**ETL Project Report**

Cristy Brust

Hemani Choksi

Olan Johnson

Stefanie Kenyon

Ning Xie

**Extract:** Original data sources were CSV files from data.world and the Guttmacher Institute.

Data representing the duration from cities to abortion clinics:

<https://data.world/the-pudding/driving-times-to-abortion-clinics/workspace/data-dictionary>

Abortion statistics by State:

<https://data.guttmacher.org/states>

**Transform:** We transformed the data using Pandas. We renamed the ‘U.S. State’ column from the data.csv to ‘state’ to match formatting in the cities.csv file and removed the ‘id’ column from the cities.csv. We then merged the two dataframes into one on “state”.

**Load:** We loaded the merged data into a SQL database using Postgres because our data was already structured in table format. But before we loaded the individual cities and healthcare data data sets from the CSVs, two columns – Number of Abortions by State Residence and Number of Abortions by State of Occurrence – the integers needed to be formatted to rid the comma for numbers in 1,000s as the data would not properly load as the presence of the comma in those numbers caused an error when uploading. The Postgres database was confusing the columns containing numbers with commas as String or VARCHAR data types. When the database tables were created, these two columns were given INT data types.