HOMEWORK 2 REPORT
CS698004 ST: TIME SERIES ANALYSIS AND FORECASTING
SURAJ KUMAR OJHA (31669171)

- **Dataset details**
  - The dataset consists of following properties:
    - Dataset is called Huawei Public and Private Cloud Release Datasets.
    - It's a univariate time series data.
    - Granularity is minutes.
    - The columns ***'day'*** and ***'time'*** represent the timestamp values.
    - All the rest columns named (***0 to 5092***) represent 5093 different serverless functions.
    - Cell value indicates the number of function invocations that occurred for a specific timestamp and a particular function.

- **Preprocessing Steps**
  - Used ***Forward fill*** method to replace missing values.
  - Used ***Min-Max scaling*** since data were in range of hundreds of thousands. After applying this transformation, data became stationary as verified by the ADF test.
  - Performed ***aggregation*** across columns named from 0 to 5092 to get total invocations.
  - Used **Augmented Dickey-Fuller (ADF)** test to check whether data is stationary or not.
  - Merged the timestamp columns (date and time) to a single column named ***datetime***.

- **Model results and Comparison**

Model	MAE	RMSE	MAPE
Naive Baseline	86.979167	161.954598	12.656594
SES (Simple Exponential Smoothing)	91.509000	153.387247	14.709435
Holt's Linear Trend	89.586084	153.015790	14.231775
Holt-Winters	146.716804	190.252477	26.668826
ARIMA	83.22808	157.832165	12.138456

- **Discussion**
  - **Model Results and Comparison**

        - ARIMA outperforms all other baseline models since data seems to be stationary already and as indicated by the values shown in the performance metric table.
    - **Strengths and weaknesses of dataset**
        - **Strength**
            - Dataset was already stationary.
        - **Weakness**
            - Data points were in the range of thousands which is rescaled [0-1] by the Min-Max scaling.
            - Timestamp values were split into two columns (day and time) in the original dataset.
    - **Insights Gained From Preprocessing and Modeling**
        - Min-Max Scaling equalized function importance, ensuring no single function skewed the total workload prediction.
        - Forward-fill avoided erroneous forecasting caused by gaps in the dataset.
        - ARIMA and Holt's Linear Trend model are suitable for forecasting serverless function invocations, as serverless workloads often exhibit an increasing trend.
- **BONUS POINTS**
    - Hyperparameter Tuning (smoothing parameter, alpha) for SES
        - SES performed best for smoothing parameter alpha = 0.7
    - VAR MODEL on multivariate time series
        - Dataset included (timestamps, CPU usage and requests_per_minute)
        - Var model with forecasting steps set to 10 doesn't perform better than ARIMA model.