A collaborative ReadMe file that explains how to direct oneself through Dark Orchid's GitHub repository and project.

CODE

- Final: The "Final" folder includes Cleaning_Data.Rmd, Final_mark_down.rmd, Final_mark_down.pdf, and dark_orchid_final_eda.rmd.

  Cleaning_Data.Rmd - The process of converting our raw files into the files that were used in our final mark down. The final markdown file is designed only to work with cleaned files that are written by the Cleaning_Data.rmd and saved into DATA/clean.

  Final_Mark_Down.rmd - The process of the entire project is summarized in this .rmd. The first section shows the feature engineering process for calculating distance between county centroids and coal plant operations. The second section of the .rmd compiles the health outcomes, air quality data, and coal data into a single tibble. The tibble is organized by county and contains 3000+ county observations for health outcomes and coal plant location data. The third section of the .rmd is a brief set of exploratory regressions on the tibble created in section two. The fourth section of the document has two important parts. First, smoking data by county is introduced as a control for our advanced regressions. Second, a set of more advanced regressions are analyzed to find relationships in the top 10% and 50% counties closest to power plants, and to investigate relationships that take smoking data into account.

  Final_Mark_Down.pdf - A knitted PDF version of the Final_Mark_Down.rmd.

  Dark_orchid_final_eda.rmd - An exploratory data analysis on our fully joined set of data that advanced regressions are run on. This EDA explores a cleaned file called health_aqi_coal_distance_smoking_CLEAN.csv that was written in the Final_Mark_Down.rmd and saved into DATA/clean. The EDA gives a more advanced breakdown of the variables used in the project and is helpful to understanding the regressions and csv's created in Final_Mark_Down.rmd.

- Individual Work: Various supporting files compiled by respective group members that clean data, create csv's, experiment with feature engineering, and provide brief EDAs.
  - Bankson
  - Hathaway
  - Leinweber
  - Norwood

DATA

- Clean: Files that were cleaned and written by either the Cleaning_Data.Rmd or Final_Mark_Down.Rmd.

  aqi_2014_CLEAN.csv - Air quality data cleaned and written by Cleaning_Data.Rmd

centroids_plants_CLEAN.csv - County centroid data combined with minimum distance from centroid to nearest coal plant. Written by Final_Mark_Down.Rmd in the feature engineering section.
coal_plants_CLEAN.csv - Coal plant data that is cleaned in Cleaning_Data.Rmd.
countycrosswalk_CLEAN.csv - Crosswalk used to join various data sets by county and fips codes. Downloaded as a clean data set.
health_aqi_coal_distance_CLEAN.csv - Compilation of health, aqi, coal plant, and minimum distance data written by Final_Mark_Down.Rmd in section two.
health_aqi_coal_distance_smoking_CLEAN.csv - Compilation of health, aqi, coal plant, minimum distance, and smoking data written by Final_Mark_Down.Rmd in section two.
health_outcomes_CLEAN.csv - Health outcomes data file that is cleaned in Cleaning_Data.Rmd
smoking_CLEAN.csv - Smoking data file cleaned in Cleaning_Data.Rmd.
uscounties_CLEAN.csv - Longitude and latitude of all US County centroids that was downloaded as a clean data file and is used in feature engineering and Final_Mark_Down.Rmd.

- Raw_Unprocessed: Folder contains .csv and .xlsx file conventions- no RMDs in the folder.
    - The file CoalPlants.csv was found from https://globalenergymonitor.org/projects/global-coal-plant-tracker/download-data. The data helped create detailed independent variables. Location of plants, longitudinal and latitudinal data, and operating status were crucial inputs from the data set.
    - The file annual_aqi_by_county_2014.csv helped provide air quality values measured by the EPA. This data set contributed to our collection of independent variables and was found from https://aqs.epa.gov/aqsweb/documents/data_api.html.
    - The folder Health Outcomes helped provide a variety of US health outcomes and was found from https://ghdx.healthdata.org/record/ihme-data/united-states-mortality-rates-county-1980-2014. These data points were helpful as dependent variables.
    - The file smoking.csv provided more independent variables. The crucial variable provided by this data set is the crude percent of adults who currently smoke in 2020. The data was found through the CDC at https://unc-policymap-com.libproxy.lib.unc.edu/newmaps#/.
    - The file uscounties.csv is a file downloaded from https://simplemaps.com/data/us-counties. It contains longitudinal and latitudinal info about all US county centroids (informed by US Census Bureau, ACS).

OUTPUTS

- ExecutiveSummary_Infographic
- Knitted_Files

OTHER

- Presentation
- Tableau - See **Explanation of Tableaus.**