**Assignment 3:** *Intro to Jupyter Notebook and ML for data analysis using scikit*

**GOAL:**

Build a predictive machine learning model that will use historic data of stock prices to make recommendations as to which stocks to invest. Along the way, you’ll get comfortable with using Jupyter Notebooks to document your work, using scikit-learn to train and test basic ML models, and start thinking about the strengths and weaknesses of models like this.

In this assignment we’ll get started with a classic ML problem: *determining the stock price of a firm based on a variety of factors*.

**Background**: You have just started your career in a local bank’s investment department. The department is new, and does not have much experience participating in the stock markets. Your manager has assigned you do conduct some financial analysis regarding the behavior of stock prices on the stock market, and provide recommendations as to which stocks to invest.

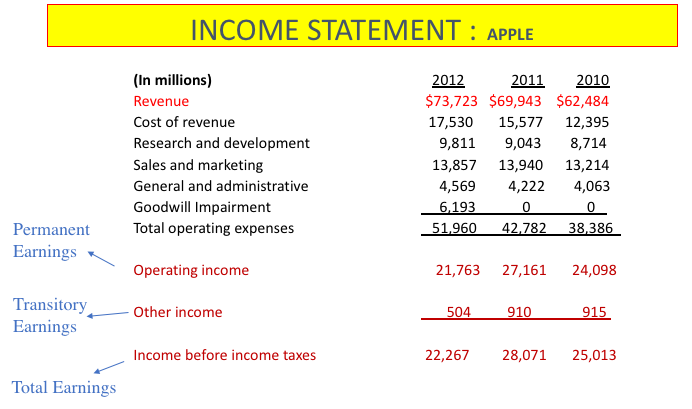
Your firm is interested in purchasing some stock with the intention of holding for 1 year and benefiting from capital gains. You are given the task of analyzing the relationship between the earnings of a firm and its stock price to provide insight to help the department decide on which firms to invest in. Your manager has provided you with the stock price and earnings information for 100 firms in the market this year for use in the analysis (stocktrainingdata.csv). Each firm is assigned a firm unique firm number. Use this firm number when recommending firms for investment. **Your report should generate insights into 4 questions: 1) What is the relation between earnings and stock price. 2) Which firms have a stronger relationship between earnings and stock price. 3) Why do these firms have stronger relationships between earnings and stock price. 4) Which firms would you recommend to invest in, and why.**

The total earnings of a firm consists of two kinds of earnings as below.

Permanent Earnings: primary earnings are profit from a firm’s primary business. For the Apple company, the primary earnings come from the selling of Apple product.

Transitory Earnings: profit from a side business. For the Apple company, besides selling products, they also profit from investing in stock or bond market. The profit from such investment other than selling the product are called as transitory earnings because of its non-recurring nature.

If you read the income statement of a firm, you could find values of these earnings correspondingly. For example, the picture below lists the income of APPLE from 2010 to 2012(note: just one example, not real).



Setup Logistics:

Checklist:

1. Install Python 3.7 or greater (ideally w/ Anaconda 64 bit)[[1]](#footnote-0)
2. Install Jupyter Notebook
3. Install packages

This will be written assuming use of Python 3.7 or greater. We **strongly** recommend using [Anaconda](https://www.anaconda.com/) for this. You’ll also need to install [Jupyter Notebook](http://jupyter.org/install) (which comes with Anaconda) and [scikit-learn](http://scikit-learn.org/stable/install.html), [matplotlib](https://matplotlib.org/), [pandas](https://pandas.pydata.org/), and [numpy](http://www.numpy.org/). We’re not going to lead you through the installation of these, but there are many YouTube tutorials that go through the process step-by-step.

If you’re new to Jupyter Notebook, there’s a decent (though slightly dated) walkthrough [here](https://www.youtube.com/watch?v=HW29067qVWk). The purpose of using Jupyter Notebook here is both to document all of your steps and to allow you to mix code, output, and commentary in the same space.

Assignment Pieces

There are five files in the Assignment 3 folder:

* **stocktraningdata.csv** - the data you will train your model on. This is data showing a variety of earnings about a list of firms and their stock prices.
* **stocktraningtest.csv** - the data you’ll use to test your model, to see how good your model is at predicting the stock price.
* **variables.txt** - a list of the variables in the files and explanations of what they mean
* **CIS8005-Fall21-Assignment3.ipynb** - a starter Jupyter Notebook that will list the rest of this assignment.

Making a model

The remainder of this assignment will be in the **CIS8005-Fall21-Assignment3.ipynb** notebook. Save a copy of this file for yourself, rename it, and fill in the missing pieces and answer the questions as directed.

Please download your final notebook as .ipynb, and submit to the iCollege site with name **CIS8005-Fall21-IA3**-yourFirstInitialLastName.**ipynb**, and a shared google link to your ipynb file in the comment.

1. If you use Anaconda 32 bit you’ll probably run into trouble later on when we get into importing big data. [↑](#footnote-ref-0)