Big data final project Report

1) Group name

Stay-at-home flyers

2) Project participants

- Xucheng Tang <xt544@nyu.edu>
- Jiaying Li <jl10919@nyu.edu>
- Ke Chen <kc4152@nyu.edu>

3) Project description: State the problem you will address and give an overview of your approach

Problem:

What's the impact of "lockdown" policy?

- Economy:
  - Stock market change in different industries
    travel / retail / online meeting / game / delivery / hotel / etc.
  - unemployment / employment in different industries

- Life:
  - Mobility:
    Mobility changes before and after lockdown policy released in different cities
  - Hot topic category on twitter

Approach:

- Data cleaning: filter missing data and noisy data
- Data transformation: normalization, integration
- Data reduction: combination, aggregation
- Data analysis
- Data visualization

4) List of datasets (together with a link to where they can be found),a description of each dataset.

- Daily Case Report from JHU and WHO
  https://github.com/CSSEGISandData/COVID-19
- Crowd-sourced COVID-19 Dataset Tracking Involuntary Government Restrictions
  https://rexdouglass.github.io/TIGR/TIGR_landing_page.nb.html
- Stock market
  https://finance.yahoo.com/
- Unemployed persons by industry and class of worker
  https://www.bls.gov/webapps/legacy/cpsatab14.htm
- Mobility changes in response to COVID-19
  https://github.com/descarteslabs/DL-COVID-19
- Apple Mobility Trends Reports
  https://www.apple.com/covid19/mobility
- Public tweets by topic category related to COVID-19
  https://github.com/citibeats-labs/covid19_bid_data
- Covid-19 Twitter dataset by topic keyword related to COVID-19
  https://github.com/thepanacealab/covid19_twitter

5) Data Cleaning and Integration: describe the steps you performed to clean and integrate the datasets, and discuss the challenges you faced. Include a link to your github repo.

Github Repo: https://github.com/HaroldTang/Covid19_LockDown_Project

Instructions:

1. Spark files need to be run in dumbo, and get merged results from HDFS

2. Jupyter Notebook results should be rerun to get the output result

Data Cleaning Step:

1. Collect data from multiple resources and decide how to interpret the data.
2. Find out the relationship between data and aggregate several datasets in the same administration level. For example, we found the connection between the reduction of outdoor activities and the government restriction on certain states.
3. Use Spark to change the data format, clean the data by deleting unnecessary whitespaces and punctuations. Then we rearrange the structure of the datasets like pivoting the columns with rows to get a more uniformed format so that the data can be processed easier later.
4. Use Pandas to deal with different types of data. Design a multi-joint and merge several datasets to get meaningful output. Jupyter notebook provides easy-accessed data visualization, improving working efficiency.
5. To sum up, we mainly deal with the following data problems:
    a. replace the null value with the proper value
    b. Delete duplicate records
    c. Perform inner and outer join among multiple tables
    d. Select valuable data (column) and drop redundant data (column)
    e. Redesign dataset format

Challenges:

1. The data format is muddled so we have to think of a unified standard output for every dataset we gathered from multiple resources. We discussed the format and finally concluded. The final output should be more concise for us to move forward.
2. The data cleaning part is confusing, we have to compare several resources to trust one of them, the data authority is very essential to us since we may have opposite results if we don't verify the data resources. We decided to trust some reliable reports from big tech companies with their GPS data gathered from people's phones rather than trust the voluntary survey online.
3. Some dataset contains multiple aspects of data, like 'google_mobility-county.csv', 'DL-us-samples-by-county.csv', which makes the logic of the datasheet ambiguous. To make it easier to understand, we redesigned the dataset format, trying to represent three-dimensional data on a two-dimensional datasheet.

4. When trying to join multiple tables via pandas, we tried concat, join, merge functions. However, we compared the results of different operations and still found them undesirable. At last, we managed to get an ideal output using reduce function.