# **SIADS 591-592 Milestone 1 Project**

## Greenhouse Gas (GHG) Emissions from Upstream and Midstream US Oil and Gas Operations

By Rafee Shaik and Greg Myers  
April-May 2020

## Project Overview

The Oil and Natural Gas industry consists of three sectors, the **Upstream** sector that focuses on Exploring and Producing (E&P) Hydrocarbons, the **Midstream** sector focuses on Transportation and storage facilities, and the **Downstream** sector will process and refine raw materials such as crude oil into consumer products like gasoline.

Enhanced hydrocarbon extraction methods, that include horizontal drilling and fracking boosted Crude Oil and Natural Gas Production in the US since 2007. The purpose of this project is to examine Greenhouse Gas (GHG) Emissions from upstream and midstream sectors of the industry and determine if there is a correlation between accelerated hydrocarbon (Crude oil and Natural gas) production in the most recent decade and GHG emissions rates. If a correlation is found, additional analysis may be able to reveal the causal source. The project will focus on three GHGs, Carbon-Dioxide (CO2), Methane (CH4), and Nitrous Oxide (N2O).

## Project Motivation

There has been an increase in US hydrocarbons production since the shale boom that started in 2007/2008 (Fig-1). Hydrocarbon production in the US increased steadily since then, this increase in production led to an increase in GHG emissions from Upstream and Midstream operations of Oil and Gas companies.

It is in every stakeholder's interest to control these emissions while optimizing the production. These stakeholders include the operating company, environmental protection agencies, and local and federal governments. The goal of the study is to find any correlation between increased hydrocarbon production and industry sector GHG emissions? Does the correlation apply to the industry as a whole or one or more individual components? This study can also lead to identifying the opportunities to improve the pipeline infrastructure and invest in pneumatic devices that can detect and prevent hydrocarbon emissions.

Both the project team members work in the Oil and Gas industry and are intrigued to find opportunities to reduce GHG emissions while optimizing the production.

## Data Sources

1. **Greenhouse Gas Emissions Data**  
   **Source:** US Oil and Gas Upstream (Exploration & Production) and Midstream (Pipelines and Storage) facilities can be accessed from the U.S. Environmental Protection Agency (EPA) FLIGHT database.  
   **Location:** <https://ghgdata.epa.gov/ghgp/main.do>  
   **Access Method:** Facility Level Information on GreenHouse gases Tool (FLIGHT) database can be accessed through the website: <https://ghgdata.epa.gov/ghgp/main.do>. Download the data in excel format after selecting appropriate filters.  
   **Format:** Excel spreadsheets  
   **Dataset Size:** Six excel spreadsheets with a total of 41K records.  
   **Period:** This data covers emissions from upstream and midstream Oil & Gas operations between 2011 and 2018.
2. **Name:** Crude Oil and Natural Gas Production volumes

U.S. Field Production of Crude Oil, and U.S. Natural Gas Gross Withdrawals; Yearly  
**Source:** Energy Information Administration (EIA) datastore.  
**Location:** <https://www.eia.gov/opendata/qb.php?category=371>  
**Access Method:** API query  
**Format:** JSON  
**Dataset Size:** Crude records: 161; Gas records: 84; ~3 kilobytes each for crude and oil  
**Period:** Crude production records from 1920 to 2019, Natural Gas production records from 1980 to 2019.

1. **Name: Emission Data from other Industrial Sectors**

# **Source:** Emissions data from other Industrial sectors can be downloaded from the **EAP data store**. We will use this data to compare GHG emissions from Oil & Gas systems and other industrial sectors.

**Location**: [**https://cfpub.epa.gov/ghgdata/inventoryexplorer/#industry/allgas/source/all**](https://cfpub.epa.gov/ghgdata/inventoryexplorer/#industry/allgas/source/all)

**Access Method**: Web Scraping if there are no popups preventing web scraping, otherwise use the manual download option.

**Format**: CSV or Web Scraping, CSV name will be IndustryWiseGHGEmissions.csv, if we’re not able to scrape it from the web.

**Dataset Size**: 243 records

**Period:** This data covers emissions from other industries between 2011 and 2018

## Data Manipulation Methods:

### **Processing Emissions Data**:

1. **Data Acquisition**:
   1. Data that was downloaded from EPA is in Excel spreadsheet format.
   2. EPA Emission reports are separated by the industry sector (upstream and midstream) and GHG gas type (CO2, CH4, and N2O). All together we have 6 excel spreadsheets.
   3. Within each excel we have a separate worksheet for each reporting year, starting from 2011 to 2018.
2. **Parsing Excel spreadsheets**: Pandas’ **read\_excel** module is useful in parsing excel spreadsheets. Read\_excel module can read multiple worksheets within an excel book. The option **"sheet\_name=’null’"** is useful to read all sheets from an excel. Read\_excel() returns a named dictionary containing worksheet name as key and corresponding data in a dataframe as value.  
   Iterate through each data frame in the dictionary, add three columns, 1. reporting\_year with the value from dictionary key, 2. GHG Gas type with the value from part of the source file name, and 3. Industry sector with the value from part of the source file name.  
   Do the same for all six excel spreadsheets.  
   Combine all dataframes into a single dataframe using pandas.concat() method.
3. **Processing the data**: Each row represents a facility operated under single company ownership or joint venture between multiple companies. So we need to separate the joint ventures and create a row for each company, calculate their portion of emission based on joint venture percentages.  
   We use regular expressions to parse out the company name and partnership percentages into separate columns. For example:

"SHELL OIL CO (51.8%); EXXONMOBIL CORP (48.2%)"

Here is the regular expression used to parse the company names and partnership percentages:

regex**=**r'(?P<PARENT\_COMPANY>[-\w\s\d,&./()#]+)([\(])(?P<CONTRIBUTION>[\(\d.]+)([%\)]\*)'

Melt this single row into multiple rows that represent a separate row for each partner company. Use the method **‘explode’** to separate the individual companies. Convert partnership percentage and emission quantity from text type to numeric type. Replace non-numeric values with zero ‘0’

1. **Aggregation**: We applied different levels of aggregations in our analysis. Data has attributes representing Company Name, Reporting Year, Gas type, and Industry sector.
   1. GHG Gas type analysis: Data were separated by GHG gas type (CO2, CH4, and N2O), all these emission quantities are represented in CO2 equivalents. We will sum them up by the Company and Gas type within the reporting year to compare the emissions of different gas types.
   2. Company level emissions: Each company can operate multiple facilities across the US, we will sum the emissions from the facilities operated by the same company within a reporting year. This data will be used to compare the emissions from different companies over the period of time.
   3. Sector level emissions: Data will be aggregated by the industry sector within the reporting year.
   4. Yearly GHG emissions: We will sum up the emissions from all companies, sectors, and gas types within the reporting year
2. **Joining Emissions data with other datasets**: Reporting\_year is be used as joining key when joining Emissions dataset with other datasets in the analysis such as Yearly Crude Oil and Natural Gas Production and Emissions from other Industries.
3. **Challenges**: Over the period of time companies report their emissions with different names, like ‘Conoco Phillips’, ‘ConocoPhillips’, ‘ConocoPhillips Company’ all these company names represent ‘ConocoPhillips’ We used a third-party library ‘cleanco’ to clean up the company names, this library helped to remove the company type suffixes, like ‘LLC’, ‘Co’, etc. We performed company name lookup to standardize the company names. We put these company name lookup table in a CSV file.
4. **Saving aggregated data for Analysis and visualizations:** Aggregated emissions data will be saved in **'Emissions\_aggregatedData.csv**' CSV file

### **Processing Crude and Natural Gas Production volumes datasets:**

1. Save API query results to a JSON file as an immutable source data reference.
2. Import JSON data (crude & gas) into Pandas data frames.
3. Perform Explode operations to separate date and production data.
4. Transform Date column into a DateTime data type.
5. Create and populate a product type column and drop unused columns.
6. Append crude and gas data frames (long-format).
7. Save the data frame to a CSV file with the name **'Processed\_AnnualProductionData.csv'** for analysis and visualizations.

### **Processing GHG Emissions from Other industries:**

1. Data is present in wide-format, unpivot the data to put it in long-format, use pandas melt method.

### **Data Integration:**

Combine the above two datasets (Emissions and Production volumes) by the Reporting year, with this combined dataset we can compare greenhouse gas emission volumes with Crude Oil and Natural Gas production volumes. Joined dataset is saved in a CSV file with name 'ProductionVsEmissionSplit.csv'

## Analysis and Visualization

The focus of the analysis is studying the impact of increased Hydrocarbon production on Greenhouse Gas (GHG) emissions from the Oil & Gas industry.

To perform this analysis we joined the annual Oil and Natural Gas Production data from EIA (Energy Information Administration, EIA.gov) with annual GHG emissions data from EPA (Environment Protection Agency EPA.gov). These datasets were joined by the Reporting Year as key. The processed and merged dataset was saved in ‘merged.csv’ file. We used this merged dataset for our analysis and visualizations.

We plotted yearly production and emissions volumes on a line chart to observe the trend in GHG Emissions and Hydrocarbon productions between the years 2011 and 2018.

Fig-8

We observed both GHG Emissions and Hydrocarbon volumes are on increasing trend between 2011 and 2018.

To understand the GHG Emissions from the Oil and Gas sectors better we did a deeper analysis by comparing GHG emissions from individual Oil&Gas sectors (Upstream and Midstream) and different hydrocarbon products (Crude Oil and Natural Gas).

Figure-7.

We calculated correlation coefficient/Pearson-R between GHG Emissions from different sectors and different hydrocarbon product production volumes.

We observed a strong-to-very strong correlation between all hyderocarbon productions and all GHG sector emissions. Particularly we observed the highest correlation between ‘Natural Gas Production’ and ‘Combined GHG Emission.

Product: Combined Production , Sector: Combined Emission , pearson-r: 0.8190218781217727

Product: Combined Production , Sector: Midstream Emission , pearson-r: 0.7307744998650313

Product: Combined Production , Sector: Upstream Emission , pearson-r: 0.5836843921994607

Product: Crude Production , Sector: Combined Emission , pearson-r: 0.7620534805062806

Product: Crude Production , Sector: Midstream Emission , pearson-r: 0.6723432310890928

Product: Crude Production , Sector: Upstream Emission , pearson-r: 0.5850816693652066

Product: Natural Gas Production , Sector: Combined Emission , pearson-r: 0.8700047336668489

Product: Natural Gas Production , Sector: Midstream Emission , pearson-r: 0.7874070889774125

Product: Natural Gas Production , Sector: Upstream Emission , pearson-r: 0.5584520095475436

This Pearson-r value may conclude that there is a strong correlation between Natural Gas Production and overall GHG emissions. Since we have only 8 years of data we will be cautious to conclude there is a causal relationship between Natural Gas Production and GHG emissions.

Linear regression:

From the above Pearson-r calculations we observed there is a strong correlation between annual Natural Gas Production volume and annual GHG Emission volume. We will use ‘Natural Gas Production’ as our independent variable (feature selection) to predict the GHG Emission volume for year 2019. Since we have only 8 data points, we will use all the datapoints to test our linear regression model.

Refer to the Liner regression code module here.

Our linear regression model predicted 337 Million metric tons of CO2 equivalent GHG emissions for the year 2019.

1. After using both plotly.graph\_object and plotly.express we feel like plotly.express has improved functionality to produce visualizations much faster.

2. Better code management: even though we used github for team coding we were working on the master branch which cause code overwrite towards the end of the project. We should have created individual branches and avoided overwriting the code.

3. Bokeh seems to a popular visualization library amount data scientists. We will try to use Bokeh in future projects.

To gain experience with technology and analysis both Rafee and Gregory worked in parallel in data acquisition, processing, and visualizations.

This project doesn’t contain any proprietary datasets and doesn’t require an NDA (Non-Disclosure Agreement) between the University of Michigan and any organization.

Rafee collaborated with his organization business unite to review the emissions project idea using the publicly available datasets. Some of the findings and tools created as part of the may benefit the organization.

Rafee Shaik has contributed to acquiring the emissions data from EPA, parsing the Emissions data excel spreadsheets, and processing data for further analysis. Rafee also collected and processed emissions data from other industries. He combined emissions and production data and prepared various interactive visualizations that compare emissions from upstream and midstream sectors; and between companies and different GHG gas types.

Rafee calculated the correlation coefficient between different Hydrocarbon production volumes and sector-wise GHG emission quantities. He prepared a linear regression model to predict future GHG emission quantity.

He collaborated with Gregory in preparing the final project report.

In this notebook we generated 8 visualizations, seven of them are time-series line+scatter plots and a Summary table (fig-2) showing emissions from each company between 2011 and 2018.

* Fig-1: Trend line of total GHG emissions from US Oil and Gas Companies between 2011 and 2018.
* Fig-2: A table showing GHG Emissions from individual US Oil and Gas Producers between 2011 and 2018
* Fig-3: An Interactive chart - Emission trends between 2009 and 2018 from the top US emitters
* Fig-4: An Interactive chart to compare emissions between different companies and sectors
* Fig-5: A 4X4 Subplots to show, Emissions by sector, GHG Gas Type, Number of Operators(companies) in each sector and Total emissions
* Fig-6: An interactive chart to compare Emissions from Upstream and Midstream Sectors VS Crude Oil and Natural Gas Production
* Fig-7: A histogram to identify the common emission quantity range from US Oil & Gas companies
* Fig-8: A line chart with predicted 2019 GHG Emission

We prepared a linear regression model between Hydrocarbon production and GHG emission volumes; We will use this model to estimate the future emissions based on production volume

## Project code pipeline

The project code and visualization are split into two jupyter notebooks.

1. Notebook **‘EmissionsDataPreparation.ipynb’** will concentrate on data acquisition and Processing. The output of this notebook will be saved in several CSV(Comma Separated Values) files. The output files include:
   1. **Emissions\_aggregatedData.csv** - This file will contain Reporting year, Company, Gas, Sector, and GHG Emission volume and 2018 Emission wise rank
   2. **Emissions\_OtherIndustries.csv** - Emissions from Other industries - columns - Reporting Year, Industrial Sector and Emission volume
   3. **ProductionVsEmissionSplit.csv** - Production and Emission numbers are split in Emission-sector and Production-Product types, Data present in Key-value types from 2011 onwards
2. Notebook **‘EmissionsProject-Visualizations.ipynb’** will take the CSV files prepared by EmissionsDataPreparation.ipynb notebook and generate required visualizations and analysis.

###### *Visualization Technique:*

We used graph\_objs library from plotly for all visualizations in this analysis.

* Most of our plots are scatter plots with line marks over time-series data.
* We used ipywidgets to add interactivity with the charts.

**Other libraries used:**  
Pandas, numpy, scipy, matplotlib, sklearn