Project Proposal: Analyzing Obesity Trends and Geographic Correlations Using Automated Data Pipeline

Name: AKM Yasar

23025328

Question:
How do obesity rates correlate with socioeconomic factors such as income and physical activity across different states in the United States?

Data Sources:

1. **Nutrition, Physical Activity, and Obesity - Behavioral Risk Factor Surveillance System (CSV)**
    - **Source:** Centers for Disease Control and Prevention (CDC)
    - **Data URL:** https://data.cdc.gov/api/views/hn4x-zwk7/rows.csv?accessType=DOWNLOAD
    - **Content:** State-level data on health metrics, including obesity rates, physical activity, and nutrition habits.
    - **License:** Open Government License (OGD). License details available at https://www.usa.gov/open.
2. **Obesity Rates and Geographic Information by State (GeoJSON)**
    - **Source:** Lake County, IL Open Data Portal
    - **Data URL:** https://services3.arcgis.com/HESxeTbDliKKvec2/arcgis/rest/services/LakeCounty_Health/FeatureServer/8/query?outFields=*&where=1%3D1&f=geojson
    - **Content:** Geographic data on obesity rates with demographic and socioeconomic variables.
    - **License:** Open Data Commons Attribution License. License details available at https://opendatacommons.org/licenses/by/1.0/.

Data Quality and Structure:

- Both datasets are state-level, ensuring compatibility for merging.

- The CSV is tabular with columns for metrics like obesity rate and physical activity.
- The GeoJSON contains spatial data with demographic attributes.
- Data quality was assessed for completeness and consistency. Minor inconsistencies in state naming conventions were resolved.

Data Pipeline:

- **Technology:** Python with pandas, sqlite3, and geopandas for data handling, cleaning, and transformation.
- **Pipeline Steps:**
    1. **Data Fetching:** Automated retrieval from API endpoints.
    2. **Data Cleaning:** Harmonized state names, removed duplicates, and normalized column names.
    3. **Data Transformation:** Converted GeoJSON to tabular format for integration.
    4. **Data Integration:** Merged datasets on state names to create a comprehensive table.
    5. **Output:** Generated a cleaned and merged CSV file for visualization.

Challenges and Solutions:

- **Issue:** Inconsistent state naming conventions between datasets.
  Solution: Standardized state names using a transformation function.
- **Issue:** Handling GeoJSON spatial data.
  Solution: Used geopandas to extract attributes and convert them into tabular format.

Meta-Quality Measures:

- **Error Handling:** Automated retries for failed API calls.
- **Data Updates:** Pipeline designed to handle new data releases without manual intervention.

Results and Limitations:

- **Output Data:**
    - **Format:** CSV file with integrated data, ensuring ease of use for visualization.
    - **Structure:** Tabular with columns for state, obesity rate, income, education, physical activity, etc.
- **Limitations:**
    - Potential bias in self-reported data from behavioral surveys.
    - Geospatial analysis limited to state-level granularity.

Reflection:
While the pipeline successfully integrates and cleans the data, future improvements could include incorporating county-level data for more detailed geographic analysis and additional factors such as healthcare access.