**ISE 365/465 – Applied Data Mining Homework 1 (Due Feb. 25 in class)**

1. Read the file “EG ViolatorsFinal” into Modeler using the appropriate source node.
2. Run the Automatic Data Preparation (ADP) node for the data using the default values of the ADP node except:
   1. Change the fixed reference date to calculate months duration from to the earliest date in the data
   2. Change the fixed reference time to calculate time duration from to midnight (0:00:00)
   3. Extract from dates the day of month and month
   4. Uncheck the “Continuous Fields: replace missing values with mean” and “Nominal Fields: replace missing values with mode” boxes for Inputs
   5. Uncheck all boxes in the “Transform Continuous Field” section
   6. Change the “Maximum Number of Values for Ordinal Fields” to 50.
3. Read the “NJWeather” file into Modeler using the appropriate source node.
   1. Remove non-month rows
   2. Convert all fields except for month into numeric fields (This will require CLEM expressions in a filler or derive node(s).)
4. Merge the EGViolators and NJWeather data together using month. Make sure to convert the months to their name before doing step 5.
5. Run a data audit node and impute any NULL values in the Time\_hours field using the algorithm method. Use sample size of 100% for this field only. (Hint: you must create a supernode from the Generate menu in the data audit node to do this. This will kick off a model run that will run around 30 seconds (depending on your computer speed) to generate the supernode.)
6. After merging the files in your stream and imputing the data in the Time\_hours field, perform each of the following actions **IN BOTH IBM SPSS MODELER and SAS ENTERPRISE MINER (You will have to create a SAS file in Modeler with your final data set to use in SAS EM):**
   1. Create a bar graph showing mean MPH over the speed limit (overlimit field)
   2. Create a table showing the same thing as in the bar graph (mean over limit by month)
   3. Add to the bar chart in 6a a color overlay for race showing what fraction of each bar is made up by each race.
   4. Create a table showing the same information as the bar chart in 6c (overlimit by month and race). Table must be sorted by Month and race (both in ascending order)
   5. Create a line chart with Time\_hour as the x axis and Mean Overlimit as the y axis.
7. Your homework submission should contain the following:
   1. **Your Modeler stream file and EM xml file uploaded to Course Site by 8:00 AM on February 25.** (35 points for general correctness)
   2. **Hand in on paper at class on February 25:**

(Note: The tables should only contain the number of rows specified.):

* + 1. A screen dump of the first 10 rows of data in a table immediately after step 3b above for the “NJWeather” data in numeric format. (10 points)
    2. A screen dump of the first 10 rows of data in a table immediately after you merged the two data sources in step 4 above. (10 points)
    3. A screen dump of the first 10 rows of the tables in 6b and 6d **for Modeler and Enterprise Miner**. (10 points)
    4. The bar charts from 6a and 6c **from both Modeler and Enterprise Miner.** (10 points)
    5. The line chart from 6e **from both Modeler and Enterprise Miner.** (5 points)
    6. Answers to the following questions:
       1. If we wanted to impute the null values for Time\_hours in another place in the stream, which node in the current stream could have been used? (5points)
       2. If we would have left the z-score transformation checked in the Automatic Data Prep node, what would have been stored in the continuous variables? Explain this in words (do not answer “the Z-score”, but rather what this means. This should only take a sentence or two at most). (5 points)
       3. When merging on the “Month” field, do you think it is better to merge using the integer value of month or the string value? Explain why. (Hint, think of possible data issues associated with integers versus strings and how easy it is to match field values exactly to allow a correct merge). (5 points)
       4. In the line chart in 6e, what hour of the day has the most speeding on average? Give a practical reason as to why this time may have high speeding. (5 points)