HOUSEPRICE

PROJECT ALPHA GROUP

1.INTRODUCTION

Group Leader

DANABAYEV KAKIM
Hanyang University,
Media & Communication,
PhD

Group member

NAM MARGARITA
Duksung Women's University,
International Trade,
bachelor degree

Group member

KHOSHIMOV ZAFARBEK
Chonnam National University,
Business Administration
Bachelor degree

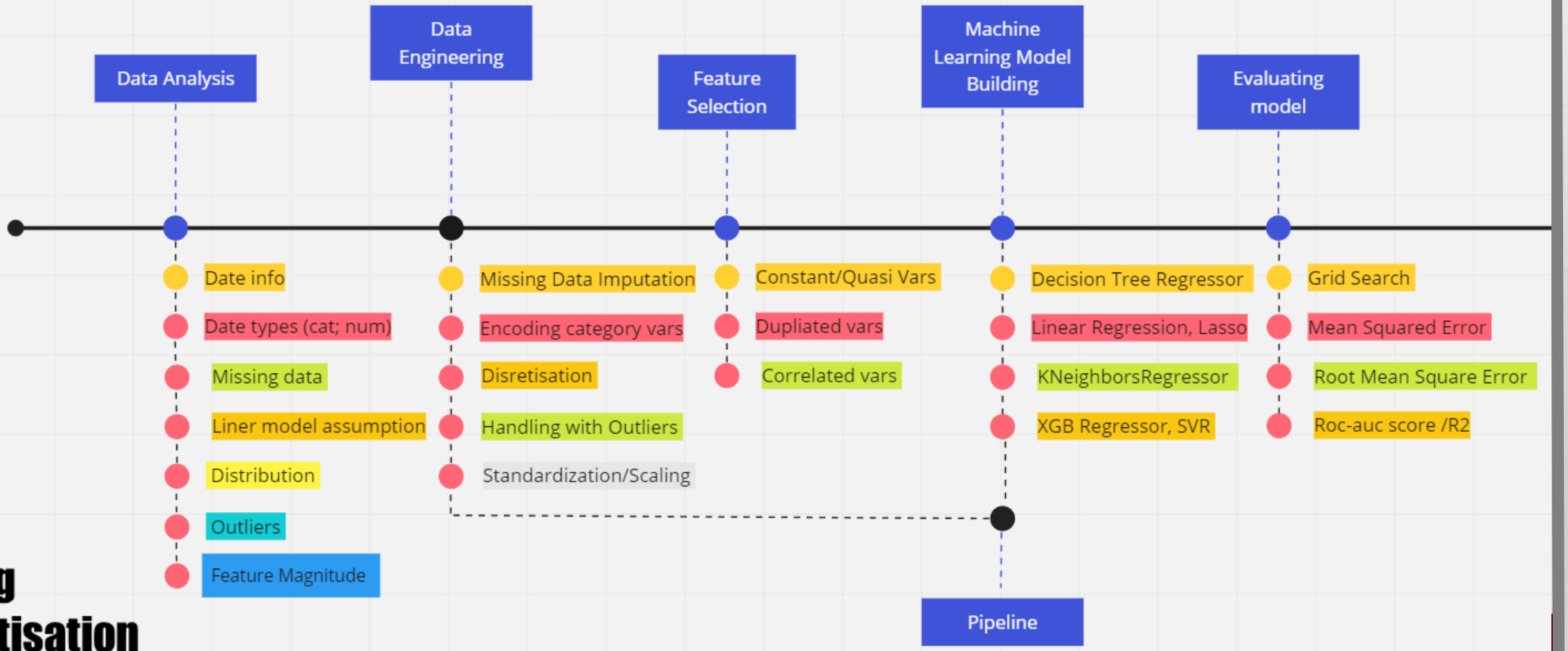
Group member

KIM VERONIKA
Moscow Civil University,
Engineering Architecture,
bachelor degree

DATE PROCESSING

Team Alpha

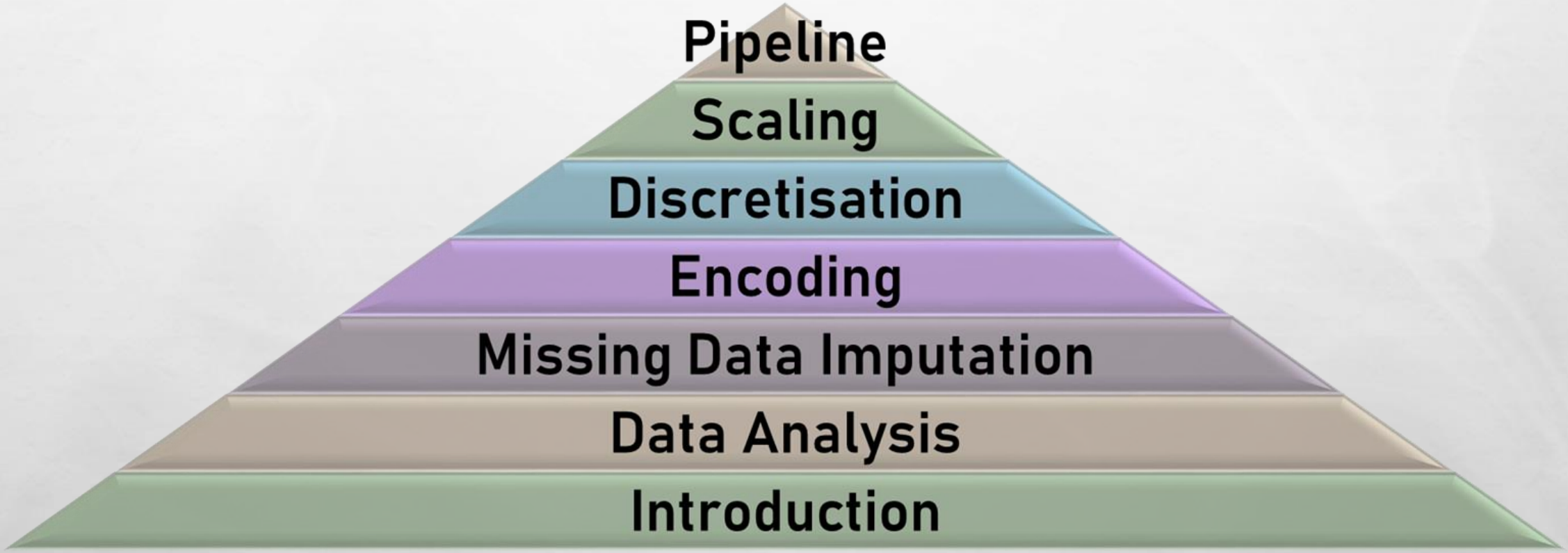
Data Analysis



*Scaling

*Discretisation

PROJECT COMPOSITION



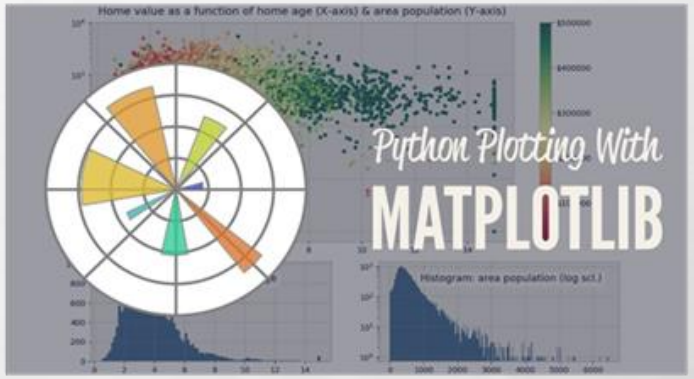
PYTHON DATA ANALYSIS LIBRARY

Pandas for processing and analysis of structured data

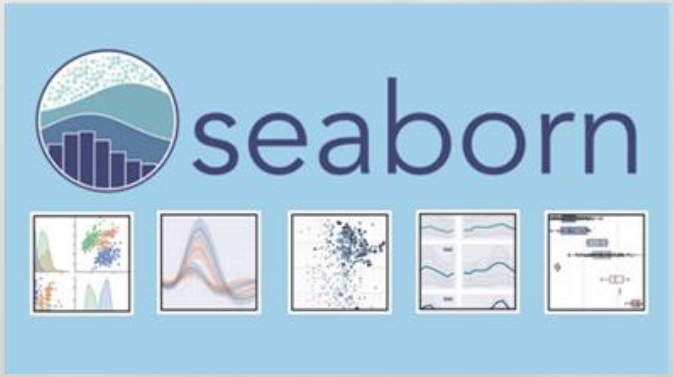
Numpy for used for mathematical calculations, statistics, matrix

Scikit-learn for data functions and algorithms using machine learning

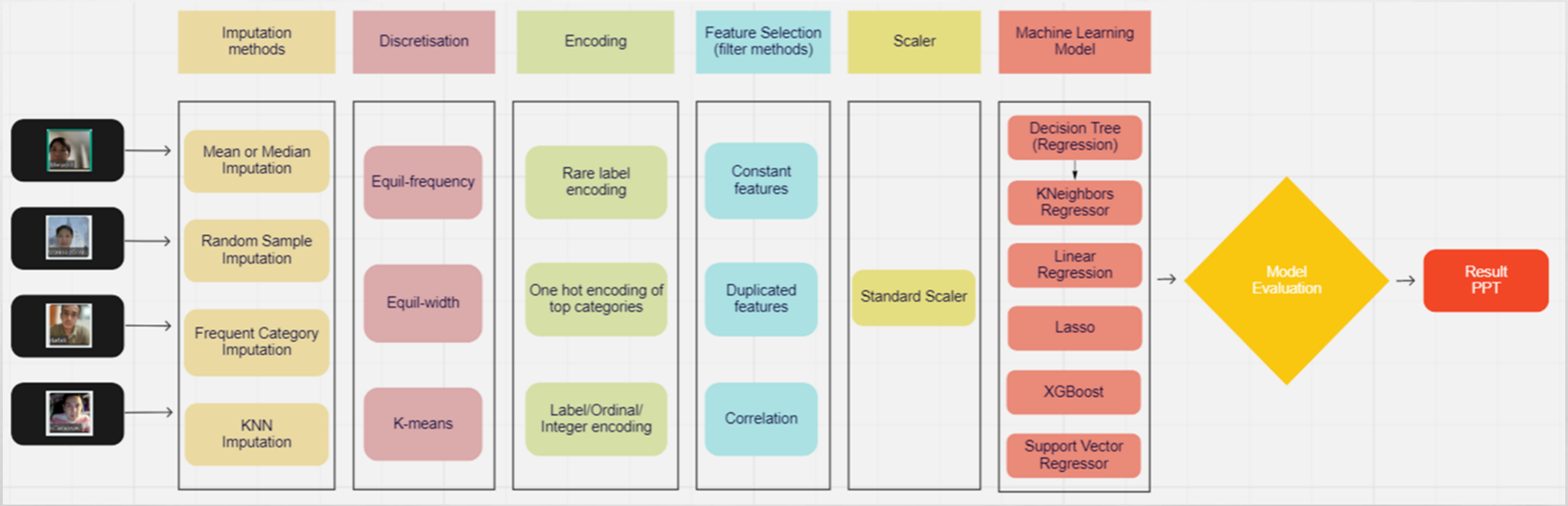
Feature-engine for a library that stores functions algorithms for data processing



Matplotlib&Seaborn for data visualization



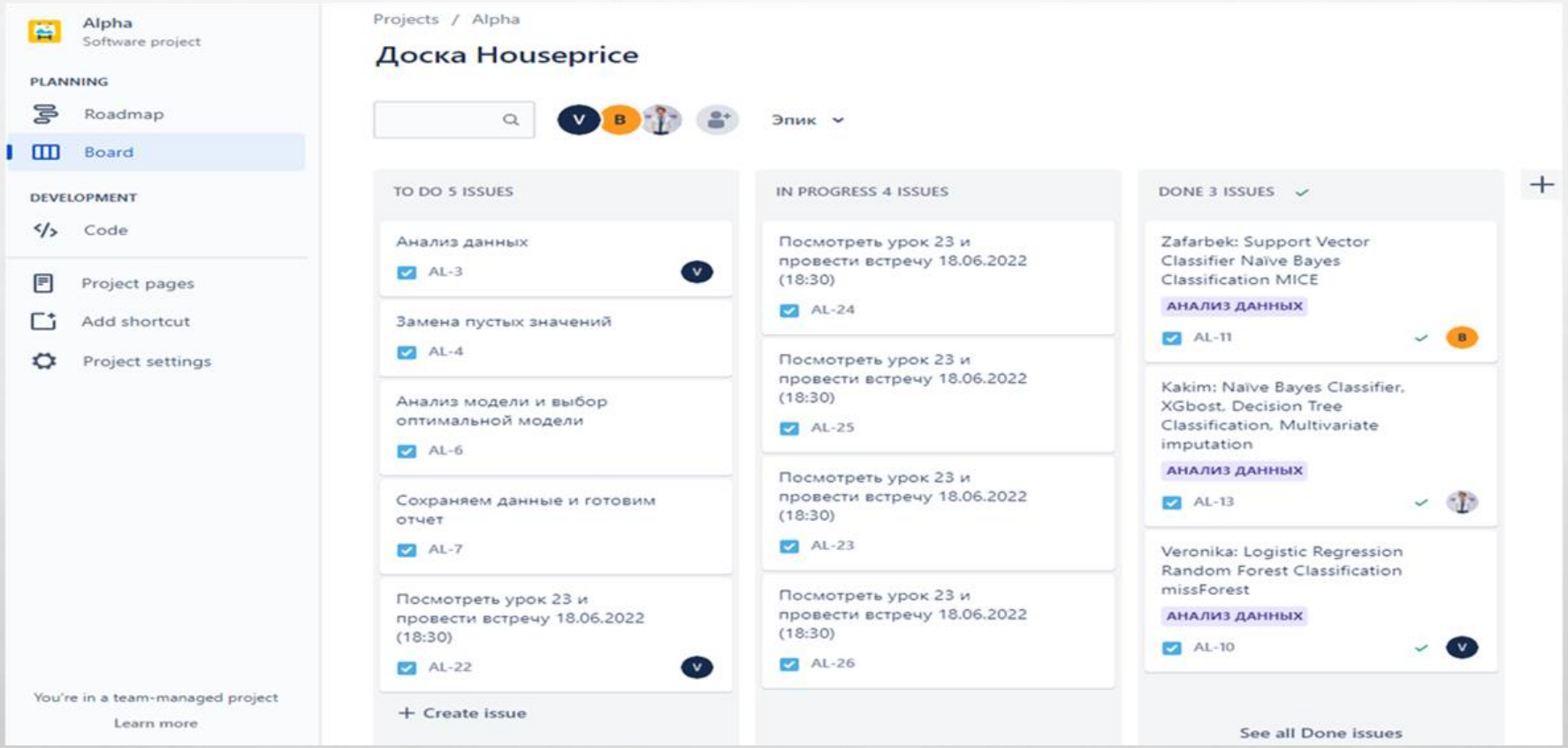
MIRO FOR STRATEGY PROJECT



ATLASSIAN

Jira Software

JIRA FOR REPORT (ON SCHEDULE IN TERMS OF TIMING)



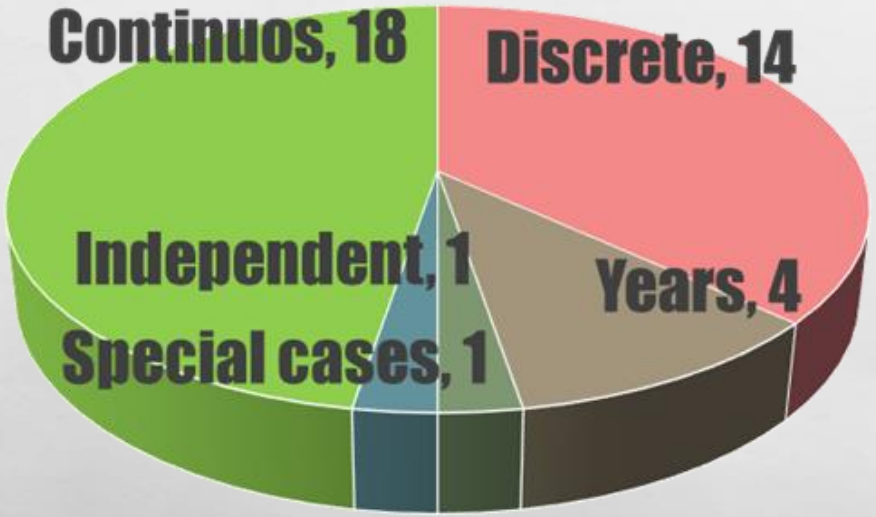
DATA ANALYSIS

- DATA INFO
- DATA TYPES (CATEGORICAL , NUMERICAL VARIABLES)
- MISSING DATA
- LINER MODEL ASSUMPTION
- DISTRIBUTION
- OUTLIERS
- DESCRETE VARIABLES
- RELATIONSHIPS WITH TARGET VALUE (CORRELATION MATRIX)
- CARDINALITY
- FEATURE SCALING

2. DATA ANALYSIS VARIABLE

Dataset : 1460 strings , 81 columns

Numerical



Categorical

TYPES OF VARIABLES

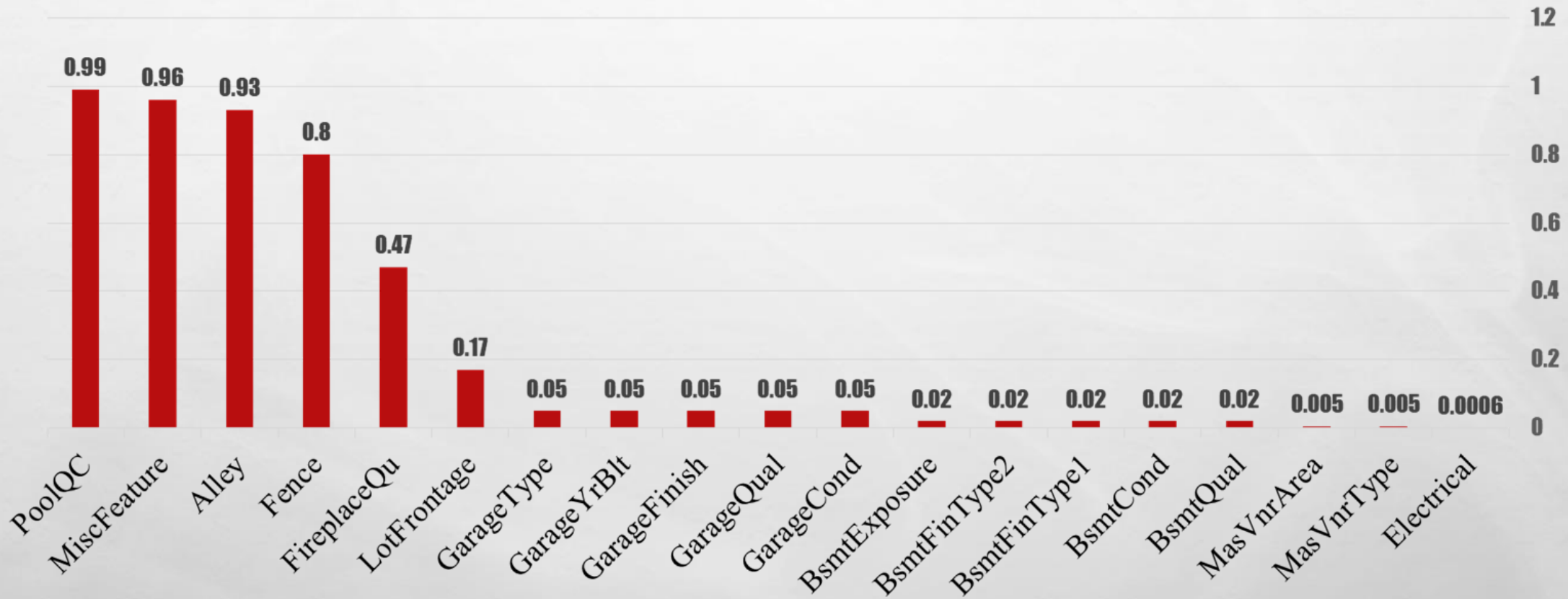
numerical

- BASEMENT: BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, BsmtFullBath, BsmtHalfBath, TotalBsmtSF
- GarageYrBlt, GarageCars, GarageArea
- OpenPorchSF, EnclosedPorch, 3SsnPorch, ScreenPorch,
- BATHROOM: FullBath, HalfBath
- GrLivArea, BedroomAbvGr, KitchenAbvGr,
- 1stFlrSF, 2ndFlrSF
- MSSubClass, LotFrontage, LotArea, OverallQual, OverallCond, MasVnrArea, LowQualFinSF,
- TotRmsAbvGrd, Fireplaces, WoodDeckSF,PoolArea, MiscVal, MoSold

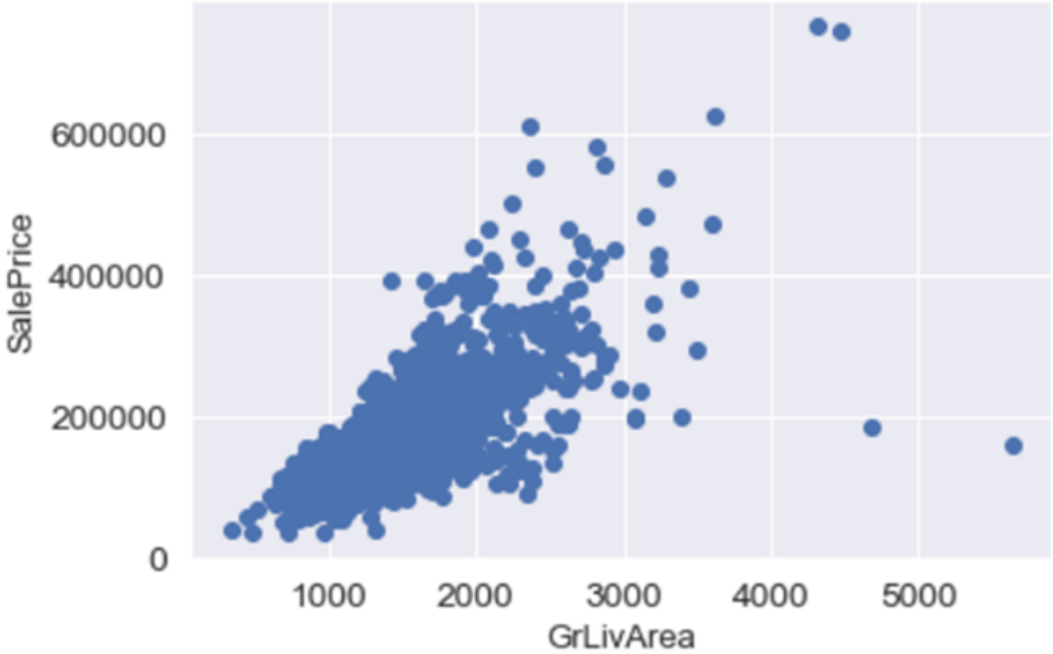
categorical

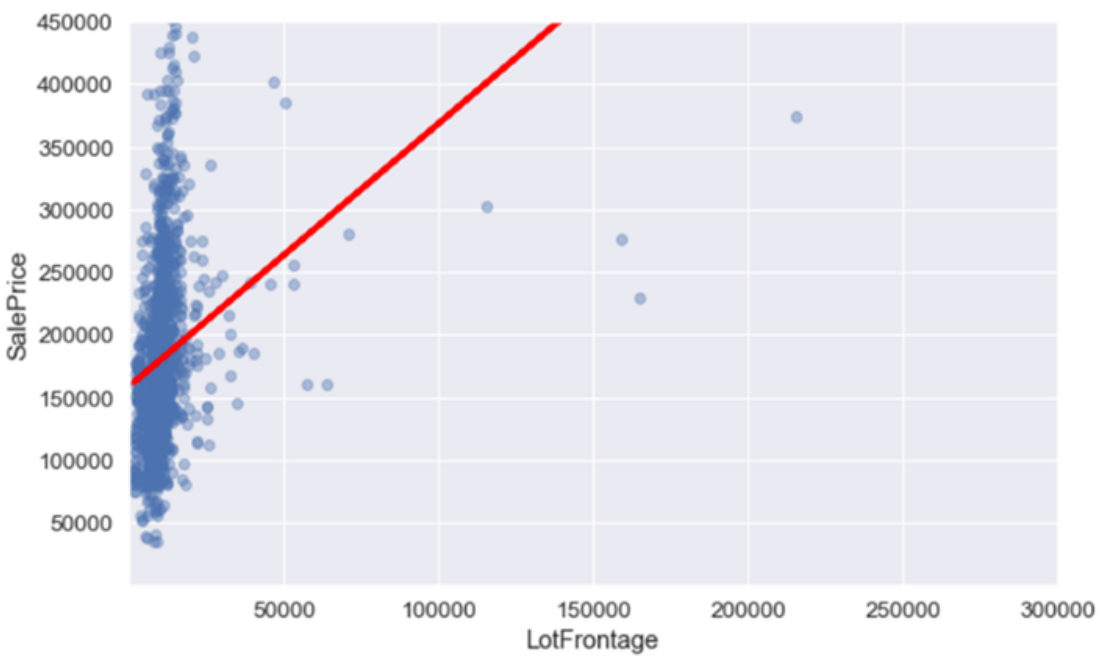
- BASEMENT: BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinType2
- GarageType, GarageFinish, GarageQual, GarageCond
- Heating, HeatingQC, CentralAir, Electrical, Utilities
- Exterior1st, Exterior2nd, ExterQual, ExterCond
- Condition1, Condition2
- BldgType, HouseStyle, RoofStyle, RoofMatl
- YearBuilt, YearRemodAdd, YrSold
- MSZoning, Street, LotShape, LandContour, LotConfig, LandSlope, Neighborhood, MasVnrType, Foundation, KitchenQual, Functional, FireplaceQu,
- PavedDrive, SaleType, SaleCondition

MISSING DATA

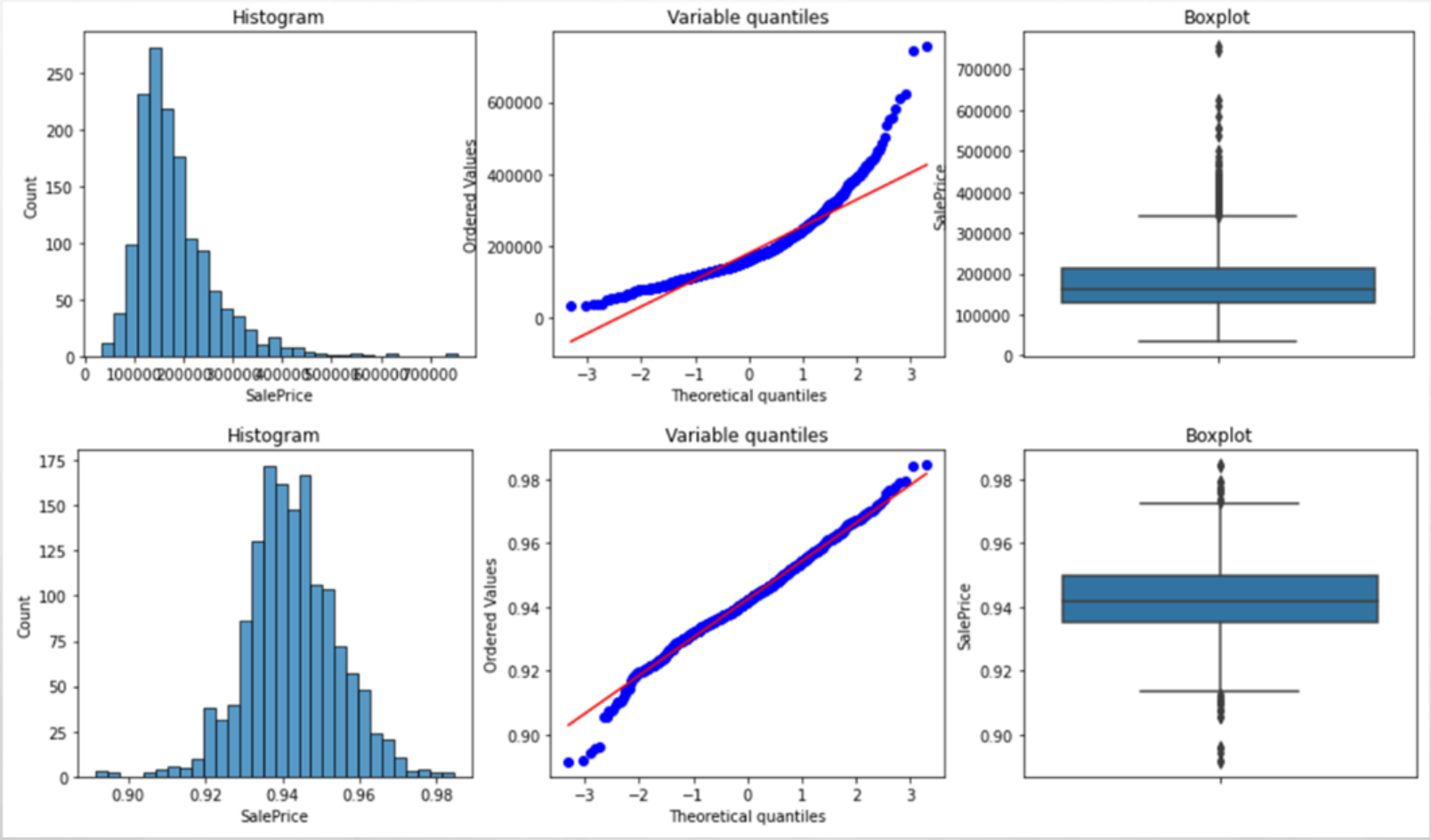


LINEAR MODEL ASSUMPTION





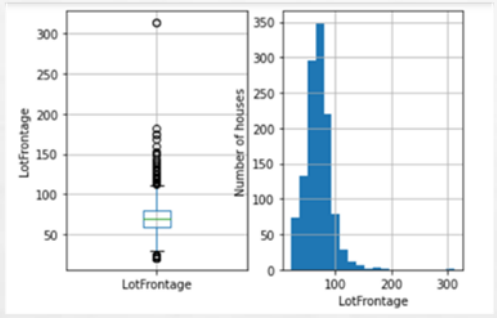
HISTOGRAM DISTRIBUTION SALE PRICE

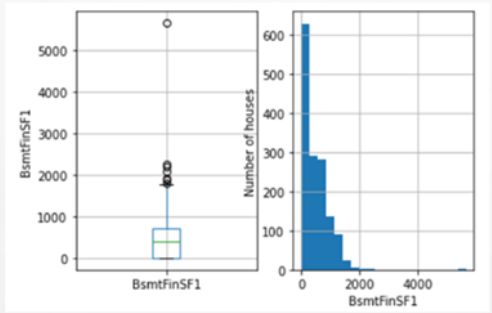


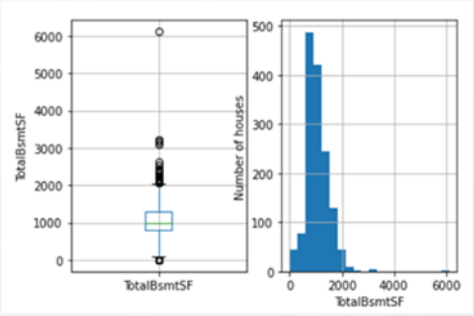
Before Logarithmetic transformation

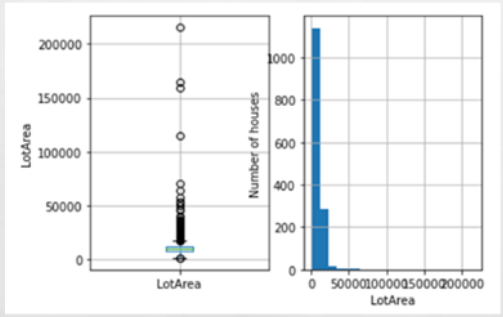
After Logarithmetic transformation

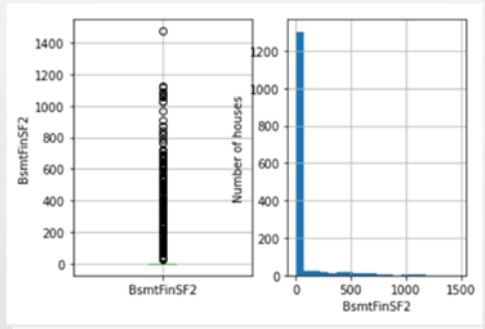
OUTLIERS AND DISTRIBUTIONS -1

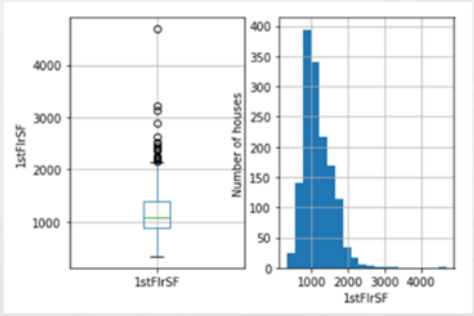


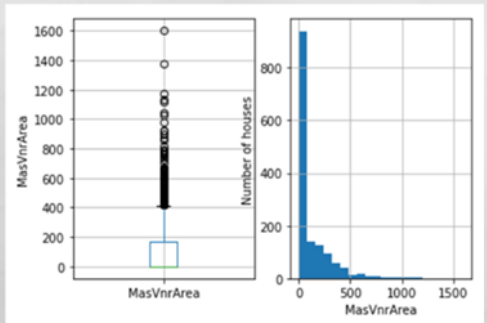


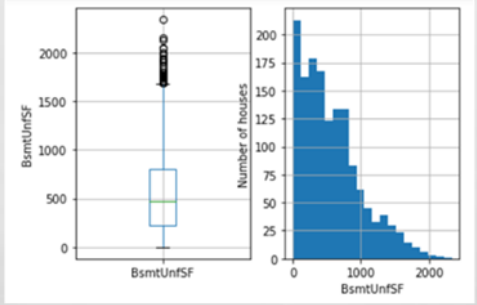


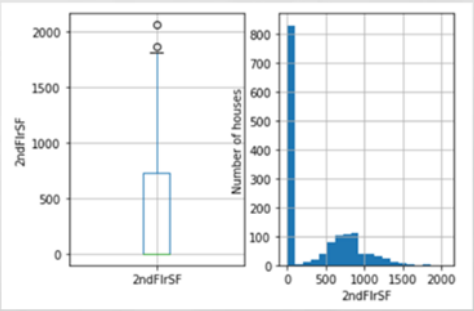




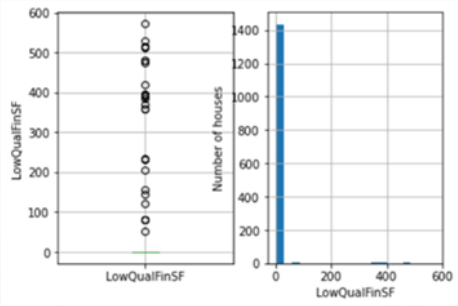


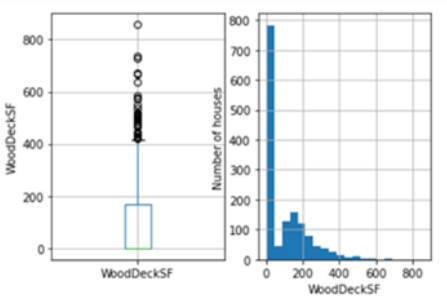


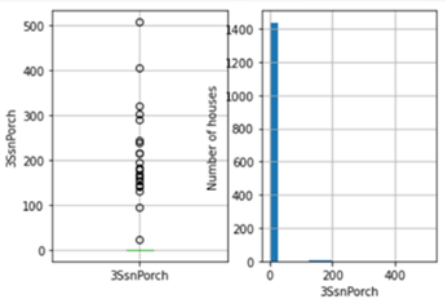


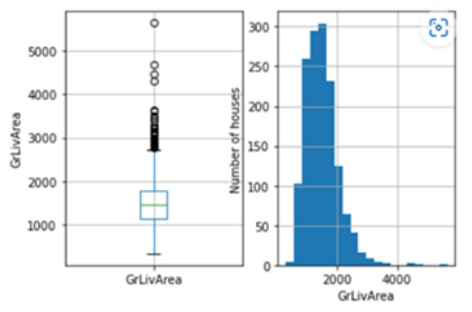


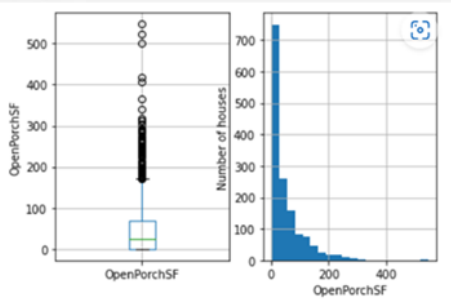
OUTLIERS AND DISTRIBUTIONS -2

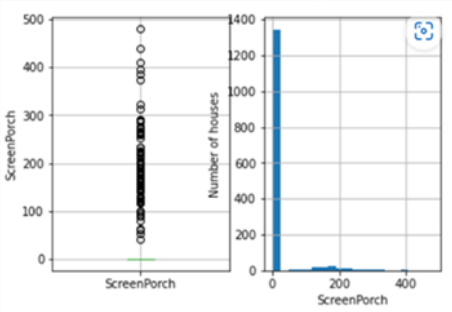


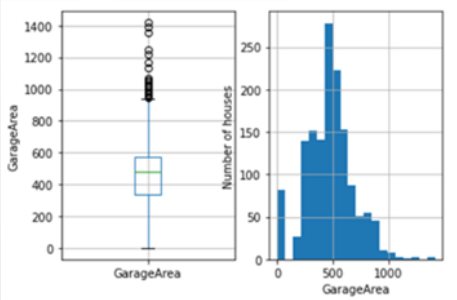


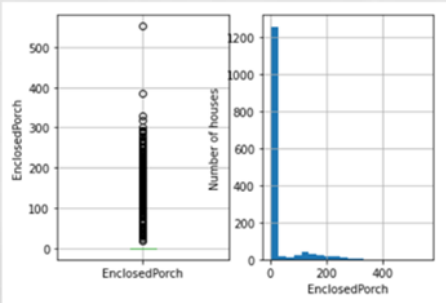


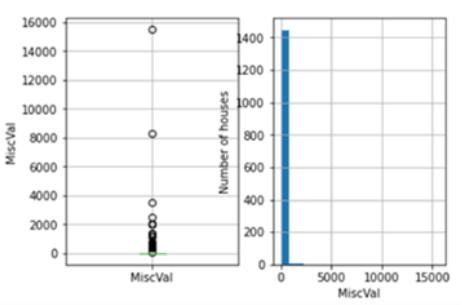




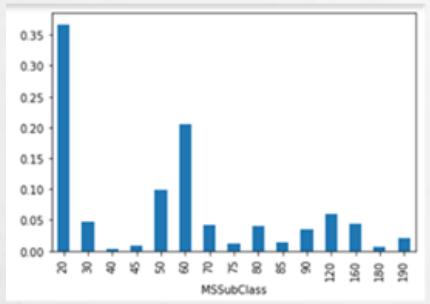


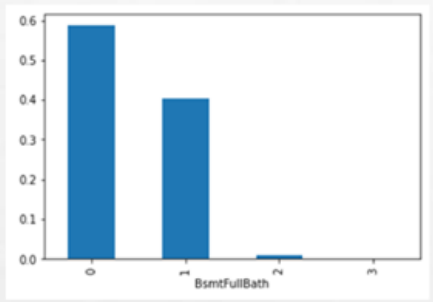


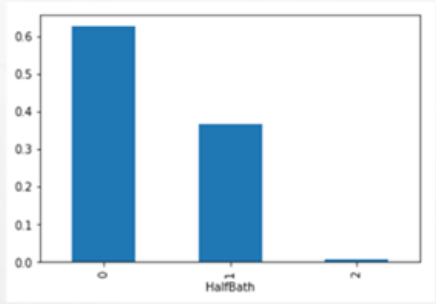


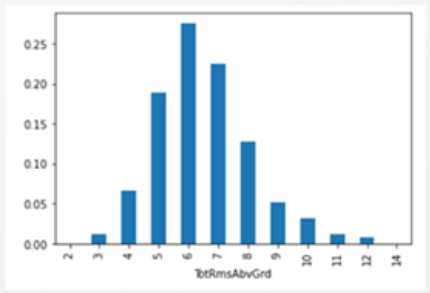


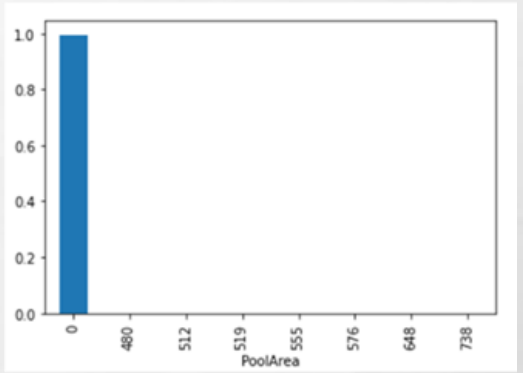
DESCRETE VARIABLES

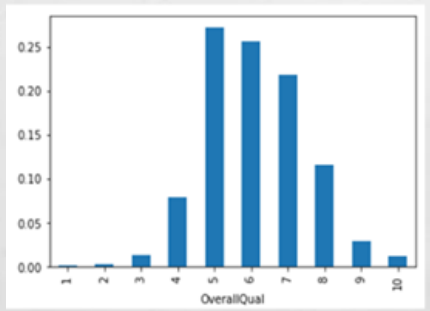


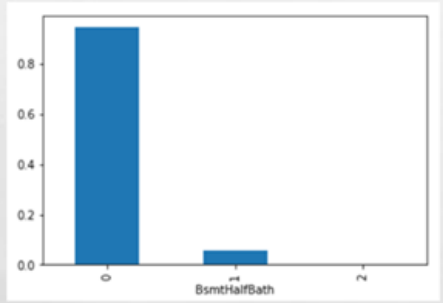


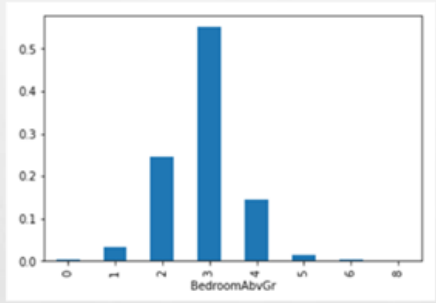


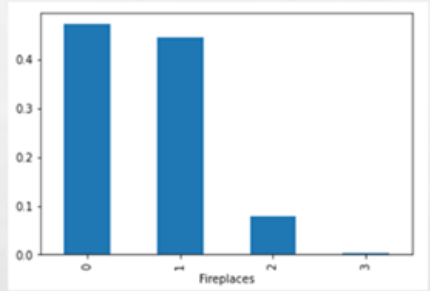


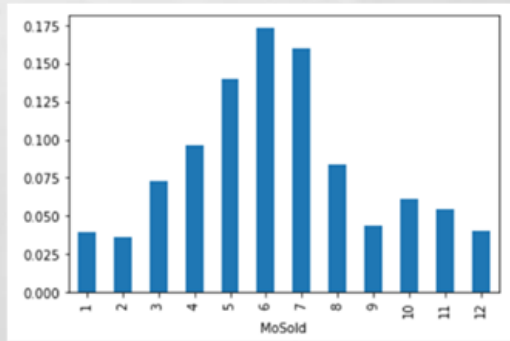


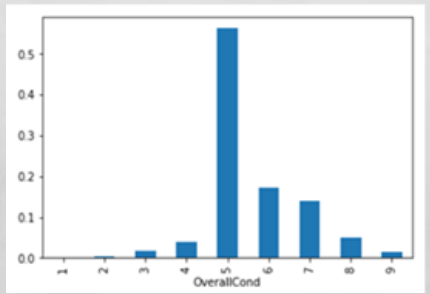


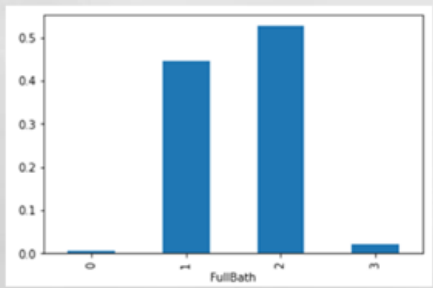


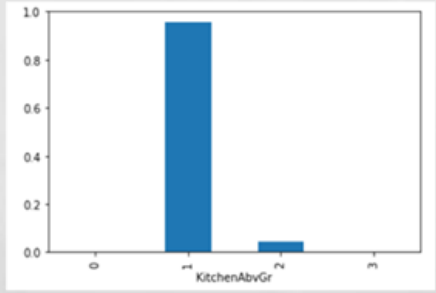


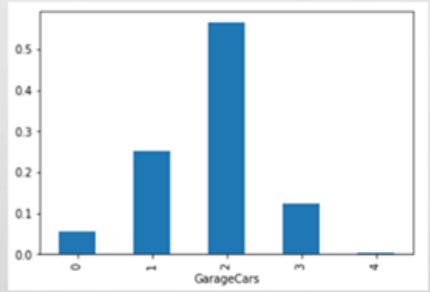




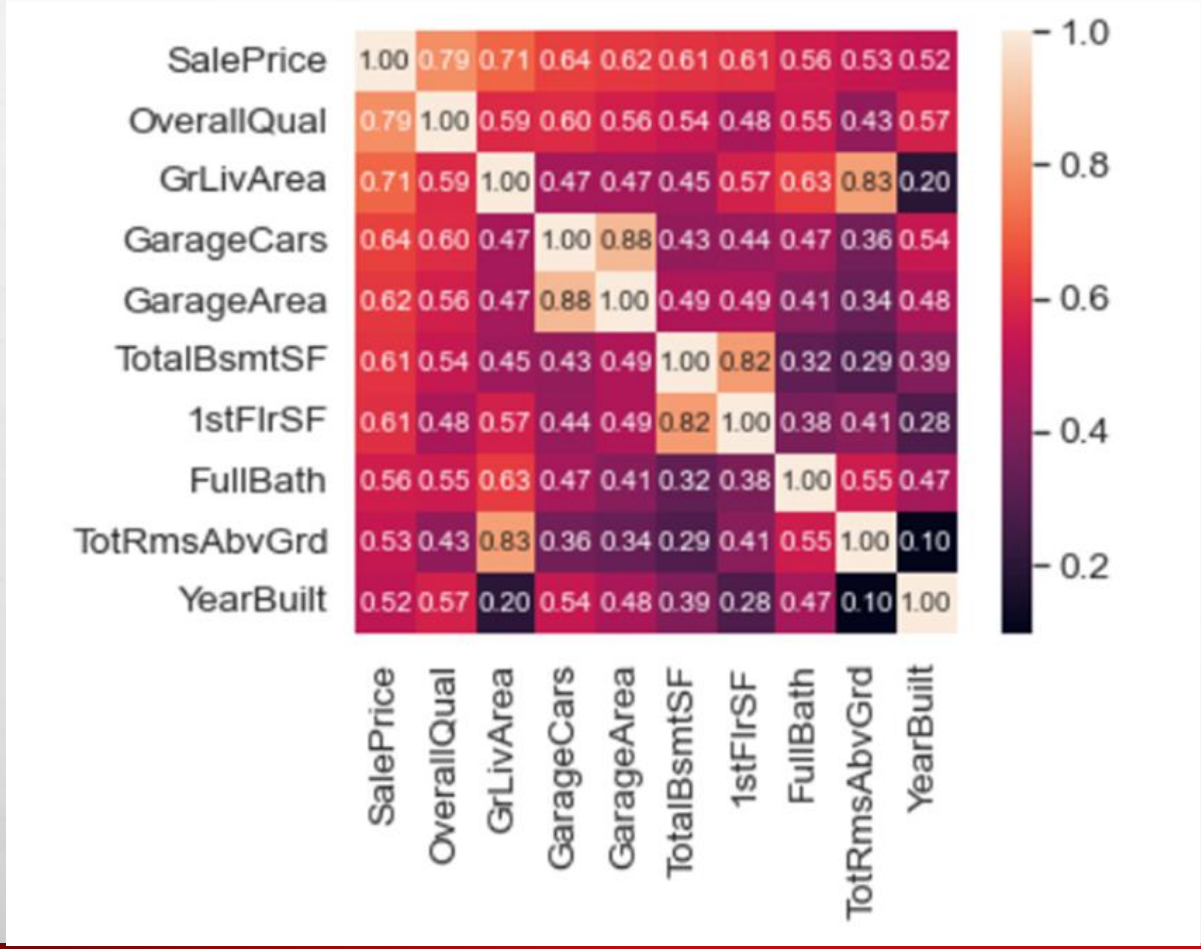








CORRELATION MATRIX



HIGHT CORRELATION FEATURES:

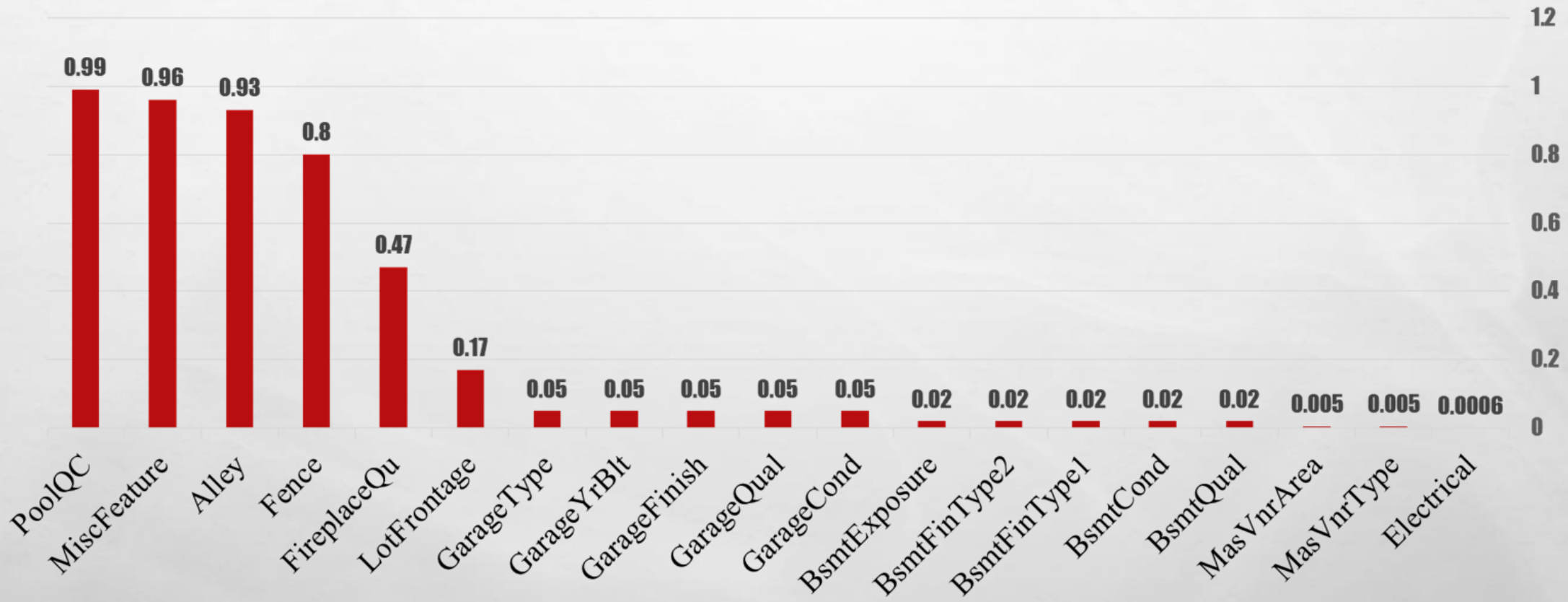
OverallQual
GrLivArea

Hight correlation features > 0.70

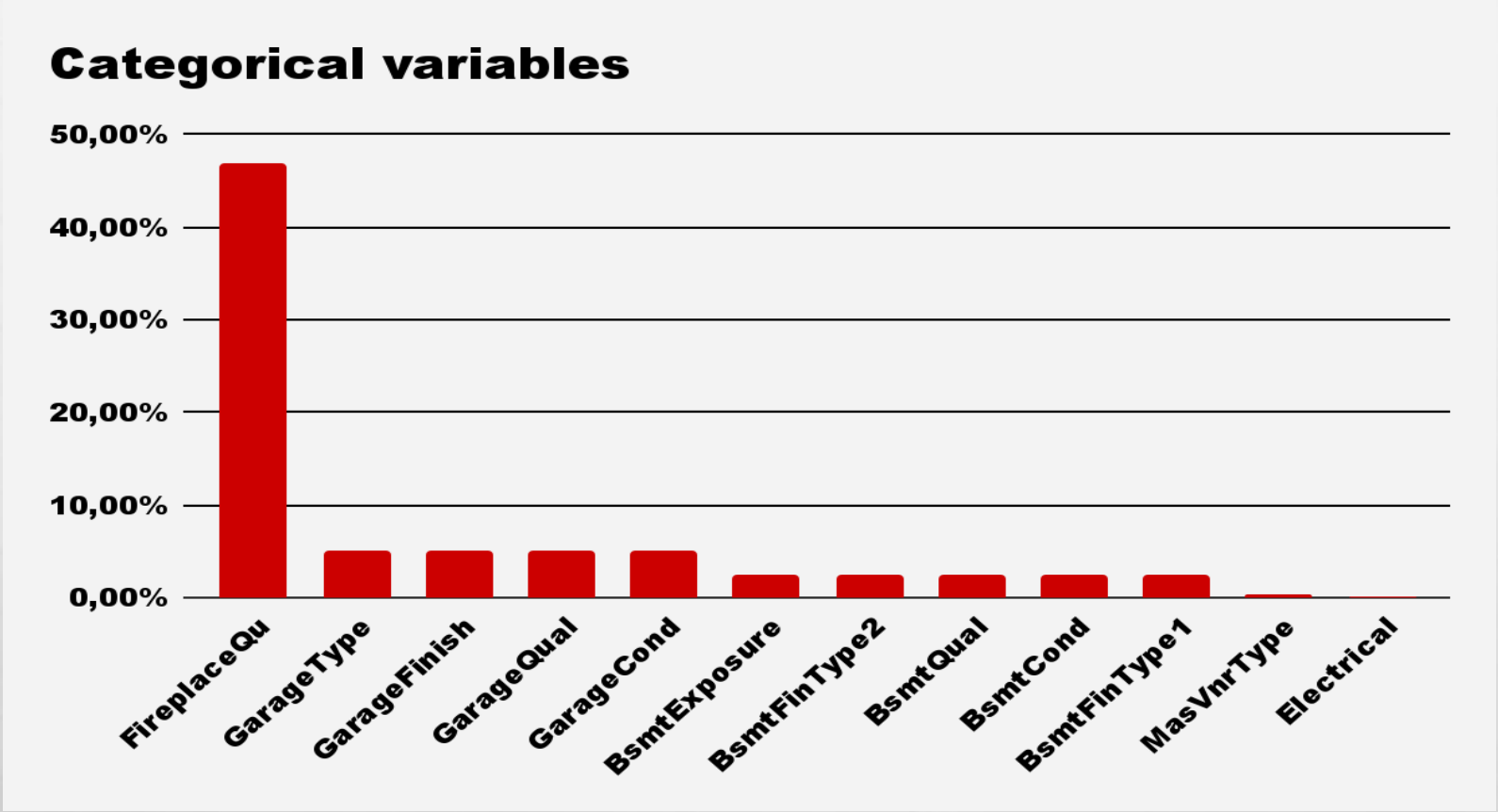
FEATURE SCALING

	MSSubClass	LotFrontage	LotArea	OverallQual	OverallCond	YearBuilt	YearRemodAdd	MasVnrArea	BsmtFinSF1	BsmtFinSF2
count	1460.000000	1201.000000	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000	1452.000000	1460.000000	1460.000000
mean	56.897260	70.049958	10516.828082	6.099315	5.575342	1971.267808	1984.865753	103.685262	443.639726	46.549315
std	42.300571	24.284752	9981.264932	1.382997	1.112799	30.202904	20.645407	181.066207	456.098091	161.319273
min	20.000000	21.000000	1300.000000	1.000000	1.000000	1872.000000	1950.000000	0.000000	0.000000	0.000000
25%	20.000000	59.000000	7553.500000	5.000000	5.000000	1954.000000	1967.000000	0.000000	0.000000	0.000000
50%	50.000000	69.000000	9478.500000	6.000000	5.000000	1973.000000	1994.000000	0.000000	383.500000	0.000000
75%	70.000000	80.000000	11601.500000	7.000000	6.000000	2000.000000	2004.000000	166.000000	712.250000	0.000000
max	190.000000	313.000000	215245.000000	10.000000	9.000000	2010.000000	2010.000000	1600.000000	5644.000000	1474.000000

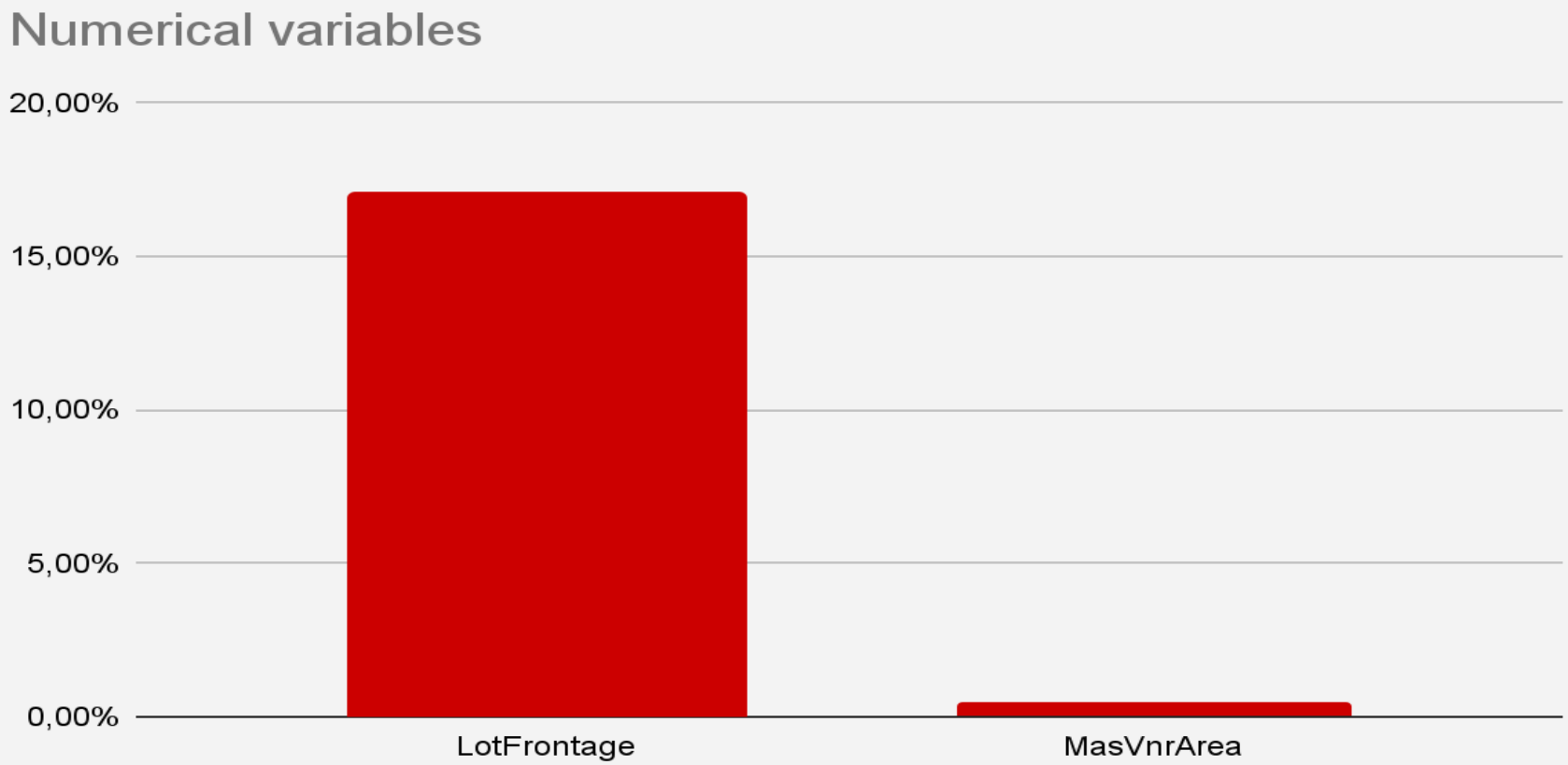
3.MISSING DATA IMPUTATION



3. Missing Data Imputation



3. Missing Data Imputation



3.Missing Data Imputation

Numerical

KNN Imputation
Mean, Median Imputation
Random Sample Imputation

Categorical

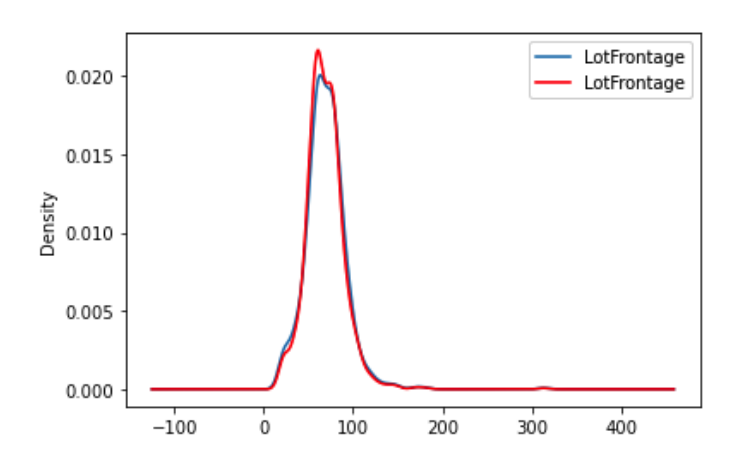
Frequent Category Imputation
Random Sample Imputation

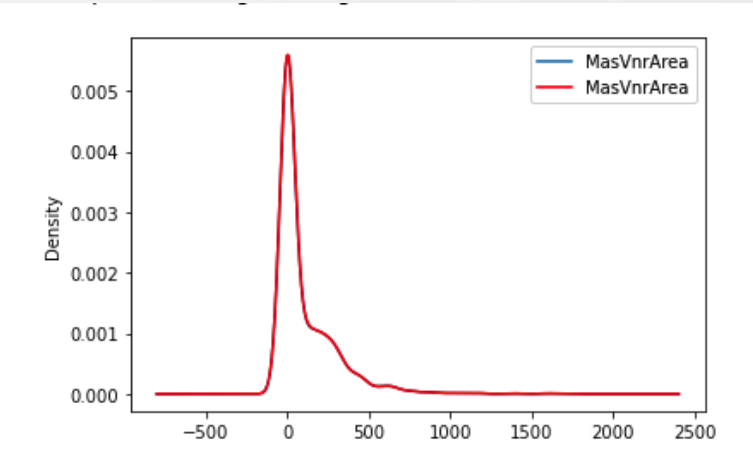
3.Missing Data Imputation

Numerical

KNN IMPUTATION

Distribution before and after imputation



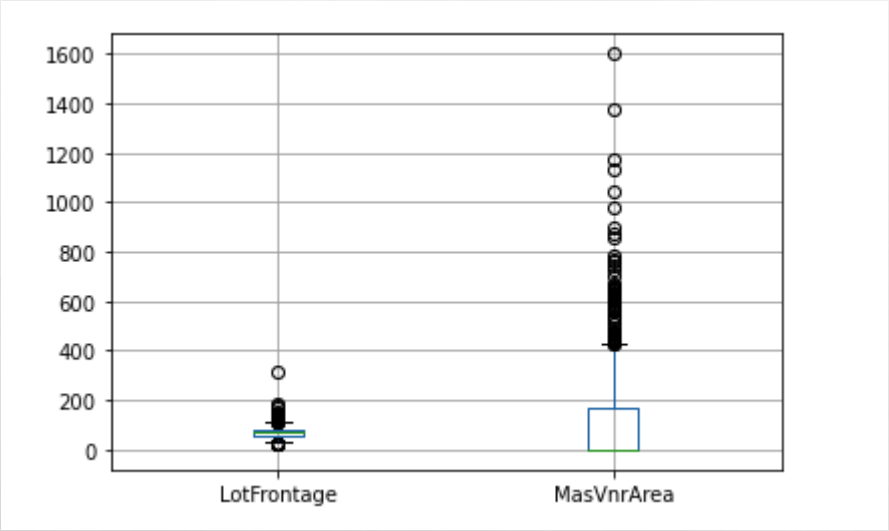


3.Missing Data Imputation

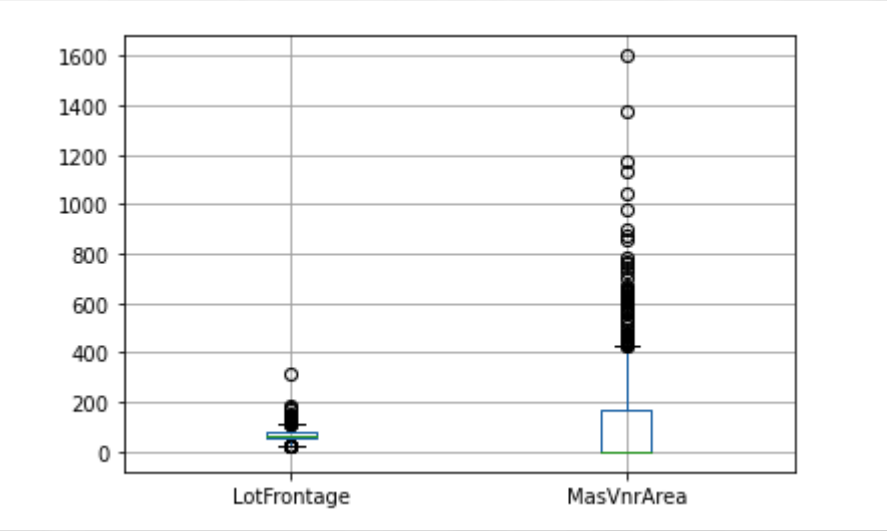
Numerical

KNN IMPUTATION

Before imputation



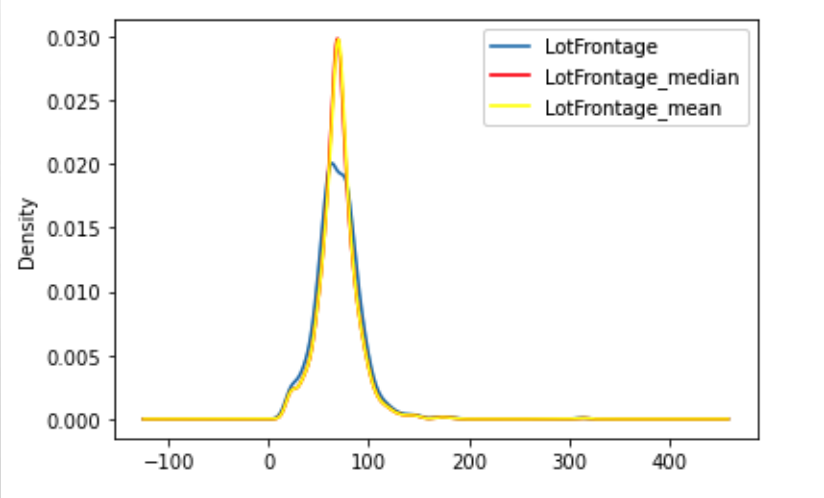
After imputation

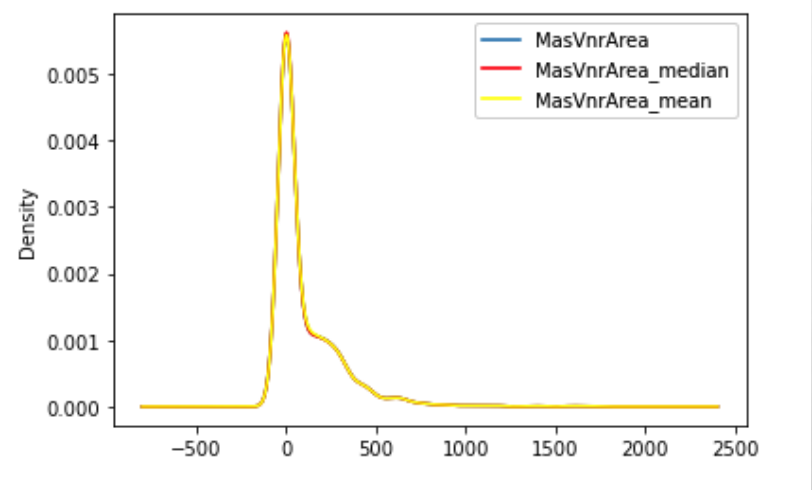


3.Missing Data Imputation

Numerical

MEAN, MEDIAN IMPUTATION

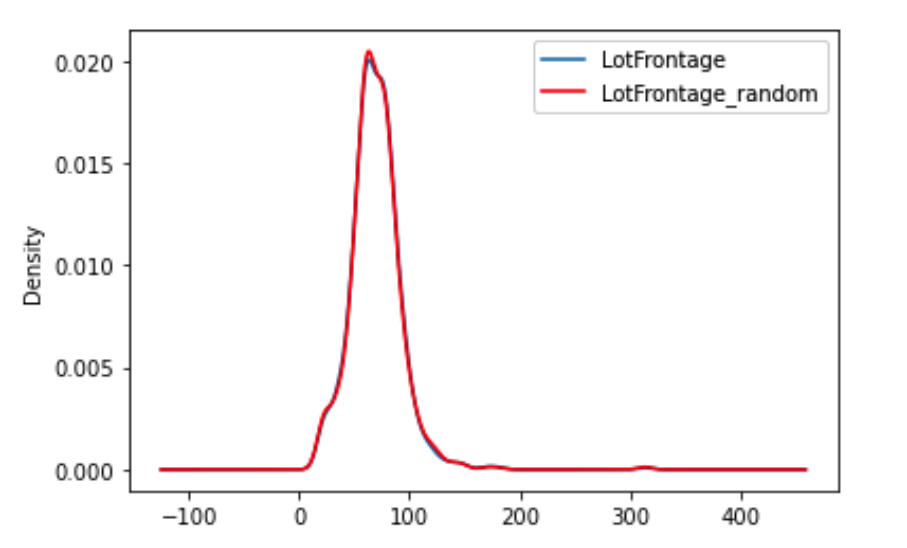


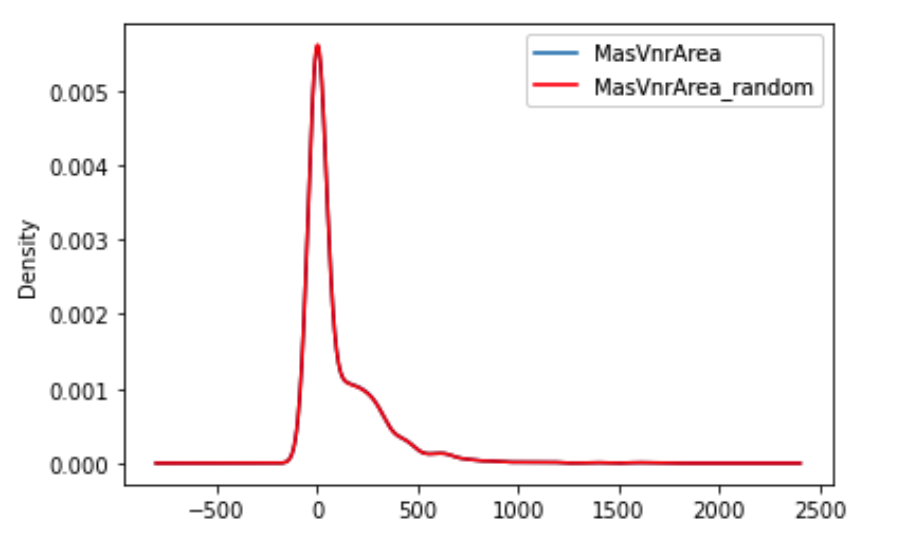


3.Missing Data Imputation

Numerical

RANDOM IMPUTATION

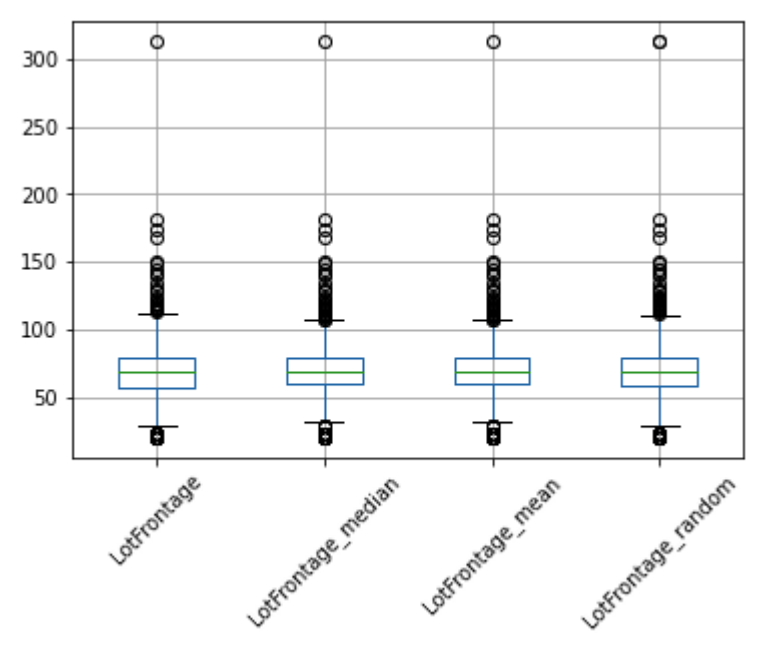


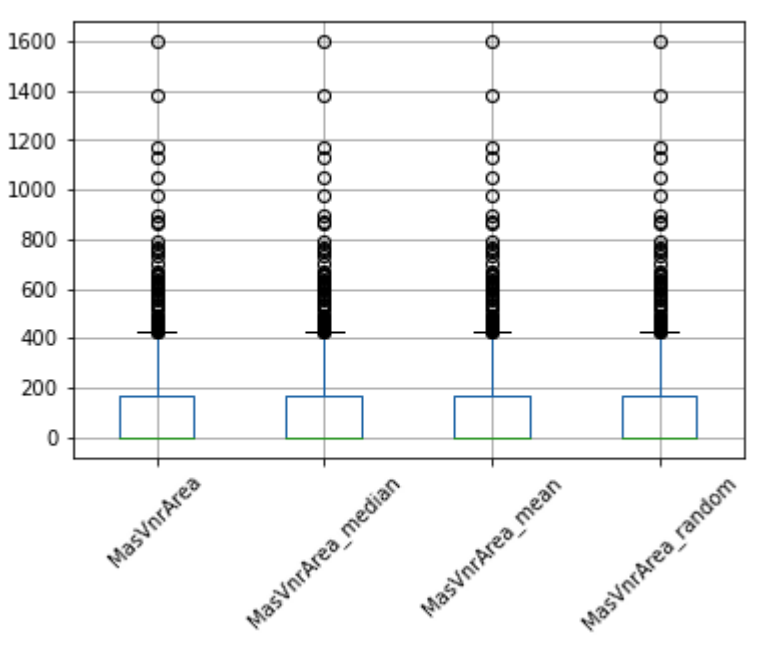


3.Missing Data Imputation

Numerical

Analysis of boxplots before and after imputation

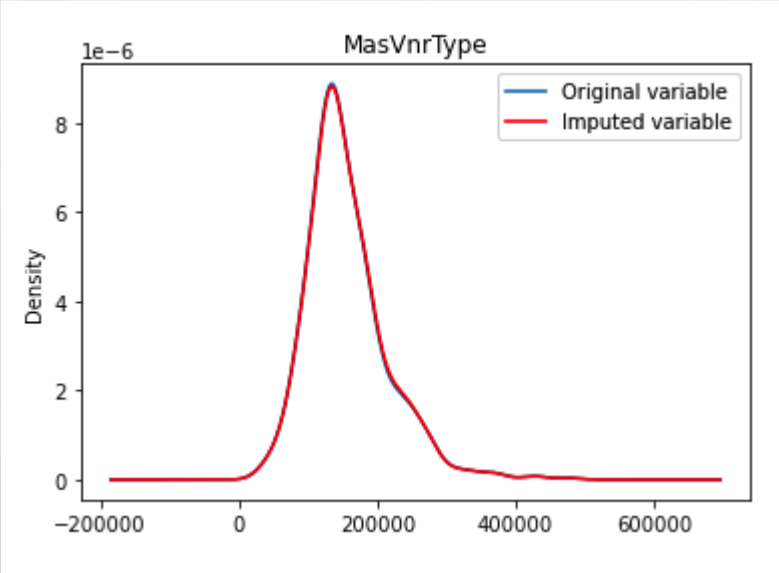


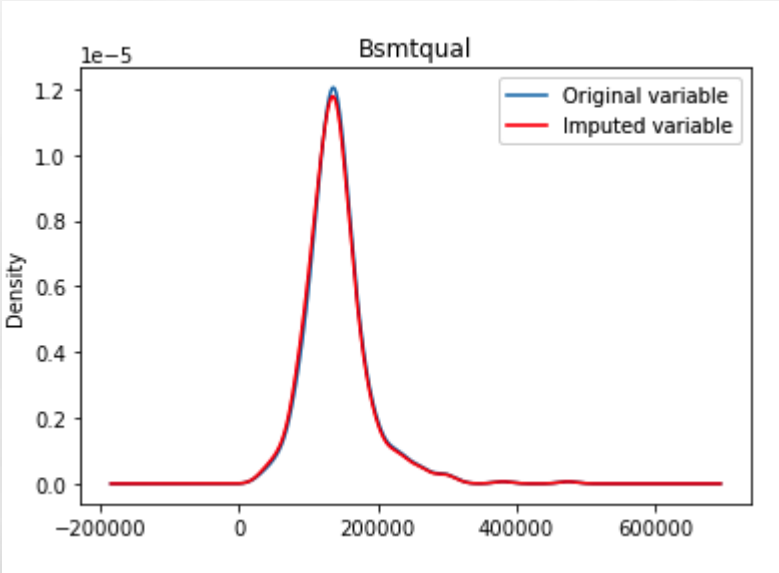


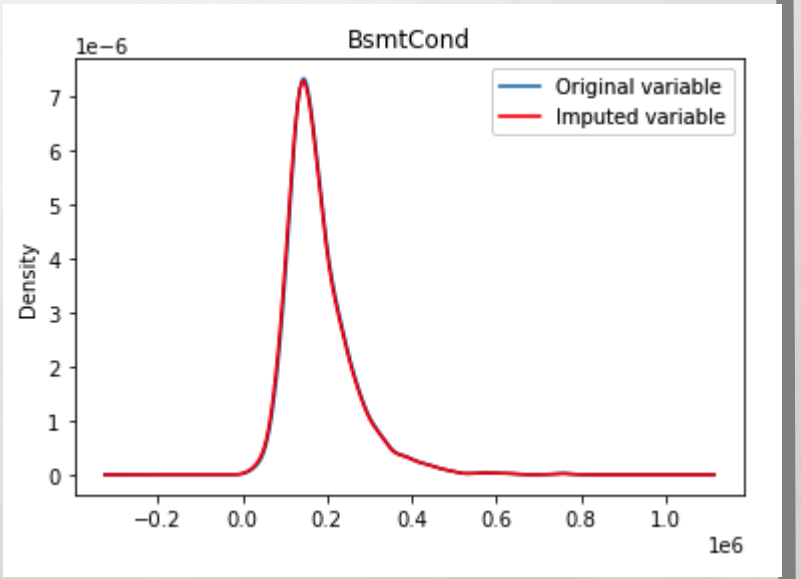
3.Missing Data Imputation

Categories

FREQUENT IMPUTATION



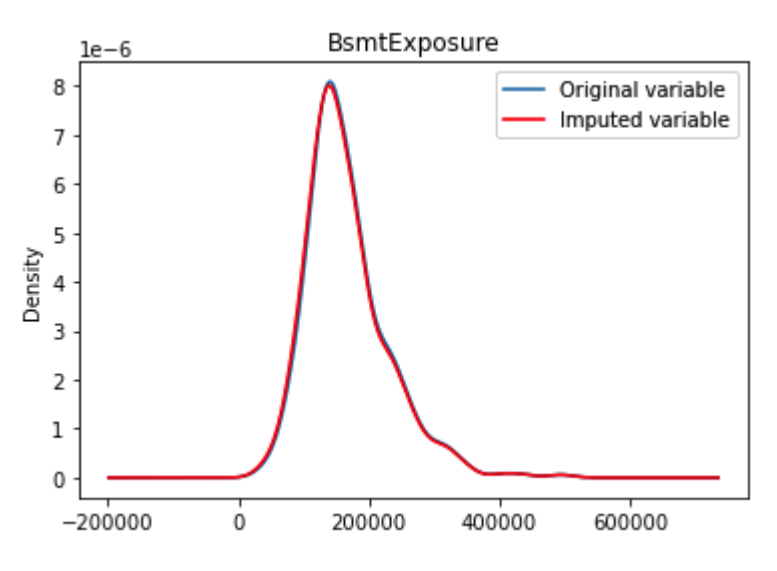


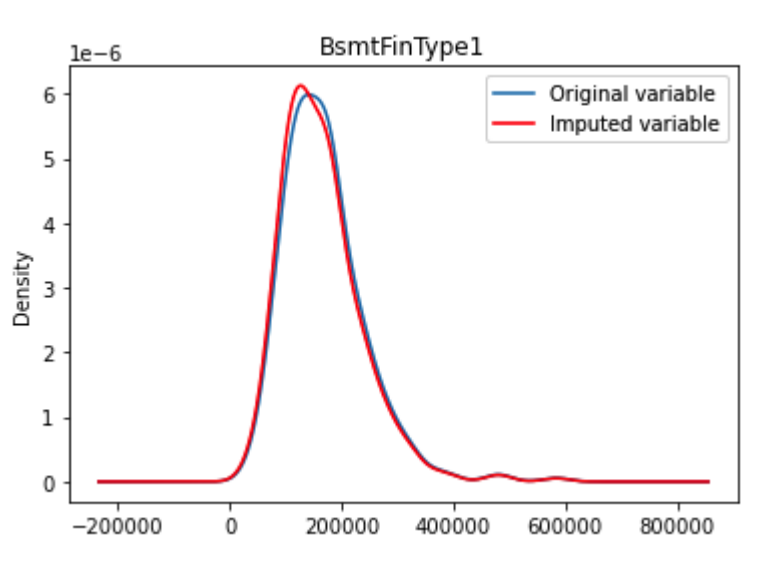


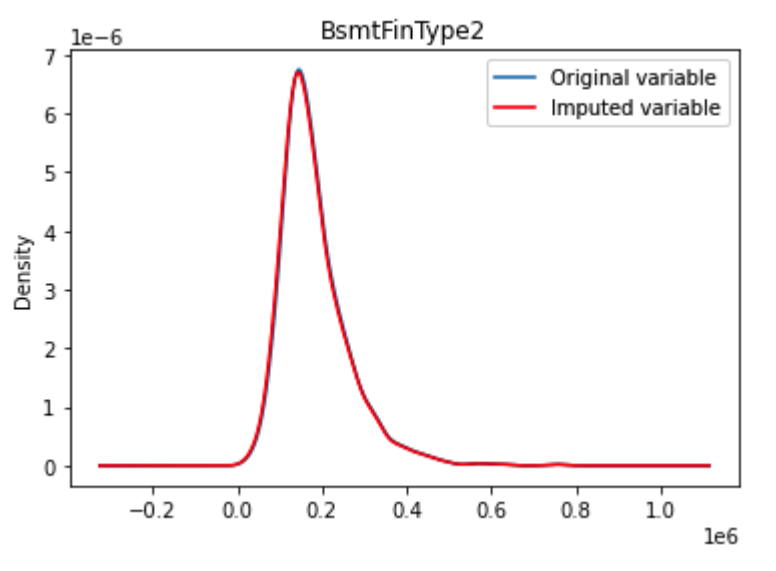
3.Missing Data Imputation

Categories

FREQUENT IMPUTATION



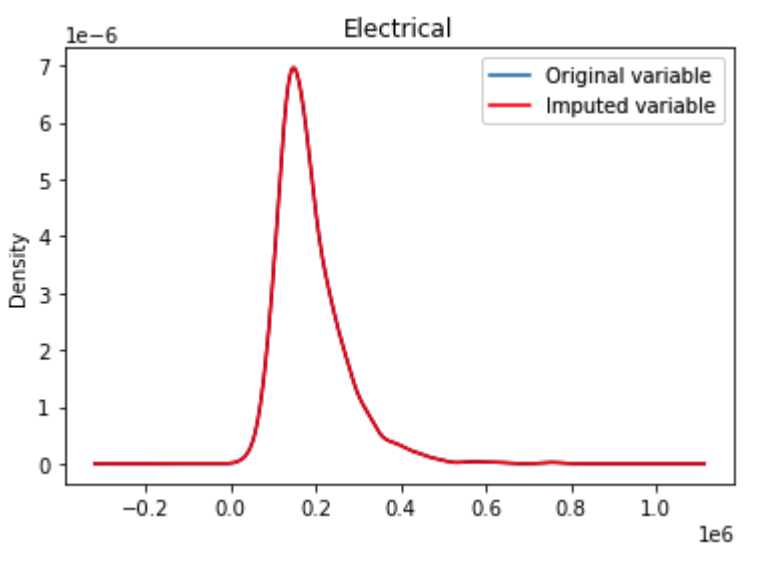


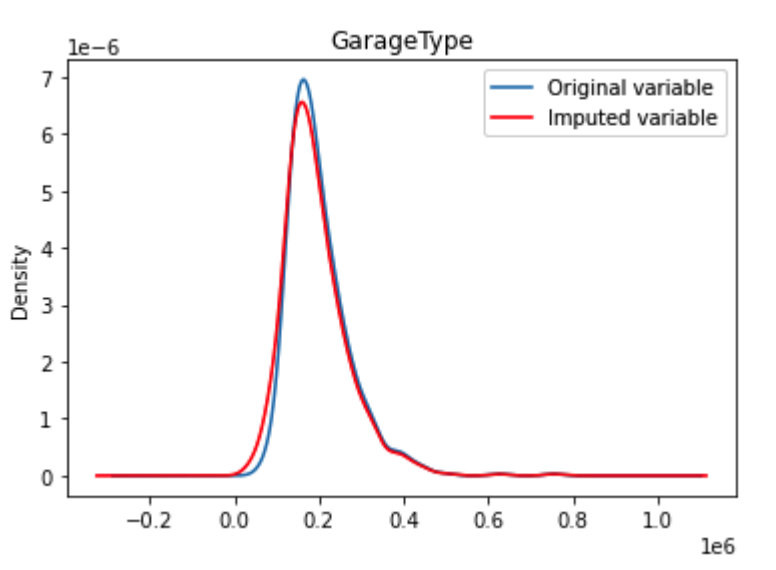


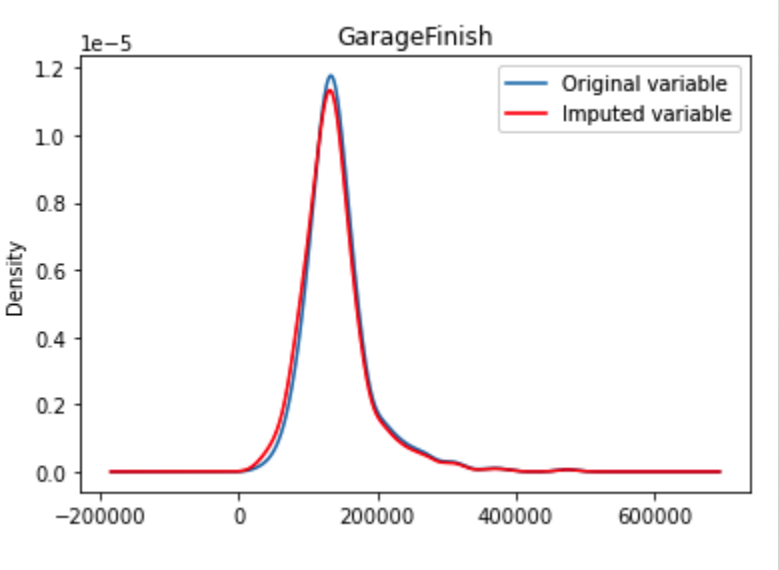
3.Missing Data Imputation

Categories

FREQUENT IMPUTATION



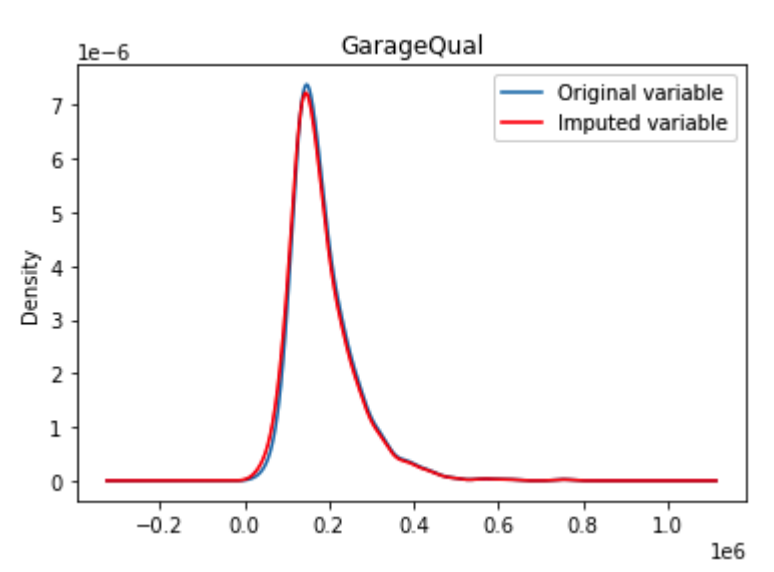


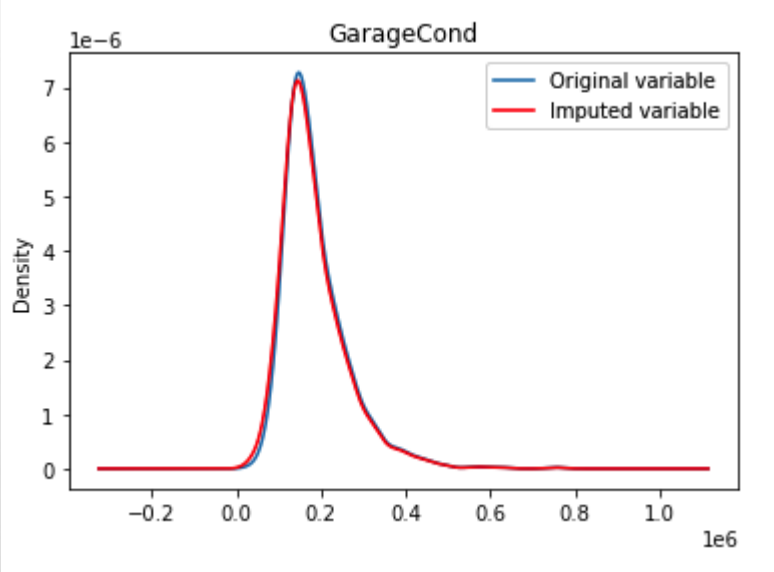


3.Missing Data Imputation

Categories

FREQUENT IMPUTATION

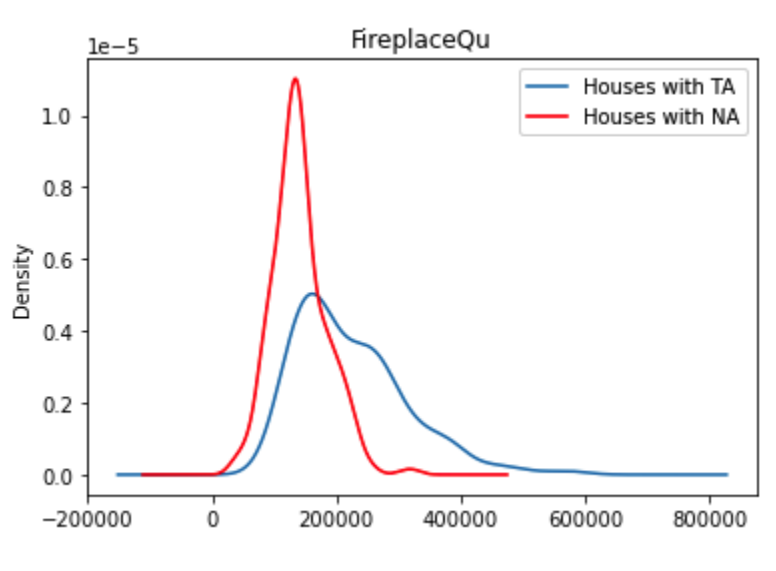


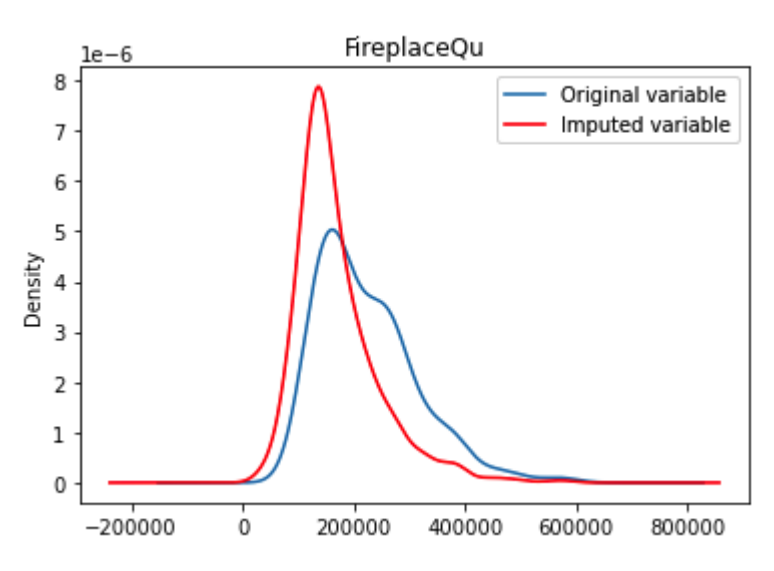


3.Missing Data Imputation

Categories

FREQUENT IMPUTATION



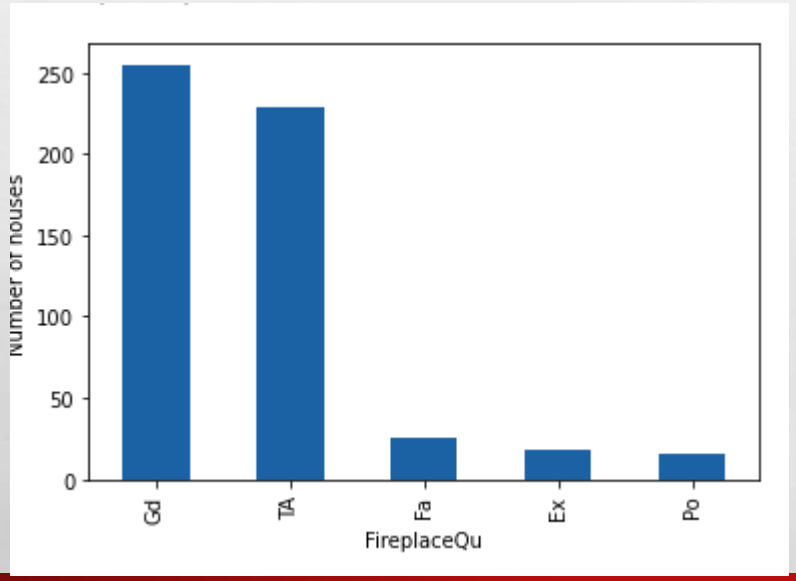


3.Missing Data Imputation

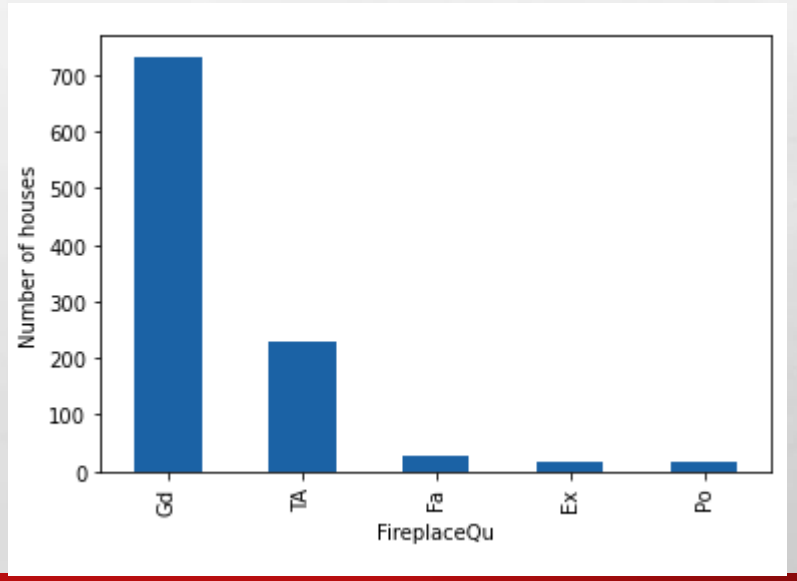
Categories

FREQUENT IMPUTATION

before imputation



after imputation



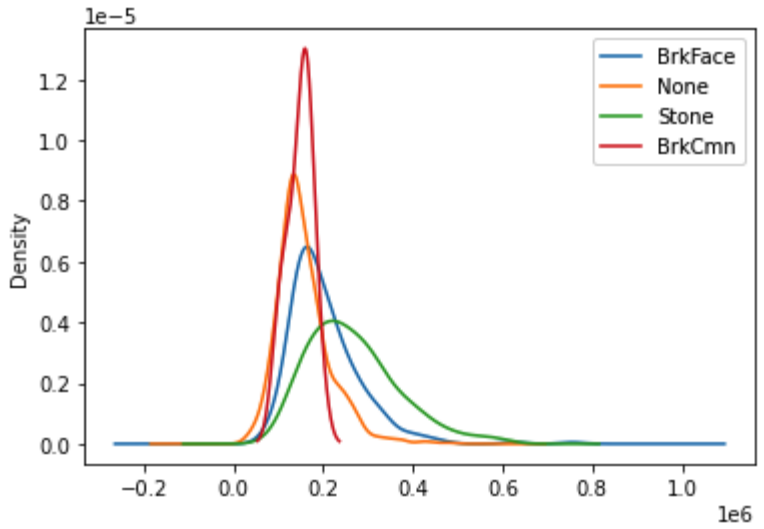
3.Missing Data Imputation

Categories

RANDOM SAMPLE IMPUTATION

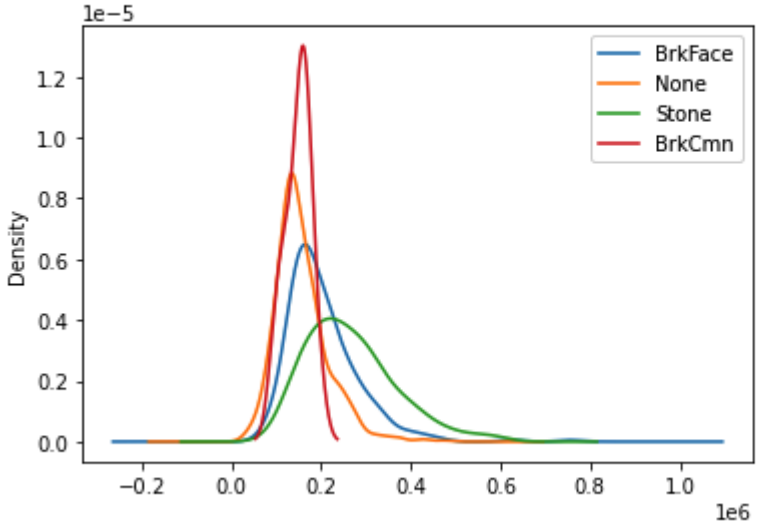
before imputation

```
automate_plot(x_train, 'MasVnrType', 'SalePrice')
```



after imputation

```
automate_plot(x_train, 'MasVnrType_imputed', 'SalePrice')
```



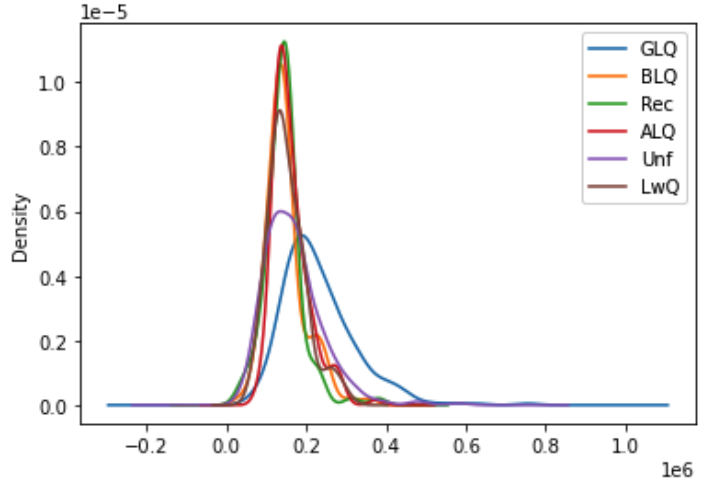
3.Missing Data Imputation

Categories

RANDOM SAMPLE IMPUTATION

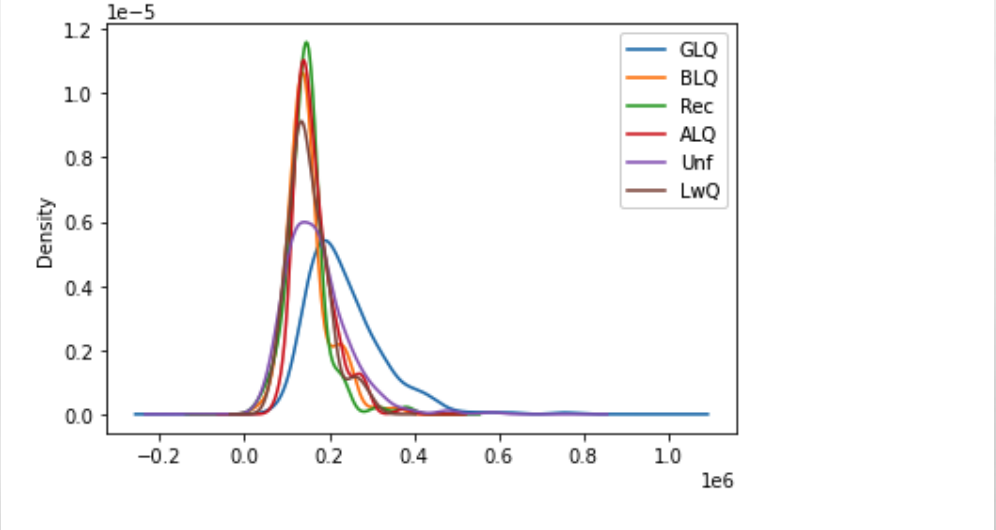
before imputation

```
automate_plot(x_train, 'BsmtFinType1', 'SalePrice')
```



after imputation

```
automate_plot(x_train, 'BsmtFinType1_imputed', 'SalePrice')
```



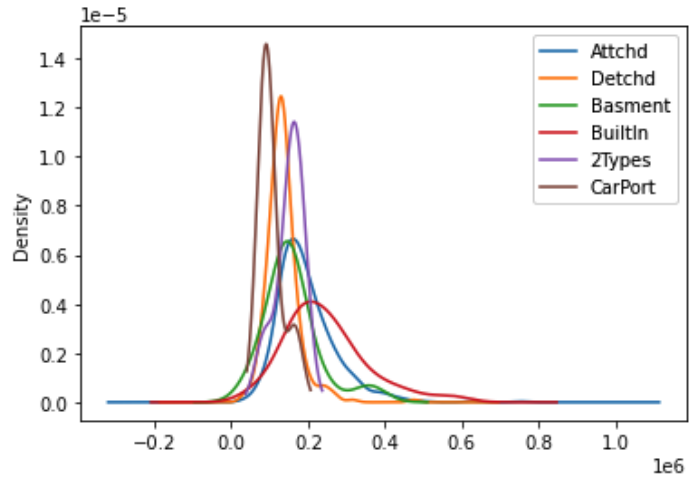
3.Missing Data Imputation

Categories

RANDOM SAMPLE IMPUTATION

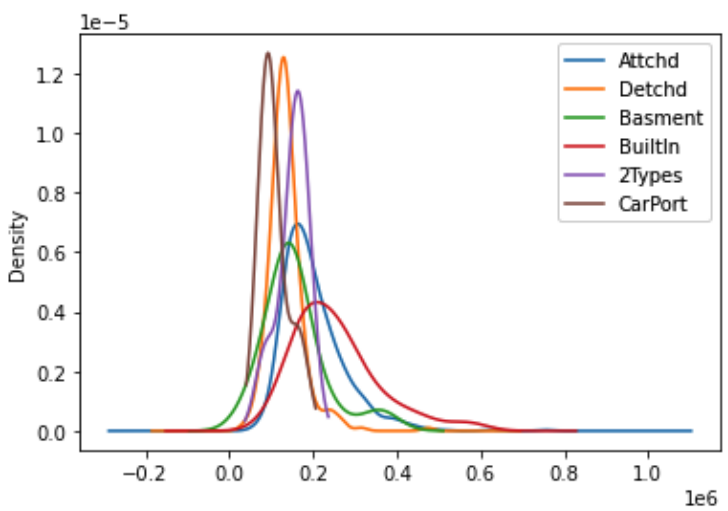
before imputation

```
automate_plot(x_train, 'GarageType', 'SalePrice')
```



after imputation

```
automate_plot(x_train, 'GarageType_imputed', 'SalePrice')
```



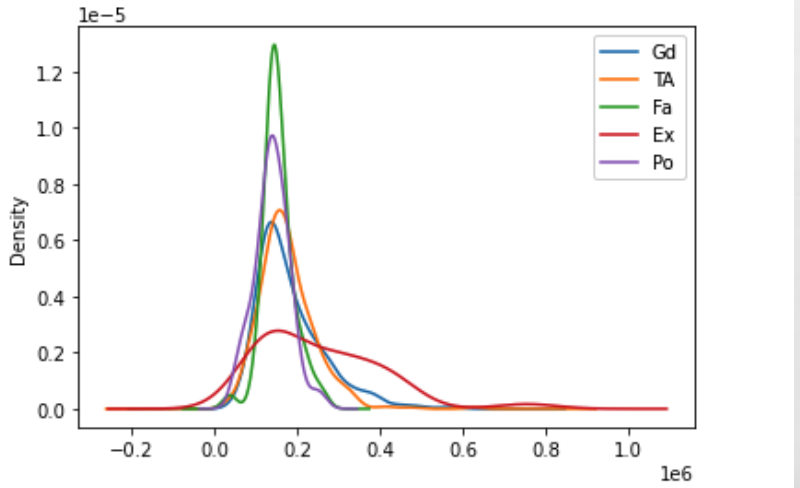
3.Missing Data Imputation

Categories

RANDOM SAMPLE IMPUTATION

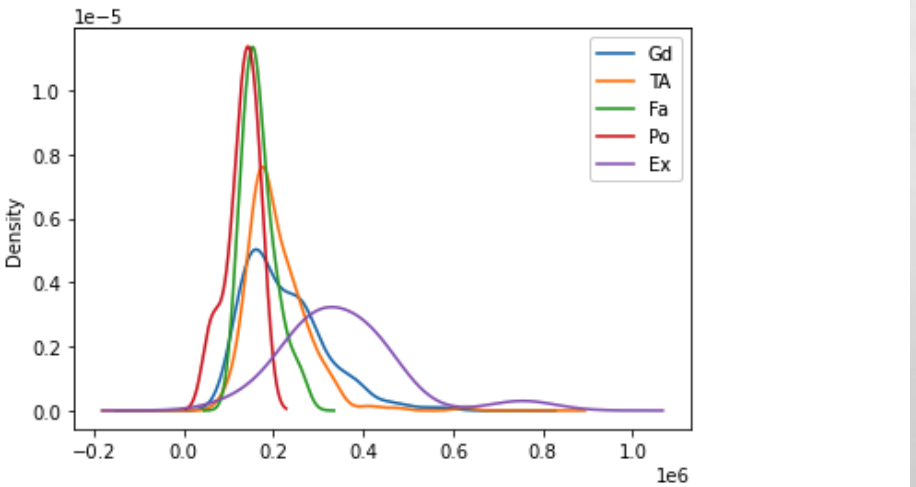
before imputation

```
automate_plot(x_train, 'FireplaceQu', 'SalePrice')
```



after imputation

```
automate_plot(x_train, 'FireplaceQu_imputed', 'SalePrice')
```



4. DISCRETISATION

- Equal - frequency Discretisation
- Equal - width Discretisation

5. ENCODING

- Rare Label Encoding
- One Hot Encoding
- Ordinal Encoding

FEATURE SELECTION

- **Constant :**

DROP CONSTANT FEATURES

- **Duplicated:**

DROP DUPLICATED FEATURES

- **Correlated:**

SMART CORRELATED SELECTION

6.PIPELINE

- **Pipeline:** Imputer_num,
  - Imputer_cat
  - Rare_label
  - Discretiser
  - Encoder
  - Scaler
  - Constant
  - Duplicated
  - Correlated
  - Algorithm

- **Grid Search:**
  - Imputer_num,
  - Imputer_cat
  - Discretiser
  - Encoder
  - Algorithm

Modeling Results

Svr grid search best params:

```
'SVR__gamma': 'scale',
'SVR__kernel': 'poly',
'discretiser__bins': 12,
'encoder__encoding_method': 'ordered',
'imputer_cat__imputation_method': 'missing',
'imputer_num__imputation_method': 'median'
```

- SVR:

SVR Train set: 97 %

SVR Test set: 86 %

THANK YOU !!!!

СПАСИБО ЗА ПРОСМОТР !!!