**[Assignment 3](https://blackboard.umkc.edu/webapps/assignment/uploadAssignment?content_id=_3149217_1&course_id=_167602_1&group_id=&mode=view)**

Be sure to mark each problem # properly and your student ID (last 4 digits) shows up - no names, remember to number your pages. The submitted file should be PDF (preferably typed). Name your file with the last four digits of your student id followed by '-A3'. For example, if the last four digits of your ID are 1234, then the file name should be the following:     1234-A3.pdf

3\_1 [30 points].  This assignment extends from Assignment-2. Q-1, which is reproduced below. Your assignment is to extend Part-b and use SVM, and provide a comparison with a discussion (note - need to do it only for 2.1 part-b).

"Q-2\_1.You’re to use the KC Weather Data (“kc\_weather\_srt.csv”, available here: [kc\_weather\_srt.csv](https://blackboard.umkc.edu/courses/1/UMKC-COMP-SCI5565-0001-45609-FS2017/content/_3027186_1/embedded/kc_weather_srt.csv) ). The data has categorized the weather for each day into three categories  (“Events”: Rain, Rain\_Thunderstorm, Snow) over the three years 2014, 2015, and 2016. You’ll note that not all dates are listed because it’s a filtered subset where other categories or no events are deleted to have a more manageable subset. The entire dataset has 366 entries. The column labels indicate the units as well such as Temp.F means temperature in Fahrenheit, Visibility.mi means Visibility in miles, etc.

You’re to do two level of analysis

1. Consider first the subset that consists only of Rain and Snow. There are 226 entries with these two categories.
   1. Apply logistic regression, LDA, QDA, and knn on this dataset to determine the accuracy, precision, and recall of these models. You’re to use randomly 180 days for the training set (approximately 80% of 226) and the rest for the test data set. Conduct your study over 100 replications, and summary the result of your analysis with your conclusion which models you’ll recommend to use based on the metrics: accuracy, precision and recall.
   2. Discuss and analyze in a systematic way you would consider eliminating some of the predictors and see if your accuracy, precision and recall improves.
2. Consider next the entire dataset consisting of 366 entries. Now logistics regression cannot be applied, but you can apply the rest of them. Repeat the above studies in i) and ii) with LDA, QDA, and knn on the entire data set (using 290 of them in a training set). Do not forget randomization and 100 replications for this study."

3\_2. [45 points] Consider the time series on Milk production data [milk-production(1).csv](https://blackboard.umkc.edu/bbcswebdav/pid-3149217-dt-content-rid-19333983_1/xid-19333983_1)

  it shows cow milk production per pound from 1962 to 1975.

 a. Try at least three different values for window size with simple moving average (SMA) for forecasting

 b. Apply exponential moving average using HoltWinters for forecasting

 c. For the above, discuss how the forecasting differs in terms of MAD and MFE and why one approach or the other is better.