Zillow's Price Feature Engineering

Chan Yu Yankai Liu Yifei Bi

Procedures Guide

1.Preparation
2.Impute Data 1
3.Creating New Features
4.Impute Data 2

Reporter : Chan Yu

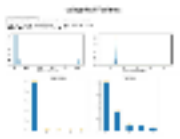
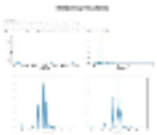
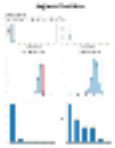
1

1.Convert Float64 to Float32

	parcelid	airconditioningtypeid	architecturalstyletypeid	basementsqft	bathroomcnt	bedroomcnt	buildingclasstypeid
0	10754147	1.0	7.0	535.0	0.0	0.0	4.0
1	10759547	1.0	7.0	535.0	0.0	0.0	4.0
2	10843547	1.0	7.0	535.0	0.0	0.0	5.0
3	10859147	1.0	7.0	535.0	0.0	0.0	3.0
4	10879947	1.0	7.0	535.0	0.0	0.0	4.0

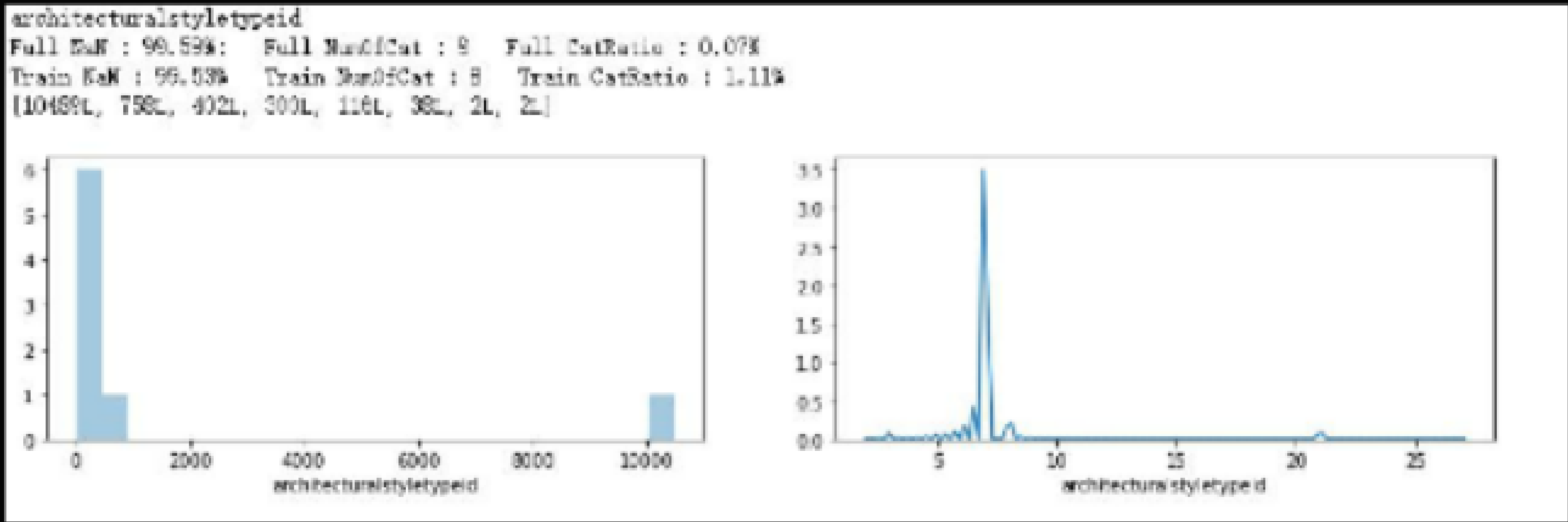
5 rows × 54 columns

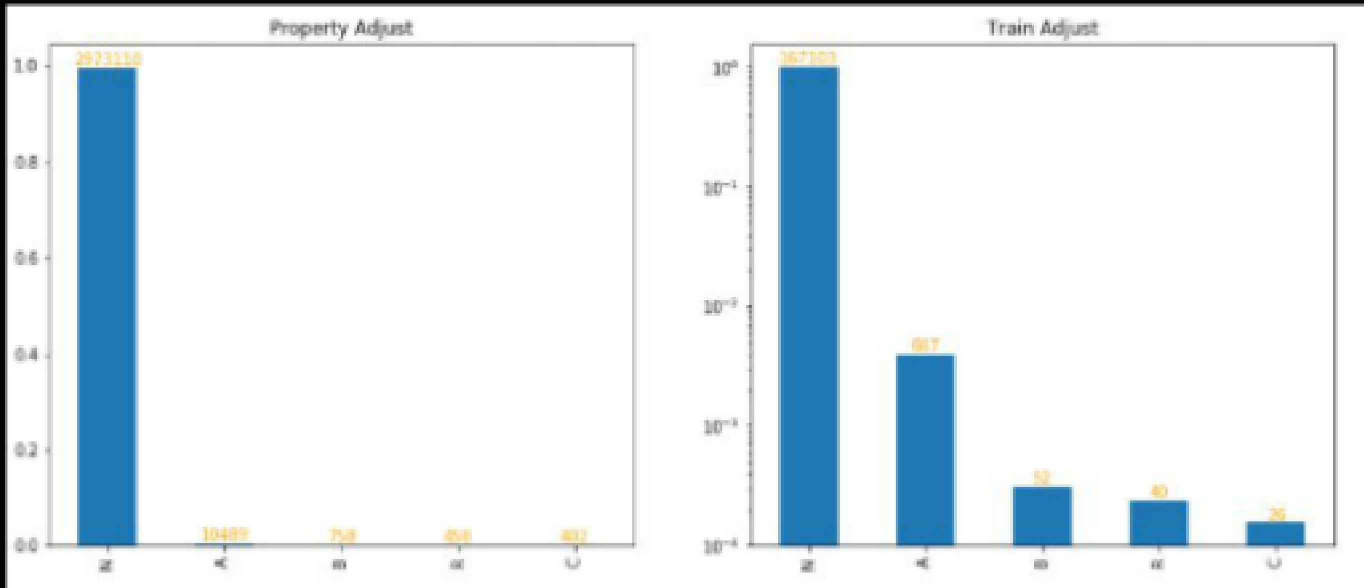
2.Group Features into Different Groups

3.Prepare Functions For Corresponding Group

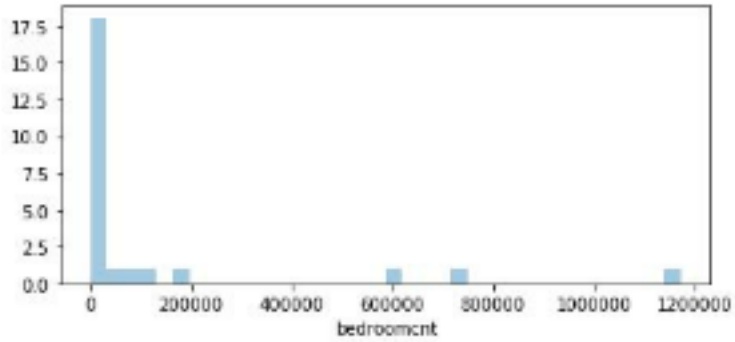
Categorical Features

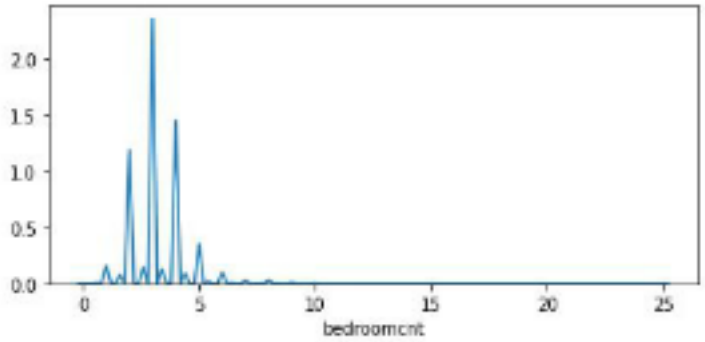


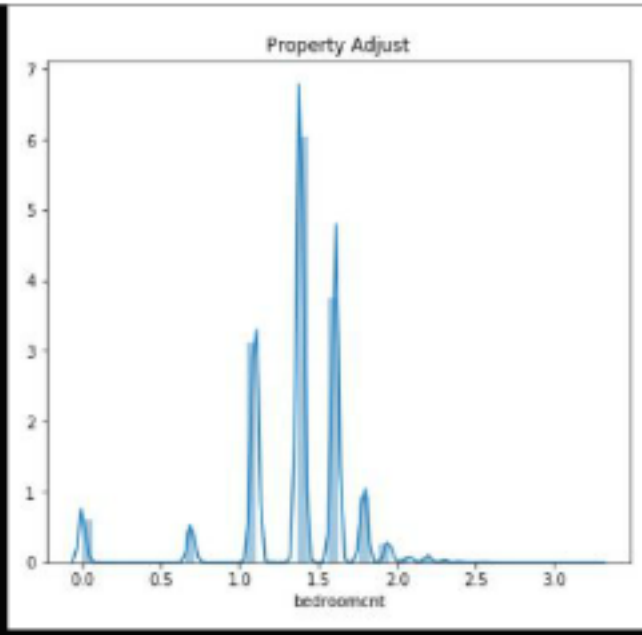


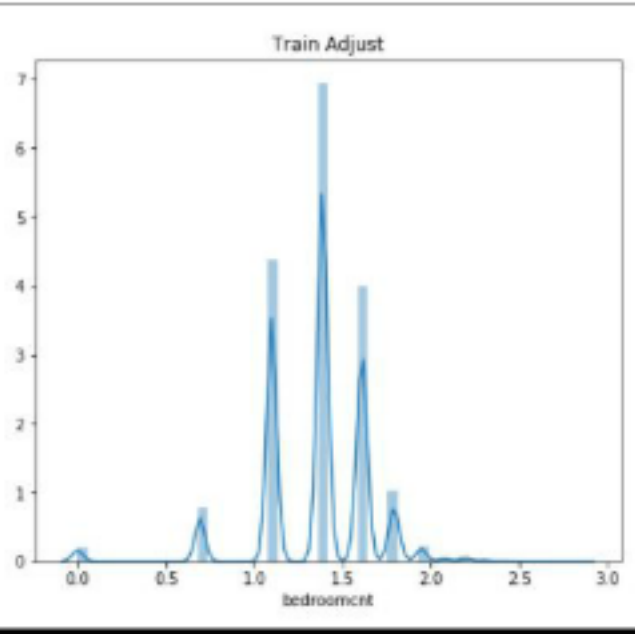
Numerical Features

```
bedroomcnt
Full NaN : 0.10%:   Full NumOfCat : 26   Full CatRatio : 0.00%
Train NaN : 0.03%   Train NumOfCat : 18   Train CatRatio : 0.01%
[1172760L, 731476L, 606787L, 182768L, 118705L, 85942L, 48915L, 13542L, 12763L, 4279L]
```

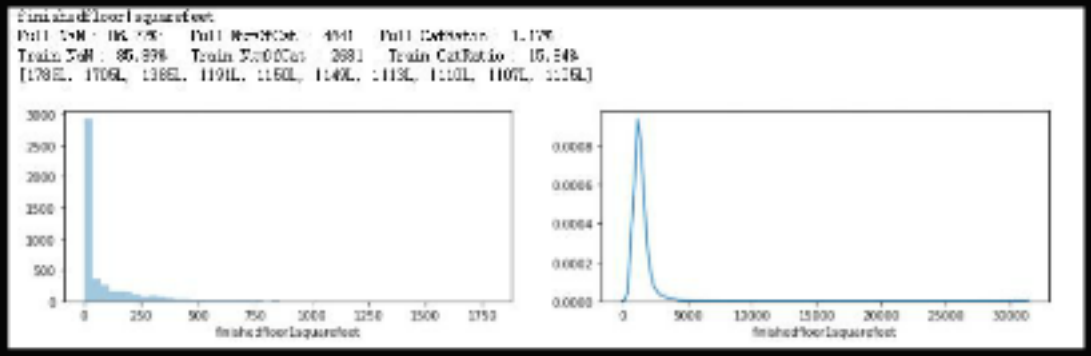


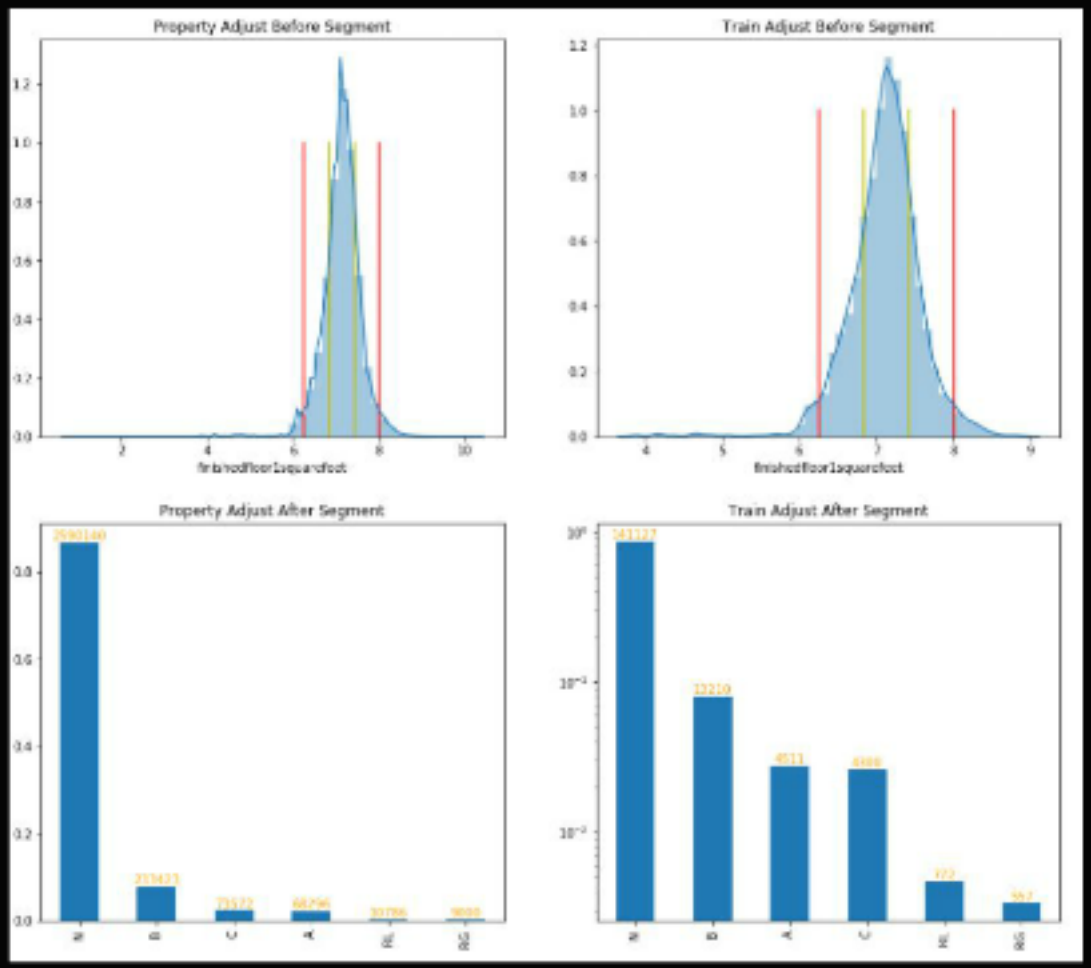






Segment Features

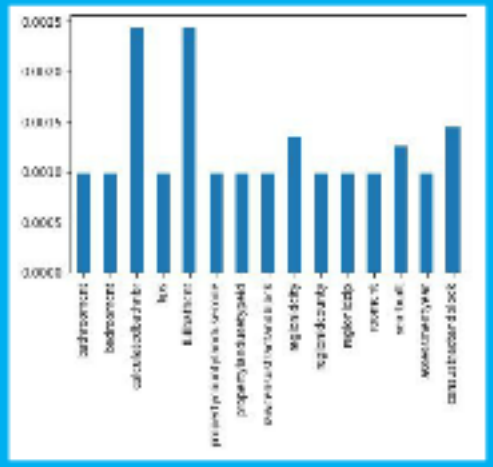




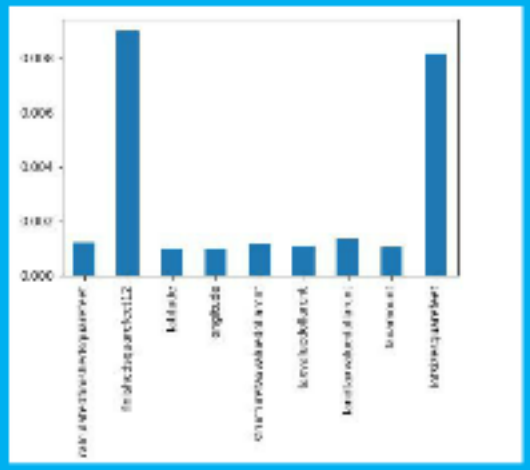
2

1.Preparation
2.Impute Data 1
3.Creating New Features
4.Impute Data 2

Impute By Median



Impute By Mode

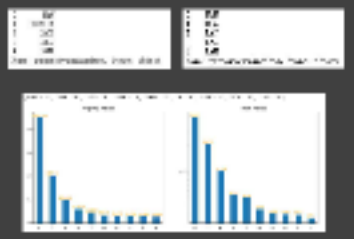


Missing ratio < 10%

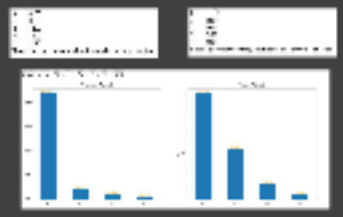
3

1.Preparation
2.Impute Data 1
3.Creating New Features
4.Impute Data 2

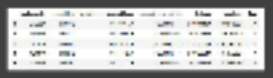
propertyzoningdesc



propertycountylandusecode

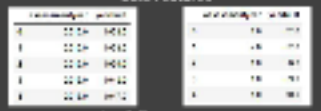


Cluster Geo Info



7 algrithms and each for 8 classes

Date Features

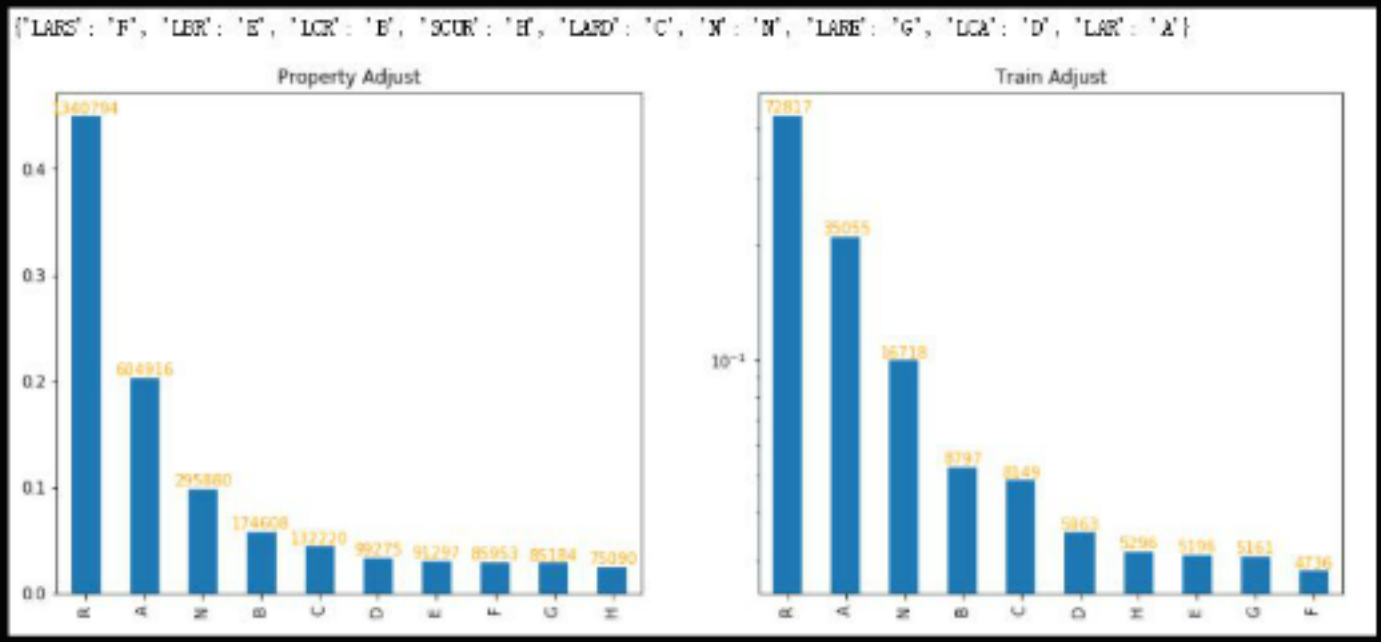


transdateyearmonthyear

propertyzoningdesc

```
0        NaN
1     LCA11*
2       LAC2
3       LAC2
4       LAM1
Name: propertyzoningdesc, dtype: object
```

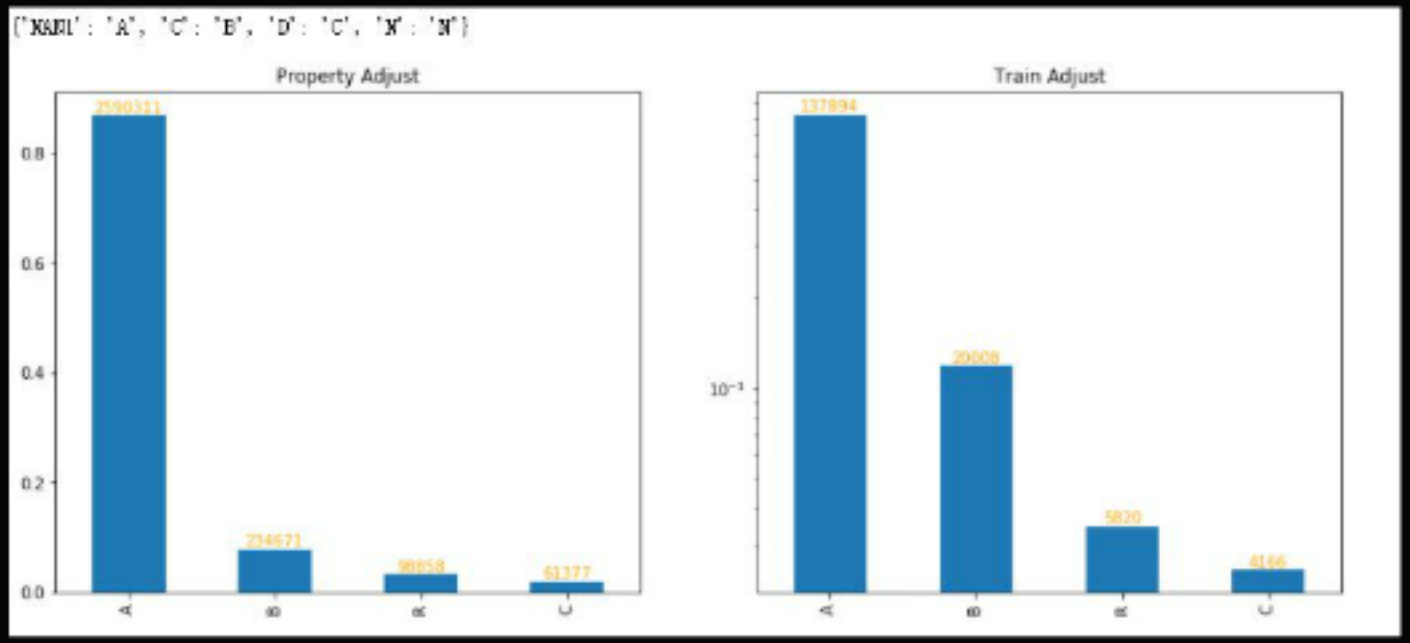
```
0     NaN
1     LCA
2     LAC
3     LAC
4     LAM
Name: propertyzoningdesc, dtype: object
```



propertycountylandusecode

```
0     010D
1     0109
2     1200
3     1200
4     1210
Name: propertycountylandusecode, dtype: object
```

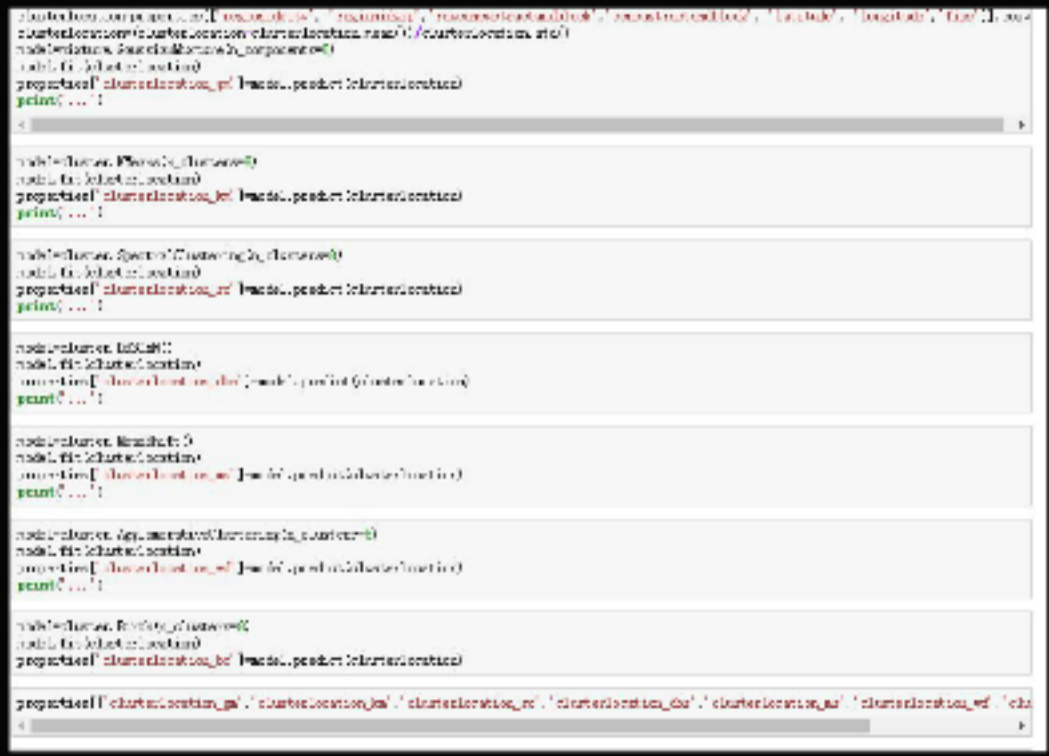
```
0        D
1     NAN1
2     NAN1
3     NAN1
4     NAN1
Name: propertycountylandusecode, dtype: object
```



Cluster Geo Info

	regionidcity	regionidzip	rawcensustractandblock	censustractandblock	latitude	longitude	fips
0	37688.0	96337.0	60378004.0	6.059032e+13	34144440.0	-118654080.0	A
1	37688.0	96337.0	60378000.0	6.059032e+13	34140432.0	-118625360.0	A
2	51617.0	96095.0	60377032.0	6.059032e+13	33989360.0	-118394632.0	A
3	12447.0	96424.0	60371412.0	6.059032e+13	34148864.0	-118437208.0	A
4	12447.0	96450.0	60371232.0	6.059032e+13	34194168.0	-118365816.0	A

7 algrithms and each for 8 classes

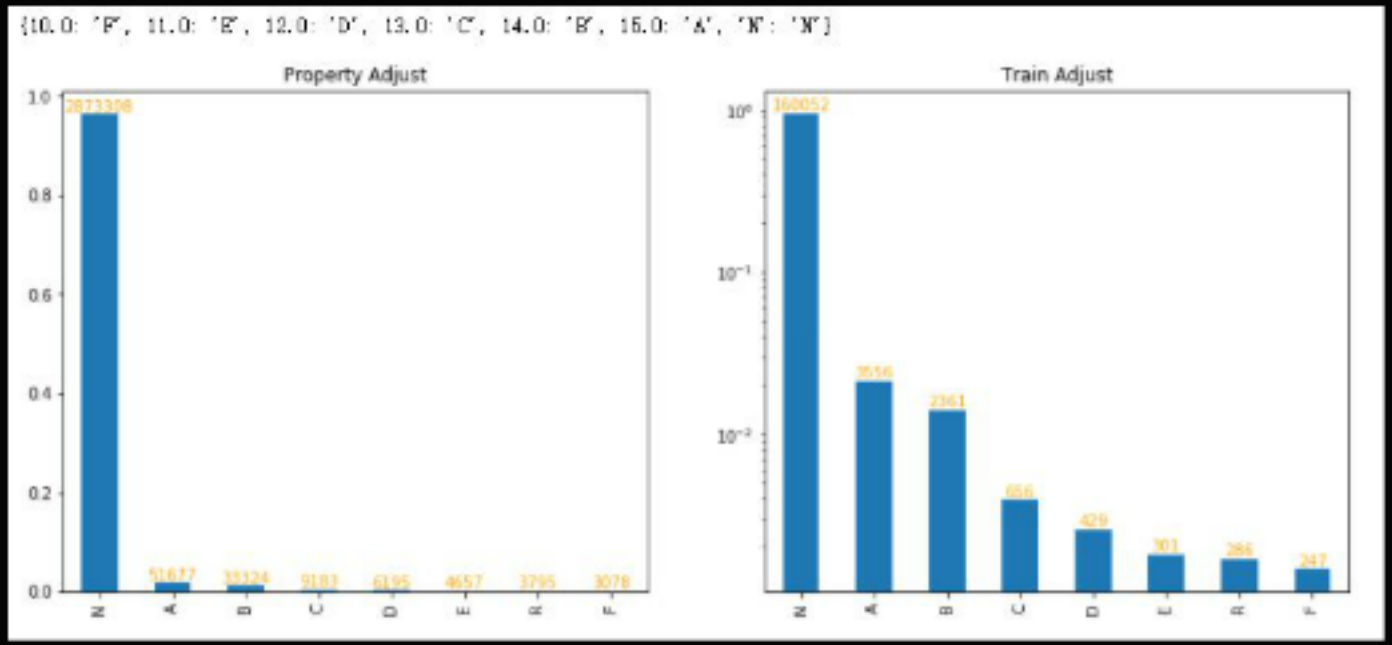


Date Features

	assessmentyear	yearbuilt
0	2016.0	1963.0
1	2015.0	1963.0
2	2016.0	1959.0
3	2016.0	1948.0
4	2016.0	1947.0

	assessmentyear	yearbuilt
0	1.0	62.0
1	2.0	62.0
2	1.0	58.0
3	1.0	69.0
4	1.0	70.0

taxdelinquencyyear

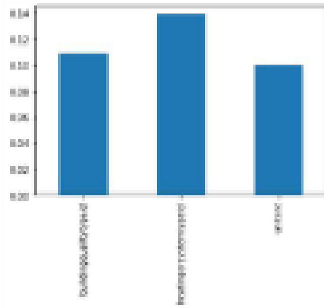


4

1.Preparation
2.Impute Data 1
3.Creating New Features
4.Impute Data 2

```
imputedatasets=imputefullsets[complete_features].copy()
for f in ['buildingqualitytypeid','heatingorsystemtypeid','unitcnt']:
    print(f)
    train_x=imputedatasets[imputefullsets[f].notnull()]
    train_y=imputefullsets[f][imputefullsets[f].notnull()]
    impute_x=imputedatasets[imputefullsets[f].isnull()]
    model = ensemble.RandomForestClassifier()
    model.fit(train_x,train_y)
    print('ok')
    scores=model.score(train_x,train_y)
    print('R2 Score : '+str(scores.mean()))
    imputefullsets.at[imputefullsets[f].isnull(),f]=model.predict(impute_x)
```

```
buildingqualitytypeid
ok
R2 Score :  0.998157804699
heatingorsystemtypeid
ok
R2 Score :  0.998199035403
unitcnt
ok
R2 Score :  0.994027609477
```



10% < Missing ratio < 30%