Energy Trends In The United States

Nihaal Pabba

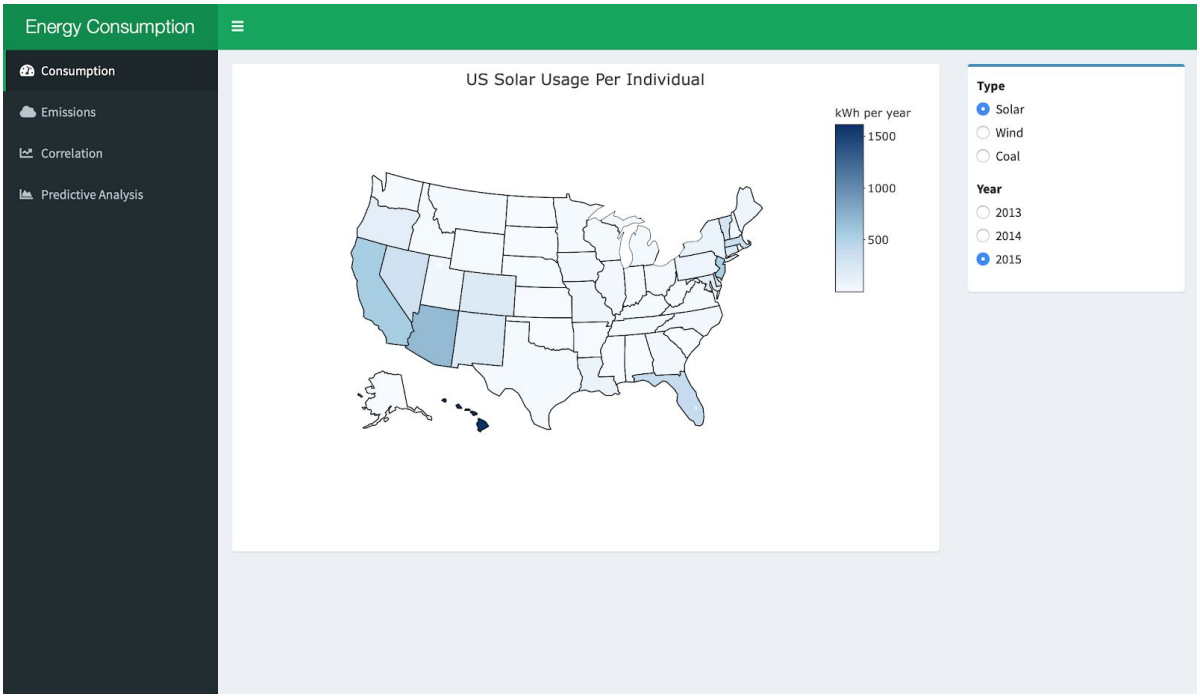
Omkar Kakade

Description

The goal of this project is to explore energy and emissions data from the United States. By identifying factors which can classify and predict this data it can be used by both environmentalists and investors interested in renewable energy.

We wish to build a tool which will help analyze factors related to energy and emissions. This will allow individuals interested in this space to explore trends among the different variables.

To accomplish this we built a dashboard with visually presents data relevant to energy and emissions.



This image for example shows the consumption of solar energy amongst the different states. Hovering over each state will show a box containing more information

as well. From this map a user can quickly see that states such as California and Arizona are consuming much more solar energy than most of the country.

Data

Data was collected from a Quandl API and U.S. Government datasets. By using a Quandl API we were able to extract energy source consumption data and GDPs per state. This data was then cleansed and restructured in order to create tables which could accurately compare the energy usage of a single source between states.

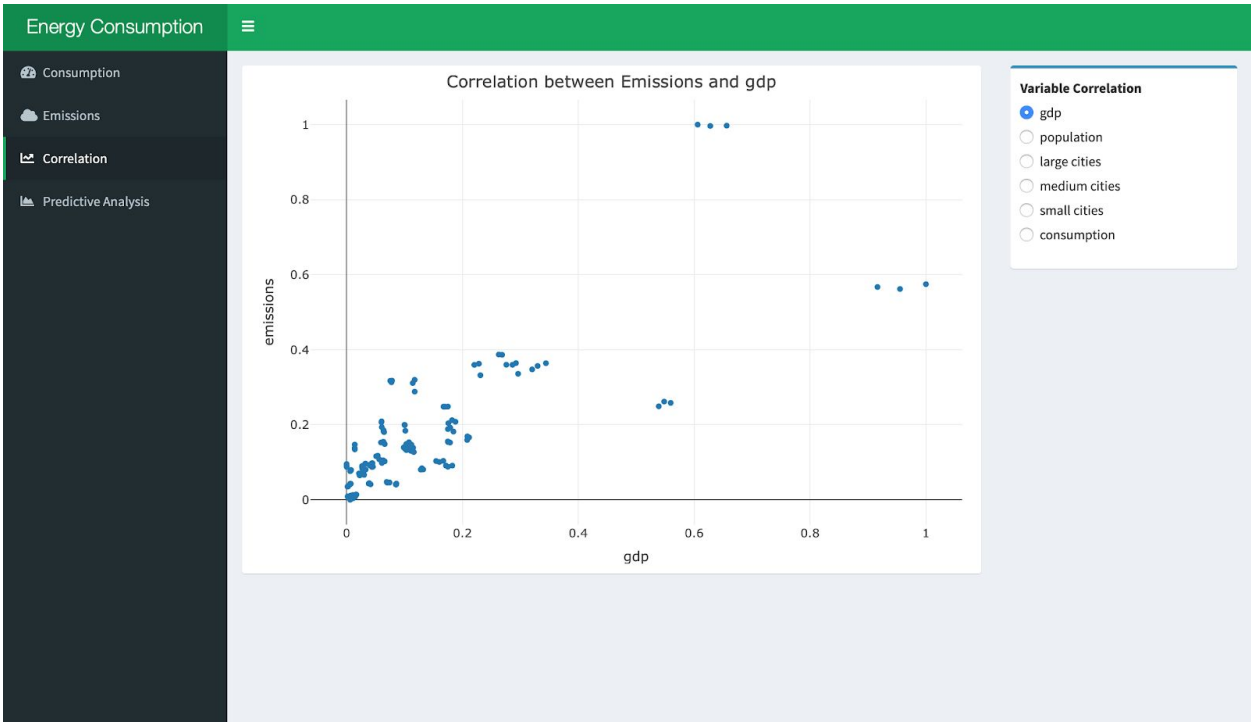
The government hosts datasets pertaining city populations, state populations, and state emissions. We were able to download excel files from the government websites and import it. We used this data to construct other factors such as number of small, medium, and large cities, the percentage of individuals living in cities, etc.

Due to the size of our data we decided it would be best to forgo a database of any type and stick with data frames. A database would simply have added extra overhead and slowed things down. Instead we store within R and cache the data within the environment.

Modeling

Using these factors we also built a model to predict the emissions for a year. We did this using a linear regression model. We added population, gdp, energy consumption, number of large, medium, and small cities, and urban distribution.

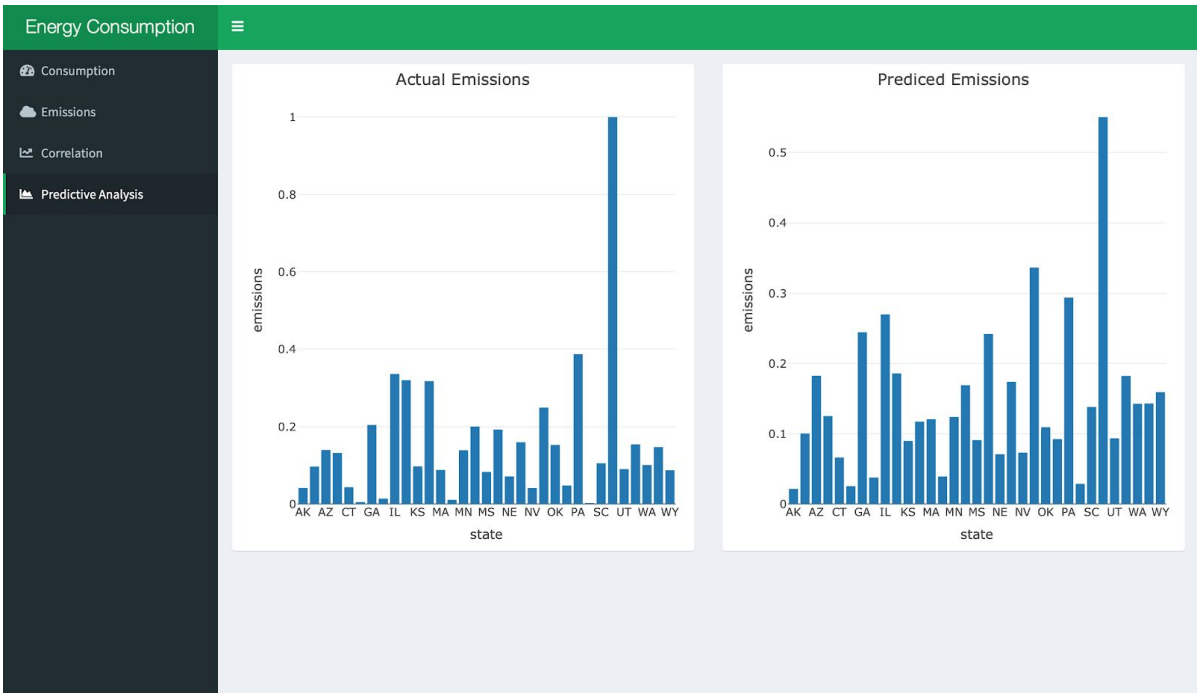
However, after trying different combinations of variables we were unable to identify a model capable of accurately predicting emissions. The correlation between all these variables and the emissions was very low. We have included a tab in the dashboard to visualize the correlations between each variable and the emissions. In future work we would be able to identify better variables and see their correlations here.



The best model we could produce had an accuracy rate about 45% given a threshold of 25%. This is likely an indicator that these factors we collected can't accurately predict emissions. Unfortunately our hypothesis proved false. The best model we could produce is shown below.

```
Call:
glm(formula = emissions ~ population + gdp + consumption, data = training)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.18106  -0.03265  -0.01886   0.01257   0.44770

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.25024    0.03710   6.744 9.16e-10 ***
population   0.28137    0.04917   5.723 1.04e-07 ***
gdp         -0.83699    0.27737  -3.018  0.00321 **
consumption  0.18642    0.06433   2.898  0.00459 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.007593814)

    Null deviance: 3.01453  on 106  degrees of freedom
Residual deviance: 0.78216  on 103  degrees of freedom
AIC: -212.63

Number of Fisher Scoring iterations: 2
```

We have included a tab in the dashboard which graphs our predictions. Hopefully in the future with a better model this would be of use for predicting emissions if given the necessary data.



Issues

There were aspects of this project which were more challenging than others. The first difficulty was the factors we wanted to analyze weren't all in a single data location. We had to utilize multiple sources which contained data of different formats and create our own standardization.

The second issue we had was finding factors which would be a good fit for our linear regression model. We attempted to find good variables with high correlations by plotting them against our emissions with ggplot. However we found that many of our variables didn't highly correlate with emissions.

Moving Forward

We certainly learned a lot from this project. We chose the scope of energy trends since it was something which interested us, however we were unable to consider what sort of analysis or prediction we could do using that data. Partially since we hadn't gone over the types of models in class yet we didn't have a good grasp on what was possible. If we were to do something like this again we would try and have a more concrete plan on what kind of analysis we would be able to do.

To improve this project we would like to explore other factors which may have an impact. Some ideas we had were looking at imports and exports data, presence of manufacturing, and modes of transportation. By exploring other factors we're confident we would be able to find variables which could better predict emissions.

Sources

Energy Data:

https://www.quandl.com/data/EIA-U-S-Energy-Information-Administration-Data

Emissions Data:

https://www.eia.gov/environment/emissions/state/analysis/

State Populations:

https://www.census.gov/newsroom/press-kits/2018/pop-estimates-national-state.html

City Populations:

https://www.census.gov/data/tables/2017/demo/popest/total-cities-and-towns.html#tables