Code Challenge

1. Data

a. For this challenge, only the data from July and August will be used for the prediction. This is because in the summer electricity is the main source of all cooling system. In the winter, the heating system uses other energy sources such as gas and heating oil and therefore the electricity demand profile is different.

b. The temperature should be the main driver for the electricity demand. Actually, it's the cooling degree hour (i.e. max(temperature - base temperature, 0)) that should be used with the base temperature set at 18 degrees C.

c. The data from 2016 to 2018 are used for training the model; data from 2019 is used for validation; and data from 2020 is used for testing.

d. A new column called WeekdayType is used to specify whether the day is a weekday or in the weekend/holiday. The holidays are July 1st and Civic holiday.

e. There are missing hours. But since the serial correlation is not being considered here, we don't need to fill in the missing values.

2. The linear model is chosen with the main driver being the cooling degree hour (CDH). The optional variables are the weekday type (weekend/holiday v. Weekday), hour, relative humidity, dew point, humidex, and wind speed, all included in different combinations. In addition, we also need to account for the demand that is independent of the weather and day/time (related to occupancy). To model this component, we take the residuals that are obtained after the data is fitted with the weather and time variables, group the residuals by hour and weekday type, and take the mean of the residuals. We end up with mean residuals by hour and weekday type. This mean residuals are then added back to the model prediction.

predicted demand = weather variables + day/time variables + intercept + mean residual due to other processes

The model does not consider serial correlation. Therefore, it can predict demand at any future points in time as long as those future points of the drivers (CDH, etc.) are available.

3. The final model obtained using the sample training and validation data is:

predicted demand = CDH + Hour + WeekdayType + Dew_Point + intercept + mean time-dependent residual due to other processes

4. For testing, all data in July and August are used. We assume all the temperature and dew point data are available for the prediction. Using the test set, the proportion of test data with precision below 500 MW is around 47%.

5. To run the entire model (including training, validation and testing), use this line:

```
model, splits, test_predicted, test_score = run()
```

The model dictionary contains the keys linreg (a linear regression object) and component (the mean time-dependent residual due to other processes).
The dictionary splits contains the training, validation, and test sets.
The test_predicted is an array of demand predicted using the test set.
The test_score is the proportion of test data points having the acceptable error < 500 MW.