**Data Analytics with Python – Assignment 2**

**Group 6 Members:**

Ma Estela Arenas

Sean Howard

Yuxiao Liu

Feng Nie

**Data Science Group Project**

**1.0 Business Understanding** – *ask relevant questions and define objectives for the problem that needs to be tackled*

It has often been observed that energy consumption tends to be at its highest on days with

hotter temperatures. In this project, our group will develop models that predict the maximum daily energy usage and pricing category based on provided weather data. The hope is that these models can be used to predict likely energy demands based on a weather forecast, which can help energy companies understand plan for future usage, and help businesses plan when to conduct energy-intensive operations.

**2.0 Data Mining** – *gather and scope the data necessary for the project*

Data sets provided were **weather\_data.csv** with 243 rows and 21 columns, has blanks and columns with both float and string, and **price\_demand\_data.csv** with 11,664 rows and 4 columns.

**3.0 Data Cleaning** – *fix the inconsistencies within the data and handle missing values*

**3.1 Changes Made**

Some data cleaning has been done on weather data.

Data cleaning on Price table, VIC column deleted, convert time into date format and then format into string format MM-dd,a,HH:mm. so that later on, we can easily grab the date and time. convert demand to INT, as the floating point number is not that important. No change to the pricing category, all data there are looking great.

The AM/PM data has been kept because there are separate temperature data on the weather table, and have yet decided its relatedness.

**3.2 Assumptions**

There were missing data on row 190 for date 07/08/2021. The minimum, maximum and 9 am temperature will be the same as the given 3 pm temperature which is 12oC. Rainfall is assumed to be 0 mm.

**3.3 Limitations**

Date range used in this project is between 1st of January and 31st of August 2021. Demand usage is within the 30-minute time interval daily.

**4.0 Data Exploration** *– form hypothesis about your defined problem by visually analyzing the data*

**5.0 Feature Engineering** *– select important features and construct more meaningful ones using the raw data that you have*

Model 1 Goal: Predict the maximum daily energy usage based on provided weather data

Get the highest usage per 30 min row (one row only) to represent the max daily usage for the day

independent variable – temperature

dependent variable – maximum daily energy usage

Model 2 Goal – Predict the maximum daily price category based on provided weather data

Get the price category based on the highest category for the day (example if we have 41 low, 0 med, 7 high we will choose HIGH)

independent variable – temperature

dependent variable – maximum daily price category

**6.0 Predictive Modelling** *– train machine for learning models, evaluate their performance, and use them to make prediction*

linear regression = demand prediction

classification = price prediction

*see model template.jpynb*

**7.0 Data Visualisation** *– communicate the findings with key stakeholders using plots and interactive visualisations*

***Questions to be answered on this assignment:***

*1. What wrangling and aggregation methods have you applied? Why have you chosen*

*these methods over other alternatives?*

*2. How have you gone about building your models and how do your models work?*

*3. How effective are your models? How have you evaluated this?*

*4. What insights can you draw from your analysis? For example, which input variables*

*are most valuable for predicting energy usage/price?*

*5. Why are your results significant and valuable?*

*6. What are the limitations of your results and how can the project be improved for future?*