**DSC 275/475: Time Series Analysis and Forecasting (Fall 2024)**

**HW #1**

**(Total points: 45)**

Note:

* Please make sure that your time-series graphs are properly labeled on both axes.
* For each of the questions, please submit a written response alongwith the results from running the code included. In addition, please attach the code. You are welcome to use Jupyter notebook (Python), Rmarkdown (R), or LiveScript (MATLAB) to combine the code and written text in one file. In such a case, please attach the original file and its PDF version.

1. *(8 pts)* The Bureau of Transportation Statistics (BTS) conducted a study to evaluate the impact of Sept 11 attacks (9/11) on U.S. air transportation. The purpose of this study is to provide a greater understanding of the passenger travel behavior patterns of persons travelling by air before and after the event. In order to assess the impact of September 11, BTS took the following approach: Using data before September 11, it forecasted future data (under the assumption of no terrorist attack). Then, BTS compared the forecasted series with the actual data to assess the impact of the event.

The data is available in the file: *BTS\_Air\_Rail\_Vehicle\_Miles.csv*.

1. Is the goal of this study descriptive or predictive ? (2 pts)
2. Create a time series plot of the *Air* data, i.e. a plot yt versus *t, where t=1,2,3 …* What would t=1, 2, 3 refer to in the time series? Which time period does t=1 refer to ? (3 pts)
3. What are the values for y1, y­2 and y3 in the time series? (3 pts)
4. *(3 pts)* In addition to air travel data, additional time series are also provided in the same data file in Question 1.

Which of these components appear in the *Vehicle* time series: i) Seasonality; ii) Trend; iii) Noise.

1. *(3 pts)* Forecasting Shampoo Sales: The file *ShampooSales.csv* contains data on the monthly sales of a certain shampoo over a 3 year period.

Which of the components (trend, seasonality, noise) are present in this series?

1. *(15 pts)* The file, *AirSensorData.xlsx*, contains the air quality data from multiple sensors. There are a total of 13 columns that contain the sensor data with the sensor name indicated at the top of the column. Answer the following questions:
2. *(4 pts)* Create a time-series plot of the first 10 sensors (i.e. 10 columns) by plotting the first 200 points in the data.
3. *(6 pts)* From the plot, you should observe there are erroneous values of “-200” at random. For each of the 13 channels, determine how many such erroneous values are present in the *entire dataset*? Report your results via a table with 2 columns where the first column contains the sensor name and the second column is the count of the erroneous values for that sensor.
4. *(5 pts)* For the time period from 3/11/2014 0:00 hrs to 3/31/2014 23:00:00 hrs (both included), plot the time series for the sensor column C6H6(GT) consisting of *daily average* values. Report the *maximum and minimum daily average value*.
5. *(6 pts)* The file, *Beverages\_Shipment\_2020.csv*, contains the US beverage product shipments data.
6. Is there seasonality in this time series? (2 pts)
7. Find the sample autocorrelation function for this data set. (2 pts)

(For Python, you can use the “plot\_acf” function in “statmodels” module. Plot at least 25 lags)

1. From the autocorrelation plot in (b), what is the seasonal period (state the units)? Explain your answer. (2 pts)
2. *(10 pts)* Data on US coal production is given in *Coal\_Production\_US\_2020.csv*.
3. Plot the coal production data and the sample autocorrelation function. (2 pts)
4. Is the time series stationary or non-stationary? (2 pts)
5. Plot the first difference of the time series and the sample autocorrelation function of the first difference. (4 pts)
6. What impact has differencing had on the time series? Comment with respect to presence or absence of stationarity (2 pts)