COMP9313 Assignment 2 Report
Name: PEIGUO GUAN **zID**: z5143964

Object Assignment2:

Main: return type Unit

1. inputFilePath: set input path
2. outputDirPath: set output path
3. input = sc.textFile(inputFilePath) get the input data in input
4. split the input data line by line and filter the value which is empty
5. map the value with first column and the last column which is the question required
6. then groupByKey to get the all value with same key
7. Define a Kb_MB_to_B function, the element in the value, processed by KB_MB_to_B function and return new value
8. sort the key
9. sort the value
10. Define a function get_mean_and_varience, use the function get_mean_and_varience to add mean and varience in value, and also get the min and max value as required
11. change the value in to required output model
12. !!! According to the forum we don't need to output csv file, we just need to output csv type txt, right? so the output type is txt => l10.coalesce(1,true).saveAsTextFile(outputDirPath)
    But if required I also can ouput as csv file, change the RDD to DF by .toDF(), and use the df attribute to ouput as csv file. Uncommend the code the last of my scala file.

Class myfunction: this is the part where define my udf function
KB_MB_to_B:

1. define a function to change KB or MB to B
2. but also change from String to Long, cause if the length is too large Integer cannot meet the requirement
3. the input type is String and Output type is Long

get_mean_and_varience:

1. define a function to calculate mean value and varience value
2. after finish calculate the result then changing to String and add "B"
3. the input type si List[Long] the output is List[Long]

to_csv_line:

1. define a function to keep all in one line
2. this can help to cahnge from RDD to DataFrame if required
3. and if the requirement is txt file, the result will be really good to csv file as well