**Description:** This is a list of assignments which is part of pySpark L1 training. This case study requires analyzing the real estate data file and generating required output.

**Technology**: pySpark (RDD API)

**Note: Use the attached source code template to submit the assignment Click** [**here**](#ExpectedOutput) **for details.**

**Input Datasets:**

RealEstate.txt

Assignment 1. Properties for given location

1. Read the RealEstate.txt file as an RDD
2. Filter out the header record
3. Create a RDD with Property ID, Size, Price for location = "La Oceana", with Size and Price columns cast to integer
4. Display content of RDD

Assignment 2. List of Unique locations

1. Read the RealEstate.txt file as an RDD
2. Filter out the header record
3. Create a RDD having unique locations
4. Display content of RDD

Assignment 3. Compute Actual Price of Property

1. Read the RealEstate.txt file as an RDD
2. Filter out the header record
3. Create a RDD having PropertyID, Location, Price, ActualPrice (= size \* Price SQ Ft)
4. Price should be cast to integer and Actual price should be cast to float and rounded off to 0 decimals
5. Sort records in descending order by Price column
6. Display content of RDD

Assignment 4. Total number of properties by location

1. Read the RealEstate.txt file as an RDD
2. Filter out the header record
3. Create a RDD having Location and property count
4. Display content of RDD

Assignment 5. Set Operations on RDDs

1. Read the RealEstate.txt file as an RDD
2. Filter out the header record
3. Create a RDD of Property ID, Location having 3 bedrooms and price < 150000
4. Create a second RDD of Property ID, Location having more than 2 bathrooms and price < 150000
5. Use an intersection operator to find properties satisfying both the above conditions
6. Display the output of rdd

**Expected Output:**

1. The source code should be submitted by filling up the attached output template.

Rename REAnalysis.template to REAnalysis.py

1. Populate the functions with the required RDD transformations
2. Test the output by executing the job in spark environment:

spark-submit REAnalysis.py

1. Upload the source code (REAnalysis.py) to a folder in sharepoint onedrive and specify the share folder URL while submitting the assignment in TopGear.

One drive can be accessed from URL: <https://wipro365-my.sharepoint.com>

**Output Template:**

