Assignment 1: Data Science Research Challenges

**Name: Tyler Conger Email: tconger1@crimson.ua.edu**

Statement:

For this assignment's preparation, the author(s) did not use any generative AI tools.

**Objective**: Develop a deeper understanding of real-world data science problems and the datasets that can potentially solve them.

**Instructions**:

1. **Research & Identify Problems**: Based on your research, pinpoint five pressing data science challenges. These can range from industry-specific issues to broader analytical dilemmas.
2. **Description**: For each problem, offer a succinct description to ensure clarity of the challenge at hand. Your description should provide context and highlight why this particular problem is significant in the realm of data science.
3. **Datasets Identification**: For every problem you've described, identify and list at least two datasets that you believe can be instrumental in addressing and possibly solving the said problem. Provide brief details of each dataset such as its source, size, and key variables.

**Submission Guidelines**:

* Compile your findings in a structured document, ensuring clear headings for each of the five problems.
* Ensure that your descriptions are concise yet comprehensive, offering enough insight for someone unfamiliar with the topic.
* For each dataset, include a link (if available) and a short annotation about its relevance to the problem.

**Evaluation Criteria**: Your assignment will be evaluated based on the originality of problems identified, clarity of descriptions, and the appropriateness of the datasets chosen for each problem.

Identified Data Science Challenges:

# Drought prediction in Forested Land

For forest official, such as rangers and park keepers, it is often important to predict what the coming year will bring. This is important to identifying the health of the ecosystem. In this data science challenge we are setting the goal of predicting drought years (years with lower than average rainfall) and wet years (years with higher than average rainfall). We will be doing this by analyzing past tree data, such as tree ring width. Each year a new tree ring is created and this can help to reflect what occurred that year, when there is better rainfall there is better growth which is reflected in the tree ring.

Your challenge is to create a model that will predict if the next year will have high, average, or low rainfall based on the tree ring data of previous years.

For this challenge some suggested data bases are linked below. These each include information on tree rings by year.

[Database One](https://datasetsearch.research.google.com/search?src=2&query=International%20Tree%20Ring%20Data%20Bank%20(ITRDB)&docid=L2cvMTF0eTc2MTQ4ZA%3D%3D)

[Database Two](https://datasetsearch.research.google.com/search?src=2&query=International%20Tree%20Ring%20Data%20Bank%20(ITRDB)&docid=L2cvMTFqOWJzY3d5aw%3D%3D)

# Energy Usage Prediction by Household

For new home owners estimating their yearly (and month-to-month) energy usage can be an important factor. Based on previous data from other homes help to give an accurate estimate their energy usage. This information can later be used by the home owner to calculate expenses, as energy cost can vary by state and region. We will be creating this prediction by analyzing data of household energy consumption and projecting an estimate based on past data.

Your challenge is to create a model that will predict the household energy usage based on household information.

For this challenge some suggested data sets are linked below. The first includes information on household electrical usage. The second includes information on household power usage.

[Database One](https://www.kaggle.com/datasets/thedevastator/240000-household-electricity-consumption-records)

[Database Two](https://data.world/databeats/household-power-consumption)

# Household Energy Efficiency

In our current world, having green houses and buildings is becoming increasing important in the fight for action against climate change. Your job is to identify households that would be best impacted by an upgrade to appliances to increase efficiency. The goal being to implement the highest level of improvement for the lowest potential cost, e.g. small fixes hat have big impact and help to reduce waste.

Your challenge is to create a model that identifies and signals out data points that would be the best for this energy improvement program, meaning they would benefit the most from improvement.

For this challenge some suggested data sets are linked below. Both of these data sets are specific to certain areas but may help your model but more data may be necessary.

[Dataset One](https://www.gov.uk/government/statistics/national-energy-efficiency-data-framework-need-consumption-data-tables-2021)

[Dataset Two](https://www.kaggle.com/datasets/thedevastator/analyzing-energy-consumption-patterns-in-chicago)

# Best Time to do Road Work Based on Motorist Trends

In the United States the classic “Road Trip” is familiar to many. However based on the seasons and local happenings there may be more or fewer drivers on the road at a certain time. This is important information for the transit administration to know so they are able to know when road work would be best and affect the fewest number of travelers.

Your challenge is to create a model that predicts what time of the year would be best for the transit administration to do necessary roadwork. The goal being to impact the fewest number of drivers, while also maintaining roadways.

You may find the below datasets useful for this task.

[Dataset One](https://datasetsearch.research.google.com/search?src=0&query=interstate%20usage&docid=L2cvMTFybXN4eHJ5cw%3D%3D)

[Dataset Two](https://www.kaggle.com/datasets/anshtanwar/metro-interstate-traffic-volume)

# Identifying the best time to buy a house

The housing market is fairly cyclical and often fluctuates. It can be important for first time buyers to be aware of what may be a good time to spring on a house. The task here is to create a model that is able to identify what time may be a good time to put an offer in for a house, based on trends and patterns in the housing market.

For this challenge one of the data sets may be useful the first of which is the average price sold in the US from 1965. The second dataset includes more information on the type of house, such as size and location.

[Dataset One](https://www.statista.com/statistics/240991/average-sales-prices-of-new-homes-sold-in-the-us/)

[Dataset Two](https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset)