Group 4 Project Technical Report 1/12/2021

Austin Barrette

Louise Colleen Cachuela

Charles Madison

Chase Lester

**Hospital Visits Data (3/1/2020 - 3/21/2020) by County**

**ETL Project**

 **E**xtract: your original data sources and how the data was formatted

Our ETL Project is compromised from 2 CSV Files found on Kaggle.com

[**https://www.kaggle.com/jieyingwu/covid19-us-countylevel-**](https://www.kaggle.com/jieyingwu/covid19-us-countylevel-)**summaries?select=hospital\_visits.csv**

hospital\_visits.csv - Hospital Visits by each day from 3/1/2020-3/21/2020 by County in the United States

counties.csv - County Demographics

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

We will be using Postgres for our database

Extract above CSVs into data frames

Transform hospital\_visits data frame

Create filtered data frame for the hospital\_visits from specific columns

Rename the column headers from int format to text format

Renamed columns to match sql

Clean the data by dropping columns and setting the index to fips

Transform counties data frame

Rename the column headers

Renamed columns to match sql

Clean the data by dropping column and setting index to fips

Use .melt function to take our wide format and stack a set of columns into single columns of data. We need to specify the data frame, id variables and the measured variables to be stacked.

Dropped all duplicates from fips column in

 **L**oad: the final database, tables, and why this was chosen.

Selected all data from hospital\_visits and us\_county, to get our table, which is joined on the fips column

In the schema file, there will be a join statement showing how the two tables were related