MIE 223 Data Science
Lab and Assignment 06:
Time-Series Data Analysis

In this lab and assignment, you will learn and gain hands-on experience working with and analyzing time-series data.

- Programming language: Python (Google Colab Environment)
- Due Date: Posted in Syllabus

Marking scheme and requirements: You are required to use the Jupyter notebook named `Assignment_06.ipynb` to provide appropriate answers to the questions in this assignment. All changes to the `Assignment_06.ipynb` notebook should be committed and pushed to your assignment repository on GitHub.

This assignment has *4 points* in total and the point allocation is shown below:

- Solution to Questions (1 point):
  - Q1(a): 0.1 point
  - Q1(b): 0.1 point
  - Q1(c): 0.05 point
  - Q2(a): 0.1 point
  - Q2(b): 0.1 point
  - Q3(a): 0.1 point
  - Q3(b): 0.05 point
  - Q4(a): 0.1 point
  - Q4(b): 0.1 point
  - Q4(c): 0.1 point
  - Q4(d): 0.1 point
- Code review (1 point)
- Quiz (2 points)

What/how to submit your work:

- All your code should be included in the provided notebook named `Assignment_06.ipynb`.
- Commit and push your work to your GitHub repository in order to submit it. Your last commit and push before the assignment deadline will be considered to be your submission. You can check your repository online to make sure that all required files have actually been committed and pushed to your repository.
- A link to create a personal repository for this assignment is posted on QUERCUS.

Notes that you should pay attention to

1. All your code should be written in the provided notebook named `Assignment_06.ipynb`.
2. All functions must *return* the specified return type.
3. Commit and push your work to your GitHub repository in order to submit it. **Your last commit and push will be considered to be your submission.** You can check your repository online to make sure that all required code has actually been committed and pushed to your repository.
4. **Your code should have output for all assignment questions unless otherwise stated**. If any of the assignment questions asks you to explain your outputs, please use markdown cells to clearly explain them.
5. As a backup plan, also upload a PDF copy of your assignment notebook to GitHub. This is useful especially when GitHub cannot render some images or outputs.
6. During the code review, TAs will go over your answer submissions to assess your understanding of the concepts relating to the assignment and the logic behind your codes. Please, ensure your codes are readable and well commented using python comments and markdown cells when necessary. Not have readable code may impact your code review score.
7. Please note the plagiarism policy in the syllabus.

1 Main Assignment

Please answer the questions below in the provided notebook named `Assignment_06.ipynb` using Google Colab. **NumPy** and **Pandas** functions should be used for all computations. Unless otherwise stated, the Pandas DataFrame '.plot()' function or **matplotlib** should be used to generate plots. **Clearly give titles and labels to all plots. Use a figure size of 15x10 for all plots in Q1-Q3. Use a figure size of 15x20 for all plots in Q4.**

You are provided with two datasets: (1) Microsoft and Apple stock prices from the last 5 years (2) wind speed (m/s) and relative humidity levels (%) in Jena, Germany from 2009-2016. As a data scientist, you want to be able to analyze time-series data to uncover useful trends and correlations over time. Some questions you might want to consider are:

- How can we select the appropriate time-scale for further analysis?
- How can we leverage data visualization to identify trends and/or data pre-processing steps?
- How can we use smoothing techniques to mitigate the impact of noise and why is this helpful?
- How can we leverage time-shifting to uncover useful correlations?

Using the **microsoft_df, apple_df,** and **jena_climate_df** DataFrame, complete the following questions.

Q1. Downsampling

(a) Plot the time series of Microsoft's and Apple's adjusted closing stock prices against time on the **same plot** and include a legend. Use 1-2 sentences to comment whether you see any trends over time and how Microsoft and Apple stock prices compare to each other.

(b) Create 5 new DataFrames that downsample Microsoft's adjusted closing stock prices on a weekly (W), monthly (M), quarterly (Q), semi-annual (6M), and annual (A) basis. Use average downsampling (as done in the lab), and the alphanumeric code you should use for resampling is provided for you in the brackets. Plot the original and all downsampled adjusted closing stock prices against time on the **same plot** and include a legend. Repeat the exercise for Apple's adjusted closing stock prices.

(c) Mention one pro and one con of downsampling. From the plots, what is the most appropriate time-scale to downsample to for this dataset?

Q2. Growth Rate

(a) Use the monthly downsampled Microsoft DataFrame from Q1(b) to compute the monthly return rate (hint: use Pandas **.pct_change()** method). Add these values as a column with the heading **monthly_return** to the monthly downsampled Microsoft DataFrame. Display the first five rows of the modified monthly downsampled Microsoft DataFrame. Repeat the exercise for the monthly downsampled Apple DataFrame. Comment on why the first monthly return rate is NaN (not a number).

(b) Plot the monthly return rate of Microsoft and Apple stocks on the **same plot** and include a legend. Use 1-2 sentences to comment on whether you see any consistent patterns over time and any insight this plot gives that is not obvious from the plot in Q1(a).

Q3. Exponential Smoothing

(a) Create 5 new DataFrames by applying exponential smoothing with $\alpha = 0.1, 0.3, 0.5, 0.7, 0.9$ to the monthly downsampled Microsoft DataFrame (refer to lab). Plot the adjusted stock price of all 6 DataFrames in the **same plot** and include a legend. Repeat the exercise using the monthly downsampled Apple DataFrame.

(b) Use 1-2 sentences to comment on the effect of decreasing the smoothing parameter α.

Q4. Correlation

(a) Provide a 3x1 subplot of the daily temperature in Calgary, Buenos Aires, and Edmonton (in that order) over time. Use 1-2 sentences to highlight the two problems that this visualization reveals (hint: look at pre-processing steps in Q4(b))?

(b) Modify **city_1_df**, **city_2_df**, and **city_3_df** to store records with dates that appear in **ALL** three DataFrames (refer to lab). Next, downsample each DataFrame on a monthly (M) basis, and apply exponential smoothing using $\alpha = 0.3$ (refer to lab). Finally, provide a 3x1 subplot of the monthly smoothed temperature for Calgary, Buenos Aires, and Edmonton (in that order) over time. Use 1-2 sentences to comment on (1) any patterns that the plots exhibit and (2) which two cities would you expect to be closest to each other and why.

(c) Provide a 3x1 subplot of the autocorrelation of the monthly smoothed temperature for Calgary, Buenos Aires, and Edmonton in that order (refer to lab). Use 1-2 sentences to comment on whether the autocorrelation plots confirm the patterns you observed in Q4(b) and if any other insights can be drawn from the autocorrelation plot (hint: consider what happens to the autocorrelation as the lag increases).

(d) Provide a 3x1 subplot of the cross-correlation between the monthly smoothed temperature of Calgary-Buenos Aires, Calgary-Edmonton, and Edmonton-Buenos Aires in that order (refer to lab). Use these plots to deduce the following: when it is summer in Edmonton, what season is it in Buenos Aires and Calgary?