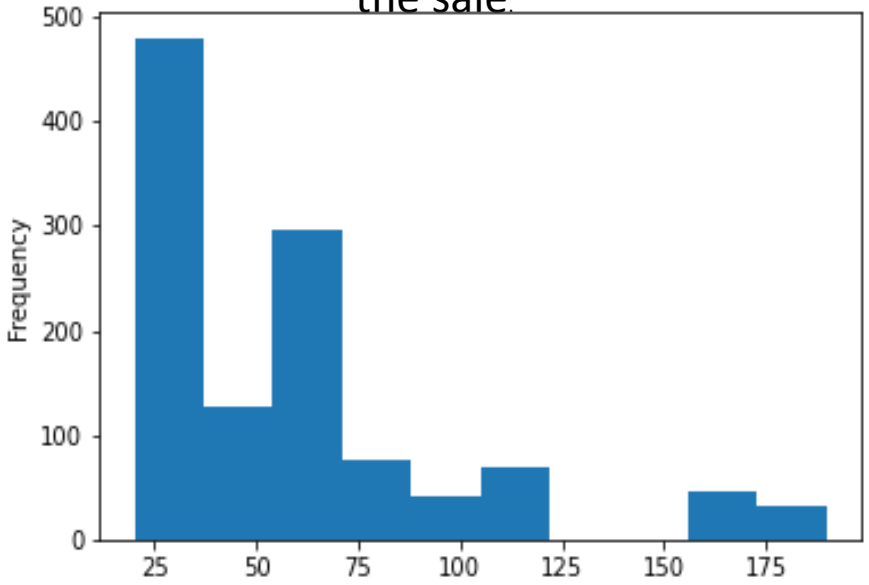
ID



As the id is unique to all, its safe to drop this column

MSSubClass

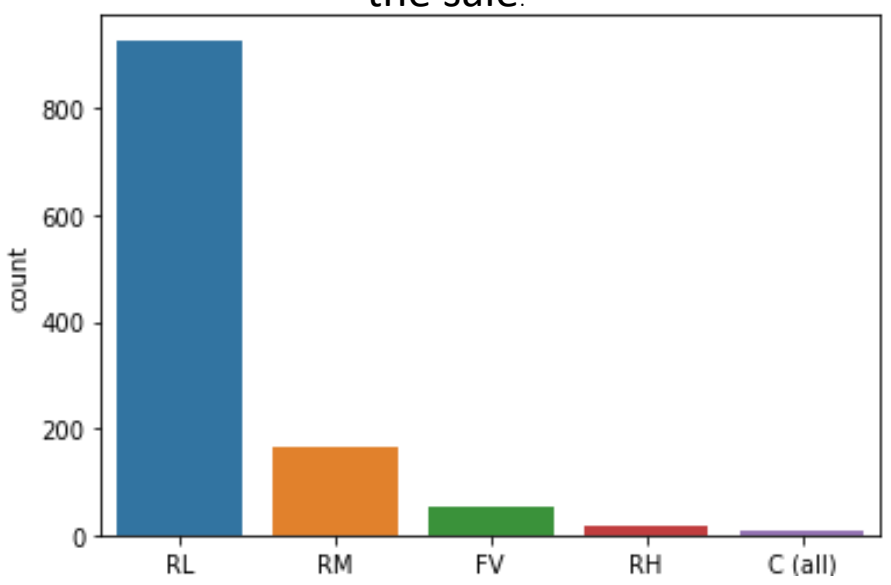
Identifies the type of dwelling involved in the sale.



- 428 properties are 20 (1-STORY 1946 & NEWER ALL STYLES)
- 244 properties are 60 (2-STORY 1946 & NEWER)
- 113 properties are 50 (1-1/2 STORY FINISHED ALL AGES)
- 69 properties are 120 (1-STORY PUD (Planned Unit Development) – 1946 & NEWER)
- 53 properties are 70 (2-STORY 1945 & OLDER)
- 52 properties are 30 (1-STORY 1945 & OLDER)
- 47 properties are 160 (2-STORY PUD - 1946 & NEWER)
- 43 properties are 80 (SPLIT OR MULTI-LEVEL)
- 41 properties are 90 (DUPLEX - ALL STYLES AND AGES)
- 26 properties are 190 (2 FAMILY CONVERSION - ALL STYLES AND AGES)
- 19 properties are 85 (SPLIT FOYER)
- 14 properties are 75 (2-1/2 STORY ALL AGES)
- 10 properties are 45 (1-1/2 STORY - UNFINISHED ALL AGES)
- 6 properties are 180 (PUD - MULTILEVEL - INCL SPLIT LEV/FOYER)
- 3 properties are 40 (1-STORY W/FINISHED ATTIC ALL AGES)

MSZoning

Identifies the general zoning classification of the sale.



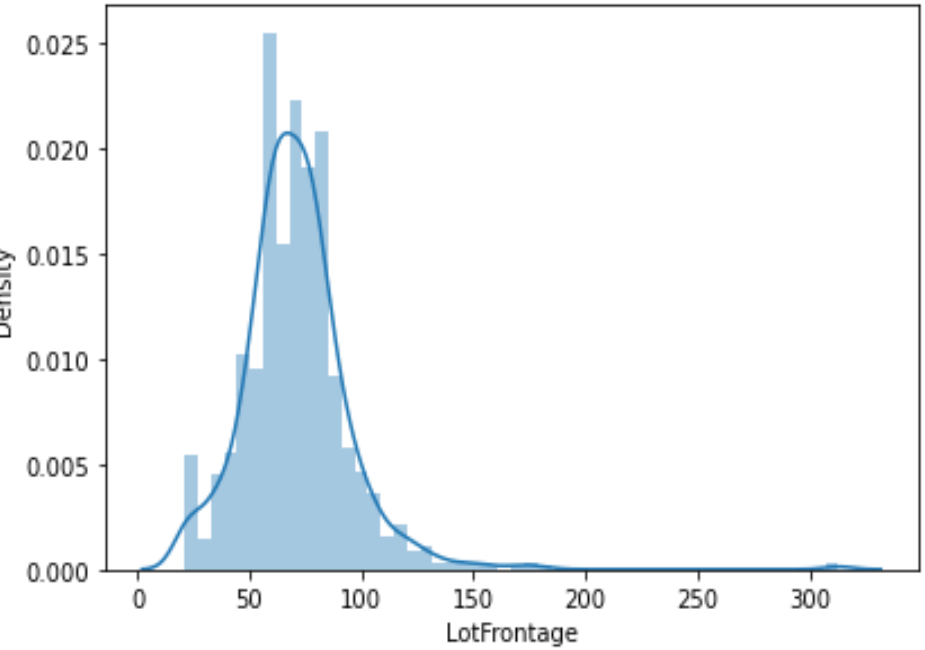
- 928 properties are RL (Residential Low Density)
- 163 properties are RM (Residential Medium Density)
- 52 properties are FV (Floating Village Residential)
- 16 properties are RH (Residential High Density)
- 9 properties are C (Commercial)

↓

Encoding object data in numeric using Label Encoder

LotFrontage

Linear feet of street connected to property

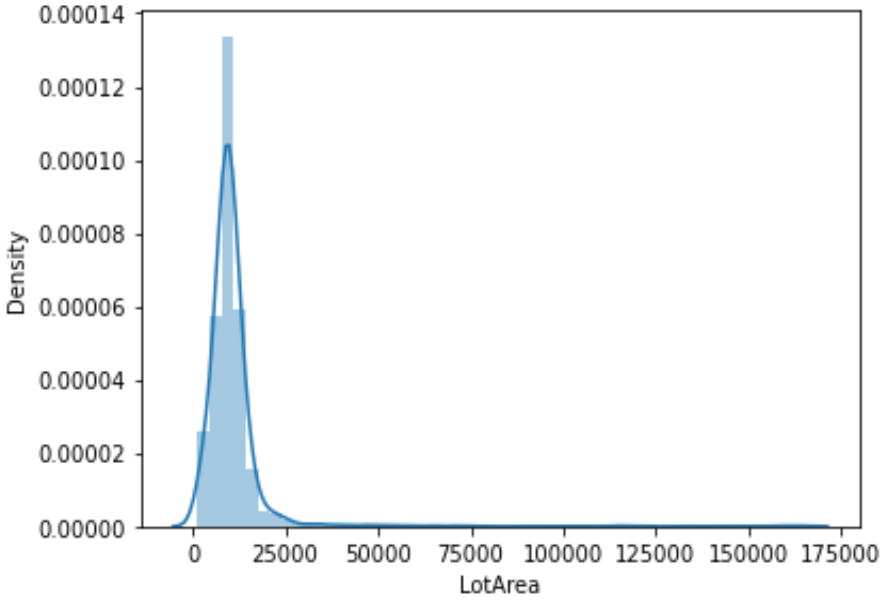


- 111 properties have LotFrontage of 60

The data is slightly skewed and will be transformed later

LotArea

Lot size in square feet

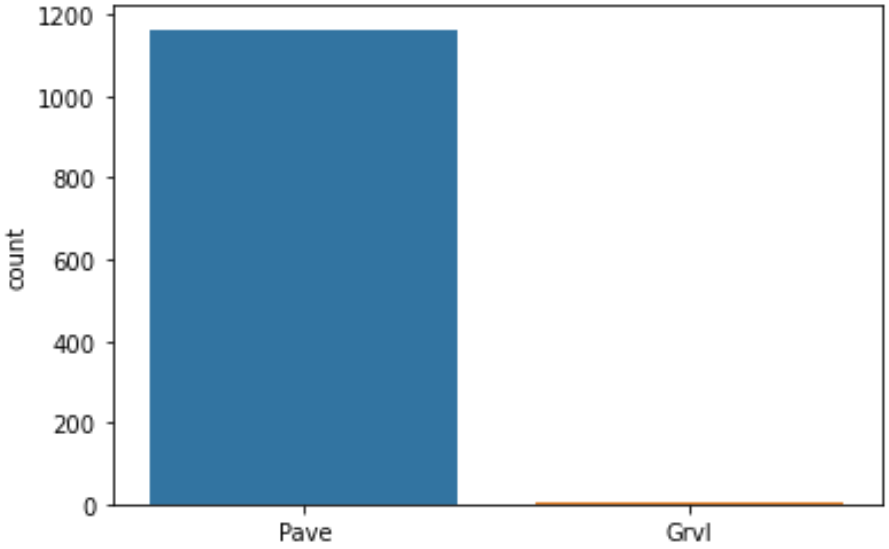


- 21 properties have LotArea of 9600

The data is skewed and will be transformed later

Street

Type of road access to property



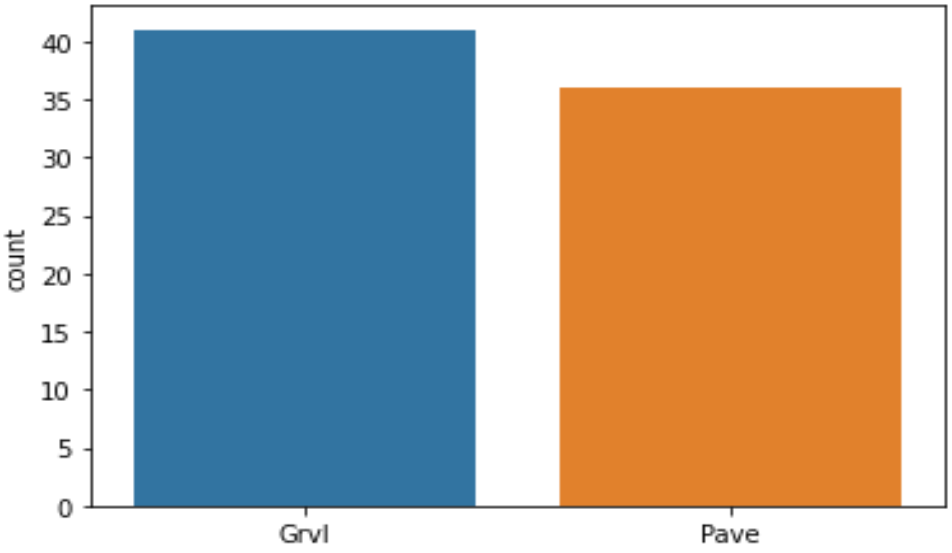
- 1164 properties have Pave (paved) road
- 4 properties have Grvl (gravel) road

↓

Encoding object data in numeric using Label Encoder

Alley

Type of alley access to property



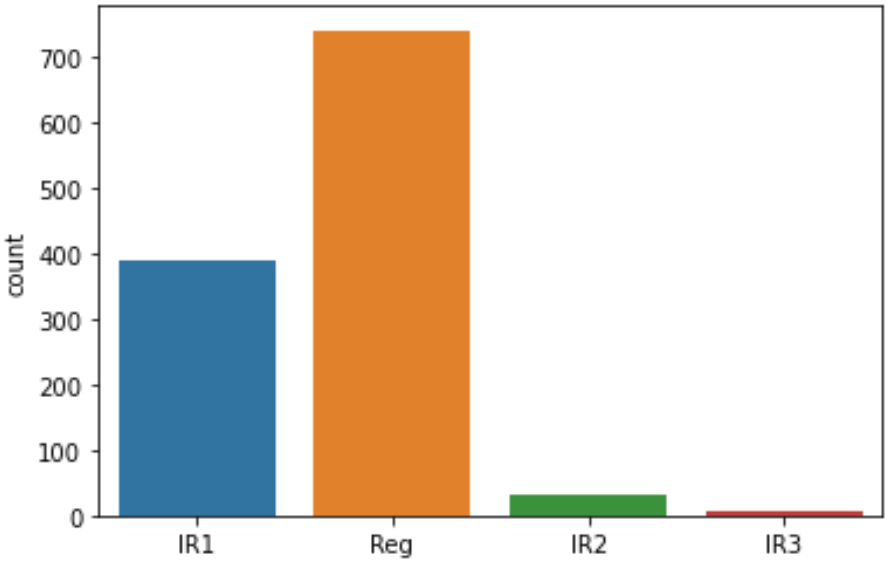
- 1091 have No alley access
- 41 have Grvl (Gravel) alley
- 36 have Pave (paved) alley

↓

Encoding object data in numeric using Label Encoder

LotShape

General shape of property

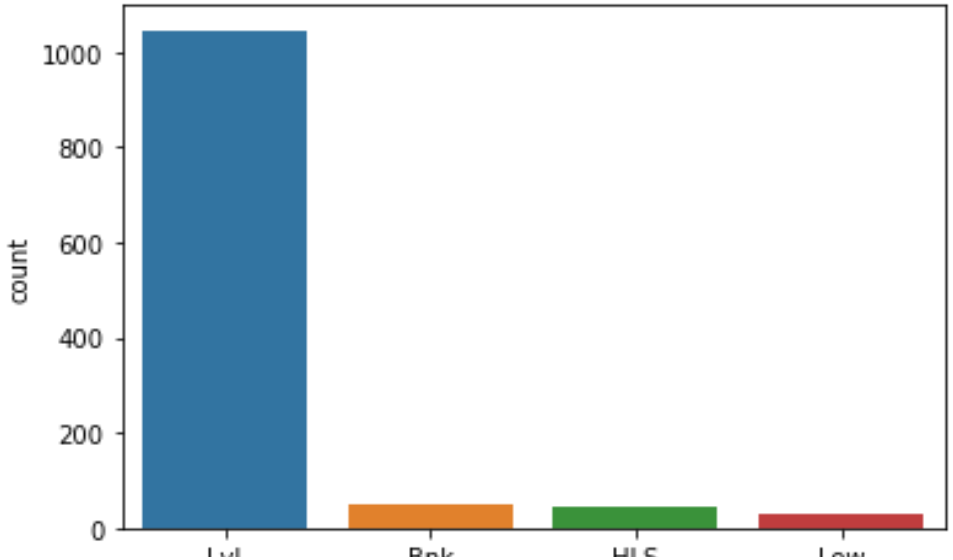


- 740 properties are Reg (Regular)
- 390 properties are IR1 (Slightly Irregular)
- 32 properties are IR2 (Moderately irregular)
- 6 properties are IR3 (Irregular)

Encoding object data in numeric using Label Encoder

LandContour

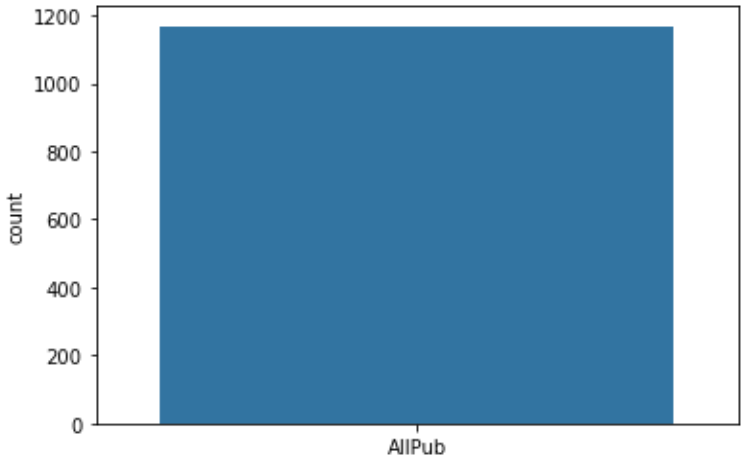
Flatness of the property



- 1046 properties are Lvl (Near Flat/Level)
- 50 properties are Bnk (Banked - Quick and significant rise from street grade to building)
- 42 properties are HLS (Hillside - Significant slope from side to side)
- 30 properties are Low depression

Encoding object data in numeric using Label Encoder

Utilities



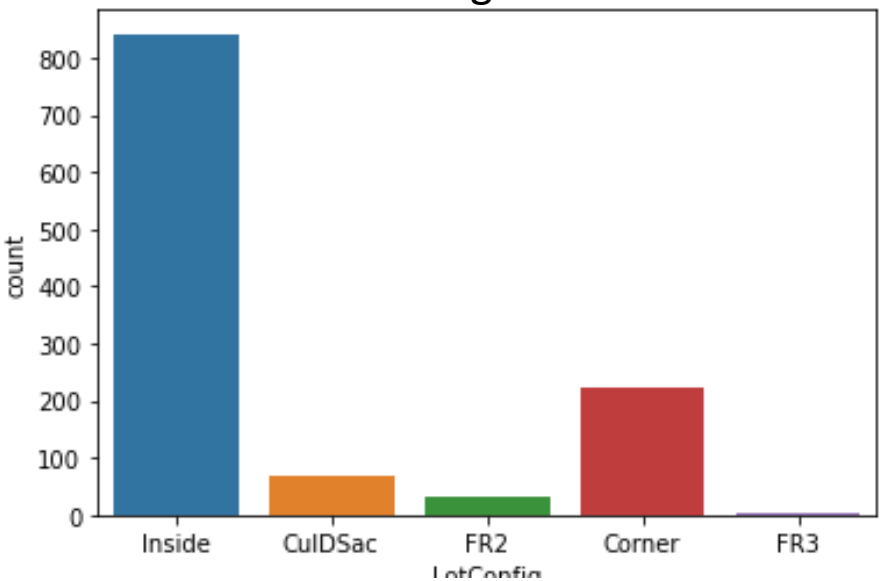
- 1168 properties have AllPub [All public Utilities (E,G,W,& S)]

↓

As all properties have the same utilities, its safe to drop this column

LotConfig

Lot configuration



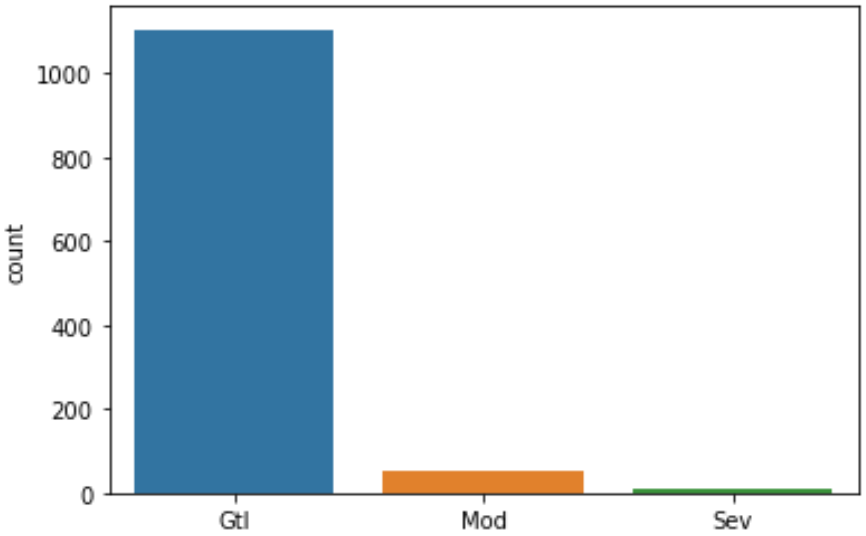
- 842 properties are inside lot
- 222 properties are corner lot
- 69 properties are Cul-de-sac
- 33 properties are Frontage on 2 sides of property
- 2 properties are Frontage on 3 sides of property

↓

Encoding object data in numeric using Label Encoder

LandSlope

Slope of property



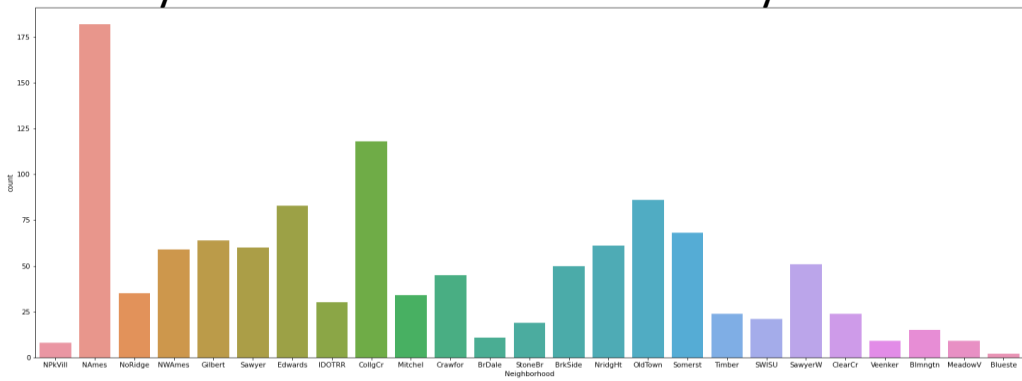
- 1105 properties have Gtl (Gentle slope)
- 51 properties have Mod (Moderate slope)
- 12 properties have Sev (Severe slopes)

↓

Encoding object data in numeric using Label Encoder

Neighborhood

Physical locations within Ames city limits



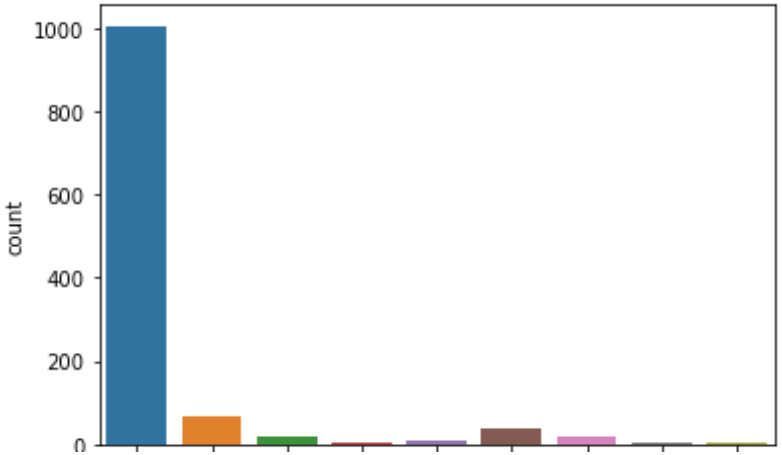
- 182 properties are in Names (North Ames)
- 118 properties are in CollgCr (College Creek)
- 86 properties are in OldTown
- 83 properties in Edwards
- 68 properties are in Somerst
- 64 properties are in Gilbert
- 61 properties are in NridgHt (Northridge Heights)
- 60 properties are in Sawyer
- 59 properties are in NWAmes (Northwest Ames)
- 51 properties are in SawyerW (Sawyer West)
- 50 properties are in BrkSide (Brookside)
- 45 properties are in Crawfor (Crawford)
- 35 properties are in NoRidge (Northridge)
- 34 properties are in Mitchel
- 30 properties are in IDOTRR (Iowa DOT and Rail Road)
- 24 properties are in Timber (Timberland)
- 24 properties are in ClearCr (Clear Creek)
- 21 properties are in SWISU (South & West of Iowa State University)
- 19 properties are in Stone Brook
- 15 properties are in Blmngtn (Bloomington Heights)
- 11 properties are in BrDale (Briardale)
- 9 properties are in MeadowV (Meadow Village)
- 9 properties are in Veenker
- 8 properties are in NPkVill (Northpark Villa)
- 2 properties are in Blueste (Bluestem)

↓

Encoding object data in numeric using Label Encoder

Condition1

Proximity to various conditions



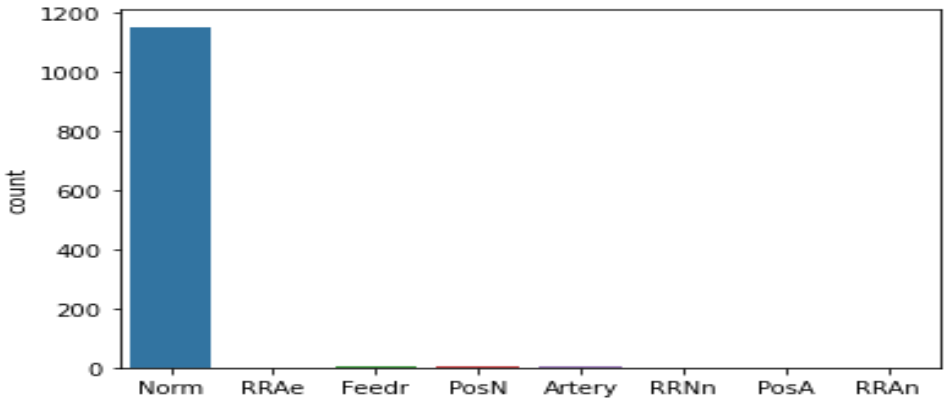
- 1005 properties are Norm (Normal)
- 67 properties are Feedr (Adjacent to feeder street)
- 38 properties are Artery (Adjacent to arterial street)
- 20 properties are RRAn (Adjacent to North-South Railroad)
- 17 properties are PosN (Near positive off-site feature--park, greenbelt, etc)
- 9 properties are RRAe (Adjacent to East-West Railroad)
- 6 properties are PosA (Adjacent to postive off-site feature)
- 4 properties are RRNn (Within 200' of North-South Railroad)
- 2 properties are RRNe (Within 200' of East-West Railroad)

↓

Encoding object data in numeric using Label Encoder

Condition2

Proximity to various conditions
(if more than one is present)



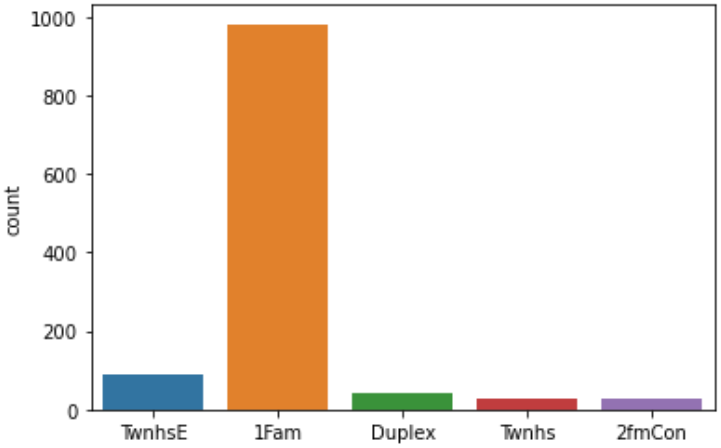
- 1154 properties are Norm (Normal)
- 6 properties are Feedr (Adjacent to feeder street)
- 2 properties are PosN (Near positive off-site feature--park, greenbelt, etc)
- 2 properties are Artery (Adjacent to arterial street)
- 1 property is RRAe (Adjacent to East-West Railroad)
- 1 property is RRNn (Within 200' of North-South Railroad)
- 1 property is PosA (Adjacent to postive off-site feature)
- 1 property is RRAn (Adjacent to North-South Railroad)

↓

Encoding object data in numeric using Label Encoder

BldgType

Type of dwelling



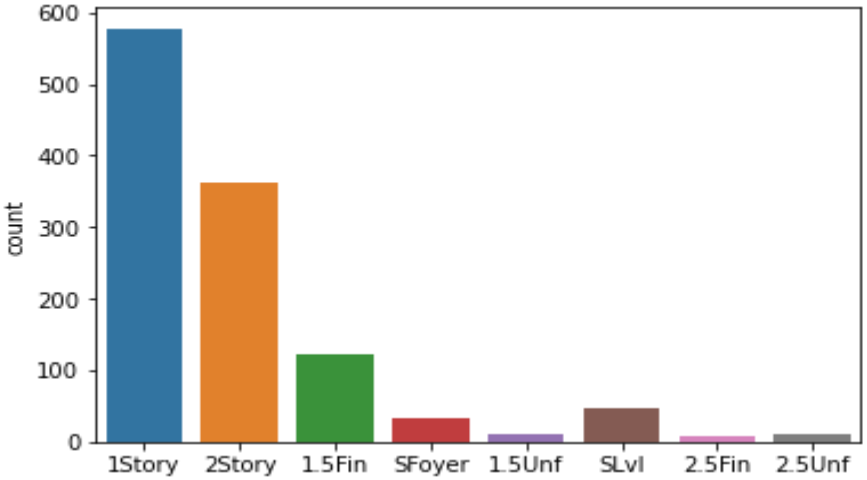
- 981 properties are 1Fam (Single-family Detached)
- 90 properties are TwnhsE (Townhouse End Unit)
- 41 properties are Duplex
- 29 properties are Twnhs (Townhouse Inside Unit)
- 27 properties are 2fmCon (Two-family Conversion; originally built as one-family dwelling)

↓

Encoding object data in numeric using Label Encoder

HouseStyle

Style of dwelling



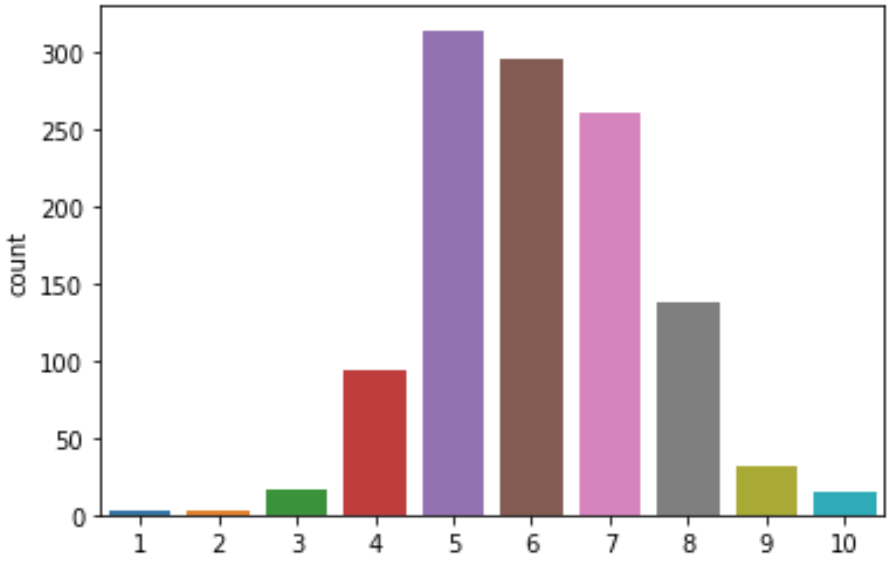
- 578 properties are 1Story (One story)
- 2Story 361
- 121 properties are 1.5Fin (One and one-half story: 2nd level finished)
- 47 properties are SLvl (Split Level)
- 32 properties are SFoyer (Split Foyer)
- 12 properties are 1.5Unf (One and one-half story: 2nd level unfinished)
- 10 properties are 2.5Unf (Two and one-half story: 2nd level unfinished)
- 7 properties are 2.5Fin (Two and one-half story: 2nd level finished)

↓

Encoding object data in numeric using Label Encoder

OverallQual

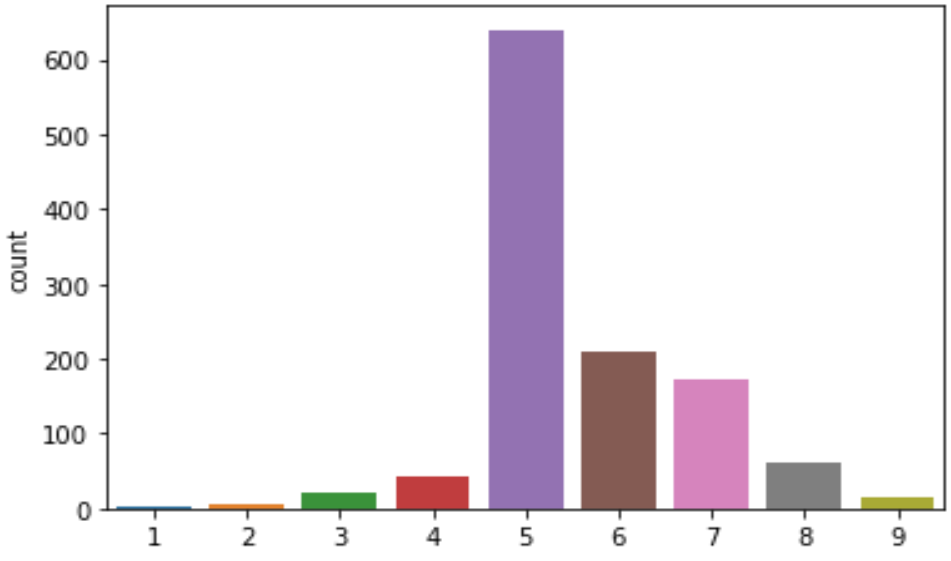
Rates the overall material and finish of the house



- 314 properties are rated 5 (Average)
- 295 properties are rated 6 (Above average)
- 260 properties are rated 7 (Good)
- 138 properties are rated 8 (Very good)
- 93 properties are rated 4 (Below average)
- 32 properties are rate 9 (Excellent)
- 16 properties are rated 3 (Fair)
- 15 properties are rated 10 (Very excellent)
- 3 properties are rated 2 (Poor)
- 2 properties are rated 1 (Very poor)

OverallCond

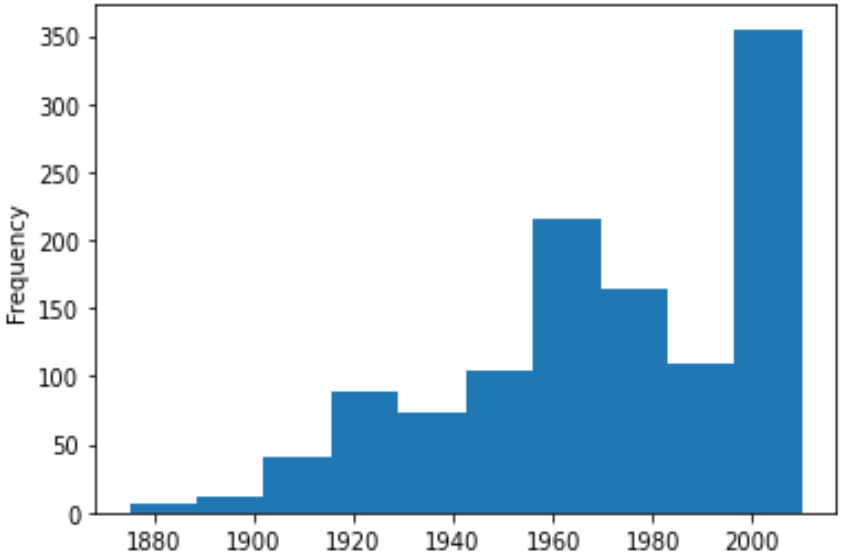
Rates the overall condition of the house



- 640 properties are rated 5 (Average)
- 209 properties are rated 6 (Above average)
- 172 properties are rated 7 (Good)
- 61 properties are rated 8 (Very good)
- 43 properties are rated 4 (Below average)
- 21 properties are rated 3 (Fair)
- 16 properties are rated 9 (Excellent)
- 5 properties are rated 2 (Very poor)
- 1 properties are rated 1 (Very poor)

YearBuilt

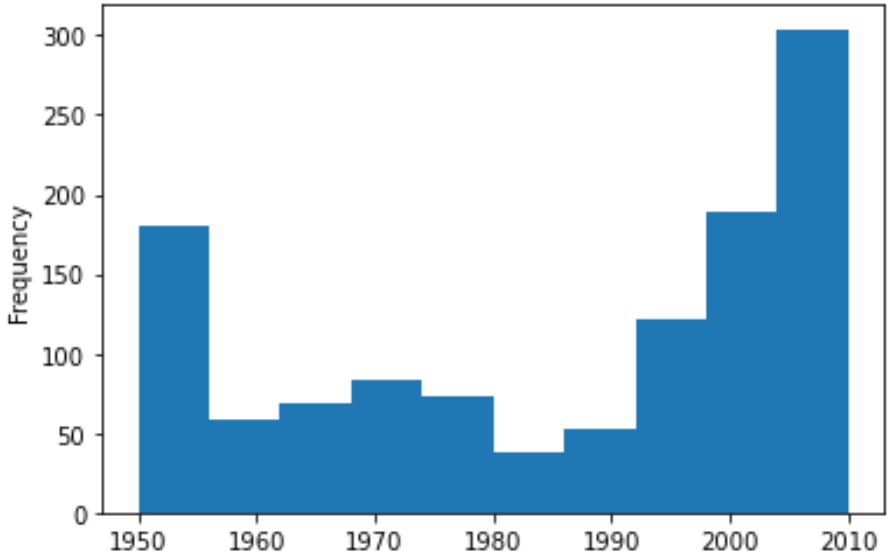
Original construction date



- Majority of the properties are built in 2000

YearRemodAdd

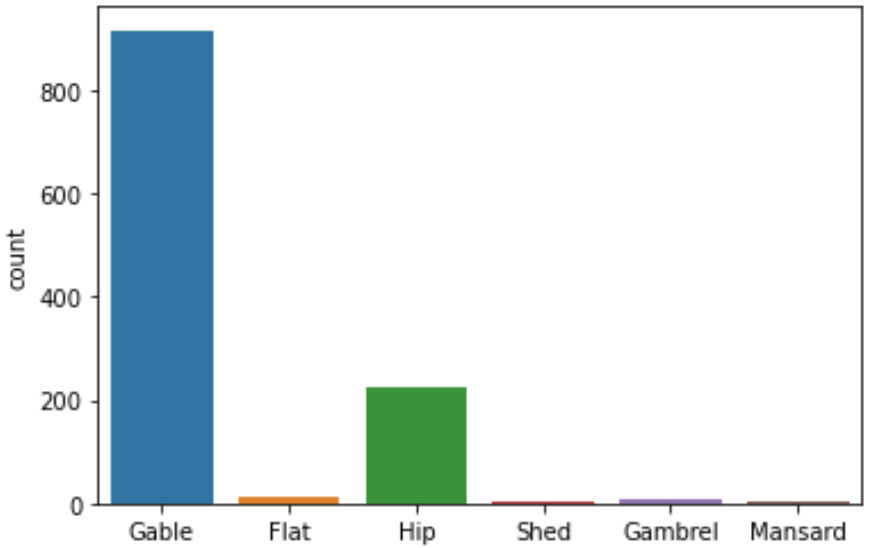
Remodel date (same as construction date if no remodeling or additions)



- Majority of the properties are built in 2010

RoofStyle

Type of roof

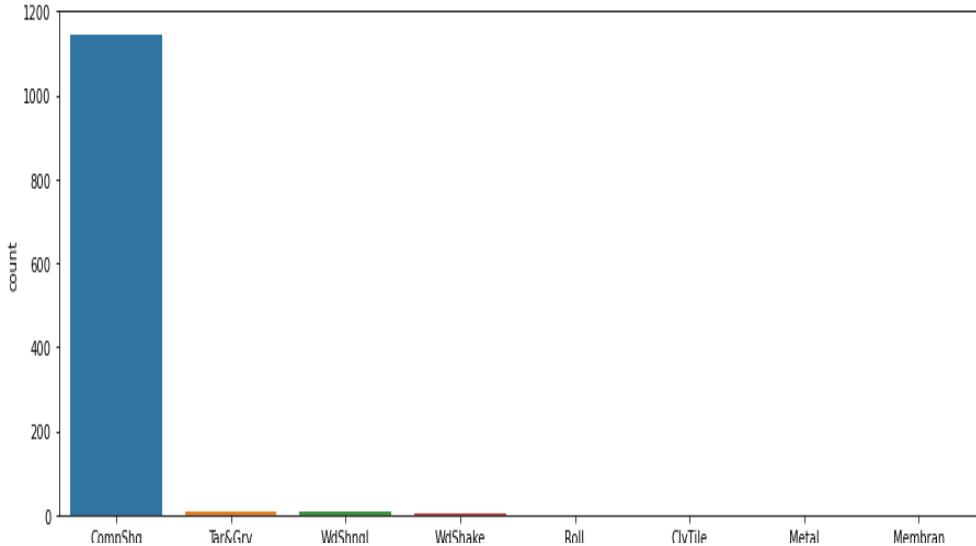


- 915 properties have gable roof
- 225 properties have hip roof
- 12 properties have flat roof
- 9 properties have gambrel roof
- 5 properties have mansard roof
- 2 properties have shed roof

Encoding object data in numeric using Label Encoder

RoofMatl

Roof material

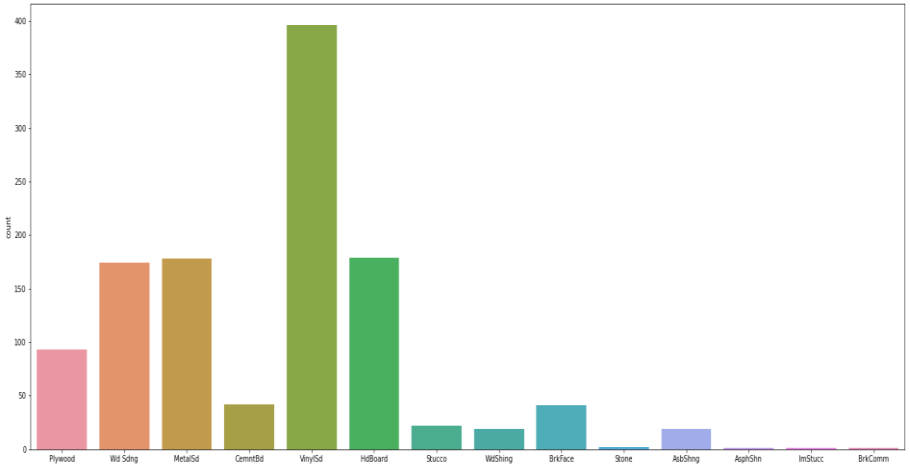


- 1144 properties have CompShg (standard(composite) shingle) roof
- 10 properties have Tar&Grv (gravel & Tar) roof
- 6 properties have WdShngl (wood shingles) roof
- 4 properties have WdShake (wood shakes) roof
- 1 property has Roll roof
- 1 property has ClyTile (clay or tile) roof
- 1 property has Metal roof
- 1 property has Membrane roof

Encoding object data in numeric using Label Encoder

Exterior1st

Exterior covering on house

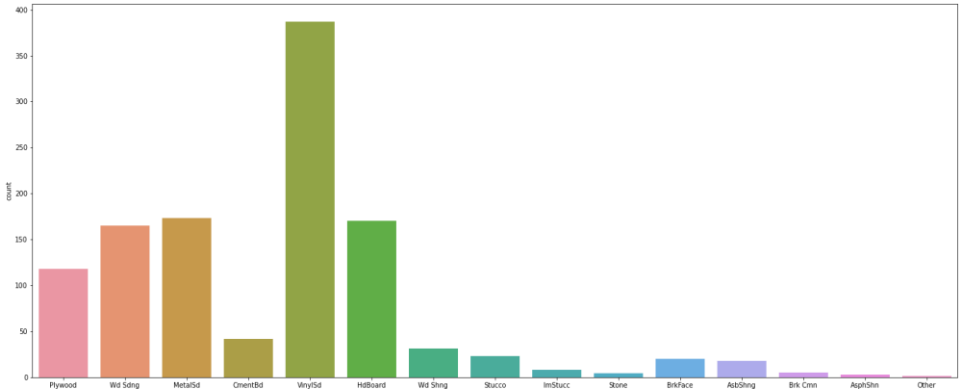


- 396 properties have VinylSd (Vinyl Siding) exterior
- 179 properties have HdBoard (Hard Board) exterior
- 178 properties have MetalSd (Metal Siding) exterior
- 174 properties have Wd Sdng (Wood Siding) exterior
- 93 properties have Plywood exterior
- 42 properties have CemntBd (Cement Board) exterior
- 41 properties have BrkFace (Brick Face) exterior
- 22 properties have Stucco exterior
- 19 properties have WdShing (Wood Shingles) exterior
- 19 properties have AsbShng (Asbestos Shingles) exterior
- 2 properties have Stone exterior
- 1 property has AsphShn (Asphalt Shingles) exterior
- 1 property has ImStucc (Imitation Stucco) exterior
- 1 property has BrkComm (Brick Common) exterior

Encoding object data in numeric using Label Encoder

Exterior2nd

Exterior covering on house (if more than one material)

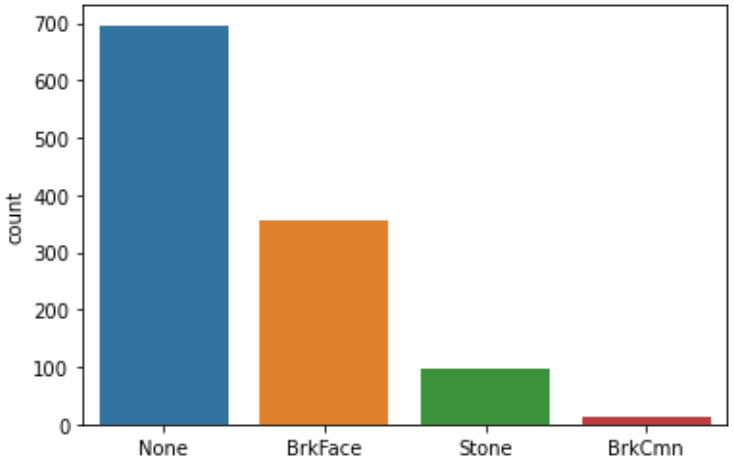


- 387 properties have VinylSd (Vinyl Siding) exterior
- 173 properties have MetalSd (Metal Siding) exterior
- 170 properties have HdBoard (Hard Board) exterior
- 165 properties have Wd Sdng (Wood Siding) exterior
- 118 properties have Plywood exterior
- 42 properties have CmentBd (Cement Board) exterior
- 31 properties have Wd Shng (Wood Shingles) exterior
- 23 properties have Stucco exterior
- 20 properties have BrkFace (Brick face) exterior
- 18 properties have AsbShng (Asbestos Shingles) exterior
- 8 properties have ImStucc (Imitation Stucco) exterior
- 5 properties have Brk Cmn (Brick Common) exterior
- 4 properties have Stone exterior
- 3 properties have AsphShn (Asphalt Shingles) exterior
- 1 property has other exterior

Encoding object data in numeric using Label Encoder

MasVnrType

Masonry veneer type



- None 696 properties lack masonry veneer
- 354 properties have BrkFace (Brick Face) masonry veneer
- 98 properties have Stone masonry veneer
- 13 properties have BrkCmn (Brick Common) masonry veneer

↓

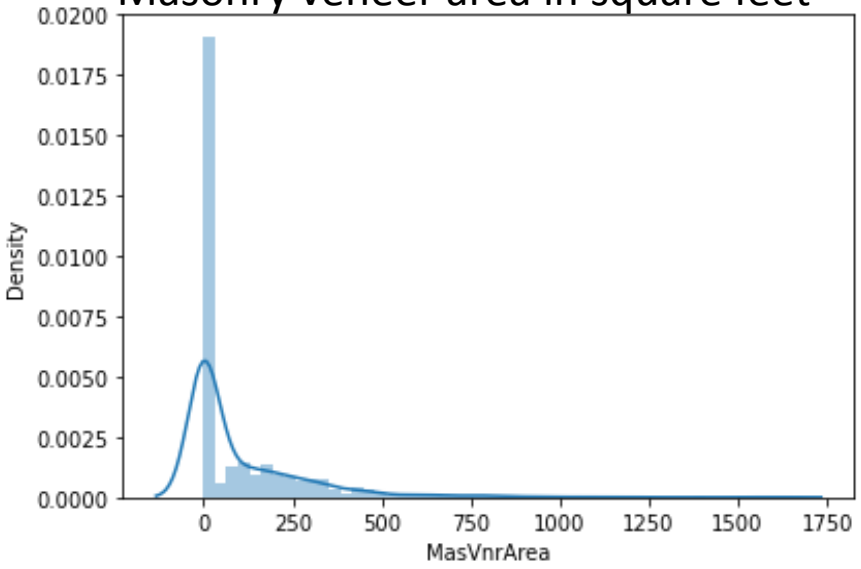
Encoding object data in numeric using Label Encoder

↓

Replacing null value with the mode of the column

MasVnrArea

Masonry veneer area in square feet



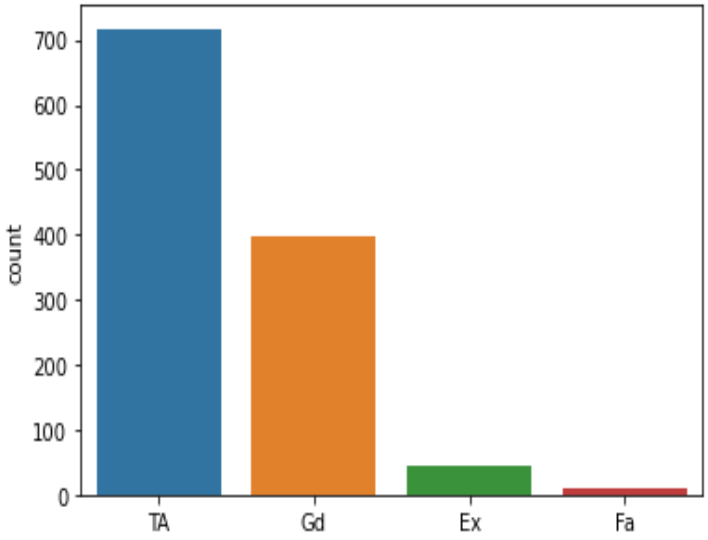
- 692 properties have lack masonries hence the area is 0

↓

The data is skewed and will be transformed later

ExterQual

Evaluates the quality of the material on the exterior

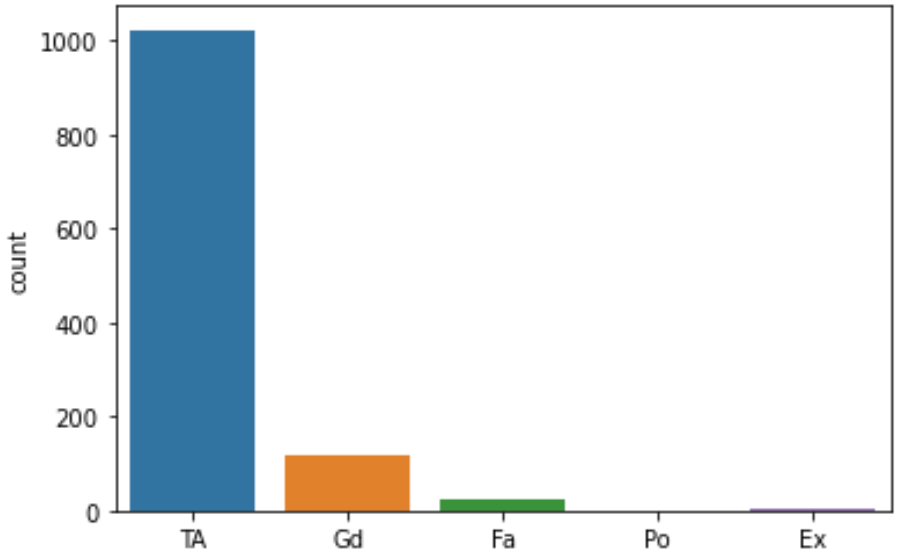


- 717 properties have TA (average/typical) exterior
- 397 properties have Gd (good) exterior
- 43 properties have Ex (Excellent) exterior
- 11 properties have Fa (fair) exterior

Encoding object data in numeric using Label Encoder

ExterCond

Evaluates the present condition of the material on the exterior

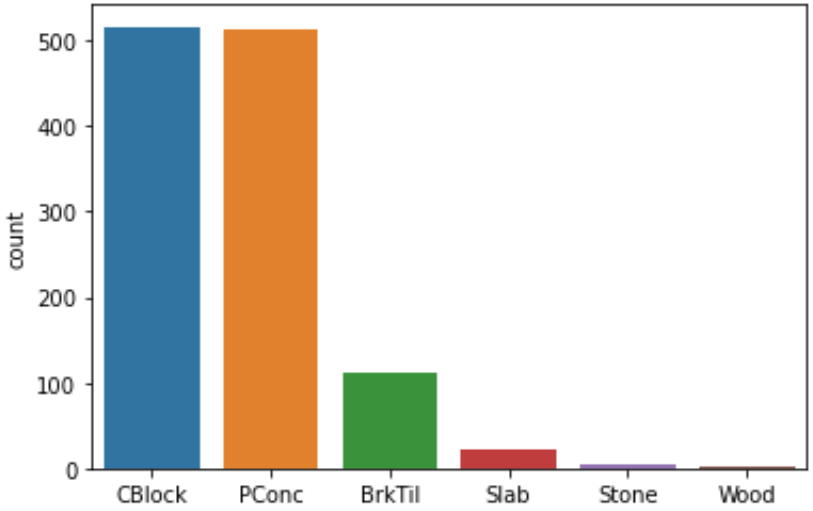


- 1022 properties have TA (average/typical) exterior
- 117 properties have Gd (good) exterior
- 26 properties have (fair) exterior
- 2 properties have Ex (excellent) exterior
- 1 property has Po (poor) exterior

Encoding object data in numeric using Label Encoder

Foundation

Type of foundation

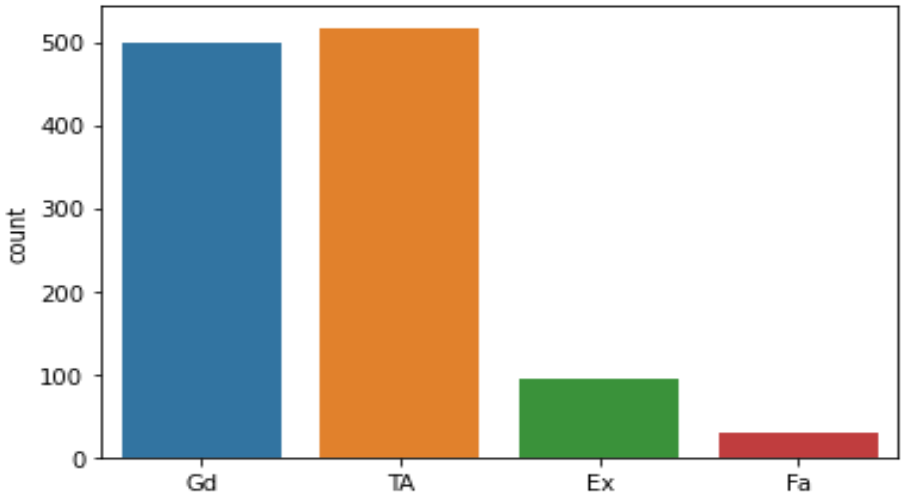


- 516 properties have Cblock (Cinder Block) foundation
- 513 properties have Pconc (Poured Contrete) foundation
- 112 properties have BrkTil (Brick & Tile) foundation
- 21 properties have Slab foundation
- 5 properties have Stone foundation
- 1 property has wood foundation

Encoding object data in numeric using Label Encoder

BsmtQual

Evaluates the height of the basement

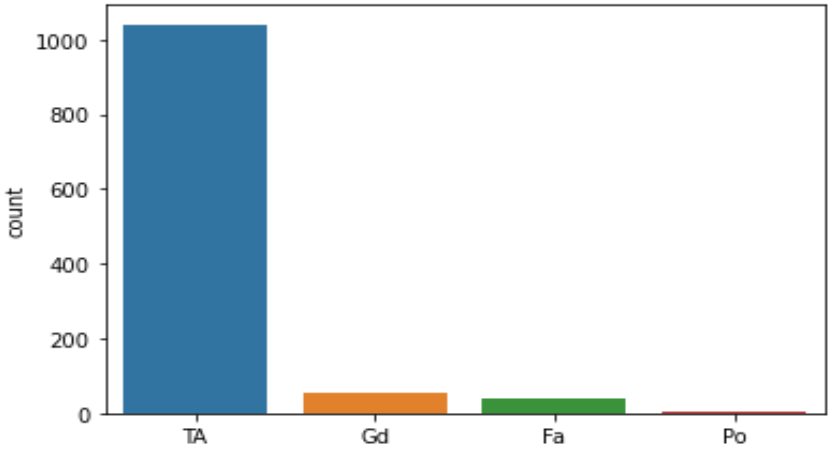


- 517 properties have TA (typical (80-89 inches)) basement height)
- 498 properties have Gd (good (90-99 inches) basement height)
- 94 properties have Ex (excellent (100+ inches) basement height)
- 29 properties have Fa (fair (70-79 inches) basement height)
- 30 properties have No basement

Encoding object data in numeric using Label Encoder

BsmtCond

Evaluates the general condition of the basement

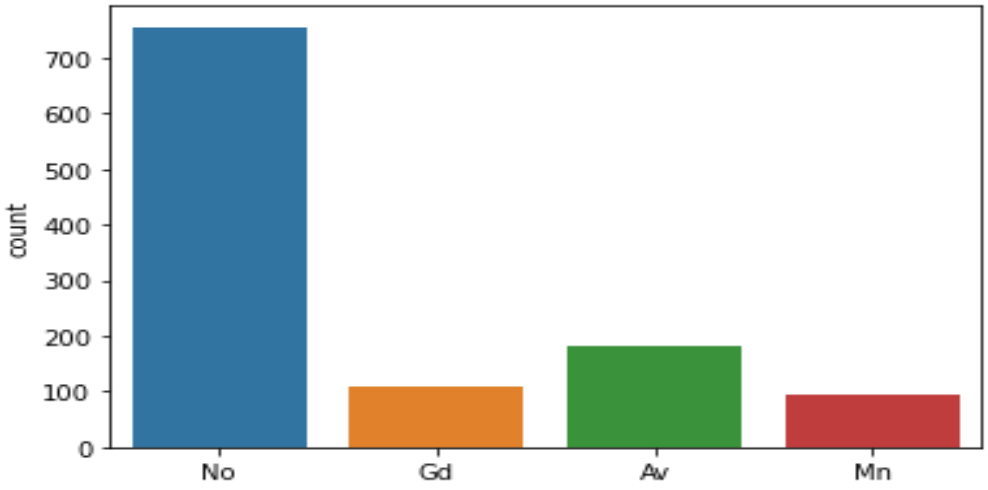


- 1041 properties have TA (Typical - slight dampness allowed) basement condition
- 56 properties have Gd (good) basement condition
- 39 properties have Fa (fair - dampness or some cracking or settling) basement condition
- 2 properties have Po (poor - severe cracking, settling, or wetness) basement condition
- 30 properties have No Basement

Encoding object data in numeric using Label Encoder

BsmtExposure

Refers to walkout or garden level walls

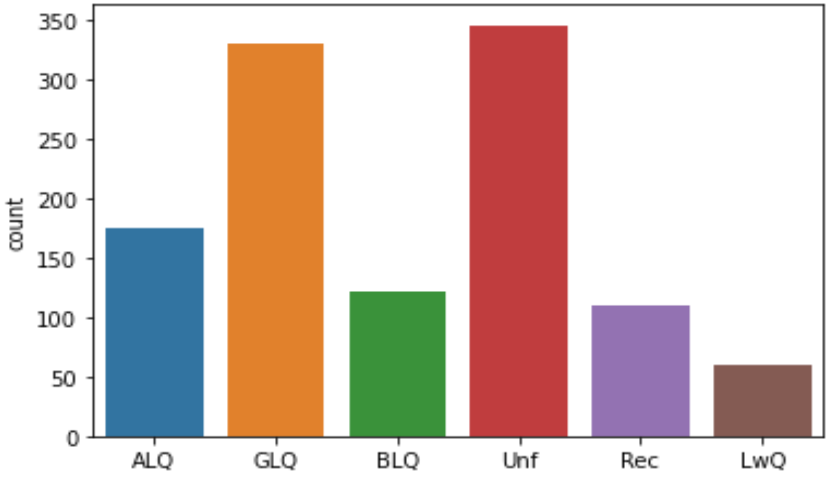


- No 756 properties No (do not have any exposure)
- Av 180 properties have Av (Average Exposure (split levels or foyers typically score average or above))
- Gd 108 properties have Gd (good) exposure
- Mn 93 properties have Mn (minimum) exposure
- 31 properties have No basement

Encoding object data in numeric using Label Encoder

BsmtFinType1

Rating of basement finished area



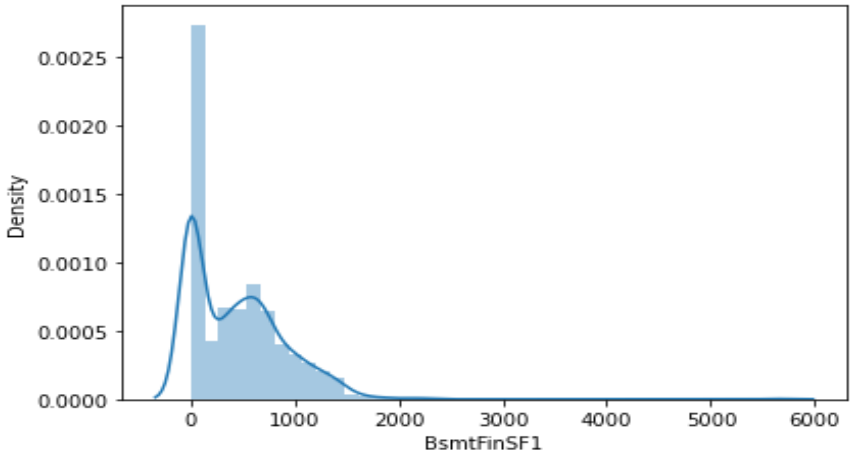
- Unf 345 properties have Unf (unfinished) basement
- GLQ 330 properties have GLQ (good living quarters) basement
- ALQ 174 properties have ALQ (average living quarters) basement
- BLQ 121 properties have BLQ (below average living quarters) basement
- Rec 109 properties have Rec (average rec room) basement
- LwQ 59 properties have LwQ (low quality) basement
- 30 properties have No basement

↓

Encoding object data in numeric using Label Encoder

BsmtFinSF1

Type 1 finished square feet



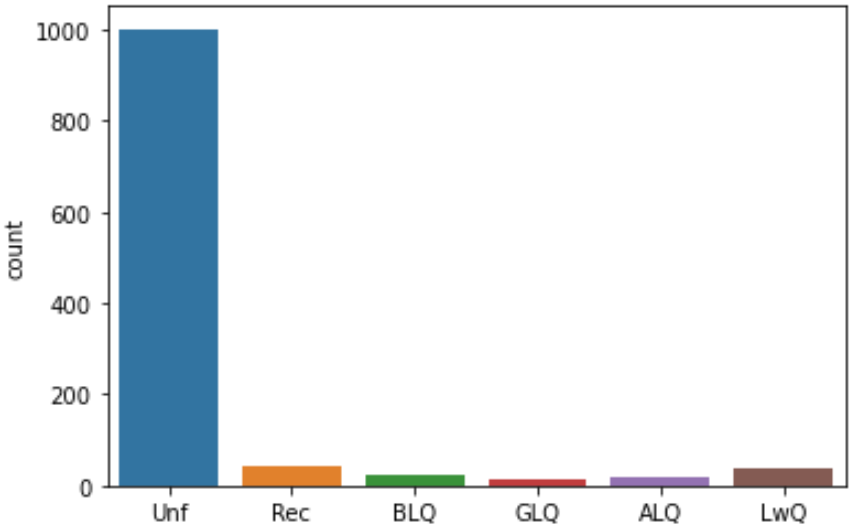
- 375 properties have no basements

↓

The data is skewed and will be transformed later

BsmtFinType2

Rating of basement finished area (if multiple types)



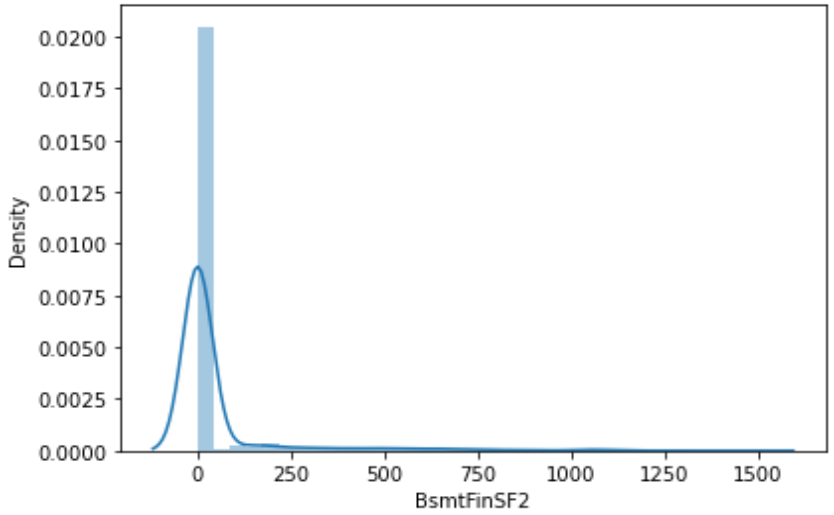
- Unf 1002 properties have Unf (unfinished) basements
- Rec 43 properties have Rec (Average Rec Room) basement
- LwQ 40 properties have LwQ (Low Quality) basement
- BLQ 24 properties have BLQ (Below Average Living Quarters) basements
- ALQ 16 properties have ALQ (Average Living Quarters)
- GLQ 12 properties have GLQ (Good Living Quarters)
- 31 properties have No basement

↓

Encoding object data in numeric using Label Encoder

BsmtFinSF2

Type 2 finished square feet



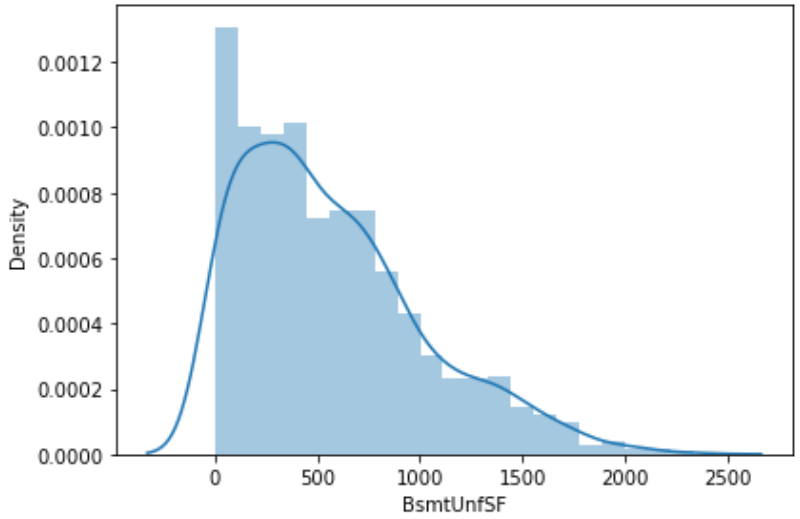
- 1032 have no basements or unfinished basements

↓

The data is skewed and will be transformed later

BsmtUnfSF

Unfinished square feet of basement area

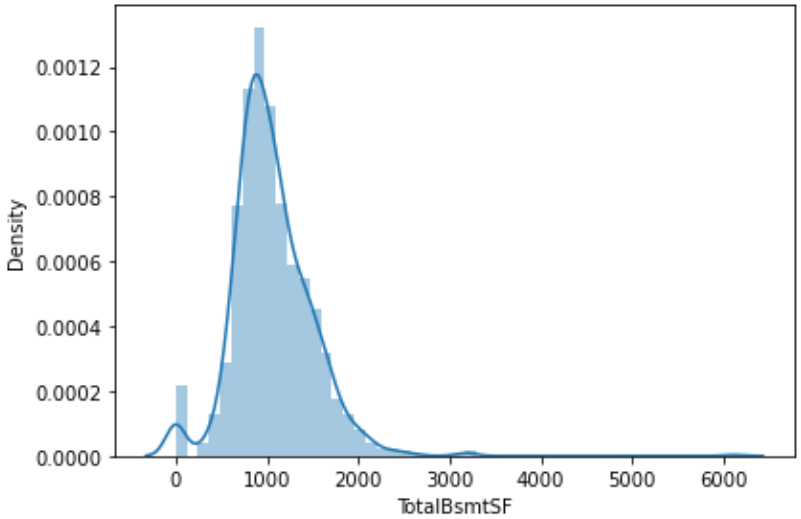


- 97 basements are finished, hence BsmtInfSF is 0

The data is slightly skewed and will be transformed later

TotalBsmtSF

Total square feet of basement area

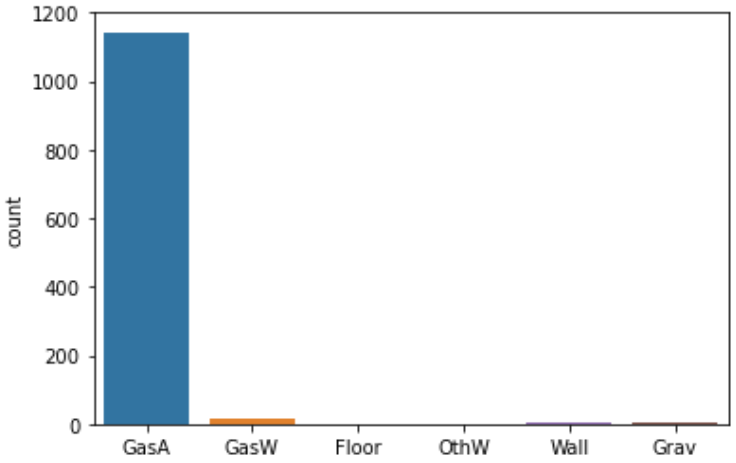


- 30 properties have no basements, hence TotalBsmtSf is 0

The data is skewed and will be transformed later

Heating

Type of heating



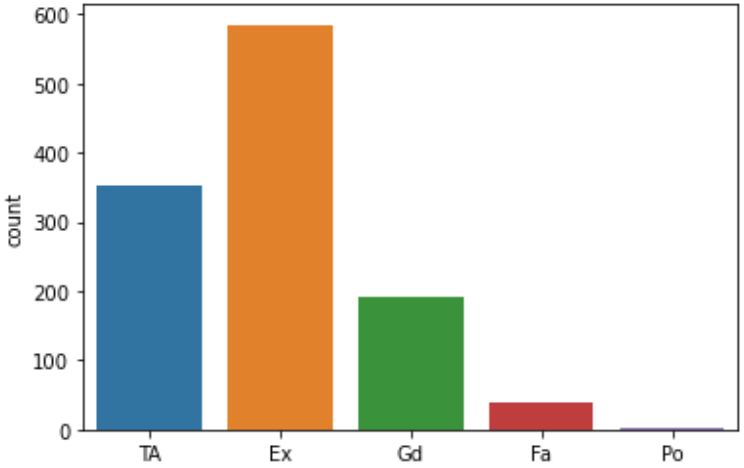
- 1143 properties have GasA (Gas forced warm air furnace) heating
- 14 properties have GasW (Gas hot water or steam heat) heating
- 5 properties have Grav (Gravity furnace) heating
- 4 properties have Wall furnace heating
- 1 property has Floor furnace heating
- 1 property has OthW (Hot water or steam heat other than gas)

↓

Encoding object data in numeric using Label Encoder

HeatingQC

Heating quality and condition

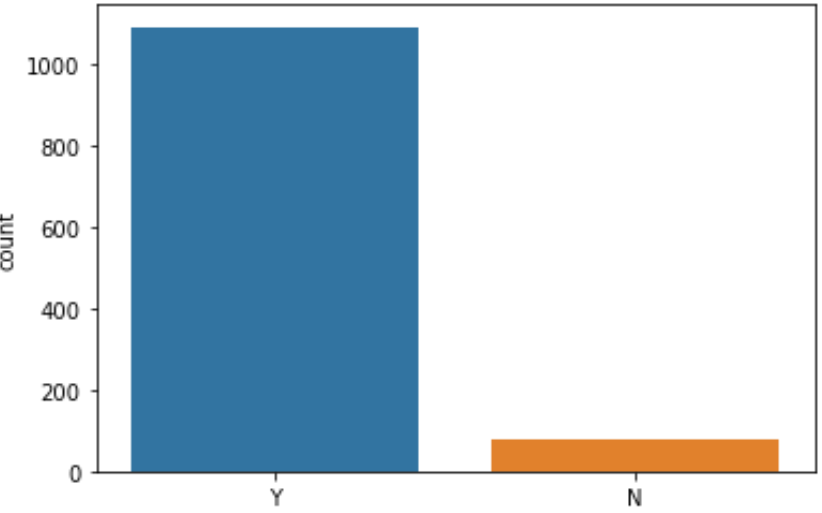


- 585 properties have Ex (Excellent) heating
- 352 properties have TA (Average/typical) heating
- 192 properties have Gd (good) heating
- 38 properties have Fa (Fair) heating
- 1 property has Po (Poor) heating

↓

Encoding object data in numeric using Label Encoder

CentralAir
Central air conditioning

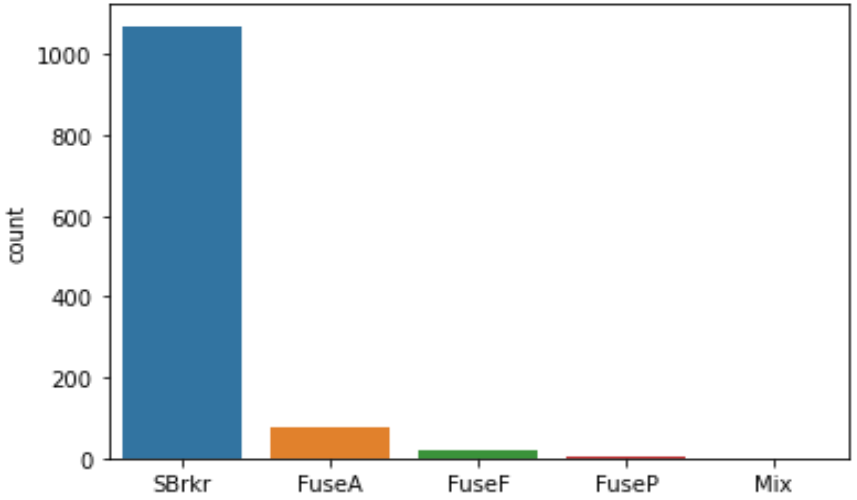


- 1090 properties have central air conditioning
- 78 properties do not have central air conditioning

↓

Encoding object data in numeric using Label Encoder

Electrical
Electrical system



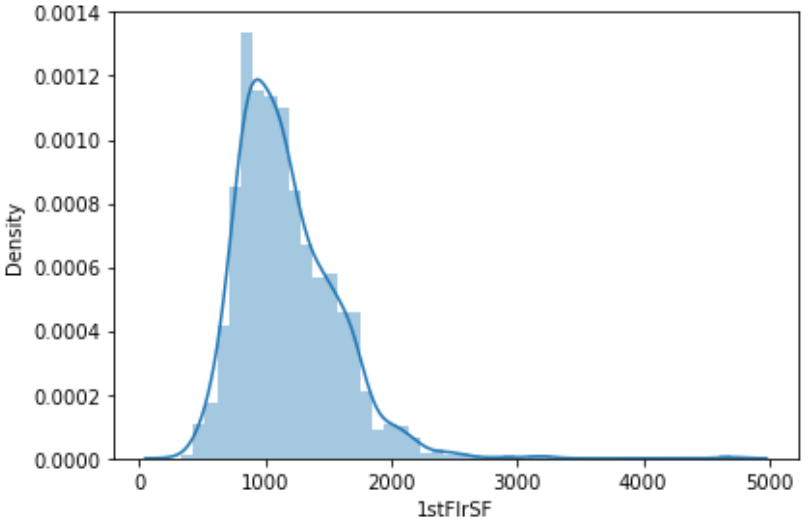
- 1070 properties have SBrkr(Standard Circuit Breakers & Romex) electrical system
- 74 properties have FuseA (Fuse Box over 60 AMP and all Romex wiring (Average)) electrical system
- 21 properties have FuseF (60 AMP Fuse Box and mostly Romex wiring (Fair)) electrical system
- 2 properties has FuseP (60 AMP Fuse Box and mostly knob & tube wiring (poor)) electrical system
- 1 property has Mixed electrical system

↓

Encoding object data in numeric using Label Encoder

1stFlrSF

First Floor square feet

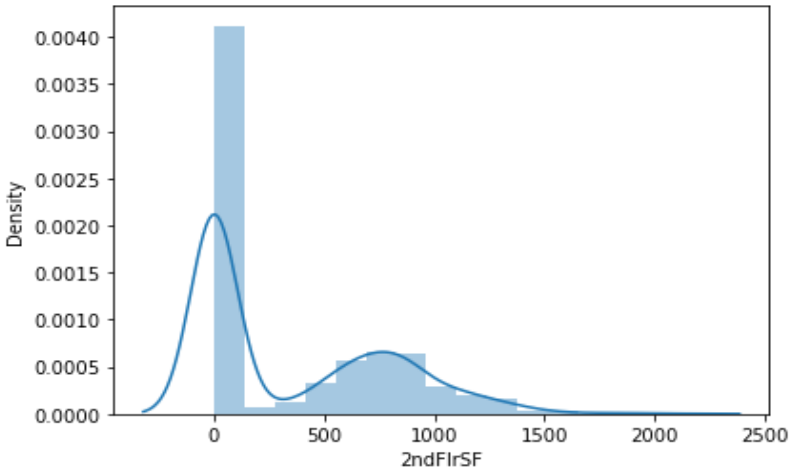


- 19 properties have 864 sq ft for first floor

The data is slightly skewed and will be transformed later

2ndFlrSF

Second floor square feet

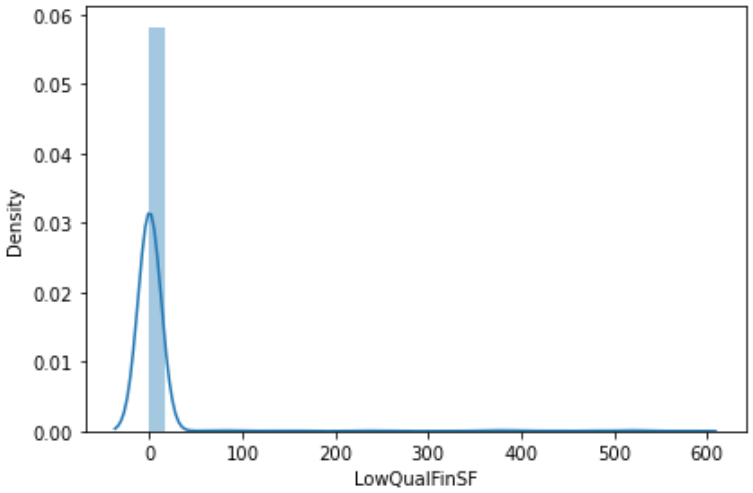


- 663 properties do not have second floor, hence 2ndFlrSF is 0

The data is skewed and will be transformed later

LowQualFinSF

Low quality finished square feet (all floors)



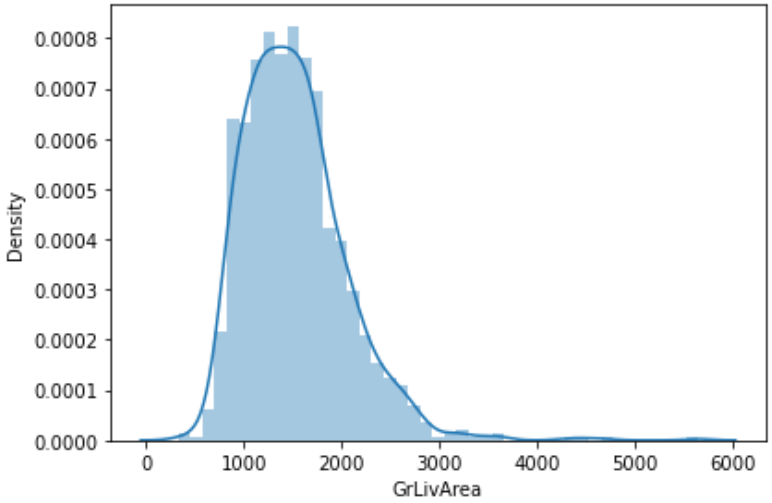
- 1145 properties do not have any low quality finished square feet

↓

The data is skewed and will be transformed later

GrLivArea

Above grade (ground) living area square feet



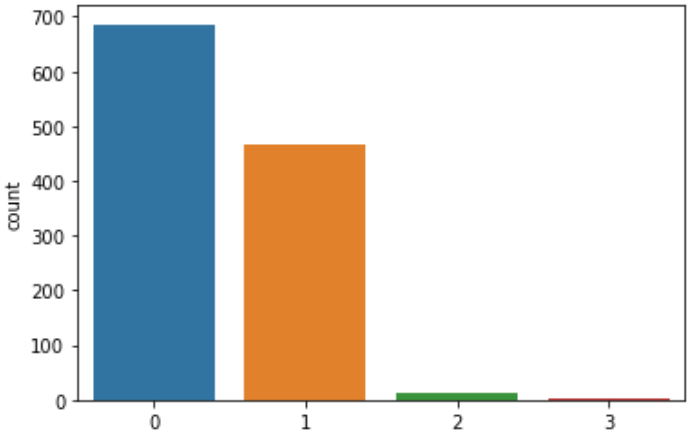
- 18 properties have 864 sq ft above grade (ground) living area

↓

The data is skewed and will be transformed later

BsmtFullBath

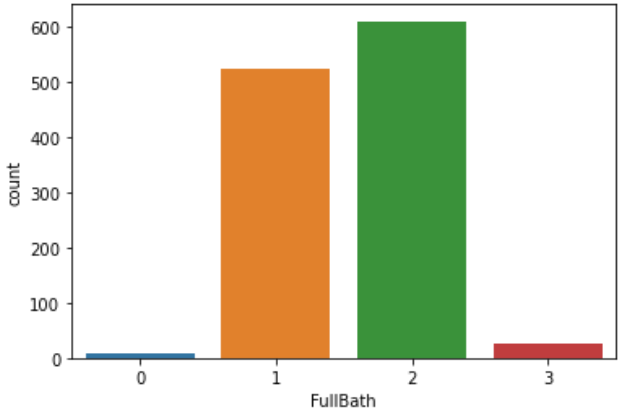
Basement full bathrooms



- 686 properties have no basement full bathrooms
- 468 properties have 1 basement full bathroom
- 13 properties have 2 basement full bathrooms
- 1 property has 3 basement full bathrooms

FullBath

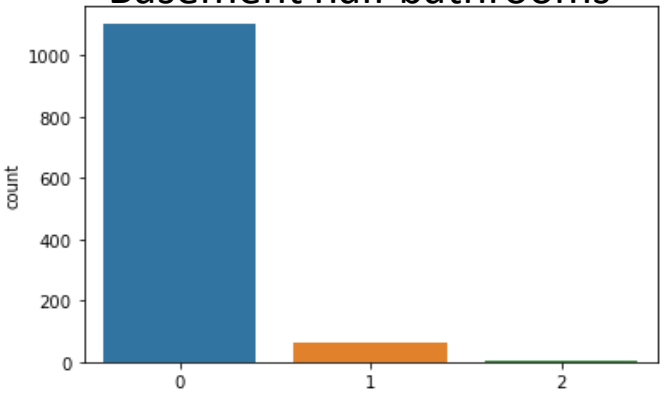
Full bathrooms above grade



- 610 properties have 2 full bathrooms above grade
- 524 properties have 1 full bathrooms above grade
- 27 properties have 3 full bathrooms above grade
- 7 properties have no full bathrooms above grade

BsmtHalfBath

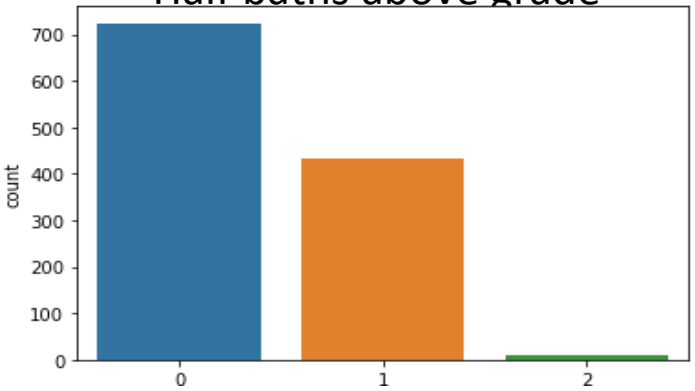
Basement half bathrooms



- 1105 properties have no basement half bathrooms
- 61 properties have 1 basement full bathroom
- 2 properties have 2 basement full bathrooms

HalfBath

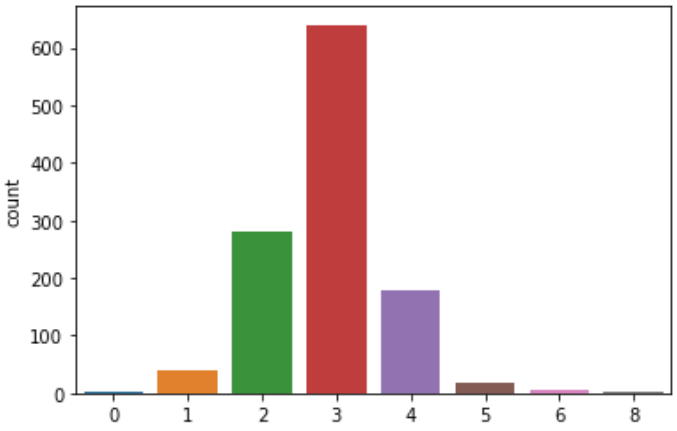
Half baths above grade



- 724 properties have no full bathrooms above grade
- 434 properties have 1full bathroom above grade
- 10 properties have 2 full bathrooms above grade

BedroomAbvGr

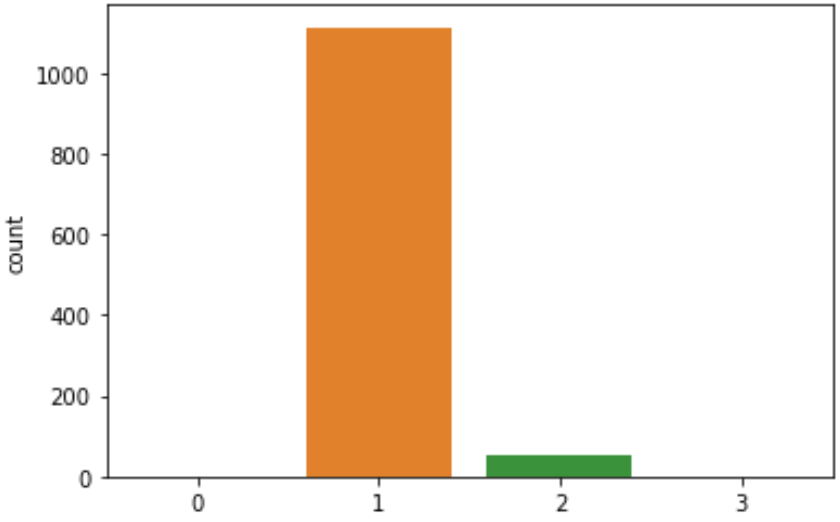
Bedrooms above grade (does NOT include basement bedrooms)



- 640 properties have 3 bedrooms above grade
- 281 properties have 2 bedrooms above grade
- 180 properties have 4 bedrooms above grade
- 39 properties have 1 bedroom above grade
- 18 properties have 5 bedrooms above grade
- 5 properties have 6 bedrooms above grade
- 4 properties have no bedrooms above grade
- 1 property has 8 bedrooms above grade

KitchenAbvGr

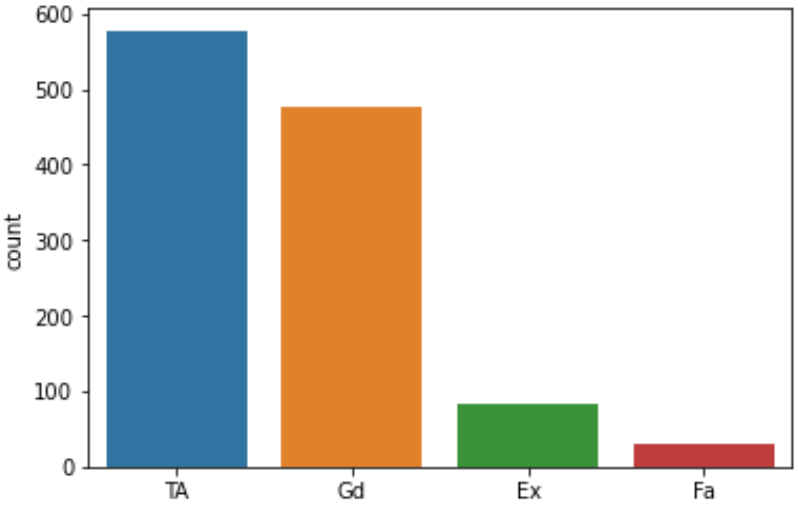
Kitchens above grade



- 1114 properties have 1 kitchen above grade
- 52 properties have 2 kitchen above grade
- 1 property has 3 kitchen above grade
- 1 properties has no kitchen above grade

KitchenQual

Kitchen quality



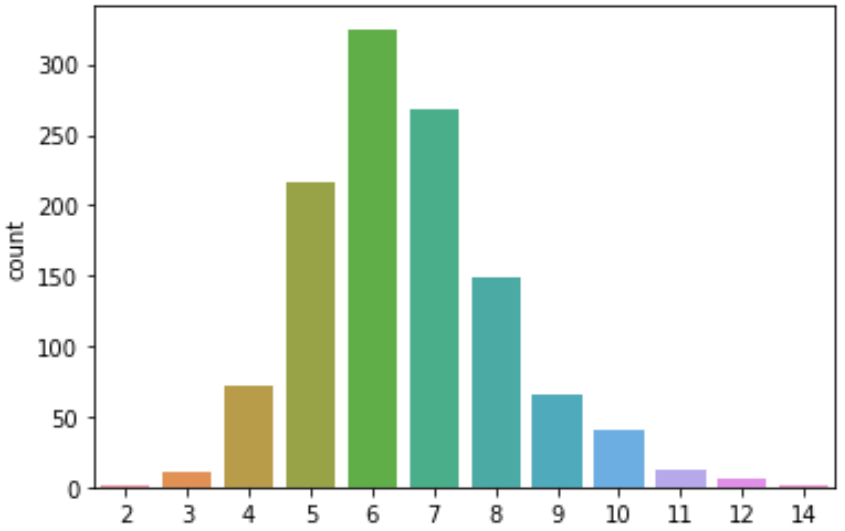
- 578 properties have TA (typical) kitchen
- 478 properties have Gd (good) kitchen
- 82 properties have Ex (excellent) kitchen
- 30 properties have Fa (fair) kitchen

↓

Encoding object data in numeric using Label Encoder

TotRmsAbvGrd

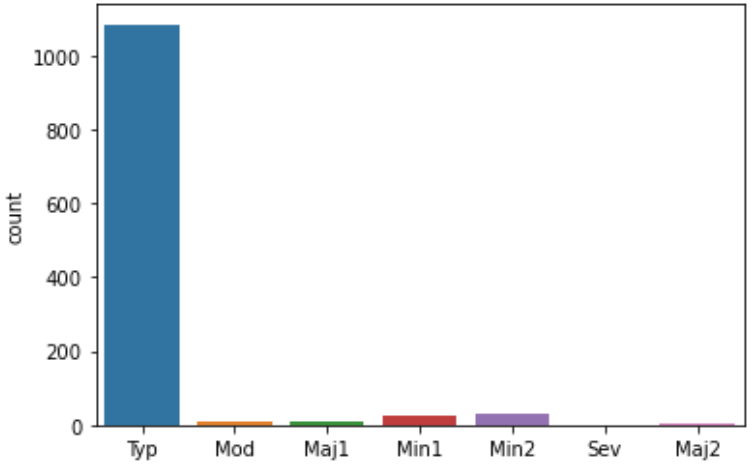
Total rooms above grade (does not include bathrooms)



- 325 properties have total 6 rooms above grade
- 268 properties have total 7 rooms above grade
- 217 properties have total 5 rooms above grade
- 148 properties have total 8 rooms above grade
- 72 properties have total 4 rooms above grade
- 65 properties have total 9 rooms above grade
- 41 properties have total 10 rooms above grade
- 13 properties have total 11 rooms above grade
- 11 properties have total 3 rooms above grade
- 6 properties have total 12 rooms above grade
- 1 property has total 2 rooms above grade
- 1 property has total 14 rooms above grade

Functional

Home functionality (Assume typical unless deductions are warranted)



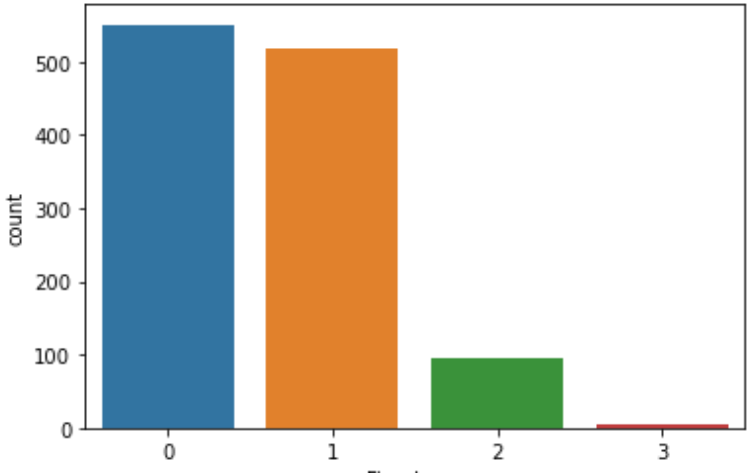
- 1085 properties have Typ (Typical Functionality)
- 30 properties have Min2 (Minor Deductions 2)
- 25 properties have Min1 (Minor Deductions 1)
- 12 properties have Mod (Moderate Deductions)
- 11 properties have Maj1 (Major Deductions 1)
- 4 properties have Maj2 (Major Deductions 2)
- 1 property has Sev (Severely Damaged)

↓

Encoding object data in numeric using Label Encoder

Fireplaces

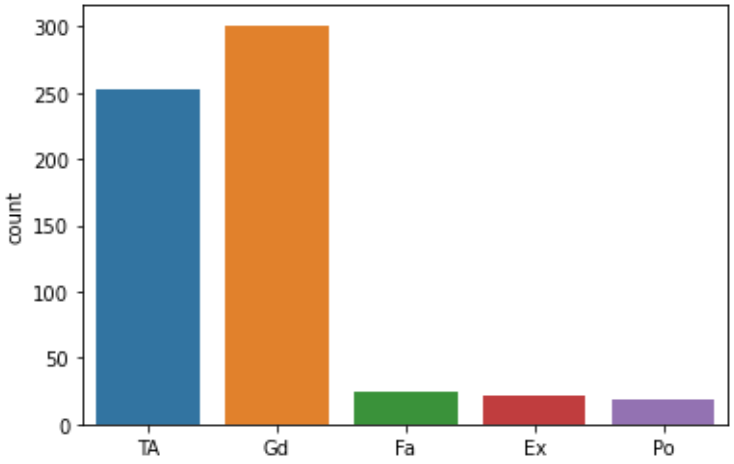
Number of fireplaces



- 551 properties have no fireplace
- 518 properties have 1 fireplace
- 94 properties have 2 fireplaces
- 5 properties have 2 fireplaces

FireplaceQu

Fireplace quality



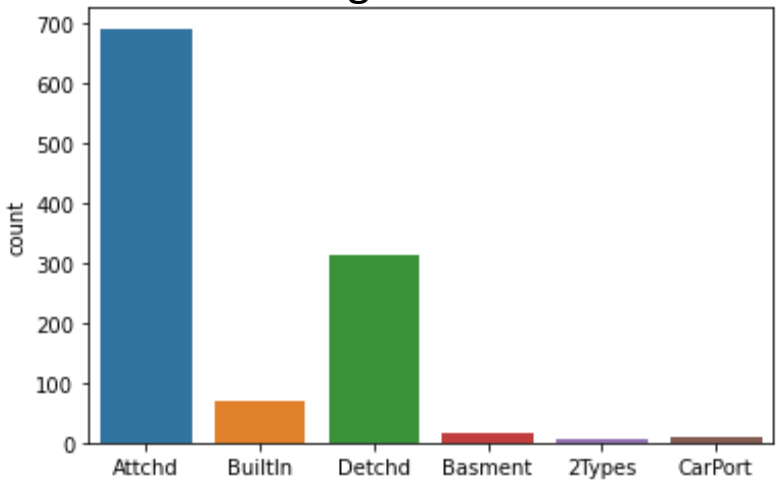
- 301 properties have Gd (good) fireplace
- 252 properties have TA (typical) fireplace
- 25 properties have Fa (fair) fireplace
- 21 properties have Ex (excellent) fireplace
- 18 properties have Po (poor) fireplace
- 551 properties have no fireplace

↓

Encoding object data in numeric using Label Encoder

GarageType

Garage location



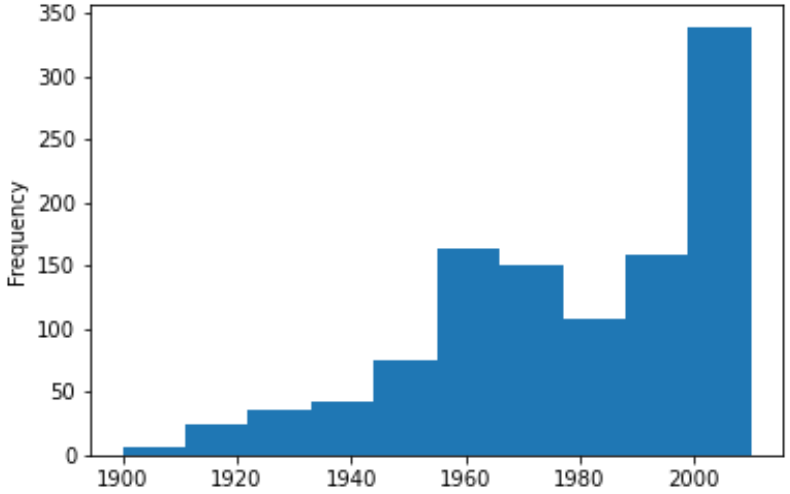
- 691 properties have Attchd (Attached to home) garage
- 314 properties have Detchd (Detached from home) garage
- 70 properties have BuiltIn (Built-In (Garage part of house - typically has room above garage)) garage
- 16 properties have Basment garage
- 8 properties have CarPort garage
- 5 properties have 2Types (More than one type of garage) garage
- 64 properties have no garage

↓

Encoding object data in numeric using Label Encoder

GarageYrBlt

Year garage was built



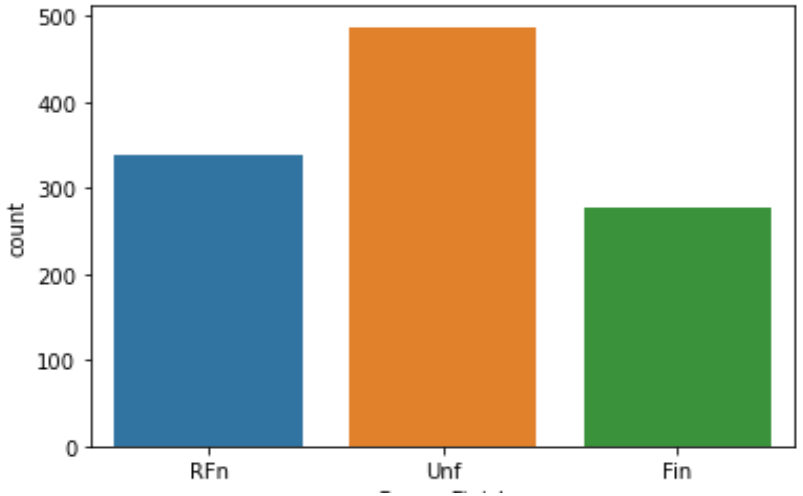
- Majority of the garages were built in the year 2000

↓

64 properties do not have garages, hence those null values are replaced by 0

GarageFinish

Interior finish of the garage



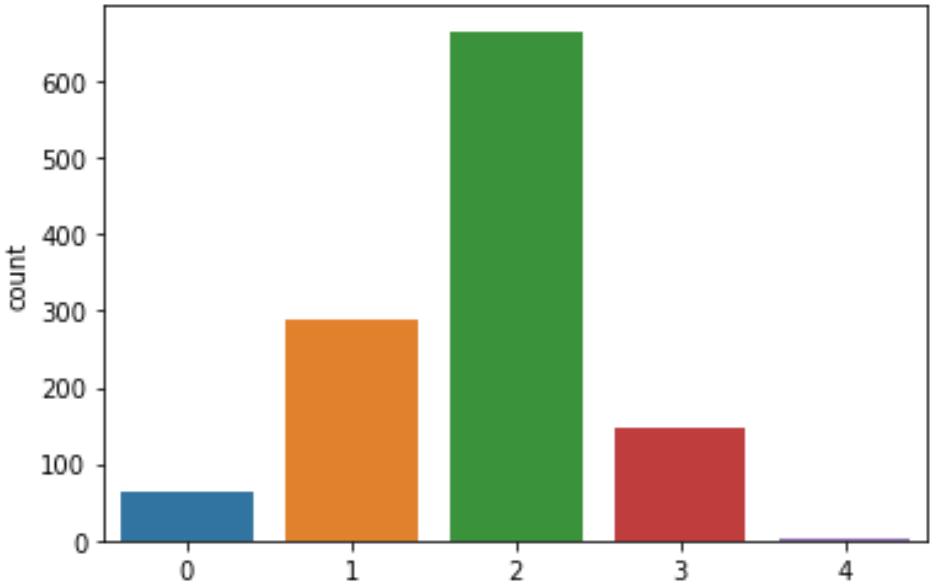
- 487 properties have Unf (Unfinished) garage
- 339 properties have RFn (Rough Finished) garage
- 278 properties have Fin (Finished) garage
- 64 properties have no garage

↓

Encoding object data in numeric using Label Encoder

GarageCars

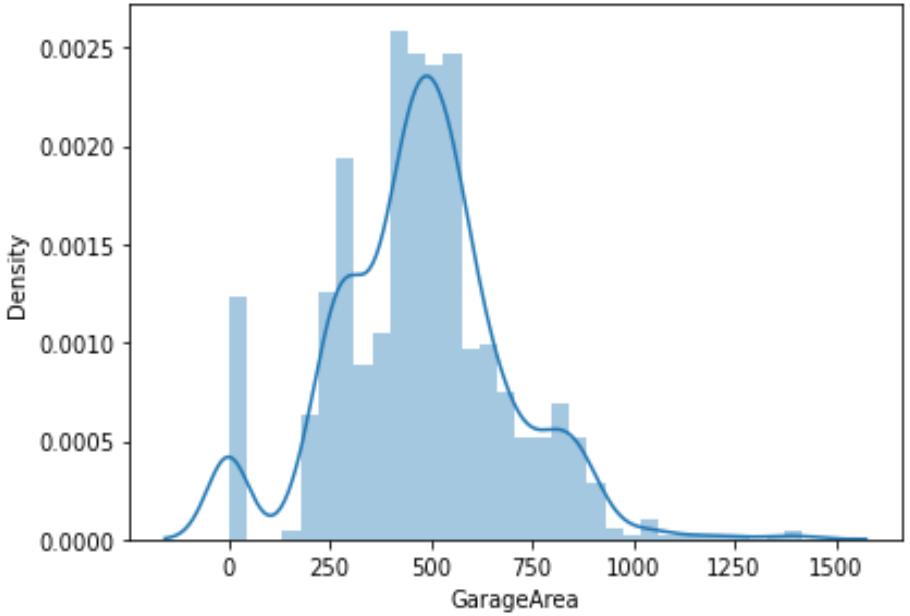
Size of garage in car capacity



- 665 properties have garage capacity for 2 cars
- 288 properties have garage capacity for 1 car
- 147 properties have garage capacity for 3 cars
- 64 properties have no garage capacity
- 4 properties have garage capacity for 4 cars

GarageArea

Size of garage in square feet



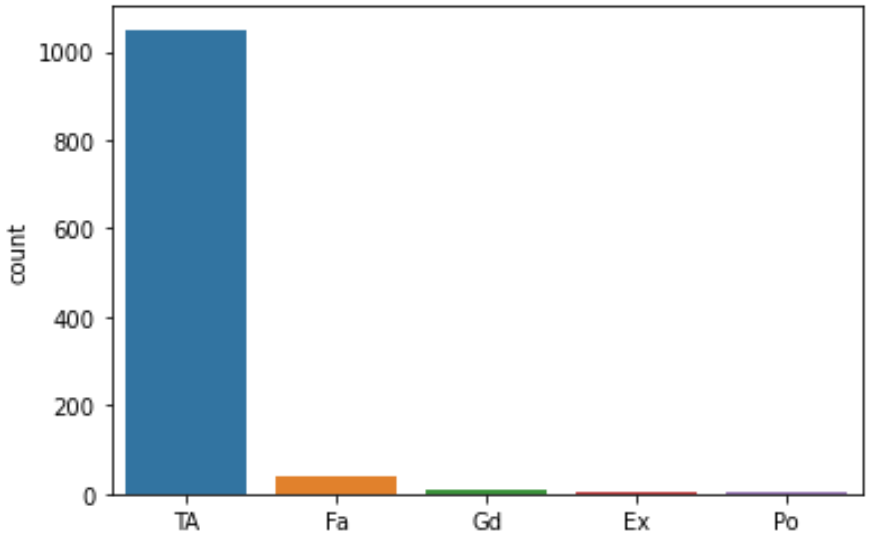
- 64 properties have no garage
- 44 properties have garage area of 440 sq ft

↓

The data is skewed and will be transformed later

GarageQual

Garage quality



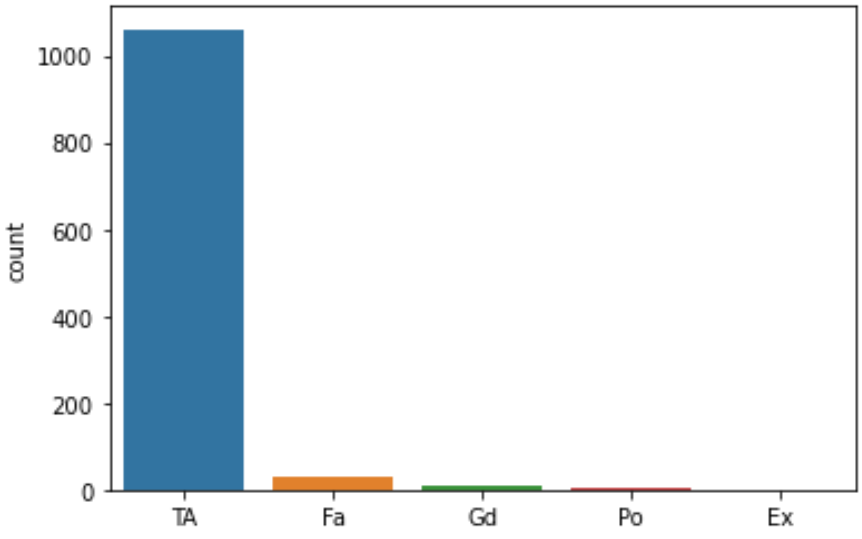
- 1050 properties have TA (typical) garage
- 39 properties have Fa (fair) garage
- 11 properties have Gd (good) garage
- 2 properties have Ex (excellent) garage
- 2 properties have Po (poor) garage
- 64 properties do not have garage

↓

Encoding object data in numeric using Label Encoder

GarageCond

Garage condition



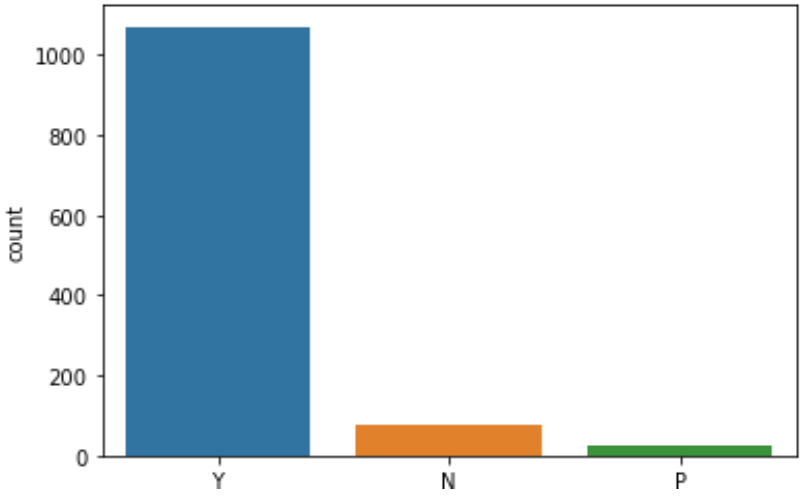
- 1061 properties have TA (typical) garage
- 28 properties have Fa (fair) garage
- 8 properties have Gd (good) garage
- 6 properties have Po (poor) garage
- 1 property has Ex(excellent) garage
- 64 properties have no garage

↓

Encoding object data in numeric using Label Encoder

PavedDrive

Paved driveway



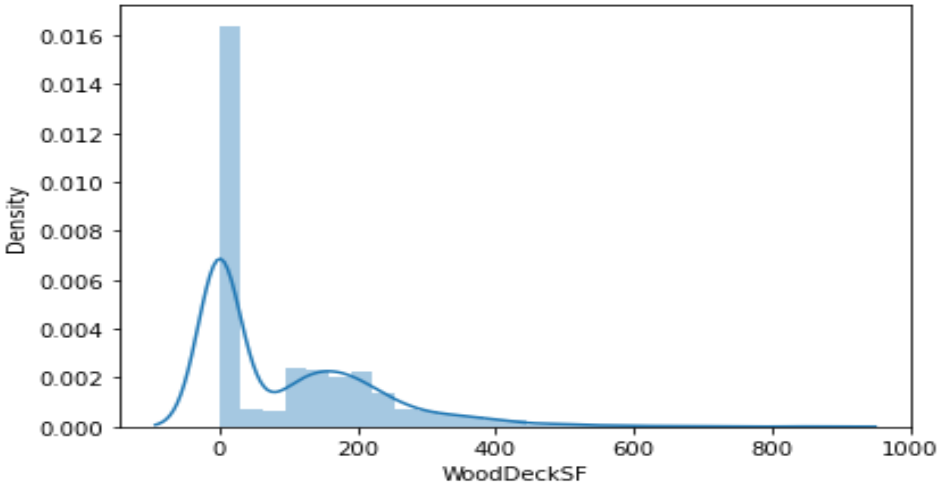
- Y 1071 properties have Y (paved) driveway
- N 74 properties have N (dirt/gravel) driveway
- P 23 properties (P) have partial pavement driveway

↓

Encoding object data in numeric using Label Encoder

WoodDeckSF

Wood deck area in square feet



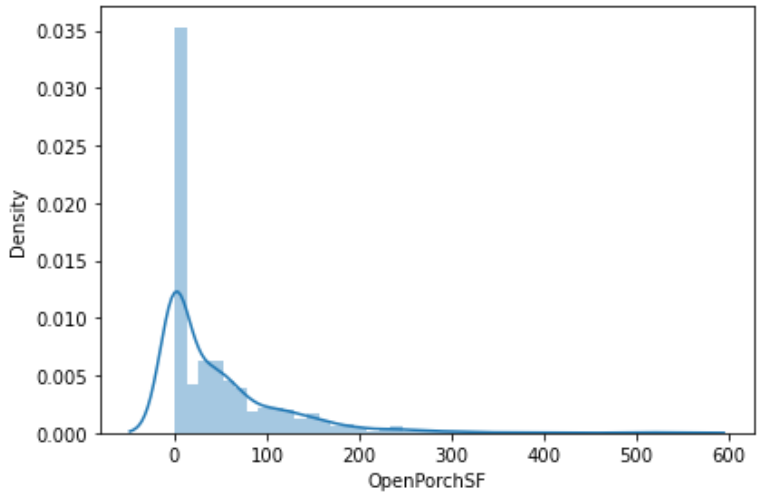
- 603 properties do not have wood deck, hence WoodDeskSF is 0

↓

The data is skewed and will be transformed later

OpenPorchSF

Open porch area in square feet

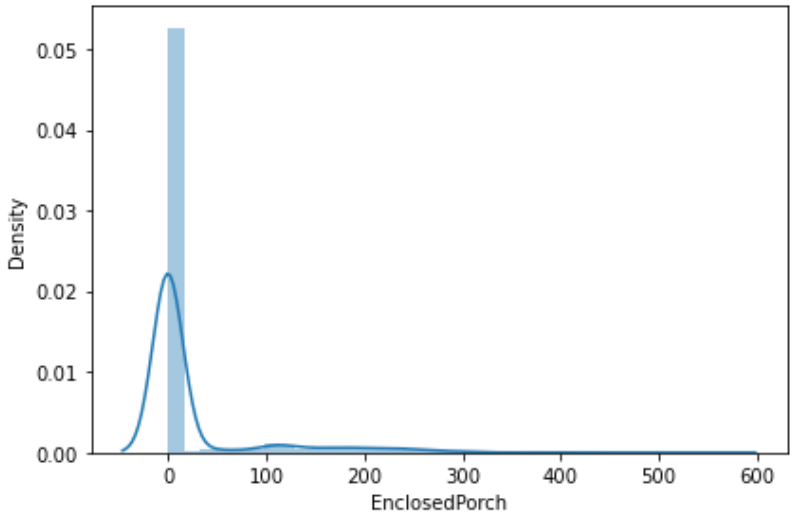


- 531 properties do not have any porch, hence OpenPorchSF is 0

The data is skewed and will be transformed later

EnclosedPorch

Enclosed porch area in square feet

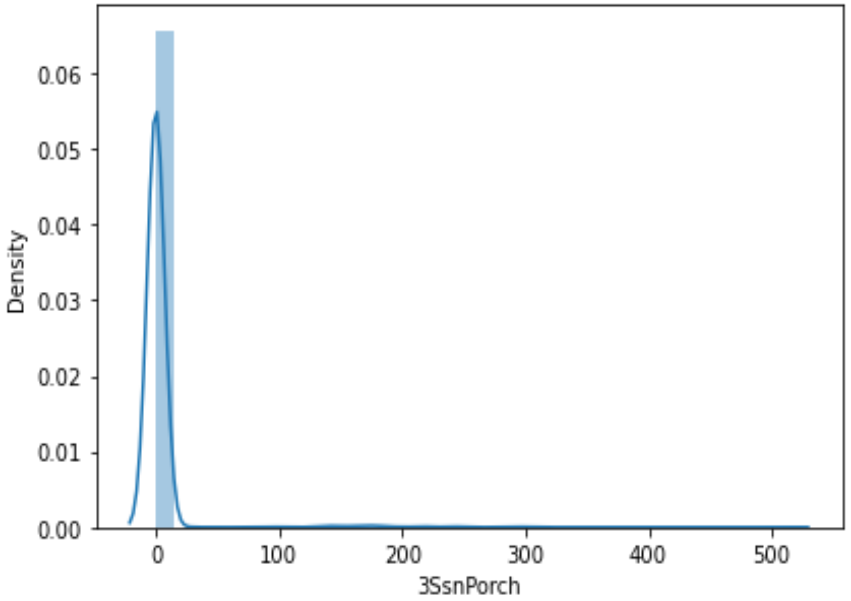


- 999 properties do not have enclosed porch, hence EnclosedPorch are is 0

The data is skewed and will be transformed later

3SsnPorch

Three season porch area in square feet



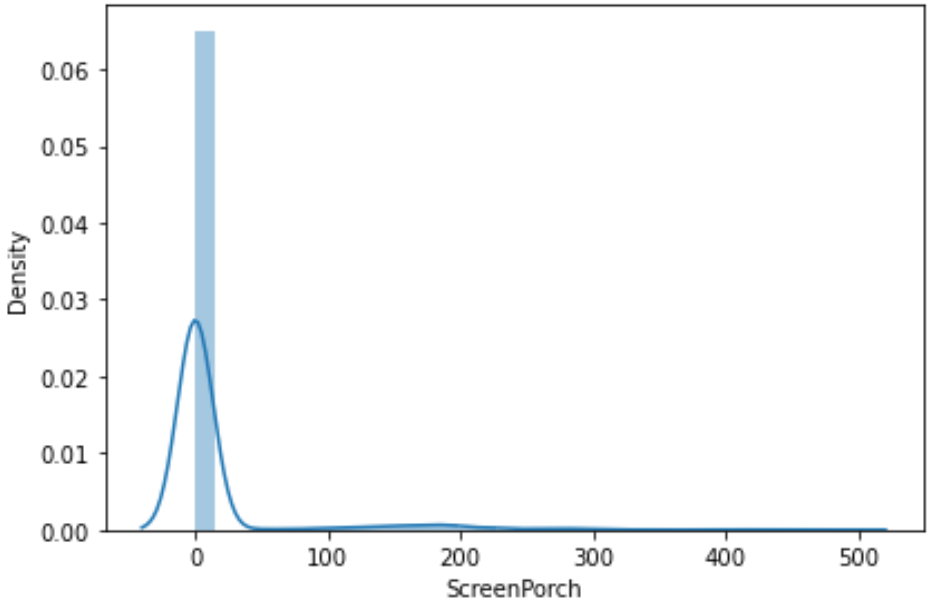
- 1146 properties do not have three season porch, hence 3SsnPorch is 0

↓

The data is skewed and will be transformed later

ScreenPorch

Screen porch area in square feet



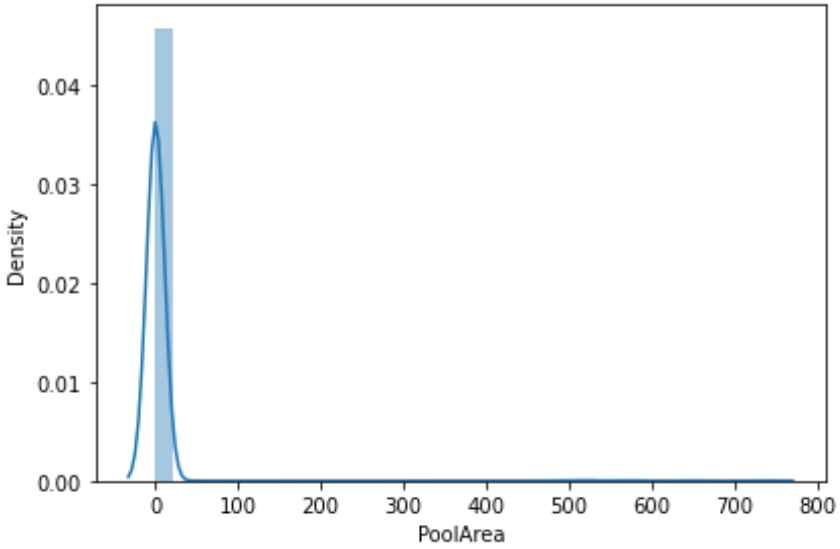
- 1073 properties do not have screenporch, hence ScreenPorch is 0

↓

The data is skewed and will be transformed later

PoolArea

Pool area in square feet

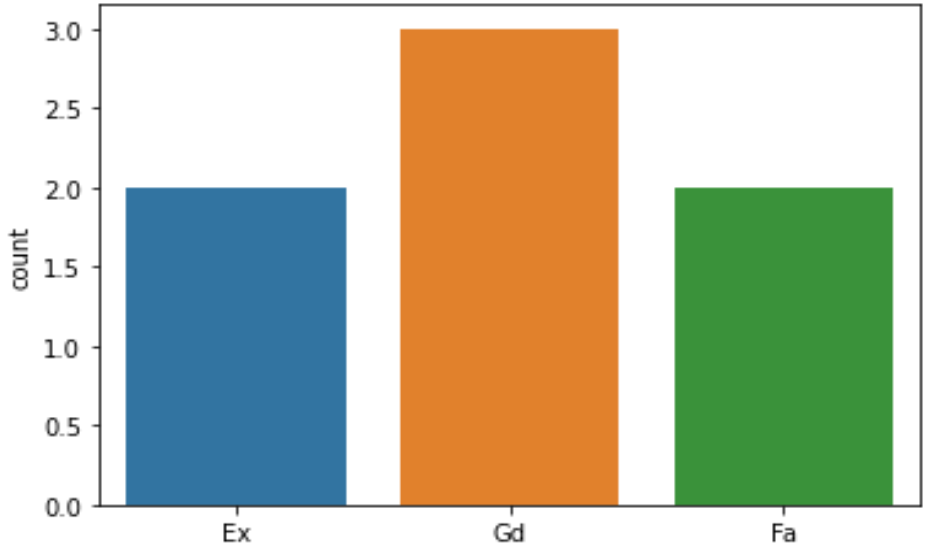


- 1161 properties do not have pool, hence PoolArea is 0

The data is skewed and will be transformed later

PoolQC

Pool quality

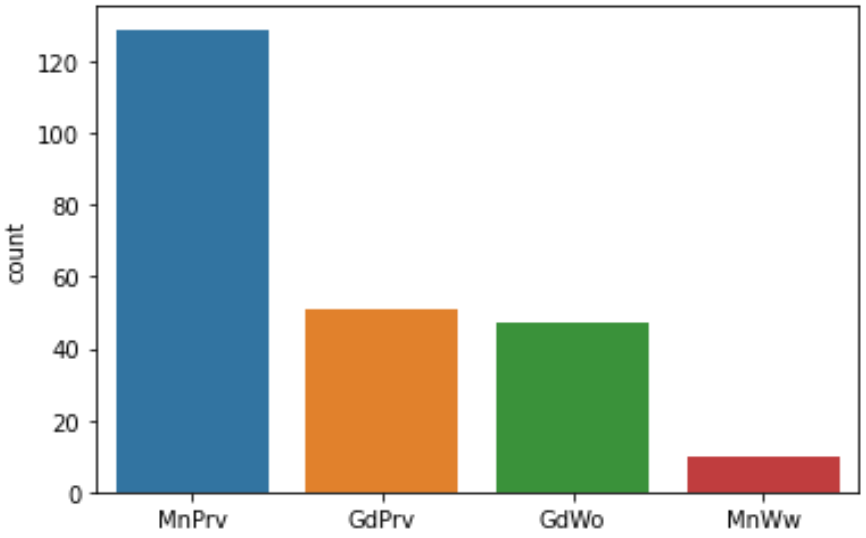


- 3 properties have Gd (Good) pool
- 2 properties have Ex(Excellent) pool
- 2 properties have Fa (fair) pool
- 1161 properties do not have a pool

Encoding object data in numeric using Label Encoder

Fence

Fence quality

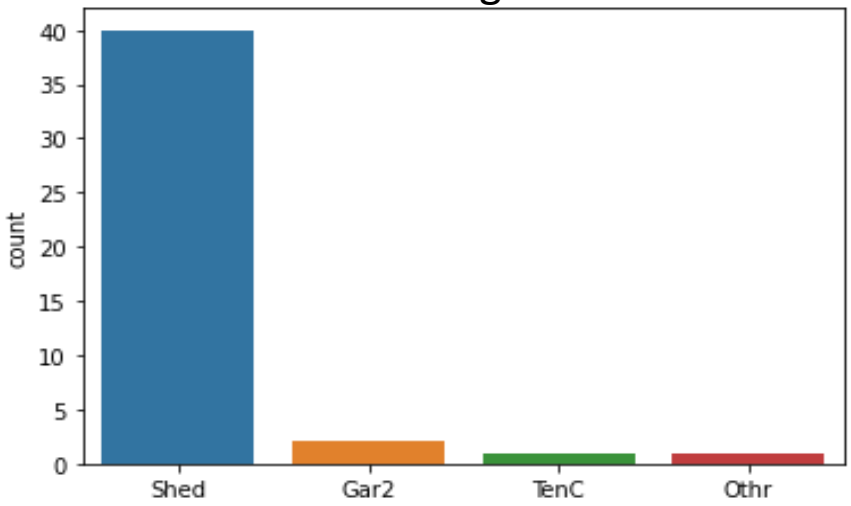


- 129 properties have MnPrv (Minimum Privacy) fence
- 51 properties have GdPrv (Good Privacy) fence
- 47 properties have GdWo (Good Wood) fence
- 10 properties have MnWw (Minimum Wood/Wire) fence
- 931 properties do not have any fence

Encoding object data in numeric using Label Encoder

MiscFeature

Miscellaneous feature not covered in other categories

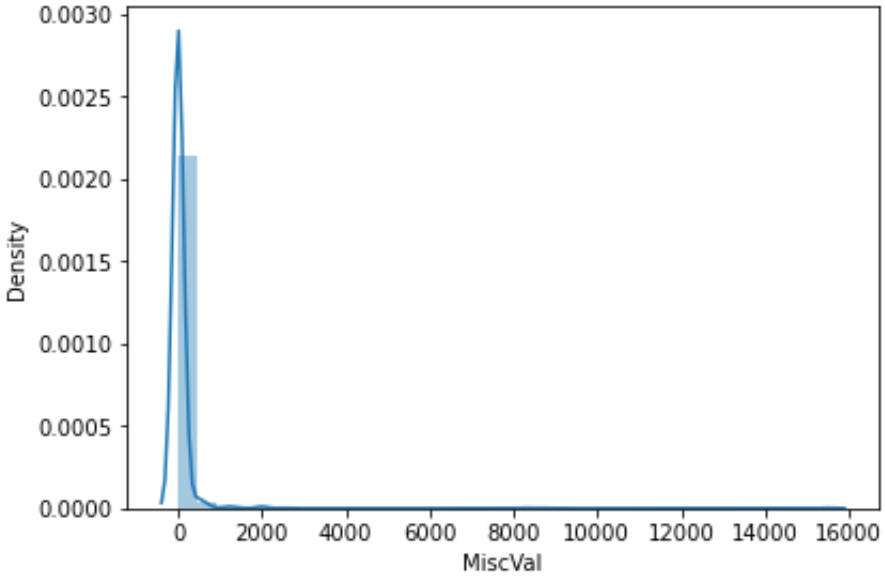


- 40 properties have Shed (over 100 SF)
- 2 properties have Gar2 (2nd Garage (if not described in garage section)
- 1 property has TenC (Tennis Court)
- 1 property has Othr
- 1124 properties do not have miscellaneous feature

Encoding object data in numeric using Label Encoder

MiscVal

$Value of miscellaneous feature



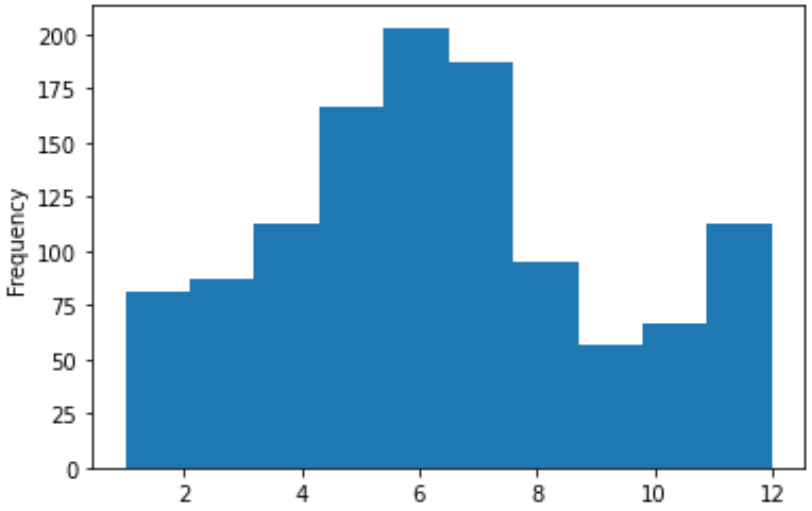
- 1126 properties do not have miscellaneous features hence MiscVal is 0

↓

The data is skewed and will be transformed later

MoSold

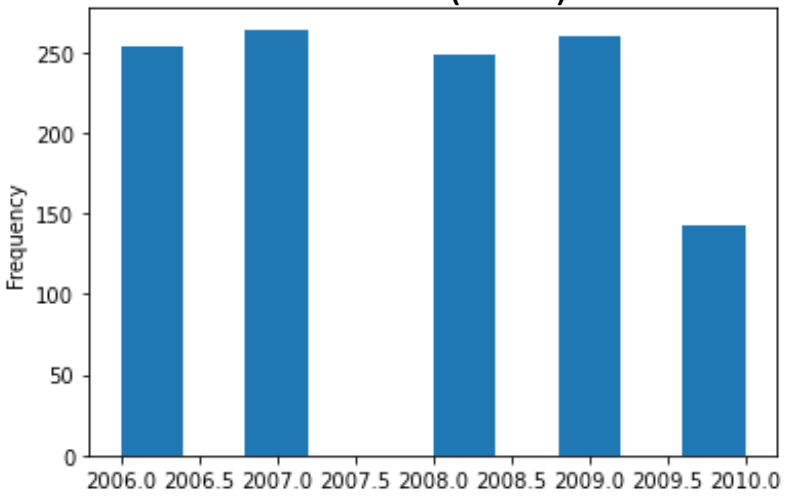
Month Sold (MM)



- Majority of the properties are sold in June

YrSold

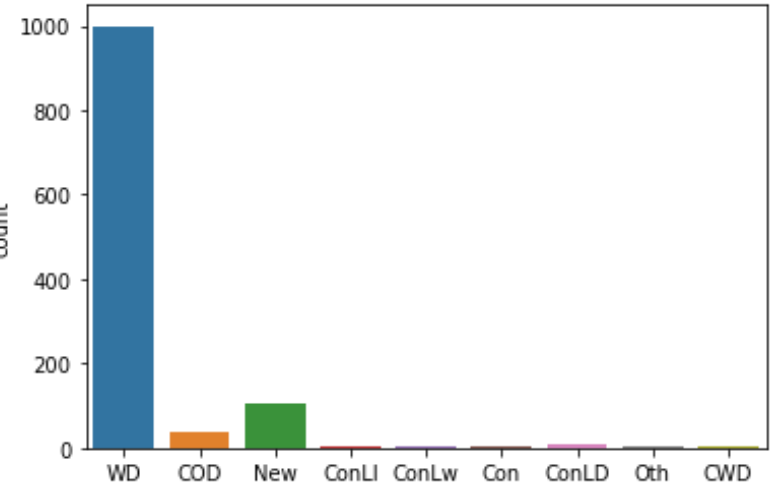
Year Sold (YYYY)



- Majority of the properties are sold between 2006-2009

SaleType

Type of sale



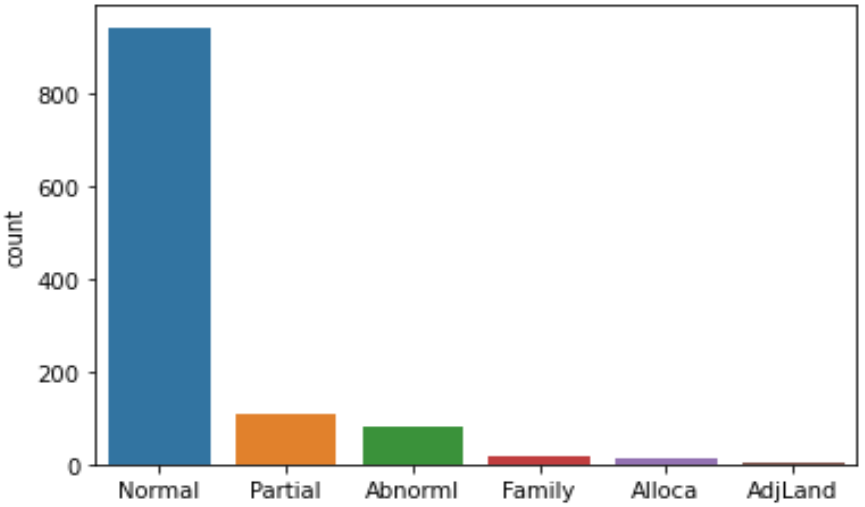
- 999 properties are WD (Warranty Deed – Conventional)
- 106 properties are New (Home just constructed and sold)
- 38 properties are COD (Court Officer Deed/Estate)
- 8 properties are ConLD (Contract Low Down)
- 5 properties are ConLI (Contract Low Interest)
- 4 properties are ConLw (Contract Low Down payment and low interest)
- 3 properties are other
- 3 properties are CWD (Warranty Deed – Cash)
- 2 properties are Con (Contract 15% Down payment regular terms)

↓

Encoding object data in numeric using Label Encoder

SaleCondition

Condition of sale



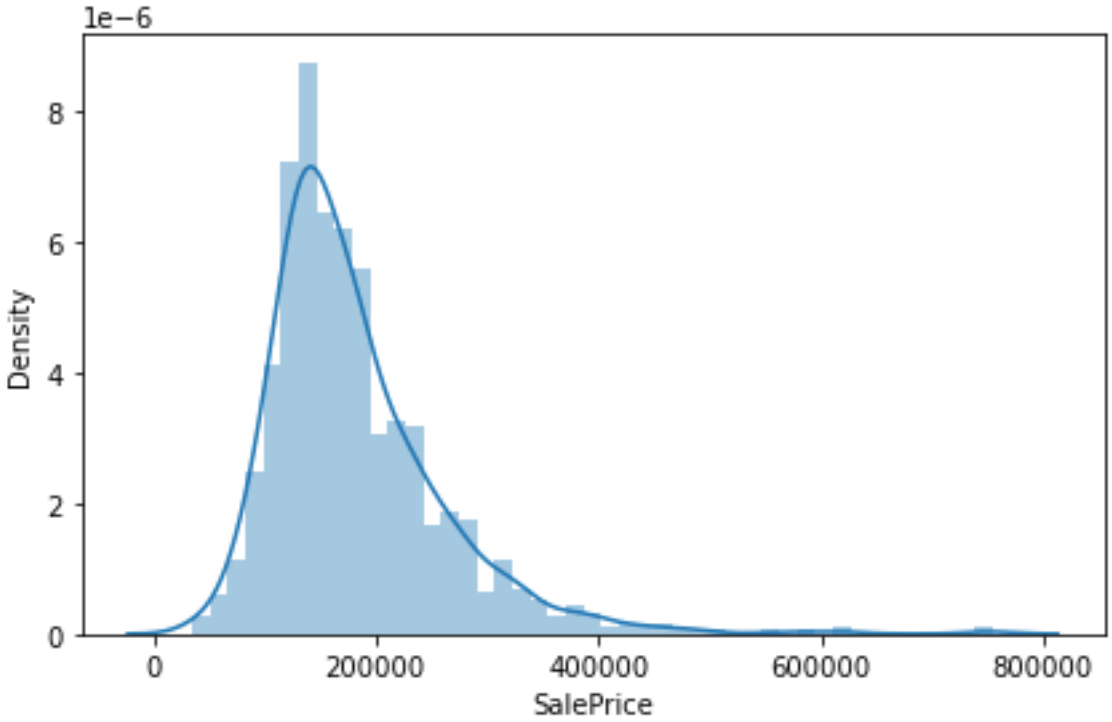
- 945 properties had Normal sale
- 108 properties had Partial (Home was not completed when last assessed (associated with New Homes)) sale
- 81 properties had Abnormal (trade, foreclosure, short sale) sale
- 18 properties had Family (Sale between family members) sale
- 12 properties had Alloca (Allocation - two linked properties with separate deeds, typically condo with a garage unit)
- 4 properties had AdjLand (Adjoining Land Purchase) sale

↓

Encoding object data in numeric using Label Encoder

LABEL

SalePrice



- 18 properties have a sale price of 140000

- The null values in 'LotFrontage‘ and 'MasVnrArea‘ are imputed using KNN Imputer.

- Statistical analysis using describe method-

MSSubClass

- count 1168.000000
- mean 56.767979
- std 41.940650
- min 20.000000
- 25% 20.000000
- 50% 50.000000
- 75% 70.000000
- max 190.000000

MSZoning

- count 1168.000000
- mean 3.013699
- std 0.633120
- min 0.000000
- 25% 3.000000
- 50% 3.000000
- 75% 3.000000
- max 4.000000

LotFrontage

- count 1168.000000
- mean 71.724315
- std 24.159328
- min 21.000000
- 25% 60.000000
- 50% 70.000000
- 75% 81.000000
- max 313.000000

LotArea

- count 1168.000000
- mean 10484.749144
- std 8957.442311
- min 1300.000000
- 25% 7621.500000
- 50% 9522.500000
- 75% 11515.500000
- max 164660.000000

Street

- count 1168.000000
- mean 0.996575
- std 0.058445
- min 0.000000
- 25% 1.000000
- 50% 1.000000
- 75% 1.000000
- max 1.000000

Alley

- count 1168.000000
- mean 1.898973
- std 0.401453
- min 0.000000
- 25% 2.000000
- 50% 2.000000
- 75% 2.000000
- max 2.000000

LotShape

- count 1168.000000
- mean 1.938356
- std 1.412262
- min 0.000000
- 25% 0.000000
- 50% 3.000000
- 75% 3.000000
- max 3.000000

LandContour

- count 1168.000000
- mean 2.773973
- std 0.710027
- min 0.000000
- 25% 3.000000
- 50% 3.000000
- 75% 3.000000
- max 3.000000

LotConfig

- count 1168.000000
- mean 3.004281
- std 1.642667
- min 0.000000
- 25% 2.000000
- 50% 4.000000
- 75% 4.000000
- max 4.000000

LandSlope

- count 1168.000000
- mean 0.064212
- std 0.284088
- min 0.000000
- 25% 0.000000
- 50% 0.000000
- 75% 0.000000
- max 2.000000

Neighborhood

- count 1168.000000
- mean 12.145548
- std 6.010364
- min 0.000000
- 25% 7.000000
- 50% 12.000000
- 75% 17.000000
- max 24.000000

Condition1

- count 1168.000000
- mean 2.032534
- std 0.871703
- min 0.000000
- 25% 2.000000
- 50% 2.000000
- 75% 2.000000
- max 8.000000

Condition2

- count 1168.000000
- mean 2.005993
- std 0.250035
- min 0.000000
- 25% 2.000000
- 50% 2.000000
- 75% 2.000000
- max 7.000000

BldgType

- count 1168.000000
- mean 0.476027
- std 1.180870
- min 0.000000
- 25% 0.000000
- 50% 0.000000
- 75% 0.000000
- max 4.000000

HouseStyle

- count 1168.000000
- mean 3.043664
- std 1.898625
- min 0.000000
- 25% 2.000000
- 50% 2.000000
- 75% 5.000000
- max 7.000000

OverallQual

- count 1168.000000
- mean 6.104452
- std 1.390153
- min 1.000000
- 25% 5.000000
- 50% 6.000000
- 75% 7.000000
- max 10.000000

OverallCond

- count 1168.000000
- mean 5.595890
- std 1.124343
- min 1.000000
- 25% 5.000000
- 50% 5.000000
- 75% 6.000000
- max 9.000000

YearBuilt

- count 1168.000000
- mean 1970.930651
- std 30.145255
- min 1875.000000
- 25% 1954.000000
- 50% 1972.000000
- 75% 2000.000000
- max 2010.000000

YearRemodAdd

- count 1168.000000
- mean 1984.758562
- std 20.785185
- min 1950.000000
- 25% 1966.000000
- 50% 1993.000000
- 75% 2004.000000
- max 2010.000000

RoofStyle

- count 1168.000000
- mean 1.402397
- std 0.832539
- min 0.000000
- 25% 1.000000
- 50% 1.000000
- 75% 1.000000
- max 5.000000

RoofMatl

- count 1168.000000
- mean 1.086473
- std 0.642848
- min 0.000000
- 25% 1.000000
- 50% 1.000000
- 75% 1.000000
- max 7.000000

Exterior1st

- count 1168.000000
- mean 8.659247
- std 3.097443
- min 0.000000
- 25% 7.000000
- 50% 11.000000
- 75% 11.000000
- max 13.000000

Exterior2nd

- count 1168.000000
- mean 9.363014
- std 3.462380
- min 0.000000
- 25% 7.000000
- 50% 11.000000
- 75% 12.000000
- max 14.000000

MasVnrType

- count 1168.000000
- mean 1.758562
- std 0.611174
- min 0.000000
- 25% 1.000000
- 50% 2.000000
- 75% 2.000000
- max 3.000000

MasVnrArea

- count 1168.000000
- mean 102.576199
- std 182.350310
- min 0.000000
- 25% 0.000000
- 50% 0.000000
- 75% 160.000000
- max 1600.000000

ExterQual

- count 1168.000000
- mean 2.530822
- std 0.699425
- min 0.000000
- 25% 2.000000
- 50% 3.000000
- 75% 3.000000
- max 3.000000

ExterCond

- count 1168.000000
- mean 3.725171
- std 0.744463
- min 0.000000
- 25% 4.000000
- 50% 4.000000
- 75% 4.000000
- max 4.000000

Foundation

- count 1168.000000
- mean 1.395548
- std 0.709379
- min 0.000000
- 25% 1.000000
- 50% 1.000000
- 75% 2.000000
- max 5.000000

BsmtQual

- count 1168.000000
- mean 2.308219
- std 0.893201
- min 0.000000
- 25% 2.000000
- 50% 2.000000
- 75% 3.000000
- max 4.000000

BsmtCond

- count 1168.000000
- mean 2.827911
- std 0.700355
- min 0.000000
- 25% 3.000000
- 50% 3.000000
- 75% 3.000000
- max 4.000000

BsmtExposure

- count 1168.000000
- mean 2.299658
- std 1.172054
- min 0.000000
- 25% 2.000000
- 50% 3.000000
- 75% 3.000000
- max 4.000000

BsmtFinType1

- count 1168.000000
- mean 2.824486
- std 1.875065
- min 0.000000
- 25% 1.000000
- 50% 2.000000
- 75% 5.000000
- max 6.000000

BsmtFinSF1

- count 1168.000000
- mean 444.726027
- std 462.664785
- min 0.000000
- 25% 0.000000
- 50% 385.500000
- 75% 714.500000
- max 5644.000000

BsmtFinType2

- count 1168.000000
- mean 4.739726
- std 0.947136
- min 0.000000
- 25% 5.000000
- 50% 5.000000
- 75% 5.000000
- max 6.000000

BsmtFinSF2

- count 1168.000000
- mean 46.647260
- std 163.520016
- min 0.000000
- 25% 0.000000
- 50% 0.000000
- 75% 0.000000
- max 1474.000000

BsmtUnfSF

- count 1168.000000
- mean 569.721747
- std 449.375525
- min 0.000000
- 25% 216.000000
- 50% 474.000000
- 75% 816.000000
- max 2336.000000

TotalBsmtSF

- count 1168.000000
- mean 1061.095034
- std 442.272249
- min 0.000000
- 25% 799.000000
- 50% 1005.500000
- 75% 1291.500000
- max 6110.000000

Heating

- count 1168.000000
- mean 1.035959
- std 0.302078
- min 0.000000
- 25% 1.000000
- 50% 1.000000
- 75% 1.000000
- max 5.000000

HeatingQC

- count 1168.000000
- mean 1.569349
- std 1.749129
- min 0.000000
- 25% 0.000000
- 50% 0.000000
- 75% 4.000000
- max 4.000000

CentralAir

- count 1168.000000
- mean 0.933219
- std 0.249749
- min 0.000000
- 25% 1.000000
- 50% 1.000000
- 75% 1.000000
- max 1.000000

Electrical

- count 1168.000000
- mean 3.688356
- std 1.042606
- min 0.000000
- 25% 4.000000
- 50% 4.000000
- 75% 4.000000
- max 4.000000

1stFlrSF

- count 1168.000000
- mean 1169.860445
- std 391.161983
- min 334.000000
- 25% 892.000000
- 50% 1096.500000
- 75% 1392.000000
- max 4692.000000

2ndFlrSF

- count 1168.000000
- mean 348.826199
- std 439.696370
- min 0.000000
- 25% 0.000000
- 50% 0.000000
- 75% 729.000000
- max 2065.000000

LowQualFinSF

- count 1168.000000
- mean 6.380137
- std 50.892844
- min 0.000000
- 25% 0.000000
- 50% 0.000000
- 75% 0.000000
- max 572.000000

GrLivArea

- count 1168.000000
- mean 1525.066781
- std 528.042957
- min 334.000000
- 25% 1143.250000
- 50% 1468.500000
- 75% 1795.000000
- max 5642.000000

BsmtFullBath

- count 1168.000000
- mean 0.425514
- std 0.521615
- min 0.000000
- 25% 0.000000
- 50% 0.000000
- 75% 1.000000
- max 3.000000

BsmtHalfBath

- count 1168.000000
- mean 0.055651
- std 0.236699
- min 0.000000
- 25% 0.000000
- 50% 0.000000
- 75% 0.000000
- max 2.000000

FullBath

- count 1168.000000
- mean 1.562500
- std 0.551882
- min 0.000000
- 25% 1.000000
- 50% 2.000000
- 75% 2.000000
- max 3.000000

HalfBath

- count 1168.000000
- mean 0.388699
- std 0.504929
- min 0.000000
- 25% 0.000000
- 50% 0.000000
- 75% 1.000000
- max 2.000000

BedroomAbvGr

- count 1168.000000
- mean 2.884418
- std 0.817229
- min 0.000000
- 25% 2.000000
- 50% 3.000000
- 75% 3.000000
- max 8.000000

KitchenAbvGr

- count 1168.000000
- mean 1.045377
- std 0.216292
- min 0.000000
- 25% 1.000000
- 50% 1.000000
- 75% 1.000000
- max 3.000000

KitchenQual

- count 1168.000000
- mean 2.328767
- std 0.832992
- min 0.000000
- 25% 2.000000
- 50% 2.000000
- 75% 3.000000
- max 3.000000

TotRmsAbvGrd

- count 1168.000000
- mean 6.542808
- std 1.598484
- min 2.000000
- 25% 5.000000
- 50% 6.000000
- 75% 7.000000
- max 14.000000

Functional

- count 1168.000000
- mean 5.742295
- std 0.987250
- min 0.000000
- 25% 6.000000
- 50% 6.000000
- 75% 6.000000
- max 6.000000

Fireplaces

- count 1168.000000
- mean 0.617295
- std 0.650575
- min 0.000000
- 25% 0.000000
- 50% 1.000000
- 75% 1.000000
- max 3.000000

FireplaceQu

- count 1168.000000
- mean 3.804795
- std 1.400665
- min 0.000000
- 25% 2.000000
- 50% 4.000000
- 75% 5.000000
- max 5.000000

GarageType

- count 1168.000000
- mean 2.499144
- std 1.935551
- min 0.000000
- 25% 1.000000
- 50% 1.000000
- 75% 5.000000
- max 6.000000

GarageYrBlt

- count 1168.000000
- mean 1869.799658
- std 451.037303
- min 0.000000
- 25% 1957.750000
- 50% 1977.000000
- 75% 2001.000000
- max 2010.000000

GarageFinish

- count 1168.000000
- mean 1.288527
- std 0.889704
- min 0.000000
- 25% 1.000000
- 50% 1.000000
- 75% 2.000000
- max 3.000000

GarageCars

- count 1168.000000
- mean 1.776541
- std 0.745554
- min 0.000000
- 25% 1.000000
- 50% 2.000000
- 75% 2.000000
- max 4.000000

GarageArea

- count 1168.000000
- mean 476.860445
- std 214.466769
- min 0.000000
- 25% 338.000000
- 50% 480.000000
- 75% 576.000000
- max 1418.000000

GarageQual

- count 1168.000000
- mean 3.927226
- std 0.645872
- min 0.000000
- 25% 4.000000
- 50% 4.000000
- 75% 4.000000
- max 5.000000

GarageCond

- count 1168.000000
- mean 3.960616
- std 0.561694
- min 0.000000
- 25% 4.000000
- 50% 4.000000
- 75% 4.000000
- max 5.000000

PavedDrive

- count 1168.000000
- mean 1.853596
- std 0.501894
- min 0.000000
- 25% 2.000000
- 50% 2.000000
- 75% 2.000000
- max 2.000000

WoodDeckSF

- count 1168.000000
- mean 96.206336
- std 126.158988
- min 0.000000
- 25% 0.000000
- 50% 0.000000
- 75% 171.000000
- max 857.000000

OpenPorchSF

- count 1168.000000
- mean 46.559932
- std 66.381023
- min 0.000000
- 25% 0.000000
- 50% 24.000000
- 75% 70.000000
- max 547.000000

EnclosedPorch

- count 1168.000000
- mean 23.015411
- std 63.191089
- min 0.000000
- 25% 0.000000
- 50% 0.000000
- 75% 0.000000
- max 552.000000

3SsnPorch

- count 1168.000000
- mean 3.639555
- std 29.088867
- min 0.000000
- 25% 0.000000
- 50% 0.000000
- 75% 0.000000
- max 508.000000

ScreenPorch

- count 1168.000000
- mean 15.051370
- std 55.080816
- min 0.000000
- 25% 0.000000
- 50% 0.000000
- 75% 0.000000
- max 480.000000

PoolArea

- count 1168.000000
- mean 3.448630
- std 44.896939
- min 0.000000
- 25% 0.000000
- 50% 0.000000
- 75% 0.000000
- max 738.000000

PoolQC

- count 1168.000000
- mean 2.988870
- std 0.157245
- min 0.000000
- 25% 3.000000
- 50% 3.000000
- 75% 3.000000
- max 3.000000

Fence

- count 1168.000000
- mean 3.475171
- std 1.112090
- min 0.000000
- 25% 4.000000
- 50% 4.000000
- 75% 4.000000
- max 4.000000

MiscFeature

- count 1168.000000
- mean 3.921233
- std 0.408514
- min 0.000000
- 25% 4.000000
- 50% 4.000000
- 75% 4.000000
- max 4.000000

MiscVal

- count 1168.000000
- mean 47.315068
- std 543.264432
- min 0.000000
- 25% 0.000000
- 50% 0.000000
- 75% 0.000000
- max 15500.000000

MoSold

- count 1168.000000
- mean 6.344178
- std 2.686352
- min 1.000000
- 25% 5.000000
- 50% 6.000000
- 75% 8.000000
- max 12.000000

YrSold

- count 1168.000000
- mean 2007.804795
- std 1.329738
- min 2006.000000
- 25% 2007.000000
- 50% 2008.000000
- 75% 2009.000000
- max 2010.000000

SaleType

- count 1168.000000
- mean 7.465753
- std 1.619459
- min 0.000000
- 25% 8.000000
- 50% 8.000000
- 75% 8.000000
- max 8.000000

SaleCondition

- count 1168.000000
- mean 3.768836
- std 1.112208
- min 0.000000
- 25% 4.000000
- 50% 4.000000
- 75% 4.000000
- max 5.000000

SalePrice

- count 1168.000000
- mean 181477.005993
- std 79105.586863
- min 34900.000000
- 25% 130375.000000
- 50% 163995.000000
- 75% 215000.000000
- max 755000.000000

- Correlation between the columns and the label ‘Sale Price’ using corr method-
    - SalePrice 1.000000
    - OverallQual 0.789185
    - GrLivArea 0.707300
    - GarageCars 0.628329
    - GarageArea 0.619000
    - ...
    - FireplaceQu -0.445910
    - GarageFinish -0.550624
    - KitchenQual -0.592468
    - ExterQual -0.624820
    - BsmtQual -0.628798

- OverallQual is 78% positively correlated to ‘Sale Price’
- GrLivArea is 70% positively correlated to ‘Sale Price’
- GarageCars is 62% positively correlated to ‘Sale Price’
- GarageArea is 61% positively correlated to ‘Sale Price’
- Fireplace is 44% negatively correlated to ‘Sale Price’
- GarageFinish is 55% negatively correlated to ‘Sale Price’
- KitchenQual is 59% negatively correlated to ‘Sale Price’
- ExterQual and BsmtQual is 62% negatively correlated to ‘Sale Price’

- Visualizing outliers using boxplot method-


TotalBsmtSF
Heating
HeatingQC
CentralAir
Electrical
1stFlrSF
2ndFlrSF
LowQualFinSF
GrLivArea
BsmtFullBath
BsmtHalfBath
FullBath
HalfBath
BedroomAbvGr
KitchenAbvGr
KitchenQual
TotRmsAbvGrd
Functional
Fireplaces
FireplaceQu
GarageType
GarageYrBlt
GarageFinish
GarageCars
GarageArea
GarageQual
GarageCond
PavedDrive
WoodDeckSF
OpenPorchSF
EnclosedPorch
3SsnPorch
ScreenPorch
PoolArea
PoolQC
Fence
MiscFeature
MiscVal
MoSold
YrSold
SaleType
SaleCondition
SalePrice

- Removing outliers using zscore method-

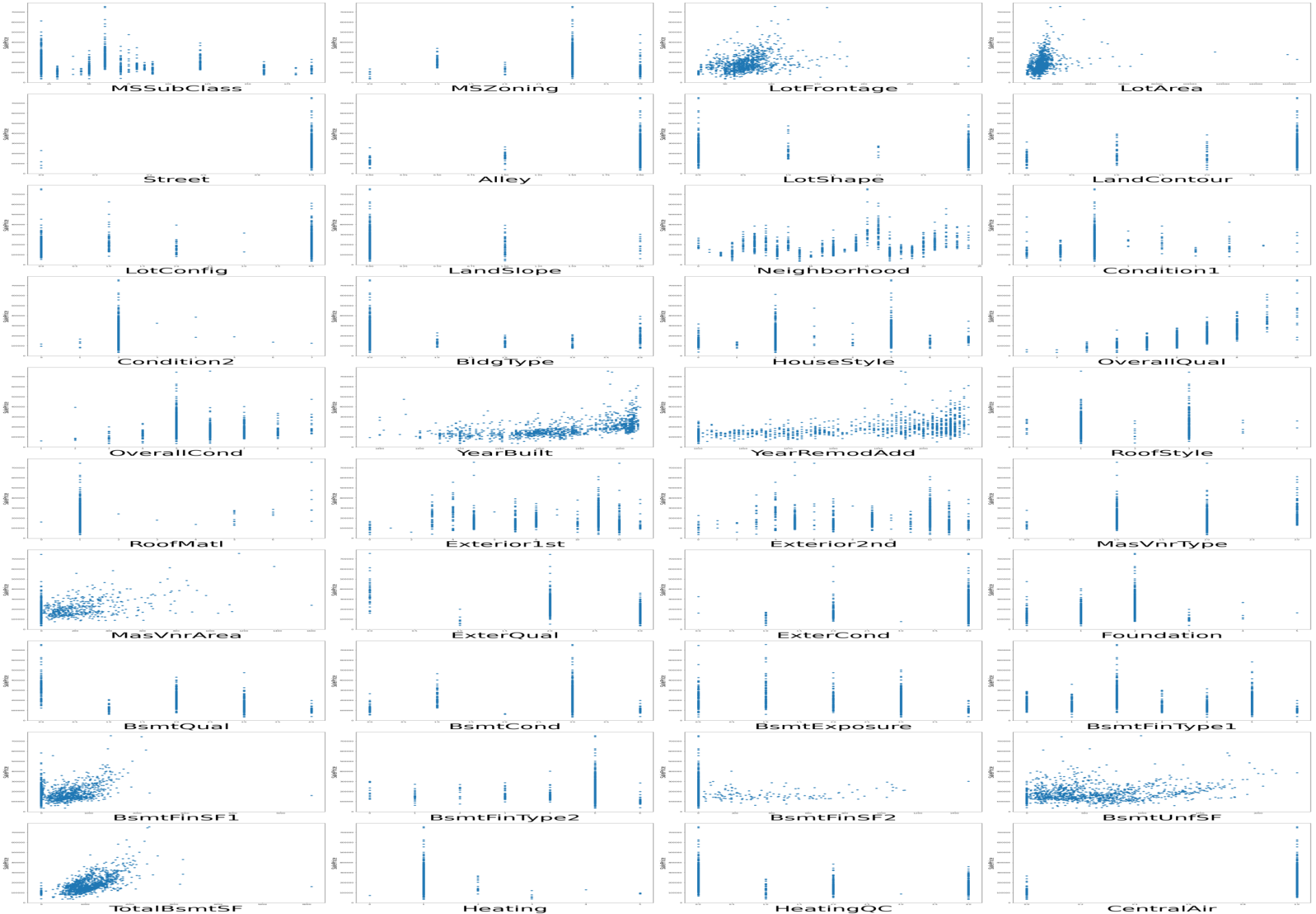
On removing the outliers the data loss is 62.58%, which is not acceptable, hence outliers are tolerated

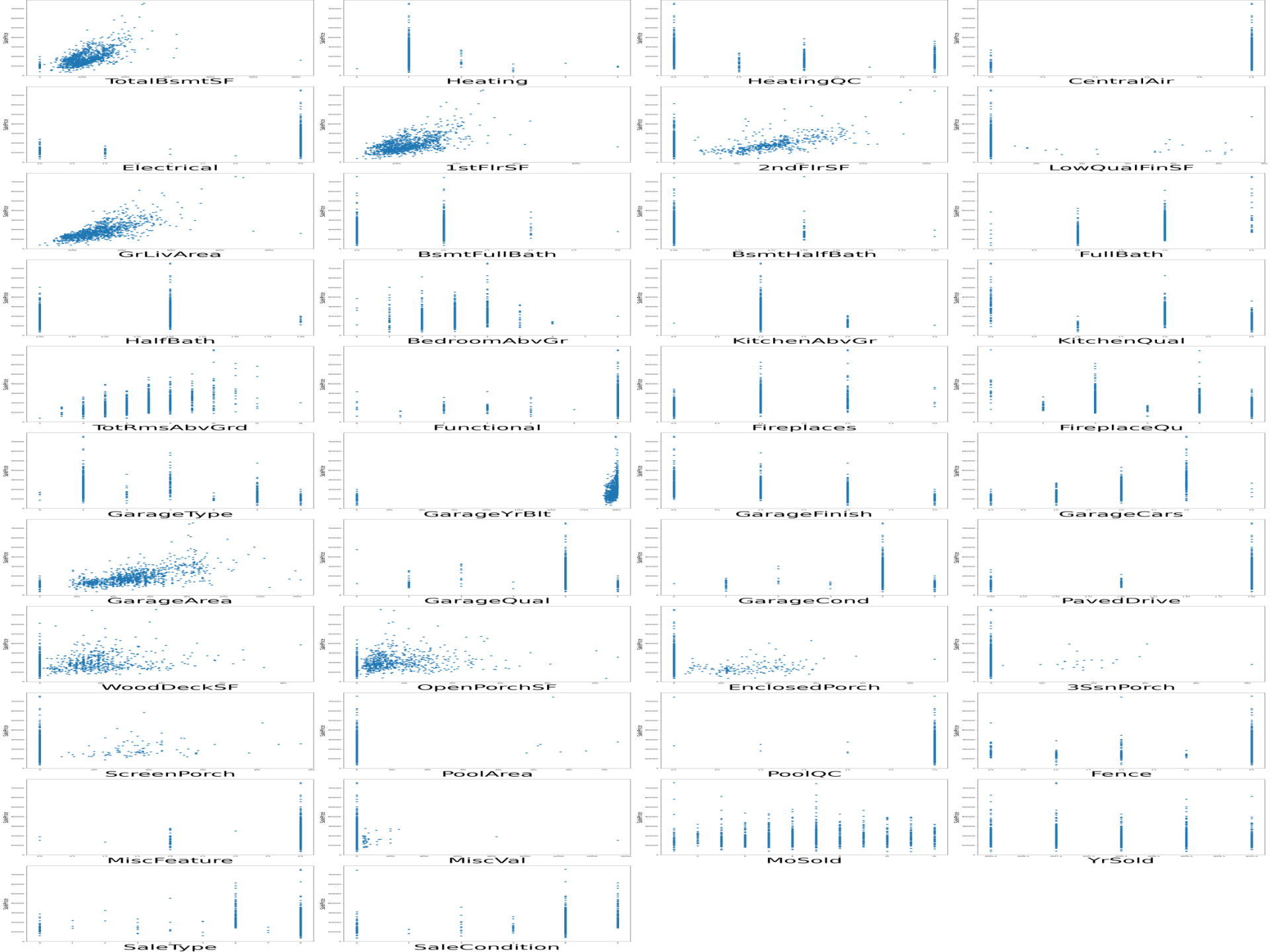
- The dataset is divided into x_train (features) and y_train (label)-

The x_train contains all the features other than the label ‘Sale Price’

The y_train contains only the label ‘Sale Price’

- Visualizing relationship between features and label-




SalePrice
TotalBsmtSF
Heating
HeatingQC
CentralAir
Electrical
1stFlrSF
2ndFlrSF
LowQualFinSF
GrLivArea
BsmtFullBath
BsmtHalfBath
FullBath
HalfBath
BedroomAbvGr
KitchenAbvGr
KitchenQual
TotRmsAbvGrd
Functional
Fireplaces
FireplaceQu
GarageType
GarageYrBlt
GarageFinish
GarageCars
GarageArea
GarageQual
GarageCond
PavedDrive
WoodDeckSF
OpenPorchSF
EnclosedPorch
3SsnPorch
ScreenPorch
PoolArea
PoolQC
Fence
MiscFeature
MiscVal
MoSold
YrSold
SaleType
SaleCondition

- Selecting Best Features using SelectPercentile and chi2-

  Hereby, top 80% features are retrieved, as follows-

- ❑ MSSubClass
- ❑ LotFrontage
- ❑ LotArea
- ❑ LotShape
- ❑ LandContour
- ❑ LotConfig
- ❑ LandSlope
- ❑ Neighborhood
- ❑ Condition1
- ❑ BldgType
- ❑ HouseStyle
- ❑ OverallQual
- ❑ OverallCond
- ❑ YearBuilt
- ❑ YearRemodAdd
- ❑ RoofStyle
- ❑ RoofMatl
- ❑ Exterior1st
- ❑ Exterior2nd
- ❑ MasVnrType
- ❑ MasVnrArea
- ❑ ExterQual
- ❑ Foundation
- ❑ BsmtQual
- ❑ BsmtCond
- ❑ BsmtExposure
- ❑ BsmtFinType1
- ❑ BsmtFinSF1
- ❑ BsmtFinSF2
- ❑ BsmtUnfSF
- ❑ TotalBsmtSF
- ❑ HeatingQC
- ❑ Electrical
- ❑ 1stFlrSF
- ❑ 2ndFlrSF
- ❑ LowQualFinSF
- ❑ GrLivArea
- ❑ BsmtFullBath
- ❑ BsmtHalfBath
- ❑ FullBath
- ❑ HalfBath
- ❑ BedroomAbvGr
- ❑ KitchenQual
- ❑ TotRmsAbvGrd
- ❑ Fireplaces
- ❑ FireplaceQu
- ❑ GarageType
- ❑ GarageYrBlt
- ❑ GarageFinish
- ❑ GarageCars
- ❑ GarageArea
- ❑ WoodDeckSF
- ❑ OpenPorchSF
- ❑ EnclosedPorch
- ❑ 3SsnPorch
- ❑ ScreenPorch
- ❑ PoolArea
- ❑ Fence
- ❑ MiscVal
- ❑ MoSold
- ❑ SaleType
- ❑ SaleCondition

- These features are then set as x_train

- The skewness observes in graphical analysis was confirmed by using the skew method-
  - MiscVal 23.065943
  - PoolArea 13.243711
  - LotArea 10.659285
  - 3SsnPorch 9.770611
  - LowQualFinSF 8.666142
  - ...
  - BsmtCond -2.927336
  - Electrical -3.104209
  - LandContour -3.125982
  - SaleType -3.660513
  - GarageYrBlt -3.898694

- This skewness was removed using the power transformer

- The x_train is now composed of 1168 rows and 62 columns

- The test datasheet and saved in another dataframe

- From the above datasheet, x_test is extracted containing only the best features obtained and stored in x_train.

- The shape of x_test is 292 rows and 62 columns

The data types of the columns are-

- MSSubClass ---------- int64
- LotFrontage ---------- float64
- LotArea ---------- int64
- LotShape ---------- object
- LandContour ---------- object
- LotConfig ---------- object
- LandSlope ---------- object
- Neighborhood ---------- object
- Condition1 ---------- object
- BldgType ---------- object
- HouseStyle ---------- object
- OverallQual ---------- int64
- OverallCond ---------- int64
- YearBuilt ---------- int64
- YearRemodAdd ---------- int64
- RoofStyle ---------- object
- RoofMatl ---------- object
- Exterior1st ---------- object
- Exterior2nd ---------- object
- MasVnrType ---------- object
- MasVnrArea ---------- float64
- ExterQual ---------- object
- Foundation ---------- object
- BsmtQual ---------- object
- BsmtCond ---------- object
- BsmtExposure ---------- object
- BsmtFinType1 ---------- object
- BsmtFinSF1 ---------- int64
- BsmtFinSF2 ---------- int64
- BsmtUnfSF ---------- int64
- TotalBsmtSF ---------- int64
- HeatingQC ---------- object
- Electrical ---------- object
- 1stFlrSF ---------- int64
- 2ndFlrSF ---------- int64 L
- owQualFinSF ---------- int64
- GrLivArea ---------- int64
- BsmtFullBath ---------- int64
- BsmtHalfBath ---------- int64
- FullBath ---------- int64
- HalfBath ---------- int64
- BedroomAbvGr ---------- int64
- KitchenQual ---------- object
- TotRmsAbvGrd ---------- int64
- Fireplaces ---------- int64
- FireplaceQu ---------- object
- GarageType ---------- object
- GarageYrBlt ---------- float64
- GarageFinish ---------- object
- GarageCars ---------- int64
- GarageArea ---------- int64
- WoodDeckSF ---------- int64
- OpenPorchSF ---------- int64
- EnclosedPorch ---------- int64
- 3SsnPorch ---------- int64
- ScreenPorch ---------- int64
- PoolArea ---------- int64
- Fence ---------- object
- MiscVal ---------- int64
- MoSold ---------- int64
- SaleType ---------- object
- SaleCondition ----------object

- The null values are present in the following columns-

Column	No of null values
LotFrontage	45
MasVnrType	1
MasVnrArea	1
BsmtQual	7
BsmtCond	7
BsmtExposure	7
BsmtFinType1	7
FireplaceQu	139
GarageType	17
GarageYrBlt	17
GarageFinish	17
Fence	248

- Encoding object data in numeric using Label Encoder
- In MasVnrType the null value is replaced by the mode of the column
- In GarageYrBlt, the null values are actually for properties lacking garages, hence null values are filled with 0
- The null values of LotFrontage and MasVnrArea are imputed using KNN Imputer

- The skewness is observed using the skew method-
  - RoofMatl 13.717569
  - MiscVal 13.264758
  - LotArea 12.781805
  - 3SsnPorch 12.277476
  - LowQualFinSF 10.929928
  - ...
  - Electrical -2.955201
  - BsmtCond -3.085864
  - LandContour -3.332422
  - GarageYrBlt -3.776995
  - SaleType -5.489874

- This skewness was removed using the power transformer

- The x_test is now composed of 292 rows × 62 columns

Sofware Requirements-

- Jupyter Notebook – Interface for the program
- Pandas – for datafram working
- Numpy – to deal with null data
- matplotlib.pyplot – for data visualization
- Seaborn - for data visualization
- Warnings- to omit warnings
- sklearn.preprocessing – to import LabelEncoder, powertransform
- Label Encoder- to encode object data to numeric data
- sklearn.impute- to import KNNImputer
- KNNImputer- to fill in the null values with meaningful data
- scipy.stats – to import zscore
- Zscore- to remove outliers
- sklearn.feature_selection- to import SelectPercentile and chi2
- SelectPercentile- to select best features
- Chi2-chi2 retrives p-values which help in filtering the best features
- power_transform- to remove the skewness in the data
- sklearn.linear_model- to import LinearRegression
- LinearRegression- to use LinearRegression model
- sklearn.metrics- to import r2_score
- R2 score- to check for the efficiency of the model
- sklearn.model_selection- to import cross_val_score
- cross_val_score- to check for overfitting and underfitting
- sklearn.model_selection- to import GridSearchCV
- GridSearchCV to enhance the working of the model by manipulating the parameters
- from sklearn.linear_model-to import Lasso
- Lasso- to regularize the linear regression model
- sklearn.ensemble – to import RandomForestRegressor, AdaBoostRegressor
- RandomForestRegressor- to use Random Forest Regressor model
- AdaBoostRegressor- to use AdaBoostRegressor

Model/s Development and Evaluation

The x_train, y_train and x_test were applied on different models as follows

Linear Regression Model

In [558]:
```
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score

lr=LinearRegression()
```

In [559]:
```
lr.fit(x_train, y_train)
pred_train=lr.predict(x_train)
print(r2_score(y_train,pred_train))
```

Cross Validation of the model

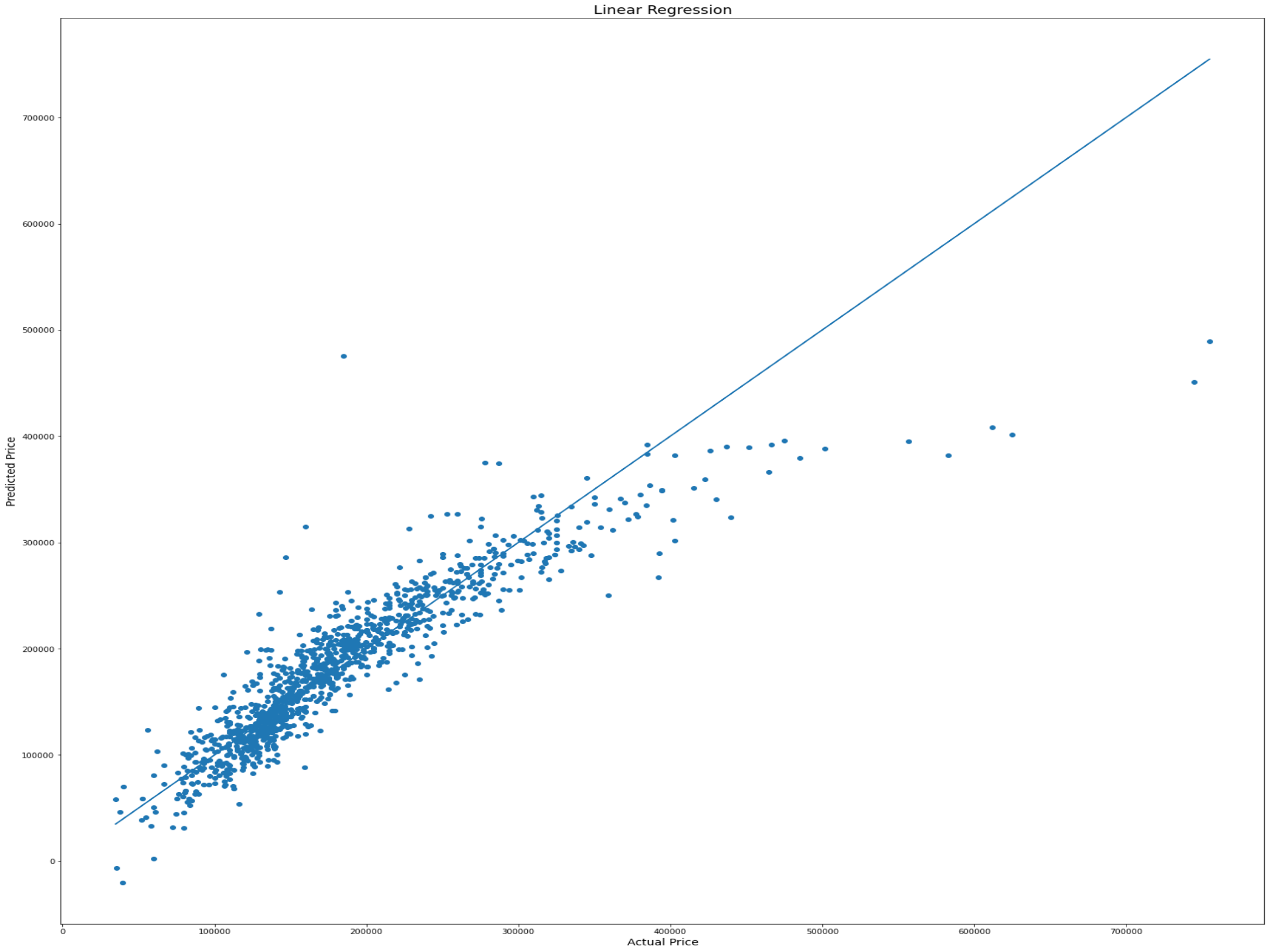
In [560]:
```
pred_train=lr.predict(x_train)
Train_accuracy=r2_score(y_train,pred_train)

from sklearn.model_selection import cross_val_score
for j in range(2,10):
    cv_score=cross_val_score(lr,x_train,y_train,cv=j)
    cv_mean=cv_score.mean()
    print("At cross fold ",j," the cv score is ", cv_mean," and accuracy score for the training is ",Train_accuracy)
    print("\n")
```

```
At cross fold  2  the cv score is  0.7467025683921329  and accuracy score for the training is  0.8313921418464786


At cross fold  3  the cv score is  0.768744857322675  and accuracy score for the training is  0.8313921418464786


At cross fold  4  the cv score is  0.777099154693031  and accuracy score for the training is  0.8313921418464786


At cross fold  5  the cv score is  0.7691533269830705  and accuracy score for the training is  0.8313921418464786


At cross fold  6  the cv score is  0.779514223048288  and accuracy score for the training is  0.8313921418464786


At cross fold  7  the cv score is  0.7768150126258332  and accuracy score for the training is  0.8313921418464786


At cross fold  8  the cv score is  0.7792116577072951  and accuracy score for the training is  0.8313921418464786


At cross fold  9  the cv score is  0.7766446215302888  and accuracy score for the training is  0.8313921418464786
```

- The **R2 score** for y_train and pred_train (data predicted on x_train) is **83.13%**

- Upon cross-validation it was observed that the number of folds did not have such impact on the accuracy and cv score. So cv=8 is selected. Here we have handled the problem of the overfitting and the underfitting by checking the training and testing score

- The graph between Actual Price and Predicted Price depicts the best fit line which passes through maximum of the points, hence suggesting that the model works well-


Linear Regression
Predicted Price
Actual Price
0
100000
200000
300000
400000
500000
600000
700000

Regularization

In [563]:
```
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import cross_val_score
from sklearn.linear_model import Lasso
parameters={'alpha':[.0001, .001, .01,.1, 1, 10], 'random_state':list(range(0,10))}
ls=Lasso()
clf=GridSearchCV(ls,parameters)
clf.fit(x_train,y_train)
print(clf.best_params_)
```

```
{'alpha': 10, 'random_state': 0}
```

In [564]:
```
ls=Lasso(alpha=10, random_state=0)
ls.fit(x_train,y_train)
ls.score(x_train, y_train)
pred_ls=ls.predict(x_train)

lss=r2_score(y_train, pred_ls)
lss
```

Out[564]:
```
0.8313907598460937
```

In [565]:
```
cv_score=cross_val_score(ls,x_train,y_train,cv=8)
cv_mean=cv_score.mean()
cv_mean
```

Out[565]:
```
0.7793974448172889
```

- The Linear Regression Model is Lasso regularized with the aid of GridSearchCV

- The best parameters for alpha and random_state are found as follows-
    - **alpha: 10**
    - **random_state: 0**

- Applying the above found best parameters on Lasso regularized Linear Regression Model, the following was obtained-

    - **R2 score for y_train and pred_train (data predicted on x_train) – 83.13%**
    - **CV score- 77.93%**

Random Forest Regressor Model

In [566]:
```
from sklearn.ensemble import RandomForestRegressor
parameters={'criterion':['mse','mae'], 'max_features':["auto", "sqrt", "log2"]}
rf=RandomForestRegressor()
clf=GridSearchCV(rf,parameters)
clf.fit(x_train,y_train)

print(clf.best_params_)
```

```
{'criterion': 'mse', 'max_features': 'sqrt'}
```

In [567]:
```
rf=RandomForestRegressor(criterion="mse", max_features="sqrt")
rf.fit(x_train, y_train)
rf.score(x_train, y_train)
pred_decision=rf.predict(x_train)

rfs=r2_score(y_train,pred_decision)
print('R2 Score: ', rfs*100)

rfscore=cross_val_score(rf,x_train,y_train,cv=8)
rfc=rfscore.mean()
print("Cross Val Score:", rfc*100)
```

```
R2 Score:  97.90844501303036
Cross Val Score: 85.41556456525264
```

- Random Forest Regressor Model is hyperparameter tuned using GridSearchCV

- The best parameters for criterion and max_features are found as follows-
    - **criterion: mse**
    - **max_features: sqrt**

- Applying the above found best parameters on Random Forest Regressor Model, the following was obtained-

    - **R2 score for y_train and pred_train (data predicted on x_train) –97.90%**
    - **CV score- 85.41%**

Ada Boost Regressor Model

In [568]:
```
from sklearn.ensemble import AdaBoostRegressor
parameters={'n_estimators':np.arange(10,100), 'learning_rate':np.arange(0.01,0.1)}
ad=AdaBoostRegressor()
clf=GridSearchCV(ad,parameters)
clf.fit(x_train,y_train)

print(clf.best_params_)
```

```
{'learning_rate': 0.01, 'n_estimators': 76}
```

In [569]:
```
ad=AdaBoostRegressor(n_estimators=76, learning_rate=0.01)
ad.fit(x_train, y_train)
ad.score(x_train, y_train)
pred_decision=ad.predict(x_train)

ads=r2_score(y_train,pred_decision)
print('R2 Score: ', ads*100)

adscore=cross_val_score(ad,x_train,y_train,cv=8)
adc=adscore.mean()
print("Cross Val Score:", adc*100)
```

```
R2 Score:  79.69514782612488
Cross Val Score: 73.73781280851995
```

- Ada Boost Regressor Model is hyperparameter tuned using GridSearchCV

- The best parameters for n_estimators and learning_rate are found as follows-
    - **learning_rate: 0.01**
    - **n_estimators: 76**

- Applying the above found best parameters on Ada Boost Regressor Model, the following was obtained-

    - **R2 score for y_train and pred_train (data predicted on x_train) –79.69%**
    - **CV score- 73.73%**

Model	R2 score for y_train and pred_train (data predicted on x_train)	CV score
Linear Regression Model	83.13%	77.93%
Random Forest Regressor Model	97.90%	85.41%
AdaBoost Regressor Model	79.69%	73.73%

- The R2 score of Random Forest Regressor 97.91% is and CV score of Random Forest Regressor is 85.41%. This is the best working model and is finalized

- The test data (x_test) is fit into the Random Forest Regressor Model, and the prices are predicted for the various properties

- Surprise Housing can now assess whether the property is good for their business or not

Predicting sale prices of properties of test data

In [572]:
```
#Using Random Forest Regressor on test data
pred_test=rf.predict(x_test)
```

In [574]:
```
print("The price of the property should be as follows :-")
pred_test
```

```
The price of the property should be as follows :-
```

Out[574]:
```
array([327148.24 , 194598.   , 259422.52 , 180974.24 , 246922.57 ,
        84716.72 , 145022.69 , 337963.69 , 244822.75 , 180794.86 ,
        83359.87 , 160211.84 , 129967.41 , 184262.05 , 325893.96 ,
       123569.93 , 109397.5  , 127372.79 , 175719.12 , 194087.28 ,
       153680.88 , 158528.08 , 157147.41 , 107082.56 , 102646.87 ,
       130363.83 , 181002.64 , 149348.07 , 172396.75 , 120422.17 ,
       138923.7  , 194715.12 , 232301.4  , 164376.5  , 113940.1  ,
       179033.63 , 199194.5  , 115032.   , 157565.54 , 147214.22 ,
       109614.13 , 317172.135, 208366.53 , 182999.51 , 145482.25 ,
       136640.38 , 136999.58 , 112569.61 , 212630.34 , 342362.28 ,
       147307.87 , 221975.17 , 105231.32 , 104862.41 , 288557.54 ,
       131224.86 , 145694.63 , 184887.6  , 128998.72 , 255893.46 ,
       106440.92 , 180985.74 , 139087.54 , 148704.37 , 206781.36 ,
       101819.59 , 156433.5  , 203109.89 , 137151.75 , 166898.48 ,
       285489.41 , 177018.5  , 177318.   , 162938.54 , 145765.56 ,
       235315.62 , 308729.18 , 194549.8  , 282675.45 , 148204.   ,
       204018.44 , 145623.95 , 145623.83 , 164266.5  , 185593.3  ,
       229294.76 , 118122.32 , 361403.97 , 154284.42 , 180213.38 ,
       247902.76 , 134617.03 , 146181.91 , 124153.69 , 199899.6  ,
       167131.   , 245526.59 , 176339.75 , 345435.57 , 127962.89 ,
       255958.73 , 104937.24 , 122965.   , 171074.97 , 189157.82 ,
       143266.4  , 270390.05 , 149700.77 , 197585.48 , 205662.02 ,
       201126.4  , 169035.5  , 199019.16 , 241697.92 , 127197.36 ,
```

CONCLUSION

- The EDA analysis of the data is essential as it helps to understand the relationship between the target and features as well as omit out unwanted columns, thereby taking care of overfitting scenario

- It is necessary to have the data encoded properly as well as the null values must be handled with care in order to avoid mis-interpretation of data

- The models should be used properly, as their regularization /hyperparamter tuning is highly advisable for the best outcome.

- The project imparted key knowledge about the retail estate market, and how beneficial it is to choose a well equipped property to enhance the company's returns

- The limitation of the solution is that it predicts the sale price of the property but does not classify it to be purchasable or not. The limitation can be handled, that post the regression model, the findings can be subjected to classification models to classify the properties to be purchased and not to be purchased