**Our Dataset and Why we Chose it - Global Data on Sustainable Energy (2000-2020)**

* 20 Year range
* 3469 data points prior to cleaning, but categories are easily relatable, and we were able to engineer additional points for analysis.
* Interesting topic, we have some background in this topic in the group as well. Easy to hit the ground running

**Slides Notes – Data Engineering**

Dropped

Renewables (% equivalent primary energy)

Population - Density\n(P/Km2)

Cleaning

Financial flows to developing countries (US $) **Dropped Null/Filled with zeros**

Renewable-electricity-generating-capacity-per-capita **Dropped Null /Filled with zeros**

Feature Engineering

Created new Dataframe to consolidate data on a regional basis. This was done for some of the visuals and leaderboards to tell the data story over the twenty years that are covered more concisely. This also helped with some of the limitations of the data that we faced.

**Regression notes and talking points**

* Regressions set to year 2020 specifically for most recent data being used
* Large left skew, due to high number of counties with a lower relative energy consumption and GDP
* Low correlation on the relationship between both Electricity from Renewables and Fossil Fuels
  + R Squared values – higher for renewables but only slightly (.05),
  + Not enough to be a predictive model – but does indicate that there are strong factors outside of GDP that would impact
* Data sets have a similarity is that the number of data points that we are considering with values not equal to zero are 147 and 148 respectively

A black background with white text

Description automatically generatedA computer screen shot of a computer screen

Description automatically generated

**Bias and Limitations**

* Timeframe of the data – 20 years is both a good sample size and a factor in the limitation of the data’s scope
* Impact of factors that cannot be refined to this data set
  + Socio-economic
  + Political
  + Public Perception
* Size of the data set, 20 columns with 3469 each datapoints prior to cleaning
* Not all counties in the world included – 148/195
* Bias in countries that are developed vs undeveloped, data set puts them on an equal basis which is helpful in a straight comparison, but there are a high number of external factors that aren’t taken into account with the data set that would impact renewable usage between the two categories of countries.

**Conclusion/Final Thoughts/Call to Action**