# Chapter 4: Data exploration and profiling

## Introduction

Contents for this lesson include an introduction and focus on the functionality available in Altair Analytics Workbench to explore and profile data.

The Dataset File Viewer and Data Profiler are introduced as well as the Chart Builder block. The use of

the SAS Language perspective for data exploration and profiling is also covered. A demonstration follows prior to outlining additional methods to explore data prior to a summary.

Figure 1: Contents

A screenshot of a computer

Description automatically generated

Profiling and exploring data is of primary importance in any analytics endeavour and leads to better data understanding, what data to use, what techniques can be applied and in general will make processing the data more efficient.

There are many ways to investigate data, the first step is to simply view contents: columns and rows and then generate statistics, tables, charts and graphs to generate insight. Modelling techniques can also be applied.

Figure 2: Introduction

A screenshot of a computer

Description automatically generated

Altair Analytics Workbench provides many capabilities to explorer and profile data including the Dataset File Viewer and Data Profiler, the Chart Builder and the SAS Language perspective.

Figure 3: Altair Analytics Workbench capabilities

A screenshot of a computer

Description automatically generated

The Dataset File Viewer and the Data Profiler provide a means to view and profile data and are available

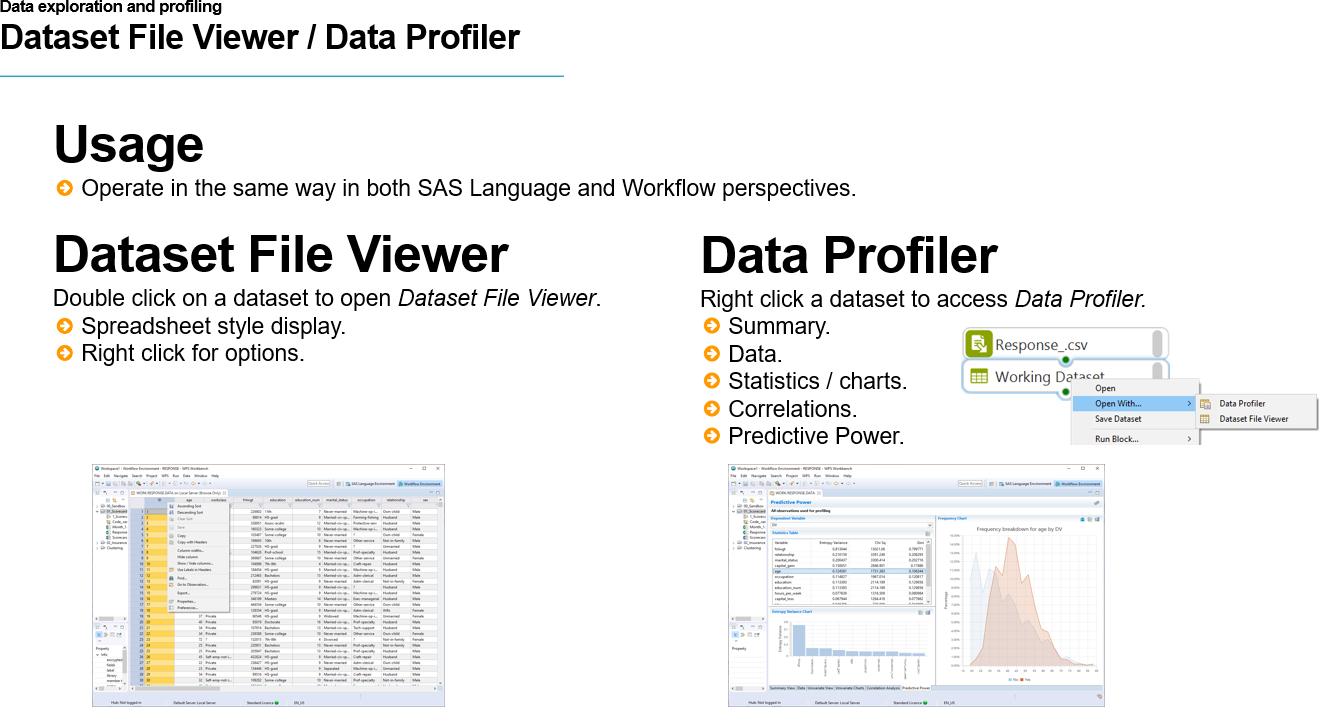
across both the SAS Language and the Workflow perspectives.

Even though both views are accessible via right click, double- clicking a dataset in either of these views

opens the Dataset File Viewer. This is a spreadsheet style display with variables in columns and observations in rows. Right-clicking any variable provides options and filtering capabilities are also included.

The Data Profiler is accessed via right-click only and includes the spreadsheet style display with additional profiling tabs containing statistics and charts to better understand and profile data.

Figure 4: Dataset File Viewer / Data Profiler



The Chart Builder is a Workflow block found in the Export group. This provides an easy mechanism to generate charts. Overlays and by variables can also be included for additional insight.

Figure 5: Chart Builder

A screenshot of a computer

Description automatically generated

The SAS Language perspective provides a wealth of capability to profile data but requires knowledge of the language of SAS to use effectively. Some simple commands can be used to generate insights

and create tables, but the data must be opened in the SAS Language perspective.

Extracting a dataset from the Workflow can be easily achieved by using a SAS code block containing a simple *libname* statement. Once complete, the data can then be easily accessed and used in the SAS Language perspective.

Figure 6: SAS Language perspective

A screenshot of a computer program

Description automatically generated

## Demonstration

So let's get on to a demonstration. This demonstration starts from the Altair Analytics Workbench Workflow perspective with a clean slate.

A project folder is added **by right-clicking the Project Explorer view selecting New and then Project**. From here, General is expanded and Project selected. Clicking Next provides the option to specify the project name, here, First\_Project. The project default location is accepted, which is the Workspace root directory but can be changed if desired.

Figure 7: New project creation

A screenshot of a computer program

Description automatically generated

Once complete, clicking Finish creates the project and the project folder is visible and appears in the Project Explorer view.

A Workflow is added to the project by **right-clicking the project folde**r **and selecting New > Workflow.**

Here the name First\_Workflow is assigned and clicking Finish adds the new Workflow to the project folder and this is automatically opened in the Workflow perspective.

The file used here is a *.csv* file called *Model\_View* and can be found in the data folder. This can be located using the File Explorer view. Double-clicking opens the file in Microsoft Excel.

As can be seen there is an id variable, some demographics and financial details as well as a variable called *DV* with two values: *0* and *1*.

This file can be imported to the Workflow using a number of methods. The first is by using an appropriate block from the import group - as this is a *.csv* file, the Text File Import block is used.

The block can be dragged from the group, or by double-clicking the Workflow canvas with the left mouse button - all blocks can be accessed.

A quick method, given this is a known file type, is to drag the file directly from the File Explorer onto the

Workflow canvas. Hovering over the block indicator shows that additional configuration is required.

Figure 8: Import not yet configured

A screenshot of a computer

Description automatically generated

Double-clicking the block opens its configuration dialog.

Figure 9: Text File Import block

A screenshot of a computer

Description automatically generated

The configuration dialog opens on the File page and as can be seen the Preview displays No preview has been generated. It can be seen from the File page that the file particulars are set correctly, and other pages also list the dataset variables and properties.

Clicking OK it can be seen that the block indicator status is now grey, meaning the block is configured correctly, but the block execution status is red and hovering displays the message: Execution failed with errors.

Figure 10: Execution failed with errors

A screenshot of a computer error message

Description automatically generated

The message does not provide any insight in terms of why there is an error so further investigation is necessary. Right-clicking the block and selecting Open Log provides more detail.

Figure 11: Block log

A screenshot of a computer program

Description automatically generated

From the log, its easy to see why the block is not executing, the file is being used by another process. This is due to the file being currently open in Microsoft Excel.

The file is closed and the block right-clicked. From here Run block > Run To Block is selected and now the block execution status is green meaning the data has been imported successfully. At this point the file is renamed to *Model\_View*.

The dataset can be opened by either double-clicking the dataset or by right-clicking and selecting Open or Open With and choosing either Data Profiler or Dataset File Viewer.

Choosing Open, or Open With > Dataset file Viewer or simply double-clicking the dataset opens the Dataset File Viewer.

Figure 12: Dataset File Viewer

A screenshot of a computer

Description automatically generated

This is a spreadsheet style display with variables in columns and observations in rows. Right-clicking provides access to sort, copy, search, and export options. Note that sort options are only available when a variable header is selected and right-clicked.

Properties provides access to variable and dataset information. The Columns page lists all dataset variables and their properties.

Figure 13: Columns

A screenshot of a computer

Description automatically generated

Other pages provide additional dataset details.

Each variable can be filtered by accessing the filtering icon and selecting an option, the filter applied is displayed as a label below the variable name, and multiple filters can be applied, and here a filter is applied to the variable age: Age ge 79.

As can be seen, only three observations meet the filter criteria of being greater than or equal to 79.

The dataset can be exported by right-clicking and selecting Export. Three options are available: Delimited, Fixed width and Excel. Note that selecting either Delimited or Fixed width provide a preview of the data and this includes the filter applied previously.

Figure 14: Export

A screenshot of a computer

Description automatically generated

Here Excel is chosen and clicking Next provides options to specify the export location. Additional options to open the file once the export is complete and overwrite an existing file are also present.

The option Open when complete is selected and clicking Next reveals the SAS language code for the export. Clicking Finish runs the export and opens the file in Microsoft Excel.

Note that the filter applied previously has still applies

## The Data Profiler

The Data Profiler is accessible from either the SAS Language or Workflow perspective by right-clicking any dataset and selecting the option: Open With > Data Profiler.

Figure 15: The Data Profiler

A screenshot of a computer

Description automatically generated

The Data Profiler opens in a new panel and contains a number of profiling views accessible by selecting an appropriate tab, one of which, the Data tab, contains the Dataset File Viewer with options and filtering capabilities as previously described.

The Summary View tab provides an overview of the dataset and its contents including its name, number of observations and number of variables, here 20000 and 10 respectively.

Dataset variables are visible with characteristics in columns including name, label, type, classification, length, format, informat, cardinality as Distinct values, the number of missing values and a frequency distribution.

The Configure data profiler icon is a shortcut to Data Profiler options available from the Windows > Preferences menu.

Figure 16: Configure data profiler

A screenshot of a computer

Description automatically generated

As can be seen there is an option to exclude the frequency distribution graphs from the Summary view and specify the Number of variables to show on the Entropy Variance chart, more on this shortly.

If we Navigate to the Data page, which is the item directly above the Data Profiler entry. This provides more insight into the default number of rows to sample on import, which informs how variables are classified.

As can be seen, the first 50 rows are read for each variable and this is used to determine the variable classification and to derive the number of distinct values in each variable.

When importing data, variable Type is determined by assessing the first 50 values for each variable. If a variable is determined to be numeric and there is less than 50 distinct values the variable Classification is Categorical and for numeric variables with more than 50 distinct values, the Classification is Continuous.

For Character variables, if there are less than 50 distinct values the variable Classification is Categorical and if there are more than 50 distinct the variable Classification is Discrete.

Bear in mind that the Classification threshold of 50 is a default value and can be changed as desired.

From the Summary View tab, clicking any column header sorts the variables. Here, Type is clicked and as can be seen all numeric variables are listed first followed by all character variables.

As can be seen, of the seven numeric variables *ID*, *age*, *hours\_worked*, *income,* and *capital\_wealth* most 9have greater than 50 distinct values and as a result are all determined to be continuous. The variable *Years\_Education* has been assigned the Classification Categorical as it has only 13 distinct values.

There are only 3 Character variables and all are classified as Categorical as a result of the number of distinct values in each not reaching the Classification threshold of 50 to be assigned the classification of Discrete.

The Summary View, among others, can be copied to the clipboard using the Copy data to clipboard icon, here contents are pasted into Microsoft Excel to illustrate.

Figure 17: Copy data to clipboard

A screenshot of a computer

Description automatically generated

The Univariate View generates univariate statistics for all numeric variables. Default statistics include the number of valid values, the number and percentage of missing values and common measures of central tendency and dispersion.

For this dataset, the variable *ID* can be ignored. As can be seen, no variable has a large number of missing values and there appears to be adequate ranges for all variables

The average age at 46 seems reasonable as do the averages for *Years\_Education*, *income* and *capital\_ wealth*. The average for *hours\_worked* looks a little low at 30.5, although there may be reasons for this.

The display can be toggled to show formatted values, and the view can be copied to the clipboard.

The configure statistics icon provides access to additional Univariate statistics. Statistics are grouped as Quantiles, Variable Structure, Measures of Central Tendency, Measures of Dispersion and Others.

Figure 18: Configure statistics

A screenshot of a computer

Description automatically generated

Some options are displayed by default as denoted by that statistic being visible in the Univariate View. From Measures of Central Tendency, the Median is selected and clicking OK shows that a new column has been added to the right showing the median for each variable.

Figure 19: Univariate View

A screenshot of a computer

Description automatically generated

The low median *capital\_wealth* of approximately 4400 in comparison to an average of 23035 supports the low average and median *hours\_worked* in that a household or individual with low average *hours\_worked* may not have the capacity to save to the same extent as those working longer hours.

The Univariate Charts view is simple to use. Select a variable from the Variable selection dropdown and a frequency table and chart are generated.

The frequency table can be copied to the clipboard and the chart type can be changed by selecting an appropriate icon. Clicking Edit chart opens the Chart Editor, from here the chart can be copied to the clipboard as a jpeg.

Figure 20: Chart Editor

A screenshot of a graph

Description automatically generated

The selected chart type is sticky and so changing the variable may require modifying the chart type applied. Here a categorical variable, *occupation* is selected and the frequency chart displays a histogram, the chart type can be changed to a pie chart by clicking the show as pie chart icon.

Viewing *Age* as a histogram it can be seen that it looks approximately normally distributed with a good spread. Viewing the variable *sex*, it can be seen the data comprises in excess of 60% females.

The variable *DV* contains two values *0* and *1*. Notice that there are predominantly more *0*'s than *1*'s.

Correlations can be generated from the Correlation Analysis tab, variables are moved from the Unselected Variables list to the Selected Variables list by either double-clicking or using the chevrons.

Figure 21: Correlation Analysis

A screenshot of a computer

Description automatically generated

The Pearson correlation coefficient is generated for each pair of variables. This can be changed to either. Spearman Rho or Kendalls Tau from the coefficient dropdown.

The results are presented in a matrix view and squares, colours and colour depth are used to convey the magnitude of the association.

The diagonal squares represent the association between a variable and itself, clicking any off-diagonal square, for example the square representing the correlation between the variables *DV* and *age*, generates statistics and a scatterplot to the right-hand side.

The correlation statistics shows the variables used, the coefficient type and magnitude and a significance value. Here the association between the variables of approximately 0.65 is significant.

The view can be changed to a table by selecting the table icon and can be sorted if necessary. Again, the option to copy to the clipboard and edit the chart are also available.

 In general, there are variables that are associated with *DV*, this is good if the focus is to use *DV* as a dependent variable to model.

There are also other pairs of variables exhibiting strong associations. For example, *age* and *income* and *income* and *capital\_wealth*. These are easy associations to understand as older people would have had time to increase income and accumulate capital wealth.

Assessment of the usefulness of each variable to model an outcome or dependent variable is available from the Predictive Power tab.

Here, *DV* is selected as the dependent variable from the dependent variable dropdown, four statistics are generated for each variable. The entropy Variance statistic is illustrated in descending order in a chart below the statistics table.

Figure 22: Predictive Power

A screenshot of a computer

Description automatically generated

Selecting any variable illustrates its distribution across the categories of the dependent variable. Here, *age* is selected. Notice that different colours are used for each category of the dependent variable.

The *age* distribution for those in the *1* category is notably higher than those in the *0* category and as a result *age* may prove to be a good predictor of the variable *DV*. The overall distribution of *age* can be toggled by clicking the chevrons.

Figure 23: Viewing all data

A screenshot of a graph

Description automatically generated

Other variables can be assessed similarly, and all output generated can be copied to the clipboard.

## The Chart Builder

Additional profiling capabilities are available by using charts to visualize data. The Chart Builder block can be dragged onto the Workflow canvas from the Export group or by double-clicking the Workflow canvas and typing ‘c‘ to locate it.

The block is then double-clicked to place it on the Workflow canvas and then dragged and dropped onto the **Model\_View** dataset.

Figure 24: Chart Builder block

A screenshot of a computer

Description automatically generated

The status indicator message: Block requires at least one plot defined, is warranted at this stage as the block has not yet been configured. Double-clicking the block or right-clicking and selecting Configure Block opens the chart builder.

Figure 25: The Chart Builder

A screenshot of a computer

Description automatically generated

There are a number of areas, first Plots, a chat type can be selected from the plots dropdown. Here, Histogram is chosen. Once selected, variables can be chosen and here *age* is selected. As can be seen from the Preview area, changes must be saved to update it and display a chart, clicking CTRL+S performs this and results are displayed.

A title and footnote can be added via title and footnote and the chart size can be set from this area also.

Depending on the chart type selected, additional options become available in the options area. For the histogram, a legend can be assigned - here distribution of *age*, CTRL+S is used to update the preview.

Figure 26: Legend label

A graph of a graph

Description automatically generated

A frequency variable can be selected as well as the y-axis scale chosen, currently this is set to percent. Bins are automatically determined but deselecting allows users defined values to be input, here 50 is used as an upper boundary limit and provides an additional level of granularity.

Figure 27: Number of bins increased to 50.

A screenshot of a graph

Description automatically generated

Additional style options are provided, and the chart axis position can be changed by selecting an option from axes. Charts can be added or deleted using the appropriate icon, here a chart is added, and this time Horizontal Bar is chosen.

Figure 28: Horizontal Bar chart

A screenshot of a computer

Description automatically generated

Only categorical variables are available for selection and here *Marital\_Status* is selected. Once saved, the preview shows the number of observations in each category with *Married* being the largest.

Additional options are available to the right-hand side including Group. Adding a Group variable colour codes each bar based on the categories of another variable, here *DV* is chosen. Each *Marital\_Status* category is colour coded to reflect the number of *0*s and *1*s in each.

Figure 29: Adding a Group variable

A screenshot of a graph

Description automatically generated

A statistic can also be chosen. Options are Mean, Sum or Frequency with frequency being the default. The option Response variable only becomes available if either Mean or Sum is selected from the Statistic drop-down.

Here, *DV* is removed as a Group variable and Mean selected from the Statistic drop-down with ­*capital\_wealth* selected as the Response variable. Saving updates the preview.

Figure 30: Adding a Response Variable

A screenshot of a graph

Description automatically generated

The chart now shows the mean *capital\_wealth* for each *Marital\_Status* category.

A by variable can be added from the chart Preferences dialog which is accessed by clicking the Configure Chart icon. The first option By Variable allows selection of one or more by variables.

Here *DV* is chosen, additional options in this menu relate to style options for both the x and y and the second x and y axes. Clicking OK and then saving using the keyboard shortcut CTRL+S updates the preview.

Figure 31: Adding a by variable

A screenshot of a graph

Description automatically generated

Given that there are two *DV* values, two charts are generated showing the mean *capital\_wealth* across *marital\_status* categories for both categories of the variable *DV* and these can be navigated by clicking Next chart or previous chart.

As can be seen there is a distinct difference between *capital\_wealth* for those in the *1* category of *DV* in comparison to the *0* category of *DV*.

For those *Married*, *capital\_wealth* is steady across the *DV* categories*.* For all other *marital\_status* categories *capital\_wealth* is higher for the *DV 1* category in comparison to the 0 category.

Charts can be exported as PDF, HTML and as images which is SVG format. A location and name can be specified. Here PDF is selected, the file named: charts, and the output location are selected as the project folder.

Using CTRL+S updates settings and outputs the PDF file. Refreshing the project shows the PDF output, which can be double-clicked to view.

Figure 32: PDF output

A screenshot of a computer

Description automatically generated

## Using the SAS Language perspective to profile data

 Additional Altair Analytics Workbench functionality to profile data is available through coding using the SAS Language perspective. Data must be made persistent to be accessible from the SAS Language perspective.

To do this, a SAS Code block is dragged from the Code Blocks group and connected to the Model\_View dataset and once complete, the SAS code block is double-clicked and opened.

Figure 33: SAS code block

A screenshot of a computer program

Description automatically generated

The block can be used to apply the language of SAS in a Workflow. 0, 1 or many datasets can be connected and output. A quick reference example of how to pass data in and out is provided.

To duplicate the connected dataset, a new output is added by clicking the Add new output dataset icon from the Outputs panel. The dataset name can be changed and it can be easily dragged into code for easy reference. The input dataset can be set to the output using the same mechanism.

Figure 34: Adding SAS Language code

A screenshot of a computer

Description automatically generated

Once complete clicking CTRL+S updates settings and applies the code. Returning to the Workflow a new dataset has been output to the Workflow canvas.

Figure 35: New dataset

A screenshot of a computer

Description automatically generated

Here there is a need to export the data so that it can be accessed via the SAS Language perspective and profiled further.

The code is cleared and a LIBNAME statement added. This will create a library called outfile at the location specified, note this is the location of the project folder, and any file sent there will be stored as a WPD file, ALTAIR’ proprietary file format.

The project folder location is easily obtained by right-clicking the project folder and selecting Show in System Explorer, the address can be copied and pasted into the SAS Code block.

A code statement is added to take the input and save it to the library as *model\_view\_copy*. No output is necessary as the file is being exported and not output to the Workflow.

Figure 36:Using LIBNAME

A screenshot of a computer

Description automatically generated

At this point, the code line beginning with libname is copied as this will come in useful shortly. CTRL+S is used to update settings and returning to the Workflow it can be seen that the process has run successfully as reflected by the green execution status indicator and the message: Completed successfully and up to date.

Figure 37: Successful execution

A screenshot of a computer

Description automatically generated

Refreshing the project, notice the file has been output to the project folder. At this point, a new code file is added from File > New > Program and named *profiling.sas*.

At the point of creating the file and as a result of being in the Workflow perspective a Change Perspective notice pops up informing that program files are best viewed with the SAS Language perspective. Would you like to switch?

There is an option to remember the decision also, this is selected here and Yes clicked.

Figure 38: Change Perspective notification

A screenshot of a computer error

Description automatically generated

Note that a similar notification appears if creating or accessing a Workflow from the SAS Language perspective.

At this point, the code file is open in the SAS Language perspective. To access the *Model\_View\_Copy* dataaset the copied libname statement is pasted and run.

A new library is created: outfile and this contains the file *Model\_View\_Copy* and can be referenced in code.

Figure 39: Accessing the dataset

A screenshot of a computer

Description automatically generated

Here some simple commands can be applied to further profile the data. First PROC CONTENTS.

The code statement is simply proc contents data and an equals sign to the data, as well as the dataset name which can be dragged from its library and dropped into the code pane. A semi-colon is added and the statement completed with a run statement and a final semicolon.

Figure 40: PROC CONTENTS

A screenshot of a computer

Description automatically generated

Once complete, the code is selected and run. The output generated can be accessed from the Output Explorer view via the HTML node. The screen can be split to show the code side by side with the output.

Figure 41: Code and output

A screenshot of a computer

Description automatically generated

As can be seen proc contents outputs dataset information, including name, size and number of variables. Three tables are generated including an alphabetic list of variables and their attributes.

Another useful procedure is PROC UNIVARIATE, it takes the same form as proc contents and outputs statistics and charts for all numeric variables. The Results Explorer to the right-hand side can be used to navigate.

Figure 42: PROC UNIVARIATE

A screenshot of a computer

Description automatically generated

Frequency tables and cross tabulations can be generated using the procedure PROC FREQ. The variables to tabulate are listed singularly. To generate cross-tabulations an asterisk is used between variables and additional options are used to declutter the output.

Figure 43: PROC FREQ

A screenshot of a computer

Description automatically generated

Here tables are created for the variables *DV, marital\_status* and *occupation* and cross-tabulations are created for *DV* with *marital\_status* and *DV* with *occupation* showing row percentages.

From the results generated, it can be seen that 45% are married and that occupations are well spread.

From the cross-tabulations it can be seen that the bulk of those in the *1* category of *DV* are *Married*, the distribution for *occupation* does not show any difference across the *DV* categories.

## Accessing Workflow code

The code for any Workflow can be accessed by right-clicking any blank area of the Workflow canvas and selecting the option Export to SAS Language Program.

A temporary directory must be set, this can be anywhere, and once complete clicking Next provides access to the code, the code can be exported to a program file or copied to the clipboard by clicking the Copy Code icon.

Figure 44: Export to SAS Language Program

A screenshot of a computer program

Description automatically generated

Here, Copy Code is clicked and the code pasted to the file created in the SAS Language perspective previously and run.

The code simply imports the data to the temporary library selected when exporting the code and is evident from the ALTAIR Server Explorer view as *Model\_View\_Copy* and can be accessed as normal.

Figure 45: Running the code

A screenshot of a computer

Description automatically generated

This has been an introduction to data exploration and profiling with Altair Analytics Workbench. Other mechanisms and methods can be used to profile data including modelling techniques such as decision trees to illustrate associations and regression to determine the strength of the association between variables.

Segmentation methods such as clustering can be used to identify groups with distinct characteristics and variable reduction methods can be further used to assess variable associations. Other capabilities such as the WoE Transform block can be used to assessing associations and applying appropriate binning to variables.

Figure 46: Other ways to explore data

A screenshot of a computer

Description automatically generated

## Summary

Figure 47: Summary

A close-up of a document

Description automatically generated

Altair Analytics Workbench capabilities for exploring and profiling data are deep and provide myriad ways to investigate any data and some methods have been demonstrated here.

This lesson introduced the Dataset File Viewer, the Data Profiler, the Chart Builder and the use of the SAS Language perspective to profile and explore data as well as introducing other ways to profile data including modelling techniques and segmentation methods.