**Produced Water Data**

Produced Water Data was pulled in from a download link from USGS. From there it was preprocessed by removing unnecessary columns and other technical cleaning. This data was used to calculate PCA data and impute missing values. All of these including the original data were exported as CSV files that were used as the sources for the tables in the database. Data was pulled from Database to be used in machine learning models

**Pressure Data**

Pressure data was provided in a text file. The text file was then converted into a CSV file. The CSV was then read into a Pandas DataFrame. From there each row of data was provided with a layer number. The time column was converted into DateTime format to show actual dates. After this the pyproj library in python was used to transform the latitude and longitude values from NAD 1927 State Plane Texas Central FIPS 4203 into the WGS84 coordinate system. This clean DataFrame was then exported as a CSV file to be used as the source of pressure data.

**Injection Volume Data**

The Injection Volumes data was imported from an api (<https://injection.texnet.beg.utexas.edu/api/Export>). Parameters were set so that only data required for the project was pulled in. Once the data was loaded, it was then transformed into a Pandas DataFrame. This Dataframe was then exported as a CSV file to be used as the source of injection volume data.

**Earthquake Data**

The Earthquake data was imported from an api (<https://maps.texnet.beg.utexas.edu/arcgis/rest/services/catalog/catalog_all_flat/MapServer/0/query>). Parameters were set so that only data required for the project was pulled in. The data was loaded in json format. The data was then appended into different lists for each feature that was pulled in using the parameters. These lists were then used to create a Pandas DataFrame. The event\_date column was then converted into DateTime format. Time was then removed from that column so that only Date was kept in that column (*year-month-date*). This clean DataFrame was then exported as a CSV file to be used as the source of earthquakes data.

**Amazon RDS**

The database that was chosen to hold all of the dataset is Amazon RDS. After the database was created, the Psycopg2 and SQLAlchemy libraries in python were used to load the data. First the connection was created between pgadmin and amazon rds. From there a table was created for each dataset that was created (earthquakes, injectionVolumes, pressureData). Each of the CSV files that were created earlier were then loaded into Pandas Dataframes. These DataFrame column names were then edited to match the created table headers. Each of the dataframes was then loaded into their corresponding tables.

**Flowchart**

Flow chart was created to show the process of our data. Data was pulled from USGS and then preprocessed. From there it was then uploaded into the SQL/AWS RDS database. The data was then pulled from the database to be used for various machine learning models as well as to create the dashboard.

