Report

1. Extract
   1. Extract data from 4 datasets: consumer complaints, states, veterans by age group, veterans by race from different sources using panda. Due to hardship of finding different data formats, we found all csv sources for this project.
2. Transform
   1. Each dataset comes in not in a well-formed format. First, we took the main dataset which is consumer complaints. Look for all columns, drop the columns with null values. Since this dataset has so many different categories, we only focus on Mortgage product and veteran sub-product.
   2. The rest 3 dataset are cleaner and straight forward. We just renamed necessary fields to make more sense on naming convention.
3. Load
   1. From postgres database, we created schema for each dataset.
   2. Using engine from sqlalcheme, each original dataset is loaded to postgres database to each table created in step 1 accordingly
   3. Then we joined all tables together
   4. The result is we have a dataset from all other dataset which has mortgage consumer complaint by veteran by age group and by races per state

|  |  |
| --- | --- |
|  | Select \* from consumer\_complaints  select \* from states |
|  | select \* from vets\_by\_age\_group |
|  | select \* from vets\_by\_race; |
|  |  |
|  | select \* from consumer\_complaints c |
|  | inner join states s on c.state = s.abbreviation |
|  | left join vets\_by\_age\_group a on s.state = a."State" |
|  | left join vets\_by\_race r on s.state = r.state |



