Amit & Olga’s Proposal:

We will analyze the coronavirusdataset from KCDC (Korea Centers for Disease Control & Prevention) available from Kaggle

Steps:

1. Load two files, patient and patient route csv files into SQL database using Postgres.
2. Define the fields desired from both files.

From Patient file:

1. Patient id
2. Sex
3. Birth year (for age)
4. Country
5. Region
6. Infection reasons
7. Contact number (number of contacts with people)
8. Confirmed date
9. Release date
10. Decease date
11. State (isolated/released/deceased)

Columns not used from patient file

1. Group
2. Order of infection
3. Infected by

From Patient route (all places visited) file:

1. Patient id
2. Date of visit
3. City
4. Visit (type of place visited)

Columns not used from patient route file:

1. Province
2. Latitude
3. Longitude

<https://github.com/amit-dhingra/ETL-Challenge>