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

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.

Data wrangling steps I took to clean the dataset:

1)Load data:

Since it is .txt file, I used pandas function pd.read_csv , with sep =" "

2)train test split

Our dataset already splited, we actually don't need do this

3)identify if there is missing data

Use .isnull() to identify the missing data, got 0 results.

So there is no missing data for our dataset, but if it has we can use 'k-NN impute' to fill missing data

4)Identify if data is balance

Use groupby().size() on y_train, we found it has 6 major classes and 6 minor classes, it is unbalanced, so we applied SMOTE to resample the data

5) Identify which columns/variables are numerical and which are categorical.

Use get_numeric_data(), we found out all the columns are numeric so we don't need do OneHotEncoder()

6)Identify outlier

Use boxplot on each column, got no outlier, but if there is outlier we can refill the outlier with 75% value if it is too larger or 25% value if it is too small

7)Plot pairs of numeric variables

The results shows some variables are linear related with each other .

8)Boxplot further investigate patten

9)Apply principal components analysis to just input variable

The result shows we can reduce the dimension to 14 components we still can keep 80%var

10)scaling the numeric field

Use standard scaling

All the codes are saved in ipython notebook