**Guidelines for ETL Project**

**Project Members:**

* Alex Ledger
* Andy Shi
* Edith Jimenez
* Rohan Kancharla
* Joseph Picca

**## Finding Data**

Your project must use 2 or more sources of data. We recommend the following sites to use as sources of data: [data.world](<https://data.world/>) or [Kaggle](https://www.kaggle.com/)

You can also use APIs or data scraped from the web. However, get approval from your instructor first.

**## Data Cleanup & Analysis**

Once you have identified your datasets, perform ETL on the data. Make sure to plan and document the following:

\* The sources of data that you will extract from.

\* The type of transformation needed for this data (cleaning, joining, filtering, aggregating, etc).

\* The type of final production database to load the data into (relational or non-relational).

\* The final tables or collections that will be used in the production database.

You will be required to submit a final technical report with the above information and steps required to reproduce your ETL process.

**## Project Report**

At the end of the week, your team will submit a Final Report that describes the following:

\* \*\*E\*\*xtract: your original data sources and how the data was formatted (CSV, JSON, pgAdmin 4, etc).

\* \*\*T\*\*ransform: what data cleaning or transformation was required.

\* \*\*L\*\*oad: the final database, tables/collections, and why this was chosen.

**Project Proposal**

**Coronavirus Datasets**

***Overview***

We plan to utilize South Korean coronavirus data from kaggle.com, with two primary datasets, one involving patient metadata and another involving patient movement data.

***Finding Data***

Data Source: [https://www.kaggle.com/kimjihoo/coronavirusdataset](https://www.kaggle.com/kimjihoo/coronavirusdataset#route.csv)

***Data Cleanup & Analysis***

* All of our data’s original format is in .csv files.
* We plan to merge our patients file with our routes file via the patient ID column
* We also plan to create new data via transformations on the columns
* *These transformations will be performed in Python/Pandas*

***Final Database***

* The joined data will be loaded to a postgres database via pandas.

***Potential Extra Work***

* Create a front-end web visualization with statistics from the dataset
* Create some visualization to show the data behavior