MIS 407 PROJECT 1

**Team MBDW**: MEYERS, WALSH, DIERCKS, BEIGHOEFER

**Data Sources**:

* Quandl.com
* Different City data portals
* Data.gov
* InfoChimps
* Social Media sites: Social Data
* The US Census Bureau

**Project Question**:

As a group we looked into the question of how employment affects high school students. We also wanted to see how the same high school students are affected by their parent’s employment/family income.

The **Data sources** we chose to use to answer this question include:

* <https://www.quandl.com/data/NCES/DROPOUT_RACE-Dropout-Rate-by-Race>
* <https://www.quandl.com/data/ECPI/MEDIAN_FAMILY_INCOME-Median-family-income-1947-2010>
* <https://www.quandl.com/data/NCES/HIGHSCHOOL_EMPLOY_TOTAL-Percentage-of-High-School-Students-aged-16-or-older-who-are-Employed-by-hours-worked-Total>

We found that these three data sets were enough to quantify the answers to the proposed question listed above.

**Reading the data into a Pandas DATAFRAME:**

Since we are using three different datasets to answer our question, we will need to create three different DATAFRAMES. We would start by importing "Pandas as pd," then creating a variable that is the file path for each excel dataset we are using. We will then use the Pandas "read\_excel" function to read the correct excel pages we want the data from, and save the result into our "DATAFRAME" variables (df\_drop, df\_income, df\_employment are the ones we will use).

**What columns should we use:**

* In the "Percentage of High School Students aged 16 or older who are Employed..." dataset, we will be using the "Date" column, along with the "Total Employed" column.
* In the "Median Family Income" dataset, we will be using both the "Date" column, and the "Dollars" column.
* In the "Dropout Rate by Race for High School Students" we will be using the "Date" and "Total" columns.

**Outline of tasks to answer the question:**

* We will need to import all three datasets into DATAFRAMES using Pandas.
* Clean up the data so we are able to use it properly (Data Munging)
* Since we want to calculate the slope of the line at different times between 1990 and 2010 (X-axis values = Dates), we need to figure out intervals we want to use. - **We will be using: 1990-1995, 1995-2000, 2000-2004, 2004-2006, and 2006-2010.**
* Once we decided what intervals (points of the line) we want to use, we can calculate the slope of the line for each dataset (we will be calculating the slope of the line 5 times for each dataset, and we are using three datasets). 5 intervals times 3 data sets is 15 slopes being calculated.
* Once we have the slopes, we can compare the three slopes of the same date interval across datasets to see if there is a correlation between them.

**- Example**: if the slope for high school employment is negative between 1990-1995 (less high school students have jobs), while the slope for high school dropout rate is also negative between 1990-1995 (less high school students dropout), and if the median family income between 1990-1995 is positive (meaning there is an increase in the average family income), there might be a correlation saying the more income a family has, there will be less need for a high school age child to bring their family an income, meaning they focus on school and lower the dropout rate. \*We might only be able to say this if the positive/negative trends continue as expected in all intervals, not just one of them.\*