

ETL PROJECT

MARCH 11TH 2020

**─**

**Team 3**

Sneha

Jalil

Monica

Erika

Clinton

Radhakishore

# Proposal:

**Show a correlation/link between daily confirmed cases of COVID-19 to stock prices of Major international hotels and Airlines.**

# Extract Process:

Original data sources were retrieved from direct websites listed below where we were able to download direct CSV files.

## Sources:

COVID-19 Confirmed Cases:

* <https://github.com/CSSEGISandData/COVID-19> ; <https://docs.google.com/spreadsheets/d/1yZv9w9zRKwrGTaR-YzmAqMefw4wMlaXocejdxZaTs6w/htmlview?usp=sharing&sle=true>

Hotels:

* HLT Hilton Worldwide Holdings Inc. Common Stock
  + <https://finance.yahoo.com/quote/HLT/history?p=HLT>
* VAC Marriott Vacations Worldwide Corporation Common Stock
  + <https://finance.yahoo.com/quote/VAC/history?p=VAC>

Airlines:

* DAL , Delta Air Lines, Inc. Common Stock
  + <https://finance.yahoo.com/quote/DAL/history?p=DAL>
* SINGF, Singapore Airlines Ltd
  + <https://finance.yahoo.com/quote/SINGY/history?p=SINGY>

# Transform (Clean Up) Process:

1. Download the stock prices to a csv file.
   1. Clean data to keep only the latest 30days
   2. Pulled in each stock ticker into its’ own dataframe
   3. Had to reformat date and bring in ticker symbol
   4. Merged all data frames together
   5. Created a new dataframe from the large merged to only bring in: adjusted closing value, ticker symbol, and date
   6. Moved Pandas over to Mongo
   7. Stock prices not available for Saturday/Sunday
2. Clean up of 30 days COVID-19 data.
   1. Convert each csv file to a data frame
   2. Loaded all 30 files as data frames then concactonated them to one dataframe
   3. Slice the column “last column” by removing the timestamp
   4. Dropped unnecessary columns to keep only “confirmed”, “death”, “last updated”, and “recovered”
   5. Rearranged and renamed the columns
   6. We then grouped by and summed by the date column
   7. Dropped all rows before the date “2/6”

# Load Process:

1. Had 2 dataframes, 1 for covid virus counts by date and the other stock adjusted closing price for hotels and airlines
2. We created a Covid database in Mongo DB and two collections; one for the stocks data frame and the other for the covid cases
3. To load the data we had to convert the individual data frames into dictionaries
4. Had to read each row from the dictionary and inserting into Mongo DB
5. Lastly, had to convert the date to string