**Time Series and Factor Analysis Project** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Part of the open source Ozone Level Detection Data Set is available on Canvas in the file “Assignment Data.csv”. Some information explaining the different variables in the dataset is available in the file “Assignment Data Details.txt” and also here: http://archive.ics.uci.edu/ml/datasets/Ozone+Level+Detection

You should open this dataset and carry out an exploratory analysis producing a descriptive summary of these data. Your exploratory analysis should focus in on the key variables you wish to predict the future state of, which are listed here:

WSR\_PK: continuous. peek wind speed -- resultant (meaning average of wind vector)   
T\_PK: continuous. Peak T   
T\_AV: continuous. Average T   
T85: continuous. T at 850 hpa level (or about 1500 m height)   
RH85: continuous. Relative Humidity at 850 hpa   
HT85: continuous. Geopotential height at 850 hpa, it is about the same as height at low altitude   
T70: continuous. T at 700 hpa level (roughly 3100 m height)   
KI: continuous. K-Index

TT: continuous. T-Totals   
SLP: continuous. Sea level pressure   
SLP\_: continuous. SLP change from previous day

Your exploratory analysis should detail any missing data and how you dealt with it. You should also discuss the treatment of outliers.

You should begin time series modelling and generate forecasts of each of the variables listed above. This should be done systematically for each variable and this should be documented in your report. Your analysis should consider simple equations for modelling, as well as more complicated models including ARIMA. You should comment on the trends and seasons that you believe are present in these datasets. At a minimum you should compare appropriate exponential smoothing models with a collection of ARIMA models.

You should propose what you believe is the appropriate model and justify this for each variable.

You should include:

* An explanation on what the model captures.
* Fit and comment on the observed with the predicted values for both models in the same plot.
* Plot and comment on the residuals of the fitted data for both models in the same plot.
* Determine and comment on the fit and measure statistics.
* Test and comment on the autocorrelation and partial autocorrelation of the residuals (using the plots and Ljung-Box Q statistic\*) up to an appropriate lag which you should determine following appropriate research.
* Test and comment on the normality of the residuals.
* State the best model.

Once you have decided on what you believe the best model, state and comment on the model parameters.

You should validate this model by an appropriate cross validation or holdout type method and forecasting

Part 2 A Factor Analysis

A dataset is provided in the file “Factor Analysis.xlsx”.

It is believed that a two factor model will explain the different variables in this dataset. You are tasked with applying an appropriate factor analysis to determine a two factor model that captures as much of the variability in the datset as possible. In doing this you should consider a variety of different models. You should use select the appropriate model and state clearly why you feel this is the appropriate model.

All of your work should be detailed in your report including appropriate figures and tables. Your codes should be included with your work. Additionally you should record a presentation of your work. This can be done with the free software screencastomatic available online. You should submit your work as an email through canvas.

You should also include a declaration as part of your submission that all work contained in the submission was your own except for others work which is clearly referenced. You will be asked to attend an interview using a conferenceing tool. In this interview, it is expected that you will be able to explain your own work. Marks will not be awarded for work that the student cannot explain.

**Relevant Papers:**

Forecasting skewed biased stochastic ozone days: analyses, solutions and beyond, Knowledge and Information Systems, Vol. 14, No. 3, 2008.   
Discusses details about the dataset, its use as well as various experiments (both cross-validation and streaming) using many state-of-the-art methods.   
A shorter version of the paper (does not contain some detailed experiments as the journal paper above) is in:   
Forecasting Skewed Biased Stochastic Ozone Days: Analyses and Solutions. ICDM 2006: 753-764