R&D Skills Tutorial #1

An Introduction to Electronic Tabular Data Management, Analysis, and Visualization using TOPCAT

PHYSICS 499-0001, 5590-0001

Practical R&D Skills (PRDS) in Astrophysical Data Analytics

Prof. Dan McIntosh (PRDS Instructor), ____________ (Graduate Teaching Asst.)
University of Missouri - Kansas City, Spring 2024

On-track completion due date: February 8, 2024

Purpose

The purpose of this tutorial is to equip students with fundamental data analytics skills specific to tabular data using TOPCAT, the intuitive data analysis and mining tool designed by UK astronomers. In today's information-intensive world, the collection of data about everything and anything is commonplace. Put simply, data analytics is the process of analyzing data, identifying trends in the data, and drawing meaningful conclusions from the data. The tutorial begins with practice opportunities (**Learning Activities**) and progresses to graded critical-thinking and problem-solving opportunities (**Assessment Tasks**) for students to demonstrate application of TOPCAT data analysis skills in realistic (i.e., practical) contexts. To maximize accessibility of this learn-by-doing tutorial and to demonstrate the broad application of the tutorial skills, the activities and tasks are applied to a table of student grades from a generic college course.

Student Learning Outcomes (SLOs): *Students who complete this tutorial with satisfactory or better competency will be able to:*

1. read in and view an electronic data table (rows & columns in CSV format) using TOPCAT
2. use TOPCAT to perform basic exploration and manipulation operations to summarize the contents of an electronic data table
3. use TOPCAT to do common 1-dimensional (single-column) data analytics operations (histogram graphs, simple statistics, identify/define subsamples)
4. use TOPCAT to do common 2-dimensional data visualizations (x vs y graphs)
5. produce professional table and figure captions

This tutorial has a total estimated completion time (ECT) of 14 hours for a typical student to complete the tutorial, ***achieve Satisfactory (or better) overall competencies***, and submit their work. This tutorial is worth 140 points. Instructions for tutorial submissions, as well as the grading process, are provided on the canvas course page in the Module related to this tutorial. Please review the 'R&D Skills Tutorial Submissions: Format Expectations' canvas page.

1. Introduction to TOPCAT – A Nifty Tool for Quick Analyses and Explorations of Data Tables

In this section, we will introduce you to TOPCAT[1] and we will use it to load, manipulate, and visualize electronic data tables. TOPCAT is a Java (a programming language) based data analysis tool. It has a strong focus on Astronomy-related data analysis, but I find it to be a very useful tool for introducing novices to data analytics operations as well as for expert coders and data sciences to perform quick explorations and preliminary analyses of new data in any discipline context.

If you wish to use TOPCAT on your own computer, installing this free software application is a straightforward process. To begin, visit the official website in footnote (1) and navigate to the Downloads section. Select the appropriate version compatible with your operating system, whether it’s Windows, macOS, or Linux. Once the download is complete, follow the installation instructions provided on the website or within the downloaded package. Typically, this involves running an installer and accepting the default settings. After successful installation, launch TOPCAT, and you’re ready to explore and analyze electronic tabular data seamlessly!

Starting a TOPCAT session will open its main Graphic User Interface (GUI) window (see Fig. 1):

- For Mac users, go to your 'Applications' folder to locate TOPCAT. Double-click it to open it. I recommend pinning to your dock once you open it so that for future use, you can open it directly from the dock by single-clicking on TOPCAT icon.
- For Windows users, double click on the TOPCAT shortcut created after installation.

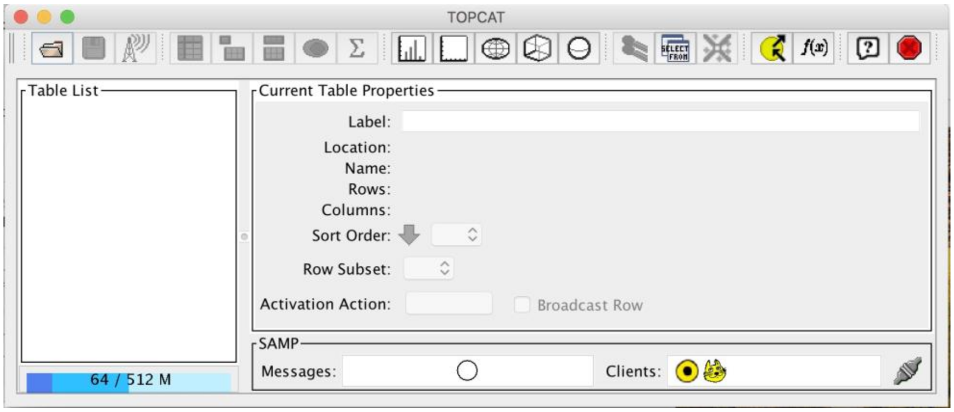


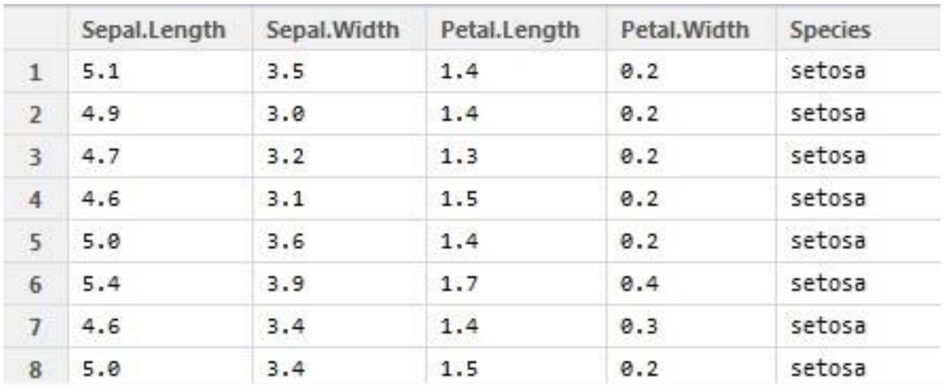
Figure 1: A new TOPCAT Graphic User Interface (GUI) session.

1.1 Introduction to Tabular Data

Electronic tabular data (hereafter, a data table) is an arrangement of information in rows and columns (or in more complex structures[2]). For example, Figure 2 shows a data table of different quantitative measurements of flowers. Electronic data tables are stored in a variety of different formats identified by their file extension, e.g., CSV (.csv), ASCII (.cat), and FITS (.fits) are common data table files. In this tutorial we will use Comma-Separated-Value (.csv) tables. Use google or your favorite gen-AI tool such as ChatGPT to understand these and other data file extensions as needed.

[1] Tool for OPerations n Catalogues And Tables: http://www.star.bris.ac.uk/˜mbt/ topcat/

[2] https://en.wikipedia.org/wiki/Table_(information)

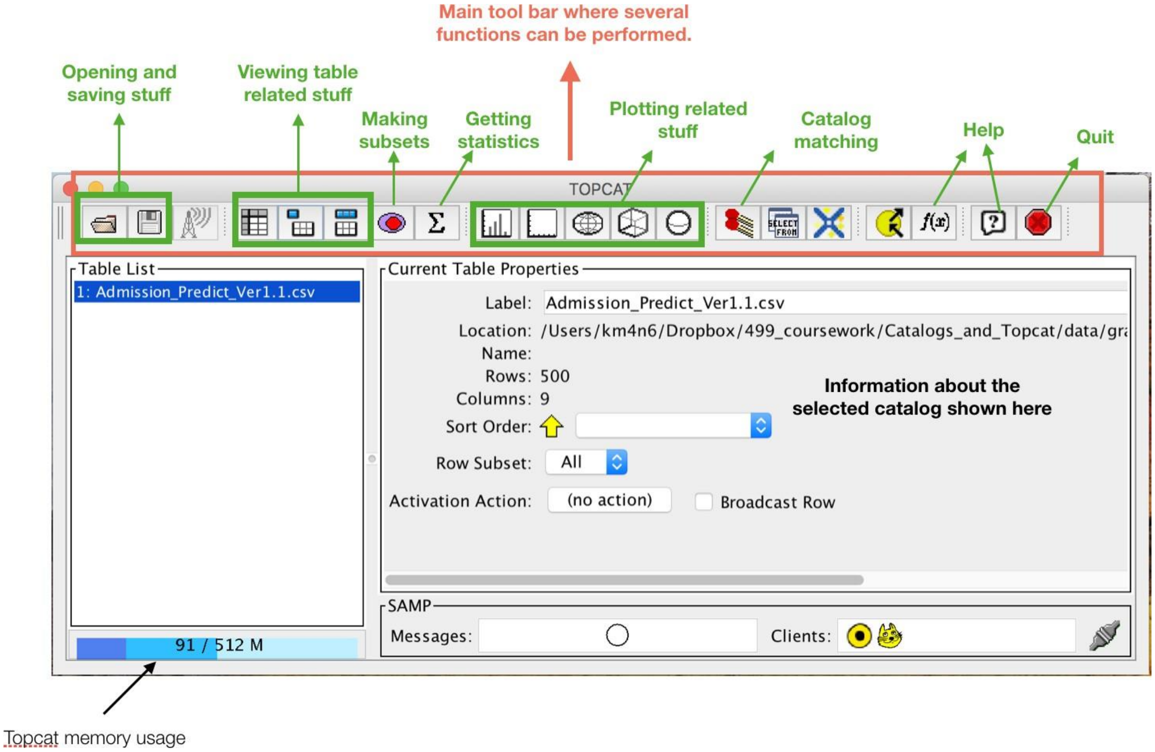


	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa

Figure 2: An example data table of basic quantitative properties of flowers by species (taken from the internet).

1.2 Tabular Data Operations

Figure 3 provides an overview of TOPCAT's first level of functionality given by a series of clickable buttons with intuitive icons. The remainder of this section will provide you opportunities to practice using many of these TOPCAT functions.



1: Admission_Predict_Ver1.1.csv

Label: Admission_Predict_Ver1.1.csv
Location: /Users/km4n6/Dropbox/499_coursework/Catalogs_and_Topcat/data/gra
Name:
Rows: 500
Columns: 9
Sort Order:
Row Subset: All
Activation Action: (no action)
□ Broadcast Row

Messages: ○
Clients:

Figure 3: A brief description of each first level function in the main GUI window.

Learning Activity 1: Opening and Viewing Data Tables Using TOPCAT

For this practice (not graded) exercise, apply the basic steps below for opening electronic data tables with TOPCAT and viewing their contents in tabular form to the **example-datatable.csv** file provided with this tutorial.

1. Open a TOPCAT session
2. Open a data table file by clicking the *opening folder* icon. This opens a Load New Table window with various functions. Select Filestore Browser to navigate to the folder (directory) containing a data table file you wish to explore, then select the appropriate Table Format from the drop-down list and click OK.
    a. Note that you can open multiple tables for analysis; each will be listed in the Table List.
3. Basic information (such as number of rows and columns) about the loaded table will display in the main TOPCAT GUI window as shown in Figure 3.
4. Click the first (left) icon in 'viewing table related stuff' options to open the Table Browser window for viewing the table contents in row-column format (as in excel).
5. If you want to view the table in a specific sorted order, use the Sort Order option in the main GUI window to select which column and which row sorting direction (increasing vs decreasing) you desire. You can exit pop-up windows by clicking the X icon.
6. If you want to rearrange the order of the viewed columns, simply click on a column header in the Table Browser and move it left or right.
7. The middle icon opens the Table Parameters window which provides a quick way to see the full system path to your data table. We will not use this.
8. The third (right) icon opens the Table Columns window which has many important uses.
    a. The most basic one is being able to visualize the different columns and their names.
    b. Use the blue check boxes to select and unselect columns shown in the table view.
    c. The column names and their corresponding $ID can be used when performing mathematical and conditional logical operations on column contents. For example, if "Col1" is the column name of the first column, then "Col1 > 0" is equivalent to $1 > 0.
    d. The Class refers to the data type of the information stored in the column, e.g., String, Float (floating-point variables), and Double (double-precision floating-point variables) are common data types. Use google or your favorite gen-AI tool to understand these and data types as you encounter them.
    e. The *green* + icon is an advanced function for adding new columns, e.g., using mathematical operations on existing columns (as in Learning Activity 2, below).
9. The *Sigma* icon provides Raw Statistics for each viewable column, i.e., each column with a blue check mark.

1.3 Mathematical Operation Functions

Example syntax for performing mathematical operations on column information include:

- Adding two columns: Col1 + Col2 is the same as $1 + $2 based on column $ID in Table Columns window
- Multiplying two columns: $1 * $2
- Adding two columns and dividing by a third: ($2 + $5) / $3

    - Note: parentheses are used to alter and bypass the standard order of operations[3]. E.g., $2 + $5 / $3 will first do Col5 divided by Col3 and then add Col2, in contrast to ($2+$5)/$3 which first sums Cols2&5 and then divides the total by Col3.
- Raising Col4 values to the 1.5 power: pow($4, 1.5)
- Taking log base-10 of Col11 values: log10($11)

This type of syntax for mathematical operations is like that used in many coding languages, such as Python, as we will learn in future tutorials. There are additional higher-order functions that you can explore by clicking the *f(x)* icon in the main GUI window and opening the Maths option.

Learning Activity 2: Manipulating and Creating New Data Tables Using TOPCAT
For this practice (not graded) exercise, apply the basic steps below to the **example-datatable.csv** file provided with this tutorial.

1. Open the Table Columns window and select the *green* + icon to add a new column based on a simple use of mathematical operation of your own design.
2. Explore adding additional new columns using different and more complex mathematical operations as you like.
3. Explore removing any columns you do not want to save, changing the order of columns, and the sorted order of rows as you like.
4. To save your newly designed table, click on the *floppy disk* icon in the main GUI window to open the Save Table(s) or Session window. Select Filestore Browser to name your new data table file and choose its format.
    a. Note, the Session option will save all your open tables and ongoing work as a ‘Session’ file so that you can quit TOPCAT and then reload your current work when you return to use this tool at a later time.

1.4 Data Table Subsets and Conditional Logical Operation Functions

The “Making subsets” icon in the main GUI window (see Fig. 3) opens a Row Subsets window. The leftmost *green* + opens a window for you to name and define a subset of the data table using mathematical and/or conditional logical operations that you enter under Expression.

Example Expression syntax for performing basic conditional logical operations include:

- Create a subset such that the Letter Grade column has the same string value: $1 == “A”
    - The size of the row subset and its relative fraction of the whole table will be given.
    - The rows belonging to this subset will be highlighted in the Table Browser.
    - To view only this subset, select it from the Row Subset menu in the main GUI window.
    - If you wish to save a single table subset to a separate data file, make sure to select the subset name first in the main GUI window before selecting the save icon.
- Create a subset such that the Letter Grade is not an A: $1 != “A”
- Create a subset such that HW#8 Float value scores are between 35 and 43: $6>=35 && $6<43
    - In this case between is defined as greater than or equal to 35 and less than 43.

[3] https://www.mathsisfun.com/operation-order-bodmas.html

    - && is the logical AND operator
- Create a subset such that HW#8 scores are either >45 or <30: $6>45 || $6<30
    - || is the logical OR operator

This type of syntax for conditional logical operations is like that used in many coding languages, such as Python, as we will learn in future tutorials. "Conditional logical operators are used in decision-making statements, which determine the path of execution based on the condition specified as a combination of multiple Boolean expressions." (www.techopedia.com).

Learning Activity 3: Selecting and Defining Row Subsets Using TOPCAT
For this practice (not graded) exercise, apply the four example subset operations in Section 1.4 to the **example-datatable.csv** file.

2. Introduction to Data Visualizations Using TOPCAT

Data visualization is an essential skill for anyone working with data. In this section, we will explore the basics of graphing one-dimensional and two-dimensional (hereafter, 1D and 2D) data distributions using TOPCAT. For any data table you load into a TOPCAT, you can make different types of graphical visualizations (aka, graphs, plots and figures) based on column-wise data. As shown in Figure 3, there are five icons[4] indicating each type of graphing function. From left to right:

1. Histogram Plot: produce a bar graph of any single column quantity in your data table.
2. Plane Plot: produce a 2D graph of an x-axis quantity vs a y-axis quantity.
3. Sky Plot: produce a 2D spherical coordinate system graph, e.g., in astronomy we use several different sky coordinate systems – we will return to this in a future tutorial.
4. Cube Plot: produce a 3D graph in x,y,z coordinate system.
5. Sphere Plot: produce a 3D graph in polar coordinate system.

By the end of this introductory data analytics tutorial, you will have a solid understanding of how to create and customize the commonplace and basic types of figures: 1D Histograms and 2D Scatter (aka x-y) Plots.

2.1 One-Dimensional (Single Column) Histogram Graphs

A **histogram** is a graphical representation that shows how data values of a given column (parameter) are distributed over the parameter space (defined by the minimum and maximum values). It is a bar chart-like graph where the horizontal axis represents the intervals, also known as bins, and the vertical axis denotes the frequency or count of data values falling into each bin. In essence, a histogram provides a simple and effective visual summary of the distribution of data and helps you identify the shape of the data distribution (e.g., symmetric, skewed, multi-modal, etc.), statistical information such as the median and spread of a distribution, and any potential or statistical outliers in a 1D dataset. Figure 4 provides an overview of TOPCAT's first level of histogram functionality given by a series of clickable buttons with intuitive icons.

[4] Newer versions of TOPCAT have additional options!

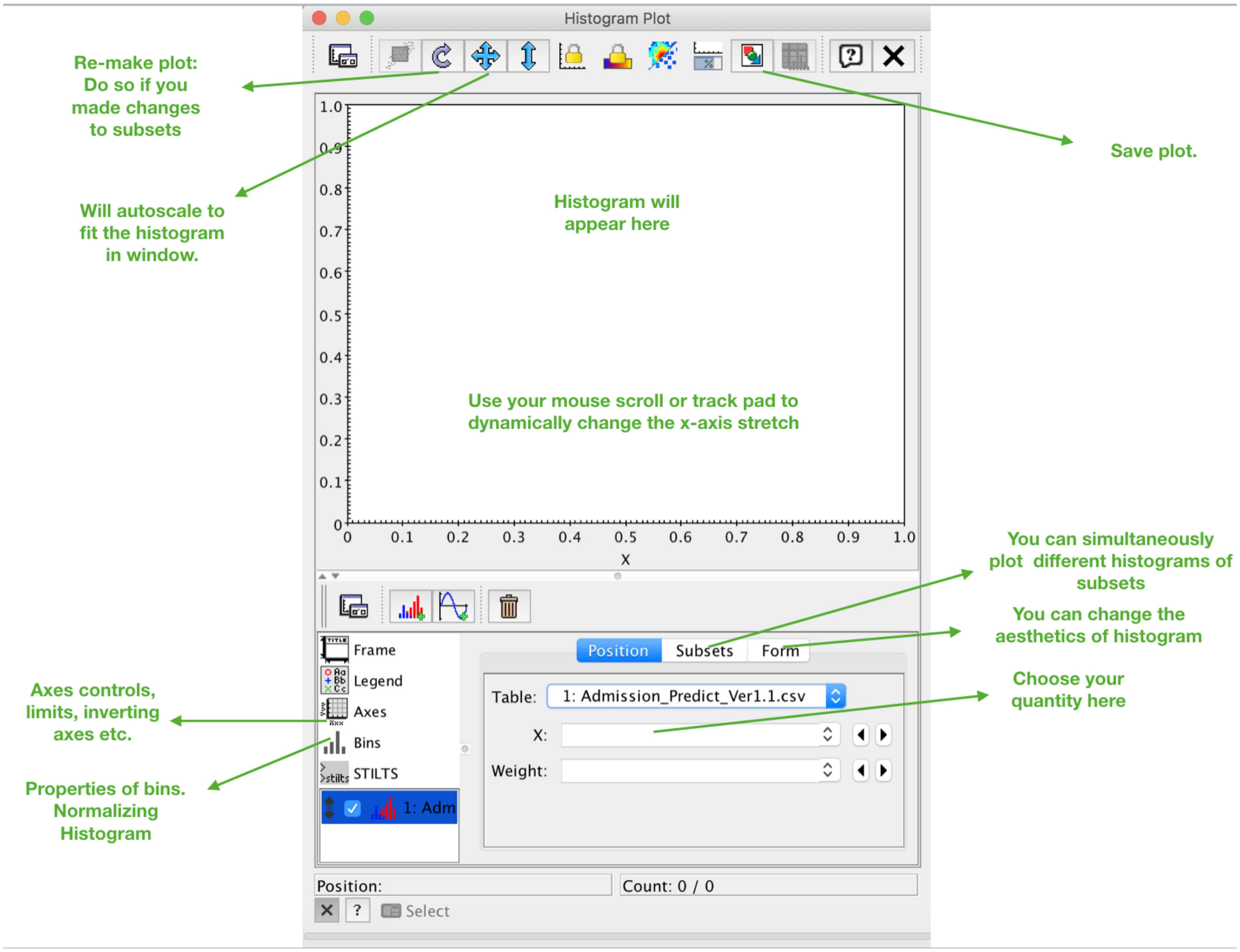


Table: 1: Admission_Predict_Ver1.1.csv

X:

Re-make plot: Do so if you made changes to subsets

Will autoscale to fit the histogram in window.

Histogram will appear here

Use your mouse scroll or track pad to dynamically change the x-axis stretch

Save plot.

You can simultaneously plot different histograms of subsets

You can change the aesthetics of histogram

Choose your quantity here

Axes controls, limits, inverting axes etc.

Properties of bins. Normalizing Histogram

Figure 4: A brief description of first level functions in the Histogram Plot window.

Learning Activity 4: Plotting Histograms Using TOPCAT

For this practice (not graded) exercise, apply the basic steps below to the **example-datatable.csv** file.

1. Click on the *histogram bar graph* icon to open a Histogram Plot window.
2. As shown in Figure 4, there are many histogram plotting functionality controls and display settings above and below where the graph will appear that allow the user to achieve the 1-dimensional data visualization they desire.
3. For a simple histogram plot of one column, select the column identifier from the X: drop-down menu found in the Position option in the lower control panel.
4. Explore changing the appearance of this histogram using the Form option functionalities, as well as other controls for axes, bins, etc.
5. To overlay a second histogram defined by a subset of the column selected in Position X:, select the Subset option and explore this functionality.

6. Now you are ready to try to reproduce an exact match of the histogram example in Figure 5. In this case, the subset “B students” was defined by $2>=80 && $2<90.

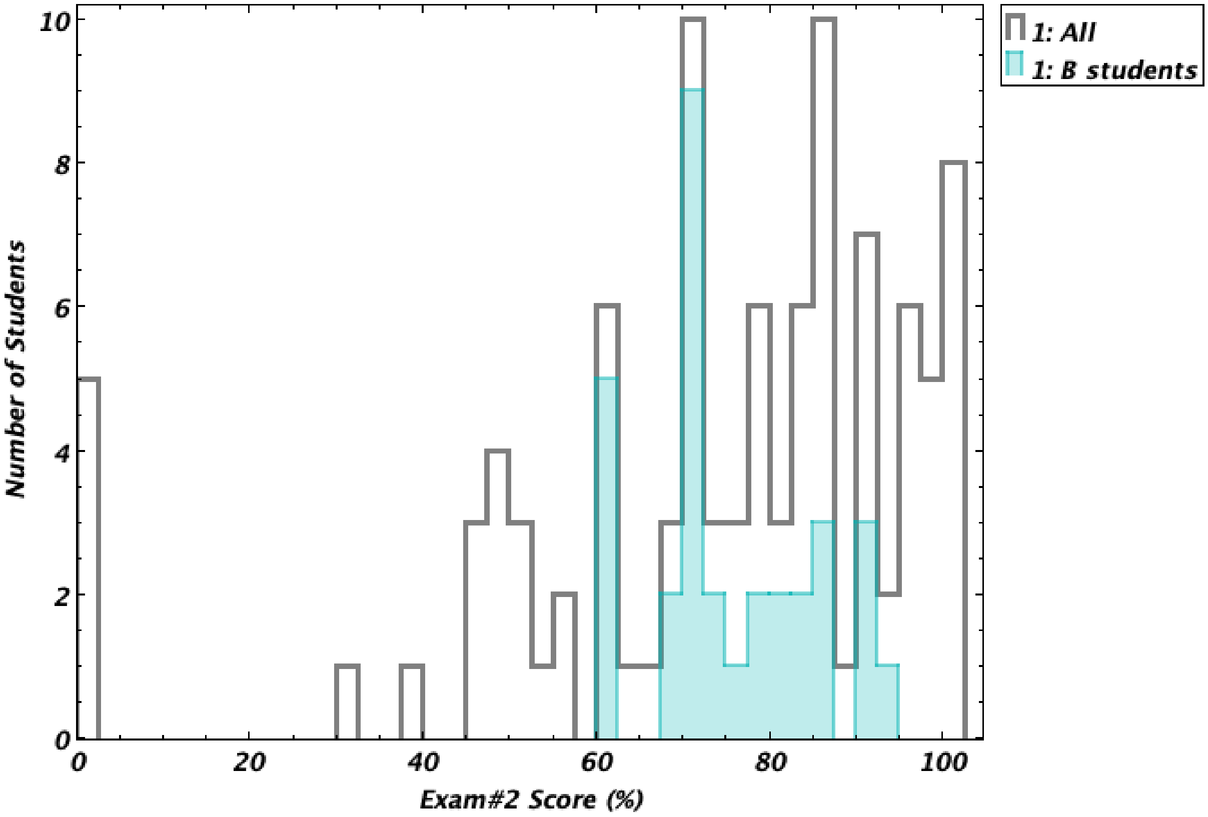


Figure 5: Learning Activity 4.

2.2 Two-Dimensional (2-Column) x-y Graphs

Often called a scatter plot or, simply, x-y plot, these types of figures are an essential and very common tool for analyzing relationships between the data values of two given parameters. Like Figure 4, there are similar x-y plotting functionality and display settings in the Plane Plot window that the user can choose from to achieve the 2D data visualization they desire.

Learning Activity 5: Making 2D x-y Plots Using TOPCAT

For this practice (not graded) exercise, once again, apply the basic steps below to the **example-datatable.csv** file.

1. Click on the *x vs y graph* icon to open a Plane Plot window.
2. For a simple x-y plot of the values from two table columns, select the column identifiers from the X: and Y: drop-down menus found in the Position option in the lower control panel.
3. Explore changing the appearance of this plot using the Form option functionalities, as well as other controls for axes, etc.
4. Explore overplotting one or more subsets in the same x-y space.
5. Last, try to reproduce an exact match of the x-y plot example in Figure 6.

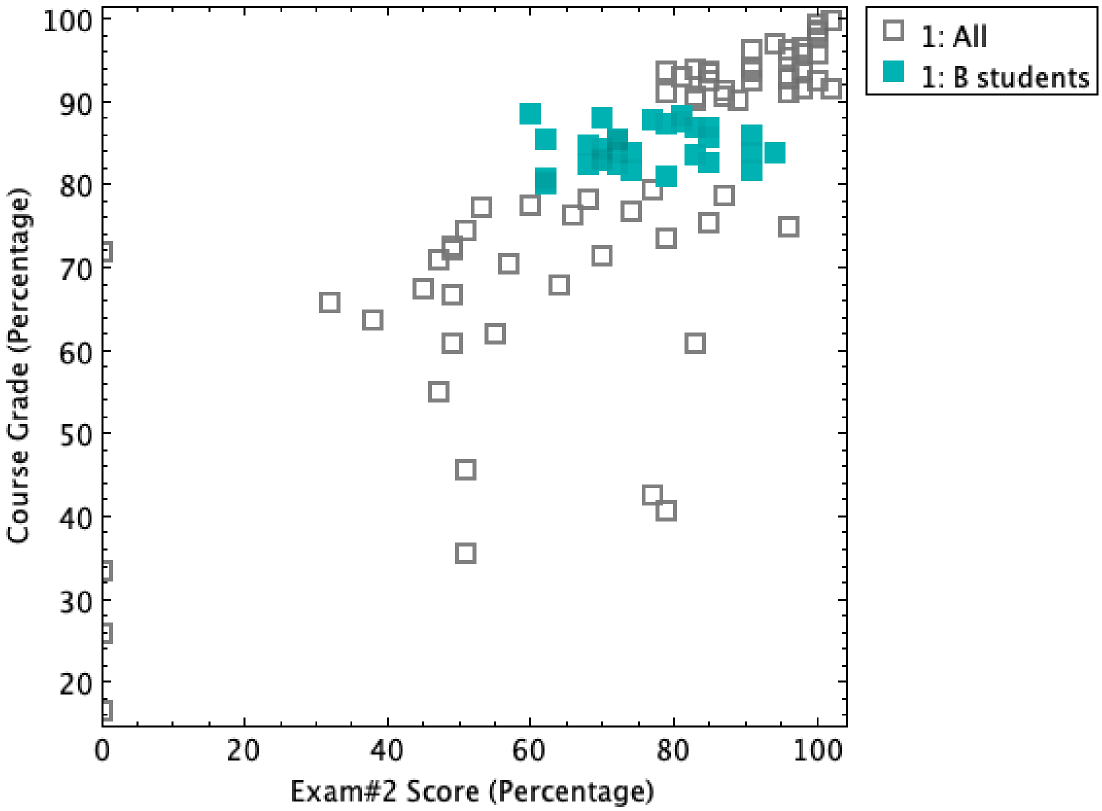


Figure 6: Learning Activity 5.

2.3 Professional Figure (and Table) Captions

A clear, informational caption is an essential aspect of creating a professional figure or table. Writing good captions requires practice. As shown in the example figure caption in Figure 7, a professional caption starts with a summary statement that describes what is plotted in the figure. Next, the caption includes all the details of the plot in a succinct fashion. A figure caption, in combination with the figure itself, should be as self-contained as possible to make it easy for the reader to understand the plot in detail. Usually, the figure caption does not need to include a statement of the interpretation or result of the figure because that will be described in the text of the publication or report itself, as will the description of the data collection and the rationale for the data analysis and visualization. These guidelines apply similarly for table captions.

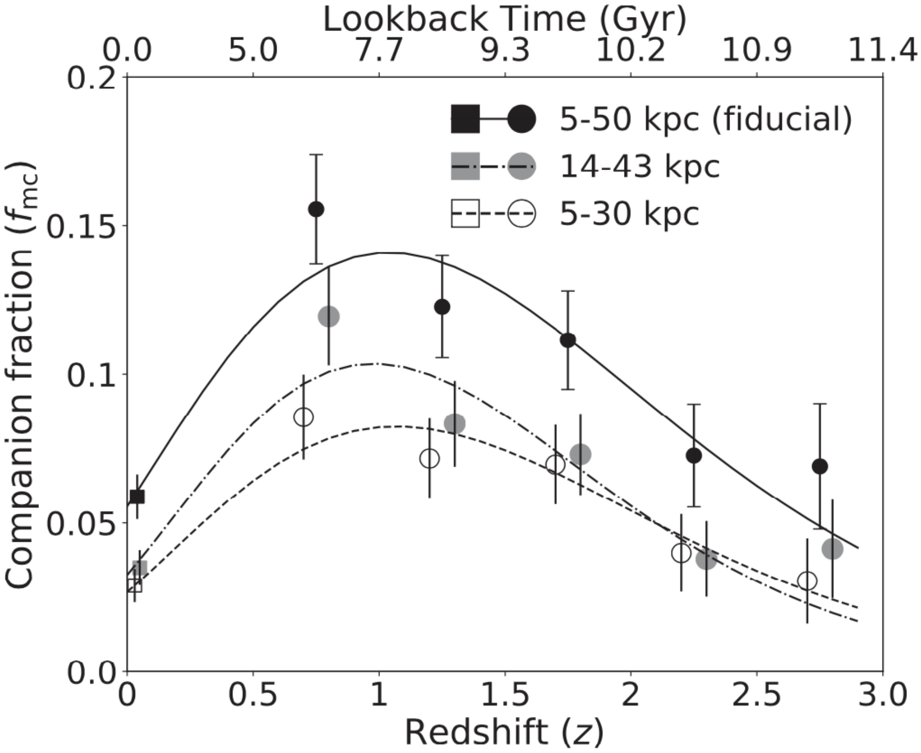


Figure 6. Comparison of the redshift evolution of the major companion fractions based on three projected separation criteria: $R_{\text{proj}} = 5$–50 kpc (fiducial; black symbols, solid line), $R_{\text{proj}} = 14$–43 kpc (grey symbols, dashed line), and $R_{\text{proj}} = 5$–25 kpc (open symbols, dot–dashed line). Best-fitting curves for each $f_{\text{mc}}(z)$ spanning $0 < z < 3$ are a modified power-law exponential (equation 7; see Section 3.3.2 for details). The low-redshift fractions are from the SDSS (squares) and the $z > 0.5$ fractions are from CANDELS (circles). The error bar on each f_{mc} data point represents 95 per cent binomial confidence limit. The non-fiducial f_{mc} data points are offset by a small amount within each redshift bin for clarity.

Figure 7: An example professional figure and caption published as Fig. 6 in Mantha, McIntosh et al. 2018[5].

3. Demonstrate Your Tutorial Knowledge and Upskilling

In this concluding section, you will demonstrate (and be graded on) your level of competency in the practical application of the tutorial concepts, data, and skills to realistic critical-thinking and problem-solving scenarios. Please review the R&D Skills Tutorial Submissions: Format Expectations as well as the Grading Rubric (in the syllabus) on the course canvas page to be clear on general expectations for achieving solid competencies.

[5] https://academic.oup.com/mnras/article/475/2/1549/4768277

Here, you are tasked with analyzing the generic course student outcomes data in the **example-datatable.csv** file. When working with data tables, remember to approach them with a critical eye and with intellectual curiosity toward gaining a thorough understanding of their informational content. Common essential questions include “what is in the table?”, “how is the table organized?”, “how was the tabulated information collected or derived?”, and “what can I learn from the table?”, for example, “what meaningful trends are found in the data?”. Developing these critical thinking skills and proficiency with tools to answer these questions is important for any and all data-intensive tasks, projects, and careers.

Assessment Task 1: Demonstrate Basic Table Manipulations Competencies Utilizing TOPCAT

The objective of this graded task is to give you the opportunity to demonstrate your level of competency with SLOs 1, 2 & 5. To achieve Satisfactory or better competencies on this task assumes that you first completed the related tutorial preparation and learning activities. ECT for this task is 1.5 hours.

Manipulate the generic student data table in TOPCAT to produce and save a new CSV file in which the table columns are rearranged in a more sensible organizational structure, two new columns have been added (one for average homework score and one for average exam score), and the table is sorted by the Percent Grade column. Reload this new table in TOPCAT to produce screenshots that demonstrate your efforts and competency. Screenshots are figures to be included in your tutorial submission. As such, they should include a professional caption (see section 2.3).

Assessment Task 2: Demonstrate Basic Data Analytics Competencies Utilizing TOPCAT

The objective of this graded task is to give you the opportunity to demonstrate your level of competency with SLOs 2, 3 & 5. To achieve Satisfactory or better competencies on this task assumes that you first completed the related tutorial preparation and learning activities. ECT for this task is 3 hours.

Create a professional slide (or slides as needed) summarizing the contents and your own statistical analysis of the student data table collected from a generic course. Include in your slide a tabular summary of the details of your statistical analysis. I recommend using word or excel or a similar common application to craft this table of your results. An excellent statistical analysis includes minimum values, maximum values, means, and standard deviations for key student outcomes. You decide which outcomes are key and include a brief rationale for your decision(s) and a brief ‘what did I learn’ concluding statement that makes reference to your analysis table. An excellent summary table includes a professional caption (see section 2.3).

Assessment Task 3: Demonstrate Basic Data Visualizations Competencies Utilizing TOPCAT

The objective of this graded task is to give you the opportunity to demonstrate your level of competency with SLOs 3, 4 & 5. To achieve Satisfactory or better competencies on this task assumes that you first completed the related tutorial preparation and learning activities. ECT for this task is 3 hours.

Add a slide (or slides as needed) that showcases one histogram figure visualizing one student outcome parameter (your choice; must be different than Fig. 5) from your analysis of the data. Include a clearly defined and labeled subset. An excellent figure will include clear and easily readable axis labels, as well as a professional caption. Include in your submission a brief statement justifying your visual analysis decisions and a brief ‘what did I learn’ concluding statement that makes reference to your figure. Repeat this task for one x-y figure visualizing two student outcome parameters (your choice; at least one axis must be different than in Fig. 6).