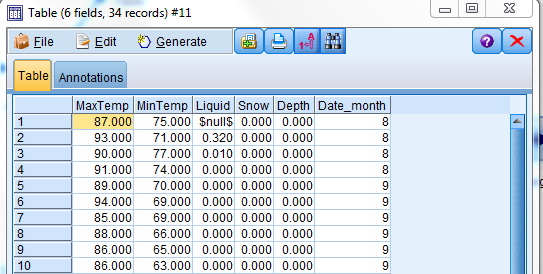
**ISE 365/465 – Applied Data Mining Homework 1**

***Name: Bolun Xu***

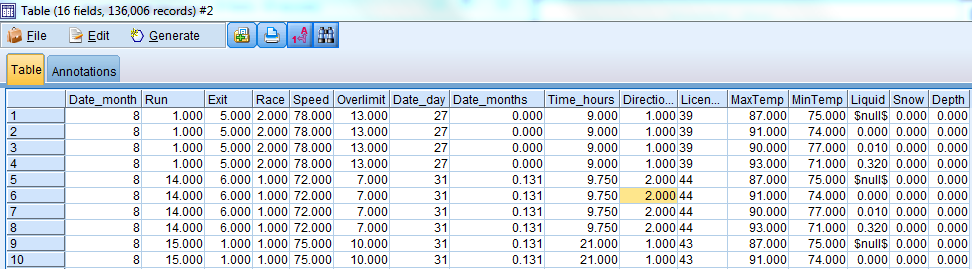
1. Your homework submission should contain the following:
   1. **Your Modeler stream file and EM xml file uploaded to Course Site by 8:00 AM on February 25.** (35 points for general correctness)
   2. **Hand in on paper at class on February 25:**

(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)

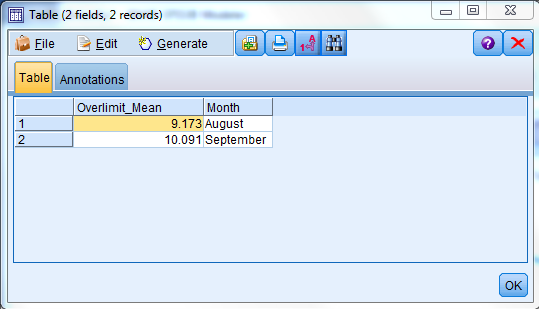


* + 1. A screen dump of the first 10 rows of data in a table immediately after you merged the two data sources in step 4 above. (10 points)

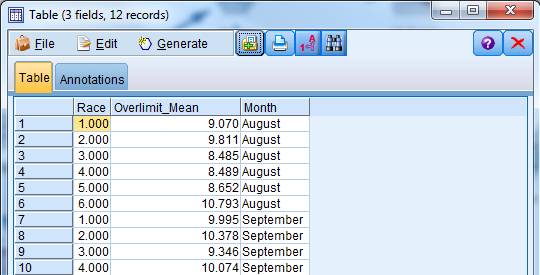


* + 1. A screen dump of the first 10 rows of the tables in 6b and 6d **for Modeler and Enterprise Miner**. (10 points)

6b：

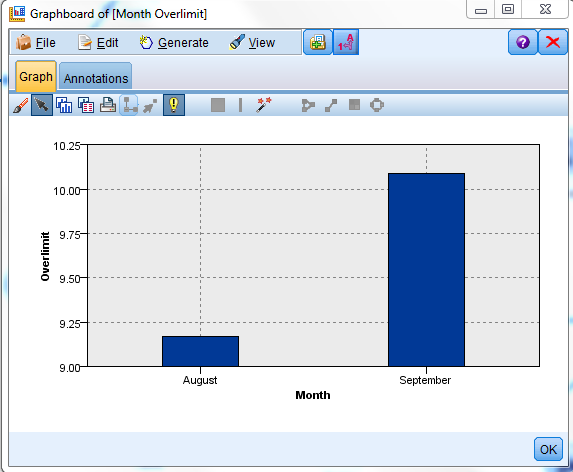


6d：

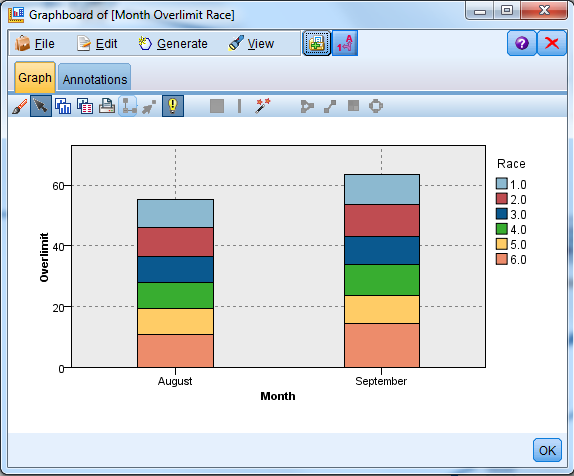


* + 1. The bar charts from 6a and 6c **from both Modeler and Enterprise Miner.** (10 points)

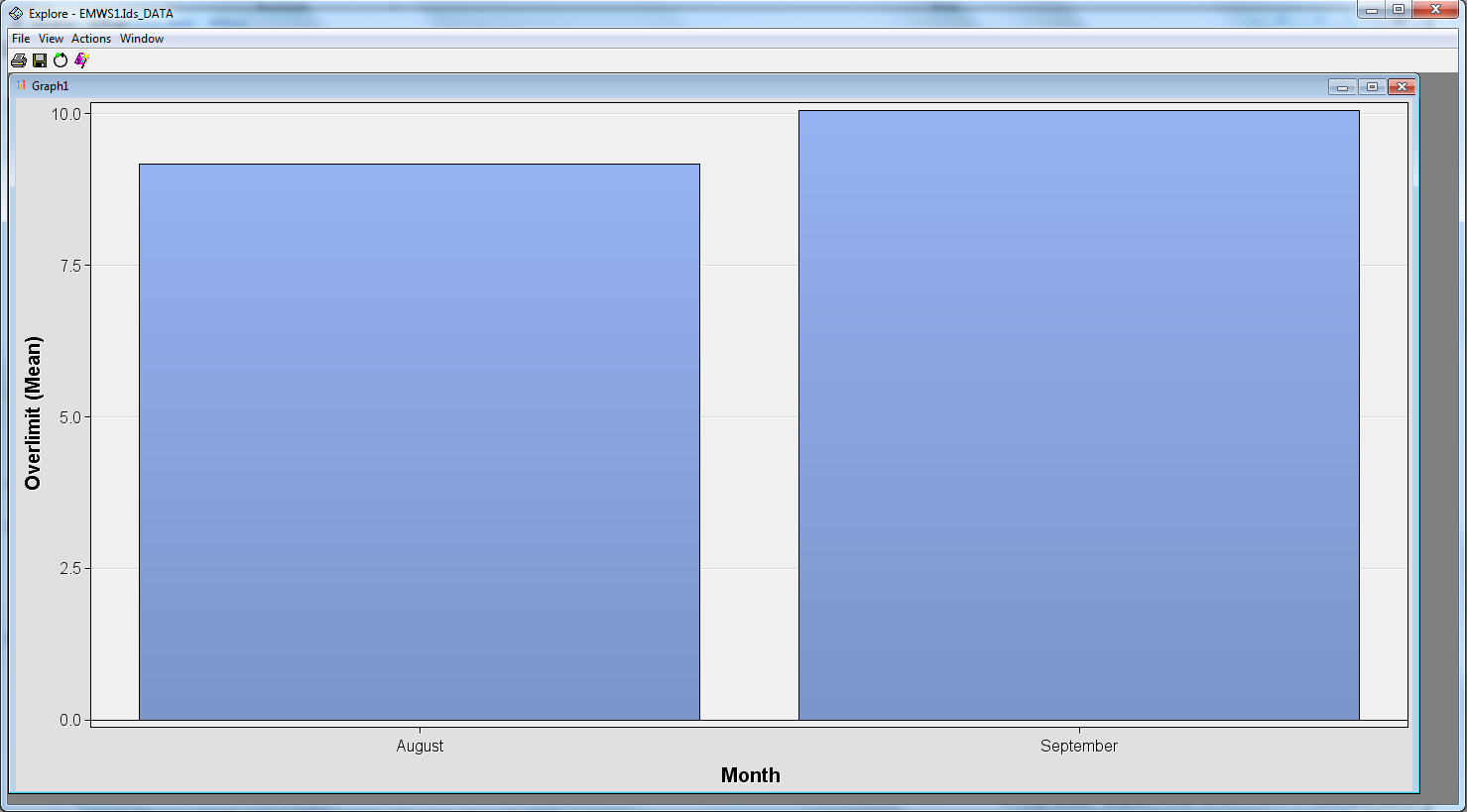
6a Modeler：



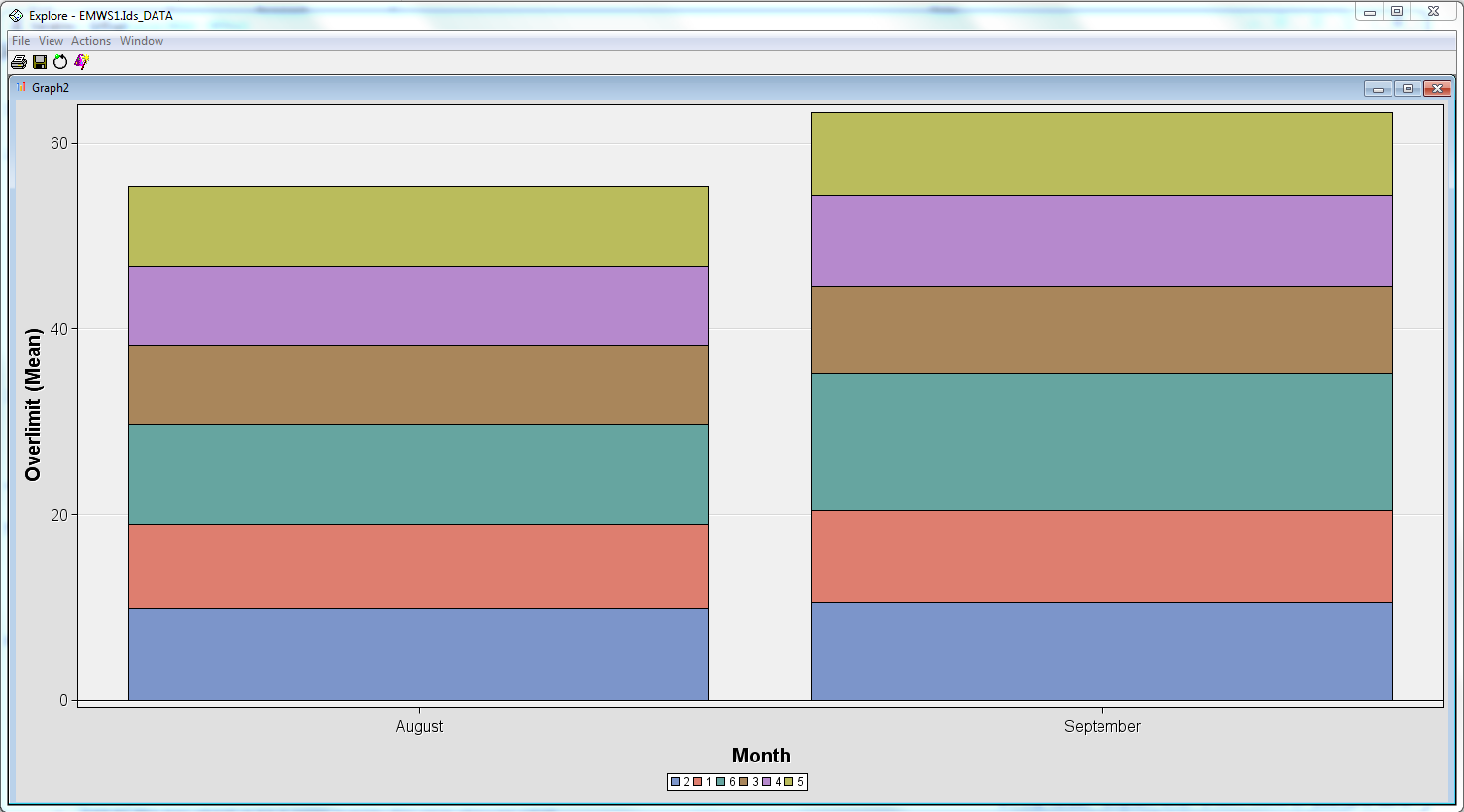
6c Modeler：



6a sas：

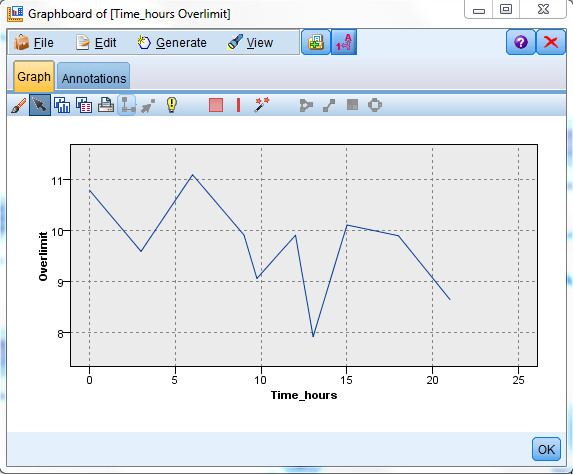


6c sas：

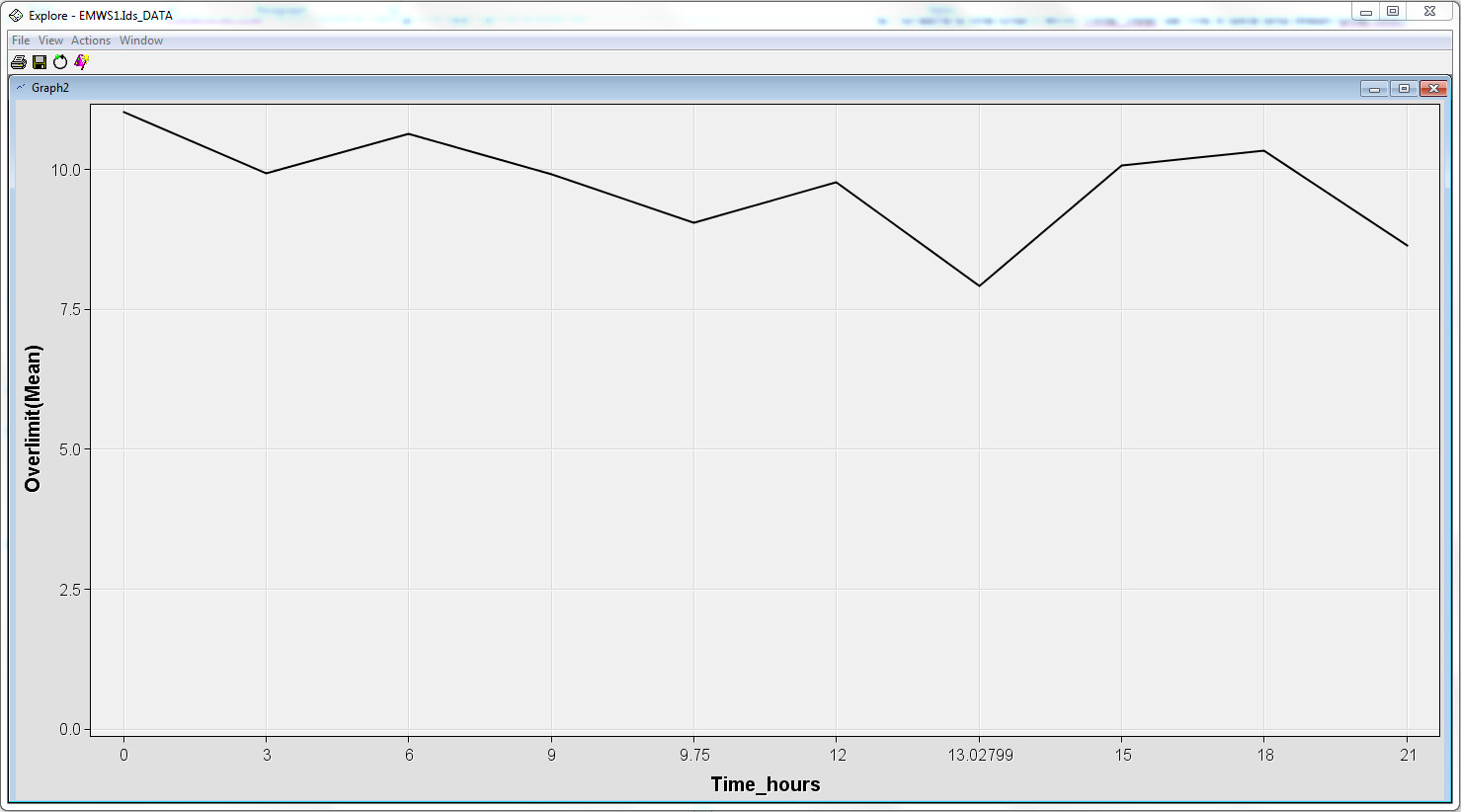


* + 1. The line chart from 6e **from both Modeler and Enterprise Miner.** (5 points)

6e Modeler：



6e sas：



* + 1. 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)

A: We use a data audit node, Analyze the quality of data, impute null values and generate a supernode.

* + - 1. 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)

A: z = (x-μ)/σ, μ is the mean of the population,

σ is the standard deviation of the population.

* + - 1. 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)

A: Integer is better. Numbers always have a better reliability than string, because any difference between spelling of Months (e.g. Capital/small letters) makes an additional record when we are merging. This situation makes errors. Otherwise, integer has less chance to make spelling errors.

* + - 1. 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)

A: Around 6:00 am. People tend to rush to work and they might not have a clear mind shortly after getting up.