Group Members
Mason Water
Edward Gates Jr.
Nikita Jones

ETL Group Project

What this project is about and what it aims to accomplish?
We will web-scrape apartment sites and census data in RVA to provide a resource for clients that compare the cost of living and public education by zip code.

Explain the problem that the project addresses specifically.
Richmond, Virginia, is quickly becoming a city of choice. Many who relocated to RVA are often tasked with searching several sights to gather information that one would need when looking for a residence. This collection of data will reduce the amount of time a client may use when seeking to gather information that would guide an informed decision about what area of RVA best suits their needs.

EXTRACT: your original data sources and how the data was formatted (CSV, JSON, pgAdmin 4, etc).

Data sources with a targeted focus on Richmond, VA

- Apartments.com
  - Web Scraping with Python

- Factfinder.census.gov (for demographic information)
  - Excel downloads to CSV conversions

Information being sought

- From Apartments.com
  - Name
  - Apartment_link
  - Address
  - Zip Code
  - Apt Phone
  - Price Range
  - Apt Rating
  - Local Education
    - School Name
    - Grade Levels
    - Student Count
    - School Phone

- From Factfinder.census.gov
    - Breakdown of zip codes by:
        - Gender
        - Age
        - Race
        - Citizens of Voting Age (18 years and older) Population

- Used bs4 to scrape apartment info from apartments.com and

- Used Pandas to extract tables from csv's on census.gov.

TRANSFORM: what data cleaning or transformation was required?

- Apartments.com
    - Used .replace to reassign labels/values to documents
    - Organized scraped information into a dictionary to be merged with census data

- Factfinder.census.gov
    - Pandas
    - Used Pandas to drop both null values to empty columns, and renamed columns.
    - Turned census csv into DataFrame and
    - Removed redundant fields,
    - Renamed fields,
    - Dropped N/A values and
    - Converted DataFrame into html table to append dictionary

LOAD: the final database, tables/collections, and why this was chosen.

- The final database
    - "Apartments_db" in MongoDB

- Table
    - "Apartments"

- Why?
    - With there not being a common denominator between the apartment information and the demographic information, a relational database would not be the proper tool to use for our loading process. Therefore, MongoDB was our best load option.