MSS Data Analytics Technical Test

This test is used to assess candidate's technical skills in data science.

Assumptions:

- Candidate is proficient in Python.

This test is split into 4 sections.

- Consuming APIs
- Querying Databases
- Exploratory Data Analysis
- Data Modelling

Type your name here:

Note to candidates:

Please install any packages that you feel would help you in the test.

This test is open-book, feel free to Google, however, please refrain from any outside communication.

All code should be well-documented and should account for edge cases, along with instructions on the expected inputs & outputs of the functions.

Emphasis would also be placed on your thought process.

Consuming APIs

An API is a software intermediary that allows two applications to talk to another. At times, we have to consume data from published REST APIs to do data modelling.

In this exercise, you will have to write a function that performs the following:

- Consumes Air Temperature data from the API (https://data.gov.sg/dataset/realtime-weather-readings)
- Returns the maximum temperature recorded for the parsed date, along with the name of location it was recorded at.

After which, please answer the following question:

- What is the maximum temperature recorded in Singapore on 16th November 2020, and the name of the corresponding location?
- If you are unable to retrieve data via the API, please use the provided JSON file [air_temperature.csv]

In [ ]:

```
# Write your code here
```

Querying from a Database

pandasql is a library that allows you to query pandas DataFrames using SQL syntax.

pandasql uses SQLite syntax, which is pretty similar to standard SQL language.

More information can be found here: https://www.sqlite.org/lang.html

Using the library pandasql, you are tasked to write an SQL query to answer the following question:

- #### I want to know the names of the top 3 items sold in terms of quantity.

Details of schemas & tables:

Tables:

- df_product = Product
- df_sales = SalesOrderDetail

In [ ]:

```
# Install package pandasql if not already installed.
!pip install pandasql
```

In [ ]:

```
# Treat each dataframe as though they are individual tables in a database
df_product = pd.read_csv('Product.csv')
df_sales = pd.read_csv('SalesOrderDetail.csv')
# Sample SQL code
output = sqldf("select * from df_product")
output
```

In [ ]:

```
# Write your code here
```

Exploratory Data Analysis

You are a data scientist working in an utilities company that manages power generation plants. These power generation plants supply electricity to the whole country. At the start of the day, the Operations team would have to determine the forecasted demand for the day, and relay that information to the power generation plants for their planning. As power generation plants take time to ramp up & down their production, having an accurate forecast is essential for ensure demand is adequately met.

Currently, the Operations team uses their experience to determine the forecasted power demand. However, the company is embarking on their Data Analytics Transformation journey and would like to augment data analytics into their workflows.

Power generation are reported in MegaWatts (MW) and are in half-hourly intervals:

- entry_date - The date (YYYY-MM-DD hh:mm:ss)
- demand_actual: Actual demand met by all Generation Registered Facilities
- demand_forecast: Forecast demand for scheduling of Generation Registered Facilities

The dataset is found in the file: **electricity_demand.csv**

You are tasked with the following:

- Performing exploratory data analysis on the dataset and share your insights (charts, tables, etc)
- In the form of a chart / figures, help the management identify how well the Operations team is forecasting the demand.
- Identify potential features that may have an impact on power consumption. (Roughly describe how you may want to obtain data for these features.)

Please include as many comments & markdown cells as possible, as we would like to know your thought process.

In [ ]:

```
# Write your code here
```

Data Modelling

After sharing your insights with the management, they are very pleased with your work and want you to produce a machine learning model that is able to forecast the power demand. Ideally, the model should have the following features:

- Accurate
- Explainable
- Minimal effort by operator to use

You have to do the following:

- Propose & justify your choice of model
- How you would test your model
- Metrics you will use to determine how accurate is your model
- Create the proposed machine learning model and assess its accuracy (If time permits)

In [ ]:

```
# Write your code here
```

Save your work!

Submit the following documents:

- Working copy of the notebook (.ipynb) with the following convention **(MSSTT_[Full Name]_[YYYYMMDD].ipynb)**
- PDF copy of the notebook (.pdf) with the following convention **(MSSTT_[Full Name]_[YYYYMMDD].pdf)**
- Text file containing the required packages (use pip freeze to generate), with the following convention **(MSSTT_[Full Name]_[YYYYMMDD].txt)**