Data Science Project Proposal: COMP 5360
Predicting Tomorrow's Power Generation: Utah vs. Colorado

John W. Muhs
john.muhs@gmail.com
U0761102

Corbett Carrell
u0502104@utah.edu
u0502104

Link to GitHub Repository: https://github.com/JohnWMuhs/2019-datascience-project

Background and Motivation

Discuss your motivations and reasons for choosing this project, especially any background or research interests that may have influenced your decision.

John: Understanding future trends in energy consumption and generation are important for several reasons. By better understanding how energy consumption will increase in the future, engineers, planners, and policy makers are better able to plan for long-term changes in power and energy infrastructure and the wide-ranging implications of a changing US electric grid. This project closely aligns with my work and interests in the power and energy industries. I am excited to explore macro-scale energy trends, and hope to apply my data science expertise to a useful and relevant issue.

Corbett: Last semester in one of my other classes I attempted to build a neural network to make predictions on how likely a person out on parole is to return to prison. I was unable to build a working network. Since then I have wanted another opportunity to try and build a working neural network. I see this project as an opportunity to further my knowledge of neural networks and gain a better understanding how to build them.

Project Objectives

Provide the primary questions you are trying to answer with your visualization. What would you like to learn and accomplish? List the benefits.

In this project, we want to look into the long-term correlations in climate and population that affect power generation and/or consumption. We plan to view this analysis through the lens of two states: Utah and Colorado. By comparing these two states, we hope to break a large analysis into something that allows users (U of U students/faculty) to gain a conceptual grasp on the energy consumption/generation trends in their communities.

To boil this project down into a three questions, we'd like to answer:
1) How does population growth and weather correlate to energy consumption and generation in Utah and Colorado?
2) What other factors might help to improve the understanding of these relationships?
3) Can we use this information to predict energy consumption/generation trends in the future?

Both Corbett and John have an interest in learning the data science process, and chose this project to cover the entire lifecycle of a realistic data science project. We wanted to work on a project that included data acquisition and scraping, cleaning, exploration and statistical analysis, and machine learning. Therefore a major objective of this project is for both team members to gain experience working with large data in several different ways.

Data

From where and how are you collecting your data? If appropriate, provide a link to your data sources.

We are looking to pull data from three main sources. Our first source is the US Energy Information Administration (EIA). We will pulling information from a monthly power plant generation database. Our second source is the National Climatic Data Center (NCDC). We will pulling weather data from weather stations in Utah and Colorado. Our third source is the US Census (or other data source). We will use the census to pull population and demographic data in Utah and Colorado. To pull data from the EIA, the NCDC, and the US Census we will be using their APIs which can be accessed at the following links:

EIA: https://www.eia.gov/electricity/data/eia923/
NCDC: https://www.ncdc.noaa.gov/cdo-web/datasets
US Census: https://www.census.gov/data/tables/time-series/demo/popest/2010s-state-total.html#par_textimage_1574439295

Ethical Considerations

The use of aggregate energy data in a model such as this could be used to drive policy around power generation or consumption. For example, some policy makers have considered enacting additional taxes associated with carbon emissions of non-renewable power plants. From my personal point of view, the dichotomy between Colorado and Utah will be interesting to see as Colorado is, in general, more progressive when it comes to integrating renewable sources of energy.

As most of this data is publicly available and provided by the US government, we don't anticipate many ways in which this data (or subsequent analysis) could be used in an unethical way. Perhaps there are political or national security concerns behind the use of this data, but we do not foresee this as an issue in our case.

Data Processing

Do you expect to do substantial data cleanup? What quantities do you plan to derive from your data? How will data processing be implemented?

Since we will be combining data from different sources into one dataframe we will have to do some cleaning. The EIA only gives energy generation breakin up by month so that means that all of our other data has to broken up into months as well. There is a possibility that we may not be able to find all our data breakin up by month. For example most population data is broken up according to year (the census is taken every decade), not month. In that case

we will probably have to identify mathematical and statistical workaround methods (e.g. rolling averages) in order to use it with the data from the EIA.

Exploratory Analysis

Which methods and visualizations are you planning to use to look at your dataset?

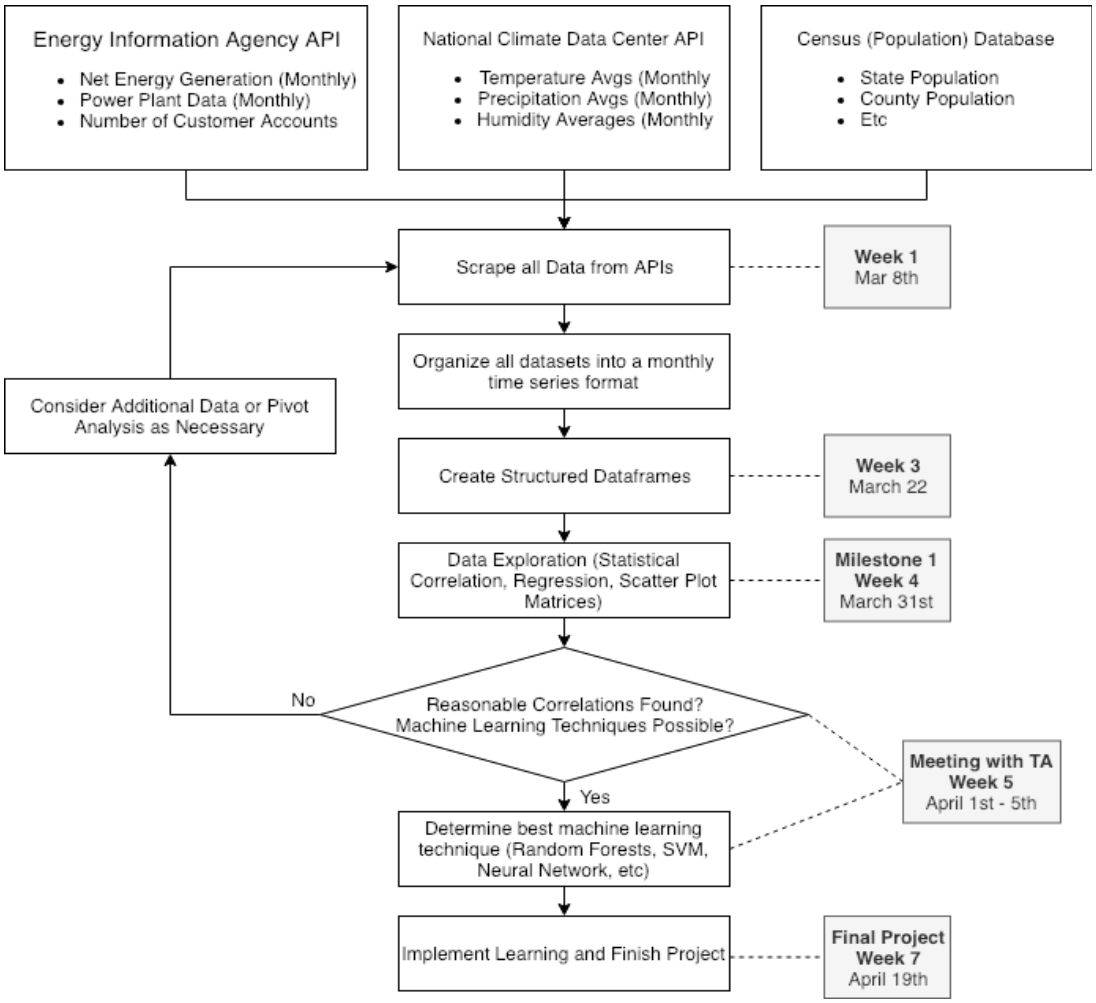
- Correlation Table/ Correlation Scatter Plot Matrix
- Different graphs to compare Utah and Colorado energy generation
- Decisions tree to visualize the different subsets of our data.

Analysis Methodology

How are you planning to analyze your data?

We will use linear regression to begin trying to build a basic predictive model. We will build a random forest to identify the variables that are most important to determining outcome. In the end we hope to build a neural network to make predictions on future energy generation in Utah and Colorado. We then hope to use the neural network to make predictions on the trend of energy generation and make a line graph to visualize this trend.

Project Schedule



Feedback from Peer Review Session -- Dylan Wootan and Teammate:

- Consider the number of y predictor points you have -- if you have monthly data
- A linear regression might be fine for more aggregate results
- A large space range (e.g. an entire state) may obscure the data -- might be helpful to break up state into regions
- Could first build a simple linear regression prediction and build a more complex
- Look at data sets related to industrial growth
- Dark Sky API can be used for weather data
- Pivot could be good even though we don't find anything