Antoine Gaton

Professor Rida Moustafa

CS356: Foundations of Big Data Analytics

14 July 2024

***Unit 5 - Individual Project***

* **Description:** The demographic data pipeline is a comprehensive program designed to process demographic data by zip code. The data is sourced from a CSV file and undergoes several processing steps to ensure data quality and generate meaningful insights.
* **Steps:**
  + *Data Ingestion*
    1. The program reads the demographic data from a provided CSV file using the pandas library.
    2. The data is loaded into a Pandas DataFrame for further processing.
  + *Data Storage*
    1. The ingested data is saved locally in a CSV file to ensure data persistence and for future access.
  + *Data Cleansing and Preprocessing*
    1. Missing values are handled using the forward-fill method to ensure data completeness.
    2. 'TotalPopulation' values of zero are treated as missing and removed to ensure data accuracy.
    3. Infinite values are replaced with NaN to maintain data consistency.
  + *Kidney Disease Analysis*
    1. The program analyzes the average kidney disease rate by state using the 'KIDNEY\_CrudePrev' column.
    2. Visualizations are created to depict the distribution of kidney disease rates by state.

| import pandas as pd *# Importing pandas for data manipulation* import numpy as np *# Importing numpy for numerical operations* import matplotlib.pyplot as plt *# Importing matplotlib for plotting* import seaborn as sns *# Importing seaborn for statistical data visualization* import warnings *# Importing warnings to handle warning messages* import os *# Importing os for file path operations* import webbrowser *# Importing webbrowser to open files in the default browser*  *# Suppress the specific FutureWarning from seaborn* warnings.filterwarnings("ignore", category=FutureWarning, module="seaborn") *# Step 1: Ingest the Data* data\_url = 'https://data.cdc.gov/api/views/mssc-ksj7/rows.csv?accessType=DOWNLOAD' *# URL for the data* df = pd.read\_csv(data\_url) *# Reading the data into a pandas DataFrame*  *# Print the columns of the DataFrame* print("Columns in the dataset:", df.columns) print("----------------------------------------------------------------------------")  *# Print columns related to kidney disease* kidney\_disease\_columns = [col for col in df.columns if 'KIDNEY' in col.upper()] print("Columns related to kidney disease:", kidney\_disease\_columns) print("----------------------------------------------------------------------------")  *# Also print the first few rows to understand the data* print(df.head()) print("----------------------------------------------------------------------------")  *# Step 2: Store the Data* df.to\_csv('demographics\_data.csv', index=False) *# Saving the DataFrame to a CSV file*  *# Step 3: Cleansing and Preprocessing* df.ffill(inplace=True) *# Handle missing values using forward fill method*  *# Ensure 'TotalPopulation' values of zero are treated as missing and remove such records* df['TotalPopulation'] = df['TotalPopulation'].apply(lambda x: np.nan if x == 0 else x) *# Replace 0 values with NaN* df.dropna(subset=['TotalPopulation'], inplace=True) *# Drop rows where 'TotalPopulation' is NaN*  *# Convert any infinite values to NaN* df.replace([np.inf, -np.inf], np.nan, inplace=True) *# Replace infinite values with NaN*  *# Step 4: Kidney Disease Analysis* *# Use 'StateDesc' for state names and 'KIDNEY\_CrudePrev' for kidney disease rates* if 'KIDNEY\_CrudePrev' in df.columns and 'StateDesc' in df.columns: *# Check if necessary columns exist*  state\_kidney\_disease = df.groupby('StateDesc')['KIDNEY\_CrudePrev'].mean().reset\_index() *# Group by state and calculate mean*  *# Sort by kidney disease rate*  state\_kidney\_disease.sort\_values(by='KIDNEY\_CrudePrev', ascending=False, inplace=True)   *# Visualize kidney disease rates by state*  plt.figure(figsize=(12, 8)) *# Set the figure size*  sns.barplot(x='KIDNEY\_CrudePrev', y='StateDesc', data=state\_kidney\_disease, palette='viridis') *# Create a bar plot*  plt.title('Average Kidney Disease Rate by State') *# Set the title*  plt.xlabel('Kidney Disease Rate (Crude)') *# Set the x-axis label*  plt.ylabel('State') *# Set the y-axis label*  kidney\_disease\_file = 'kidney\_disease\_by\_state.png' *# Filename for the plot*  plt.savefig(kidney\_disease\_file) *# Save the plot to a file*  plt.show() *# Display the plot*   *# Open the kidney disease rate plot*  webbrowser.open('file://' + os.path.realpath(kidney\_disease\_file)) *# Open the saved plot* |
| --- |

* **Future Applications**
  + The demographic data pipeline has several potential applications based on the analysis of kidney disease rates and population distribution:
    - *Public Health Planning*
      1. Resource Allocation: By identifying states with higher rates of kidney disease, public health officials can allocate healthcare resources more effectively to areas with the greatest need.
      2. Preventative Programs: The analysis can help in planning and implementing preventative health programs targeting states with high kidney disease rates.
    - *Healthcare Services*
      1. Facility Planning: Hospitals and clinics can use this data to determine where to establish new facilities or expand existing ones to better serve populations with higher disease prevalence.
      2. Patient Outreach: Healthcare providers can develop targeted outreach programs to educate communities in states with high kidney disease rates about risk factors and management strategies.
    - *Policy Making*
      1. Legislation: Lawmakers can use the insights from this data to create policies that address the healthcare needs of states with higher kidney disease rates.
      2. Funding: Government agencies can allocate funding more effectively to support healthcare initiatives in states with greater health challenges.
    - **Insurance Industry**
      1. Risk Assessment: Insurance companies can use kidney disease prevalence data to assess risk and adjust premiums accordingly for different states.
      2. Healthcare Plans: Insurers can design specialized healthcare plans that offer better coverage for conditions prevalent in specific regions.
    - **Community Health Initiatives**
      1. Local Programs: Community organizations can develop local health initiatives based on the specific needs highlighted by the data, such as kidney disease prevention and treatment programs.
      2. Awareness Campaigns: Grassroots campaigns can be organized to raise awareness about kidney disease in states with high prevalence rates.
    - **Research and Development**
      1. Medical Research: Researchers can use the data to identify trends and correlations in kidney disease prevalence, leading to potential breakthroughs in treatment and prevention.
      2. Clinical Trials: Pharmaceutical companies can target states with higher disease rates for clinical trials to ensure diverse and relevant participant pools.
    - **Population Health Management**
      1. Health Monitoring: Public health authorities can continuously monitor population health trends using the data, allowing for timely interventions and updates to public health strategies.
      2. Data-Driven Decisions: Health departments can make data-driven decisions to improve overall community health and reduce the burden of kidney disease.
* **Conclusion**
  + The demographic data pipeline effectively processes and analyzes demographic data by zip code, focusing on kidney disease rates and population distribution by state.
  + By ensuring data integrity through comprehensive cleansing and preprocessing, the pipeline provides reliable insights that can be utilized in various sectors. Public health officials and healthcare providers can leverage this data to allocate resources and design targeted interventions, while policymakers and insurance companies can make informed decisions to address regional health challenges.
  + Community organizations and researchers can also benefit from this data to develop localized health programs and advance the understanding of kidney disease.
  + Overall, this pipeline is a valuable tool for enhancing public health initiatives, optimizing healthcare services, and informing policy decisions, ultimately contributing to improved community health and well-being.