**Case Problem**

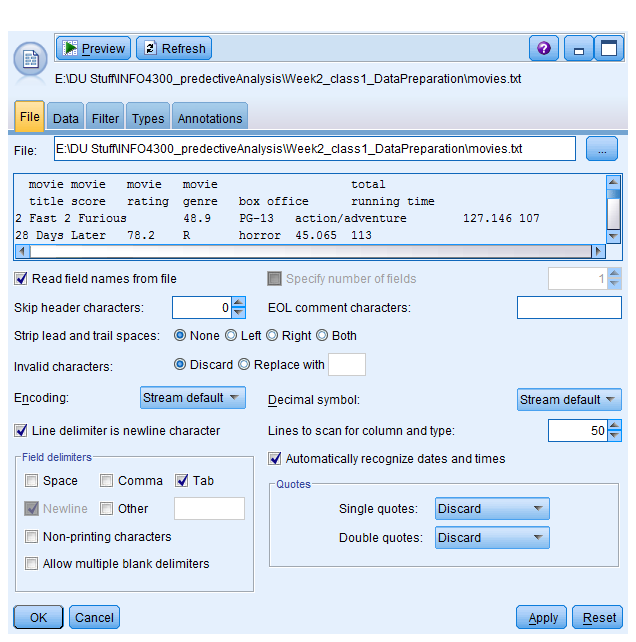
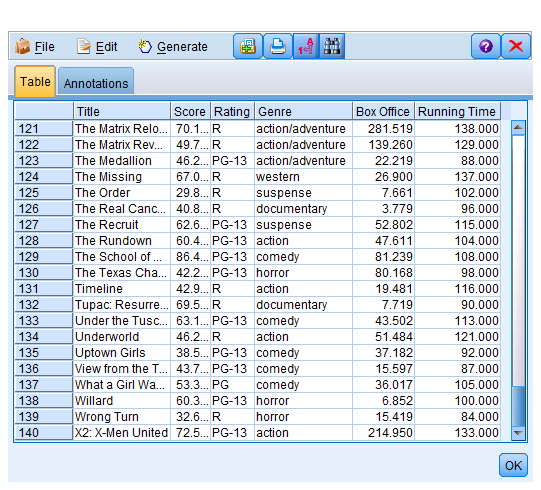
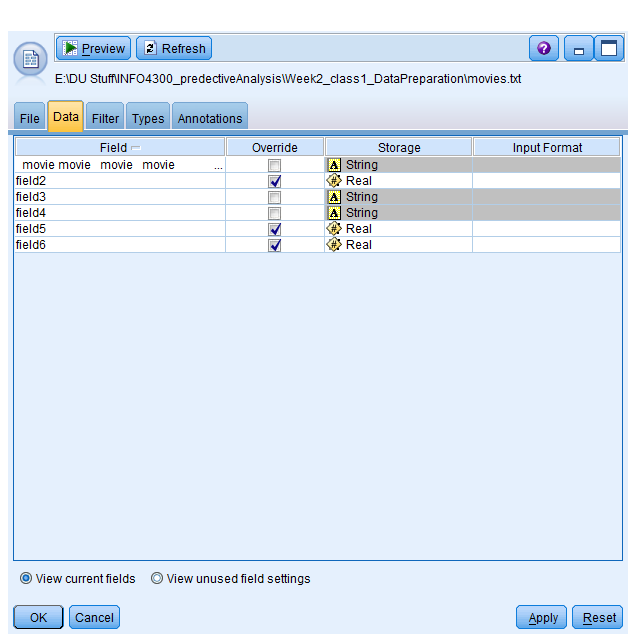
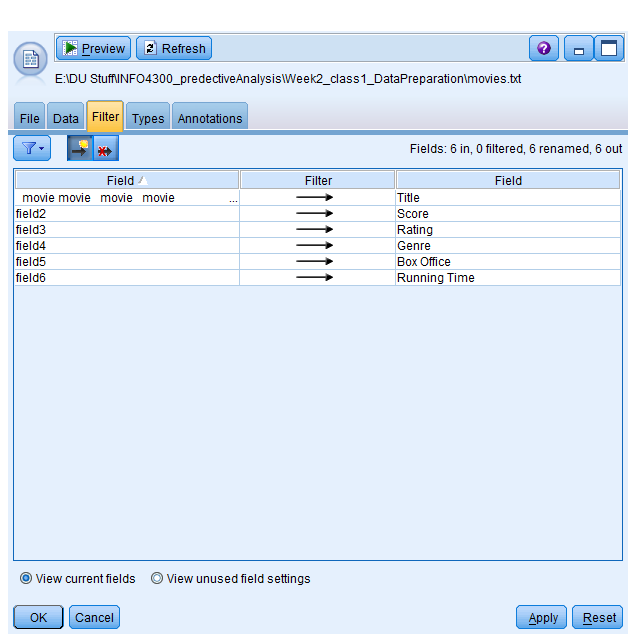
**Predictive Analytics**

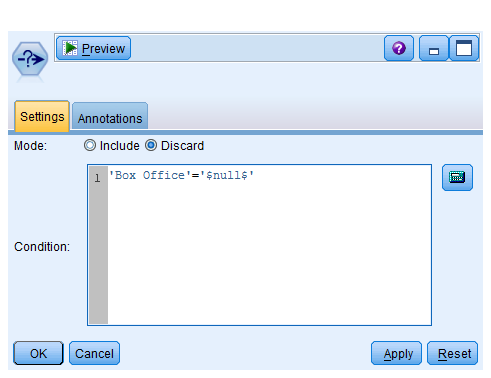
**Module 1 – Lesson 3**

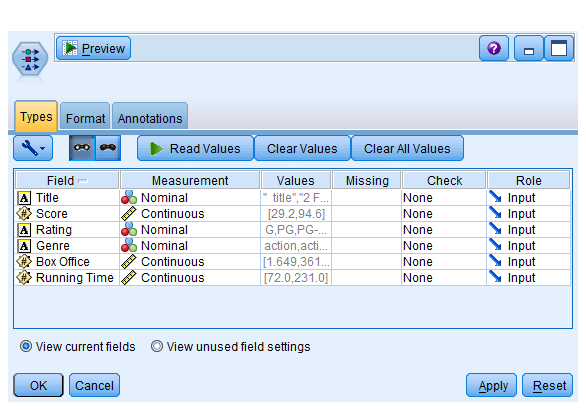
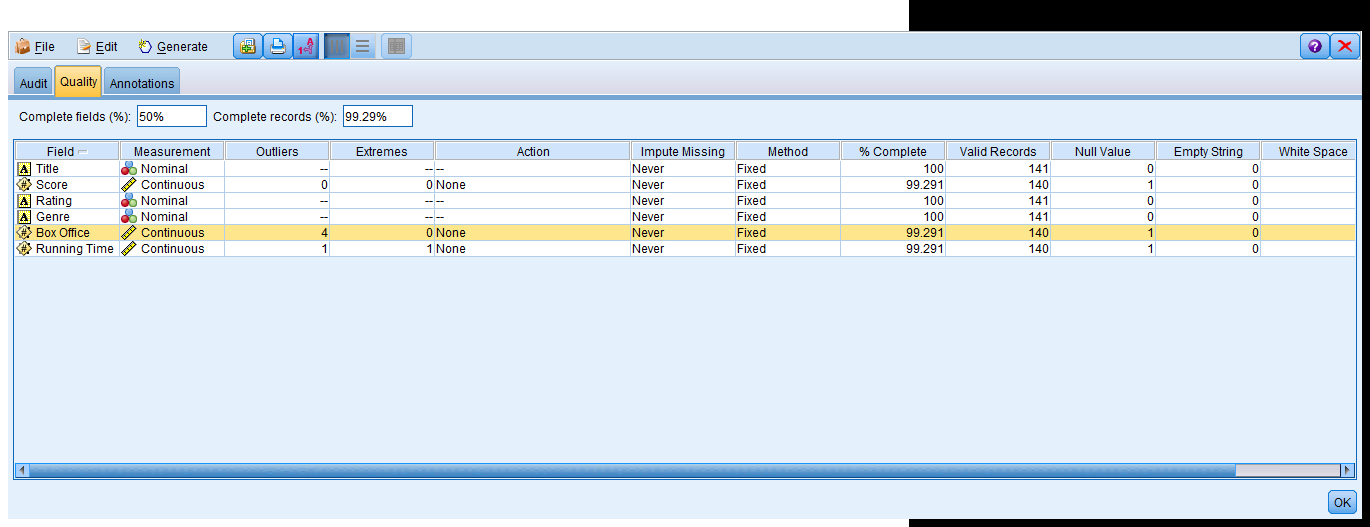
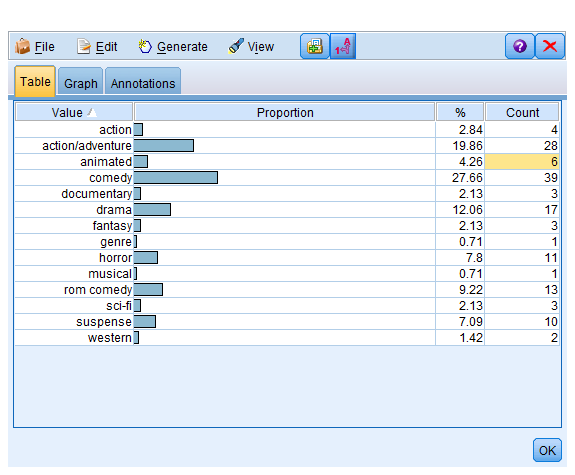
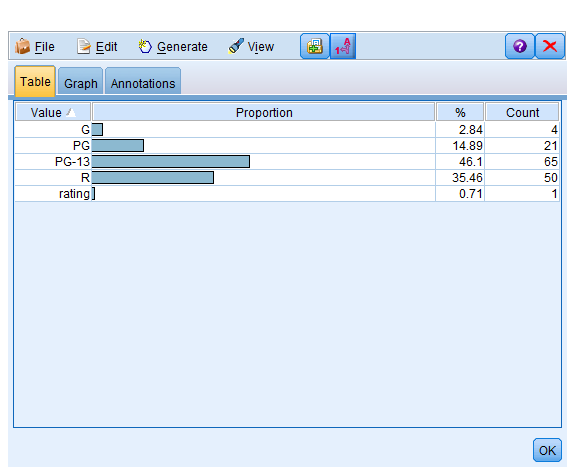
**Professor: Lee**

Rajiv Kumar

**Dr. K enjoys tracking the Rotten Tomatoes scores of movies and he has collected a dataset from the RossmanChance.com website in movies.txt. In this file, the “target” variable that we are examining is the Rotten Tomatoes Score.**

* + **View this file first to see what it looks like. What is the ‘delimiter’ between each variable?**
    - It is a tab delimiter file
  + **Open the text file using the Source Node: Var. Check the box for ‘Line delimiter is newline character.’ Check the box for Field delimiter: Tab.**
    - ****
  + **The columns should be “Title”,”Score”,”Rating”,”Genre”,”Box Office”,”Running Time.” Make sure you have 140 total cases (rows/records)**
  + ****
    - **Use the Data Tab to indicate the correct types (Check Override to change the default. Use either String or Real)**
    - ****
    - **Use the Filter Tab to change the column/variable names.**
    - ****
  + **Add a Select Node to select only rows that don’t have the header row ‘mess’ in them. What would be a good way to test if it is the header row??**

****

* + **Add a Types Node to set up the correct Types and preview your data.   
    **
  + **Perform a Data Audit for your Variables.** 
    - **Assess your Categorical & Continuous Variables**
    - ****
    - ****
    - ****
    - **Create Multiple Variable Summaries of your data. State your conclusions about which variables you think will provide predictive ability for the rotten tomatoes score.**

**Score**

**Statistics**

**Count 140**

**Mean 55.983**

**Min 29.200**

**Max 94.600**

**Range 65.400**

**Variance 206.610**

**Standard Deviation 14.374**

**Standard Error of Mean 1.215**

**Box Office**

**Statistics**

**Count 140**

**Mean 61.051**

**Min 1.649**

**Max 361.119**

**Range 359.470**

**Variance 3966.102**

**Standard Deviation 62.977**

**Standard Error of Mean 5.323**

**Running Time**

**Statistics**

**Count 140**

**Mean 107.557**

**Min 72.000**

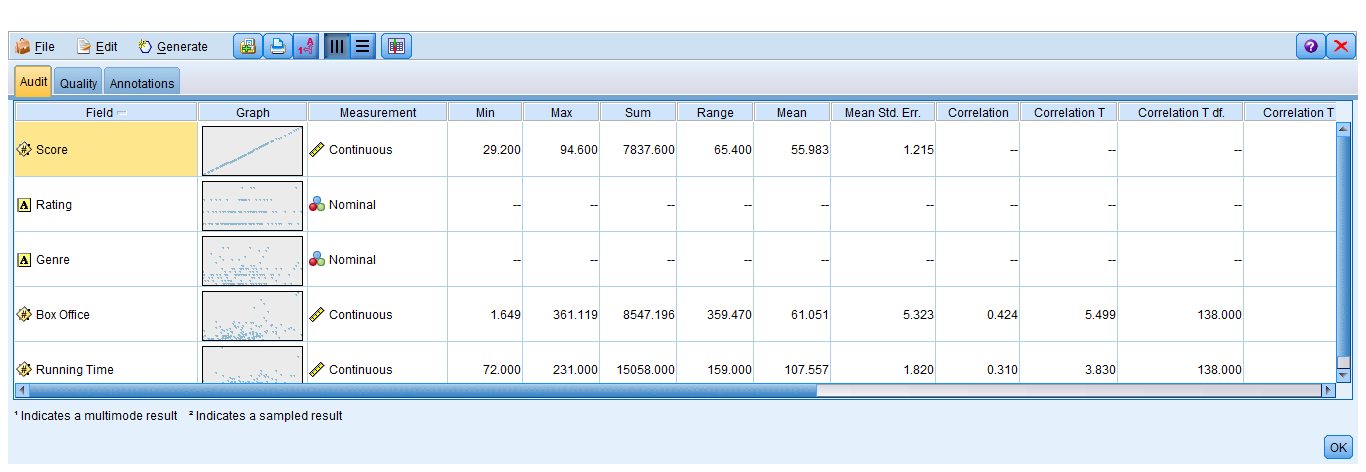
**Max 231.000**

**Range 159.000**

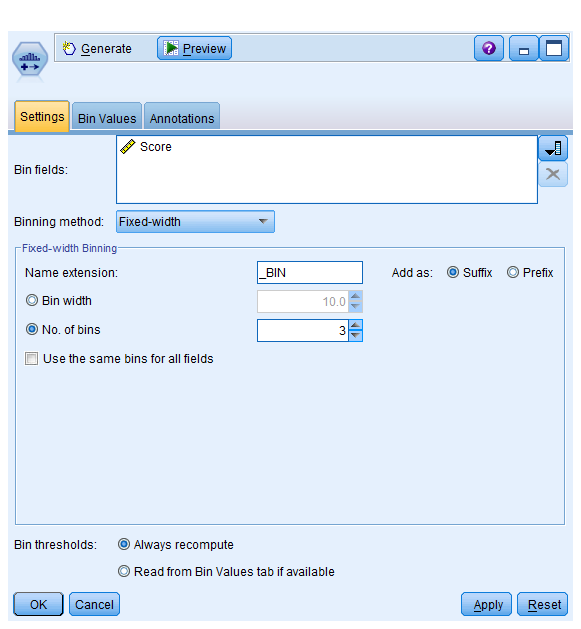
**Variance 463.745**

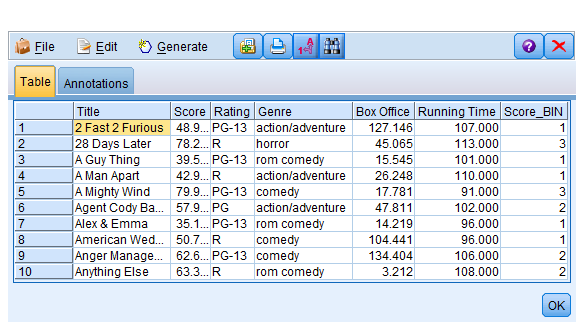
**Standard Deviation 21.535**

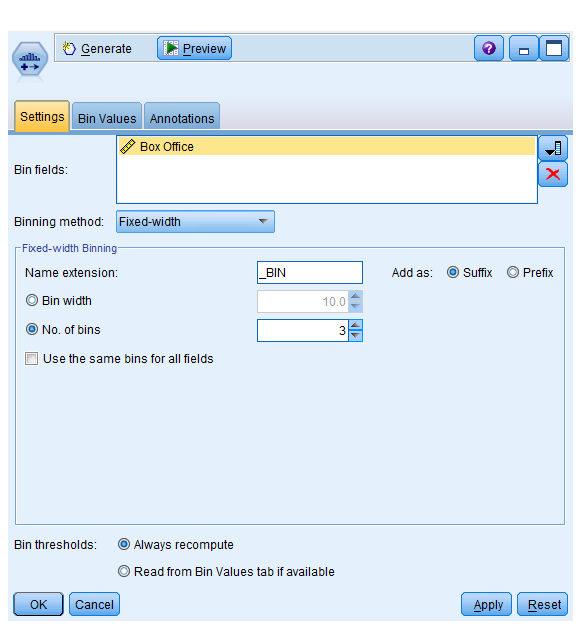
**Standard Error of Mean 1.820**

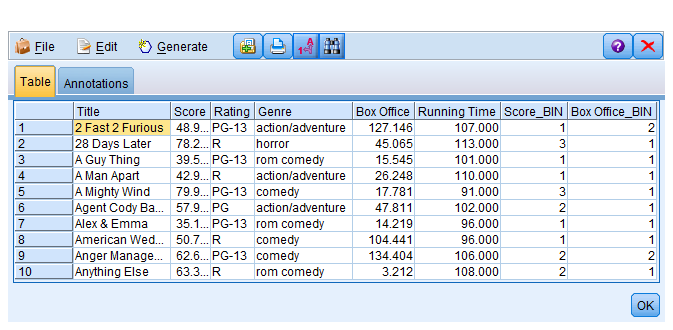
* + - ****

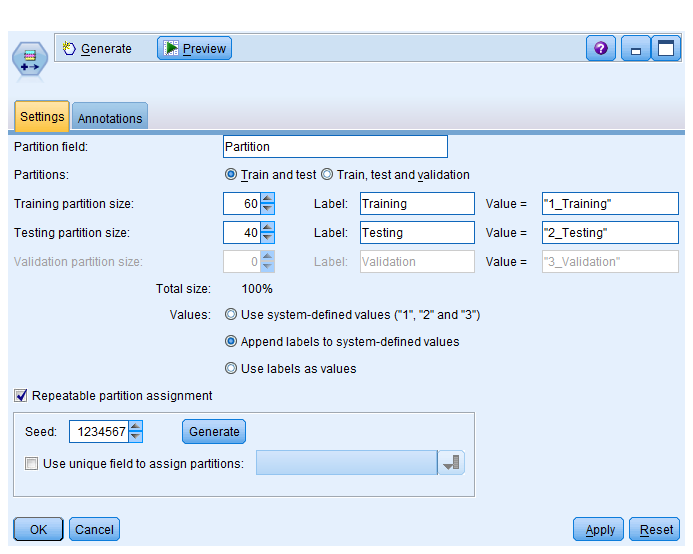
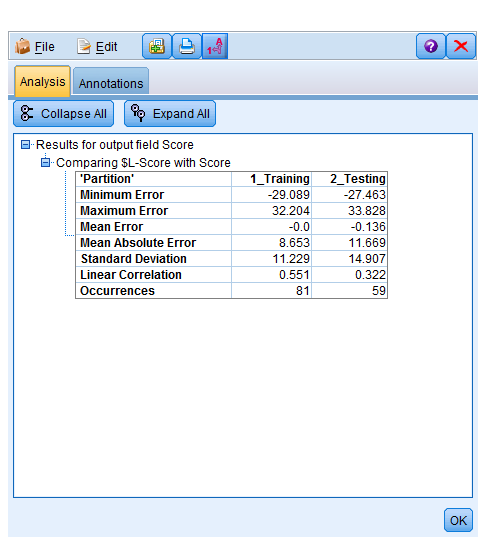
Correlation coefficient for Box office and running time are relatively higher ,.42 and .32 respectively. so these two variables can be part of our input features for the model. We will have to further analyze other variable to see how they perform in the model.

* + **Derive two variables from your audit. (One should be to derive a new column “Score Category” that divides the Rotten Tomatoes score into 3 values: High, Medium, and Low. Use our own scheme for determining the cutoffs for this variable. You may want to use a two-step procedure (first define low and high, then break up high into medium and high due to set up of the IF THEN ELSE condition). Make sure you document it in your data audit.)**
  + ****

****

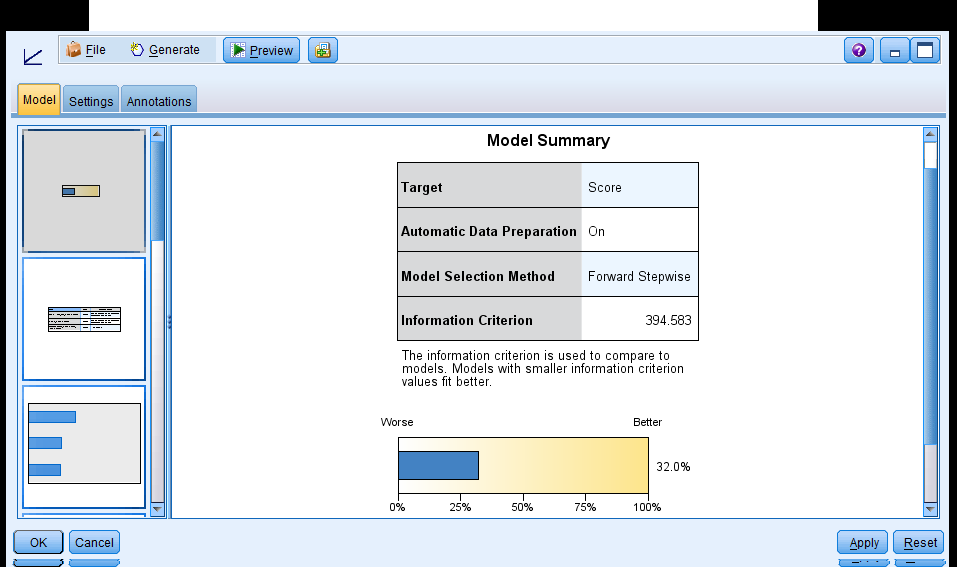
****

****

* + **Create a new variable using the Partition node to divide your data into a training and test data set.**
  + ****
  + **Create a regression model and use the analysis node to examine your results. How well did the model perform? Do you see any evidence of overfitting from the Analysis node?**
  + ****

**Model Accuracy:**

I was able to get only 27.7% model accuracy after trying several options. I finally chose Genre and Box office as my input features for the model. I don’t see any evidence of over-fitting as model perform equally well on test and training data set.

****

**Submit your completed Word file and your stream file on Canvas. In the comments, put the names of the folks who worked on this assignment.**