DSC 200 Data Wrangling

Fall 2024 Final Project

Title: "Urban Housing Dataset Integration by Region"

Objective

Students will work in pairs to acquire, clean, and integrate datasets related to housing trends in a U.S. region or state of their choice. The focus will be on building a unified, clean, and ready-to-analyze dataset, demonstrating mastery of data wrangling techniques.

Project Context

You are part of a data team tasked with preparing housing-related data for future analysis. Your group will select a U.S. region or state (e.g., California, Texas, New York City) and focus on acquiring, cleaning, and integrating relevant datasets from various sources specific to your chosen area.

Project Tasks

Task 1: Data Acquisition

1. Choose a Region/State

Each group will select a specific region or state in the U.S. as their focus. Examples include California, Texas, New York City, or smaller regions like Seattle or Austin. No two groups should choose the same region or state. For this reason, indicate which state your group has chosen in the linked Canvas Discussion thread. A member of the group should state their group number and the region they have chosen for this project.

2. Acquire Data from Multiple Sources

- Excel Files: Import neighborhood demographic data specific to your chosen region.
  Download Demographics Data (Excel): https://www.census.gov/data.html
- CSV Files: Import median housing prices and rental costs for your chosen region.
  Download Housing Prices Data (CSV): https://www.zillow.com/research/data/
  Download Rental Costs Data (CSV): https://www.kaggle.com/datasets
- Web Scraping: Extract rental listings and descriptions from a real estate website such as Craigslist or Apartments.com for your selected area.
- API Access: Retrieve recent housing market trends specific to your region using Zillow or a similar platform API.
  Zillow API Documentation: https://www.zillow.com/howto/api/APIOverview.htm
- PDF Files: Extract relevant data from government housing policy reports specific to your chosen area. Download Housing Policy Report Example (PDF): https://www.huduser.gov/portal/home.html

Task 2: Data Cleaning and Integration

1. Data Cleaning

- Standardize column names and formats across datasets.
- Handle missing data (e.g., imputation, removal, or marking).
- Convert numerical data to consistent types.
- Resolve inconsistencies in date and time formats.
- Remove duplicate records from scraped data.

2. Data Integration

- Merge Datasets: Combine the provided data on common keys such as zip code or neighborhood.
- Add Calculated Fields: Include calculated columns like the price-to-income ratio or average rental cost per square foot.
- Validate Results: Ensure that the merged dataset contains no missing or inconsistent values.

Deliverables

1. Clean Dataset: Submit a single CSV file with the merged, cleaned, and prepared data for your chosen region.
2. Python Code: Submit Python scripts or Jupyter Notebooks with documented steps for:
   - Data acquisition (Excel, CSV, PDF, scraping, and API).
   - Data cleaning (missing values, standardization, and formatting).
   - Data integration (merging and validation).
3. Brief Summary: Provide a short summary (200–300 words) describing the region selected, the data used, and the steps taken to clean and integrate the datasets.

Skills Applied

- Python Basics: File handling, data structures, and scripting.
- Working with Files: Reading and writing Excel, CSV, and PDF files.
- Web Scraping: Acquiring data from websites using Python libraries.
- API Access: Retrieving and parsing data from web APIs.
- Data Cleaning: Handling missing values, standardizing formats, and resolving inconsistencies and outliers.
- Data Integration: Merging and aligning datasets into a single, cohesive structure.

Starter Kit

- Download Demographics Data (Excel): https://www.census.gov/data.html
- Download Housing Prices Data (CSV): https://www.zillow.com/research/data/
- Download Rental Costs Data (CSV): https://www.kaggle.com/datasets

- Download Housing Policy Report Example (PDF): https://www.huduser.gov/portal/home.html
- Zillow API Documentation: https://www.zillow.com/howto/api/APIOverview.htm

Notes for Students

- Be mindful of website terms of service when scraping data.
- Document any challenges you face and the steps you took to resolve them.
- Ensure your final dataset is clean, consistent, and well-structured for future analysis.
- Focus on regional specificity when acquiring and preparing data.

Rubric

1. Technical Proficiency (40%)

Criteria:
- Correct use of Python libraries for data acquisition, cleaning, and integration.
- Clean, efficient, and well-documented code.
- Appropriate handling of missing data, inconsistencies, and duplicates.

2. Data Cleaning and Integration (30%)

Criteria:
- Datasets are successfully cleaned and standardized.
- Integration across datasets is seamless and logically executed.
- Validation checks ensure data integrity and consistency.

3. Documentation and Process (20%)

Criteria:
- Clear, concise, and well-structured documentation of steps taken.
- Detailed summary explaining data cleaning and integration process.
- Proper explanation of challenges encountered and solutions applied.

4. Final Dataset Submission (10%)

Criteria:
- Final dataset is clean, complete, and well-structured.
- Dataset is ready for analysis and includes calculated fields where appropriate.
- Dataset submission is in the correct format and adheres to the instructions.