Do socioeconomic factors correlate with wind energy adoption in the United States?

Data Report

Handed in by:	Christoph Sommermann
Matriculation number:	22687554
Course:	Methods of Advanced Data Engineering

Research Questions:

1. Which state has the highest number of wind turbines in the United States?

2. Which states have the most favorable socioeconomic conditions (e.g., lowest poverty rates, highest education levels)?

3. **Main question:** Do favorable socioeconomic conditions correlate with higher wind turbine adoption in the United States?

1. Data Sources

1.1 United States Wind Turbine Database

1.1.1 Dataset description and Motivation

The U.S. Wind Turbine Database provides comprehensive and up-to-date information on wind energy infrastructure in the United States, covering turbine installations from 1982 to 2024. It allows for state-level analysis of wind turbine adoption, including installation dates, capacities, and technical specifications. This dataset is crucial for identifying which states have the highest number of wind turbines.

The Wind Turbine Database comprises 29 columns and over 70,000 rows, including key descriptor columns such as *case_id*, *t_state*, and *t_county*, which identify and locate each turbine. Additional columns provide technical specifications and operational details, including *p_year* (year of installation), *t_cap* (capacity in kilowatts), *t_hh* (hub height in meters), *t_rd* (rotor diameter in meters), and geographic coordinates (xlong, ylat). The dataset also includes information on turbine models, manufacturers, and retrofit details, offering a comprehensive overview of wind energy infrastructure in the United States.
Link to the data

1.1.2 Data Origin and Quality

The Database for U.S. Wind Turbine, available from the U.S. Geological Survey via the Data.gov platform, combines records from multiple sources, including the Federal Aviation Administration (FAA), the American Clean Power (ACP) Association, Lawrence Berkeley National Laboratory (LBNL), and the United States Geological Survey (USGS).

Accuracy. The dataset's accuracy relies on the provider, with a reported locational error tolerance of ±10 meters and specifications based on verified sources, though some uncertainties remain.
Completeness. The dataset is largely complete, with most columns having less than 6% missing data. However, specific fields, such as *usgs_pr_id* (50% missing), which represents project IDs, and *t_retro_yr* (89% missing), indicating turbine retrofit years, exhibit significant gaps.
Consistency. The dataset is consistent, with standardized formats and defined units, ensuring reliable integration and analysis.

1.1.3 Data License

The Data.gov catalog states that the U.S. Wind Turbine Database is intended for public access and use. While no specific license is provided, the dataset is considered a U.S. Government Work, placing it in the public domain. To fulfill obligations, the data source will be properly acknowledged in all reports and outputs. The statement that public use of the dataset is permitted can be found under the link in 1.1.1.

1.2 ACS 5YR Socioeconomic Estimate Data by State

1.2.1 Dataset description and Motivation

The ACS 5-Year Socioeconomic Estimate Data provides detailed and aggregated socioeconomic information at the state level in the United States. It includes key indicators such as poverty rates, median household income, workforce participation, and educational attainment, offering valuable insights into regional socioeconomic conditions. This dataset is critical for analyzing the socioeconomic factors on state level in the US.

This Dataset contains 52 rows and 144 columns, representing socioeconomic metrics for each U.S. state and territory. Key attributes include *GEOID* (state identifiers), *STUSAB* (state abbreviations), *NAME* (state names), and various socioeconomic indicators such as *B08013EST1* (aggregate income) and *B24021EST28* (employment statistics). This data set can therefore be used to investigate possible correlations between socio-economic factors and the adaptation of wind power.
Link to the data

1.2.2 Data Origin and Quality

The ACS 5-Year Socioeconomic Estimate Data is provided by the U.S. Department of Housing and Urban Development (HUD) through the ArcGIS Open Data platform. It aggregates detailed socioeconomic metrics for U.S. states and territories, offering a robust basis for analyzing regional trends.

Accuracy. The accuracy of the dataset cannot be explicitly verified, as it is derived from aggregated 2016-2020 ACS 5-Year estimates compiled by the U.S. Census Bureau.
Completeness. The dataset is complete, with all columns containing data for the 52 rows representing U.S. states and territories. No missing values were identified.
Consistency. The dataset is consistent, featuring standardized column names and data formats. Key attributes, such as state identifiers and socioeconomic metrics follow uniform conventions, facilitating smooth analysis and integration.

1.2.3 Data License

The ArcGIS Open Data platform states that the ACS 5-Year Socioeconomic Estimate Data is intended for public access and use. While no specific license is provided, the dataset is published by the U.S. Department of Housing and Urban Development (HUD) and is considered a U.S. Government Work, placing it in the public domain. To fulfill obligations, the data source will be properly acknowledged in all reports and outputs. The statement confirming public use can be found under the link in 1.2.1.

2. Data Pipeline

Data Pipeline Overview. The data pipeline was implemented in Python to automate the collection, transformation, and storage of datasets. Key python libraries include *requests* for downloading windmill data via a direct URL and fetching socioeconomic data through an API, *pandas* for cleaning and transforming data, and *SQLAlchemy* for saving the processed data into an SQLite database. The pipeline follows a structured flow: fetching data, storing raw files,

preprocessing, saving cleaned data, and preparing it for analysis. The steps are visualized in Figure 1. A key aspect of the pipeline in *pipeline.py* is modularity, achieved by separating tasks into *preprocessor.py* for data cleaning and *downloader.py* for data retrieval. This design simplifies maintenance, enables reusability of components, and allows for easier debugging and updates.

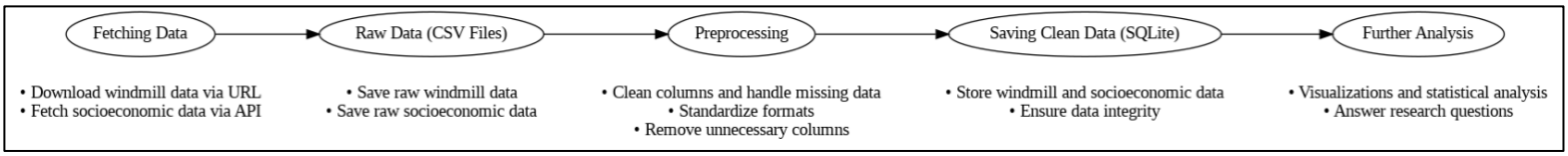

Figure 1: Pipeline Overview

Transformation and Cleaning. The pipeline ensures data usability by standardizing column names, removing irrelevant metadata (e.g., *usgs_pr_id*, *t_retrofit*), and dropping rows with critical missing values (e.g., *p_year*, *t_cap*) in the windmill dataset. For the socioeconomic dataset, geospatial columns (e.g., *Shape__Area*, *Shape__Length*) were removed, numeric formats standardized, and only relevant indicators retained. These steps streamlined both datasets for compatibility and analysis.

Challenges and Solutions. The pipeline encountered three primary challenges. First, API timeouts and incomplete downloads were addressed by adding retry logic to requests. Second, large windmill CSV files were processed using chunking in pandas to manage memory constraints. Third, merging datasets required aligning state-level keys, resolved by standardizing identifiers such as GEOID across datasets. These solutions ensured a smooth pipeline execution.

Meta-Quality Measures. Meta-quality measures were implemented primarily in the *DataRetriever* class within the *downloader.py* file, where *requests.get()* uses *response.raise_for_status()* to detect and stop processing on API failures (e.g., 404 or 500 errors), the *download_csv* method checks *os.path.exists(file_path)* to prevent overwriting files or redundant downloads, and the timeout parameter ensures the pipeline avoids indefinite waits during data fetching.

3. Results and Limitations

The output of the data pipeline consists of two cleaned and processed datasets: windmill data and socioeconomic data, stored in an SQLite database. The data structure ensures compatibility, with tables containing standardized column formats, consistent naming, and filtered attributes relevant to the analysis. This structure enhances query efficiency and supports downstream analysis. Additionally, the SQLite Viewer extension in Visual Studio Code was used to directly analyze the SQL file, allowing for seamless exploration and validation of the processed data.
The SQLite format was chosen for its simplicity, lightweight design, and integration capabilities with Python, allowing flexible data manipulation and analysis. The processed datasets are complete for most fields, with missing values addressed during preprocessing, ensuring high overall quality.
However, limitations remain. While the windmill data and socioeconomic data are both up to date (ending in 2024 and 2020, respectively), the temporal mismatch may still affect the alignment of trends. Additionally, uncertainties in some turbine specifications and the state-level aggregation of socioeconomic data may introduce biases or limit granularity in insights. These challenges will be addressed through careful interpretation in the final report.