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 were used for training the model; data from 2019 was used for validation; and data from 2020 was used for testing.

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

e. There were missing hours. But since the serial correlation was not being considered here, they were not filled in.

2. The linear model was chosen with the main driver being the cooling degree hour (CDH). The optional variables were 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 needed to account for the demand that was independent of the weather and day/time (related to occupancy). To model this component, we took the residuals that were obtained after the data was fitted with the weather and time variables, grouped the residuals by hour and weekday type, and took the mean of the residuals. We ended up with mean residuals by hour and weekday type. This mean residuals were then added back to the model prediction.

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

The final 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 was:

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

The test score used to evaluate the model's performance was the proportion of the differences between predicted and actual demands that were below 500 MW. The model that yielded the highest proportion is chosen to be the final model. Here, the final model yielded a test score of around 50%

4. For testing, all data in July and August were used. We assumed 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 was 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.

6. To obtain a 24-hour demand forecast from the test set, run the function summerDemandForecast:

```
predicted, actual, difference = summerDemandForecast(model, date, splits[“test”])
```

Where the date is a string i.e. “2020-07-05”. The test set consists of all data in July and August in the year 2020.