```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline

import warnings
warnings.filterwarnings("ignore")
```

```
train = pd.read_csv("train.csv")
train.head()
```

	ID	y	X0	X1	X2	X3	X4	X5	X6	X8	X10	X11	X12	X13	X14	X15	X16	X17	X
0	0	130.81	k	v	at	a	d	u	j	o	0	0	0	1	0	0	0	0	
1	6	88.53	k	t	av	e	d	y	l	o	0	0	0	0	0	0	0	0	
2	7	76.26	az	w	n	c	d	x	j	x	0	0	0	0	0	0	0	1	
3	9	80.62	az	t	n	f	d	x	l	e	0	0	0	0	0	0	0	0	
4	13	78.02	az	v	n	f	d	h	d	n	0	0	0	0	0	0	0	0	

```
train.shape
```

Automatic saving failed. This file was updated remotely or in another tab. Show diff

```
test = pd.read_csv("test.csv")
test.head()
```

	ID	X0	X1	X2	X3	X4	X5	X6	X8	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19
0	1	az	v	n	f	d	t	a	w	0	0	0	0	0	0	0	0	0	0
1	2	t	b	ai	a	d	b	g	y	0	0	0	0	0	0	0	0	0	1
2	3	az	v	as	f	d	a	j	j	0	0	0	0	1	0	0	0	0	0
3	4	az	l	n	f	d	z	l	n	0	0	0	0	0	0	0	0	0	0
4	5	w	s	as	c	d	y	i	m	0	0	0	0	1	0	0	0	0	0

```
test.shape
```

```
(4209, 377)
```

```
for i in train.columns:
    if i not in test.columns:
        print("Output variable is {}".format(i))
```

```
Output variable is y
```

Understand your data

```
train.shape
```

```
(4209, 378)
```

```
train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4209 entries, 0 to 4208
Columns: 378 entries, ID to X385
dtypes: float64(1), int64(369), object(8)
memory usage: 12.1+ MB
```

We've three different type of data

- 1 Float variables
- 369 Integer variables

Automatic saving failed. This file was updated remotely or in another tab. Show diff

```
pd.set_option('display.max_rows', None)
pd.set_option('display.max_columns', None)
```

Check Whether Variance is Zero

```
from sklearn import preprocessing
```

```
variance_with_zero = train.var()[train.var()==0].index.values
variance_with_zero
```

```
array(['X11', 'X93', 'X107', 'X233', 'X235', 'X268', 'X289', 'X290',
       'X293', 'X297', 'X330', 'X347'], dtype=object)
```

12 Variables are there. Removing all variables whose variance is zero

```
train = train.drop(variance_with_zero,axis=1)
```

```
train = train.drop('ID',axis=1)
```

```
test_var_with_zero = test.var()[test.var()==0].index.values
test_var_with_zero
```

```
    array(['X257', 'X258', 'X295', 'X296', 'X369'], dtype=object)
```

```
test = test.drop(test_var_with_zero,axis=1)
```

```
test = test.drop('ID',axis=1)
```

Check for null and unique values

```
train.isna().sum().sum()
```

```
    0
```

```
test.isna().sum().sum()
```

```
    0
```

```
train.nunique()
```

Automatic saving failed. This file was updated remotely or in another tab. Show diff

```
    X292        2
    X294        2
    X295        2
    X296        2
    X298        2
    X299        2
    X300        2
    X301        2
    X302        2
    X304        2
    X305        2
    X306        2
    X307        2
    X308        2
    X309        2
    X310        2
    X311        2
    X312        2
    X313        2
    X314        2
    X315        2
    X316        2
```

```
X316      2
X317      2
X318      2
X319      2
X320      2
X321      2
X322      2
X323      2
X324      2
X325      2
X326      2
X327      2
X328      2
X329      2
X331      2
X332      2
X333      2
X334      2
X335      2
X336      2
X337      2
X338      2
X339      2
X340      2
X341      2
X342      2
X343      2
X344      2
X345      2
X346      2
X348      2
X349      2
X350      2
```

Automatic saving failed. This file was updated remotely or in another tab. Show diff

Label Encoding

Train Data

```
for i in train.columns:
    a=train[i].dtype
    if a == 'object':
        print(i)
```

```
X0
X1
X2
X3
X4
X5
X6
X8
```

```
le = preprocessing.LabelEncoder()
```

```
train['X0']= le.fit_transform(train['X0'])
```

```
train['X0'].unique()
```

```
array([32, 20, 40,  9, 36, 43, 31, 29, 39, 35, 19, 27, 44, 45,  7,  8, 10,
       46, 37, 15, 12, 42,  5,  0, 26,  6, 25, 13, 24,  1, 22, 14, 30, 38,
       21, 18, 23, 41,  4, 16, 34, 33, 17, 11,  3, 28,  2])
```

```
train['X1']= le.fit_transform(train['X1'])
train['X2']= le.fit_transform(train['X2'])
train['X3']= le.fit_transform(train['X3'])
train['X4']= le.fit_transform(train['X4'])
train['X5']= le.fit_transform(train['X5'])
train['X6']= le.fit_transform(train['X6'])
train['X8']= le.fit_transform(train['X8'])
```

```
train.head()
```

	y	X0	X1	X2	X3	X4	X5	X6	X8	X10	X12	X13	X14	X15	X16	X17	X18	X19
0	130.81	32	23	17	0	3	24	9	14	0	0	1	0	0	0	0	1	0
1	88.53	32	21	19	4	3	28	11	14	0	0	0	0	0	0	0	1	0
2	76.26	20	24	34	2	3	27	9	23	0	0	0	0	0	0	1	0	0
																0	0	0
																0	0	0

Automatic saving failed. This file was updated remotely or in another tab. Show diff

```
train['y'].nunique()
```

```
2545
```

Test Data

```
for i in test.columns:
    a=test[i].dtype
    if a == 'object':
        print(i)
```

```
X0
```

```
    X1
    X2
    X3
    X4
    X5
    X6
    X8
```

```
test['X0']= le.fit_transform(test['X0'])
test['X1']= le.fit_transform(test['X1'])
test['X2']= le.fit_transform(test['X2'])
test['X3']= le.fit_transform(test['X3'])

test['X4']= le.fit_transform(test['X4'])
test['X5']= le.fit_transform(test['X5'])
test['X6']= le.fit_transform(test['X6'])
test['X8']= le.fit_transform(test['X8'])
```

```
test.head()
```

	X0	X1	X2	X3	X4	X5	X6	X8	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19	X2(
0	21	23	34	5	3	26	0	22	0	0	0	0	0	0	0	0	0	0	(
1	42	3	8	0	3	9	6	24	0	0	0	0	0	0	0	0	0	1	(
2	21	23	17	5	3	0	9	9	0	0	0	0	1	0	0	0	0	0	(
3	21	13	34	5	3	31	11	13	0	0	0	0	0	0	0	0	0	0	(
4	45	20	17	2	3	30	8	12	0	0	0	0	1	0	0	0	0	0	(

Automatic saving failed. This file was updated remotely or in another tab. Show diff

PCA For Train Data

```
X_train = train.drop("y", axis=1)
y_train = train["y"]
```

```
X_train.shape
```

```
    (4209, 364)
```

```
y_train.shape
```

```
    (4209,)
```

```
from sklearn.decomposition import PCA
```

```
train_pca = PCA(n_components=0.95)
```

```
Xtrain_pca = train_pca.fit_transform(X_train)
```

```
Xtrain_pca.shape
```

```
     (4209, 6)
```

```
pca_train = pd.DataFrame(Xtrain_pca, index=X_train.index, columns=["PC1", "PC2","PC3", "PC4",
```

```
pca_train.shape
```

```
     (4209, 6)
```

```
train_pca.explained_variance_ratio_*100
```

```
     array([38.33478209, 21.38803259, 13.2618659 , 11.82664248,  9.20600842,
             1.59060433])
```

PCA For Test Data

```
test.shape
```

Automatic saving failed. This file was updated remotely or in another tab. Show diff

```
test_pca = PCA(n_components=0.95)
```

```
Xtest_pca= test_pca.fit_transform(test)
```

```
Xtest_pca.shape
```

```
     (4209, 6)
```

```
pca_test = pd.DataFrame(Xtest_pca, index=test.index, columns=["PC1", "PC2","PC3", "PC4","PC5'
```

```
pca_test.shape
```

```
     (4209, 6)
```

```
test_pca.explained_variance_ratio_*100
```

```
    array([43.51510223, 17.67089683, 13.64629223, 10.97791165,  8.62220781,
            1.43396216])
```

XG_Boost

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(pca_train,y_train, test_size = 0.1,random
```

```
from xgboost import XGBRegressor
```

```
xgb = XGBRegressor(objective="reg:linear",learning_rate=0.5)
```

```
xgb.fit(X_train, y_train)
```

```
    [17:35:46] WARNING: /workspace/src/objective/regression_obj.cu:152: reg:linear is now d
    XGBRegressor(learning_rate=0.5)
```

```
pred_xgb = xgb.predict(X_test)
pred_xgb
```

```
    array([ 96.677025,  89.24108 , 100.43104 , 109.19301 ,  91.60834 ,
            97.33993 ,  91.90733 ,  93.67694 ,  98.8307  ,  94.62247 ,
            98.45339 , 101.671616, 109.627335, 104.961494, 110.62873 ,
           108.62807 , 106.53911 ,  96.93742 ,  95.02451 , 102.47036 ,
```

Automatic saving failed. This file was updated remotely or in another tab. Show diff

```
           112.86675 ,  98.55632 ,  95.26648 , 101.83872 ,  90.8998  ,
           102.73306 ,  94.46457 ,  97.55571 ,  96.1337  ,  81.290504,
           102.42054 ,  95.2727  , 103.47246 ,  98.83295 , 103.61694 ,
           101.54705 , 108.80643 ,  76.04421 ,  96.05677 ,  88.67628 ,
            93.987495, 113.107956, 116.37802 , 103.7517  , 104.72315 ,
            91.66554 , 101.37777 ,  95.744225,  93.63236 ,  92.20367 ,
           106.02717 ,  91.14704 ,  96.79794 ,  96.73966 , 111.65975 ,
            99.62692 ,  93.93386 ,  97.167755,  96.28038 ,  93.64416 ,
           111.94243 ,  99.984764, 105.62146 ,  78.28211 ,  98.44162 ,
           109.244064,  93.572914,  95.78864 ,  97.27851 ,  98.81077 ,
           109.72914 ,  94.030334, 113.70408 ,  78.80711 , 105.50059 ,
            93.10872 ,  91.95896 , 105.21229 ,  90.51801 , 101.05882 ,
            92.08666 , 114.68566 , 110.2027  , 101.72687 ,  94.227844,
           110.36757 ,  94.31439 , 102.11966 , 110.66194 , 110.545654,
           109.329704, 108.95114 , 103.345825,  94.84276 , 104.98302 ,
            98.62796 ,  96.65266 ,  94.74956 , 111.17852 ,  99.491325,
           111.43373 ,  96.33035 , 106.29449 ,  98.32257 , 101.204216,
           105.53293 ,  95.89699 ,  96.65266 , 101.60975 ,  94.53765 ,
           108.2031  , 105.3902  ,  98.37849 , 105.03179 ,  95.03175 ,
            95.65686 , 110.61664 ,  91.62252 ,  96.35483 ,  78.36273 ,
            97.80674 , 102.57756 ,  96.43804 , 118.72537 ,  98.08389 ,
           102.59251 , 105.18584 , 113.46639 , 108.30829 ,  91.429306,
           111.728325, 106.56453 ,  99.54356 , 104.2002  , 100.48177 ,
            98.06534 , 106.850914, 100.504684, 107.3871  , 104.16318 ,
            78.121284, 109.57519 , 101.82887 , 108.49518 , 104.832375,
            92.72431 ,  94.13557 , 100.731674, 108.99372 , 101.7765  ,
            93.7157  , 103.78064 , 100.51735 , 103.81624 , 100.98854 ,
            94.10373 , 107.96189 , 108.79763 ,  87.78747 , 108.23189 ,
           111.06655 , 110.56345 ,  94.63027 , 100.5408  , 113.65731 ,
            96.89276 , 110.52625 ,  93.73564 ,  92.96491 ,  82.30502 ,
            94.40658 ,  95.08634 , 107.16149 , 104.99619 , 103.49964 ,
           111.93336 , 110.83525 , 109.7473  ,  88.2186  ,  94.48949 ,
            97.664955, 101.978424, 109.56851 , 107.9143  ,  96.00706 ,
            84.863976, 111.27362 ,  91.99808 , 115.04045 , 113.27181 ,
           110.69929 , 109.36672 ,  97.66214 ,  97.81973 , 103.16315 ,
           100.0719  ,  93.3702  ,  96.559685, 103.82577 , 101.11321 ,
           109.49825 ,  98.642265, 102.34473 ,  93.6259  ,  93.51254 ,
           102.54329 ,  95.46213 ,  92.45855 ,  96.733864,  79.66513 ,
            92.27564 ,  98.938576, 103.0508  ,  90.47591 ,  95.80722 ,
            94.44601 , 113.337906,  94.30125 , 111.901215,  95.3619  ,
            81.412506, 101.55886 ,  94.66011 , 110.54164 ,  97.69184 ,
           101.32527 ,  95.156204, 107.50836 ,  94.00897 ,  94.222916,
            94.726036,  95.281685, 107.78557 , 113.31062 , 108.028435,
            96.19138 ,  94.24291 , 111.86142 ,  91.58949 , 118.98313 ,
           111.619194,  98.66714 , 114.201744,  88.190186, 101.84383 ,
           102.426926, 103.03613 ,  96.38758 , 107.737656, 107.23494 ,
            95.61202 ,  94.12057 ,  73.05708 ,  98.63693 , 102.84685 ,
           107.18926 ,  95.27108 ,  95.065994,  92.53257 ,  94.98839 ,
           106.60193 ,  87.97817 , 114.35699 , 118.412224, 110.07702 ,
            77.46064 ,  90.83843 ,  94.55281 ,  93.53385 , 110.051544,
           106.53911 , 104.441124, 111.197495,  91.56613 ,  96.754074,
            90.79647 , 106.08779 , 118.99239 ,  97.653694, 106.734985,
```

```
r2_score(y_test,pred_xgb)
```

Automatic saving failed. This file was updated remotely or in another tab. Show diff

```
mean_squared_error(pred_xgb, y_test)
```

```
     69.66731100421458
```

```
from sklearn.metrics import mean_squared_error
from sklearn.model_selection import KFold
from sklearn.model_selection import cross_val_score
```

```
kfold = KFold(n_splits=7)
results = cross_val_score(xgb, X_train, y_train, cv=kfold)
y_test_pred = xgb.predict(X_test)

mse = mean_squared_error(y_test_pred, y_test)
```

```
y_pred = xgb.predict(X_test)
```

```
[17:35:46] WARNING: /workspace/src/objective/regression_obj.cu:152: reg:linear is now d
[17:35:46] WARNING: /workspace/src/objective/regression_obj.cu:152: reg:linear is now d
[17:35:47] WARNING: /workspace/src/objective/regression_obj.cu:152: reg:linear is now d
[17:35:47] WARNING: /workspace/src/objective/regression_obj.cu:152: reg:linear is now d
[17:35:47] WARNING: /workspace/src/objective/regression_obj.cu:152: reg:linear is now d
[17:35:47] WARNING: /workspace/src/objective/regression_obj.cu:152: reg:linear is now d
[17:35:47] WARNING: /workspace/src/objective/regression_obj.cu:152: reg:linear is now d
```

```
mse
```

```
69.66731100421458
```

Automatic saving failed. This file was updated remotely or in another tab. Show diff

Colab paid products - Cancel contracts here

✓ 0s completed at 11:05 PM ● ×