**Project Title 1:** Projection of CO2 Emissions from Electricity Production in the US

**Project Theme** The purpose of this project is to forecast CO2 emissions related to electricity generation in USA. Statistical models such ARIMA will be used to analyze emissions data for US

**Problem Statement & Business Case**

Forecasting future changes in climate is a very complicated task and requires a wide breadth of information to even come close to an accurate prediction of future changes. However, there are some major influencers that are useful for providing rough estimations of future increases or decreases in the global temperature. Of these, one of the most effective is greenhouse gasses, which cause light to be trapped in the atmosphere longer due to the way in which they scatter light. This means that increased levels of greenhouse gasses cause more energy to be transferred to the Earths atmosphere throughout the 'random walk' of incoming photons from the sun, since the increased presence of these gasses cause light to stay in the atmosphere longer. Therefore, being able to forecast future increases or decreases in greenhouse gasses such as methane (CH4), carbon dioxide (CO2), water vapor (H2O), and nirous oxide (N2O) is a vital part of being able to accurately predict future changes in the Earths temperature. That is why for this project I attempted to model the trends in CO2 emissions from a wide range of sources using ARIMA modeling in order to determine if this type of model could be used to help in identifying future climate trends.

**Data Collection**

For this project the dataset I selected is the [carbon emissions dataset](https://www.kaggle.com/txtrouble/carbon-emissions/metadata) which contains data corresponding to the monthly and annual CO2 emissions from electricity generation, which was acquired from the US Energy Information Administration and contains data from 1973-2016. The data is broken down by fuel type and the values below correspond to million metric tons of CO2.

**Reference**

[Electricity Data - U.S. Energy Information Administration (EIA)](https://www.eia.gov/electricity/data.php#elecenv)

[Carbon Emissions | Kaggle](https://www.kaggle.com/txtrouble/carbon-emissions/metadata)

**Project Title 2:** King County House Price Predictions

**Problem Statement & Business Case**

Local real estate company serving King County in Washington, US, asked to build a house price prediction model. The company wants to utilize the model to provide their house price estimations to their customers, house sellers and buyers. The goal of this project is collecting data with house prices and finding the best model that can predict house sale prices with the least amount of errors. The goal of this part of project was finding the best models that can predict house sale prices using various machine learning techniques. The followings are the summary of what I have done and found from the In-depth machine learning analysis.

**Data Collection**

The main dataset, House Sales in King County, USA, was from Kaggle.

The data contains records of house sales for King County included in the Seattle–Tacoma–Bellevue metropolitan area in Washington, US. The dataset has one year of records for houses sold between May 2014 and May 2015, but this project assumed that it is recent data. It consists of 19 house features, house id numbers, dates sold and sale prices for 21,613 house sales. The house features include the number of bedrooms and bathrooms, square footages of home and lot, year built, zip code,latitude and longitude of a house associated with each sale.

**Reference**

<https://www.kaggle.com/harlfoxem/housesalesprediction>

**Project Title 3:** Yelp Sentiment Analysis

**Problem Statement & Business Case**

Positive yelp reviews are crucial for a business. Yelp not only drives traffic to

business but easily helps users make decisions about which establishment to visit.

These decisions are often made based on reviews. A restaurant owner may want to tailor the customer experience to items which lead to high reviews and make business decisions to eliminate/reduce anything that can lead to negative reviews. The goal of project will be to use data science methods particularly Supervised ML that will help establish a connection between the rating levels and what might drive them.

**Data Collection**

The data is provided by Yelp for as an open dataset( https://www.yelp.com/dataset ). This large subset data from yelp that contains 5 different JSON encoded files. These files contain data on the business, the customer review, the user, the datetime of visit, and additional tips left by the user.

**Reference**

https://www.yelp.com/dataset

**Project Title 4: Twitter User Gender Classification**

Project Theme Sentiment analysis is famous among major brands. It is the way to identify the tone and emotions expressed through written or spoken online communication.

Problem Statement & Business Case

The goal of this project is to simply view a Twitter profile text data and judge whether the user was a male, a female, or a brand (non-individual). NLP is used and its different methods pave the way for achieving a solution for performing the analysis. This is useful for the prospective client in determining a particular gender in analyzing well, based on seeing a user’s tweet or a profile.

**Data Collection**

This data set was used to train a CrowdFlower AI gender predictor. This dataset was from a Kaggle competition. The dataset contains 20,000 rows, each with a username, a random tweet, account profile and image, location, and even link and sidebar color.

**Reference**

[Twitter User Gender Classification | Kaggle](https://www.kaggle.com/crowdflower/twitter-user-gender-classification)