**Hello!**

Thank you for your interest in the Data Scientist position at PSE. In this role, you will be exploring data and using machine learning techniques to design and build algorithms. We have prepared a dataset that is representative of the data we work with, and we’d like to see how you extract insights and apply machine learning techniques to it. Since we know that there are many ways to approach a problem, we view this as a very open-ended assignment.

**The problem:**

PSE’s marketing team is considering offering a rebate to our customers to upgrade their gas furnace to a new, more fuel-efficient model. But first we need to understand which customers own gas furnaces. In a typical Pacific Northwest home, gas furnaces consume the most gas compared to any other appliance. They are used primarily on cold days to heat a house. Analyzing gas usage patterns as a function of temperature can be good way to differentiate customers with and without gas furnaces. For example, we might expect a customer with a gas furnace to have higher usage on a cold day compared to a hot day. But we don’t expect the same to be true for customers that don’t use gas heating.

*How would you approach looking for customers with and without gas furnaces?*

*Use* *data visualization, statistical methods, machine learning techniques, etc. to design and build a proof-of-concept gas furnace detector.*

**The data:**

The dataset is a ~10MB file with about 750K records. Each record represents a single observation of usage and temperature for one customer on one day. The records represent a year of data for about 2000 PSE natural gas, residential customers. There are three fields:

* CUSTOMER\_ID – The customer identification number. String.
* DAILY\_USAGE – The gas usage for a single day. Float.
* AVG\_TEMP – The average temperature for the customer's location on the same day that DAILY\_USAGE was measured. Float.

**The deliverable:**

Please prepare a Jupyter notebook showing your work (documentation is very important to us!) and a short (<15 minute), informal presentation on your findings. Assume that the audience is a group of business stakeholders who are very familiar with the problem statement and dataset but not as familiar with data science techniques and terminology. Be prepared to answer questions about your thought process and machine learning choices.

**Notes:**

* There is no right way of solving this problem. There are many possible methods, all with different advantages.
* Use the internet (we do!)
* We expect you to spend a few hours on this. We don’t expect a complete, perfectly polished solution. However, we would like to know what additional steps you would take if you had more time.