Speaker Notes – DATA606 Presentation

**Slide 1 – DATA 606 Project**

**Slide 2 – Research Question**

This project explores if there is a relationship between GDP and CO2 emissions within US States, and if we can use GDP to predict CO2 emissions.

**Slide 2 – Background**

Before covering the data analysis portion – I wanted to give background on what carbon dioxide is – many of you probably know this – but I wanted to quickly cover it. From here on out I will just refer to CO2 or Carbon Dioxide as carbon.

Carbon is a type of greenhouse gas that is known for trapping heat in the atmosphere, and it is emitted when we burn fossil fuels, or through other natural processes, like volcanic eruptions.

So basically, carbon is warming the planet, and causes global warming (also known as climate change). Our dependence on fossil fuels raises carbon levels in the atmosphere – so I thought it would be interesting to explore states’ behavior in releasing CO2 from energy production.

**Slide 3 – Background Cont.**

The data used in this project comes from the Department of Energy’s annual Electric Power Report and the US Department of Commerce annual summary statistics.

As a heads up I will be referring to the department of energy as DOE and the department of commerce as DOC.

**Slide 4 – Dataset Characteristics**

To develop the final dataset, I joined the DOE Annual Electric Power Report and the DOC annual summary statistics for state economies.

Each row in the final dataset, represents the respective state’s annual energy usage and GDP for the reporting year.

There are 1326 rows in the data set and eight columns. It’s important to note that the final dataset uses the total energy production figures – as the DOE report provided data by energy type. Additionally, due to missing data, DC was excluded from this analysis.

**Slide 5- Data Exploration (Energy Production)**

One of the first things I wanted to do was categorize the energy types by renewable and non-renewable energy. I thought it might be interesting to include it as an explanatory variable in a multiple linear regression.

Unfortunately – this was not possible because the `Other` category (seen in this bar chart) included renewable and nonrenewable types – and the data wasn’t really categorized in a way that made this possible.

**Slide 7 – Data Exploration (Producer Type) Visualization**

Here is a bar chart of the count of energy producer by type. It is basically binned by Commerical, Utility, Industrial, and then Independent Power Producers, separated by if they are “co gen or not” Co-gen stands for the ability to cogenerate heat and power.

On the final slide, there’s language outlining where these producer types are defined in the DOE report.

**Slide 8 – Data Exploration CO2 and GDP**

So now we’ll get into the exploratory analysis for determining if there is a relationship between CO2 and GDP.

First, I used the summary function to review a few statistics like the minimum value, first quartile, median, mean, 3rd quartile and maximum value.

For CO2 and GDP, the median is less than the mean, so this indicates that the data is positively skewed or right-skewed. We can look into the distribution of these variables further by using a histogram.

**Slide 9 – Data Exploration CO2 Distribution**

So here is an interactive histogram I developed with plotly. If you hover you can see which state is the outlier, which in this case it is Texas. This is something we want to be mindful of, since this data is right skewed to begin with.

**Slide 10 – Data Exploration GDP Distribution**

Here is another interactive histogram through plotly! In this case we can cover again to see who are the tails at the end, and it is CA. Texas also has a higher GDP compared to other states.

**Slide 11 – CO2 and GDP Scatterplot**

To determine if there is a relationship between the CO2 emitted from energy production in the respective state, and the state GDP, a scatter plot can be used to visualize the linear relationship.

**Slide 12 – CO2 and GDP Scatterplot (cont)**

Based on the plot produced, there is there is evidence that there is a relationship, but it shouldn’t be surprising that there seems to be a few outliers. You’ll notice that there are multiple values for the states, this is because there is 26 years of data captured.

**Slide 13 – GDP and Year Scatterplot**

I also wanted to review the relationship between GDP and Year, to see if it would be possible to incorporate Year as an explanatory variable in a regression. This indicates that the relationship is not linear and seems exponential – this means that the assumption of independence of independent variables may not be met.

**Slide 14- Year and CO2 Scatterplot**

This graph shows that there isn’t a linear relationship between year and CO2. This also means if we were to add Year to a regression, as an independent variable, the linearity assumption would not be met.

**Slide 15- Dealing with Skewness and Outliers**

So there seems to be a few assumptions not met for running a linear or multiple regression:

1. Independence – our independent variables are not independent, which may mean we will run into issues with collinearity.
2. Linear relationship between independent and dependent variables – Year and CO2 do not seem to have a linear relationship, and GDP and CO2 seem to have outliers.

Due to this, I decided to run a log transformation for the data. So here I created a new dataset that I called trans\_state\_gdp\_emissions and created a new column which included the logged values for GDP and CO2. Then, I created two models, one with only the GDP as the explanatory variable, and CO2 as the response variable, and then a second multiple linear regression adding in Year.

**Slide 16- Scatterplot with Least Squares Line**

So on this slide – we are looking at the least squares regression for GDP and CO2 – which looks decent (considering the assumptions that weren’t met). But, there are clearly outliers – this means that for certain values, the predictions may be very off.

**Slide 17 – Residuals Histogram Plot**

Then we will check if the residuals are normal (or close to normal), which for the most part it is – with the exception for those few outliers.

**Slide 18 -Residuals vs. Fitted (Predicted) Values**

So in this plot, we want to check that there is constant variability of points, which for the most part is true, except for the negative residuals for the smaller fitted values.

**Slide 19 – Curvature Test**

So this is also checking for variability but it is also incorporating the curvature test – using the CAR package. This shows slight curvature, which can be a red flag.

**Slide 20 – QQ Plot**

For this graph, we’d ideally like to see a linear relationship. However, the heavy tail indicates there may be skewnewss in the residuals – this is signaling that yep our residuals are not totally normally distributed.

**Slide 21 – Predict CA 2023**

Although the last slides were using the first model, I wanted to incorporate the multiple linear regression or log\_m2 in this presentation too, so using the modeled results we can predict what the CO2 emissions from energy produced in California were for 2023, then we can see how far off we were! Based on this result I can say that it is pretty far off – but California was an outlier so that is to be expected.

**Slide 22 – Predict PA 2023**

Here we can use Pennsylvania to see how far off we are – but measure this proportionally. Here it is showing we are about 25% off – which considering all the assumptions we failed to meet for a regression – kinda surprising.

**Slide 23 – Predicted vs. Real Values for 2023**

I wanted to see how far off we were and in which states so I made a map with a tool tip that covers the relative residual, then binned it into colors to classify how “off” the results were. So in conclusion – I would not trust this model, and likely would want to explore other ways to model this data.

**Slide 24- References**

* Here are links to the reports I used as references and the data used in this presentation.

**Slide 25- Quick note**

This is the note I mentioned about the definitions provided in the annual electric power report, on this slide I just have reference to where to find the definitions for producer type and energy source.

To Dos:

Send Angela Lui presentation: [angela.lui@cuny.edu](mailto:angela.lui@cuny.edu)

Because I am a tab hoarder ☹

Plotly configuration: <https://plotly.com/r/configuration-options/>

Revealjs instructions: <https://quarto.org/docs/presentations/revealjs/>

Revealjk Options (instructions cont.): <https://quarto.org/docs/reference/formats/presentations/revealjs.html>

Quarto Presentation Github Code: <https://github.com/quarto-dev/quarto-web/blob/main/docs/presentations/revealjs/demo/index.qmd>

Demo presentation: <https://quarto.org/docs/presentations/revealjs/demo/#/code-animations>

Example Presentation from JB: <https://fall2024.data606.net/slides/Example_Project_Slides.html#1>

Medium Article: <https://anshikaaxena.medium.com/how-skewed-data-can-skrew-your-linear-regression-model-accuracy-and-transfromation-can-help-62c6d3fe4c53>