# **Data Gathering**

The data for the project was chosen from three different types of sources:

1. **Oil and gas production data**: The historical oil and gas production data for all states under consideration was gathered. This was a bit challenging, as the data is not available in a single place. For Texas, the Railroad Commission (RRC) provides historical oil and gas data. However, this is not trivial to extract from their website, even though it is supposed to be free. Here, the Enigma website proved to be very helpful. They also don’t have the Texas data on their website, but they were willing to provide me with the raw data. For the other states, Enigma had the oil and gas production data organized by well, and it was straightforward to download it from their website.
2. **Unemployment data**: The Bureau of Labor Statistics (BLS) provides local area unemployment data for the past several decades. Getting data from their website is also not straightforward. One can download it by county and year, one by one, but for the large number of years I was interested in, this was too tedious and time consuming. The BLS provides a public data API, and I wrote my own code using that.
3. **Income tax returns**: The Internal Revenue Service (IRS) provides annual income tax collections by county on their website, which can be downloaded easily.
4. **Oil price**: The historical oil price by month was downloaded from the Energy Information Administration (EIA) website. Though they have an API, it is more straightforward to just download it.

# **Data Cleaning**

All the data sources contained mostly clean data. In some cases, there were missing values of oil price for a day. We are not interested in daily oil prices, just the aggregate monthly average oil price for comparing against monthly changes in oil production, and national averages for comparing against income tax returns. The other data sources were already well curated at source, and did not need much cleaning.

# **Data Manipulation**

All three types of data needed to be handled differently:

1. **Oil and gas data**: The oil and gas data needed to be organized by county. So the oil and gas production by well needed to be cross referenced with the well information, which gives the location of the well. This can help allocate the county for each well, and hence the oil and gas production can be aggregated by county, for each month, since the starting date to the end date.
2. **Unemployment data**: The unemployment data is extracted in the form of county codes, reported by month and year. The county codes should be modified to actually report the county, and the month and year columns were modified to be a single time column. This made plotting the data much easier.
3. **Income Tax returns**: This data is reported annually, and the only manipulation necessary is for which data to compare against.
4. **Oil price**: The oil prices need to be averaged over a given month.