Energy Production, Sector and Price: A Historical Account

Group members- Tiffany, Pablo, Paul S., Peter

Problem Statement- Over the past 50 years, how have different states changed in their volume of energy production, allocation of production by sectors, export and import of energy (to other states), and energy cost? We aim to identify relationships and patterns between these variables to understand both historical and current trends across the country.

What open data sets might you be working with? We will be using the Energy Information Agency's US Energy API bulk download, primarily looking at two high level datasets:

- SEDS (State Energy Data System), which includes production by sector (including renewables), prices, and demographic information for each state (such as GDP and population).
- Electricity > Net Generation by Fuel Type. Not only are we focused on electrical generation (as opposed to metrics purely on petroleum, natural gas or coal, which are also high level categories), but the EIA API breaks down electrical production by sectors that include granularity of different renewables in a similar way to SEDS, allowing for comparisons between the two datasets.

Three other possible high level datasets that we might look into are Petroleum, Natural Gas and Coal. These three are also available from the EIA API download. We would need these three in case we require more details on price and consumption.

Possible ties to Wikipedia (or any Wikimedia related projects)? None as of this point.

What challenges do you anticipate running into?

Refining our scope will be challenging. As we explore the data we are coming up with more and more questions that we are interested in and could answer. Our success in balancing out exploratory data analysis that guides what questions we seek to answer with a specific, refined goal will go a long way in determining how effective our project is.

A further, related challenge is selecting the data and correctly interpreting the variables that best answer our questions. The EIA bulk download data sets contain a deep hierarchy of diverse types of information that we will have to go through and understand before doing our analysis. This will require some of the advanced pandas data manipulation skills that we have been

practicing in the class. Further, the relationship between high-level variables in the SEDS dataset and other high level categories (such as Electricity) is unclear in some places, particularly as we are not energy experts and many of the terms used are ambiguous or have little supporting documentation that we have discovered at this point. We will have to work outside of the API through documentation in other parts of the EIA's site to understand some of these variables.

A final challenge will be identifying the best ways to visualize the variables that inform on our project question.

What skills will a team need to develop to solve this problem?

- Data import/export (accessing JSON files and bringing them into the appropriate data structures in Python)
- Pandas data manipulation skills, to clean, organize, and analyze data
- Javascript, in order to pull the data into an accessible format in a web browser
- Data Visualization, including basic design principles and technical skills. Two team members are currently in Marti Hearst's Information Visualization class, and we will identify and apply whatever visualization technique we determine to be most accessible and useful for our final output, possibly using Javascript tools, such as D3 or Leaflet, Python tools, such as matplotlib, or a combination like mpld3.
- Basic statistics to identify relationships between variables and quantify changes over time. If necessary, we would explore the statsmodel Python tool to generate regressions and statistical analysis.