clean_files.sh

```
sed -i '34, 38d' /home/Documents/Hadoop/Insurance_Data_Set/data/*.csv.0
sed -i '1, 9d' /home/Documents/Hadoop/Insurance_Data_Set/data/*.csv.0

sed -i '73, 77d' /home/Documents/Hadoop/Insurance_Data_Set/data/*.csv.1
sed -i '1, 9d' /home/Documents/Hadoop/Insurance_Data_Set/data/*.csv.1

sed -i '33, 37d' /home/Documents/Hadoop/Insurance_Data_Set/data/*.csv.2
sed -i '1, 9d' /home/Documents/Hadoop/Insurance_Data_Set/data/*.csv.2

sed -i '52, 56d' /home/Documents/Hadoop/Insurance_Data_Set/data/*.csv.3
sed -i '1, 9d'/home/Documents/Hadoop/Insurance_Data_Set/data/*.csv.3

sed -i '40, 43d' /home/Documents/Hadoop/Insurance_Data_Set/data/*.csv.4
sed -i '1, 9d' /home/Documents/Hadoop/Insurance_Data_Set/data/*.csv.4

sed -i '73, 77d' /home/Documents/Hadoop/Insurance_Data_Set/data/*.csv.5
sed -i '1, 9d' /home/Documents/Hadoop/Insurance_Data_Set/data/*.csv.5
```

insurance_clean_data.py

```
import xlrd
import csv
import glob
import pandas as pd

def List_All_CSV_Files_From_Folder(filePath, fileType):
      allFiles = glob.glob(filePath + fileType)
      list_ = []
      for file_ in allFiles:
            #Call Append_Year_To_CSV method
            #Append_Year_To_CSV(file_, file_)
            clean_data(file_)
#Add the root data to all child data for processing
def clean_data(inputFileName):
      content = []
      total_per = []
      dataframe = pd.read_csv(inputFileName, usecols=[0,1], header=None)
      privious_data=''
      for index, row in dataframe.iterrows():
            data, per, i = row[0], row[1], index
            if data[0] == ' ':
                  data = privious_data + data.strip()
            else :
                  privious_data = data
            content.append(data)
            total_per.append(per)
      #Create DataFrame with required columns only
      CleanedDataFrame = pd.DataFrame({'Insurance_Description':content,
'Total_Per': total_per})
      #Create the csv file using required columns
      #print 'outputFileName : ' + outputFileName
```

```
      CleanedDataFrame.to_csv(inputFileName+'.csv', index=None)

def Process_File():
      #print 'Started Processing files..'
      filePath =r'/home/Documents/Hadoop/Insurance_Data_Set/data' # use your path
      fileType = '/*.csv.3'
      List_All_CSV_Files_From_Folder(filePath, fileType)

Process_File()
```

insurance_assignment1.pig

```
insuranceData2001 = LOAD
'/home/Documents/Hadoop/Insurance_Data_Set/data/cleandata/2001.csv.3.csv' USING
PigStorage(',') AS (description:chararray, total: double);
--Fetch data for People aged 18-64..Privately insured (alone or  in combination)
data2001 = FILTER insuranceData2001 BY description == 'People aged 18-64..Privately
insured (alone or  in combination)';
--dump data2001;

insuranceData2002 = LOAD
'/home/Documents/Hadoop/Insurance_Data_Set/data/cleandata/2002.csv.3.csv' USING
PigStorage(',') AS (description:chararray, total: double);
--Fetch data for People aged 18-64..Privately insured (alone or  in combination)
data2002 = FILTER insuranceData2002 BY description == 'People aged 18-64..Privately
insured (alone or  in combination)';
--dump data2002;

insuranceData2003 = LOAD
'/home/Documents/Hadoop/Insurance_Data_Set/data/cleandata/2003.csv.3.csv' USING
PigStorage(',') AS (description:chararray, total: double);
--Fetch data for People aged 18-64..Privately insured (alone or  in combination)
data2003 = FILTER insuranceData2003 BY description == 'People aged 18-64..Privately
insured (alone or  in combination)';
--dump data2003;


insuranceData2004 = LOAD
'/home/Documents/Hadoop/Insurance_Data_Set/data/cleandata/2004.csv.3.csv' USING
PigStorage(',') AS (description:chararray, total: double);
--Fetch data for People aged 18-64..Privately insured (alone or  in combination)
data2004 = FILTER insuranceData2004 BY description == 'People aged 18-64..Privately
insured (alone or  in combination)';
--dump data2004;

insuranceData2005 = LOAD
'/home/Documents/Hadoop/Insurance_Data_Set/data/cleandata/2005.csv.3.csv' USING
PigStorage(',') AS (description:chararray, total: double);
--Fetch data for People aged 18-64..Privately insured (alone or  in combination)
data2005 = FILTER insuranceData2005 BY description == 'People aged 18-64..Privately
insured (alone or  in combination)';
--dump data2005;
```

```
insuranceData2009 = LOAD
'/home/Documents/Hadoop/Insurance_Data_Set/data/cleandata/2009.csv.3.csv' USING
PigStorage(',') AS (description:chararray, total: double);
--Fetch data for People aged 18-64..Privately insured (alone or  in combination)
data2009 = FILTER insuranceData2009 BY description == 'People aged 18-64..Privately
insured (alone or  in combination)';
--dump data2009;

insuranceData2010 = LOAD
'/home/Documents/Hadoop/Insurance_Data_Set/data/cleandata/2010.csv.3.csv' USING
PigStorage(',') AS (description:chararray, total: double);
--Fetch data for People aged 18-64..Privately insured (alone or  in combination)
data2010 = FILTER insuranceData2010 BY description == 'People aged 18-64..Privately
insured (alone or  in combination)';
--dump data2010;

insuranceData2011 = LOAD
'/home/Documents/Hadoop/Insurance_Data_Set/data/cleandata/2011.csv.3.csv' USING
PigStorage(',') AS (description:chararray, total: double);
--Fetch data for People aged 18-64..Privately insured (alone or  in combination)
data2011 = FILTER insuranceData2011 BY description == 'People aged 18-64..Privately
insured (alone or  in combination)';
--dump data2011;

sub_data2001 = FOREACH data2001 GENERATE total;
sub_data2002 = FOREACH data2002 GENERATE total;
sub_data2003 = FOREACH data2003 GENERATE total;
sub_data2004 = FOREACH data2004 GENERATE total;
sub_data2005 = FOREACH data2005 GENERATE total;
sub_data2009 = FOREACH data2009 GENERATE total;
sub_data2010 = FOREACH data2010 GENERATE total;
sub_data2011 = FOREACH data2011 GENERATE total;

result = UNION data2011, data2010, data2009, data2005, data2004, data2003,
data2002, data2001;
grp= Group result all;

insurance_average = FOREACH grp GENERATE (double)AVG(result.total) AS
Total_Average;
dump insurance_average;
```

Insurance_assignment2.pig

```
insuranceData2001 = LOAD
'/home/Documents/Hadoop/Insurance_Data_Set/data/cleandata/2001.csv.3.csv' USING
PigStorage(',') AS (description:chararray, total: double);
--Fetch data for People aged 65 and older..Publically insured (no private)
data2001 = FILTER insuranceData2001 BY description == 'People aged 65 and
older..Publically insured (no private)';
--dump data2001;

insuranceData2002 = LOAD
'/home/Documents/Hadoop/Insurance_Data_Set/data/cleandata/2002.csv.3.csv' USING
PigStorage(',') AS (description:chararray, total: double);
```

```
--Fetch data for People aged 65 and older..Publically insured (no private)
data2002 = FILTER insuranceData2002 BY description == 'People aged 65 and
older..Publically insured (no private)';
--dump data2002;

insuranceData2003 = LOAD
'/home/Documents/Hadoop/Insurance_Data_Set/data/cleandata/2003.csv.3.csv' USING
PigStorage(',') AS (description:chararray, total: double);
--Fetch data for People aged 65 and older..Publically insured (no private)
data2003 = FILTER insuranceData2003 BY description == 'People aged 65 and
older..Publically insured (no private)';
--dump data2003;


insuranceData2004 = LOAD
'/home/Documents/Hadoop/Insurance_Data_Set/data/cleandata/2004.csv.3.csv' USING
PigStorage(',') AS (description:chararray, total: double);
--Fetch data for People aged 65 and older..Publically insured (no private)
data2004 = FILTER insuranceData2004 BY description == 'People aged 65 and
older..Publically insured (no private)';
--dump data2004;

insuranceData2005 = LOAD
'/home/Documents/Hadoop/Insurance_Data_Set/data/cleandata/2005.csv.3.csv' USING
PigStorage(',') AS (description:chararray, total: double);
--Fetch data for People aged 65 and older..Publically insured (no private)
data2005 = FILTER insuranceData2005 BY description == 'People aged 65 and
older..Publically insured (no private)';
--dump data2005;

insuranceData2009 = LOAD
'/home/Documents/Hadoop/Insurance_Data_Set/data/cleandata/2009.csv.3.csv' USING
PigStorage(',') AS (description:chararray, total: double);
--Fetch data for People aged 65 and older..Publically insured (no private)
data2009 = FILTER insuranceData2009 BY description == 'People aged 65 and
older..Publically insured (no private)';
--dump data2009;

insuranceData2010 = LOAD
'/home/Documents/Hadoop/Insurance_Data_Set/data/cleandata/2010.csv.3.csv' USING
PigStorage(',') AS (description:chararray, total: double);
--Fetch data for People aged 65 and older..Publically insured (no private)
data2010 = FILTER insuranceData2010 BY description == 'People aged 65 and
older..Publically insured (no private)';
--dump data2010;

insuranceData2011 = LOAD
'/home/Documents/Hadoop/Insurance_Data_Set/data/cleandata/2011.csv.3.csv' USING
PigStorage(',') AS (description:chararray, total: double);
--Fetch data for People aged 65 and older..Publically insured (no private)
data2011 = FILTER insuranceData2011 BY description == 'People aged 65 and
older..Publically insured (no private)';
--dump data2011;
```

```
sub_data2001 = FOREACH data2001 GENERATE total;
sub_data2002 = FOREACH data2002 GENERATE total;
sub_data2003 = FOREACH data2003 GENERATE total;
sub_data2004 = FOREACH data2004 GENERATE total;
sub_data2005 = FOREACH data2005 GENERATE total;
sub_data2009 = FOREACH data2009 GENERATE total;
sub_data2010 = FOREACH data2010 GENERATE total;
sub_data2011 = FOREACH data2011 GENERATE total;

result = UNION data2011, data2010, data2009, data2005, data2004, data2003,
data2002, data2001;
grp= GROUP result ALL;

insurance_average = FOREACH grp GENERATE (double)AVG(result.total) AS
Total_Average;
DUMP insurance_average;
```