Project 2: Extract Transform and Load

**Group Members:**

* Gustavo Gyotoku
* Rasika Bhonsale
* Jeanie Wu
* Pooja Mahajan

**Topic:**

Health Care - Breast Cancer in the United States

**Data sources:**

• <https://www.kaggle.com/datasets/amandam1/breastcancerdataset?resource=download>

• [https://www.statecancerprofiles.cancer.gov/incidencerates/index.php?  
stateFIPS=00&areatype=county&cancer=055&race=00&sex=2&age=001&stage=999&year=0  
&type=incd&sortVariableName=rate&sortOrder=default&output=0#results](https://www.statecancerprofiles.cancer.gov/incidencerates/index.php?)

**Project Description:**

Extracted data from sources of breast cancer demographics in the United States.   
Dataset contained different data types like integers, strings, and floats.

The “National Cancer Institute” provides a table lookup options such as area, area type, cancer, race & ethnicity, sex, age, stage and year for querying and web-scraping.

The second source comes from Kaggle, which contained research data from different patients with different demographics.

**Process Flow:**

Extract data from CSV file into Pandas

Transform Data

Load Dataframe into Postgres

Extract data from web page into Pandas

Transform Data

Load Dataframe into Postgres

**Process Description:**

* The breast cancer dataset was extracted from two sources:
* Kaggle – (CSV file)
* National Cancer Institute (through web scrapping)
* Data was loaded into a pandas DataFrame and cleaned where necessary. Dropped unnecessary columns and removed null values from the DataFrame.
* Reading/Transforming the CSV file:
* Removed the null values:

1. Patient\_ID
2. Age
3. Tumour\_Stage
4. ER status
5. Date\_of\_Surgery
6. Patient\_Status
   1. Dropped 13 null values from the record

* Web scrapping to DataFrame:
  + Renamed the columns:

1. “Met Healthy People Objective of \*\*\*?” to “objective”
2. “Age-Adjusted Incidence Rate cases per 100000(95% Confidence Interval)” to “age-adjust rate”
3. “CI\*Rank fork;(95% Confidence Interval)" to “CI”
4. “Recent 5-Year Trend‡ in Incidence Rates (95% Confidence Interval)" to “Incidence Rates”

* Dropped columns:

1. objective
2. CI
3. Incidence Rates

* Created a database: “breastcancerdb” in Postgres, then loaded both data sets into the database.
* Created a connection to the database with SQLAlchemy.
* Loaded both DataFrames to Postgres database: “breastcancerdb” in the table “Patient\_record” and “Breastcancer\_record”.
* Loaded the data and verified integrity with Pandas sql functions. No issues were found.