**Project 3 – Data Engineering**

**Project Members:**

Manal Bayoumi, Rocio Cantu, and Melissa Mosby.

**Project Overview:**

Our project will analyze historical stock data for five companies: Alphabet Inc. (aka Google), Apple Inc., Intel Corporation, Microsoft Corporation, and Oracle Corporation. We selected to do this analysis with the data engineering track. We will employ jupyter notebook and pandas to capture and modify our data using ELT workflows. Next, we will feed our data into SQLite. Also, we will generate an ERD diagram to show the relationships of our database tables. We will create a method to read our data from the database with pandas DataFrame.

**Our project will determine various questions when analyzing these stocks:**

* How do investors pick their stocks?
* When buying and selling stocks, do investors depend on the opening or the closing prices of stocks?
* What factors make stock prices rise or fall for a company like Apple, Google, Intel, Oracle, Microsoft?

**Required Package Download**

*Please run the following command in jupyter or bash terminal*

!pip install yfinance

!pip install pandas-datareader

!pip install streamlit --upgrade

!pip install plotly

**Instructions**

**Part One: ETL.ipynb**  
This is a jupyter notebook file, with this file, user can download stock data from Yahoo finance API and then save to SQLite3 database.

Data from API includes:

* Data
* Open
* Close
* High
* Low
* Volume

We used ETL to transform the data by:

* Adding the following columns:
  + stock\_ticker;
  + normalization, which compares the different stocks gain and loss;
  + percentage\_change, which determines the best and worst day to invest; and
  + accumulative\_return, which determines the investment returns from the market.
* Changing datatype for database to record.
* Removing a redundant floating point.

Here is an example of Apple’s original data:

<<KEEP THE SCREENSHOT>>

Here the example of Apple’s transformed data:

<<KEEP THE SCREENSHOT>>

Here is the ERD diagram:

<<KEEP THE SCREENSHOT>>

**Part Two: Database to Pandas**

We used a jupyter notebook file, where the data was loaded into a database Group\_Project\_3.db. The user can extract data using query tools or python sql packages.  
**With the sample code below, the user should use pandas and the sqlite library to extract data from the database and display the data with the dataframe.**

**<<KEEP THE CODE>>**

**<<KEEP THE DATAFRAME SAMPLE>>**

**<<KEEP THE PLOTTING SAMPLE>>**

**From the pictures above, some conclusions can be made:**

* **Google stock had a total increase in 2023.**
* **The greatest gains of 7%, occurred on February 2, 2023.**
* **The greatest loss of 9%, occurred on October 25, 2023.**

**Part Three: Groupwebsite.py**

For a better user experience, our group built an interactive webpage with a streaming package called streamlit. This package was not covered in class. This package uses the javascript plotly package build web.

In order to install the package, please install the following package:

<<KEEP THE CODE>>

Then use the following code:

<<KEEP THE CODE>>

**There are several reasons why my group decided to use SQLite:**

* SQLite is a SQL based database that can create a single database file that can easily be uploaded to GitHub and shared with other group members.
* SQLite facilities the work when using multiple databases.
* We can DB Browser read SQLite file and check if the file performs well.
* SQLite is small like a paper; however, the other SQL is a big library.
* Download DB Browser at this link <https://sqlitebrowser.org>.

**Why SQL is better than MongoDB?**

* Using SQL over NoSQL in that project is better because stock databases are saved in a table form. SQL can perform analytical queries, such as filters, joins, merges, and aggression on the data.
* SQL databases are better for multi-row transactions, while NoSQL is better for unstructured data (Smallcombe, 2023)

**Project Ethical Considerations:**

* Our project used public historical stock data. When considering ethics, we should handle the data with integrity. We should be transparent about any factors that could influence their analysis, interpretation, or reporting of results.

**References**

* “Alphabet Inc. (GOOG) Stock Historical Prices & Data.” Yahoo! Finance, Yahoo!, 16 Jan. 2024, <https://finance.yahoo.com/quote/GOOG/history?p=GOOG>
* “Apple Inc. (AAPL) Stock Historical Prices & Data.” Yahoo! Finance, Yahoo!, 16 Jan. 2024, <https://finance.yahoo.com/quote/AAPL/history?p=AAPL>
* “Intel Corporation Stock. (INTC) Stock Historical Prices & Data.” Yahoo! Finance, Yahoo!, 16 Jan. 2024, <https://finance.yahoo.com/quote/INTC/history?p=INTC>
* “Microsoft Corporation Stock. (MSFT) Stock Historical Prices & Data.” Yahoo! Finance, Yahoo!, 16 Jan. 2024, <https://finance.yahoo.com/quote/MSFT/history?p=MSFT>
* “Oracle Corporation Stock. (ORCL) Stock Historical Prices & Data.” Yahoo! Finance, Yahoo!, 16 Jan. 2024, <https://finance.yahoo.com/quote/ORCL/history?p=ORCL>
* Smallcombe, Mark. “SQL VS NOSQL: 5 Critical Differences.” Integrate.Io, 9 Nov. 2023, [www.integrate.io/blog/the-sql-vs-nosql-difference/#two](http://www.integrate.io/blog/the-sql-vs-nosql-difference/#two).