Project 2 Report

Introduction:

I would like to build a single-index model for stocks price changes to evaluate the stocks' risk. The financial concept "high risk high return" means that the investors will be compensated for bearing high risk by higher return. However, some risks are firm specific risk that can be eliminated through portfolio diversification. Investors will only be compensated for the risk they bear that can not be eliminated by diversification, which is the systematic risk(also called market risk).

Basically, the project is running a regression of the return of stocks against the market return(approximated by S&P500) to find the correlation coefficient of a certain stock's return and the market changes. The correlation coefficient represents the systematic risk of the stock. By generating graphs of regression, one can judge the market risk from the slopes.

Details:

Step 0: I will get the historical price data for interested stocks, S&P500 and T-bills in Yahoo Finance. (Nice clean excel files)
Step 1: The daily return rate list(Rt=Pt-Pt-1/Pt) will be calculated for each of the stocks and S&P500.
Step 2: Run the regression for each stock with S&P500.
Step 4:Generate regression graphs as well as a table presenting the correlation coefficient for each stocks.

Data structure:

The data was stored in a big Dictionary, which is a dictionary of dictionaries.
Each sub dictionary in the big Dictionary has the same structure that has dates as keys and return rate as it's value.
I chose dictionary to store the data. I want to store the dates of the data, so I can mix and relate data from the same date, instead of separate them apart.

Difficulties I met:

1.The most difficult problem I met is matching data form from different source.
At first place, I planned to deduct a risk free rate for both stock specific return and market return to make it fit financial theory better. I planned to build two dictionaries with "dates" as their keys. The dates looks the same in excel file("1/2/14"). After painful debugging, I found out one of them is "1/2/14" while the other one is "2014-01-02" in TestEdit. I tried hard to change them by functions like:

- s.split(/) or s.split(-)
  Because strings are immutable while lists are mutable

```
#This function take in an individual distorial data file and the risk free dictionary
#The function returns a dictionary which has the date as keys, and the corresponding ecessive return as it's value
def individualDict(line,RF):
    subDict={}
    file=open(line)
    file.readline()
    List=[]
    for k in file:
        k=k.split(",")
        List.append(k)
    for i in range(len(List)-1):
        a=changeName(List[i][0])
        subDict[List[i][0]]=(float(List[i+1][5])-float(List[i][5]))/float(List[i][5])-RF[a]
        print(subDict[List[i][0]])
    return subDict

def changeName(s):
    NameList=s.split(-)
    mm=NameList[1].lstrip(0)
    dd=NameList[2].lstrip(0)
    yy=NameList[2][1:]
    key="mm"+"/"+"dd"+"/"+"yy"
return key
```

- s.lstrip(0)
  I use this function to turn the month/date(like 01/02/2014) into the other form without changing type from string and integer back forth.

However, it still didn't work out for unknown reasons. So I cannot connect the daily risk free return data with return data for each stock.

2.I still have difficulty in modifying the code from scikit.

I tried to input data from the dictionaries I built to array needed to run regression but it didn't work out.

How to run?

import matplotlib.pyplot as plt
import numpy as np
from sklearn import datasets, linear_model
from sklearn.metrics import mean_squared_error, r2_score