HomeWork3_Preprocessing

March 6, 2022

[623]:
```
import numpy as np
import pandas as pd
# ^^^ pyforest auto-imports - don't write above this line
import numpy as np
import pandas as pd
```

[624]:
```
train_data=pd.read_csv('G:
↪\Documents\ReferenceBooks\MachineLearning\Rohban\Homework\HW3\Train.csv')
test_data=pd.read_csv('G:
↪\Documents\ReferenceBooks\MachineLearning\Rohban\Homework\HW3\Test.csv')
```

[625]:
```
train_data.drop(["Id"], axis = 1, inplace=True)
test_data.drop(["Id"], axis = 1, inplace=True)
```

[626]:
```
train_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1460 entries, 0 to 1459
Data columns (total 80 columns):
MSSubClass       1460 non-null int64
MSZoning         1460 non-null object
LotFrontage      1201 non-null float64
LotArea          1460 non-null int64
Street           1460 non-null object
Alley            91 non-null object
LotShape         1460 non-null object
LandContour      1460 non-null object
Utilities        1460 non-null object
LotConfig        1460 non-null object
LandSlope        1460 non-null object
Neighborhood     1460 non-null object
Condition1       1460 non-null object
Condition2       1460 non-null object
BldgType         1460 non-null object
HouseStyle       1460 non-null object
OverallQual      1460 non-null int64
OverallCond      1460 non-null int64
YearBuilt        1460 non-null int64
```

```
YearRemodAdd     1460 non-null int64
RoofStyle        1460 non-null object
RoofMatl         1460 non-null object
Exterior1st      1460 non-null object
Exterior2nd      1460 non-null object
MasVnrType       1452 non-null object
MasVnrArea       1452 non-null float64
ExterQual        1460 non-null object
ExterCond        1460 non-null object
Foundation       1460 non-null object
BsmtQual         1423 non-null object
BsmtCond         1423 non-null object
BsmtExposure     1422 non-null object
BsmtFinType1     1423 non-null object
BsmtFinSF1       1460 non-null int64
BsmtFinType2     1422 non-null object
BsmtFinSF2       1460 non-null int64
BsmtUnfSF        1460 non-null int64
TotalBsmtSF      1460 non-null int64
Heating          1460 non-null object
HeatingQC        1460 non-null object
CentralAir       1460 non-null object
Electrical       1459 non-null object
1stFlrSF         1460 non-null int64
2ndFlrSF         1460 non-null int64
LowQualFinSF     1460 non-null int64
GrLivArea        1460 non-null int64
BsmtFullBath     1460 non-null int64
BsmtHalfBath     1460 non-null int64
FullBath         1460 non-null int64
HalfBath         1460 non-null int64
BedroomAbvGr     1460 non-null int64
KitchenAbvGr     1460 non-null int64
KitchenQual      1460 non-null object
TotRmsAbvGrd     1460 non-null int64
Functional       1460 non-null object
Fireplaces       1460 non-null int64
FireplaceQu      770 non-null object
GarageType       1379 non-null object
GarageYrBlt      1379 non-null float64
GarageFinish     1379 non-null object
GarageCars       1460 non-null int64
GarageArea       1460 non-null int64
GarageQual       1379 non-null object
GarageCond       1379 non-null object
PavedDrive       1460 non-null object
WoodDeckSF       1460 non-null int64
OpenPorchSF      1460 non-null int64
```

```
EnclosedPorch    1460 non-null int64
3SsnPorch        1460 non-null int64
ScreenPorch      1460 non-null int64
PoolArea         1460 non-null int64
PoolQC           7 non-null object
Fence            281 non-null object
MiscFeature      54 non-null object
MiscVal          1460 non-null int64
MoSold           1460 non-null int64
YrSold           1460 non-null int64
SaleType         1460 non-null object
SaleCondition    1460 non-null object
SalePrice        1460 non-null int64
dtypes: float64(3), int64(34), object(43)
memory usage: 912.6+ KB
```

[627]:
```
dict(train_data.dtypes)
```

[627]:
```
{'MSSubClass': dtype('int64'),
 'MSZoning': dtype('O'),
 'LotFrontage': dtype('float64'),
 'LotArea': dtype('int64'),
 'Street': dtype('O'),
 'Alley': dtype('O'),
 'LotShape': dtype('O'),
 'LandContour': dtype('O'),
 'Utilities': dtype('O'),
 'LotConfig': dtype('O'),
 'LandSlope': dtype('O'),
 'Neighborhood': dtype('O'),
 'Condition1': dtype('O'),
 'Condition2': dtype('O'),
 'BldgType': dtype('O'),
 'HouseStyle': dtype('O'),
 'OverallQual': dtype('int64'),
 'OverallCond': dtype('int64'),
 'YearBuilt': dtype('int64'),
 'YearRemodAdd': dtype('int64'),
 'RoofStyle': dtype('O'),
 'RoofMatl': dtype('O'),
 'Exterior1st': dtype('O'),
 'Exterior2nd': dtype('O'),
 'MasVnrType': dtype('O'),
 'MasVnrArea': dtype('float64'),
 'ExterQual': dtype('O'),
 'ExterCond': dtype('O'),
 'Foundation': dtype('O'),
```

```
'BsmtQual': dtype('O'),
'BsmtCond': dtype('O'),
'BsmtExposure': dtype('O'),
'BsmtFinType1': dtype('O'),
'BsmtFinSF1': dtype('int64'),
'BsmtFinType2': dtype('O'),
'BsmtFinSF2': dtype('int64'),
'BsmtUnfSF': dtype('int64'),
'TotalBsmtSF': dtype('int64'),
'Heating': dtype('O'),
'HeatingQC': dtype('O'),
'CentralAir': dtype('O'),
'Electrical': dtype('O'),
'1stFlrSF': dtype('int64'),
'2ndFlrSF': dtype('int64'),
'LowQualFinSF': dtype('int64'),
'GrLivArea': dtype('int64'),
'BsmtFullBath': dtype('int64'),
'BsmtHalfBath': dtype('int64'),
'FullBath': dtype('int64'),
'HalfBath': dtype('int64'),
'BedroomAbvGr': dtype('int64'),
'KitchenAbvGr': dtype('int64'),
'KitchenQual': dtype('O'),
'TotRmsAbvGrd': dtype('int64'),
'Functional': dtype('O'),
'Fireplaces': dtype('int64'),
'FireplaceQu': dtype('O'),
'GarageType': dtype('O'),
'GarageYrBlt': dtype('float64'),
'GarageFinish': dtype('O'),
'GarageCars': dtype('int64'),
'GarageArea': dtype('int64'),
'GarageQual': dtype('O'),
'GarageCond': dtype('O'),
'PavedDrive': dtype('O'),
'WoodDeckSF': dtype('int64'),
'OpenPorchSF': dtype('int64'),
'EnclosedPorch': dtype('int64'),
'3SsnPorch': dtype('int64'),
'ScreenPorch': dtype('int64'),
'PoolArea': dtype('int64'),
'PoolQC': dtype('O'),
'Fence': dtype('O'),
'MiscFeature': dtype('O'),
'MiscVal': dtype('int64'),
'MoSold': dtype('int64'),
```

```
 'YrSold': dtype('int64'),
 'SaleType': dtype('O'),
 'SaleCondition': dtype('O'),
 'SalePrice': dtype('int64')}
```

0.1 Fix all numerical columns and null values in categoricals

[628]:
```
#fixing all numeric values
for column in train_data.columns:
    if train_data[column].dtype is np.dtype('O'):
        train_data[column]=np.where(train_data[column].
↪isnull(),"_Unknown_",train_data[column])
    else:
        train_data[column]=np.where(train_data[column].
↪isnull(),train_data[column].mean(),train_data[column])
```

[629]:
```
#fixing all numeric values
for column in test_data.columns:
    if test_data[column].dtype is np.dtype('O'):
        test_data[column]=np.where(test_data[column].
↪isnull(),"_Unknown_",test_data[column])
    else:
        test_data[column]=np.where(test_data[column].
↪isnull(),train_data[column].mean(),test_data[column])
```

[630]:
```
train_data.LandSlope.value_counts(dropna=False)
```

[630]:
```
Gtl    1382
Mod      65
Sev      13
Name: LandSlope, dtype: int64
```

[631]:
```
commons=list(set(train_data.columns) & set(test_data.columns))
len(commons)
```

[631]:
```
79
```

[632]:
```
test_data = test_data[commons]
commons.append('SalePrice')
train_data = train_data[commons]
commons
```

[632]:
```
['RoofStyle',
 'BsmtExposure',
 'BsmtQual',
```

```
'Condition2',
'HalfBath',
'EnclosedPorch',
'1stFlrSF',
'MasVnrArea',
'MasVnrType',
'FireplaceQu',
'LotConfig',
'CentralAir',
'RoofMatl',
'ScreenPorch',
'WoodDeckSF',
'PoolArea',
'KitchenAbvGr',
'BldgType',
'SaleType',
'YrSold',
'OverallQual',
'BedroomAbvGr',
'GarageYrBlt',
'FullBath',
'LandContour',
'BsmtCond',
'BsmtHalfBath',
'2ndFlrSF',
'Fence',
'TotalBsmtSF',
'YearBuilt',
'GarageQual',
'ExterCond',
'ExterQual',
'BsmtFinType1',
'Fireplaces',
'PavedDrive',
'BsmtFinSF2',
'LotArea',
'MSSubClass',
'MSZoning',
'Alley',
'BsmtFullBath',
'BsmtUnfSF',
'3SsnPorch',
'Heating',
'YearRemodAdd',
'TotRmsAbvGrd',
'GarageCond',
'Foundation',
```

```
'LotShape',
'KitchenQual',
'Street',
'HeatingQC',
'LandSlope',
'LowQualFinSF',
'BsmtFinSF1',
'GarageFinish',
'Electrical',
'Exterior2nd',
'HouseStyle',
'MiscFeature',
'GarageType',
'Neighborhood',
'Functional',
'OpenPorchSF',
'BsmtFinType2',
'OverallCond',
'Exterior1st',
'GarageCars',
'GarageArea',
'Utilities',
'MoSold',
'PoolQC',
'GrLivArea',
'LotFrontage',
'SaleCondition',
'Condition1',
'MiscVal',
'SalePrice']
```

0.2 Nominal Variables

[633]:
```
def mapping(orderedlist):
    i=0
    ordered = {}
    for item in orderedlist:
        ordered[str(item)] = i
        i = i+1
    return ordered
```

[634]:
```
def convert(data,column,mappingList):
    if column not in data.columns:
        return
    if data[column].dtype is np.dtype('O'):
        data[column] = data[column].replace(mapping(mappingList))
```

```
        if any(data[column] == '_Unknown_'):
            data[str(column)+'_IsNull'] = np.where(data[column] ==␣
↪'_Unknown_',True,False)
            data[column] = np.where(data[column]=='_Unknown_',-1,data[column])
            data[column] = np.where(data[column]==-1,data[column].
↪mean(),data[column])
            data[str(column)+'_IsNull'] = pd.
↪to_numeric(arg=data[column+'_IsNull'])
```

[635]:
```
convert(train_data,'LotShape',["Reg", "IR1", "IR2","IR3"])
convert(train_data,'LandContour',["Lvl","Bnk","HLS","Low"])
convert(train_data,'Utilities',["ELO","NoSeWa","NoSewr","AllPub"])
convert(train_data,'LandSlope',["Sev","Mod","Gtl"])
convert(train_data,'ExterQual',["Ex","Gd","TA",'Fa','Po'])
convert(train_data,'ExterCond',["Ex","Gd","TA",'Fa','Po'])
convert(train_data,'BsmtQual',["Ex","Gd","TA",'Fa','Po'])
convert(train_data,'BsmtCond',["Ex","Gd","TA",'Fa','Po'])
convert(train_data,'BsmtExposure',["Gd","Av","Mn",'No','NA'])
convert(train_data,'BsmtFinType1',["GLQ","ALQ","BLQ",'Rec','LwQ','Unf','NA'])
convert(train_data,'BsmtFinType2',["GLQ","ALQ","BLQ",'Rec','LwQ','Unf','NA'])
convert(train_data,'HeatingQC',["Ex","Gd","TA",'Fa','Po'])
convert(train_data,'CentralAir',['No','Yes'])
convert(train_data,'KitchenQual',["Ex","Gd","TA",'Fa','Po'])
convert(train_data,'Functional',["Typ","Min1","Min2",'Mod','Maj1','Maj2','Sev','Sal'])
convert(train_data,'FireplaceQu',["Ex","Gd","TA",'Fa','Po','NA'])
convert(train_data,'GarageFinish',["Fin","RFn","Unf",'NA'])
convert(train_data,'GarageQual',["Ex","Gd","TA",'Fa','Po','NA'])
convert(train_data,'GarageCond',["Ex","Gd","TA",'Fa','Po','NA'])
convert(train_data,'PavedDrive',["Y","P","N"])
convert(train_data,'PoolQC',["Ex","Gd","TA",'Fa','NA'])
convert(train_data,'Fence',["GdPrv","MnPrv","GdWo",'MnWw','NA'])
convert(train_data,'CentralAir',['Y','N'])
train_data
```

```
C:\ProgramData\Anaconda3\lib\site-packages\pandas\core\ops\__init__.py:1115:
FutureWarning: elementwise comparison failed; returning scalar instead, but in
the future will perform elementwise comparison
  result = method(y)
```

[635]:

	RoofStyle	BsmtExposure	BsmtQual	Condition2	HalfBath	EnclosedPorch	\
0	Gable	3	1	Norm	1.0	0.0	
1	Gable	0	1	Norm	0.0	0.0	
2	Gable	2	1	Norm	1.0	0.0	
3	Gable	3	2	Norm	0.0	272.0	
4	Gable	1	1	Norm	1.0	0.0	
…	…	…	…	…	…	…	
1455	Gable	3	1	Norm	1.0	0.0	

```
1456     Gable            3        1       Norm       0.0            0.0
1457     Gable            3        2       Norm       0.0            0.0
1458       Hip            2        2       Norm       0.0          112.0
1459     Gable            3        2       Norm       1.0            0.0

      1stFlrSF  MasVnrArea MasVnrType FireplaceQu  … BsmtCond_IsNull  \
0        856.0       196.0    BrkFace    0.339041  …           False
1       1262.0         0.0       None           2  …           False
2        920.0       162.0    BrkFace           2  …           False
3        961.0         0.0       None           1  …           False
4       1145.0       350.0    BrkFace           2  …           False
…          …          …         …          …  …             …
1455     953.0         0.0       None           2  …           False
1456    2073.0       119.0      Stone           2  …           False
1457    1188.0         0.0       None           1  …           False
1458    1078.0         0.0       None    0.339041  …           False
1459    1256.0         0.0       None    0.339041  …           False

      BsmtExposure_IsNull BsmtFinType1_IsNull  BsmtFinType2_IsNull  \
0                   False               False                False
1                   False               False                False
2                   False               False                False
3                   False               False                False
4                   False               False                False
…                   …                   …                    …
1455                False               False                False
1456                False               False                False
1457                False               False                False
1458                False               False                False
1459                False               False                False

      FireplaceQu_IsNull  GarageFinish_IsNull  GarageQual_IsNull  \
0                   True               False              False
1                  False               False              False
2                  False               False              False
3                  False               False              False
4                  False               False              False
…                  …                   …                  …
1455               False               False              False
1456               False               False              False
1457               False               False              False
1458                True               False              False
1459                True               False              False

     GarageCond_IsNull PoolQC_IsNull  Fence_IsNull
0                False          True          True
1                False          True          True
```

```
2                False          True          True
3                False          True          True
4                False          True          True
…                …              …             …
1455             False          True          True
1456             False          True         False
1457             False          True         False
1458             False          True          True
1459             False          True          True

[1460 rows x 91 columns]
```

[636]:
```
convert(test_data,'LotShape',["Reg", "IR1", "IR2","IR3"])
convert(test_data,'LandContour',["Lvl","Bnk","HLS","Low"])
convert(test_data,'Utilities',["ELO","NoSeWa","NoSewr","AllPub"])
convert(test_data,'LandSlope',["Sev","Mod","Gtl"])
convert(test_data,'ExterQual',["Ex","Gd","TA",'Fa','Po'])
convert(test_data,'ExterCond',["Ex","Gd","TA",'Fa','Po'])
convert(test_data,'BsmtQual',["Ex","Gd","TA",'Fa','Po'])
convert(test_data,'BsmtCond',["Ex","Gd","TA",'Fa','Po'])
convert(test_data,'BsmtExposure',["Gd","Av","Mn",'No','NA'])
convert(test_data,'BsmtFinType1',["GLQ","ALQ","BLQ",'Rec','LwQ','Unf','NA'])
convert(test_data,'BsmtFinType2',["GLQ","ALQ","BLQ",'Rec','LwQ','Unf','NA'])
convert(test_data,'HeatingQC',["Ex","Gd","TA",'Fa','Po'])
convert(test_data,'CentralAir',['No','Yes'])
convert(test_data,'KitchenQual',["Ex","Gd","TA",'Fa','Po'])
convert(test_data,'Functional',["Typ","Min1","Min2",'Mod','Maj1','Maj2','Sev','Sal'])
convert(test_data,'FireplaceQu',["Ex","Gd","TA",'Fa','Po','NA'])
convert(test_data,'GarageFinish',["Fin","RFn","Unf",'NA'])
convert(test_data,'GarageQual',["Ex","Gd","TA",'Fa','Po','NA'])
convert(test_data,'GarageCond',["Ex","Gd","TA",'Fa','Po','NA'])
convert(test_data,'PavedDrive',["Y","P","N"])
convert(test_data,'PoolQC',["Ex","Gd","TA",'Fa','NA'])
convert(test_data,'Fence',["GdPrv","MnPrv","GdWo",'MnWw','NA'])
convert(test_data,'CentralAir',['Y','N'])
test_data
```

[636]:
```
     RoofStyle BsmtExposure BsmtQual Condition2  HalfBath  EnclosedPorch  \
0        Gable            3        2       Norm       0.0            0.0
1          Hip            3        2       Norm       1.0            0.0
2        Gable            3        1       Norm       1.0            0.0
3        Gable            3        2       Norm       1.0            0.0
4        Gable            3        1       Norm       0.0            0.0
…        …          …      …        …       …          …              …
1454     Gable            3        2       Norm       1.0            0.0
1455     Gable            3        2       Norm       1.0            0.0
1456     Gable            3        2       Norm       0.0            0.0
```

```
1457     Gable            1        1       Norm       0.0            0.0
1458     Gable            1        1       Norm       1.0            0.0

      1stFlrSF  MasVnrArea MasVnrType FireplaceQu  … BsmtFinType1_IsNull  \
0        896.0         0.0       None    0.287183  …               False
1       1329.0       108.0    BrkFace    0.287183  …               False
2        928.0         0.0       None           2  …               False
3        926.0        20.0    BrkFace           1  …               False
4       1280.0         0.0       None    0.287183  …               False
…          …          …         …          …  …                 …
1454     546.0         0.0       None    0.287183  …               False
1455     546.0         0.0       None    0.287183  …               False
1456    1224.0         0.0       None           2  …               False
1457     970.0         0.0       None    0.287183  …               False
1458     996.0        94.0    BrkFace           2  …               False

      BsmtFinType2_IsNull KitchenQual_IsNull  Functional_IsNull  \
0                   False              False              False
1                   False              False              False
2                   False              False              False
3                   False              False              False
4                   False              False              False
…                   …                …                  …
1454                False              False              False
1455                False              False              False
1456                False              False              False
1457                False              False              False
1458                False              False              False

      FireplaceQu_IsNull  GarageFinish_IsNull  GarageQual_IsNull  \
0                   True                False              False
1                   True                False              False
2                  False                False              False
3                  False                False              False
4                   True                False              False
…                  …                  …                  …
1454                True                 True               True
1455                True                False              False
1456               False                False              False
1457                True                 True               True
1458               False                False              False

     GarageCond_IsNull PoolQC_IsNull  Fence_IsNull
0                False          True         False
1                False          True          True
2                False          True         False
3                False          True          True
```

```
4                False          True          True
…                …          …          …
1454              True          True          True
1455             False          True          True
1456             False          True          True
1457              True          True         False
1458             False          True          True

[1459 rows x 93 columns]
```

[ ]:

0.3 Categorical Variables

0.3.1 Check to see if has null values

[637]:
```
train_data['SaleCondition'].value_counts(dropna=False)
```

[637]:
```
Normal     1198
Partial     125
Abnorml     101
Family       20
Alloca       12
AdjLand       4
Name: SaleCondition, dtype: int64
```

0.3.2 encoding categorical with one-hot-encoding

[ ]:

[638]:
```
train_data_new=pd.
↪get_dummies(data=train_data,columns=['MSSubClass','MSZoning','Street',
                                      ␣
↪'LotConfig','Neighborhood','Condition1',
                                     ␣
↪'Condition2','BldgType','HouseStyle','RoofStyle','RoofMatl','Exterior1st',
                                    "Alley","MiscFeature" ,␣
↪'Exterior2nd',
                                       'MasVnrType',
                                   'Foundation','Heating',
                                     'Electrical',
                                    'GarageType',
                                     ␣
↪'SaleType','SaleCondition'])
```

```
train_data_new
```

[638]:

	BsmtExposure	BsmtQual	HalfBath	EnclosedPorch	1stFlrSF	MasVnrArea
0	3	1	1.0	0.0	856.0	196.0
1	0	1	0.0	0.0	1262.0	0.0
2	2	1	1.0	0.0	920.0	162.0
3	3	2	0.0	272.0	961.0	0.0
4	1	1	1.0	0.0	1145.0	350.0
…	…	…	…	…	…	…
1455	3	1	1.0	0.0	953.0	0.0
1456	3	1	0.0	0.0	2073.0	119.0
1457	3	2	0.0	0.0	1188.0	0.0
1458	2	2	0.0	112.0	1078.0	0.0
1459	3	2	1.0	0.0	1256.0	0.0

	FireplaceQu	CentralAir	ScreenPorch	WoodDeckSF	…	SaleType_ConLw
0	0.339041	0	0.0	0.0	…	0
1	2	0	0.0	298.0	…	0
2	2	0	0.0	0.0	…	0
3	1	0	0.0	0.0	…	0
4	2	0	0.0	192	…	0

	SaleType_ConLw
0	0
1	0
2	0
3	0
4	0

```
…                       …                       …                       …
1455                    0                       0                       0
1456                    0                       0                       0
1457                    0                       0                       0
1458                    0                       0                       0
1459                    0                       0                       0

     SaleCondition_Normal  SaleCondition_Partial
0                       1                       0
1                       1                       0
2                       1                       0
3                       0                       0
4                       1                       0
…                     …                       …
1455                    1                       0
1456                    1                       0
1457                    1                       0
1458                    1                       0
1459                    1                       0

[1460 rows x 249 columns]
```

[639]:
```
test_data_new=pd.
 ↪get_dummies(data=test_data,columns=['MSSubClass','MSZoning','Street',
                                                                  ␣
 ↪'LotConfig','Neighborhood','Condition1',
                                                                  ␣
 ↪'Condition2','BldgType','HouseStyle','RoofStyle','RoofMatl','Exterior1st',
                                                                  ␣
 ↪'Exterior2nd',"Alley","MiscFeature",
                                                     'MasVnrType',
                                                   'Foundation','Heating',
                                                     'Electrical',
                                                    'GarageType',
                                                                  ␣
 ↪'SaleType','SaleCondition'])
test_data_new
```

[639]:
```
     BsmtExposure BsmtQual  HalfBath  EnclosedPorch  1stFlrSF  MasVnrArea  \
0               3        2       0.0            0.0     896.0         0.0
1               3        2       1.0            0.0    1329.0       108.0
2               3        1       1.0            0.0     928.0         0.0
3               3        2       1.0            0.0     926.0        20.0
4               3        1       0.0            0.0    1280.0         0.0
…             …      …       …              …       …          …
1454            3        2       1.0            0.0     546.0         0.0
1455            3        2       1.0            0.0     546.0         0.0
```

```
1456            3        2       0.0            0.0    1224.0         0.0
1457            1        1       0.0            0.0     970.0         0.0
1458            1        1       1.0            0.0     996.0        94.0

     FireplaceQu  CentralAir  ScreenPorch  WoodDeckSF  …  SaleType_New  \
0       0.287183           0        120.0       140.0  …             0
1       0.287183           0          0.0       393.0  …             0
2              2           0          0.0       212.0  …             0
3              1           0          0.0       360.0  …             0
4       0.287183           0        144.0         0.0  …             0
…          …          …           …          …  …           …
1454    0.287183           0          0.0         0.0  …             0
1455    0.287183           0          0.0         0.0  …             0
1456           2           0          0.0       474.0  …             0
1457    0.287183           0          0.0        80.0  …             0
1458           2           0          0.0       190.0  …             0

      SaleType_Oth  SaleType_WD  SaleType__Unknown_  SaleCondition_Abnorml  \
0                0            1                   0                      0
1                0            1                   0                      0
2                0            1                   0                      0
3                0            1                   0                      0
4                0            1                   0                      0
…            …          …                 …                    …
1454             0            1                   0                      0
1455             0            1                   0                      1
1456             0            1                   0                      1
1457             0            1                   0                      0
1458             0            1                   0                      0

      SaleCondition_AdjLand  SaleCondition_Alloca  SaleCondition_Family  \
0                         0                    0                     0
1                         0                    0                     0
2                         0                    0                     0
3                         0                    0                     0
4                         0                    0                     0
…                     …                 …                   …
1454                      0                    0                     0
1455                      0                    0                     0
1456                      0                    0                     0
1457                      0                    0                     0
1458                      0                    0                     0

     SaleCondition_Normal  SaleCondition_Partial
0                       1                      0
1                       1                      0
2                       1                      0
```

```
3                       1                       0
4                       1                       0
…                       …                       …
1454                    1                       0
1455                    0                       0
1456                    0                       0
1457                    1                       0
1458                    1                       0

[1459 rows x 240 columns]
```

[640]:
```
train_data_new.isnull().sum().sum()
```

[640]:
```
0
```

[641]:
```
columns = list(set(train_data_new.columns)& set(test_data_new.columns))

y = train_data_new.SalePrice

train_copy = train_data_new[columns]
test_copy = test_data_new[columns]

mean = train_copy.mean(axis = 0)
std = train_copy.std(axis = 0)

train_copy -= mean
train_copy /= std

test_copy -= mean
test_copy /= std
```

[ ]:
```

```

[642]:
```
importance = {}
for item in columns:
    cor = np.corrcoef(train_copy[item].
 ↪astype(float),train_data_new['SalePrice'])[0,1]
    if np.abs(cor)>0.35:
        importance[item]=cor

importantList=list(importance.keys())

test_copy = test_copy[importantList]
# importantList.append('SalePrice')
train_copy = train_copy[importantList]
```

[643]:
```
train_copy['SalePrice'] = train_data['SalePrice']
```

[644]:
```
train_copy.shape
```

[644]:
```
(1460, 28)
```

0.4 Now we can save the data

[645]:
```
train_copy.to_csv('G:
↪\Documents\ReferenceBooks\MachineLearning\Rohban\Homework\HW3\TrainPreprocessed.
↪csv',index=False)
```

[646]:
```
test_copy.to_csv('G:
↪\Documents\ReferenceBooks\MachineLearning\Rohban\Homework\HW3\TestPreprocessed.
↪csv',index=False)
```

[ ]:

[ ]: