**ETL Project:   
*Reddit and its influence on the GME and AMC stocks***

**Participants:**   
Fernando Núñez   
Uriel Arriaga

Carlos Mendoza   
María Inés Oñate

Overview/abstract:   
  
WallStreetBets (r/wallstreetbets, also known as WSB), is a subreddit where participants discuss stock and option trading. This subreddit recently came into visibility given the influence this subreddit had in stock prices.   
  
We wanted to discover if the volume of posts mentioning specific stocks directly impacted the fluctuation these stocks had each day. To do this, we necessarily had to migrate raw data into a production database, use an API to extract the closing and opening prices for the stocks we wanted to analyze and correlate them with the number of mentions that particular stock had each day.

Extraction

The CSV we opted for includes Reddit posts that span from January 28th 2021 to May 21st 2021 and contain information or mentions related to WallStreetBets (WSB).

We used Yahoo Finance to establish an API connection and thus download the daily opening and close for each day, in order to correlate mentions with stock fluctuation. This can be found within the *Tests* folder in the ipynb named Yahoo \_Finance.ipynb

We obtained raw data from

* Reddit Wall Street Bet posts // Source: Kaggle https://www.kaggle.com/gpreda/reddit-wallstreetsbets-posts
* Selected stock information from Google Finance (GME, AMC)

Source: Yahoo Finance data reader

Transform

Given that our main data source had unfiltered and excessive information, we wanted to filter and clean the existing posts by those who directly mention the stocks GME and AMC. With this data, we then paired posts that mentioned the stocks with the day in which they had been posted. After doing this, we were able to use an API call to gather information from the opening and closing prices for said stocks for each day in which posts mentioning these stocks had been published.   
  
**How did we do it? *Step by step***   
  
In order to find the stock mentioned within the CSV, we created a list that contained the tickers we needed (GME and AMC) and we used a *for* to associate the list with a respective timestamp. We then created a list of the number of comments per day for each ticker.

For the API call we used Yahoo Finance and Pandas Data Reader to import the data related to the stocks we were interested in. So we created a list with the tickers we needed and created a list for the opening and closing prices and. We limited the API call to the starting and end dates that we are considering for the Reddit posts within the CSV and used date as an index.   
  
We substituted empty spaces with NaN and created a single data frame that shows us the stock price each day for the opening and closing and its associated volume. We then had to change strings to numbers in order to round the stock prices and make the data more manageable. Finally, we had to align and join the closing data with our new data set and add opening prices for the stock.   
  
  
Load  
  
We chose MongoDB given that we have a non-relational database.

Using pymongo connection from pandas to connect to mongodb (load.ipynb)

mongodb+srv://<username>:<password>@cluster0.skvrw.mongodb.net/test  