DATA SCIENCE CHALLENGE

*Thank you for your interest in SparkCognition. The following DS Challenge allows our team to better understand your technical abilities and gives you some insight into the type of work SparkCognition’s data scientists encounter on a daily basis.*

*You will have three (3) days to complete the exercise which is estimated to take approximately two (2) hours. Please aggregate your responses to a single notebook or file with associated comments and include your full name in the file name. Acceptable submission formats include .ipynb, .html, and .pdf.*

*Each section will be weighted equally in the evaluation. In addition, please note that you will be evaluated on your quality of coding, thought process, and clarity in communication.*

# Exercise 1: Data Preparation and Preprocessing

*Please complete the following exercise, using comments in your code to explain your reasoning.*

Your customer is a power trading organization. You have just received a dataset from them and are trying to clean raw data in preparation for modeling. The data set includes hourly weather metrics as well as power output values in kWh.

Please clean and prepare the dataset for modeling based on the following instructions. Use comments to explain your reasoning along the way.

## Datasets:

Power\_data\_preprocessing.csv Weather\_data\_preprocessing.csv

## Instructions:

1. The customer has shared two datasets with inconsistent quality.
   1. Some of the weather data is missing. Using the information in the column header, please determine the appropriate way to handle the missing data and **explain your reasoning** in the comments.
2. The Power data is particularly useful for modeling purposes but needs some preprocessing.
   1. Power usage can be affected by the day, month, or season. Please choose one of these timeframes and create a categorical variable for it.
   2. Please determine if any feature scaling is required and, if so, choose an appropriate scaling method. **Briefly explain your reasoning**.
   3. The company experienced a power outage due to high winds. On days where the wind was 30mph or above customers lost power. Please create a filter to remove the dates when customers lost power.
3. In the weather dataset,
   1. Can you find any strong relationships between any of the features? If so, which ones and what do you propose to do with them? Thoroughly **explain** your methods and findings.

# Exercise 2: Model Building

*Please complete the following exercise in the same file as the other exercises, using comments in your code to explain your reasoning.*

An auto-insurance company is revamping its pricing model. The analyst developing the new price model believes that the best approach is to develop 2 models: one for customers who are likely to file an insurance claim within the first year of their contract and another one for all other customers. The analyst has prepared a clean dataset consisting of 10,000 customers and 10 engineered features which capture driving behavior. The data has already been preprocessed for you (i.e., no missing data, no outliers, data is scaled, no correlated features, and the classes are fairly balanced).

The data is contained in *claim\_prediction.csv*, where CLAIM = 1 means the customer filed a claim in the first year and CLAIM = 0 means the customer did not.

Develop a model to predict if customers will file a claim in their first year based on their driving behavior.

In addition to submitting your code, use comments to **explain** the decisions you made and how well you expect this model to perform on new data from a similar customer pool (and why).

Note that you are being evaluated on your model building and validation workflow, rather than on the complexity of your solution.