Exercise 03

Clone the project https://github.com/Gayani91/Training

Read data from data/RealEstate.csv
Columns: MLS,Location,Price,Bedrooms,Bathrooms,Size,Price SQ Ft,Status

Implement following using Scala and Spark RDDs

1. Number of houses located in Santa Maria-Orcutt
2. List the location and the price of houses which are priced over 500000
3. List the houses which have 3 bedrooms and available for short sale
4. Find the highest priced house in Cayucos which has more than 3 bedrooms and 2 bathrooms
5. What is the average price of houses to be sold in each city
6. Apply below aprivAvg equation and get the avg prices in a new column
   avgPrice = (price+PriceSQFt*Size)/2

Exercise 04

Implement above using Scala and Spark DataFrames

Exercise 05

1. Create a maven project
2. Add dependencies
3. Implement wordcount with spark
4. Create a fatrjar will all dependencies included
5. Copy jar to demo environment
6. Copy data (to remote location) and then to hdfs
7. Run script and load data