**ETL - Heart Disease**

Dataset Source: ~ Kaggle https://www.kaggle.com/dileep070/heart-disease-prediction-

using-logistic-regression?select=framingham.csv

file name: framingham.csv

**Read and Examine the Dataset**

**Part A ~ Initial ETL of dataframe columns**

Initial Number of Rows:4238

Initial Number of Columns:16

These are the columns and their corresponding data types:

* male int64
* age int64
* education float64
* currentSmoker int64
* cigsPerDay float64
* BPMeds float64
* prevalentStroke int64
* prevalentHyp int64
* diabetes int64
* totChol float64
* sysBP float64
* diaBP float64
* BMI float64
* heartRate float64
* glucose float64
* TenYearCHD int64

Data was filtered to show only Females [male column 0= female 1 = male]

* Number of Rows before filtering: 4238
* Current Number of Rows:2419
* Current Number of Columns:16

645 Null Values were found in the following columns:

Education, cigsPerDay, BPMeds, totChol, BMI, HeartRate, Glucose

* Number of Rows before dropping rows with null values: 2419
* Current Number of Rows after dropping null values:2034
* Current Number of Columns:16

The index is converted as a column, which will eventually be used as a primary key for the database.

* Index was initially converted to a column as index1 , but was eventually deleted as numbering/indexing also included the male population in the dataset.
* Index was converted again as a column, renamed ‘female\_patient\_id’, and only reflects the current number/index of the female population.

“Clean” dataset is currently saved as heart\_disease\_females.csv

File name: heart\_disease\_females.csv

* Current Number of Rows:2034
* Current Number of Columns:15

Subdivided the different columns into four different dataFrames, with all “female\_patient\_id” present in all the dataFrames. The different groupings of the columns are demographics, habits, pre-existing conditions and parameters.

Four other columns were added to the parameters dataframe .

Proposed columns to be dropped prior to MLM

Education

Target: TenYearCHD

**AWS – pgAdmin string connection**

* AWS Management Console

Created RDS

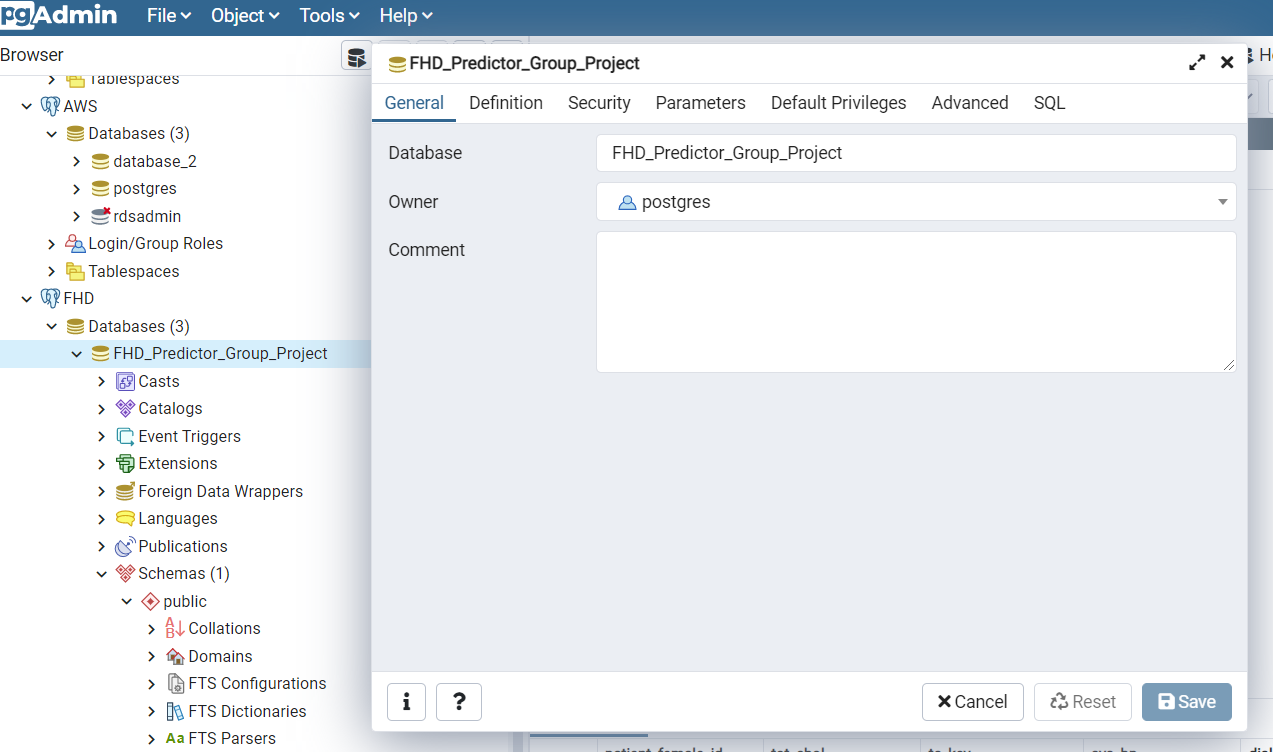
DB Identifier: fhd-Predictor

Initial DB Name FHD\_Predictor\_Group\_Project

* pgAdmin connection

Created Server

Server Name: FHD\_Predictor\_Group+Project

****

* S3 Bucket

Created S3 Bucket

Bucket Name: fhd-bootcamp2021-project

