Capstone Mini-Project: Data Analysis on Human activity prediction based on smartphone data sets

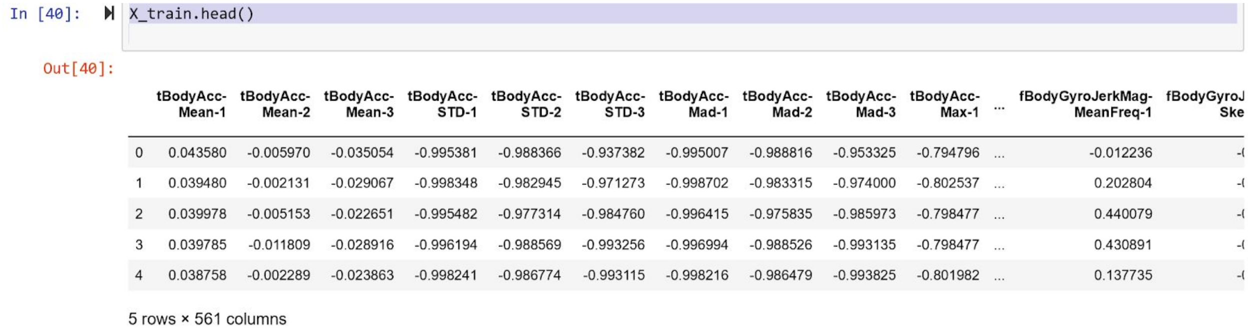
Problem:
Nowadays a lot of people wear smart devices which can track people's activity, those collected sensor signal data can be used to predict human activity(walking,sitting,lying,walking stairs and also pose transition activities). In some situation based on their patten, we may even able to identify different age group activity, or further the individual participants from their activity styles , which may help to detect early signal of sickness.and guide people to live in an healthy style. In this Analysis ,we only try to resolve the first step, predict human activity. Below is the dataset we are going to use.

Data: UCI
http://archive.ics.uci.edu/ml/datasets/Smartphone-Based+Recognition+of+Human+Activities+and+Postural+Transitions
Sensor signals from smartphones of a group of 30 people performed a protocol of activities(such as walking,sitting ,lying and walking stairs) were collected and pre-processed.

Phase one: Data wrangling

1)Load Data

we will download the data from the above link,used pandas function pd.read_csv to load it



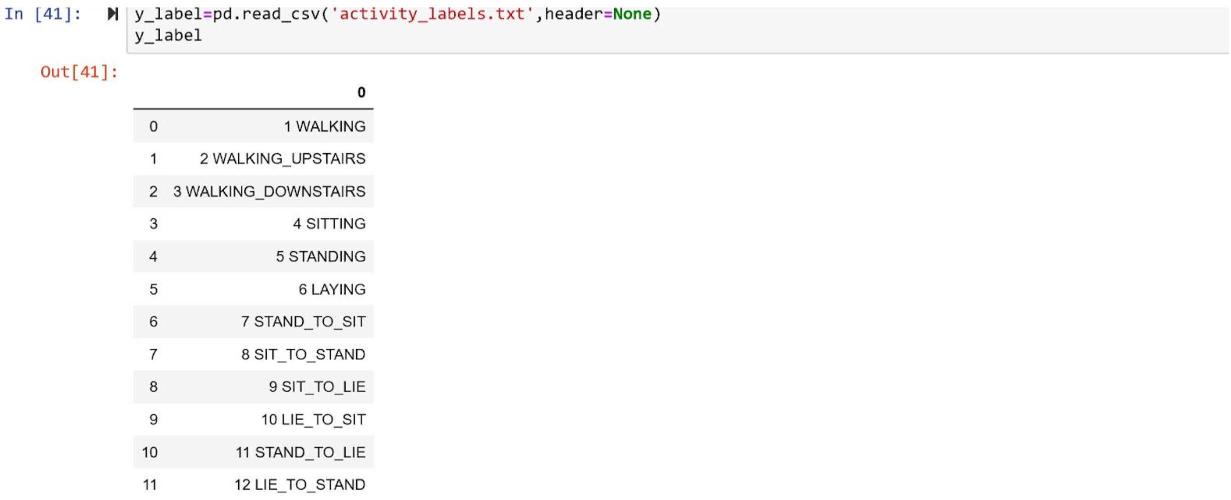
```
In [40]:  X_train.head()
```

Out[40]:

	tBodyAcc-Mean-1	tBodyAcc-Mean-2	tBodyAcc-Mean-3	tBodyAcc-STD-1	tBodyAcc-STD-2	tBodyAcc-STD-3	tBodyAcc-Mad-1	tBodyAcc-Mad-2	tBodyAcc-Mad-3	tBodyAcc-Max-1	...	fBodyGyroJerkMag-MeanFreq-1	fBodyGyroJ Ske
0	0.043580	-0.005970	-0.035054	-0.995381	-0.988366	-0.937382	-0.995007	-0.988816	-0.953325	-0.794796	...	-0.012236	-
1	0.039480	-0.002131	-0.029067	-0.998348	-0.982945	-0.971273	-0.998702	-0.983315	-0.974000	-0.802537	...	0.202804	-
2	0.039978	-0.005153	-0.022651	-0.995482	-0.977314	-0.984760	-0.996415	-0.975835	-0.985973	-0.798477	...	0.440079	-
3	0.039785	-0.011809	-0.028916	-0.996194	-0.988569	-0.993256	-0.996994	-0.988526	-0.993135	-0.798477	...	0.430891	-
4	0.038758	-0.002289	-0.023863	-0.998241	-0.986774	-0.993115	-0.998216	-0.986479	-0.993825	-0.801982	...	0.137735	-

5 rows × 561 columns

X_train has 7767 records and 561 columns, y_train has 12 different label



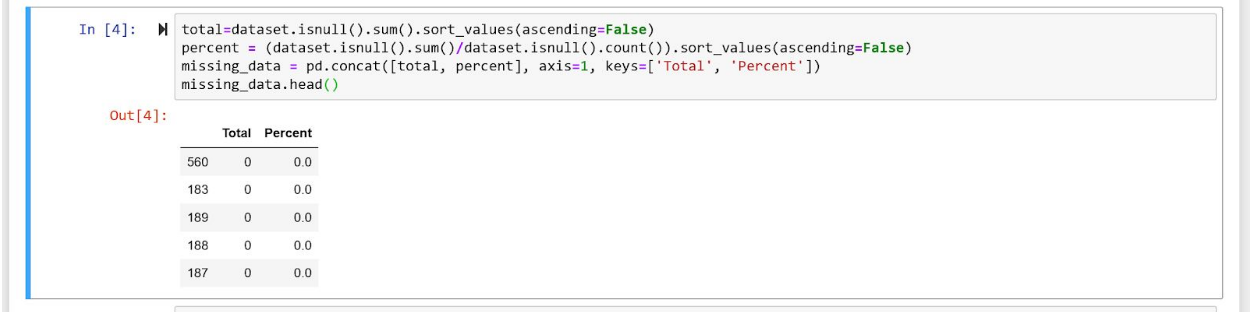
In [41]:
```
y_label=pd.read_csv('activity_labels.txt',header=None)
y_label
```

Out[41]:

	0
0	1 WALKING
1	2 WALKING_UPSTAIRS
2	3 WALKING_DOWNSTAIRS
3	4 SITTING
4	5 STANDING
5	6 LAYING
6	7 STAND_TO_SIT
7	8 SIT_TO_STAND
8	9 SIT_TO_LIE
9	10 LIE_TO_SIT
10	11 STAND_TO_LIE
11	12 LIE_TO_STAND

2)Missing data

The null analysis shows there is no missing data for this dataset,so we don't need to fill the missing data



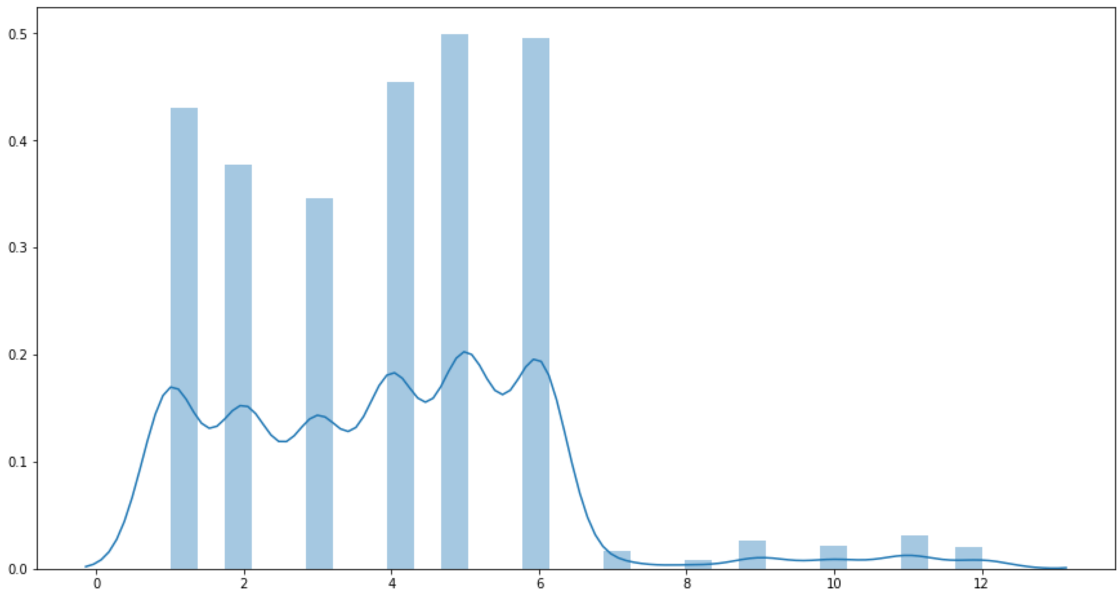
In [4]:
```
total=dataset.isnull().sum().sort_values(ascending=False)
percent = (dataset.isnull().sum()/dataset.isnull().count()).sort_values(ascending=False)
missing_data = pd.concat([total, percent], axis=1, keys=['Total', 'Percent'])
missing_data.head()
```

Out[4]:

	Total	Percent
560	0	0.0
183	0	0.0
189	0	0.0
188	0	0.0
187	0	0.0

3) **Data balance**

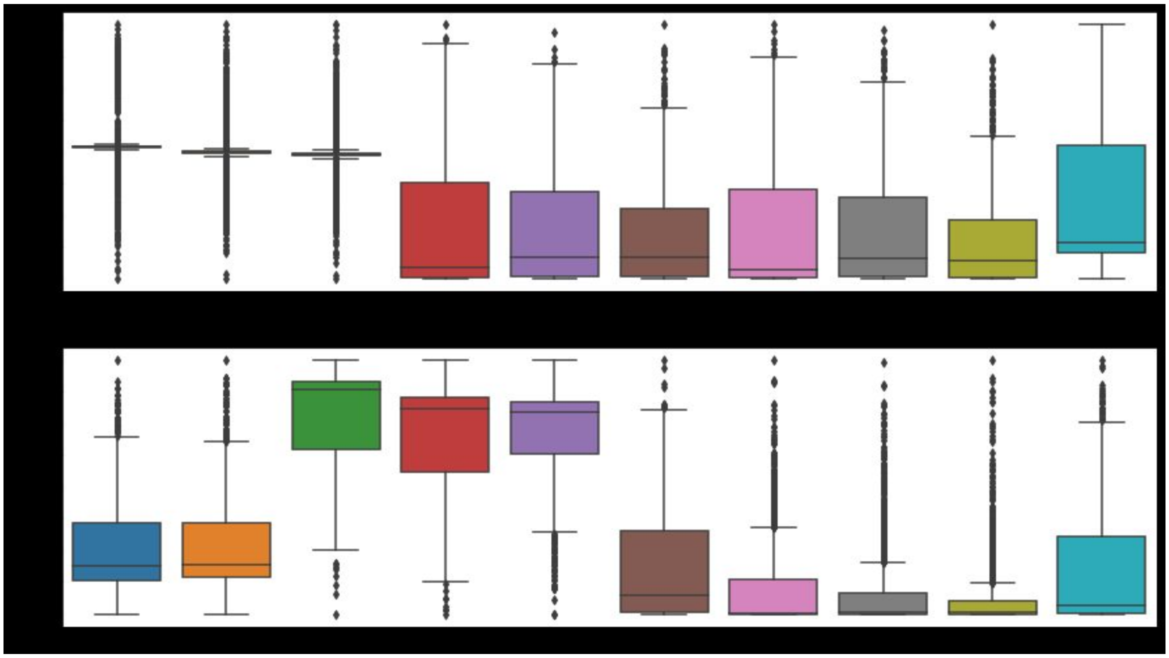
The barplot shows our dataset has 6 major classes and 6 minor classes, it is unbalanced, so we applied SMOTE to resample the data



4) Outlier

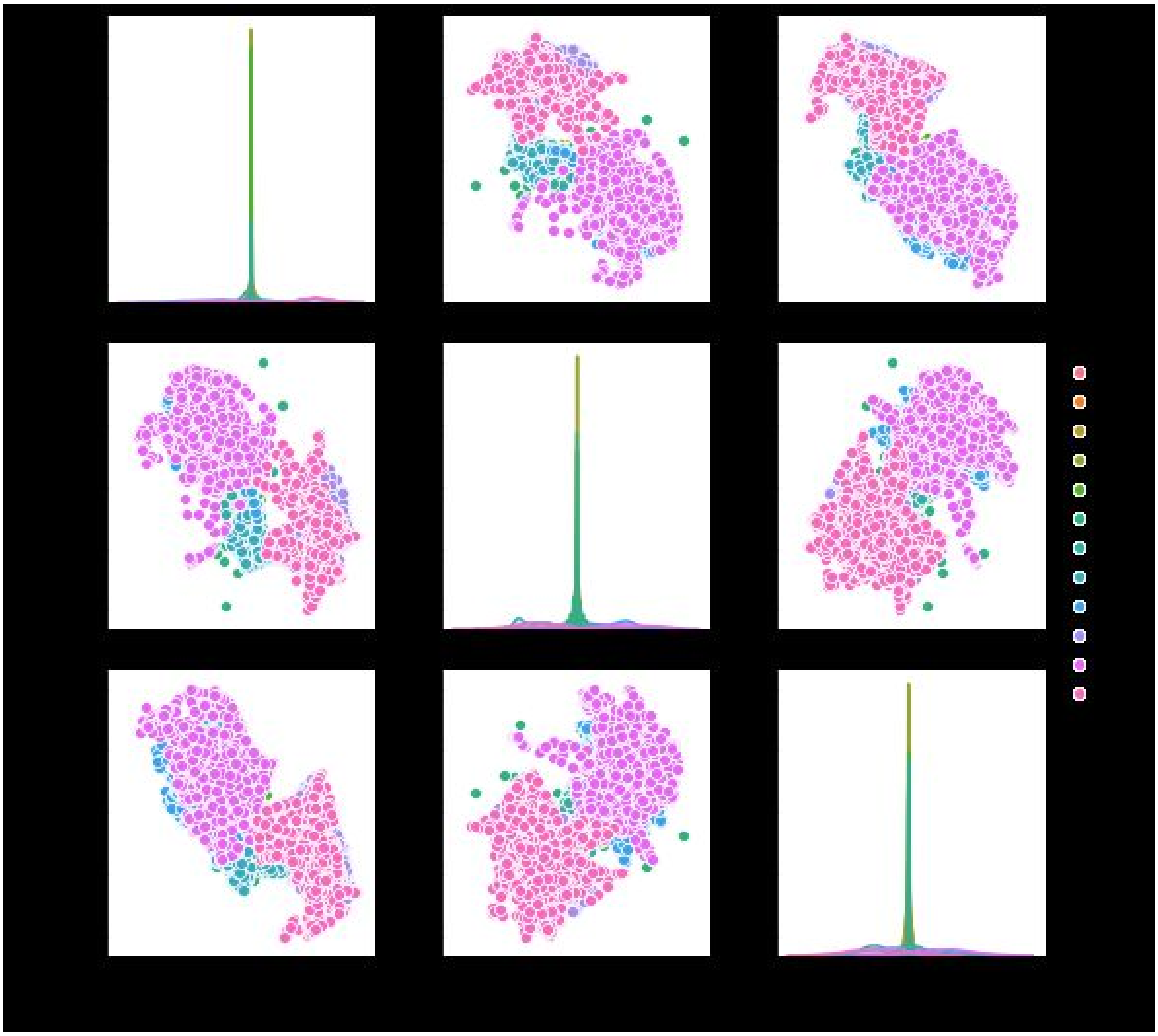
According to the boxplot , we can see that there are a lot of features having outliers

For example feature 8 has a lot of larger outliers,we refill the outlier with 75% value ,feature 13 has smaller outliers ,we will refill those with 25% value.



5)**Correlation**

The results of plot pairs of the first 3 numeric variables shows some variables are correlated with each other , for example feature 0 and feature 1



6)Null hythophis

There is no significenct difference between below two sets:

Set a) feature 2 value larger than 0.5, y belong to class 4

Set b) feature 2 value less than 0.5, y not belong to class 4

```
new_dfyy=len(new_df[(new_df[2]>0.5)&(new_df['0_y']==4)])
new_dfyn=len(new_df[(new_df[2]>0.5)&(new_df['0_y']!=4)])
new_dfny=len(new_df[(new_df[2]<0.5)&(new_df['0_y']==4)])
new_dfnn=len(new_df[(new_df[2]<0.5)&(new_df['0_y']!=4)])

import scipy.stats as stats
oddsratio, pvalue = stats.fisher_exact([[new_dfyy,new_dfyn], [new_dfny,
new_dfnn]])
pvalue
```

```
Out[79]:
7.363006721969925e-25
```

The pvalue is 7.363006721969925e-25 < 0.05, so the null hythophis is rejected, there is a big chance if feature 2>0.5 ,it will belong to class 4.

Summary

After some basic analysis on our datasets,we know that there is no missing data for our dataset, and it is all numeric fields. we resample it into a more balance dataset using smoke technique. We also refill the outlier with 25% or 75% value. The pairbox shows some features are correlated with each other. When feature 2>0.5, there is a big chance ,the final class will be 4.

All the codes are saved in ipython notebook