Basic Data Analysis Using JMP in Windows

Table of Contents:

I.   Getting Started with JMP

The JMP software can be launched by clicking on the Start button located on the bottom left corner of the screen. Next, move the arrow onto Programs and click on **JMP 12**. Your initial view of JMP will be a menu bar, a tool bar, a Tip of the Day window, and the JMP starter window.
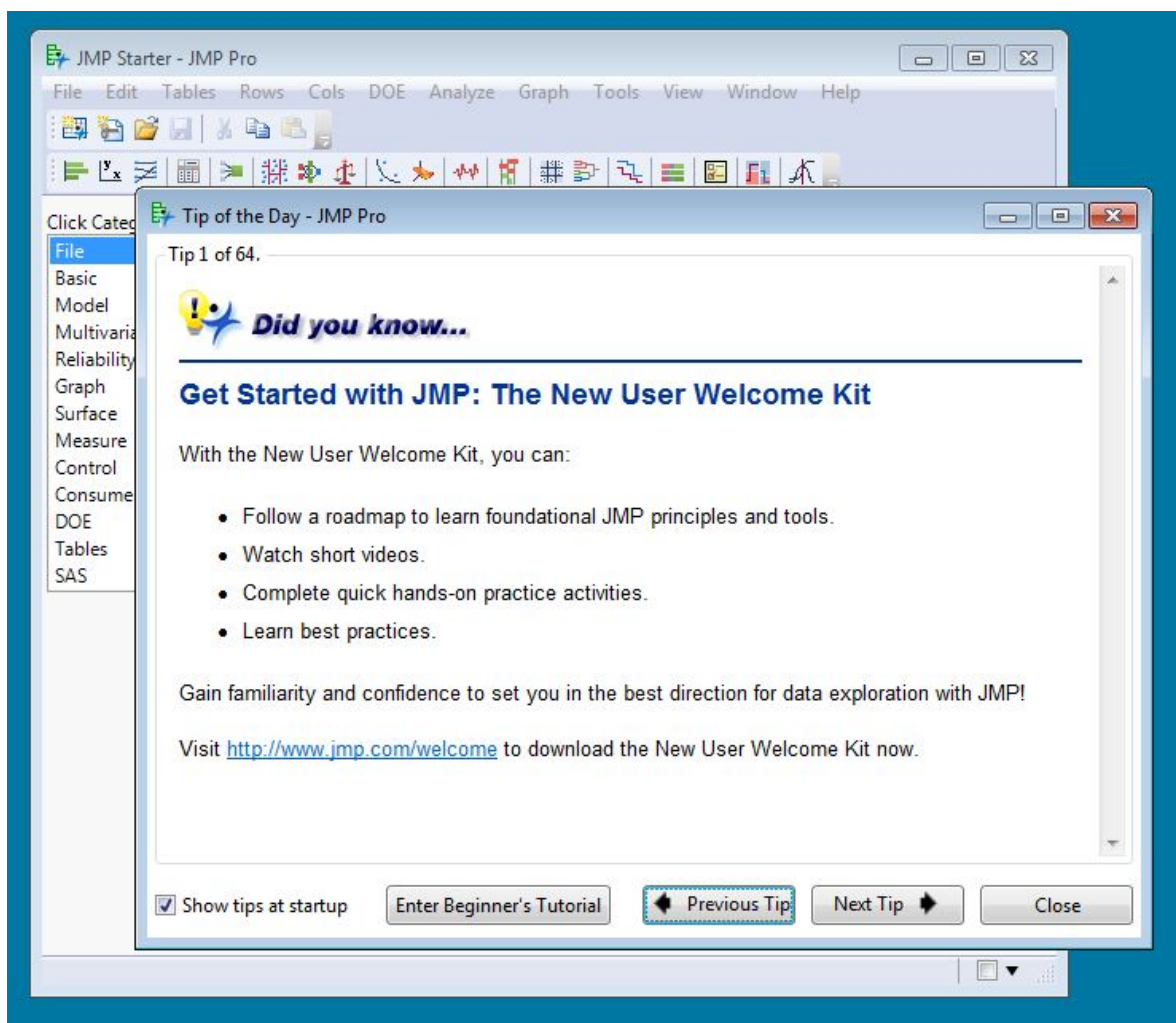
Close the "Tip of the Day" window and now you will have the JMP starter window where you can create or open a Data Table and Script file.
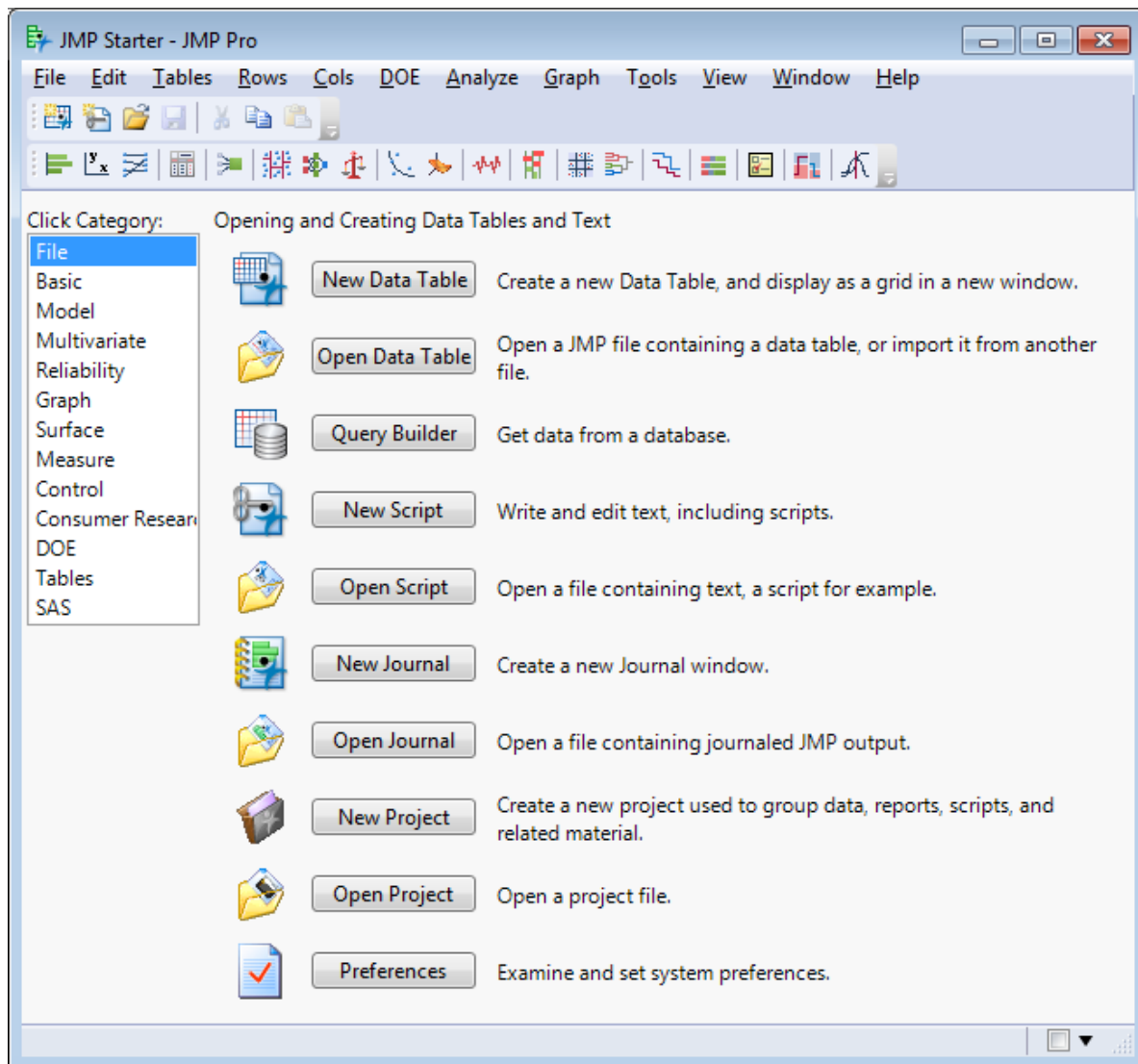
II Entering Data in JMP

Selecting **File > New** > **Data Table** (or clicking the New Data Table button on the JMP Starter window) creates and displays a data table with an empty data grid, as shown below. The count of table rows and columns appears in the corresponding panels to the left of the data grid. In the data grid, a row number identifies each row, and each column has a column name. Rows and columns are referred to respectively as *observations* and *variables*.
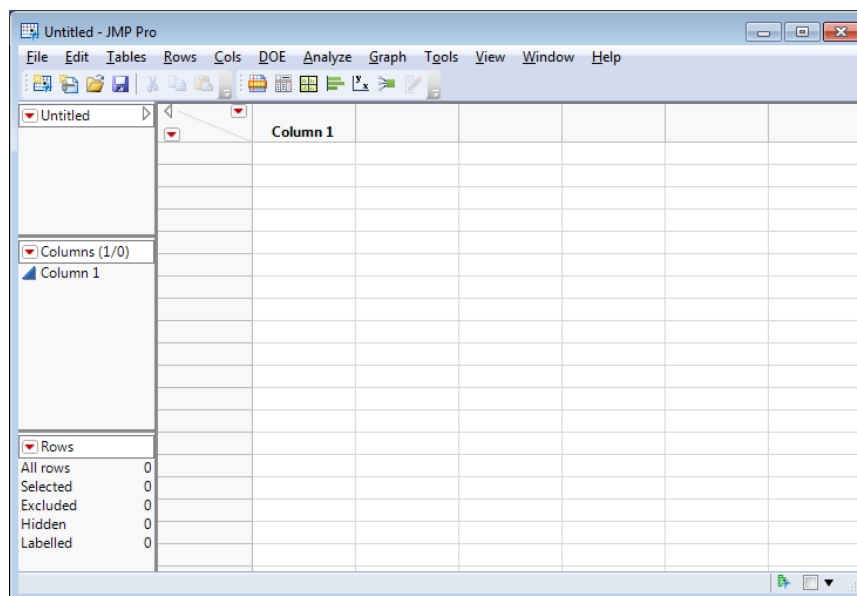
 Double clicking on the heading of the first column ("Column 1") will produce the following screen where you can specify the variable name and the corresponding attributes.
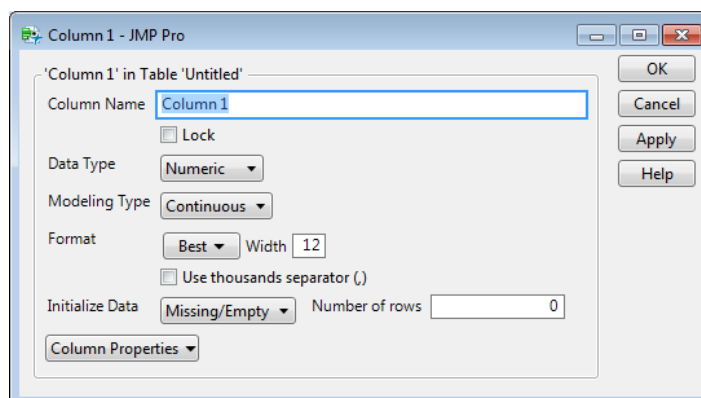
The following specifications can be made for each variable:

Column Name: The name of the variable specified. A default name will be assigned to a variable unless a name is specified. The column name should not contain any punctuation. An underscore (_) is permitted.

Data type: Data type can be numeric, character, row state, or expression. If you choose "row state", which stores information about whether the rows are excluded, hidden, labeled, colored, or marked, then this column has its own data type and it does not have a modeling type because its values are not used in analyses. Most users will work with numeric or character data type.

Modeling type: JMP uses three modeling types to determine how to analyze the column's values:
• **Continuous** (▲) Values are numeric measurements.
• **Ordinal** (▮) Values are ordered categories, which can have either numeric or character values.
• **Nominal** (▮) Values are numeric or character classifications.

Format: The default will usually be all you need. If you wish you can chose a more specific format of the variable; for example, "date", "time", "currency" etc.

Initial data values: You can specify initial values for this variable. By default it is "Missing/Empty".

Once you have specified the variables, click OK to get back to the datasheet. Now, you can enter your data into the corresponding cells.

III. Saving JMP Data File

1. Before exiting JMP, save your data file by clicking **File** on the menu bar and then **Save As** (or **Save**)
2. In the Save Data As window, navigate to the proper folder and type in a **File name**.
3. Click **Save** and file will be saved by default with the extension **".jmp"**. You can also save your data file in the other available format listed under "**type**" in the save menu.

IV. Opening an Existing Data File

Selecting **File > Open** (or clicking the Open Data Table button on the JMP Starter window) presents a file selection window with a list of existing tables. For JMP data files, select the file and click **Open**. For files from other software, i.e SPSS data sets with the .sav extension, change the **Data Files** pull down to match the file type that you have.
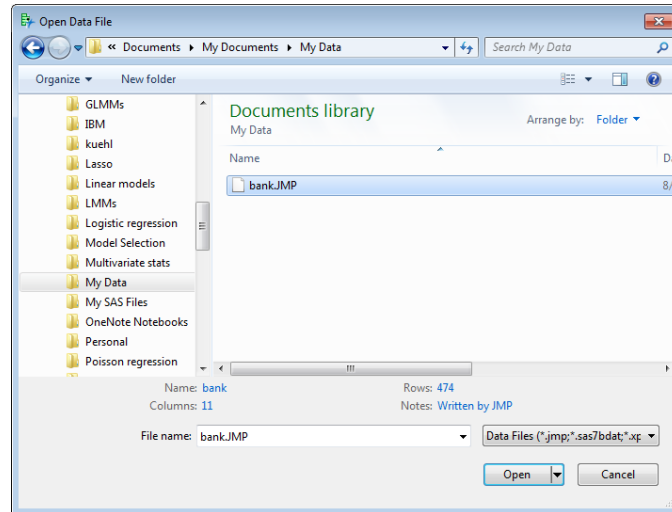
Once the dataset appears in the window (as shown below), you can change a variable's characteristic by double clicking at the top of the corresponding column.

You can also add new variables or additional rows of data to an existing dataset. In order to add variables, select **Cols** on the menu bar and then click **New column**. It will ask you to specify the characteristics of the new column. Once you are done, click **OK.** In order to add rows, select **rows** on the menu bar and then click **Add rows**. It will ask you to input the amount of rows to add, then click **OK**.

V. Transforming and Manipulating Dataset

1. **Sorting Data**
   - Click **Tables > Sort**.
   - The sort window will appear (as shown below). Click the variable name in the box on the left hand side. You can select multiple variables for sorting. The top variable listed will be used for the primary sort; the second variable down will be used for the secondary sort and so on.
   - Select either an ascending or descending sort order (by default it is ascending).
   - Click **OK**.
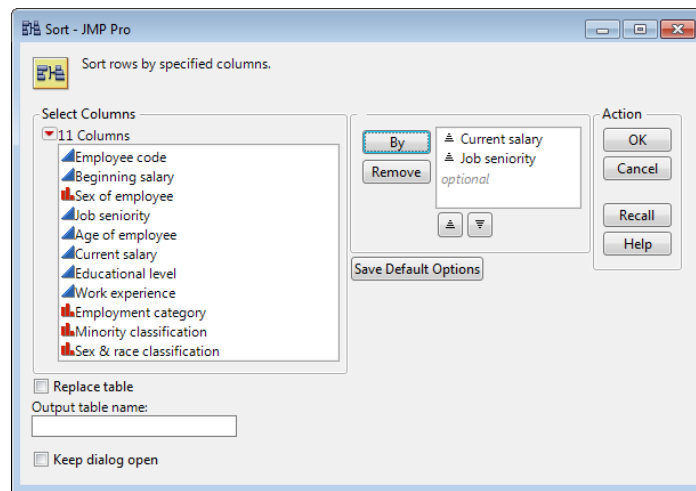
A new data table will pop up with the sorted changes. To make a sorted dataset permanent, save the dataset after sorting.

2. **Creating New Variable Based on Existing Variable**

The two most commonly used procedures to create new variables from existing ones are computing and/or recoding a variable.

- To compute a new variable:

   1. Click **cols> New column.**
   2. Enter the name of the new variable in the **column name** box and specify the attributes of the variable.
   3. Under the **column properties** choose **Formula**.
   4. Click **OK**.
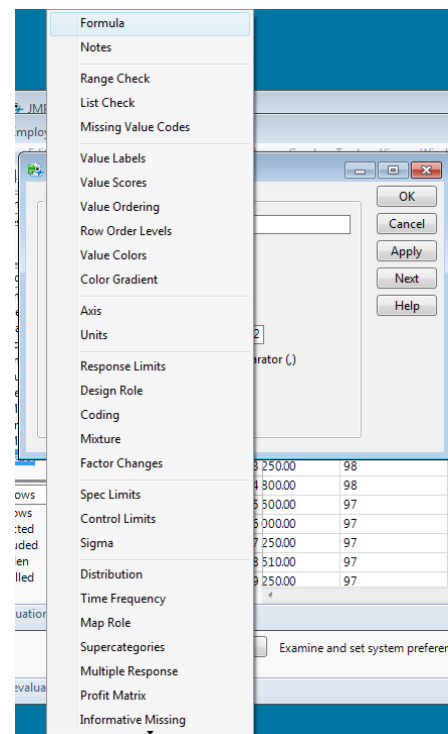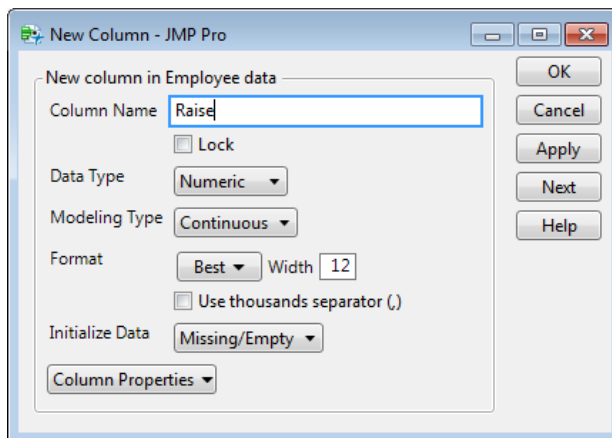
In this particular example "**Raise**" is the new variable created which is the difference between the current salary and the beginning salary. Once the "**formula**" option is chosen from the **column properties**, the following window will appear. Select the first variable "current salary" from the **table column** box, following by the **operator** "-" and then the second variable "Beginning Salary". Click **OK** and in the New Column box click **OK**.
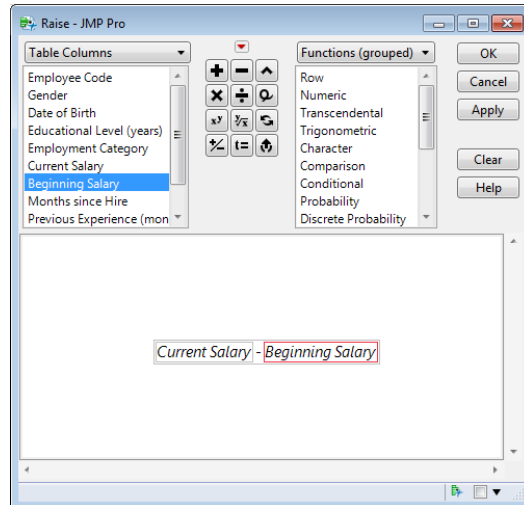
- To recode a variable (in this example changing a continuous variable such as "Education Level" into a categorical variable with 3 categories).

  1. Click **Cols > New column.**
  2. Enter a variable name and specify the properties. For example in this case "Educational_category"
  3. Select **formula** from the **column properties.**
  4. Select **"Educational level"** from the **table column box** and then click **conditional** from the **function group box** on the right and select **"if"** from the **Functions (grouped).** (shown below)
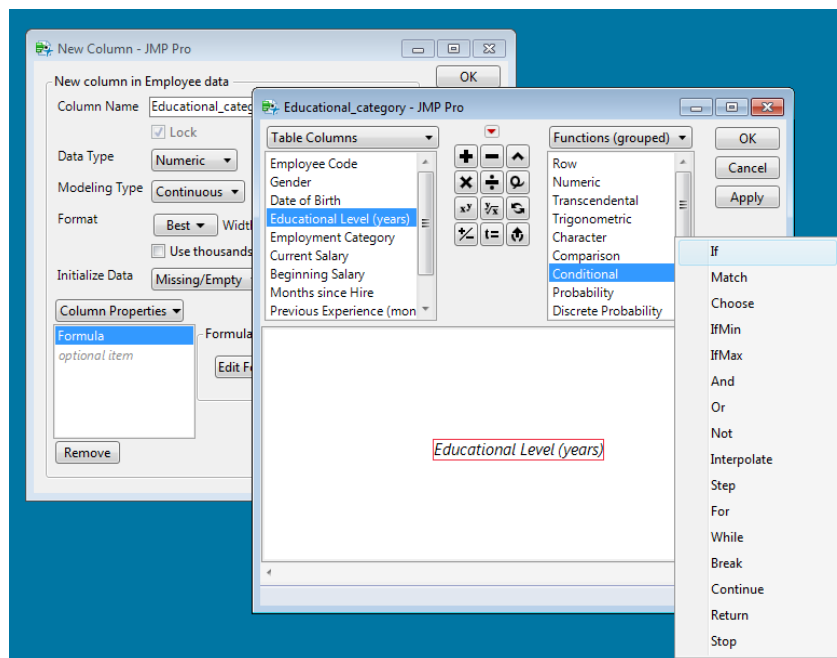
5. Specify the conditions:
   a. Click in the box of the variable name ("Educational level") and type your condition ("< 12") followed by enter.
   Then type the value of "No high School"(including the quotes) in the box labeled "then clause".
   b. Click in the box labeled "else clause", highlight the variable "Educational level", then click **conditional** from the **function group box** on the right and select **"if"** from the **Functions (grouped).** Click in the box of "Educational level" and type then type "=12". Then type the value of "High School"(including the quotes) in the box labeled "then clause".
   c. Lastly assign a value of "Some College", in the case when "Educational level" > 12. The final menu should look as follows:
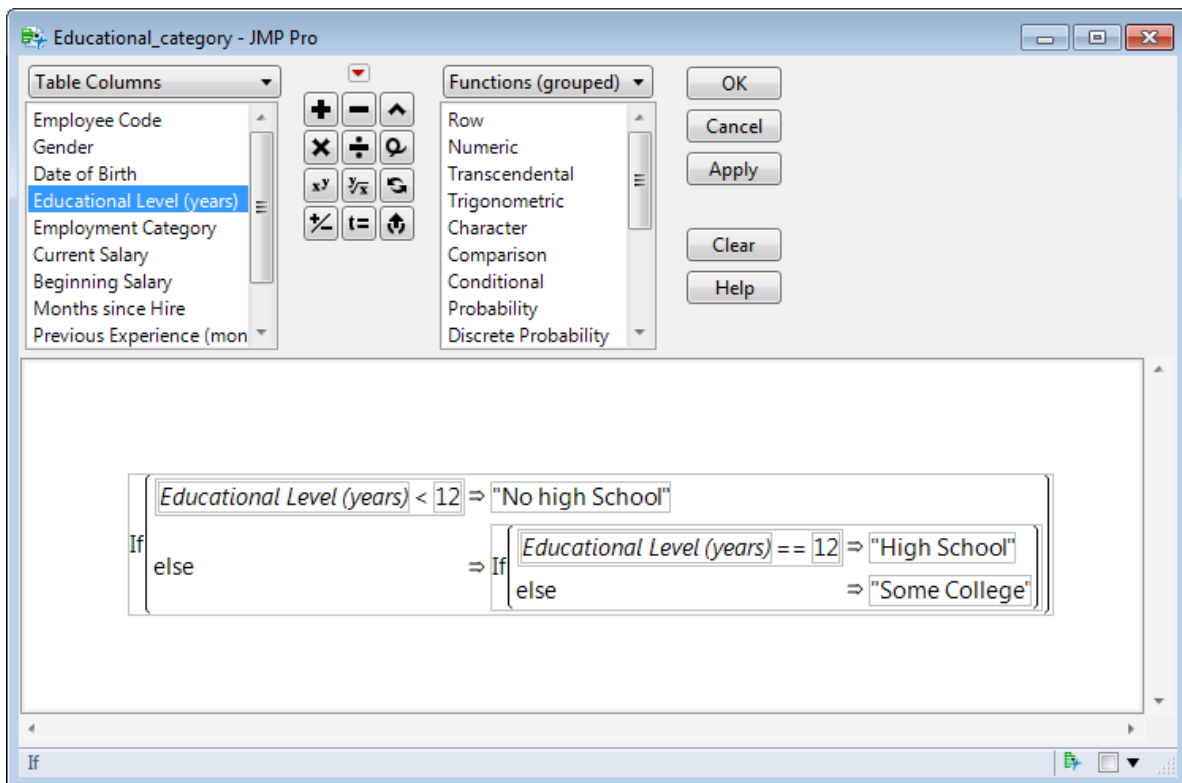

6. Click **Apply**.
7. Click **OK**.
8. In the New Column box click **OK**.

Now in the data sheet, you have two new columns; one called "Raise" which is computed as the difference between current and beginning salary and another called "Educational_category" which is based on the variable "educational level".

Another option for recoding variables when cascading if/else statements get cumbersome is the following:

    Right click on the variable in the area to the left OR select the column and use the "Cols" menu.
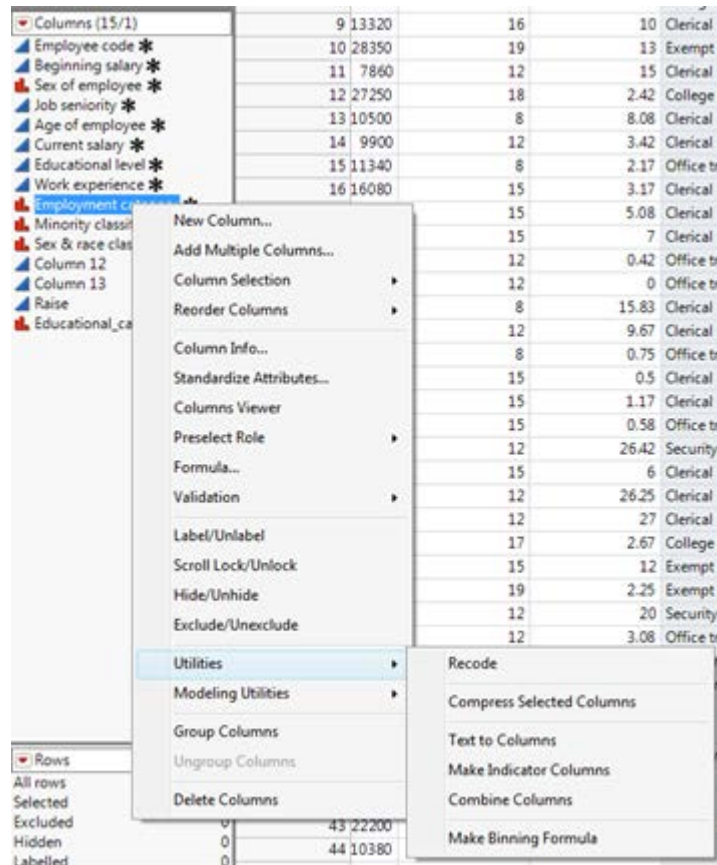
Under "Utilities" there is a "Recode" option which opens this dialogue:
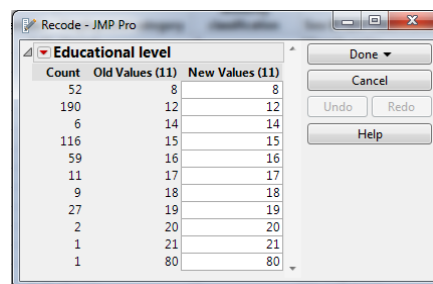

By selecting done, you can save "in place" which will replace the values in the column (Not reccomended), save into a "New Column" or "Formula Column" which will make a formula out of the way you fill out this table.
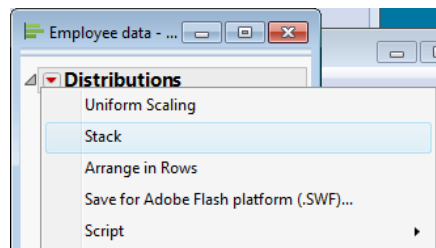
VI. Basic Data Analysis:

1. **Describing Categorical Variables**
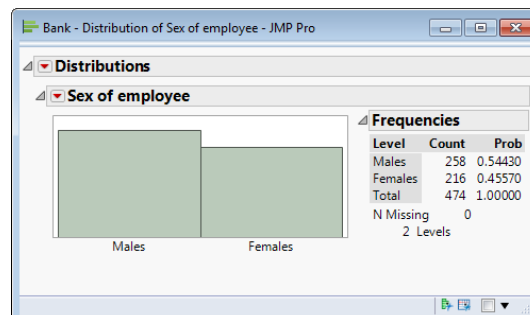
i   Click **Analyze > Distribution**.

ii.   Select the categorical variable of interest (in this example "Sex of employee") and enter it in the **Y, Columns** field.

iii.   Click **OK** and the output window will display the following.
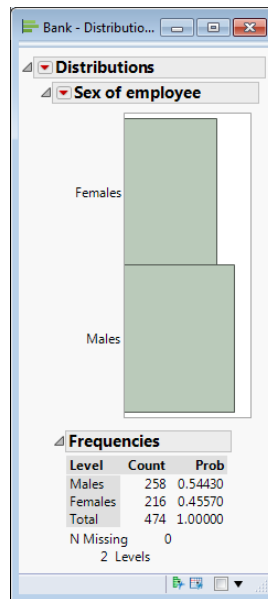
iv.   Click on the red triangle next to **Distributions** and select **Stack**

to obtain the bar chart shown horizontally.

The "**frequencies**" displays each level of the categorical variable (in this case female and male) and the corresponding count and proportion (which is also termed as probabilities). The graph displays the count for each category.

Note: A small red triangle displayed next to a title in the output indicates that addition analysis/output/options can be requested. A small grey triangle displayed next to a title in the output indicates that the output can either be expanded or hidden.

2. **Describing Continuous Variables**

   i. Click **Analyze > Distribution**.
  ii. Select the continuous variable (in this example "Job seniority").
 iii. Click **OK**.
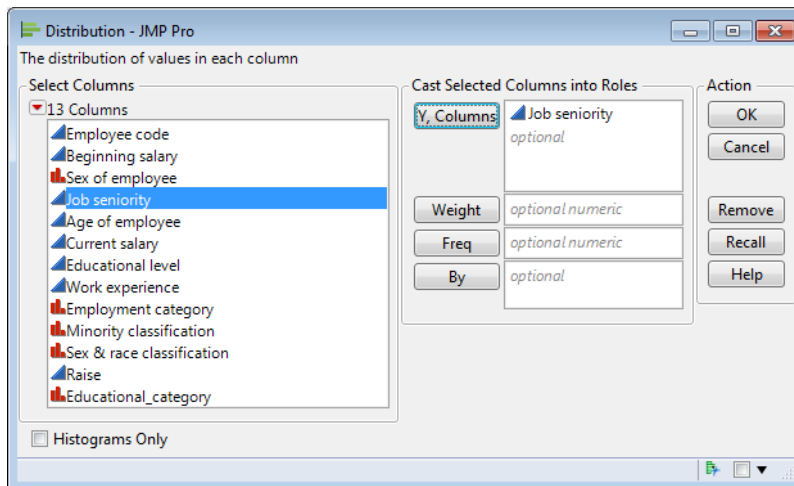
The output window will give you a frequency histogram, boxplot, and summary statistics including the mean and standard deviation of the variable. Click on the red triangle next to **Distributions** and select **Stack** will give the output displayed horizontally
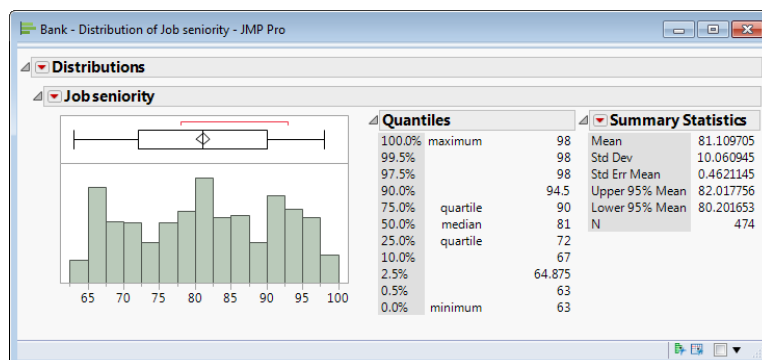
3. **Describing Two Categorical Variables Simultaneously**

A cross tabulation allows you to present the number of observations available in subgroups defined by two categorical variables simultaneously. To obtain a crosstabulation:

i.   Click **Analyze > Fit Y by X**
ii.   Select "Sex of employee" for **row (X, factor)** and "Employment Category" for the column **(Y, factor)**.
iii.   Click **OK**.
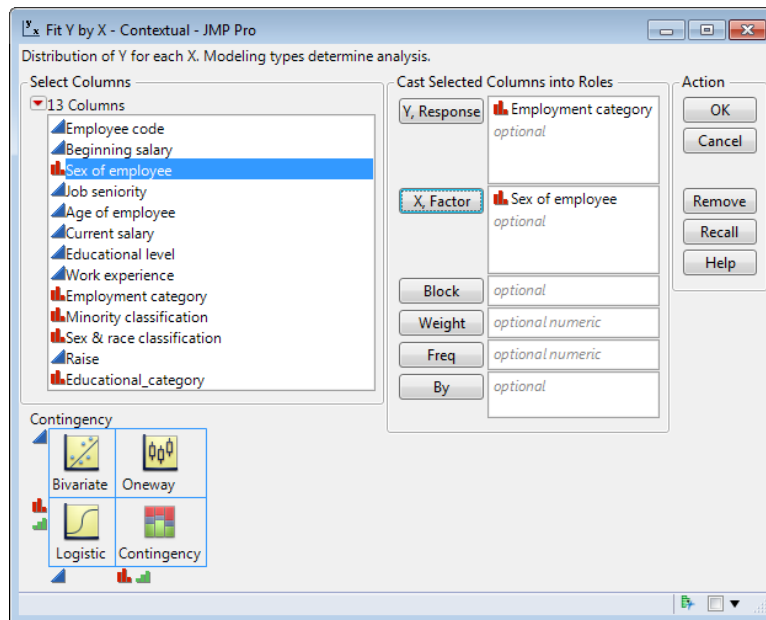
This will produce the following output for the Sex of employee versus Employment category cross tabulation:
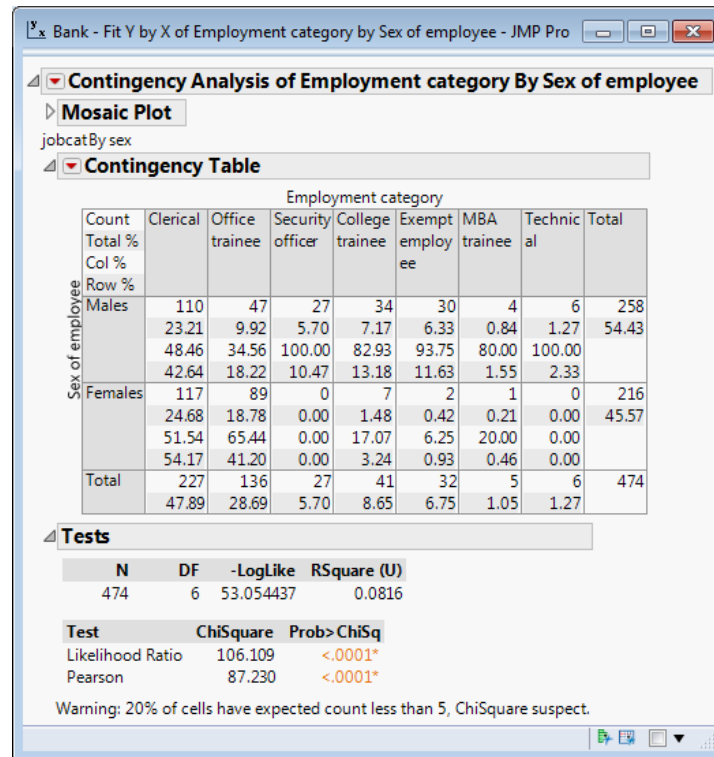
The output screen also has the Pearson Chi-Square test and the respective p-values, which is used to assess the independence of the variables.

Besides the mosaic plot that you obtain in the output a clustered bar chart is an appropriate way to graph two categorical variables simultaneously. To obtain a clustered bar chart go to **Graph>Chart** and under the "Categories, X, Levels" first bring "Employment category" and then "Sex of employee" variables. The order of the variables influences the graph: the second variable will be nested within the first variable.
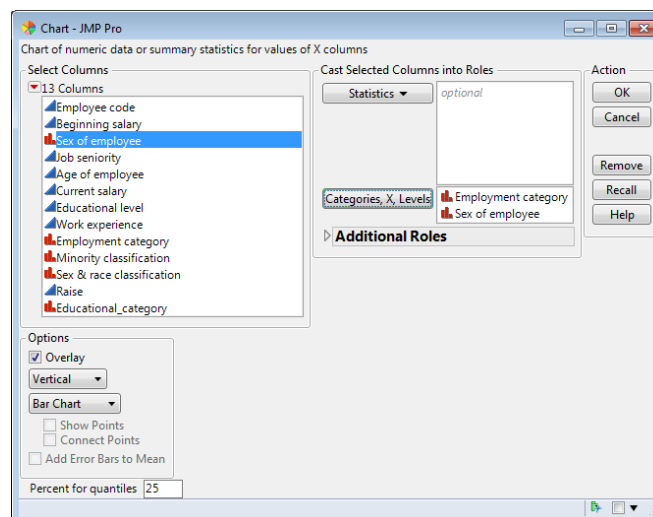
The output window will be displayed as:
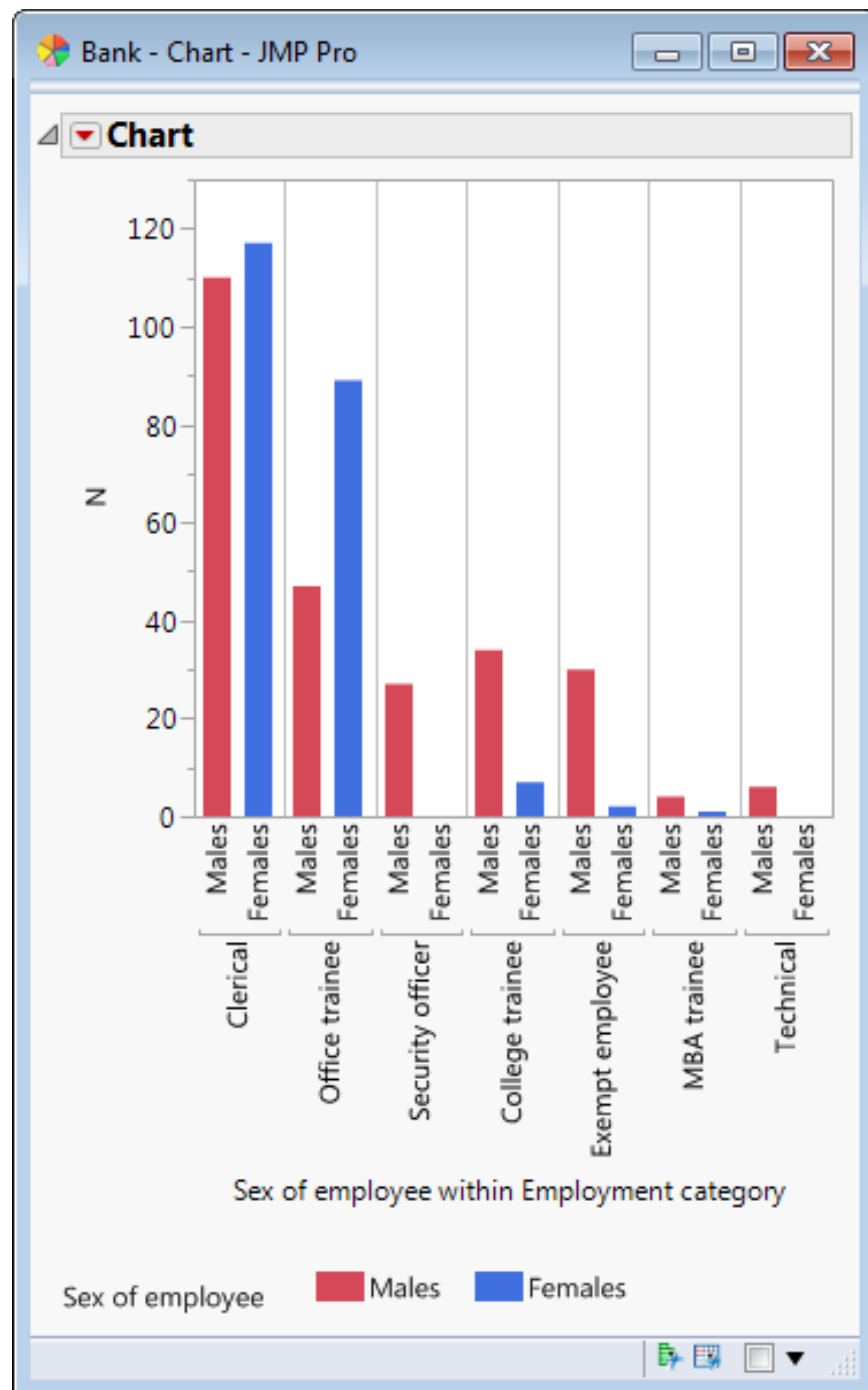
4. **Describing Two Continuous Variables Simultaneously**

To assess how two continuous variables vary together you can ask for the correlation:

To obtain a correlation go to the menu: **Analyze>Multivariate methods>multivariate.**
Select the continuous variables of interest from the **Select Columns** box and click onto
Y, Columns (as shown below).
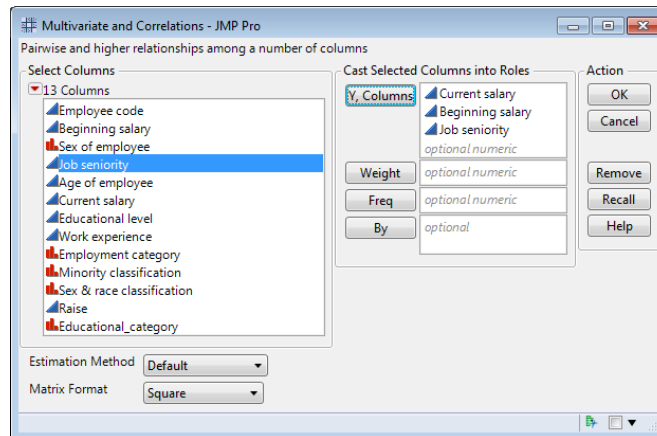
Click OK.
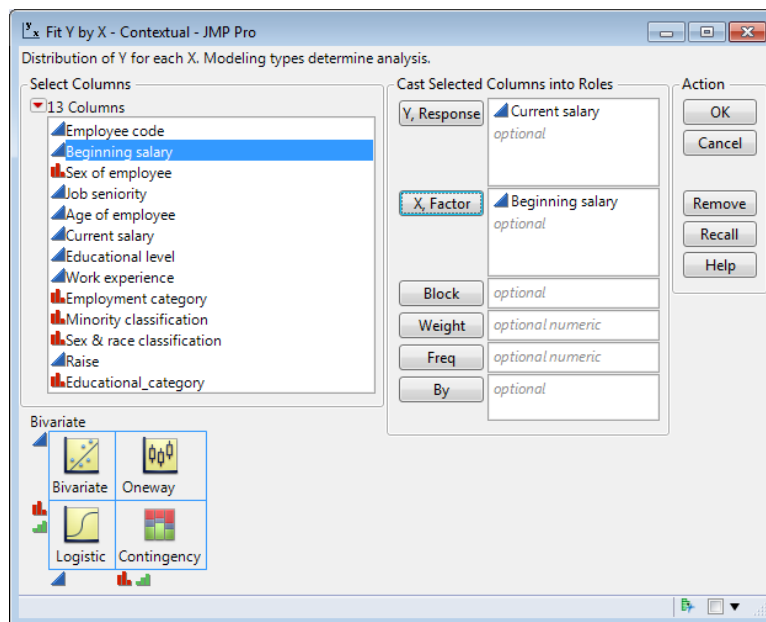
The output screen will appear as:
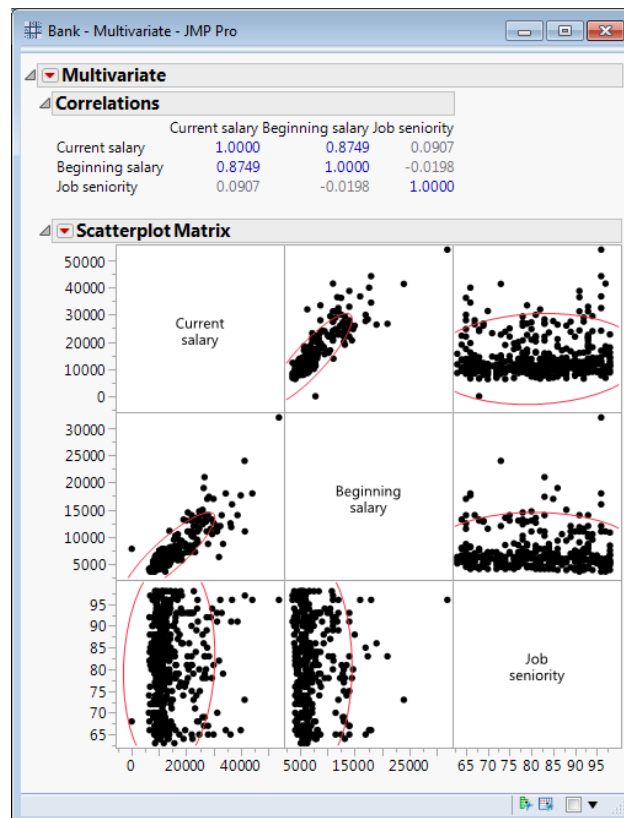
The "Correlations" table shows the correlation coefficient for each variable against each other variable. Note the diagonal is always 1, as each variable is perfectly correlated to itself.

The "Scatterplot Matrix" shows scatterplots for each variable plotted against each other variable. For example, the scatterplot in the middle of the top row shows "Beginning Salary" on the x-axis and "Current salary" on the y-axis. This is the same plot as the scatterplot in the middle of the left column, but the x and y axes have been switched.

To obtain a regression to assess how one continuous variable (the independent variable) predicts another continuous variable (the dependent variable) with a linear relationship:

i.   Click **Analyze > Fit Y by X**
ii. Choose the continuous variables of interest (In this case, "current salary" has been chosen as dependent and "beginning salary" has been chosen as the predictor).
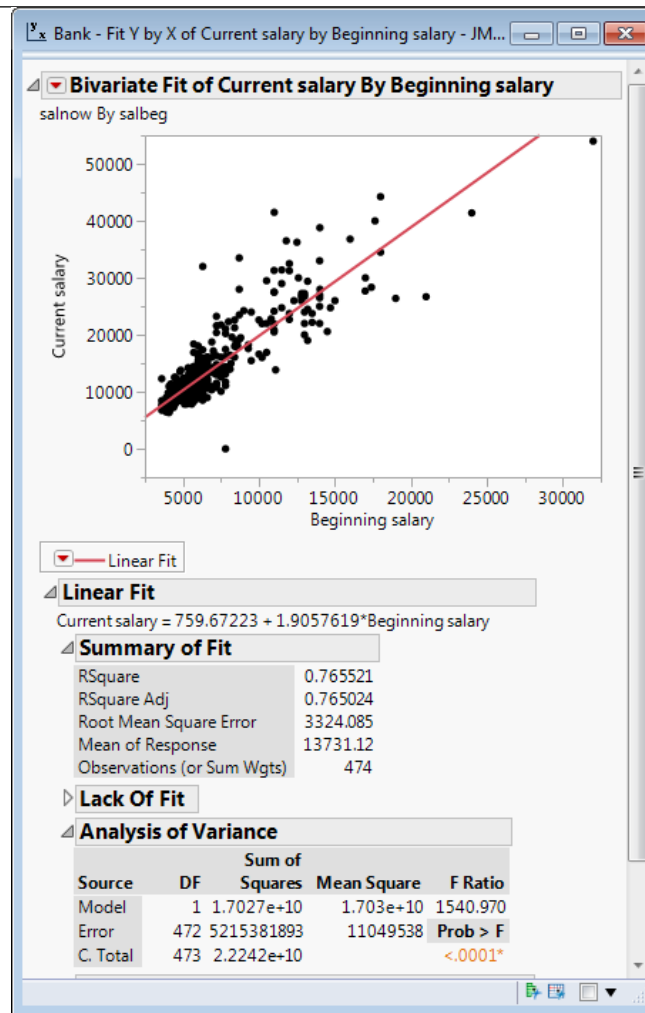iii.

iv. Click **OK** and the output screen will be displayed as:
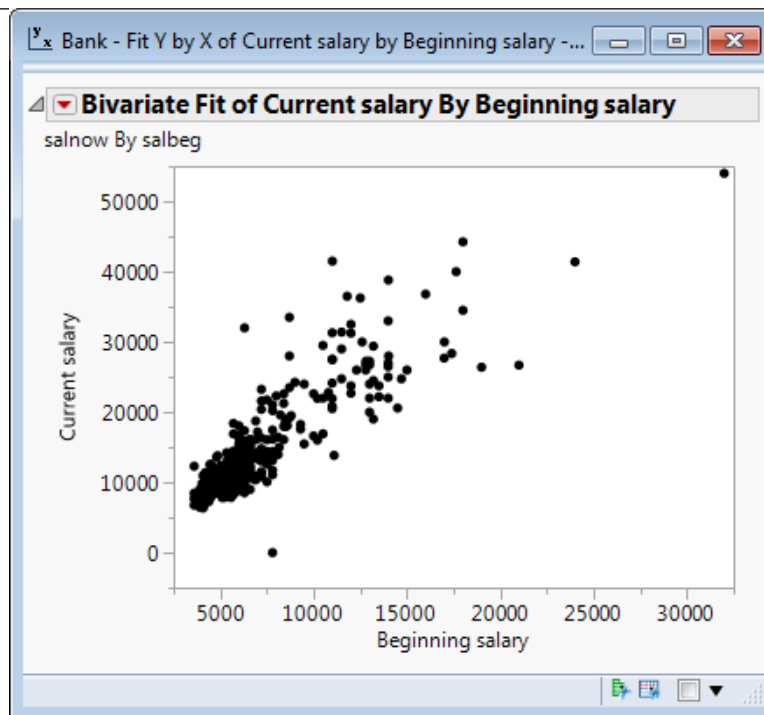
To overlay a regression line on the plot click the red triangle icon to the left of "Bivariate fit" and choose the "fit line" option (as shown below).
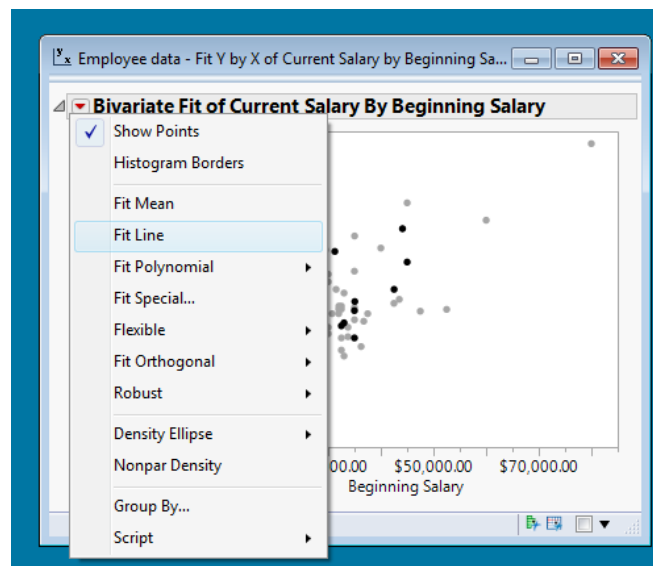

The output window will appear as:

5.   Describing One Categorical and One Continuous Variable Simultaneously

To summarize a continuous and a categorical variable you can calculate for example the mean of the continuous variable for each level of categorical variable:

i.   Click **Analyze > Fit Y by X**
ii. Choose a continuous variables of interest for the dependent list i.e. "**Y, response**" (in this case, "current salary") and categorical variable for the independent i.e. "**X, factor**" (in this example "Employment category")
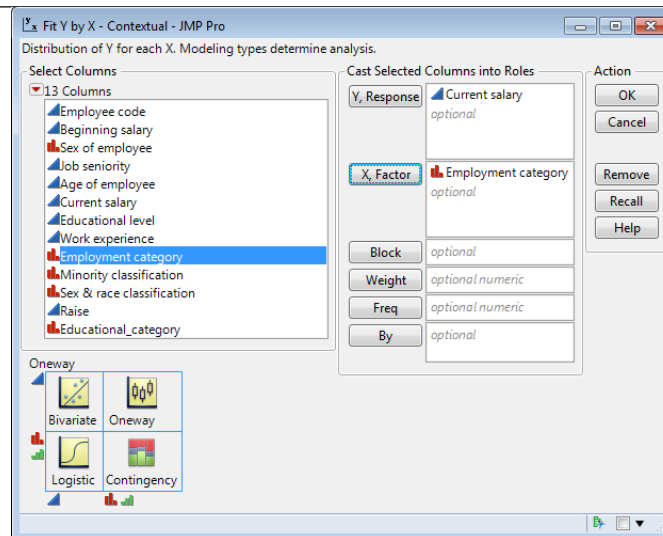
Click **OK**. The above command will generate a dotplot showing the observations of current salary by employment category.
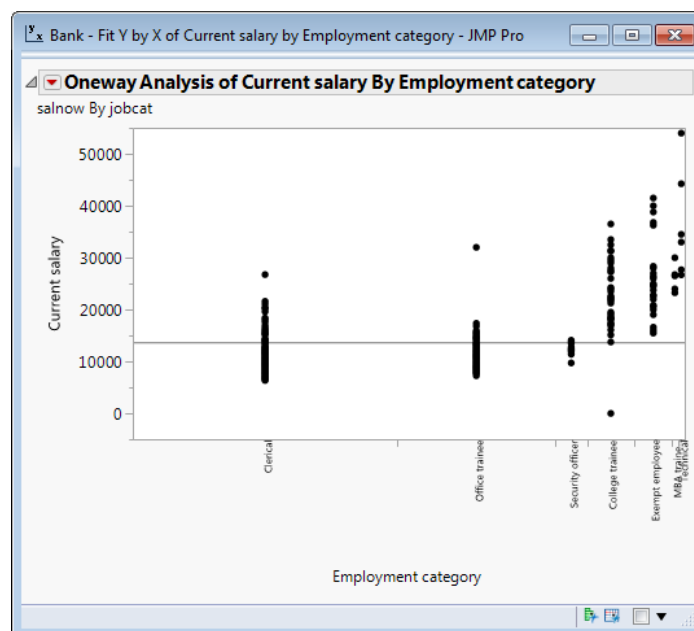
iii. To overlay boxplots on this display, clicking on the red triangle next to Oneway Analysis … and select **Display options** and **Boxplots**
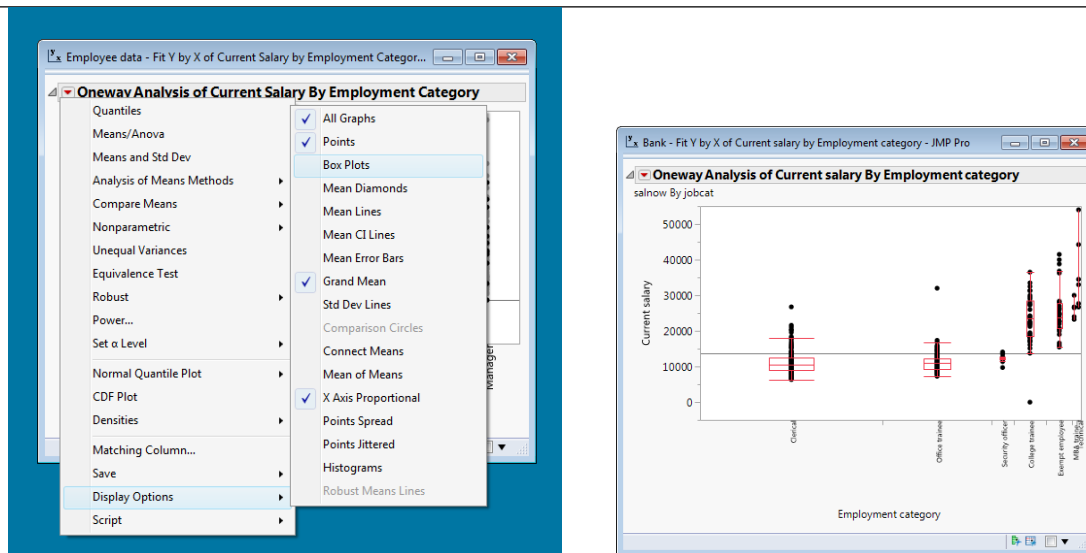
iv. Clicking on the red triangle next to Oneway Analysis … and select **Means/Anova**. In the output for **Means for Oneway Anova** you have output detailing the mean current salary and corresponding confidence interval for each level of the job category.

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
Clerical	227	11134.8	264.6	10615	11655
Office trainee	136	11136.4	341.8	10465	11808
Security officer	27	12375.6	767.1	10868	13883
College trainee	41	23476.7	622.5	22253	24700
Exempt employee	32	25595.6	704.7	24211	26980
MBA trainee	5	26100.0	1782.7	22597	29603
Technical	6	36691.7	1627.3	33494	39889

Std Error uses a pooled estimate of error variance

Level: lists the name of each group.
Number: Number of observation in each group.
Mean: is the mean of each group.
Std Error: is the standard error of each group mean.
Lower 95%: is the lower 95% confidence interval of the group mean.
Upper 95%: is the upper 95% confidence interval of the group mean.

To represent graphically, a continuous and a categorical variable simultaneously, you can use a bar chart where the heights of the bar display the mean of the continuous variable for each level of the categorical variable:

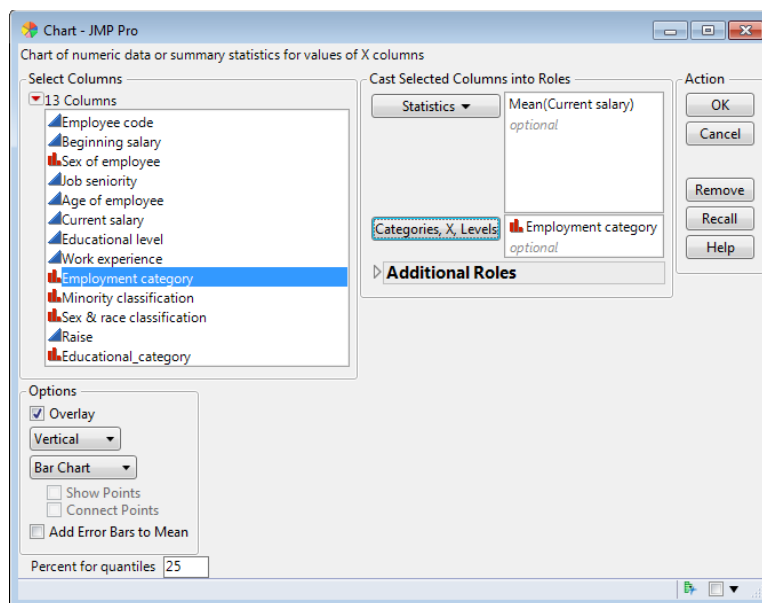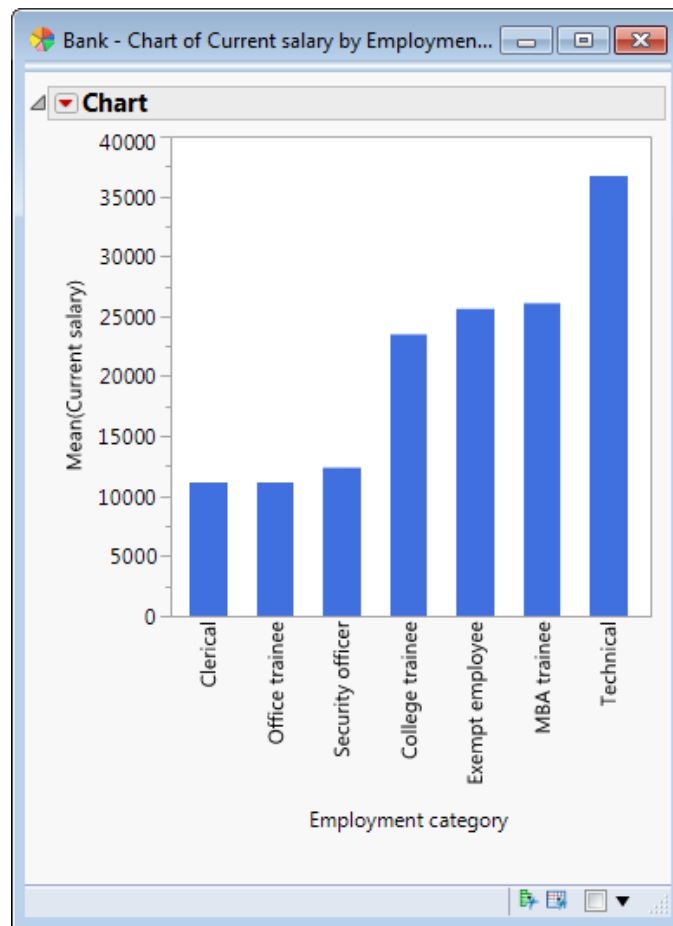i.   Click **Graphs > Chart…**
ii. In the **Select Columns** field, select the continuous variable (In this case "Current Salary"). Click the "**Statistics**" button and select mean.
iii. Select **Categories, X, levels** for the categorical variable. (In this example Employment Category)

iv. Click **OK.**

The output screen will be displayed as:

To add error bars, check the box next to "Add Error Bars to Mean" and use the pull-down menu to choose the summary statistic for your error bars.

Additional analyses you might want to consider when you work with a categorical variable and a continuous variable simultaneously are a T-test or ANOVA. Both of these analyses can be done by clicking **Analyze > Fit Model**. We strongly encourage you to read about these procedures in a Statistics book to understand the underlying concepts.

VII. Saving and Quitting JMP

Before exiting JMP, be sure to save your data file. This can be done

1. By clicking **File** >**Save As** (or **Save**).
2. Choose the appropriate folder to **Save in**: and the desired **File name**.
3. Click **Save.**
4. **To save the analyes or charts, hover or click your mouse over the blue bar at the top of the window to open up the menubar.**
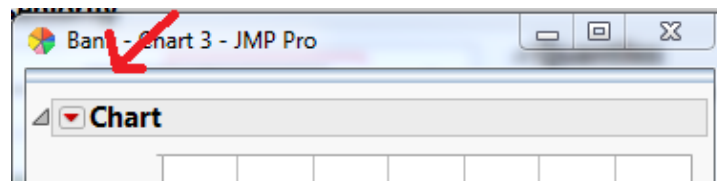
   You may save as a jump report (to work with later) or as an image file. You can also save charts as images by selecting File>Export.

5. Close the JMP session.
6. To quit without saving anything click **File>Exit.**

   .