

EDA_Final_Project_2020

This repository is for the Environmental Data Analytics final project in data visualization and analysis at Duke University in Spring 2020.

Summary

This project is to analyze the migration patterns of residents of coal counties in central Appalachia from the years 2000 to 2017. Two “coal counties” serve as a case study for migration in central Appalachia: Harlan County, Kentucky and Dickenson County, Virginia. Harlan County serves as a “Boom Bust” county, meaning the county experienced a coal production boom in 2000 and a coal production bust in 2010. Dickenson, on the other hand, is a “Bust Bust” county, where the county experienced a coal production bust in 2000 and did not recover by 2010.

An important portion of understanding migration in Appalachia is investigating the mine closures within the coal counties. As such, this project creates and analyzes a visualization of mine closures in Harlan and Dickenson counties between the years 2000 and 2011. The data analysis of the number of mine closures as well as the location of the closures could be indicative of migration within the central Appalachian region.

Investigators

Hannah Smith

Duke University

Nicholas School of the Environment

hannah.g.smith@duke.edu

Master’s of Environmental Management Student

Keywords

coal mining, Appalachia, mine closures, coal mining trends

Database Information

The data in this repository is coal mine data compiled by the Coal and America Bass Connections team at Duke University from the Energy Information Administration (EIA) online database throughout the fall of 2019.

Folder structure, file formats, and naming conventions

Folder Structure

- **Data/Raw:** The raw data folder holds the files with the data drawn directly from the EIA database.
- **Data/Processed:** The processed data folder houses the files wrangled and processed in R.
- **Code:** The code folder contains the code ran in R for the project.
- **Output:** The output folder contains the tables and graphs produced from the processed data.

File Formats

- **.csv:** spreadsheets of data
- **.pdf:** output files

Naming Conventions

Files are named according to the following naming convention: `databasename_datatype_details_stage.format`, where:

databasename refers to the database from where the data originated

datatype is a description of data

details are additional descriptive details, particularly important for processed data

stage refers to the stage in data management pipelines (e.g., raw, cleaned, or processed)

format is a non-proprietary file format (e.g., .csv, .txt)

Metadata

EIA_MineData_Harlan_Raw.csv EIA_MineData_Dickenson_Raw.csv

- **year** (char): the year the mine data represents
- **mine.name** (char): the name of the mine
- **mine.state** (char): the state in which the mine operated
- **countystr** (char): the county in which the mine operated
- **mine.basin** (char): the geological region in which the mine operated
- **mine-status** (char): details where the mine was active, active with men employed but no production, temporarily closed, or permanently closed
- **mine.type** (char): whether the mine was an underground or surface (strip) mine
- **company.type** (char): the filing status of the company that owns the mine
- **operation.type** (char): whether the mine is a mine and/or preparation plant
- **operating.company** (char): company that owns and operates the mine
- **operating.company.address** (char): the address of the company that owns and operated the mine (may not be in the same location as the mine itself)
- **union.code** (numeric): the code representing the union of which the miners are members, may or may not be applicable
- **production.stons** (numeric): annual coal production in tons of the mine
- **average.employees** (numeric): annual average of employees (mine-wide) employed by the mine
- **labor.hours** (numeric): annual number of man hours performed
- **ARC** (numeric): whether the mine.state is in the Appalachian Regional Commission or not (1 = yes, 0 = no -> will change to logical)

Scripts and code

<list any software scripts/code contained in the repository and a description of their purpose.>

Quality assurance/quality control

<describe any relevant QA/QC procedures taken with your data. Some ideas can be found here:> <https://www.dataone.org/best-practices/develop-quality-assurance-and-quality-control-plan> <https://www.dataone.org/best-practices/ensure-basic-quality-control> <https://www.dataone.org/best-practices/communicate-data-quality> <https://www.dataone.org/best-practices/identify-outliers> <https://www.dataone.org/best-practices/identify-values-are-estimated>