**Exercise 6a**

*Unstructured Spark exercise*

**Prior Knowledge**

Unix Command Line Shell

Simple Python

Spark Python  
Simple SQL syntax

**Learning Objectives**

Pulling together your skills from previous exercises

**Software Requirements**

(see separate document for installation of these)

* Apache Spark 2.0.0
* Python 2.7.12
* Nano text editor or other text editor

**Aim**

There is a file on your VM that contains some data from the Land Registry:

~/datafiles/landreg/landreg.csv

The CSV file has a header line. Full details are available here:

<https://www.gov.uk/guidance/about-the-price-paid-data>

But you don’t need to know all that.

The aim is simple:

I’d like you to calculate the average price paid by postcode for the data.

I want you to only base this on the first part of the postcode not the full postcode.

Please tell me the average price paid for the postcode areas: OX1, SW11.

**There are some hints overleaf.**

Hints:

1. Create a directory to hold your code and files
2. Use the same databricks CSV reader to load the data in
3. Be warned that some Postcodes are empty and need to be “cleaned” before you can start doing things like averaging.
4. You should know enough to do this as a set of Map/ReduceByKey operations.
5. Alternatively, you can do this all in SQL if you like SQL.
6. If you like to mix and match you can also go back to SQL from an RDD:  
   <https://spark.apache.org/docs/latest/sql-programming-guide.html#interoperating-with-rdds>
7. Ask me or David if you get stuck.