Lab 2

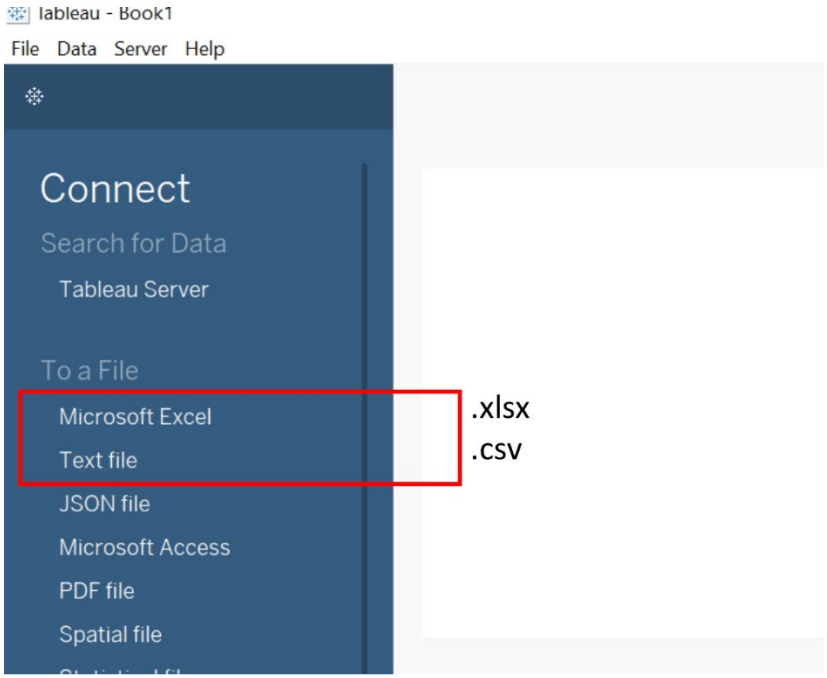
Learning outcomes:

1. Able to generate and interpret box-plot
2. Able to generate and interpret histogram
3. Able to generate and interpret Scatter plot

Exercise 1 – Introduction to Tableau

Launch Tableau.

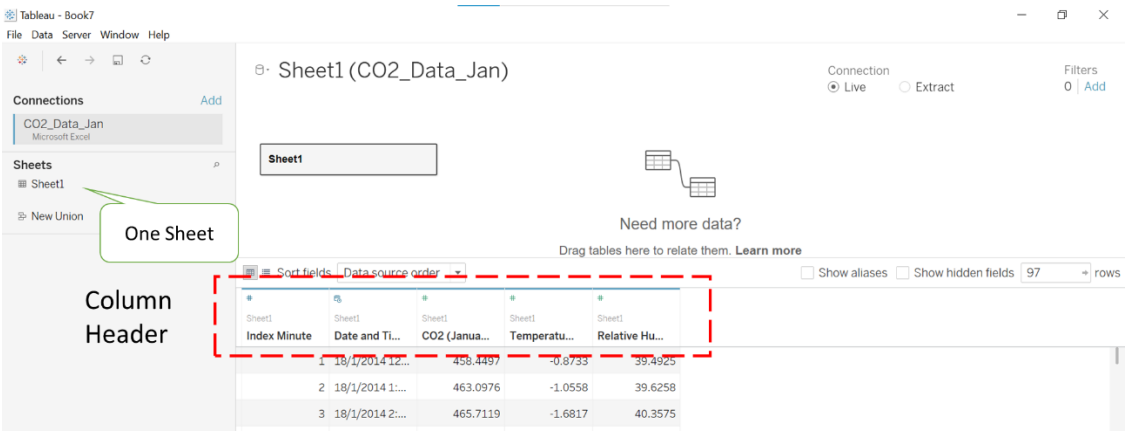
Read File



Depend on the type of file that you want to read.

For this exercise, we are going to read ***CO2_data_Jan.xlsx***

Hence, select "Microsoft Excel", and browse for the data.



This is the first page after loading the file.

The data is displayed with rows and columns. Each column represents a variable.

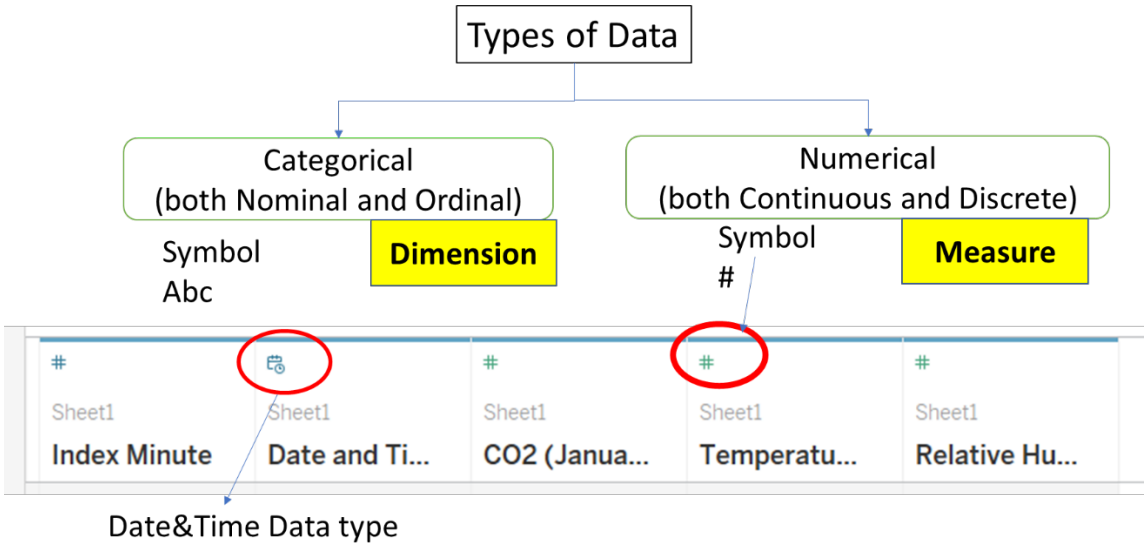
Each row is an observation. In this case, each row is a time stamp.

Data arranged in rows and columns is called **Structured**.

There are 3 Numerical variables, one date&time data and Row Index. There is no categorical type of data.

*(*categorical data refers to text input, example Gender)*

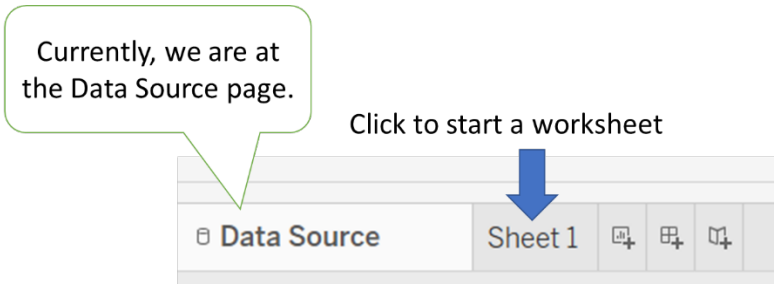
Note the special symbol for each variable in TABLEAU:



Dimension and Measure are the names of the data type used by TABLEAU.

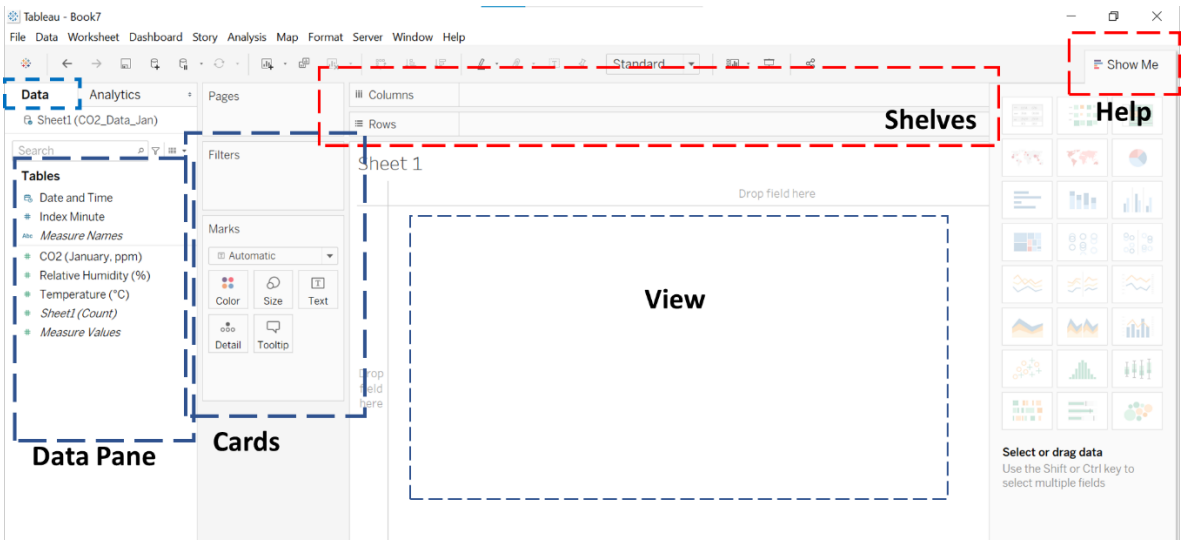
Dimension – represented by symbol **Abc.** For categorical (both nominal and ordinal) types of data

Measure – represented by symbol **#.** For Numerical (both continuous and discrete)



Proceed by starting a new worksheet.

This is the layout in a Sheet.



Note in TABLEAU, there is no workflow. It is just plotting a chart.

Column – can think it as the variable for the X-axis

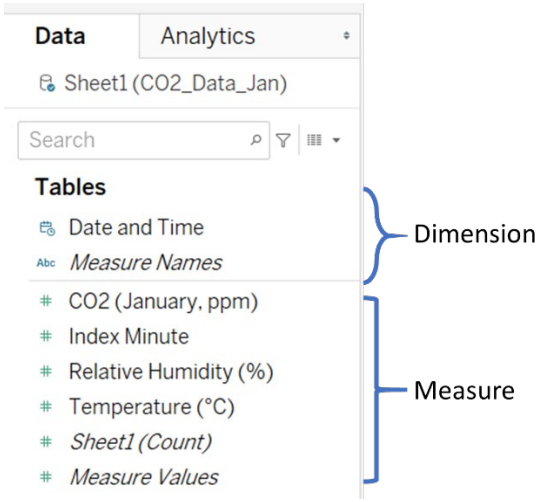
Row – can think it as the values assign to the Y-axis

You can drag and drop any variables in the "Data Pane" to the column and row to create a plot.

The options in the "Cards" pane, are display options. We will use them when we need it.

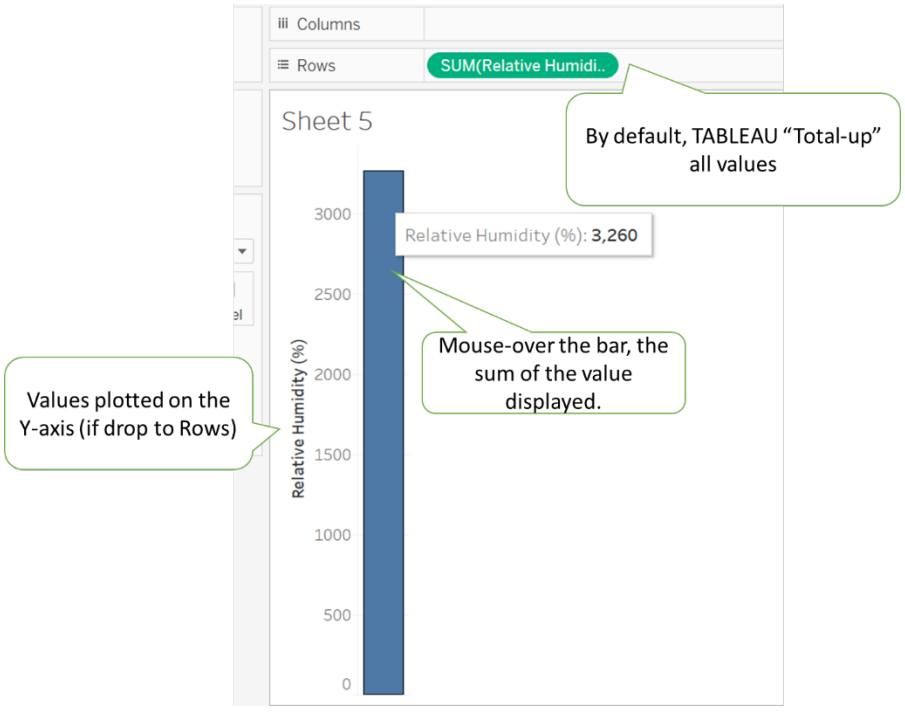
It is important to understand the arrangement in the Data Pane.

Data Pane

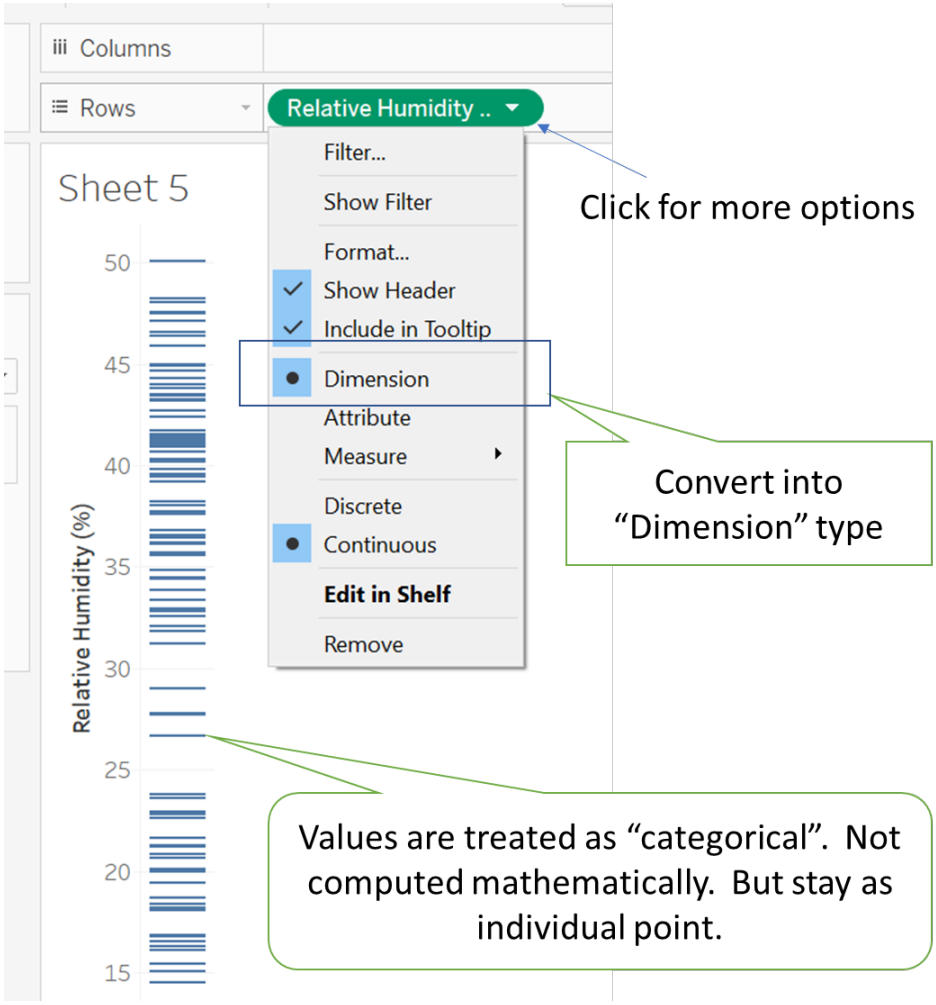


If you drag data from "Dimension" to columns/rows shelve, each value is plotted as individual mark on the axis. In this data set, there is only the special date&time data, no categorical (or you can think it as qualitative data).

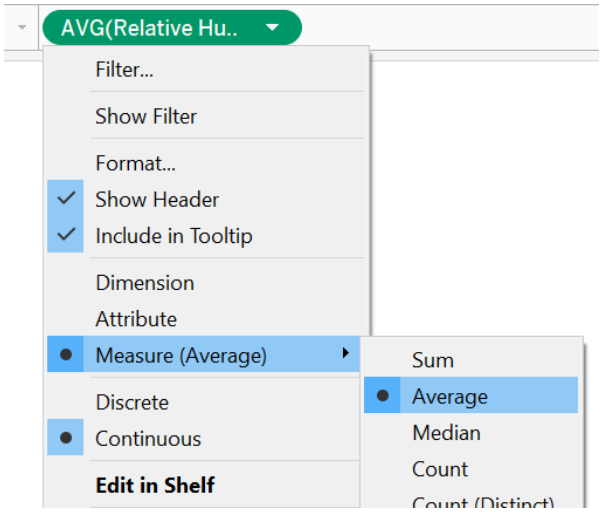
If you drag any data from the "Measure" to columns/rows shelves, values will be plotted along the X or Y-axis (as continuous reading). And display a horizontal or vertical bar showing the SUM of values of that variable. Example:



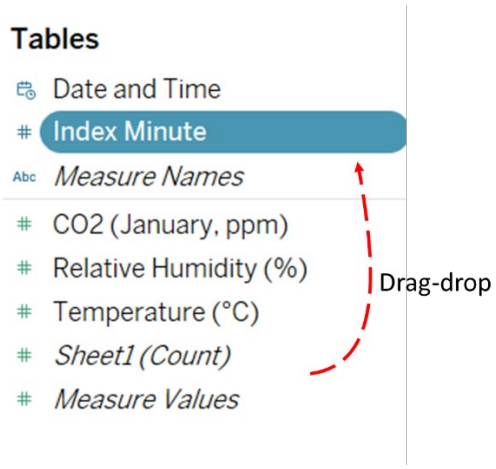
Something interesting about TABLEAU: Data type is easily converted to "Dimension"



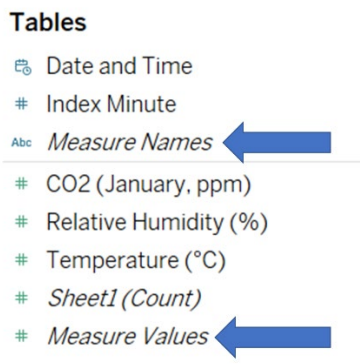
Let's convert back to "Measure", and find the "Average"



Note that "Index Minute" is in the Measure Field. As this data is just an indexing, not meant to be calculated. Drag the variable to the "Dimension" Field. (This is also a way to switch data type)



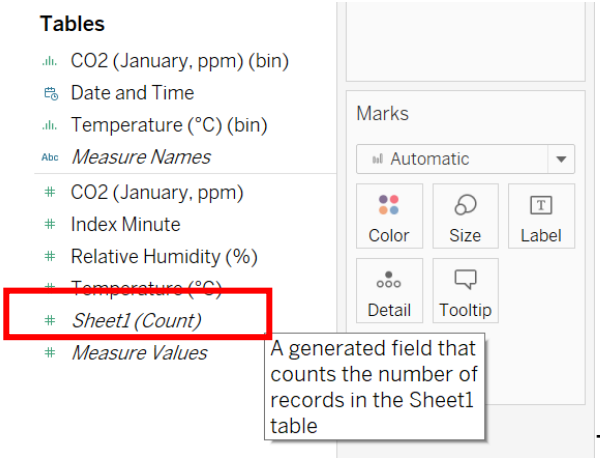
Note, in TABLEAU, variable is easily change to Dimension by drag-drop.



There are two additional variable called "**Measure Names**" and "**Measure Values**". This is generated by TABLEAU. It will show the name of the variable and the value when drop into the view.

If you drag any of these into the "Filter" card, it allows you to manually select. Very flexible!

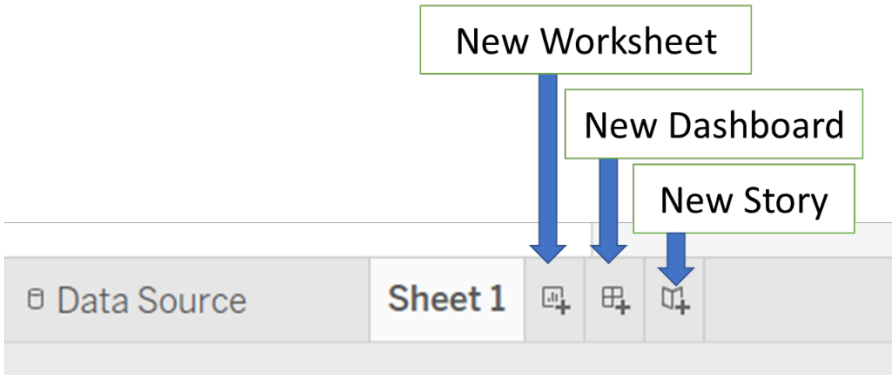
Other from Measure Names and Measure Values, the (Count) variable is also generated by Tableau.



This is used quite frequently.

In this course, we may not be able to cover all the features in TABLEAU. Hence, at times, you need to explore on your own.

There are 3 different options beside the "Data Source" at the lower left corner:



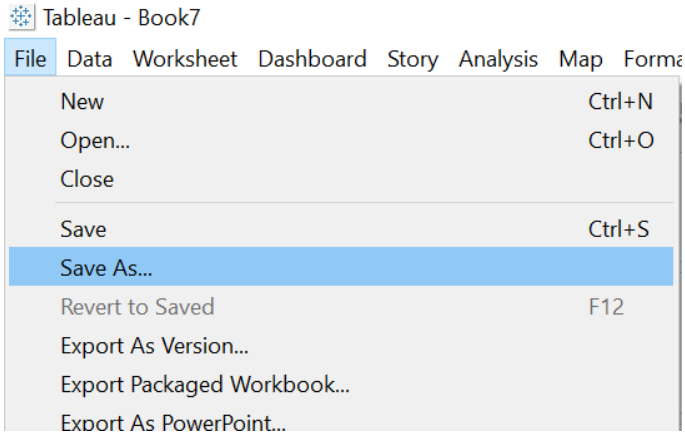
Most of the times, we add “New worksheet”. Every chart is a new worksheet.

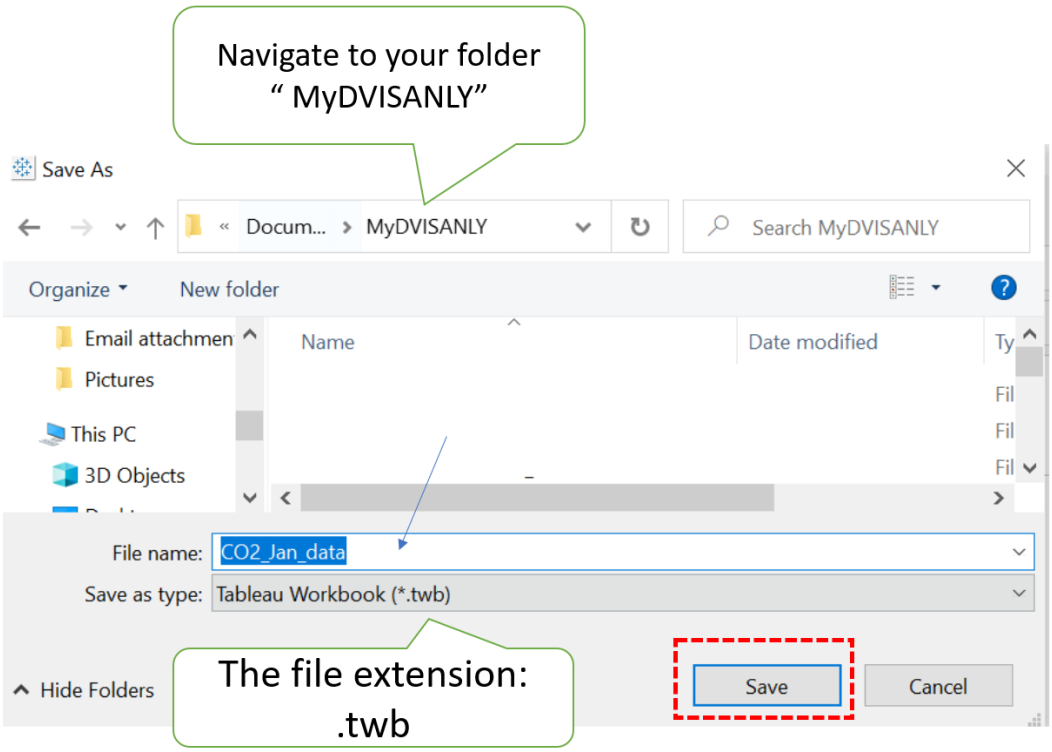
How to save a file?

Before we go further, let’s learn how to save our work in TABLEAU.

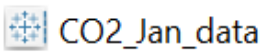
A file in TABLEAU is called a “workbook”. Currently “workbook” is not saved. When you saved a “workbook”, all “Sheets” are automatically saved.

Save the “workbook” as “*CO2_Jan_data*”.



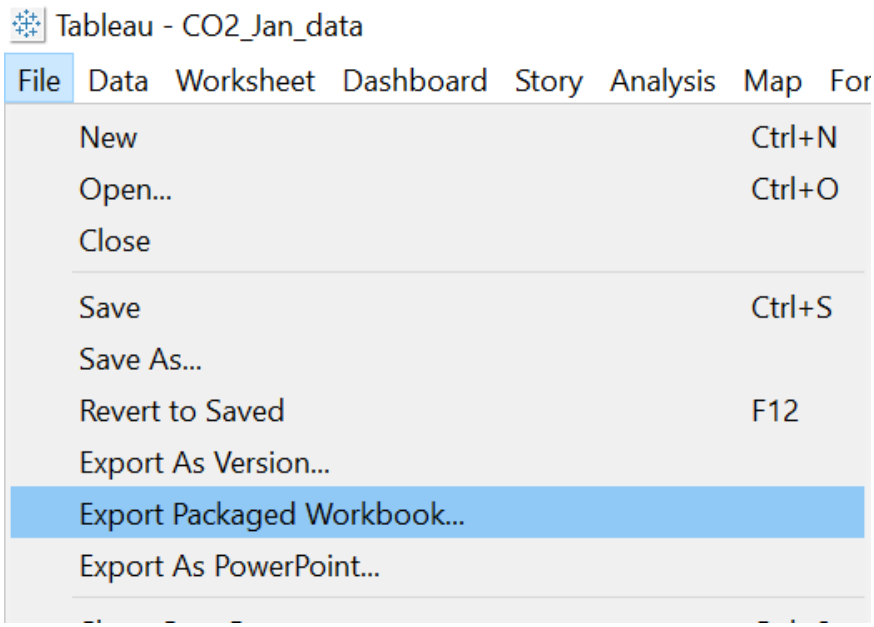


If you open your file explorer, you can find the workbook in your chosen folder:



Note that this workbook does not embed the data file. So, your data file and workbook must be in the same folder, when open this workbook.

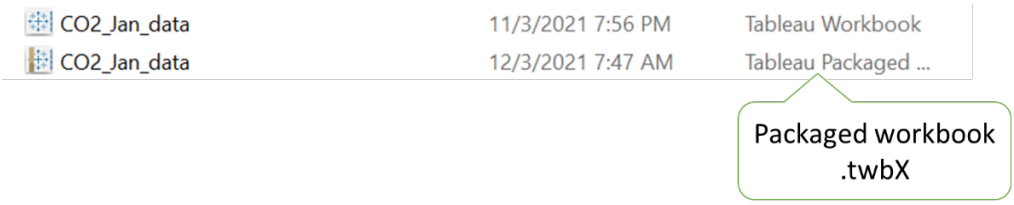
When you submit your work in LMS, you must EXPORT your workbook, so that the data files are embedded,



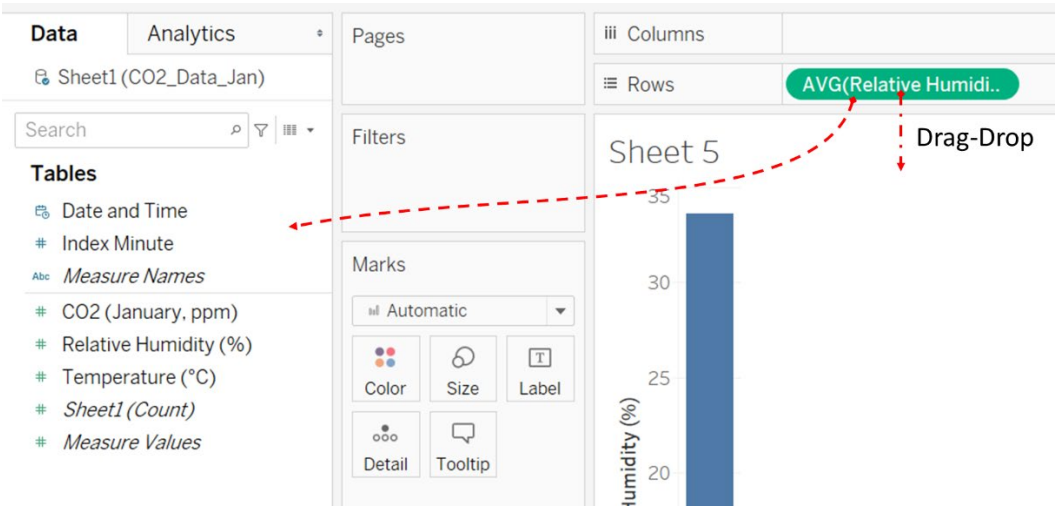
File > Export Packaged Workbook

Important to use this method to save your file for submission.

The file extension of a packaged workbook is **.twbx**

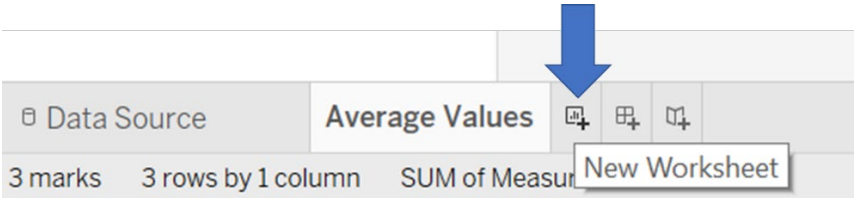


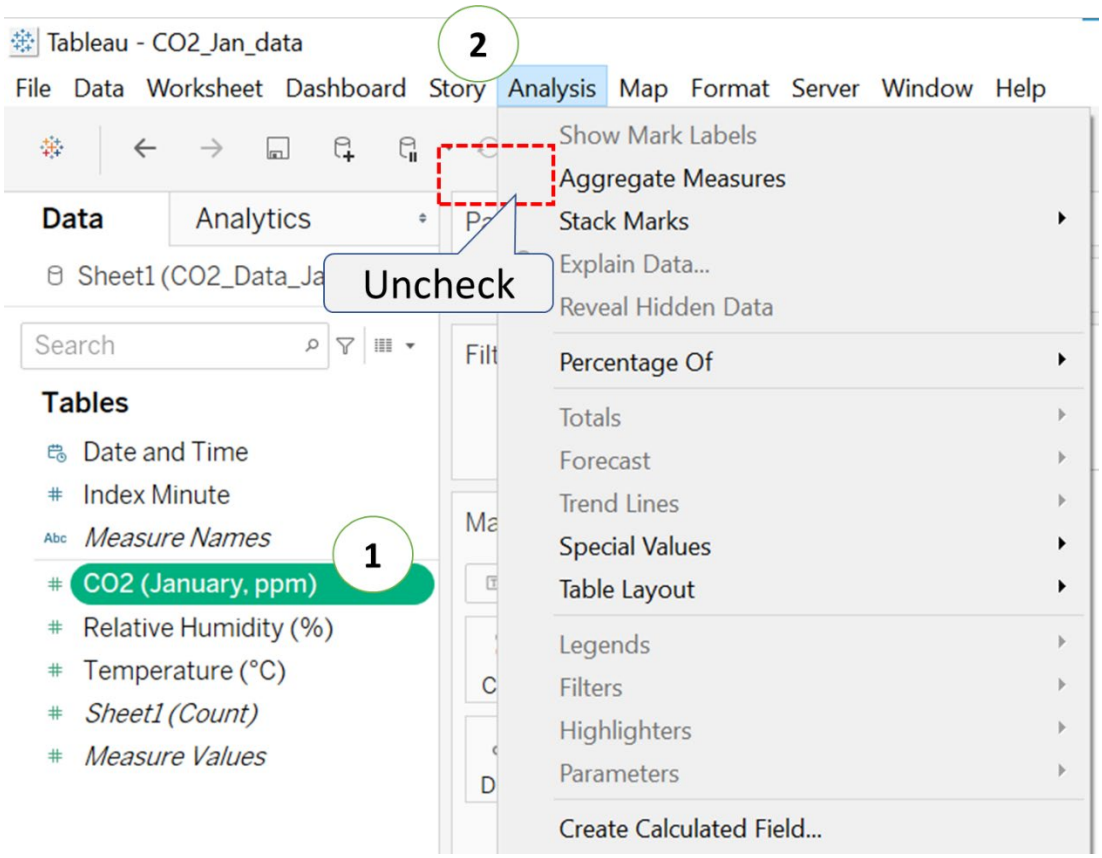
Let's clear the "shelves", if you have added anything to it, by dragging the "pills" and drop anywhere (into the View or Data Pane)



Exercise 2 - How to create boxplot?

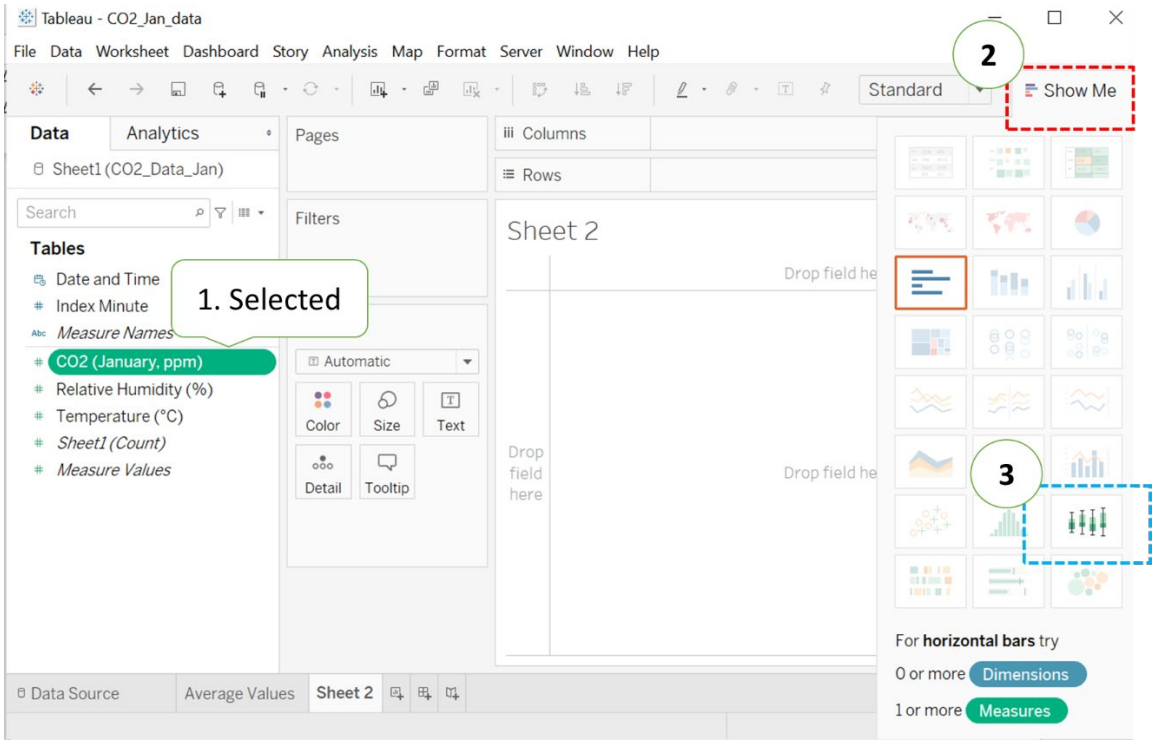
Let's start a new Worksheet.

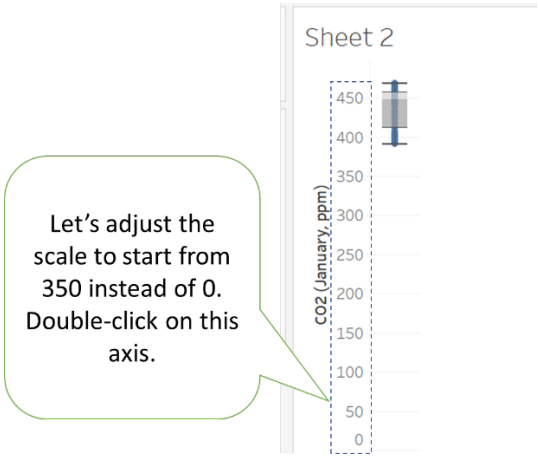


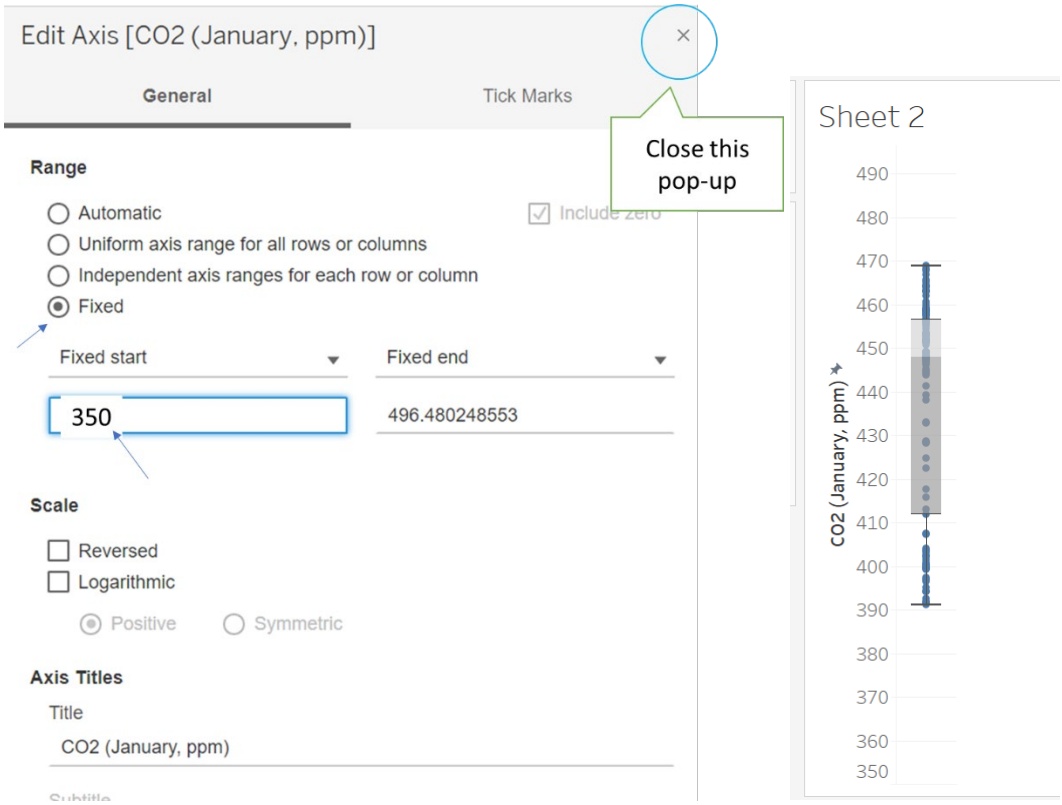


Very important: uncheck “Aggregate Measure”, if you want to create boxplot.

By default, TABLEAU will always “Aggregate” the values, like add-up and compute the Mean. To create boxplot, this option must “uncheck” so that all points will be plotted.

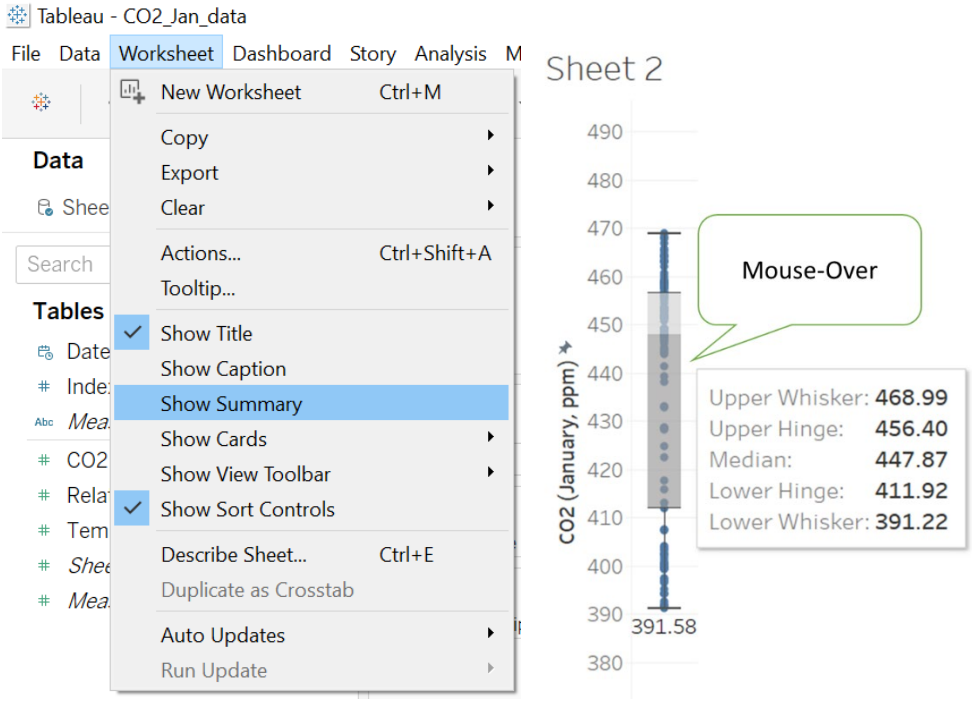






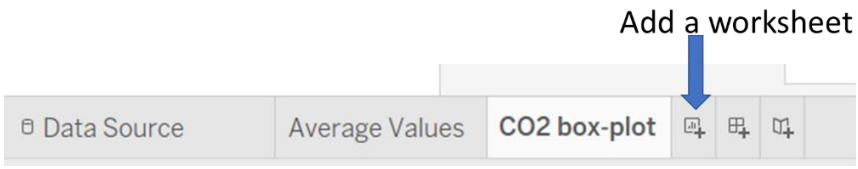
Mouse over the box-plot and you can read the statistics.

Or, top-bar menu: worksheet > Show Summary



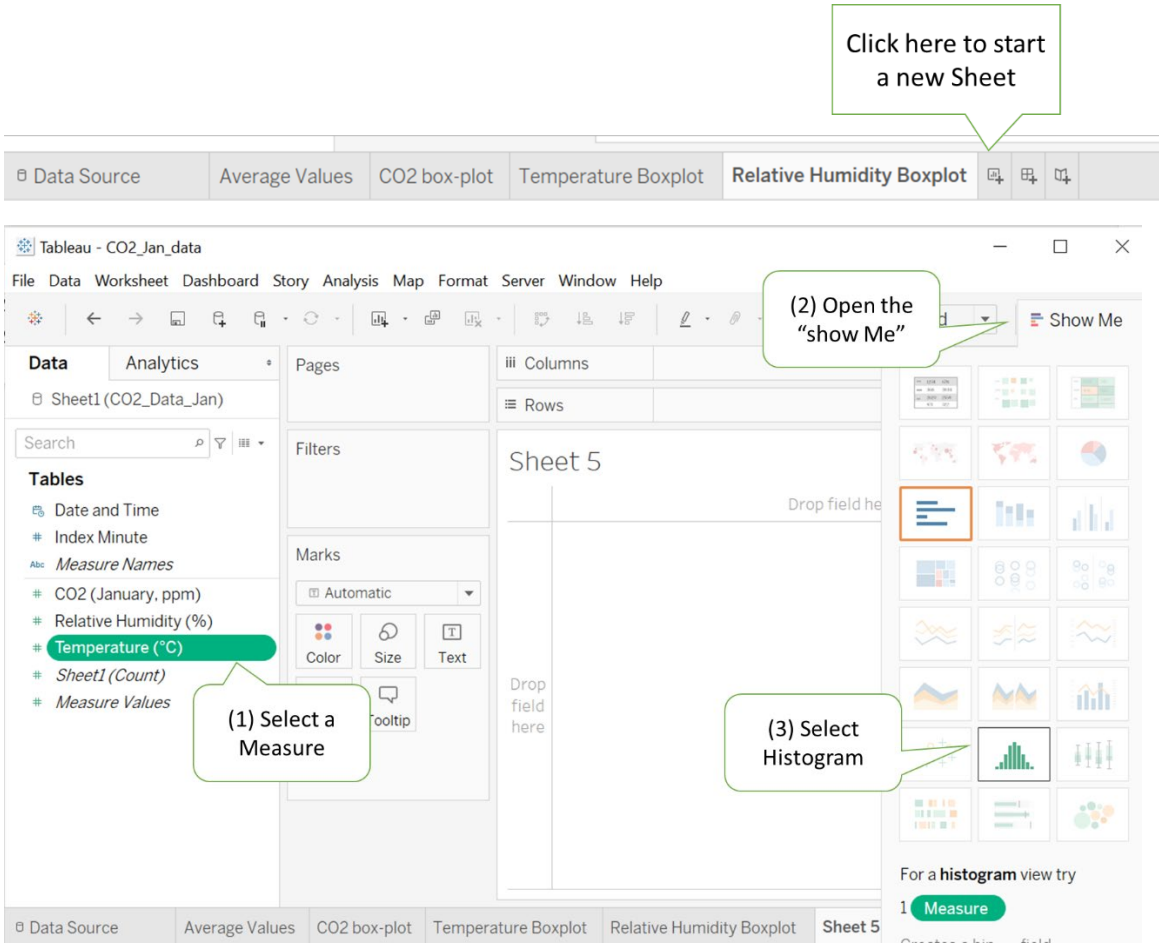
The Lower Hinge and Upper Hinge are the Lower Quartile (Q1) and Upper Quartile (Q3) respectively. The Lower Whisker is the minimum and Upper Whisker is the maximum.

Name the worksheet as “CO2 box-plot”

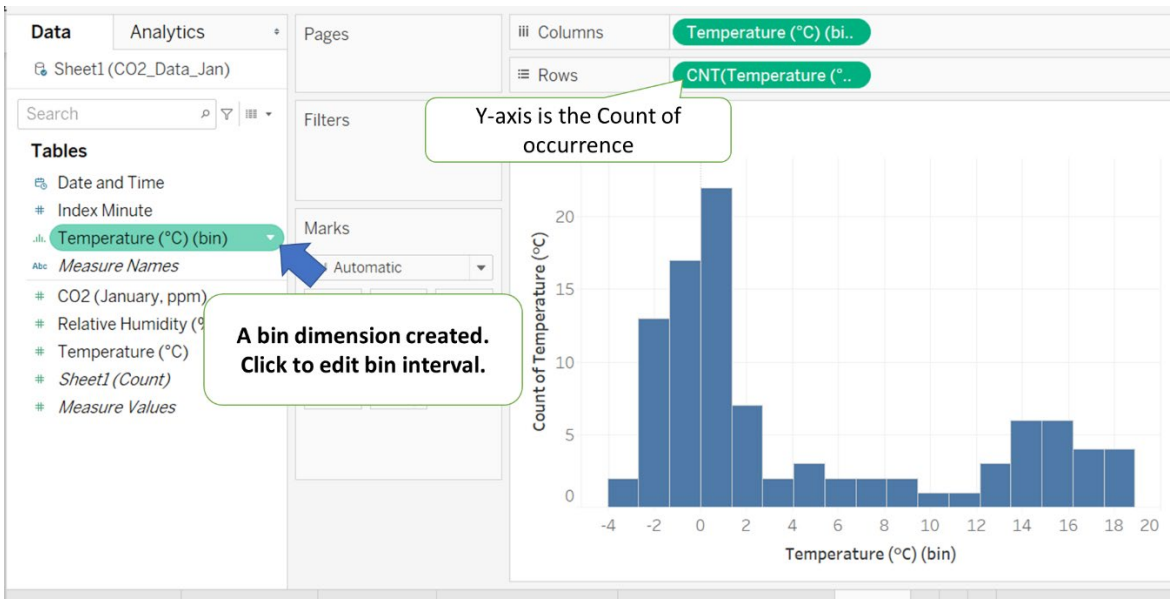


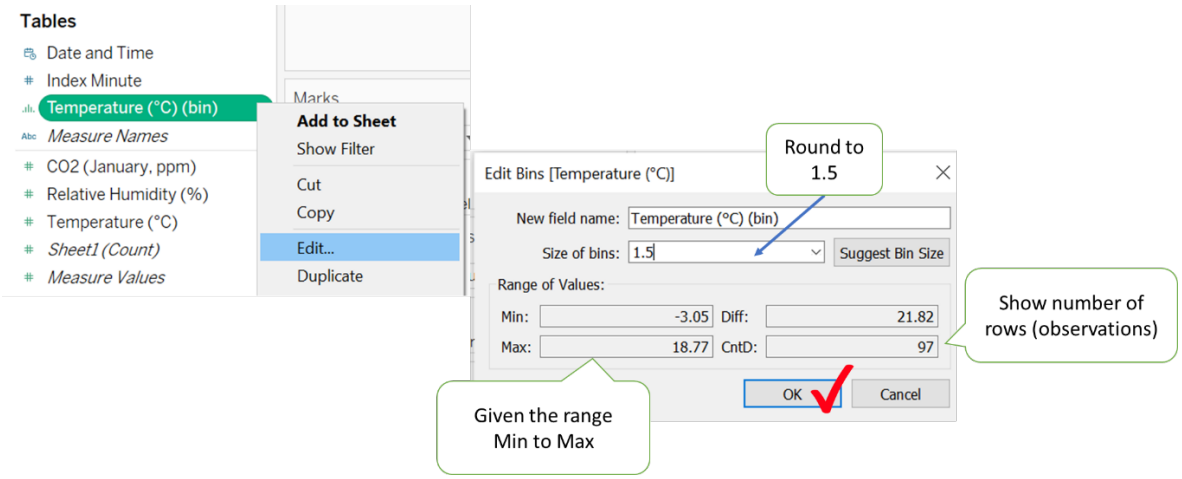
Exercise 3 – How to create a histogram?

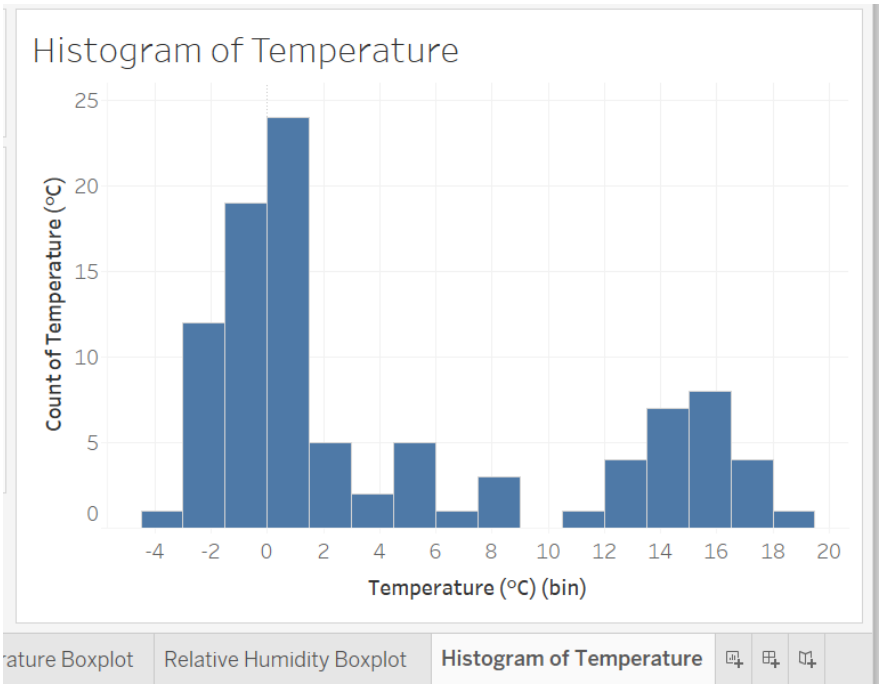
Start a new Sheet.



Click on the "Show Me" again to close the list.







Name the worksheet.

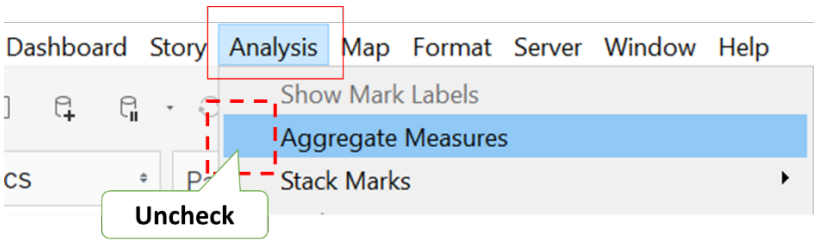
Insights:
There were 2 distinct distributions.
Temperature < 9, the mode is between 0 – 1.5, happening 24 times.
Temperature > 10, the mode is 15 – 16.5, happening 8 times.

Exercise 4 – Create Scatter Plot

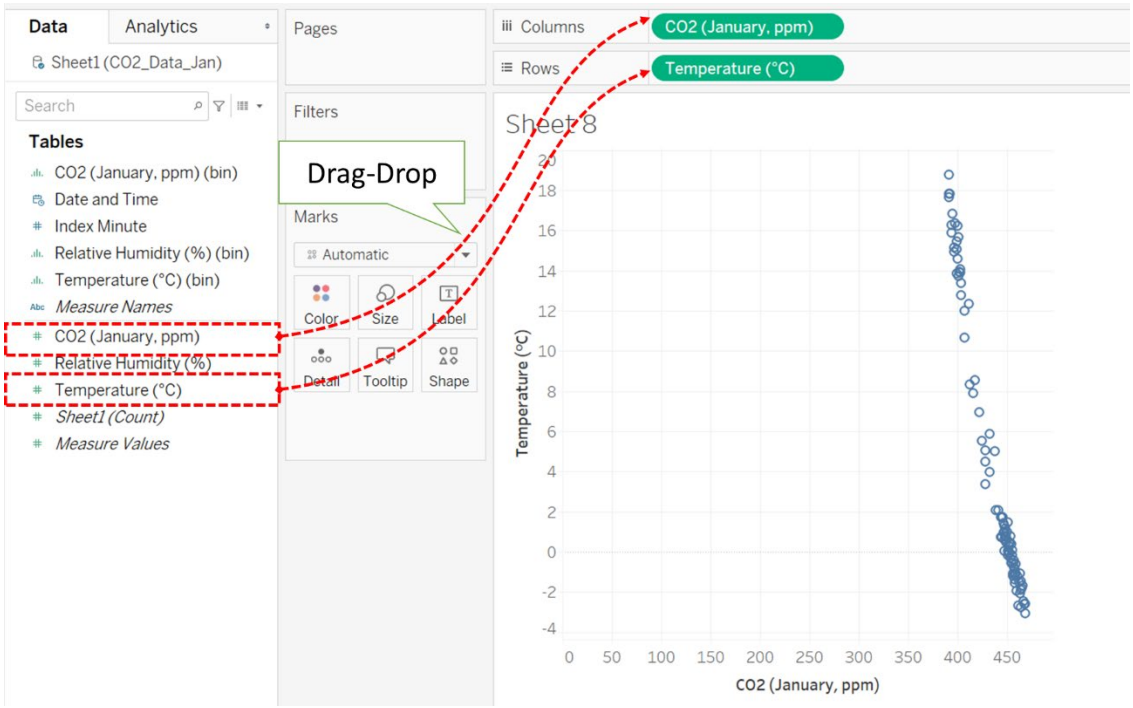
Histogram reveals some insights to the data distribution. What about the association between the variables?

Start a new Sheet.

"Aggregate" option must be OFF for scatter plot, so that each data point is displayed (else TABLEAU will "add" all values up to give a point only)

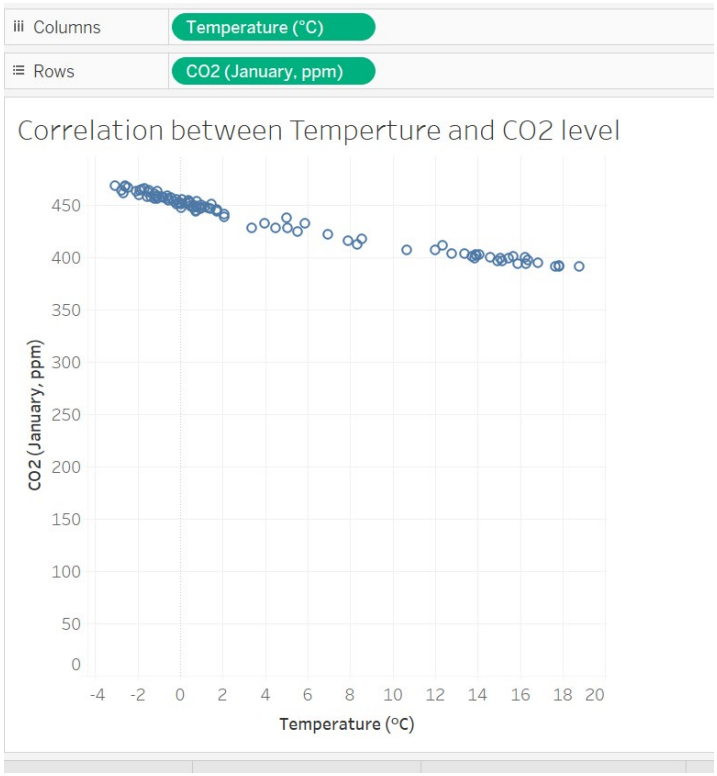


Let's compare CO2 level and Temperature:



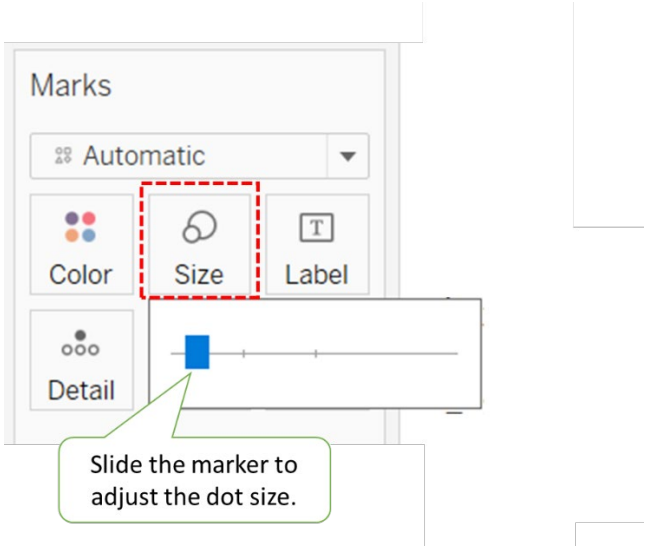
Drag the variable to either columns/rows shelves.

OR, the other way:



Both scatter plots show Strong negative correlation.

You may find the “dot” size too big. You can adjust the dot size smaller:



Review

Data arranged in rows and columns is called Structured.

Each row is an observation.

Each column is a variable.

Variable can be categorical or numerical.

Box plot, histogram and scatter plot can only be created with numerical data type.

You have learned how to:

Create a box plot

Create a histogram

Create a scatter plot

Remenber, you need to practice more (go through the steps a few times) to get familiar with the steps.

Lab 3

Learning Outcomes:

1. Able to normalise data by z-transformation in TABLEAU
2. Able to use KNIME to read CSV files
3. Able to explain the data structure and data types
4. Able to use Statistics node and interpret the Mean, Median, Standard Deviation, Max, and Min
5. Able to use CrossTab node to gather insights
6. Able to use Linear Correlation node to generate correlation matrix

Exercise 1 : Data Normalisation

Data set: *FictionIceCream_Sale.csv*

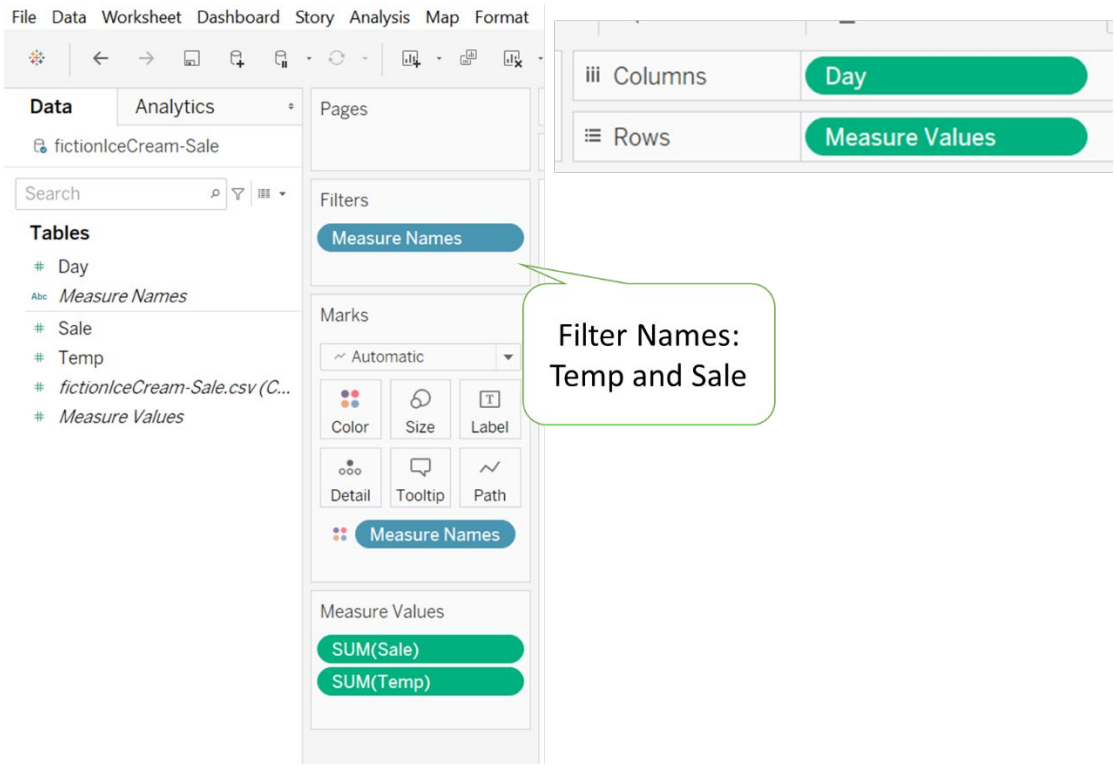
Question:
Does ice-cream sales increase when weather temperature decreases?

Download the data and copy it to your subject folder.

Launch Tableau

Read in the file as “text”.

Start a new sheet. Proceed to plot “Temp” and “Sale” on the same Y-axis (drop to rows), and “Day” on the X-axis (drop to Columns)



(Show me how)

What do you observe? Does the “Sale” increases, when the “Temp” increases?

To have meaningful comparison of 2 variables of different units, we must "normalise" to the same reference.

We will apply Z-transform method. Transform the data to have mean=0 and SD=1.

$$Z_i = \frac{x_i - \bar{x}}{SD}$$

Mean X: $\bar{x} = (\Sigma x_i) / n$

Standard Deviation

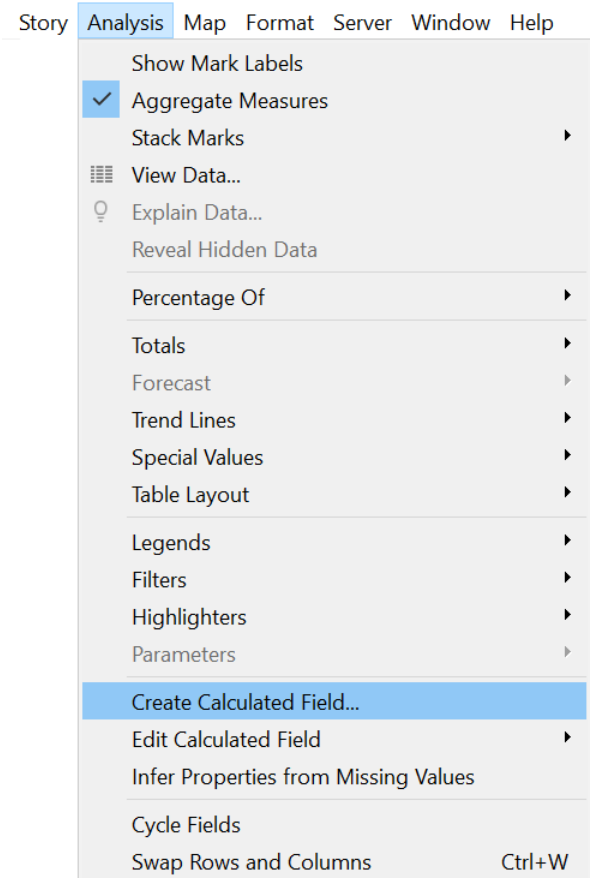
This mean, we will create two new fields; one to transform the "Temperature" and another to transform the "Sale".

One named "NormTemp", which is the new Z-transform of the "Temperature".

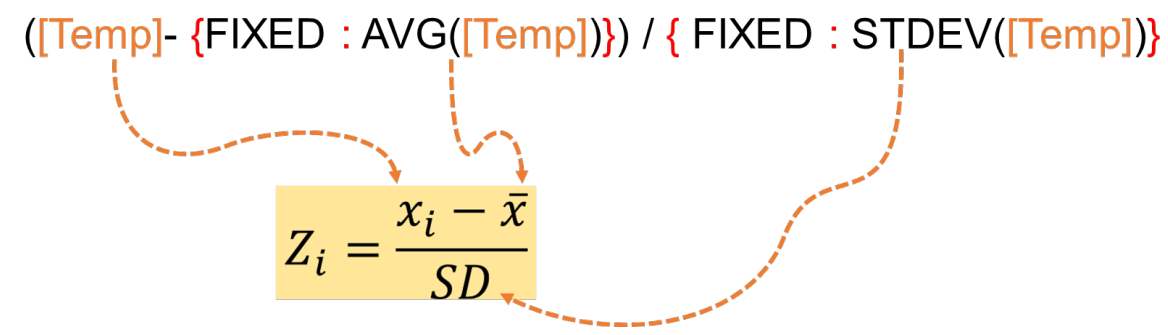
Another named "NormSale" which is the new Z-transform of the "Sale".

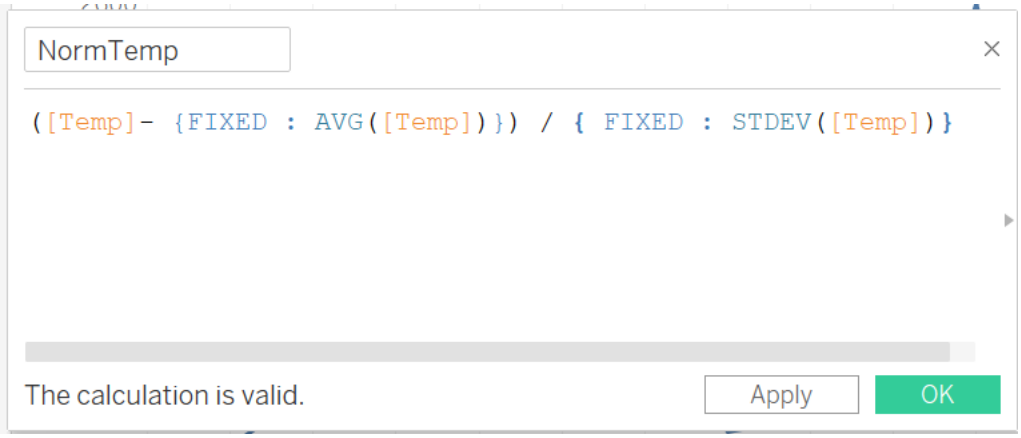
(show me how)

Create a New Calculated Field;



Enter the following formula for "**NormTemp**":





Click “OK”.

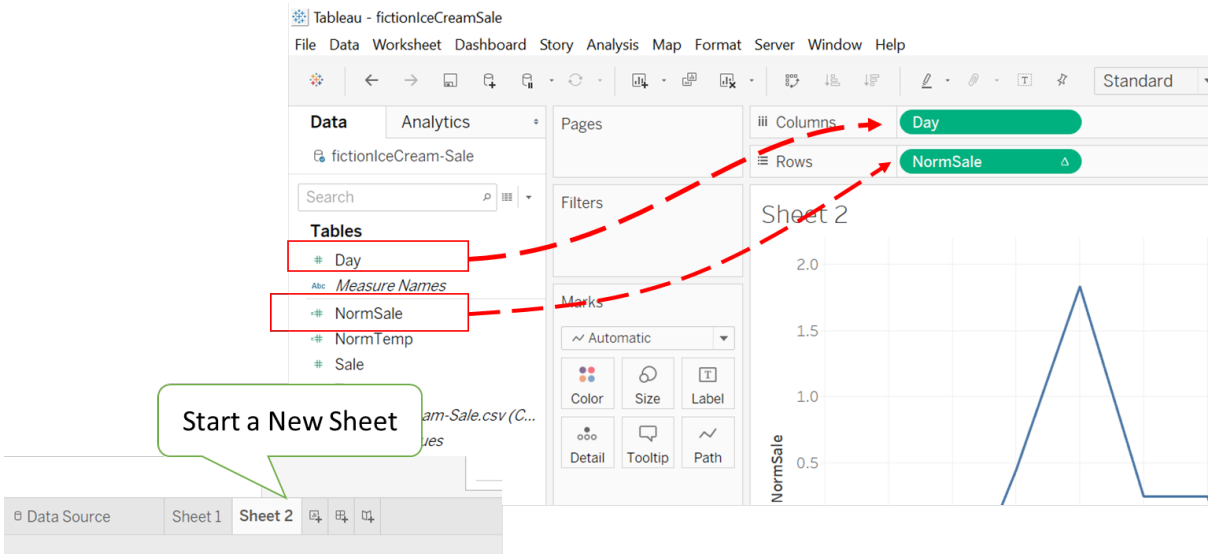
Create a new calculated field for “NormSale”:

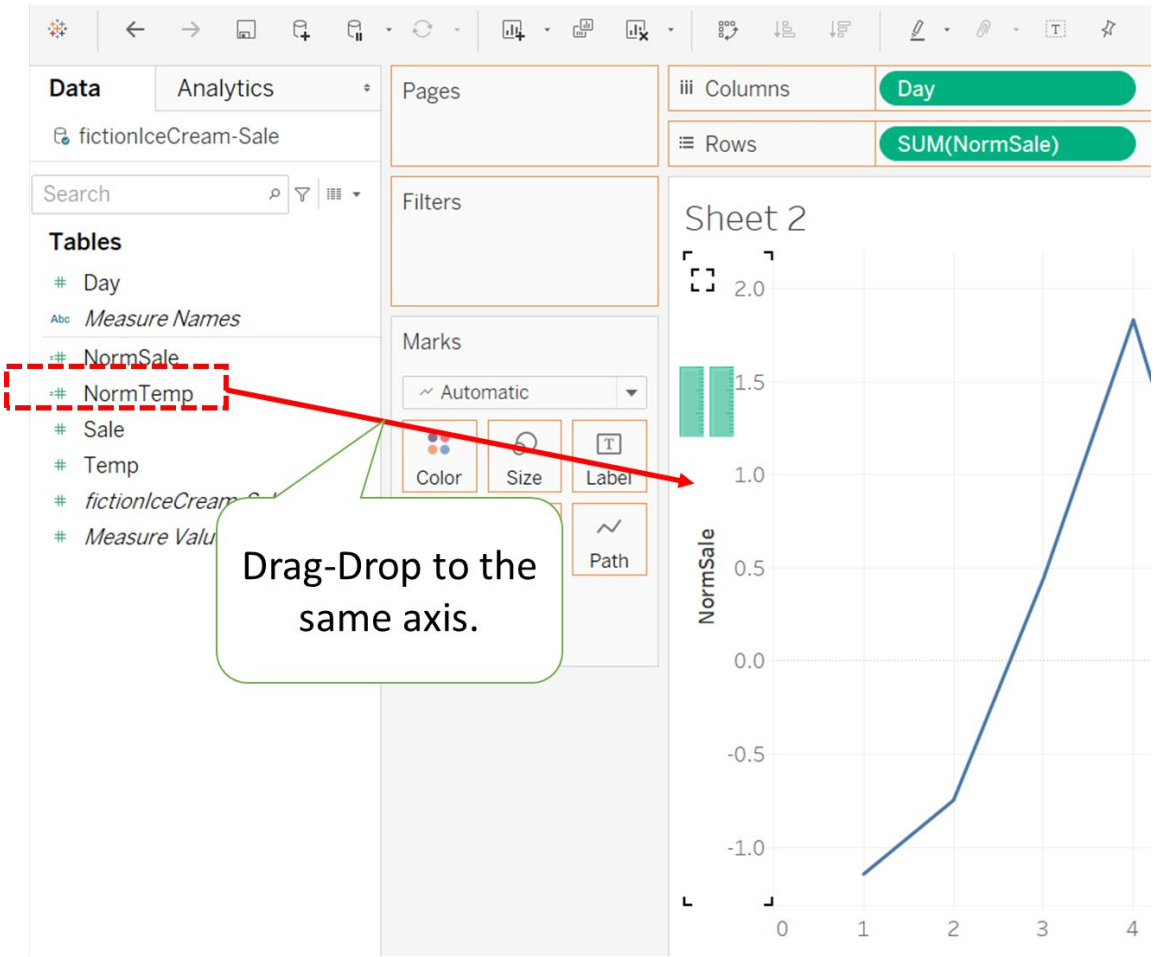
Enter the following formula for “**NormSale**”:

([Sale] - { FIXED : AVG([Sale])}) / { FIXED : STDEV([Sale])}

Click “OK” to continue.

Create a new Sheet.





What do you observe?

Exercise 2

We will make use of a dataset containing video game sales data for various genres of video games on different gaming platforms. For more details, you can visit this link (https://www.kaggle.com/gregorut/videogamesales)

Data set: ***vgsales_small.csv***

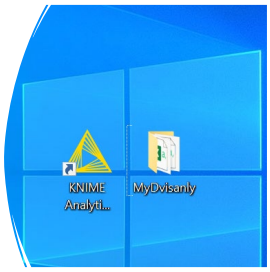
Appendix

Variable	Description
Rank	Ranking of overall sales
Name	Name of the game
Platform	Platform of the games release (eg. PC, PS4, etc)
Year	Year of the game's release
Genre	Genre of the game
Publisher	Publisher of the game
NA_Sales	Sales in North America (in millions)
EU_Sales	Sales in Europe (in millions)
JP_Sales	Sales in Japan (in millions)
Other_Sales	Sales in the rest of the world (in millions)
Global_Sales	Total worldwide sales

Questions for this data:

1. Which region (NA_Sale, JP_Sale, EU_Sale, Other_Sale) has the best video game sales performance?
2. What is the probability of finding a "shooter" game on a X360 platform?
3. Are there any linear correlation among the sales in the 4 regions?

Prepare your folder and data

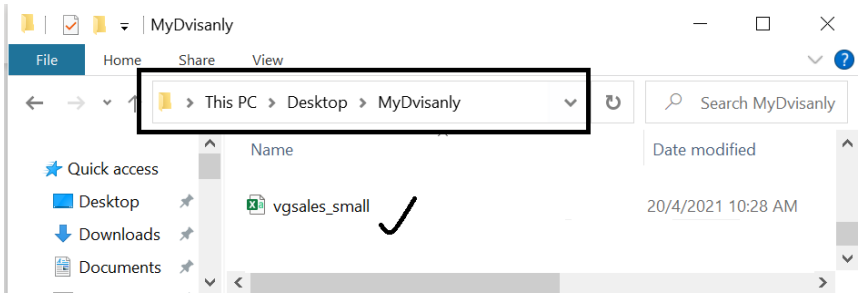


Create a folder on your "desktop" (or document folder), and name this folder "***MyDvisanly***".

All files will be saved in this folder.

Download the data ("*vgsales_small.csv*") from LMS > Week 3.

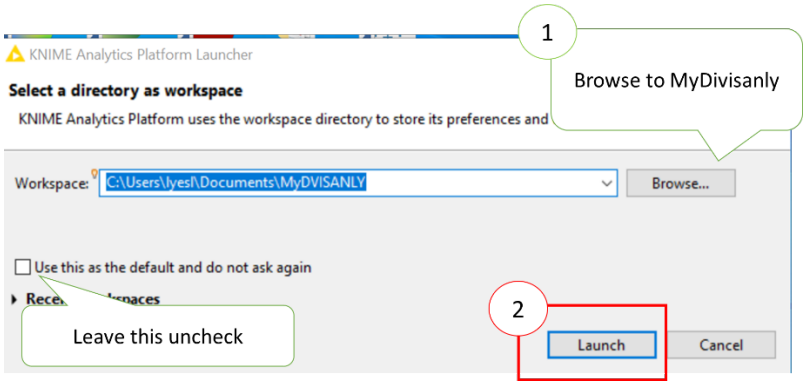
Copy the file to your folder:



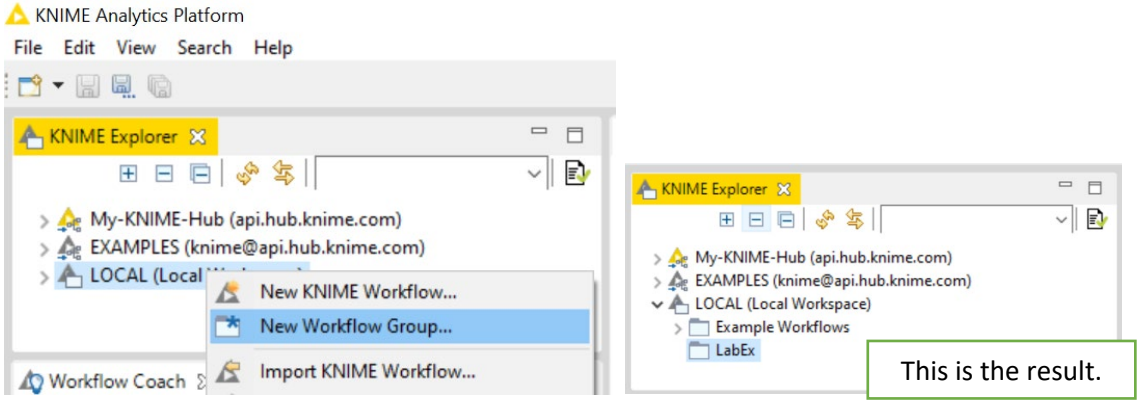
You are ready to launch KNIME.

Getting Started with KINME

1. Launch KNIME.
2. Browse to the folder created for this subject:

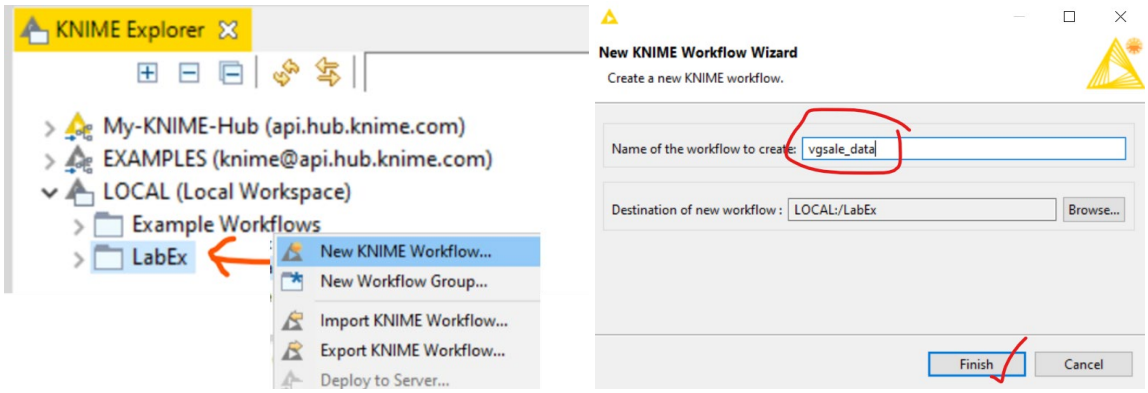


3. Right-click on the "LOCAL" to create **New Workflow Group**, and name it "**LabEx**".



Future workflow for the lab exercises will be saved here.

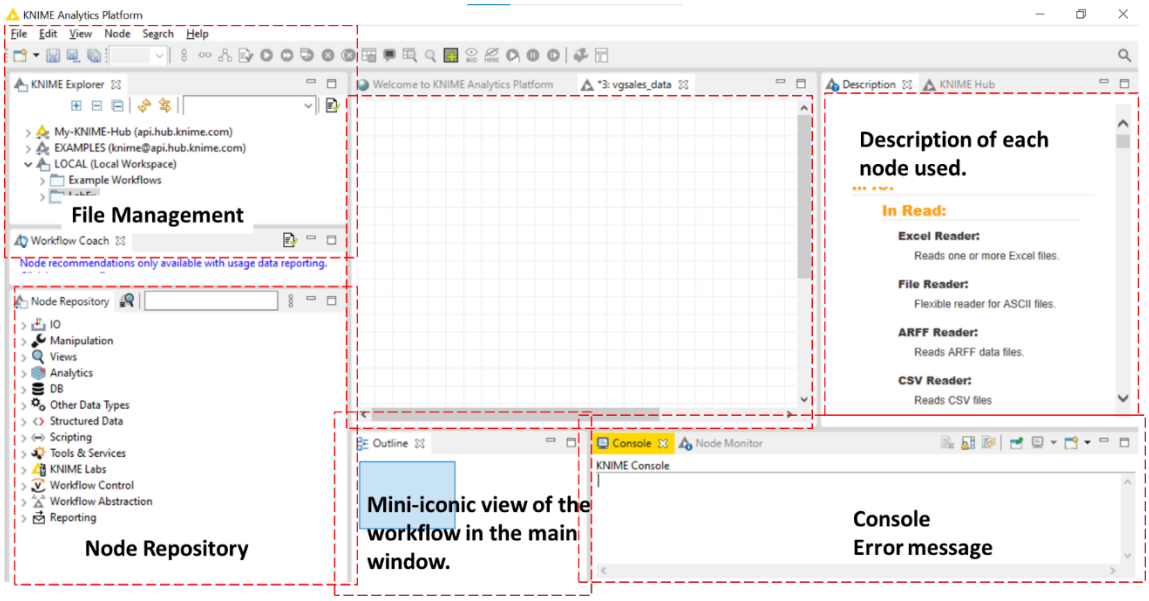
4. Right-click on the "LabEx" folder and create a new **WorkFlow**, name it as "**vgsale_data**".



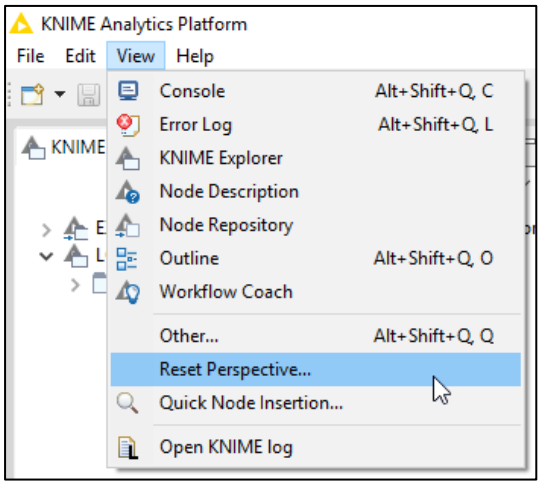
This is the KNIME file that we will work on.

(Show me how)

5. The KNIME workspace is divided into several panels.
   The top left corner shows your various KNIME workflows and workflow groups.
   The bottom left corner contains all the nodes found in KNIME.
   The top right corner shows the description of the selected node. Right now, we do not have any chosen nodes; hence it is empty.
   The centre is the main canvas, where we will add nodes later on.
   You can choose to show/hide any of the panels, but we will leave it as the default view for now.



If, at any point, you wish to restore the default view, click on **View > Reset Perspective.**

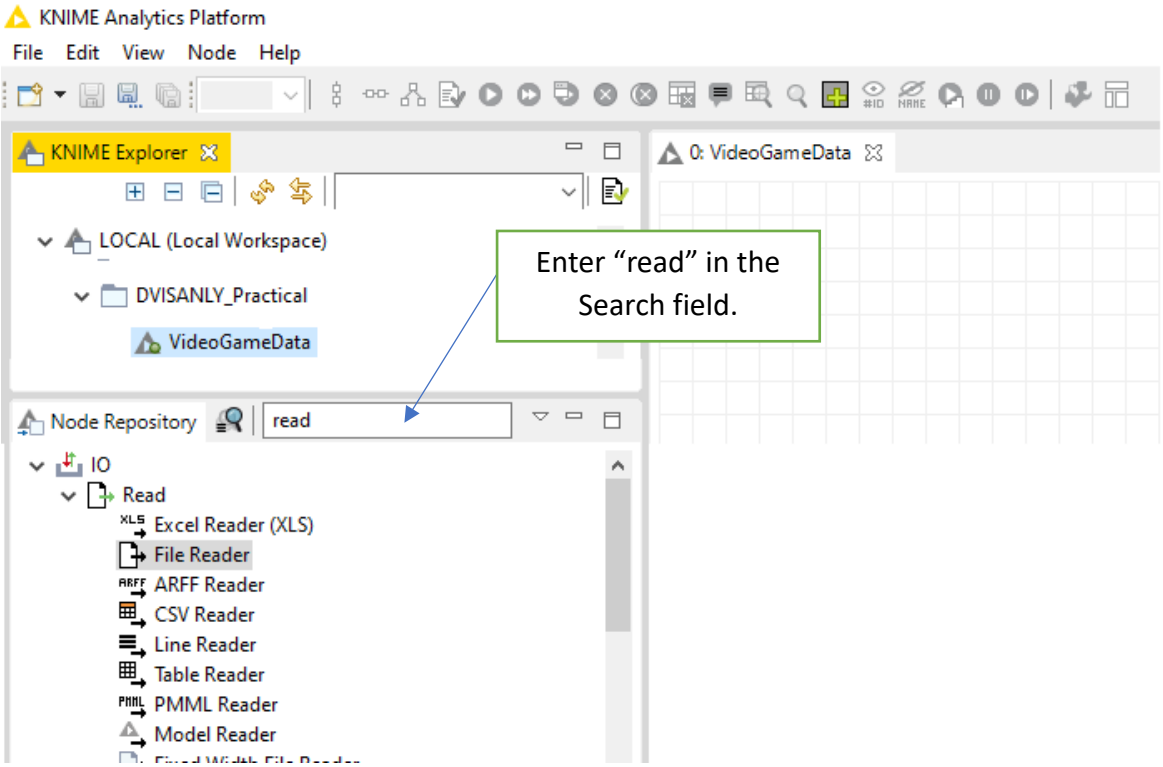


To reset your layout

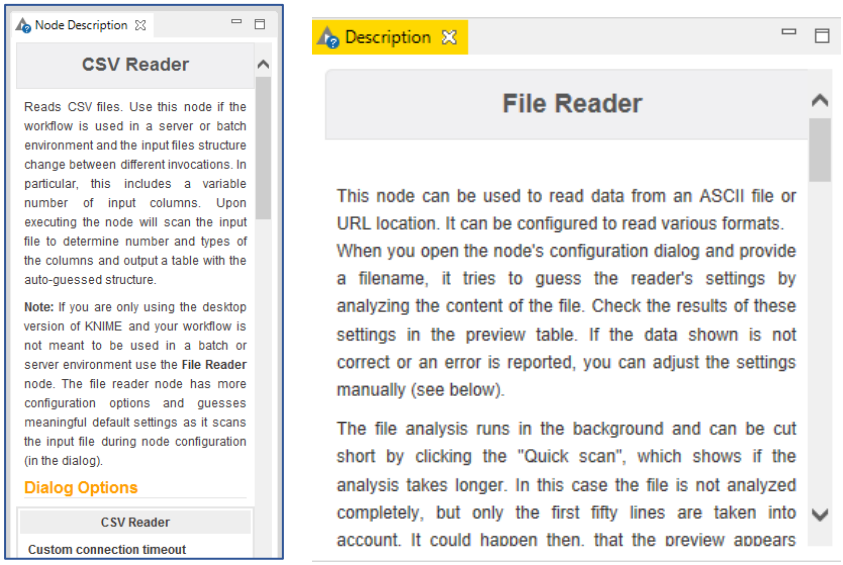
File Reader Node

We are now ready to add nodes to our first Workflow.

In the Node Repository panel (bottom left), enter "*read*" into the search box and press Enter.

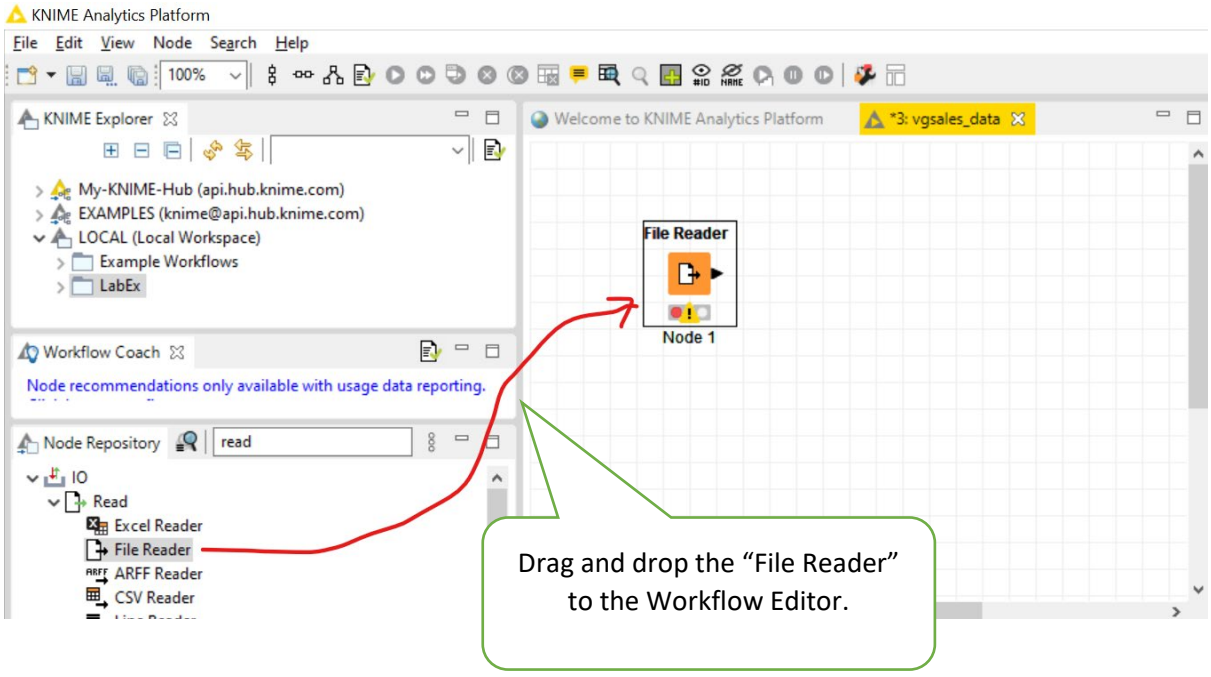


Both CSV and File Reader nodes can be used.



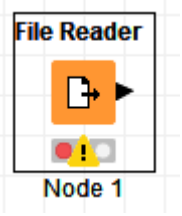
Select the CSV Reader node. As you click on the node, you should notice that the Node Description panel (top right) being populated with details of your selected node. This is where you can find out more about the various configuration options of your selected node.

Our preferred choice is to use the "File Reader" which is more robust than the CSV Reader.



(Show me how)

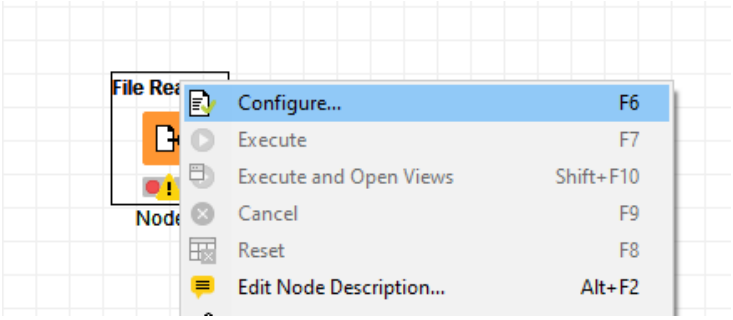
Interesting fact about a NODE:



You would have noticed the traffic light (red, yellow, green) below the node. In KNIME, all nodes would have this traffic light status below it.

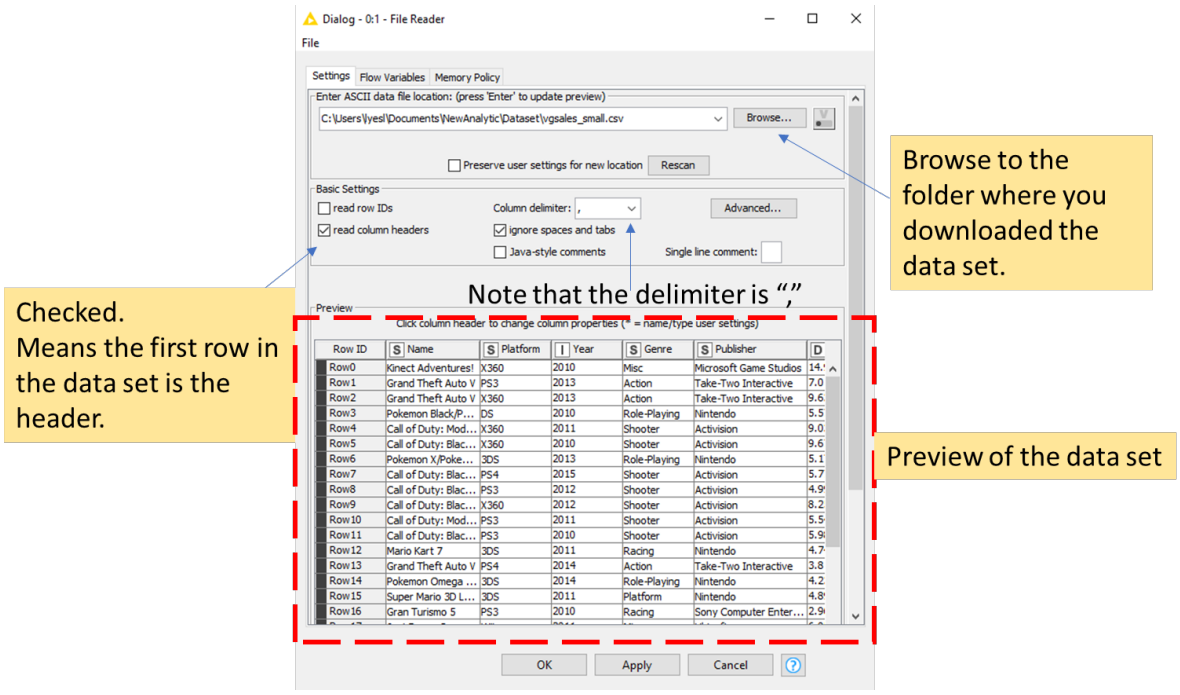
- Green - successful execution,
- Yellow - ready but not executed,
- Red - not ready or unsuccessful execution

Our node's status is Red because we have not configured it.



(Right-click on the node or double-click on the node)

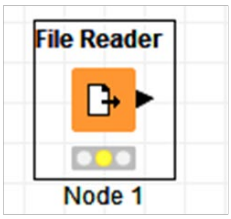
Select the *vgsales_small.csv* file.



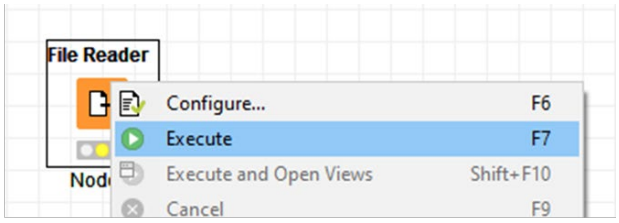
Ensure that the **read row IDs** checkbox is unchecked. (why?)

Because the given file does not have RowID column. So, KNIME will create a RowID column.

Click “OK” to continue.

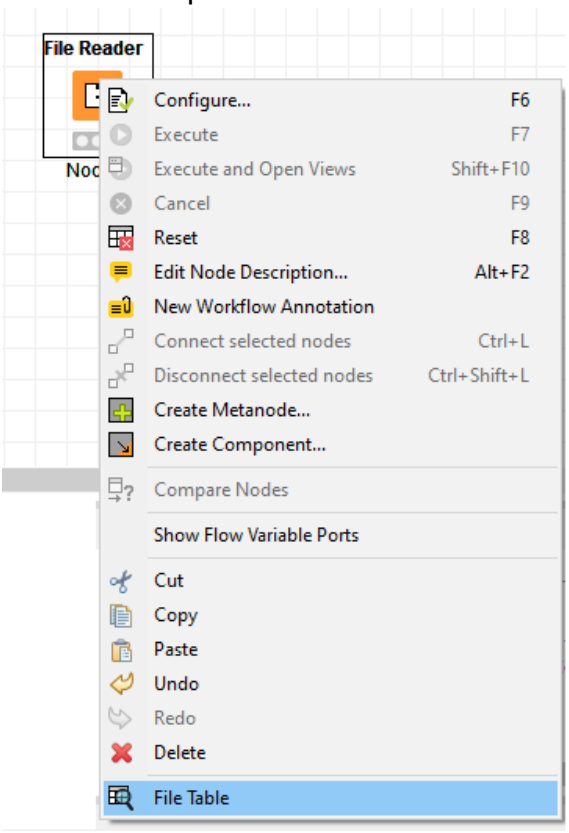


Node is ready to execute.

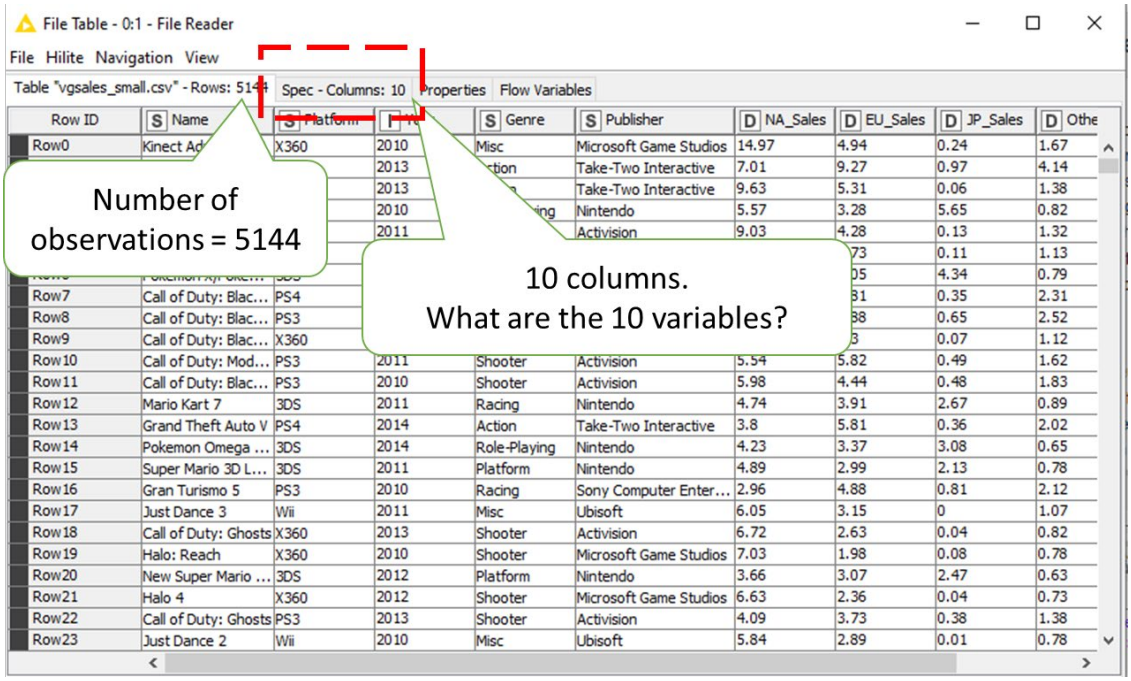


Right-click and select “Execute”

We can now proceed to view the table.

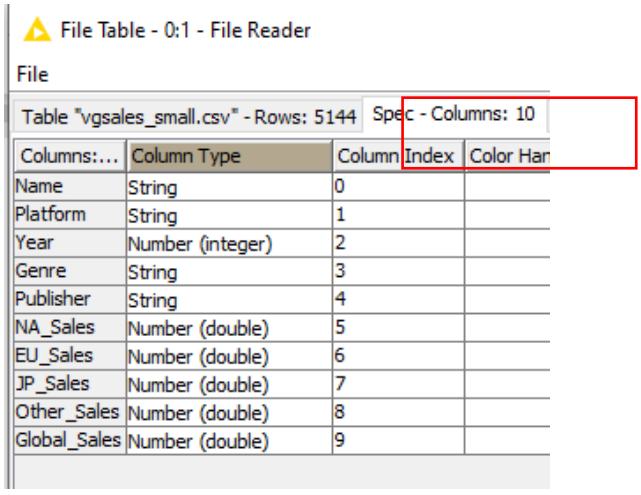


(Right-click and select “File Table”)



Row ID	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Othe
Row0	Kinect Ad	X360	2010	Misc	Microsoft Game Studios	14.97	4.94	0.24	1.67
			2013	ction	Take-Two Interactive	7.01	9.27	0.97	4.14
			2013		Take-Two Interactive	9.63	5.31	0.06	1.38
			2010	ing	Nintendo	5.57	3.28	5.65	0.82
			2011		Activision	9.03	4.28	0.13	1.32
							73	0.11	1.13
							05	4.34	0.79
Row7	Call of Duty: Blac...	PS4					31	0.35	2.31
Row8	Call of Duty: Blac...	PS3					38	0.65	2.52
Row9	Call of Duty: Blac...	X360					3	0.07	1.12
Row10	Call of Duty: Mod...	PS3	2011	Shooter	Activision	5.54	5.82	0.49	1.62
Row11	Call of Duty: Blac...	PS3	2010	Shooter	Activision	5.98	4.44	0.48	1.83
Row12	Mario Kart 7	3DS	2011	Racing	Nintendo	4.74	3.91	2.67	0.89
Row13	Grand Theft Auto V	PS4	2014	Action	Take-Two Interactive	3.8	5.81	0.36	2.02
Row14	Pokemon Omega ...	3DS	2014	Role-Playing	Nintendo	4.23	3.37	3.08	0.65
Row15	Super Mario 3D L...	3DS	2011	Platform	Nintendo	4.89	2.99	2.13	0.78
Row16	Gran Turismo 5	PS3	2010	Racing	Sony Computer Enter...	2.96	4.88	0.81	2.12
Row17	Just Dance 3	Wii	2011	Misc	Ubisoft	6.05	3.15	0	1.07
Row18	Call of Duty: Ghosts	X360	2013	Shooter	Activision	6.72	2.63	0.04	0.82
Row19	Halo: Reach	X360	2010	Shooter	Microsoft Game Studios	7.03	1.98	0.08	0.78
Row20	New Super Mario ...	3DS	2012	Platform	Nintendo	3.66	3.07	2.47	0.63
Row21	Halo 4	X360	2012	Shooter	Microsoft Game Studios	6.63	2.36	0.04	0.73
Row22	Call of Duty: Ghosts	PS3	2013	Shooter	Activision	4.09	3.73	0.38	1.38
Row23	Just Dance 2	Wii	2010	Misc	Ubisoft	5.84	2.89	0.01	0.78

This file contains 5144 rows of observations. Each observation is a game record.



Columns:...	Column Type	Column Index	Color Har
Name	String	0	
Platform	String	1	
Year	Number (integer)	2	
Genre	String	3	
Publisher	String	4	
NA_Sales	Number (double)	5	
EU_Sales	Number (double)	6	
JP_Sales	Number (double)	7	
Other_Sales	Number (double)	8	
Global_Sales	Number (double)	9	

There are 10 variables.

They are : Name, Platform, Year, Genre, Publisher, NA_Sales, EU_Sales, JP_Sales, Other_Sales and Global Sales.

What is the difference between Number (double) and Number (integer)?

Number (double) refers to decimal value.

In statistics, they are called Numerical Continuous.

Number (integer) refers to whole number (no decimal places).

In statistics, they are called Numerical Discrete.

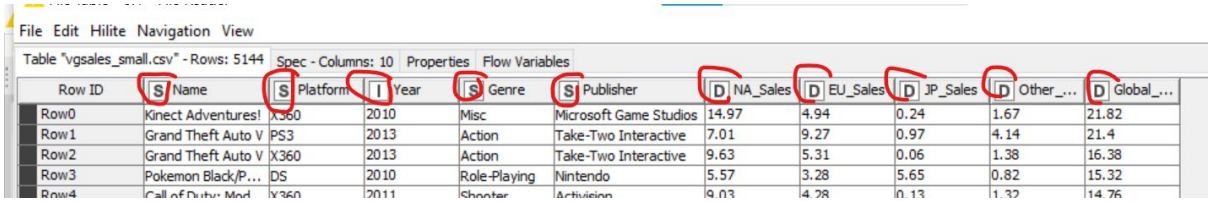
Why is it important to define the data type of a variable?

Different data types use different functions or operations.

For example, for numerical type, we can determine the statistics information like mean, median, and Standard Deviation.

For String, addition of 2 String type, means "joining" the text into one single String. This function is known as "concatenate".

Notice, there is a symbol of "S", "I" and "D" in front of each variables' name:



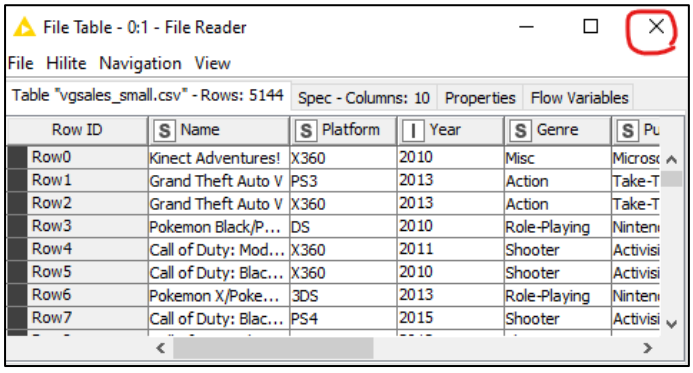
Row ID	S Name	S Platform	I Year	S Genre	S Publisher	D NA_Sales	D EU_Sales	D JP_Sales	D Other_...	D Global_...
Row0	Kinect Adventures!	X360	2010	Misc	Microsoft Game Studios	14.97	4.94	0.24	1.67	21.82
Row1	Grand Theft Auto V	PS3	2013	Action	Take-Two Interactive	7.01	9.27	0.97	4.14	21.4
Row2	Grand Theft Auto V	X360	2013	Action	Take-Two Interactive	9.63	5.31	0.06	1.38	16.38
Row3	Pokemon Black/P...	DS	2010	Role-Playing	Nintendo	5.57	3.28	5.65	0.82	15.32
Row4	Call of Duty: Mod...	X360	2011	Shooter	Activision	9.03	4.28	0.13	1.32	14.76

"S" type refers to STRING, imply categorical type.

"I" refers to Integer, imply numerical, discrete type.

"D" refers to Decimal, imply numerical, continuous type.

You can close the Table view by clicking on the "X" on the right corner.



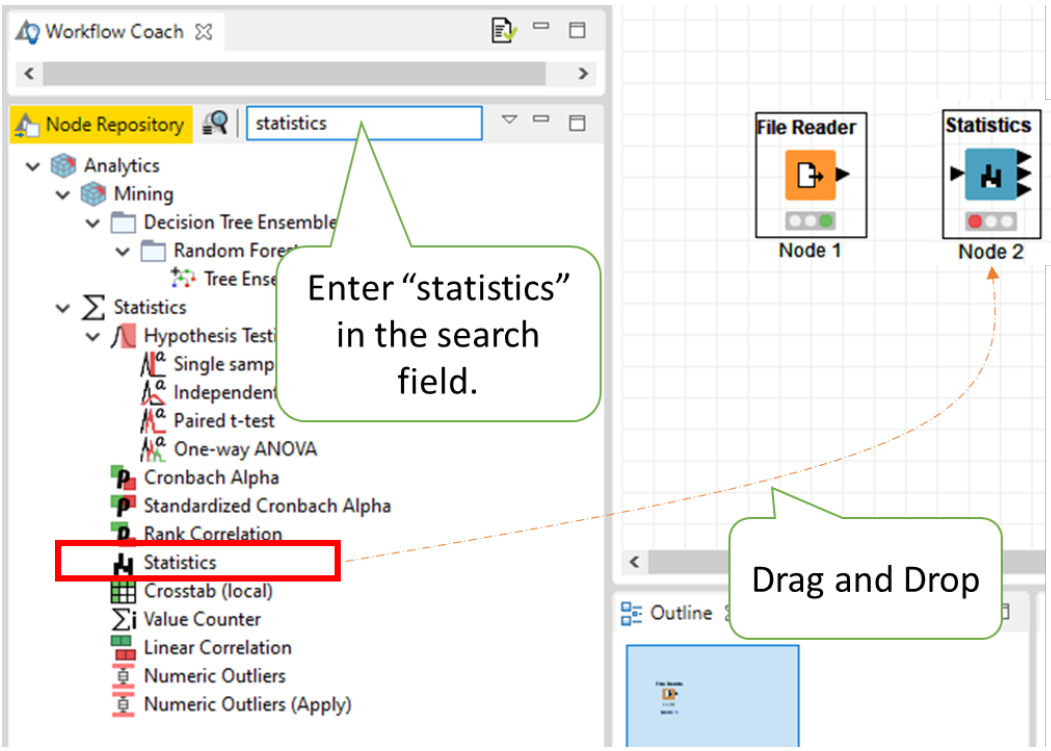
Row ID	S Name	S Platform	I Year	S Genre	S Pu
Row0	Kinect Adventures!	X360	2010	Misc	Microsc
Row1	Grand Theft Auto V	PS3	2013	Action	Take-T
Row2	Grand Theft Auto V	X360	2013	Action	Take-T
Row3	Pokemon Black/P...	DS	2010	Role-Playing	Ninten
Row4	Call of Duty: Mod...	X360	2011	Shooter	Activisi
Row5	Call of Duty: Blac...	X360	2010	Shooter	Activisi
Row6	Pokemon X/Poke...	3DS	2013	Role-Playing	Ninten
Row7	Call of Duty: Blac...	PS4	2015	Shooter	Activisi

Statistic Node

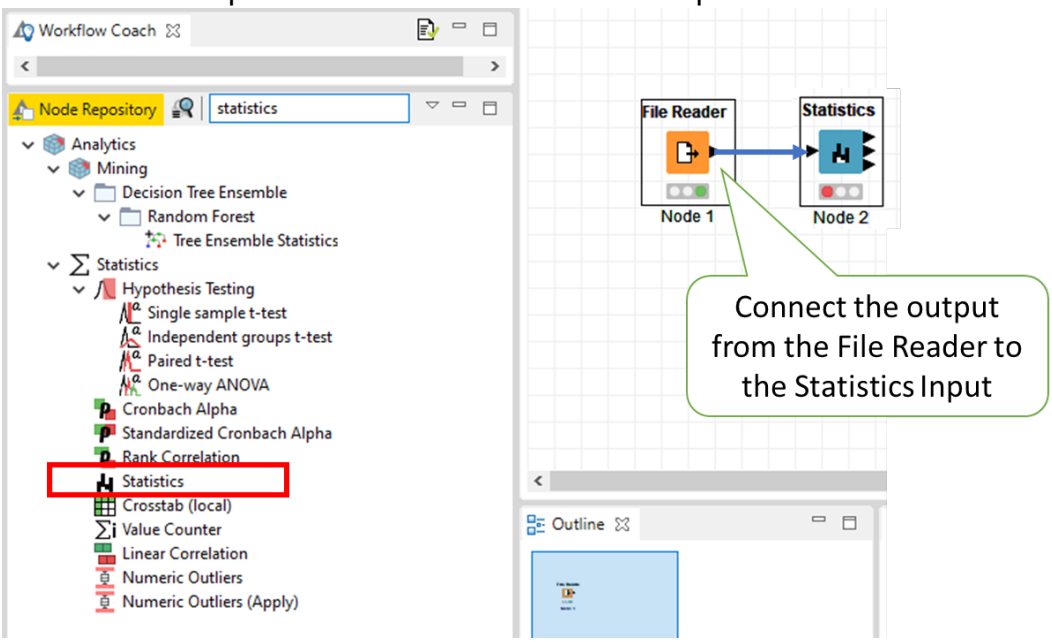

Now, we will proceed to uncover the statistic of the data set.

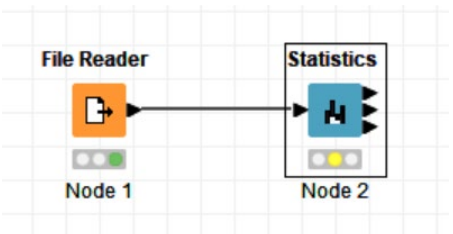
Add a Statistics node downstream from the File Reader node.

(Show me how)



Connect the output from the File Reader to the input of the Statistic node.

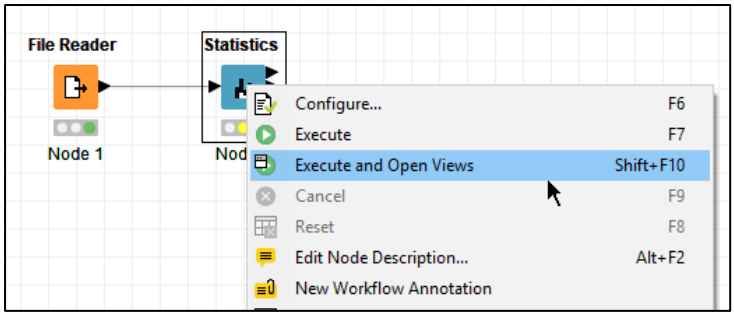




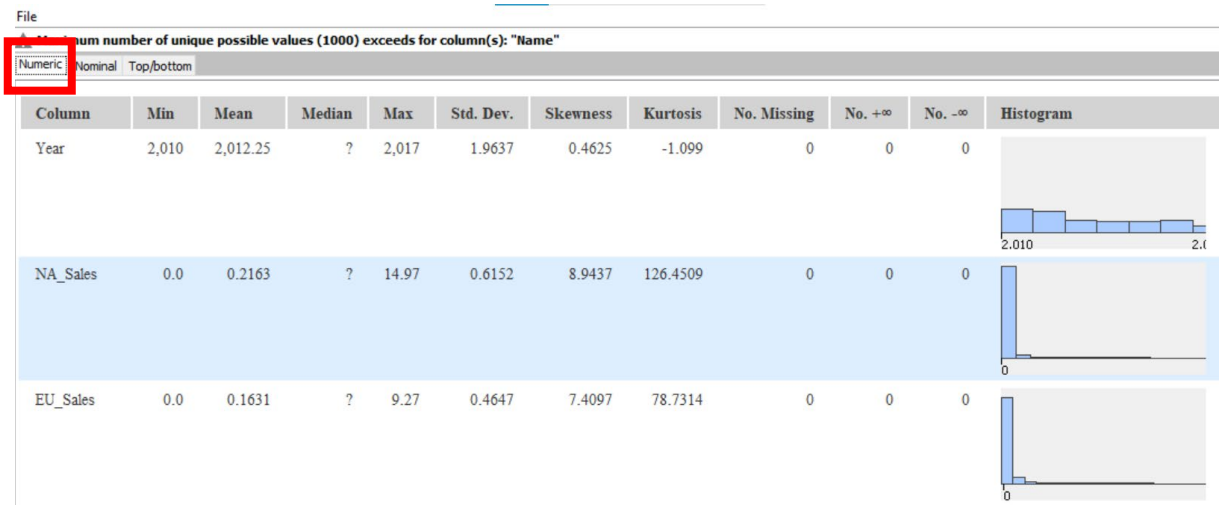
Notice the Statistics node is "yellow".

We can skip the configuration for now. Let's execute and study the result.

Right-click on the Statistics node, select **Execute and Open Views**



The result consists of 3 tabs (Numeric; Nominal; top/bottom).
Let us first look at the Numeric Tab.



Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
Year	2,010	2,012.25	?	2,017	1.9637	0.4625	-1.099	0	0	0	
NA_Sales	0.0	0.2163	?	14.97	0.6152	8.9437	126.4509	0	0	0	
EU_Sales	0.0	0.1631	?	9.27	0.4647	7.4097	78.7314	0	0	0	

Summary Statistics of the 6 numeric variables are automatically calculated.

Wait! Why is there Mean value for "Year"? Though "Year" is an integer, statistical information is not relevant.

Strictly speaking, "Year" is NOT a numerical type, it is a categorical (it should be a "String" type.)
Other examples that a number is interpreted as "String" are Identification number, block number, bus number, etc.

The Statistics node computes the summary statistics (min, max, median, mean, standard deviation, etc.) of each of the numeric fields. It also produces the histogram to show the spread of the values for the field. There's also a column showing you the number of missing values for each field, a useful feature!

File

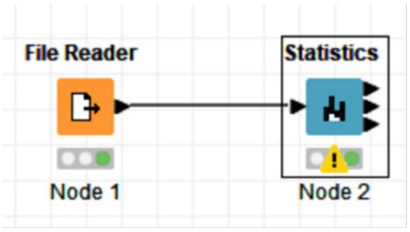
⚠ Maximum number of unique possible values (1000) exceeds for column(s): "Name"

Numeric | Nominal | Top/bottom

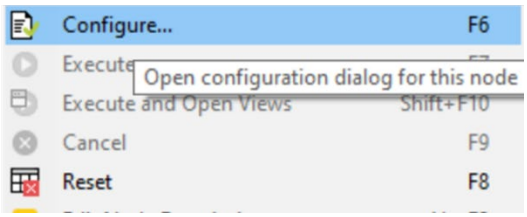
Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
Year	2,010	2,012.25	?	2,017	1.9637	0.4625	-1.099	0	0	0	
NA_Sales	0.0	0.2163	?	4.97	0.6152	8.9437	126.4509	0	0	0	
EU_Sales	0.0	0.1631	?	9.27	0.4647	7.4097	78.7314	0	0	0	


Why is there missing values for Median?
Is not missing. Is just that we did not select the option in the statistics configuration to compute the Median (remember, we skip the configuration.)

Now, close the statistics file dialog, and configure the Statistics node.

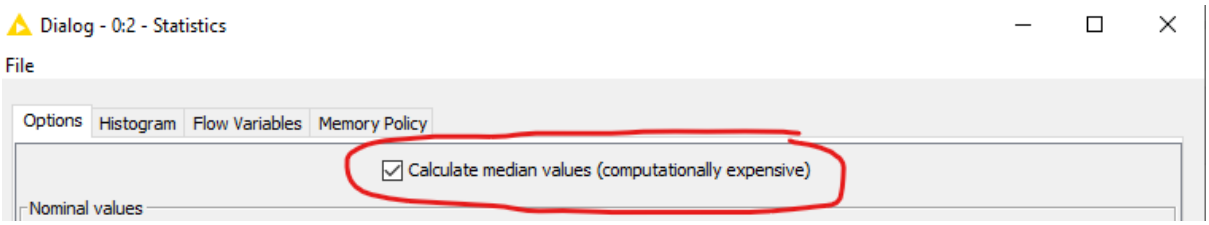


Right-click

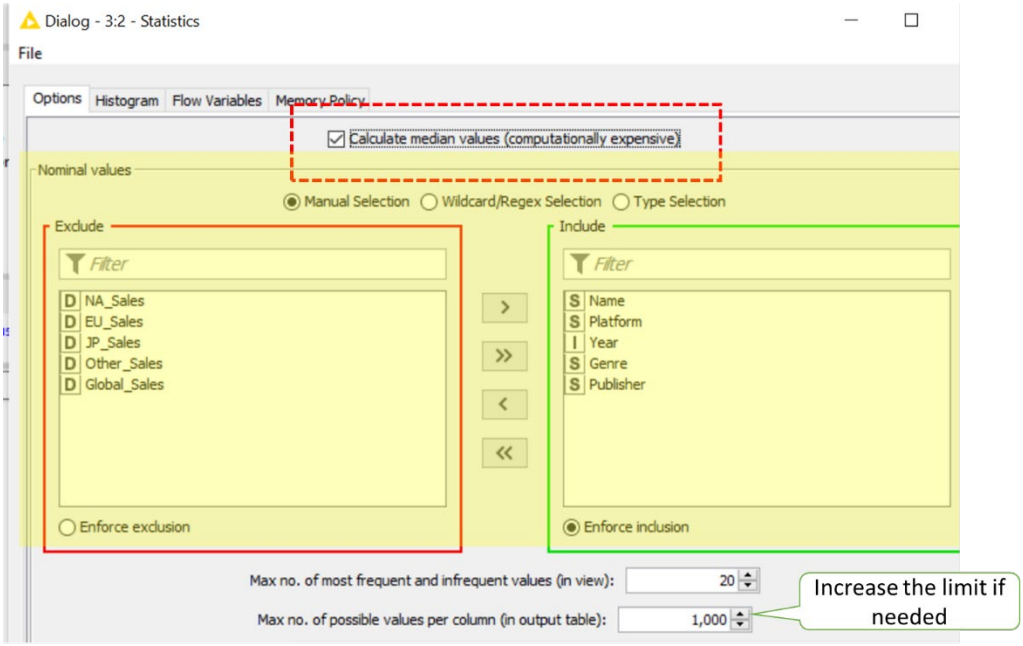


Check on the option to calculate the Median


Why do you think computing Median is a computationally expensive operation?



Calculation of Median is computationally expensive because the data need to be sorted in increasing order first to determine the value in the middle position.

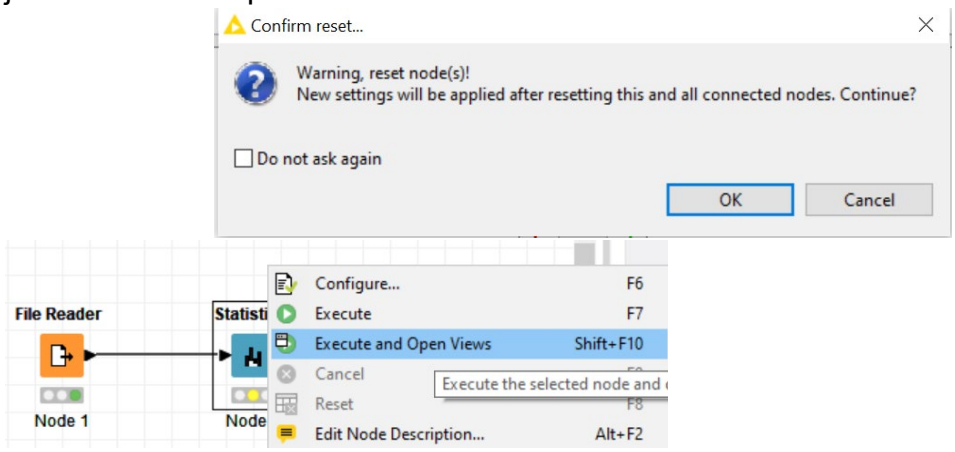


The central region is the option to select Nominal variables. The program has selected all STRING type (see the [S] symbol), but smart enough to include the "Year" in the nominal.
(refer to earlier explanation, that "nominal" refers to non-numeric type of values, and is also called Categorical in statistics term)

Note also there is a limit setting to read the number of values per column. The "Name" variable has number of possible values exceed 1000, as it is unique for every game. For this analysis, we don't have to increase it, as we are not checking the statistics for Name.

Proceed to click "OK" and execute the node to see the result.

You will always see this pop-up dialog box every time changes are done. We will just click "OK" to proceed.



What insights can we gather from the statistics?

Question 1

Which region (NA_Sale, JP_Sale, EU_Sale, Other_Sale) has the best video game sales performance?

We can compare the Mean, Median, Max among these 4 regions.

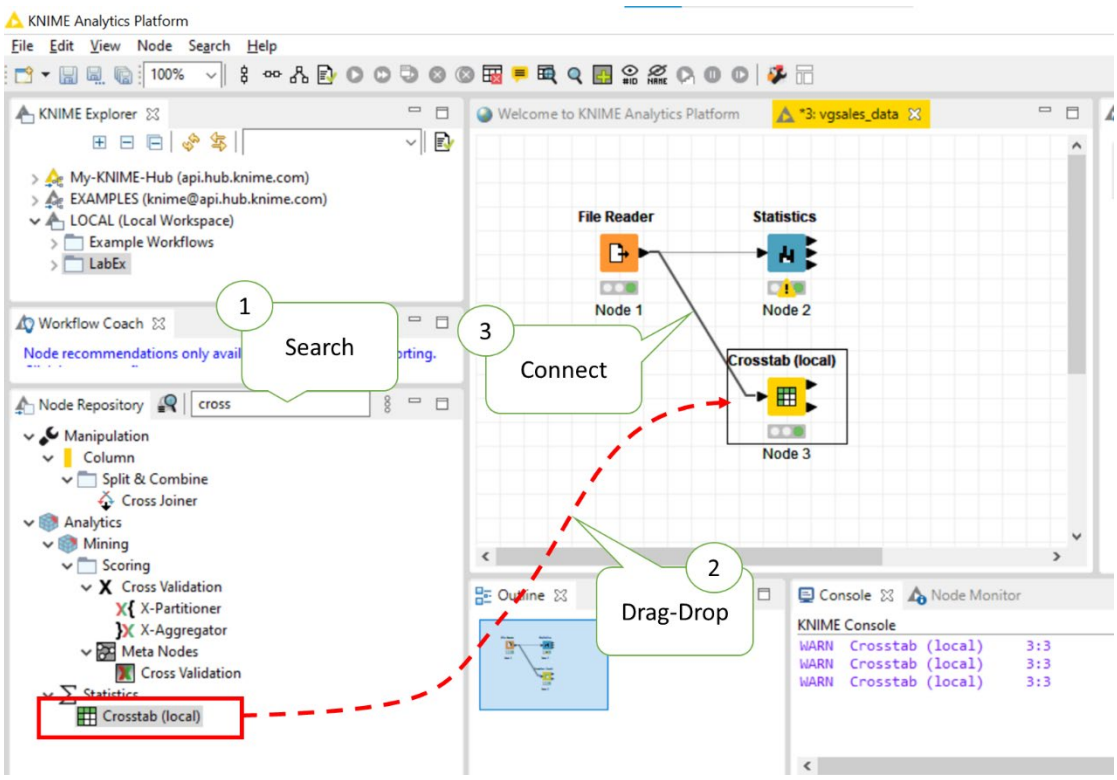
Question 2:

What is the probability of finding a "shooter" game on X360 platform?

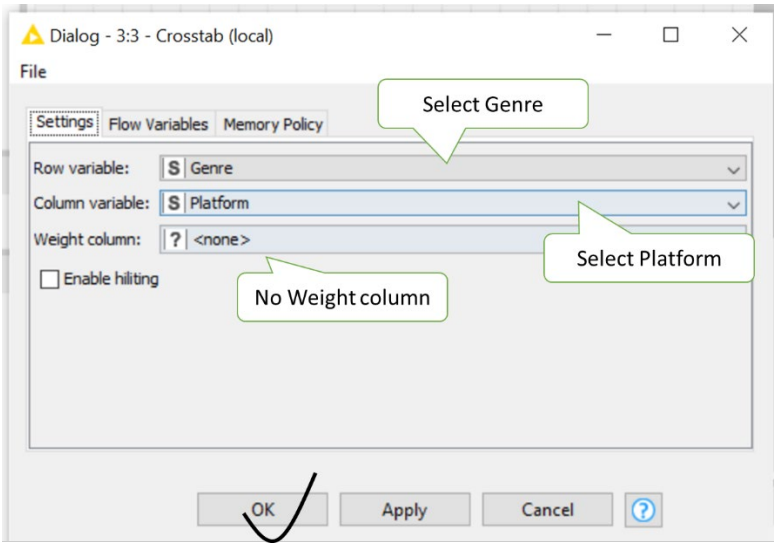
For this question, we can find the answer from cross tabulation (or contingency table).

(show me how)

Let's search for the crossTab node:

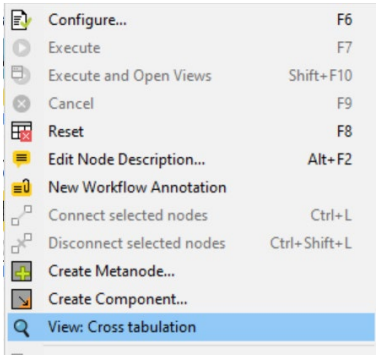


Configure the CrossTab node:

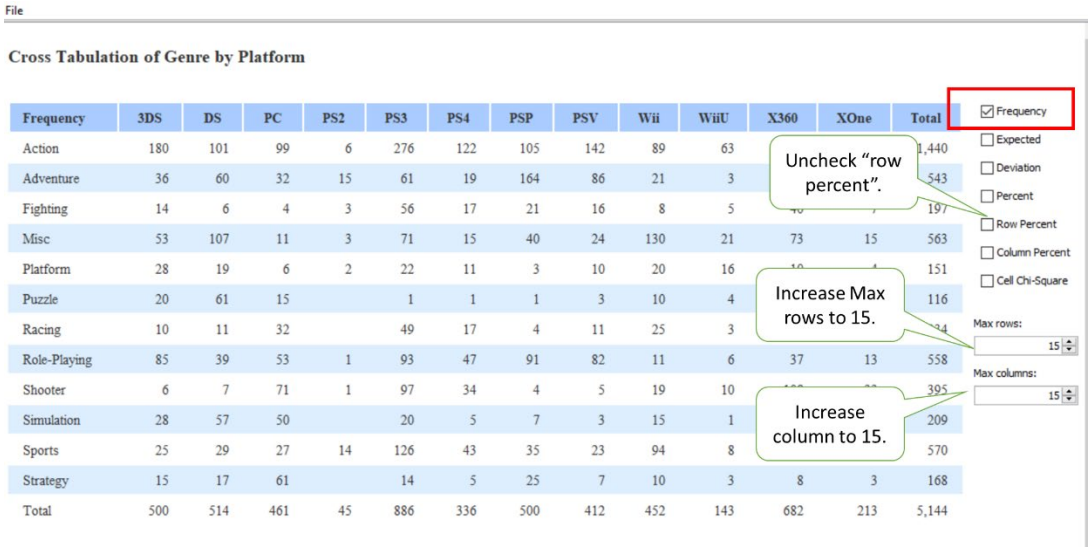


Shooter game is a type of Genre, and X360 is a type of platform. So, we select Genre and Platform as the 2 variables.

Execute the node. Right-click on the "CrossTab" node to view the result:



Edit the setting on the right menu:



Frequency	3DS	DS	PC	PS2	PS3	PS4	PSP	PSV	Wii	WiiU	X360	XOne	Total
Action	180	101	99	6	276	122	105	142	89	63			1,440
Adventure	36	60	32	15	61	19	164	86	21	3			543
Fighting	14	6	4	3	56	17	21	16	8	5			197
Misc	53	107	11	3	71	15	40	24	130	21	73	15	563
Platform	28	19	6	2	22	11	3	10	20	16			151
Puzzle	20	61	15		1	1	1	3	10	4			116
Racing	10	11	32		49	17	4	11	25	3			
Role-Playing	85	39	53	1	93	47	91	82	11	6	37	13	558
Shooter	6	7	71	1	97	34	4	5	19	10			395
Simulation	28	57	50		20	5	7	3	15	1			209
Sports	25	29	27	14	126	43	35	23	94	8			570
Strategy	15	17	61		14	5	25	7	10	3	8	3	168
Total	500	514	461	45	886	336	500	412	452	143	682	213	5,144

We just want to see the "frequency". The number of rows and columns needed is at least 12, as there are 12 categories from each variable.

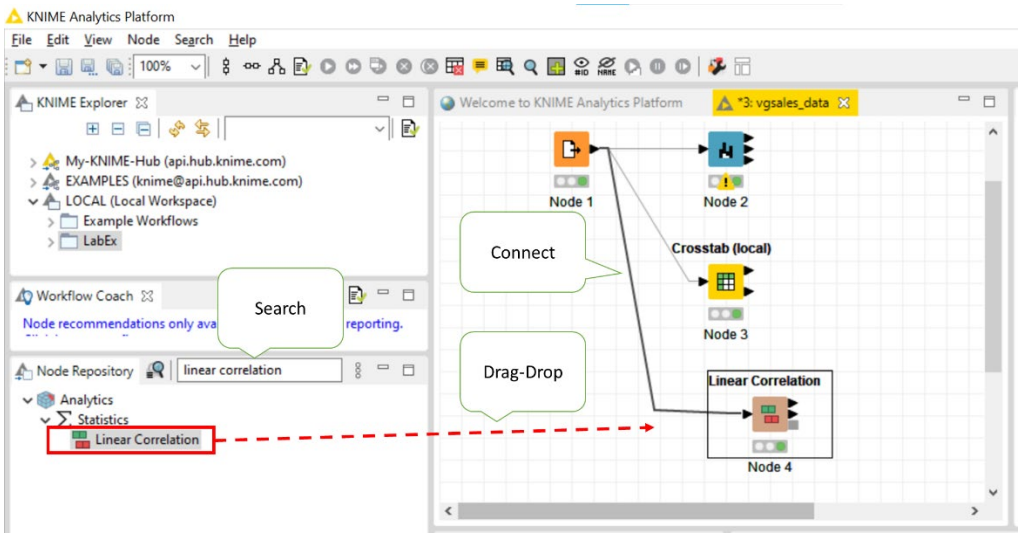
Answer the question:

What is the probability of finding a "shooter" game on X360 platform?

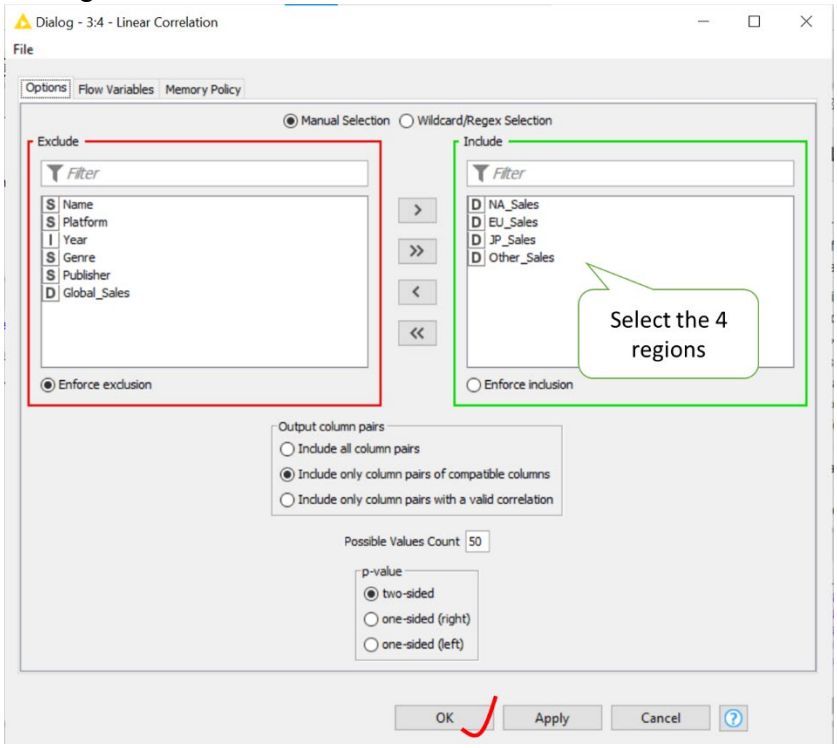
Question 3:
Are there any linear correlations among the sales in the 4 regions?

We need a Linear Correlation node to help us.

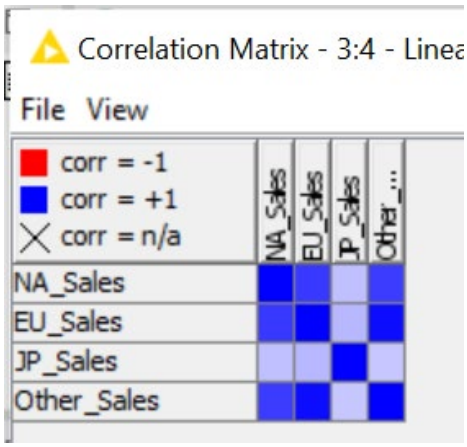
(Show me how)



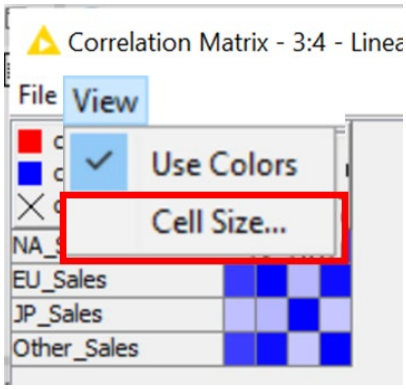
Configure the Linear Correlation node:

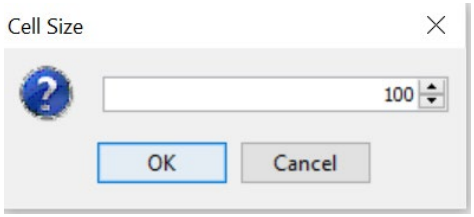


Execute and open view.



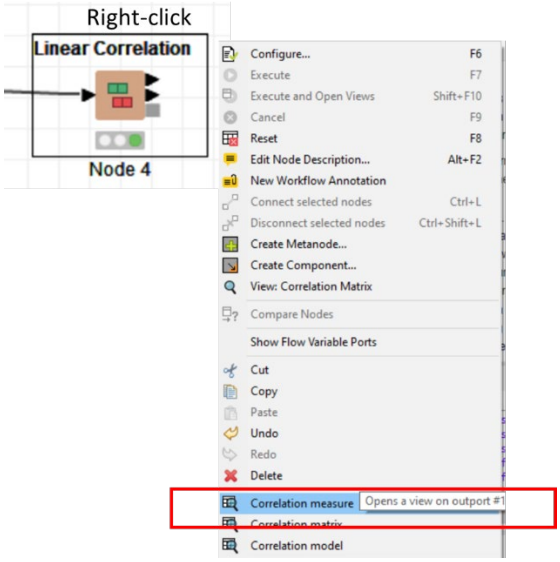
Too small. We can enlarge it by setting the cell size under "View".

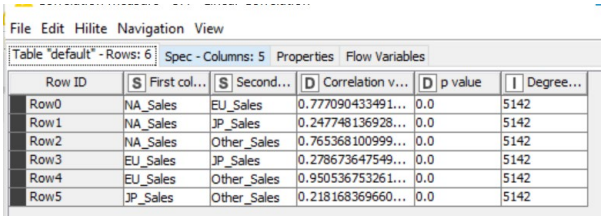




Mouse over each cell to see the Linear correlation coefficient, r.

Alternately, you can also check the linear correlation coefficient from the *correlation measure*:



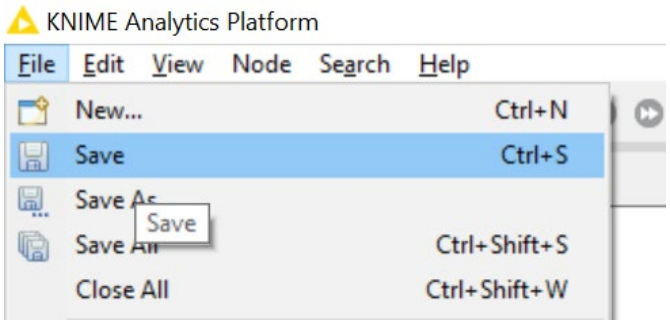


Row ID	First col...	Second...	Correlation v...	p value	Degree...
Row0	NA_Sales	EU_Sales	0.777090433491...	0.0	5142
Row1	NA_Sales	JP_Sales	0.247748136928...	0.0	5142
Row2	NA_Sales	Other_Sales	0.765368100999...	0.0	5142
Row3	EU_Sales	JP_Sales	0.278673647549...	0.0	5142
Row4	EU_Sales	Other_Sales	0.950536753261...	0.0	5142
Row5	JP_Sales	Other_Sales	0.218168369660...	0.0	5142

Answer the question:

Are there any linear correlations among the sales in the 4 regions?

Save the KNIME workflow.



Additional questions that can be asked from the data:
4) What is the probability of finding a “puzzle” genre?
5) What is the probability of finding an “action” genre among the PS3 platform?

Self-check point:

	Check List (Y/N)
You understood the concept of	
Normal distribution	
Central Limit Theorem	
Normalisation using	
- Z-Transform	
- Feature Scaling (Min-Max)	
Probability	
Conditional Probability	
Interpreting Cross Tabulation (Contingency Table)	
You understood how to:	
Use Calculated Field in TABLEAU	
Transform data to Z-score	
Plot the Z-score of 2 variables on the same timeline.	
You know how to use : (in KNIME)	
Statistics Node to check Mean, Max, Min and Median	
Linear Correlation Node to show Correlation Matrix	
Crosstab Node to answer categorical data	

Lab 5

Learning outcomes:

1. Able to create visualisation charts for categorical data analysis

Introduction

For categorical data types, the visual charts are:

- Pie-Chart
- Bar Chart
- Contingency Table (Cross Tabs)

Exercise 1: Visualisation for categorical data

Data Set: *mental_health_small.csv*

This was a survey returns of IT Professionals from various countries on their views of their employer's attitudes towards and support for the mental health of employees. For more details (e.g., the survey questions for each column), you can visit this link (https://www.kaggle.com/osmi/mental-health-in-tech-survey/data)

Number of observations: 982

Number of variables (excluding date&time): 25

Only numerical variable is AGE, the rest are categorical type.

Information that can be gathered from categorical data are related to counts, frequency, and percentage.

Use suitable visual charts to support your answer to the following questions:

1) What is the percentage that the IT professionals are female? *Use Pie-chart.

   (ans: 21.89%)

   (Show me how)

2) What is the breakdown of survey respondents by country and gender?

   *Use Stack bar-chart

   (Show me how)

3) What is the percentage of IT professional, who are females (C), and willing to discuss mental issues with their Supervisor (A) , and have family history of mental illness (B)? P(A∩B∩C)

   (ans: 36/982 = 0.0367 or 3.67%)

   (Show me how)

4) Among the female IT professionals (C), what is the percentage that they are willing to discuss mental issues with their Supervisor (A), and have family history of mental illness (B)? P (A|B|C)
   (ans: 36/215 = 0.167 or 16.7%)

   (Show me how)

5) Most participants took the survey at which time of the day?

6) Ask 2 further questions and present your insights.

Appendix

Data File: *mental_health_small.csv*

This dataset contains the following data fied:

Variable	Survey questions
Timestamp	
Age	
Gender	
Country	
State	If you live in the United States, which state or territory do you live in?
self_employed	Are you self-employed?
family_history	Do you have a family history of mental illness?
treatment:	Have you sought treatment for a mental health condition?
work_interfere	If you have a mental health condition, do you feel that it interferes with your work?
no_employees	How many employees does your company or organization have?
remote_work	Do you work remotely (outside of an office) at least 50% of the time?
tech_company	Is your employer primarily a tech company/organization?
benefits	Does your employer provide mental health benefits?
care_options	Do you know the options for mental health care your employer provides?
wellness_program	Has your employer ever discussed mental health as part of an employee wellness program?
seek_help	Does your employer provide resources to learn more about mental health issues and how to seek help?

anonymity	Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources?
Leave	How easy is it for you to take medical leave for a mental health condition?
mental***health***consequence	Do you think that discussing a mental health issue with your employer would have negative consequences?
phys***health***consequence	Do you think that discussing a physical health issue with your employer would have negative consequences?
coworkers	Would you be willing to discuss a mental health issue with your coworkers?
supervisor	Would you be willing to discuss a mental health issue with your direct supervisor(s)?
mental***health***interview	Would you bring up a mental health issue with a potential employer in an interview?
phys*health*interview	Would you bring up a physical health issue with a potential employer in an interview?
mental***vs***physical	Do you feel that your employer takes mental health as seriously as physical health?
obs_consequence	Have you heard of or observed negative consequences for coworkers with mental health conditions in your workplace?
comments	Any additional notes or comments

Guide notes:

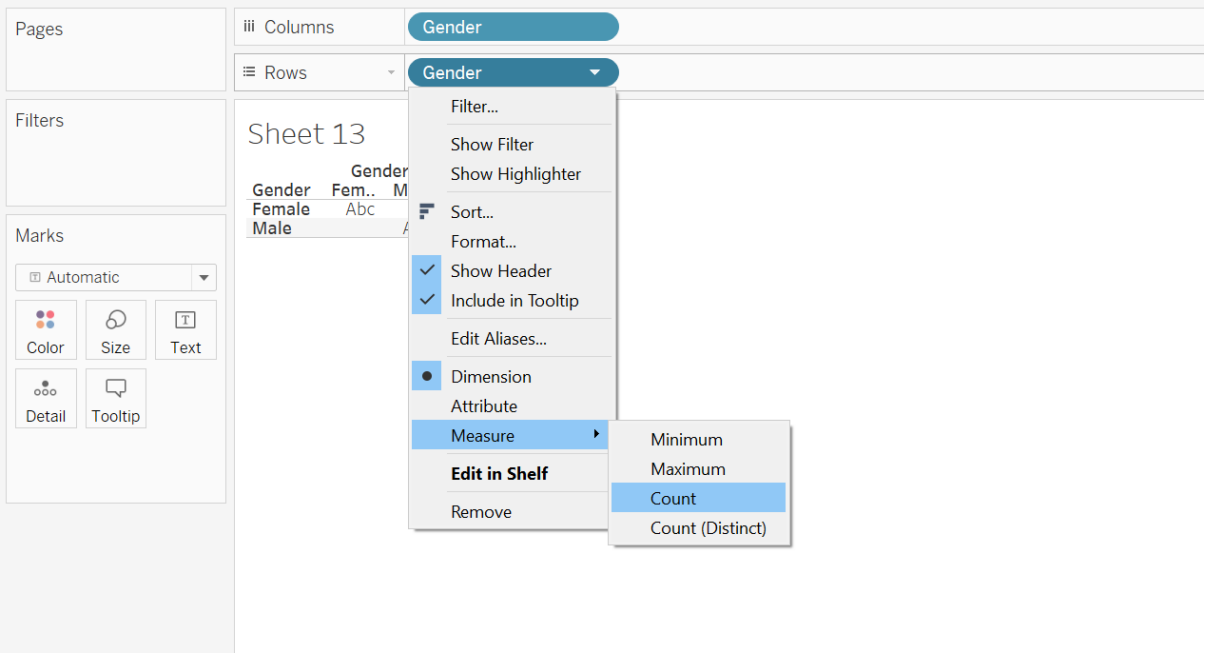
How to create a Pie-chart in Tableau:

Example, we want to find the distribution of males and females in the survey.

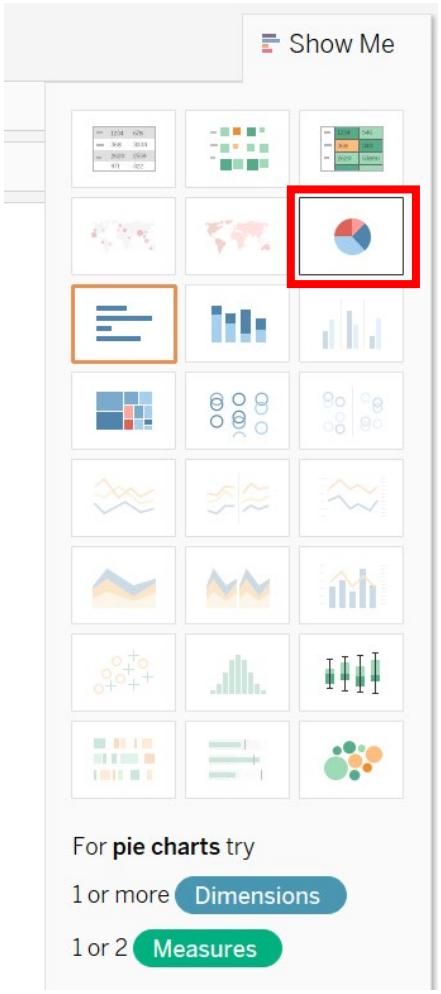
First, is to create a bar chart:

Drag the variable “Gender” to both columns and rows.

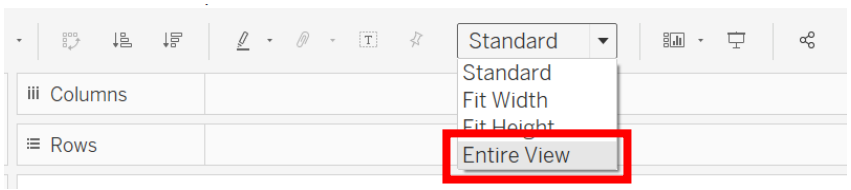
Edit the option for the “Gender” in Rows as “Count”.



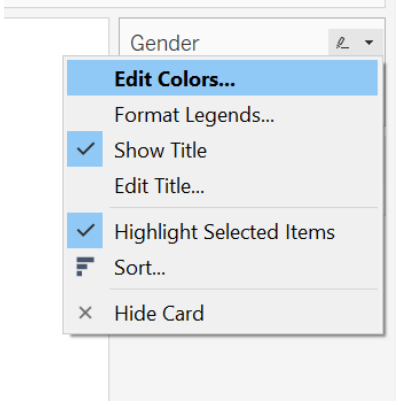
Then, open-up the “Show Me” tab (on the right), and select “pie-chart”:



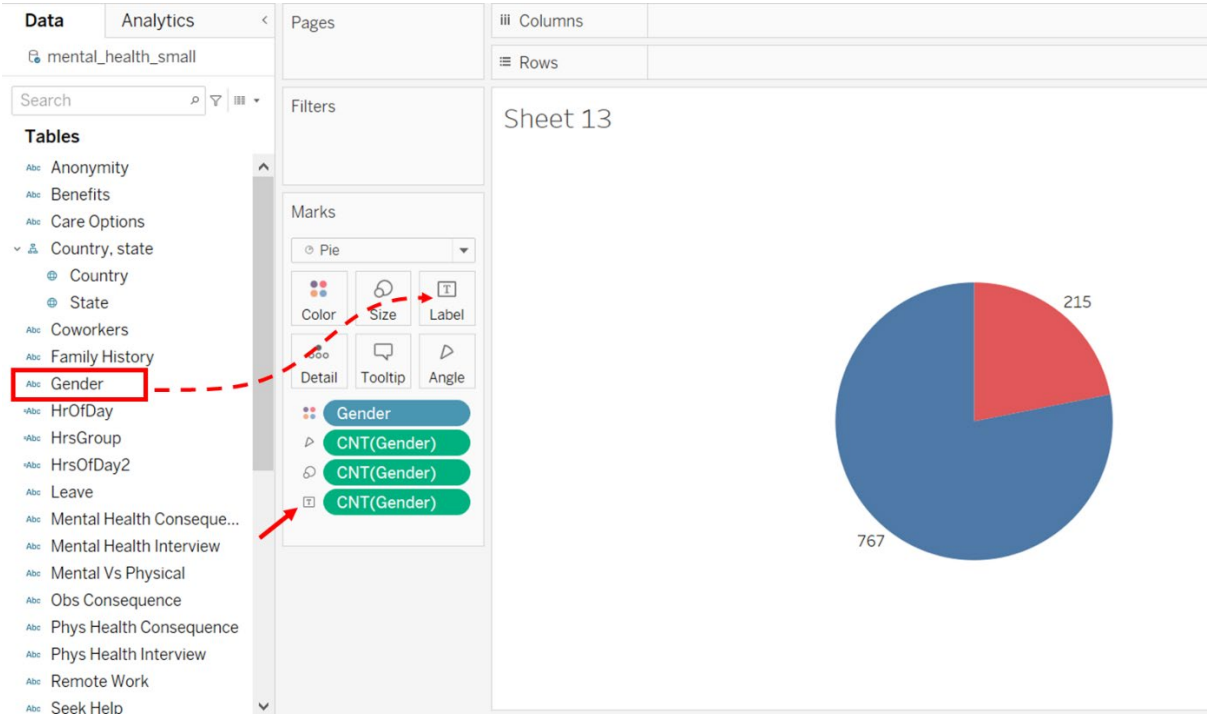
Extend the chart to full view:



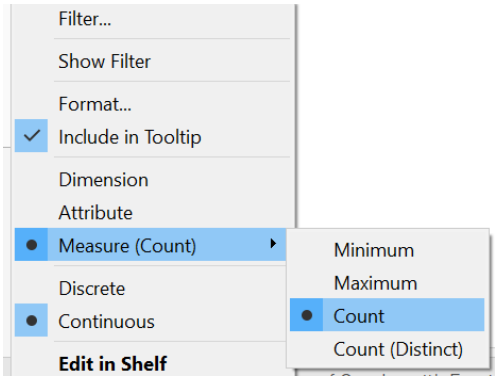
You change the legend color from:



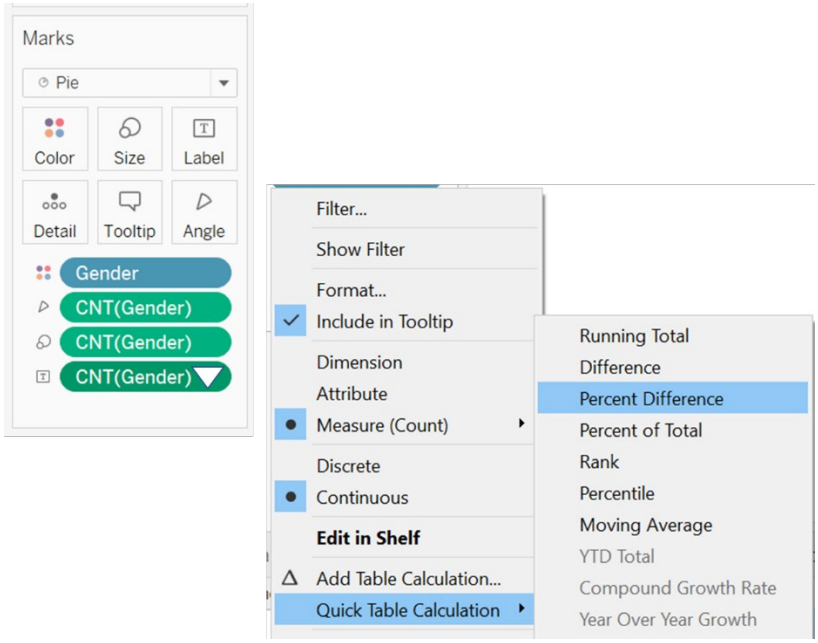
To display the numbers on the chart: Drag “Gender” to Label marks.



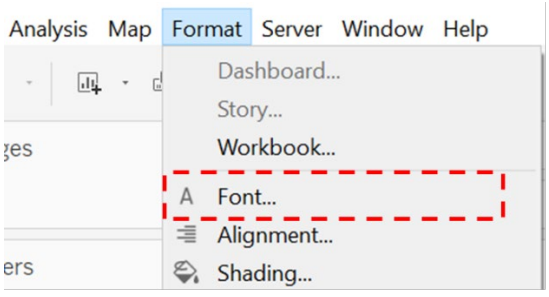
Change the measure to “count”.

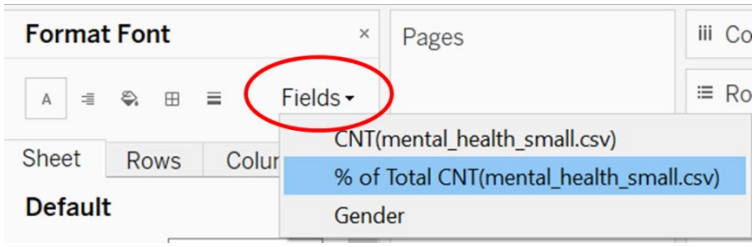


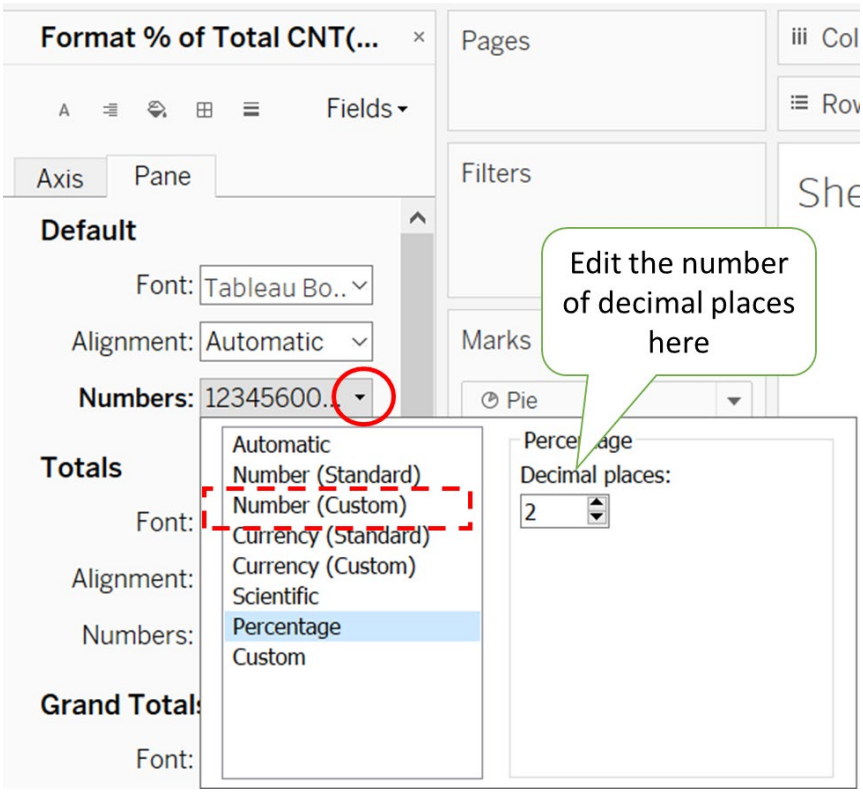
To show as Percentage:

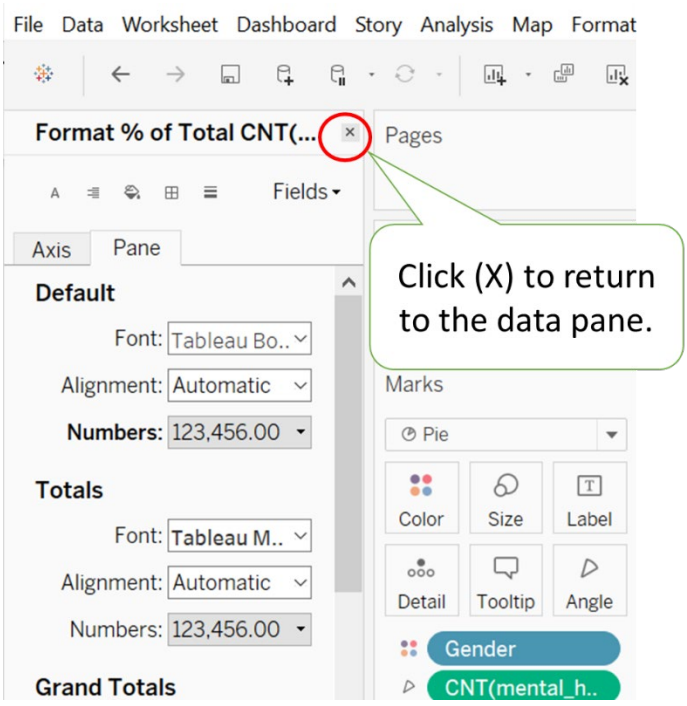


To edit the number of decimal places:



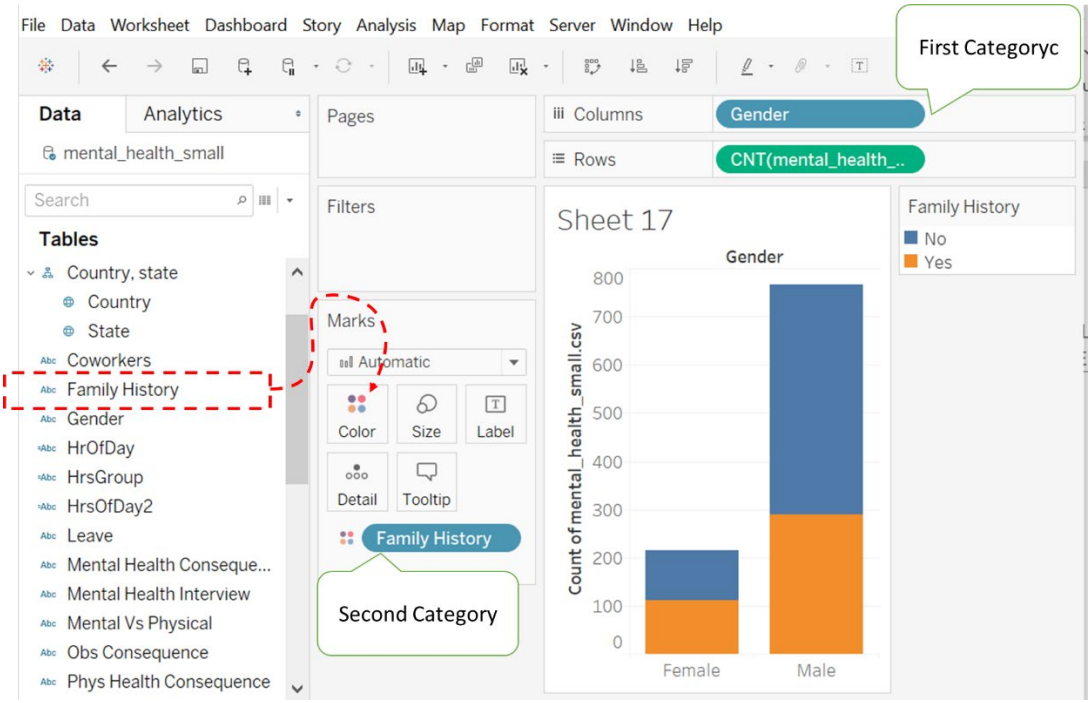



Format % of Total CNT(...
Pages
Filters
Marks
Pie
Fields
Axis
Pane
Default
Font: Tableau Bo..
Alignment: Automatic
Numbers: 12345600..
Totals
Font:
Alignment:
Numbers:
Grand Totals
Font:
Automatic
Number (Standard)
Number (Custom)
Currency (Standard)
Currency (Custom)
Scientific
Percentage
Custom
Decimal places:
2
Edit the number of decimal places here


File Data Worksheet Dashboard Story Analysis Map Format
Format % of Total CNT(...
Pages
Fields
Axis
Pane
Default
Font: Tableau Bo..
Alignment: Automatic
Numbers: 123,456.00
Totals
Font: Tableau M..
Alignment: Automatic
Numbers: 123,456.00
Grand Totals
Marks
Pie
Color
Size
Label
Detail
Tooltip
Angle
Gender
CNT(mental_h..
Click (X) to return to the data pane.

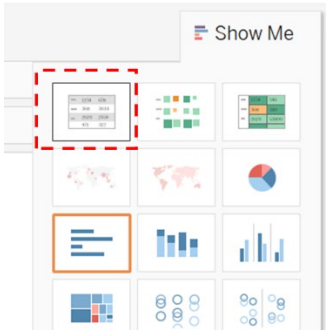
How to create stack bar-chart?

Stack-bar is used to show 2 categories on the same chart.

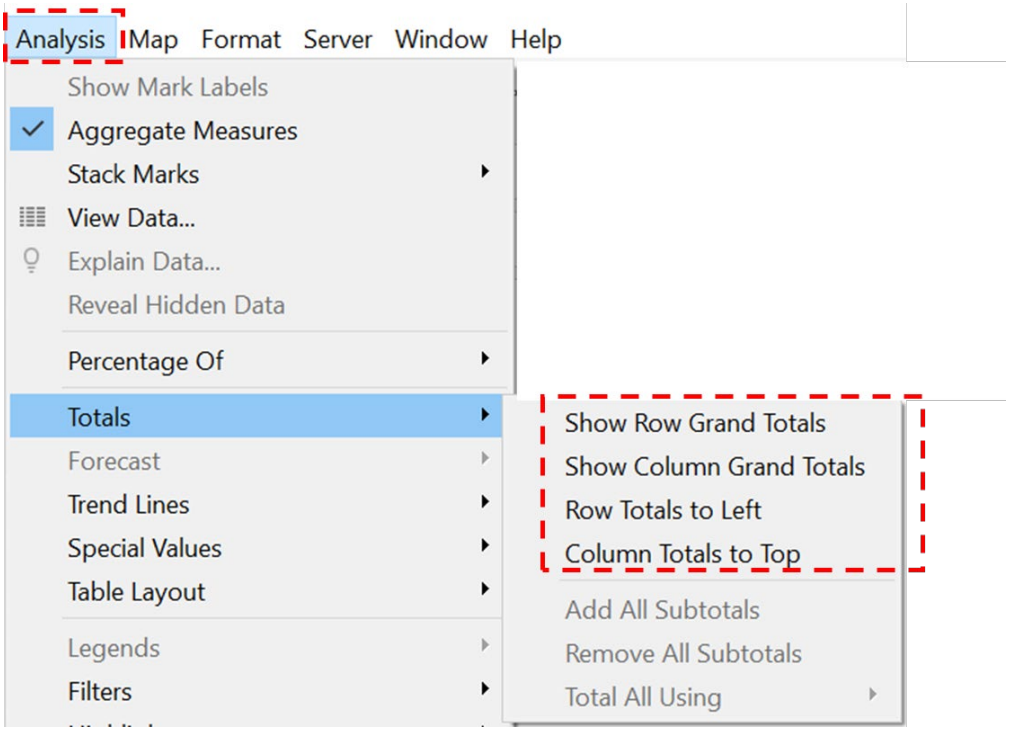


Tips:

Stack bar can be easily displayed into contingency table:



You can compute the total for each rows and columns, as well as grand total from "Analysis" menu.

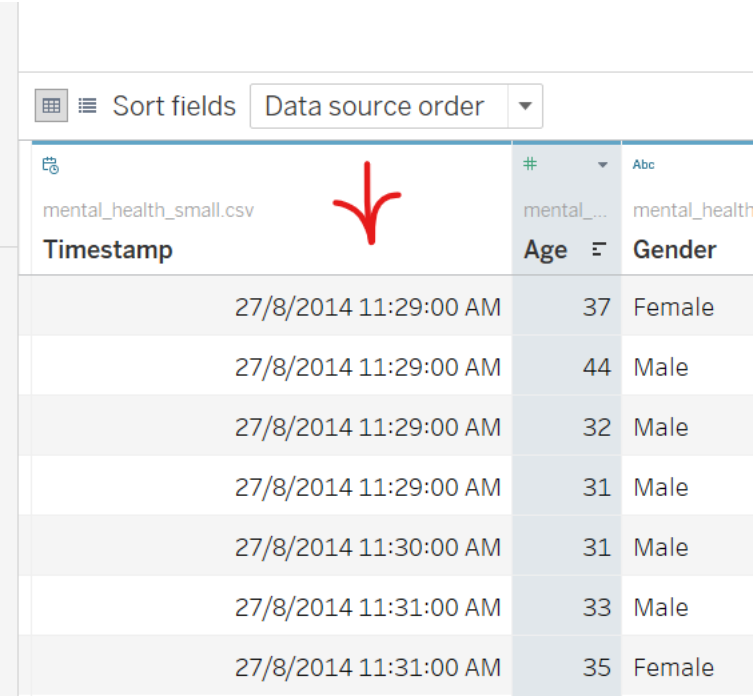


This will give the Cross Tabs:

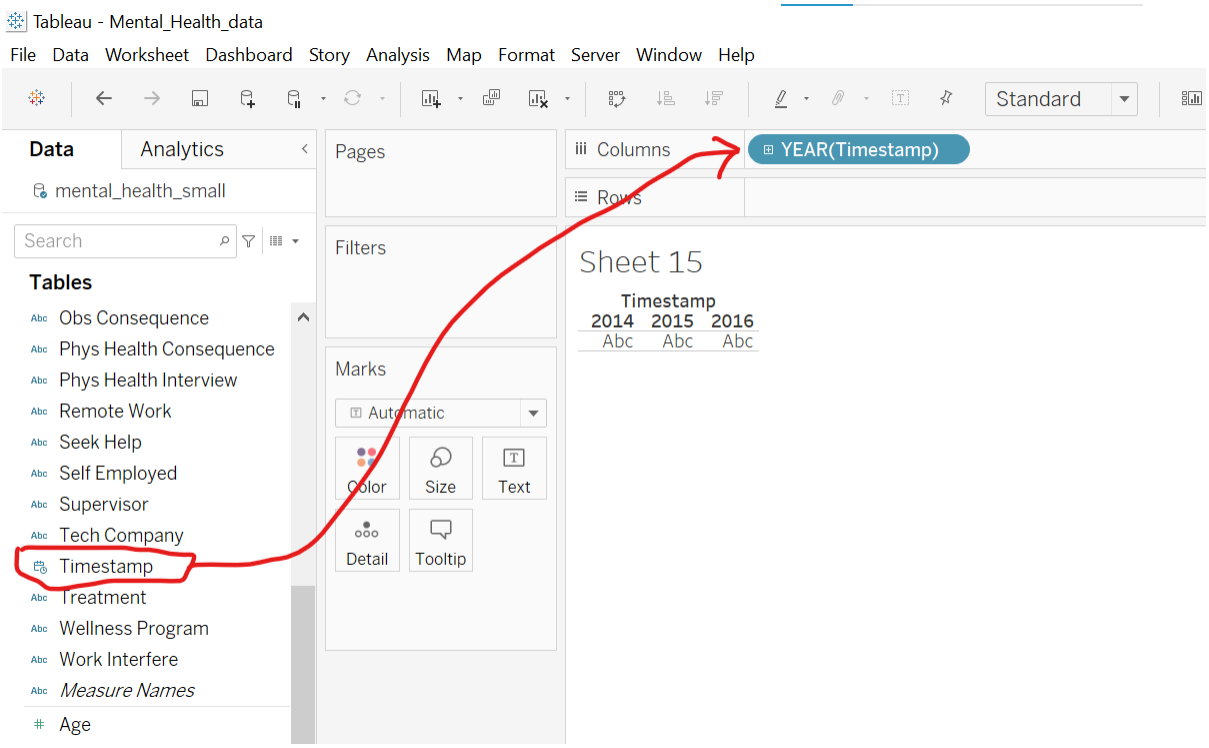
	Family History		
Gender	No	Yes	Grand Total
Female	103	112	215
Male	478	289	767
Grand ..	581	401	982

Plotting variable Vs timeline

The timestamp shows the date and time that participant submitted his/her survey.

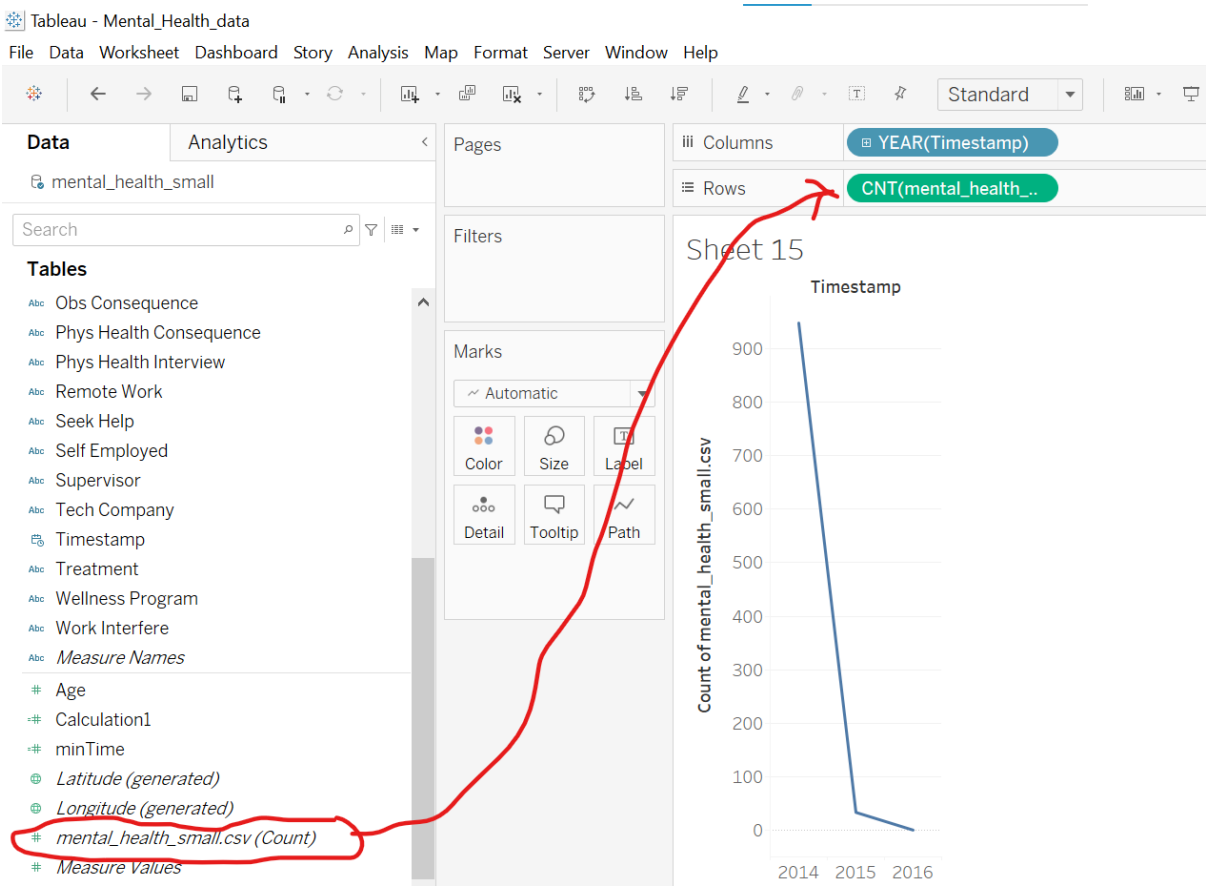


We are interested to know which time of the day when most participants submit their survey. The time that we are interested will be which hour of the day (instead of minutes or seconds). We can drag the Timestamp to the "column", which will be the label for the X-axis.



The table shows there are 3 years, 2014, 2015, 2016.

We can drag the generated count field to the “row”:

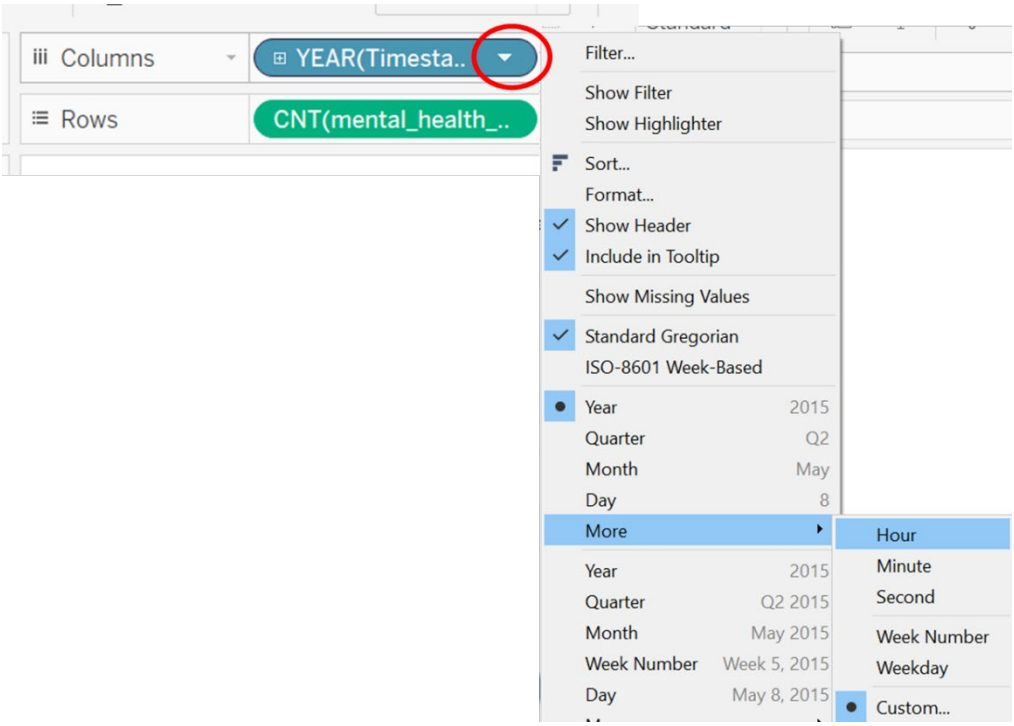


The X-axis shows the 3 years mark, and the Y-axis shows the count from each year.

2014 has the most submissions.

We are interested in which time of the day has the most submissions. The time units will be in hour (not day, minutes nor seconds).

Edit the “timestamp” to “hour”:



Now, you will be able to find which time of the day has the most submissions.

Lab 6

Learning outcomes:

1. Able to handle missing values in the data set
2. Able to perform exploratory data analysis

Exercise 1

Data set: *Tomslee_airbnb.csv*

Create a new workflow, named as "*Airbnb_data*"

This data is retrieved from : http://tomslee.net/category/airbnb-data

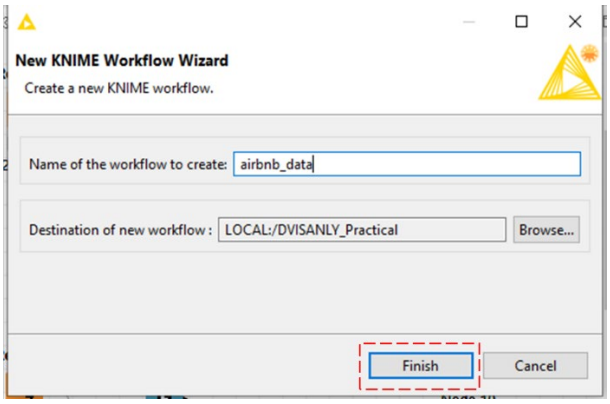
(You can refer to this site for the explanation of the variable used in the data)

Appendix A (next page) explains the variables used in the data.

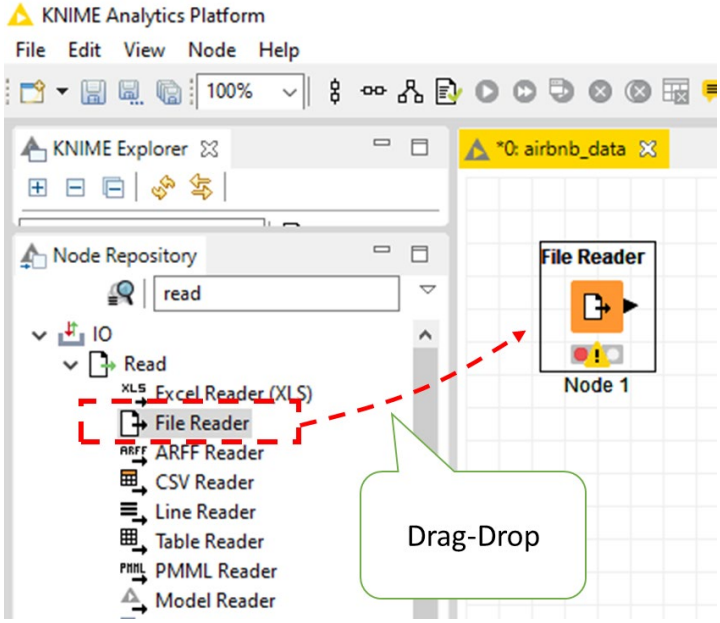
Attributes description of data file **tomslee_airbnb.csv**.

room_id:	A unique number identifying an Airbnb listing. The listing has a URL on the Airbnb web site of http://airbnb.com/rooms/room_id
host_id:	A unique number identifying an Airbnb host. The host's page has a URL on the Airbnb web site of http://airbnb.com/users/show/host_id
room_type:	One of "Entire home/apt", "Private room", or "Shared room"
borough:	A sub region of the city or search area for which the survey is carried out. The borough is taken from a shapefile of the city that is obtained independently of the Airbnb web site. For some cities, there is no borough information; for others the borough may be a number. If you have better shapefiles for a city of interest, please send them to me.
neighbourhood:	As with borough: a sub region of the city or search area for which the survey is carried out. For cities that have both, a neighbourhood is smaller than a borough. For some cities there is no neighbourhood information.
reviews:.	The number of reviews that a listing has received. Airbnb has said that 70% of visits end up with a review, so the number of visits. Note that such an estimate will not be reliable for an individual listing (especially as reviews occasionally vanish from the site), but over a the number of reviews can be used to estimate city as a whole it should be a useful metric of traffic
overall_satisfaction:	The average rating (out of five) that the listing has received from those visitors who left a review.
accommodates:	The number of guests a listing can accommodate.
bedrooms:	The number of bedrooms a listing offers.
price:	The price (in $US) for a night stay. In early surveys, there may be some values that were recorded by month.
minstay:	The minimum stay for a visit, as posted by the host.
latitude and longitude:	The latitude and longitude of the listing as posted on the Airbnb site: this may be off by a few hundred metres. I do not have a way to track individual listing locations with
last_modified:	the date and time that the values were read from the Airbnb web site

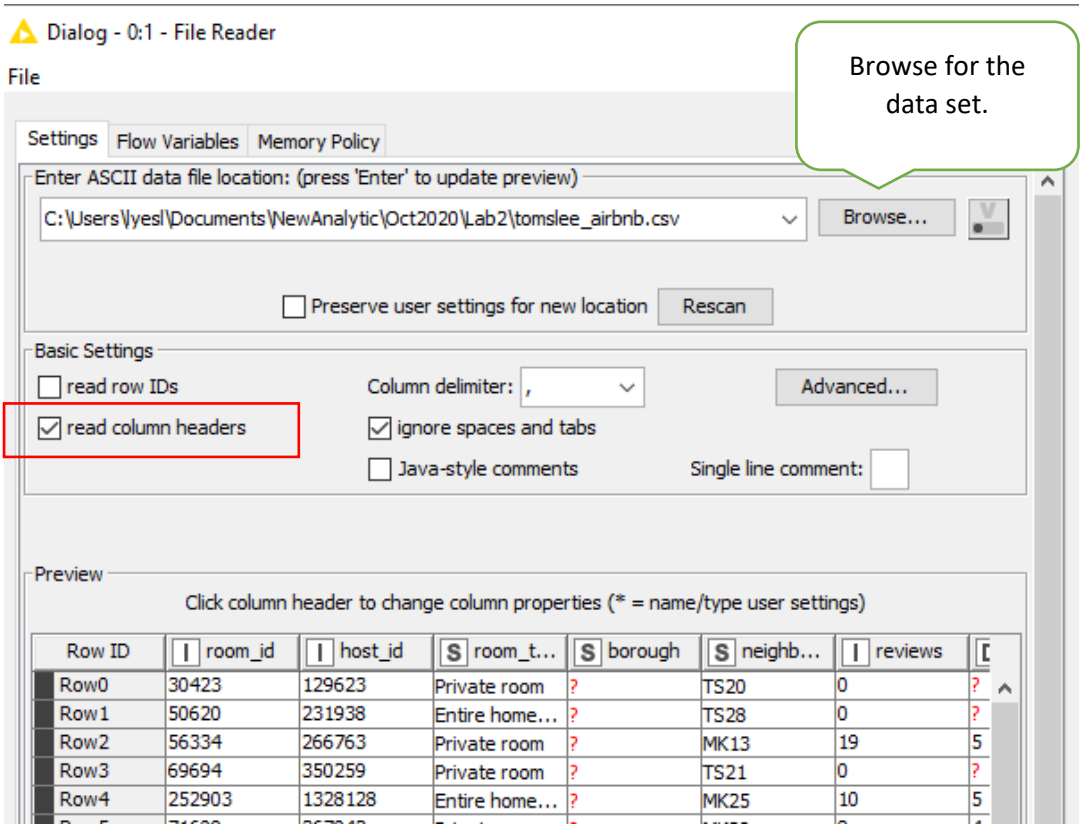
Let's create a new workflow in KNIME.



What is the first node to use?

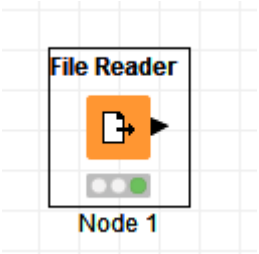


Double-click on the "File Reader" node to configure.



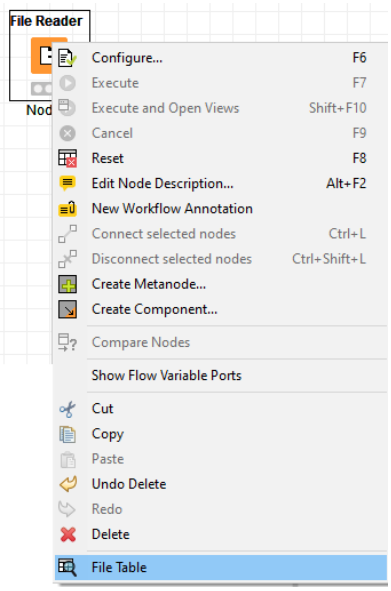
Notice a lot of "?" marks appear in our data. The "?" marks indicate missing value. That is, there are no value in that particular column and row. There seem to be a lot of "missing values".

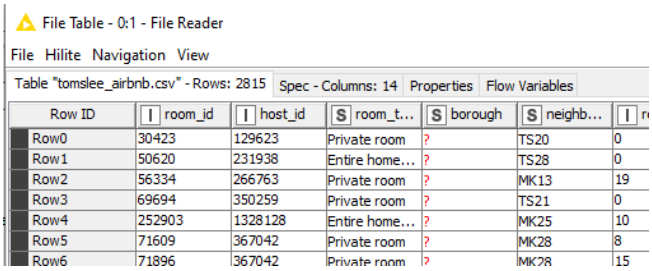
Click "**OK**" and proceed to "**Execute"** the node.



Check that your node should be in "Green" status now.

Let's study the File Table.



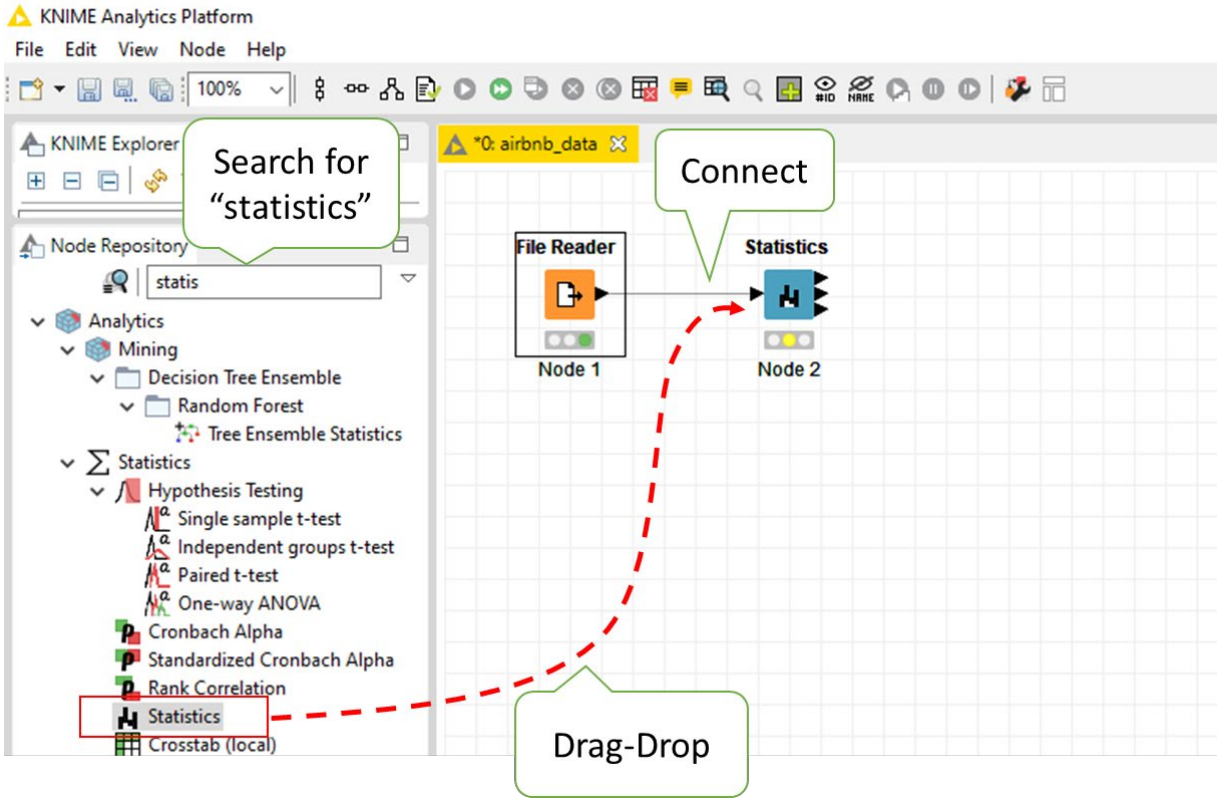


Row ID	room_id	host_id	room_t...	borough	neighb...	
Row0	30423	129623	Private room	?	TS20	0
Row1	50620	231938	Entire home...	?	TS28	0
Row2	56334	266763	Private room	?	MK13	19
Row3	69694	350259	Private room	?	TS21	0
Row4	252903	1328128	Entire home...	?	MK25	10
Row5	71609	367042	Private room	?	MK28	8
Row6	71896	367042	Private room	?	MK28	15

How many observations are there? _____

How many variables? _____

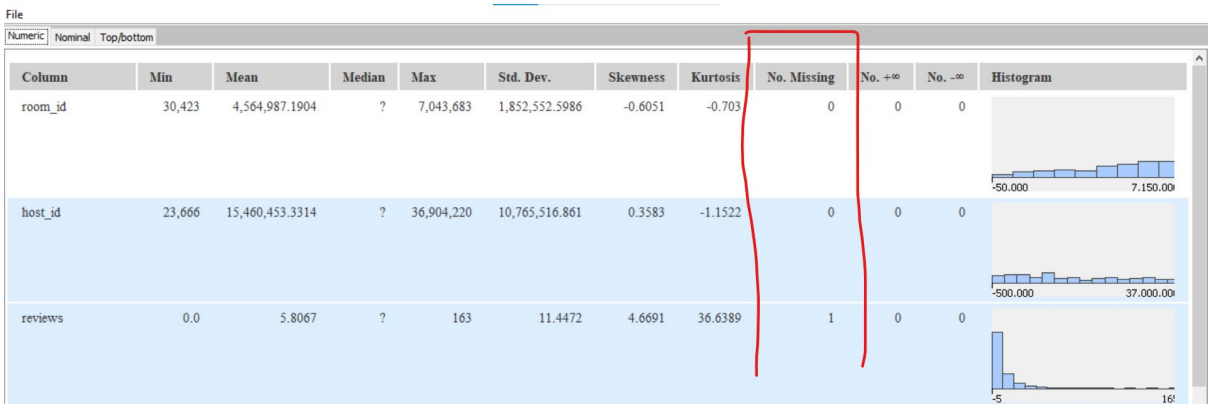
We need to know in total how many missing values, to decide our strategy to discard the columns or rows. Statistics node should be able to show how many missing values.



There is no need to configure the Statistics.

Proceed to "**Execute and Open Views**".

Check the column labelled: No. Missing Values (number of missing values)



Identify which variables have missing values:

Name of the variable	Number of missing values

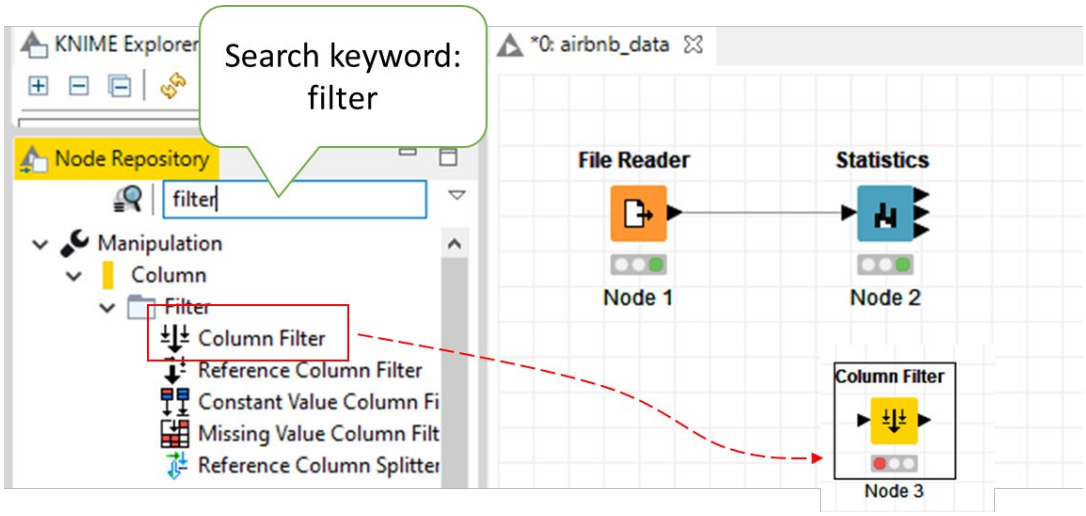
What happen if we perform remove rows containing missing values?

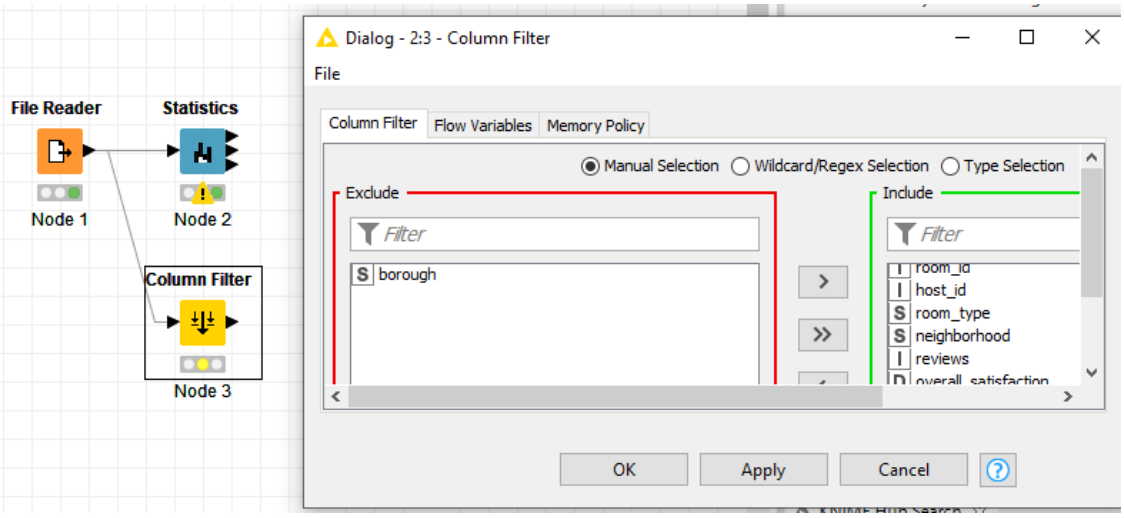
If we remove rows containing missing values, we will be left with no rows. No data.

All the rows will be removed because all the rows contain missing value in Borough variable.

So, we must first remove the column "borough".

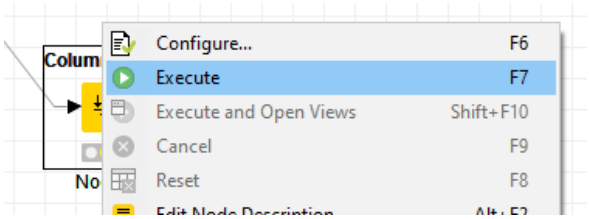
Use Column filter node.



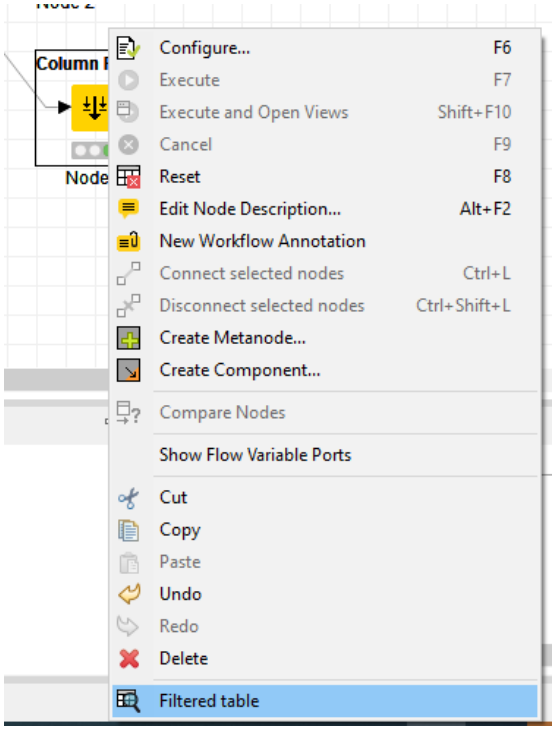


Configure Column Filter node. Click "**OK**" to continue.

Execute the Column Filter node.

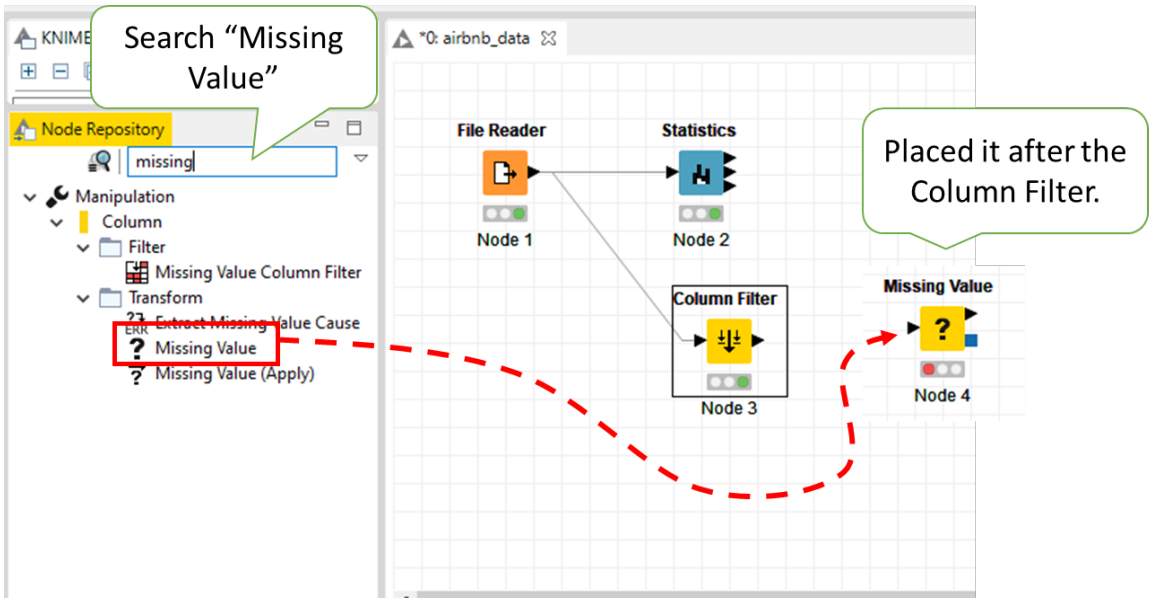


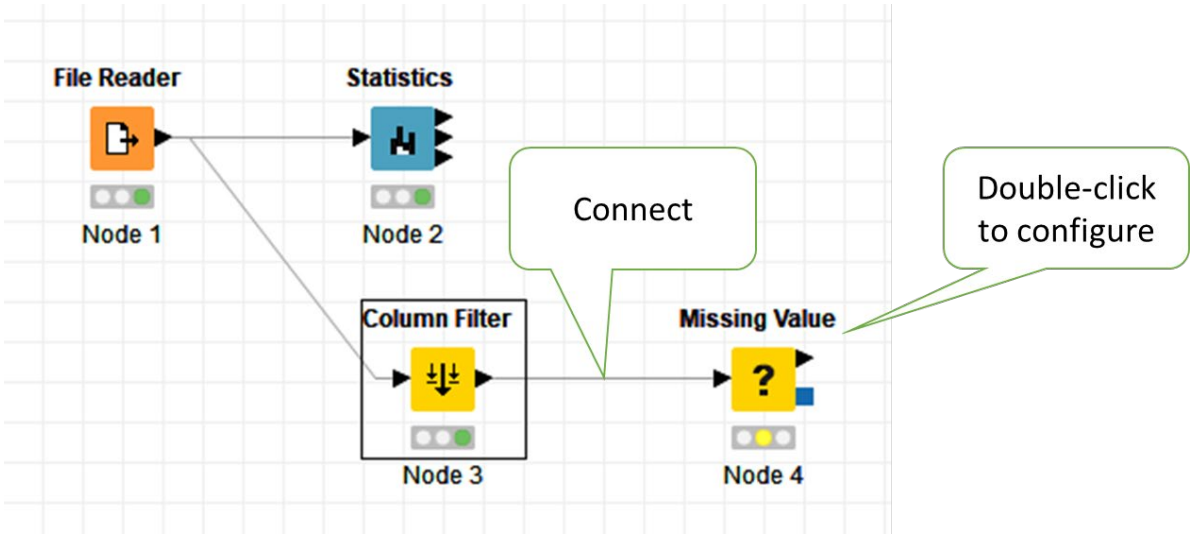
You can check the outcome by right-click on the "Column Filter" and select "**Filtered Table**":

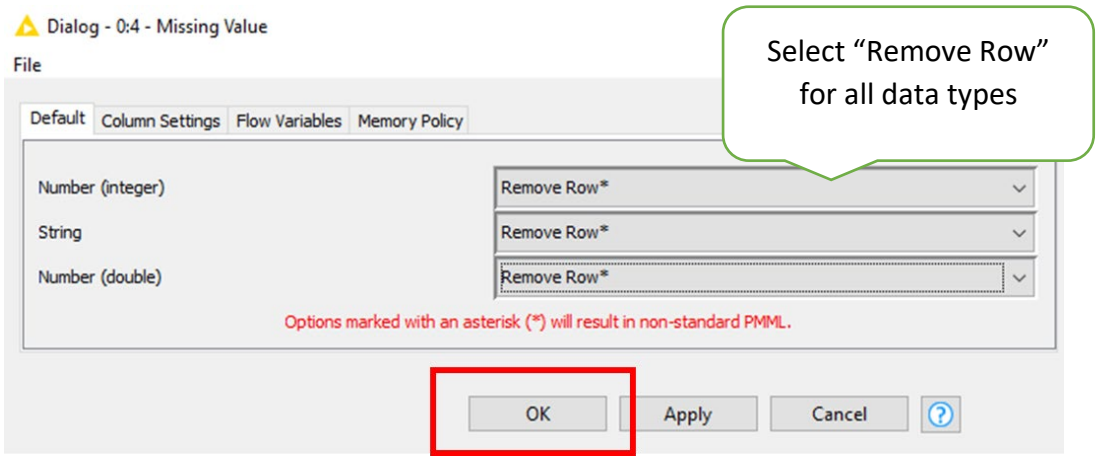


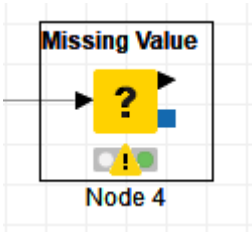
The number of columns should reduce to 13 now.

Proceed to add "Missing Value" node.


Search "Missing Value"
Node Repository
missing
Manipulation
Column
Filter
Missing Value Column Filter
Transform
Missing Value
Missing Value (Apply)
*0: airbnb_data
File Reader
Node 1
Statistics
Node 2
Column Filter
Node 3
Missing Value
Node 4
Placed it after the Column Filter.


File Reader
Node 1
Statistics
Node 2
Column Filter
Node 3
Missing Value
Node 4
Connect
Double-click to configure


Dialog - 0:4 - Missing Value
File
Default
Column Settings
Flow Variables
Memory Policy
Number (integer)
Remove Row*
String
Remove Row*
Number (double)
Remove Row*
Options marked with an asterisk (*) will result in non-standard PMML.
OK
Apply
Cancel
Select "Remove Row" for all data types



Note the error in the console.

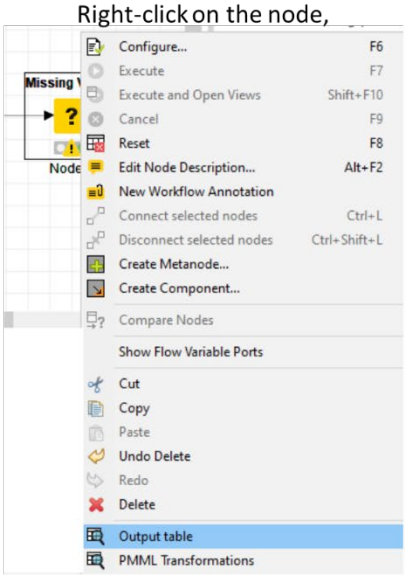
This missing value handler does not produce standard PMML 4.2!

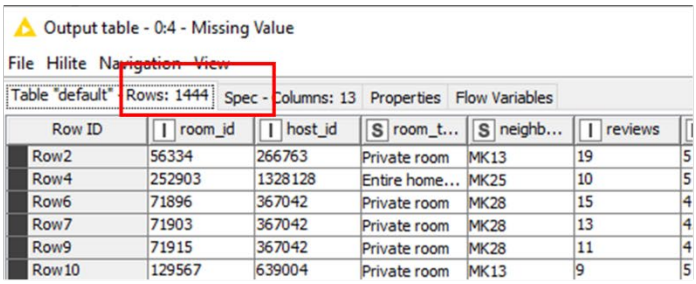
As we are not going to export to PMML format, is fine. Ignore the error.

**PPML – Predictive Model Markup Language*

You can check the output table:

Right-click on the node,





Row ID	room_id	host_id	room_t...	neighb...	reviews
Row2	56334	266763	Private room	MK13	19
Row4	252903	1328128	Entire home...	MK25	10
Row6	71896	367042	Private room	MK28	15
Row7	71903	367042	Private room	MK28	13
Row9	71915	367042	Private room	MK28	11
Row10	129567	639004	Private room	MK13	9

The number of observations (rows) is 1444.

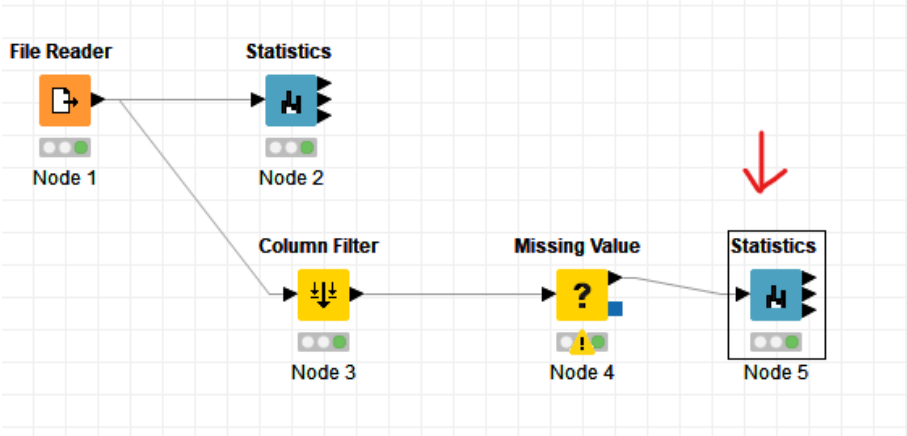
Reduced from 2815 to 1444, lose 1371 rows of observations.

Final number of observations is still large enough for further analysis.

Check:

How can we verify that there are no more missing data?

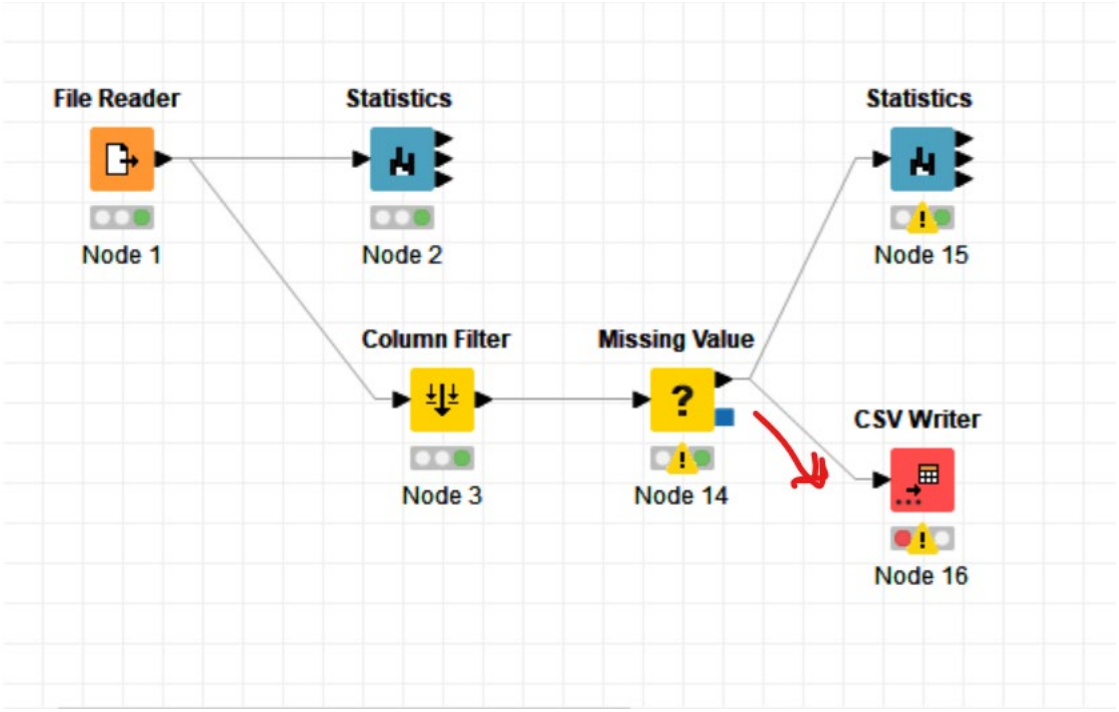
Yes, add statistics node after the “Missing Value” node to check.

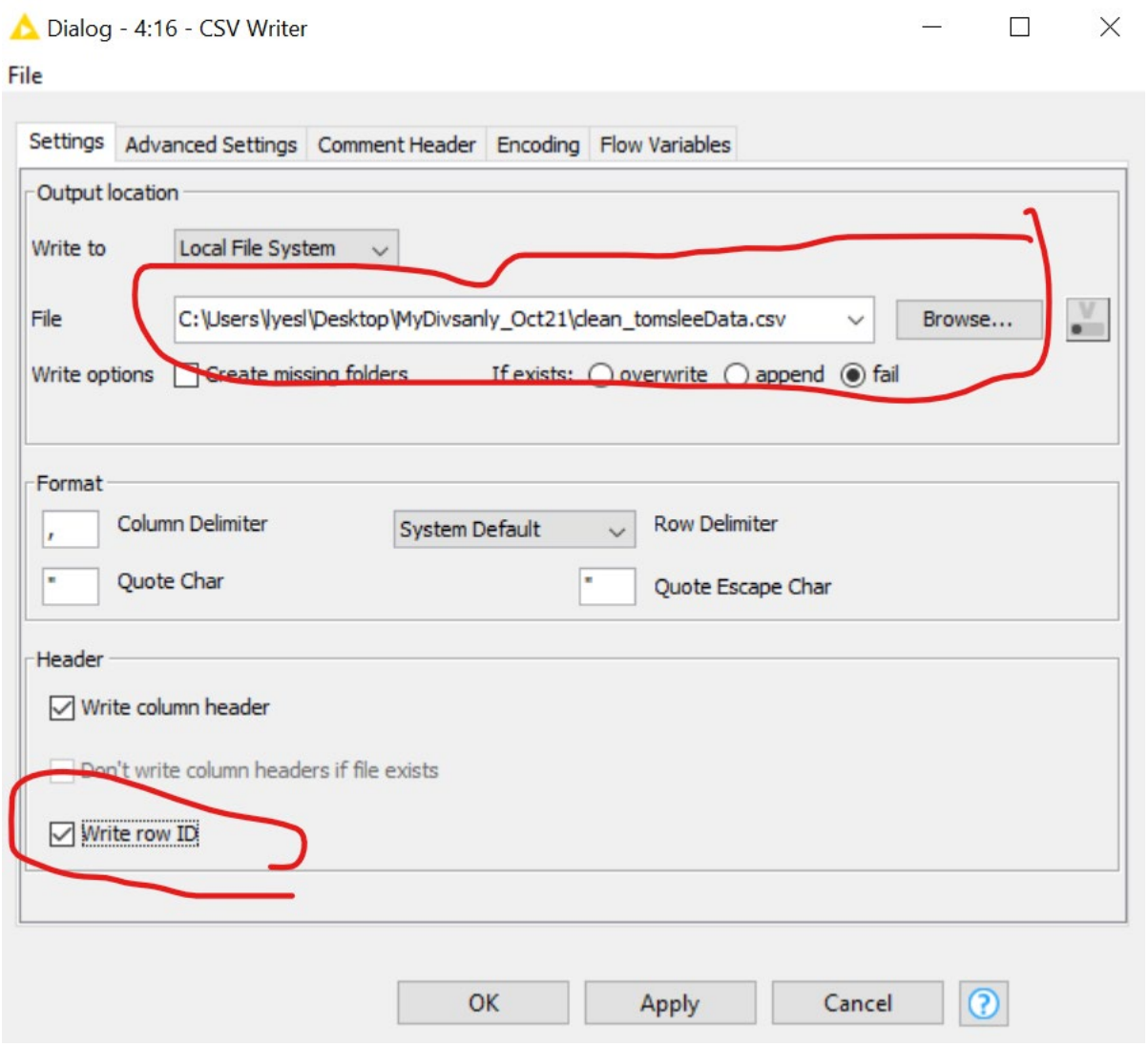


The column under “No. Missing value” should show “0” for all variables.

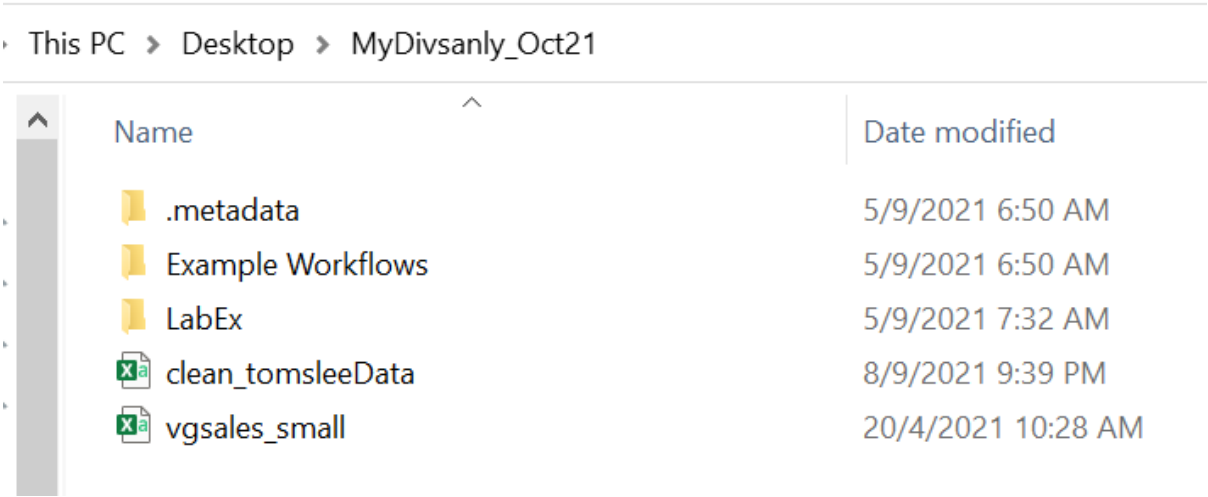
Next, export the clean data to a new CSV file.

Save the new file as “***clean_tomsleeData***”





Execute the **CSV Writer** node.



Read this clean data into TABLEAU.

The data is ready for Exploratory Data Analysis.

Answer the following questions:

1. **How many surveyance gave "Overall Satisfaction" ≥ 4?**
   (*plot bar chart of Count Vs overall_satisfactioin in TABLEAU)

2. **What is the percentage of responses with "Overall Satisfaction"= 5?**
   (*plot a pie-chart)

3. **What is the distribution of the "price"?**

   (hint: support with box-plot and histogram)

Data Driven Decision:

Mr Lim (Finding a room)

How would you advice?

Lab 7

Learning outcomes:

Able to perform the following data transformation methods:

1. Use **String to Date&Time** node to convert variable to correct date&Time format
2. Use **Rule Engine** node to replace values with alternate String
3. Use **Concatenate** node to append data
4. Use **Calculated Field** in TABLEAU to reshape data

Exercise 1

Data set: *SensorData_setA.txt*

This dataset represents ambient data collected from smart homes occupied by volunteer residents. Data are collected continuously while residents perform their normal daily routines. Ambient PIR motion sensors, door/temperature sensors, and light sensors are placed throughout the homes at locations that are related to specific activities of daily living of the volunteers.

Refer to Appendix A to understand what do the variables measure.

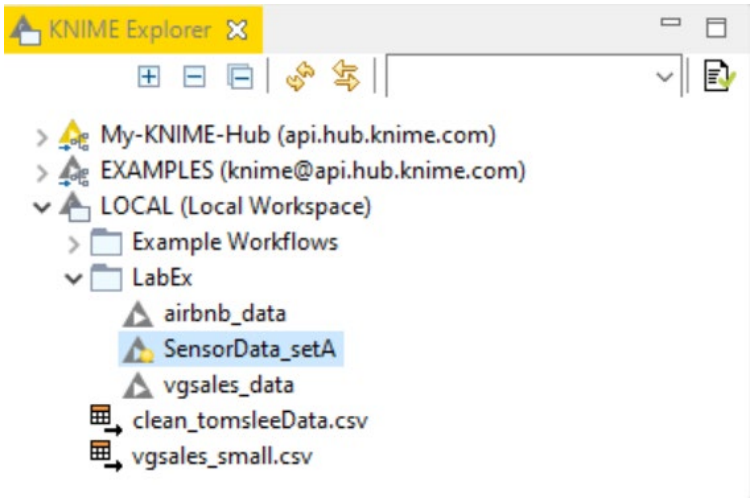
Appendix A

Attribute	**Attribute description**
DateTime	YYYY-MM-DD HH:MM:SS.XXXXXX 24-hour format
SensorID	ID which identify the sensor: • DXXX Door Sensor • LSXXX Light Sensor • MXXX Motion Detector • MAXXX Motion Area Detector • TXXX Temperature
Translate01	Room-level sensor location: • Kitchen • Ignore • LivingRoom • Bedroom • Bathroom • DiningRoom • OutsideDoor • WorkArea • Hall
Translate02	Location which sensor is aimed at: • Kitchen • Ignore • LivingRoom • Bathroom • DiningRoom • Closet • Hall • Bedroom • WorkArea • Bed • FrontDoor • Refrigerator • BathroomTemp • LivingRoomTemp • FrontDoorTemp

Value	Value from the sensor depending on the type: • Control4-Door : OPEN or CLOSE • Control4-LightSensor : Integer values ranging from 0 to 100 (pitch black to very bright) • Control4-Motion : ON or OFF • Control4-MotionArea : ON or OFF • Control4-Temperature : a decimal in Celsius with 0.5 degrees Celsius accuracy
Sensor Type	Various sensor type : • Control4-Door • Control4-LightSensor • Control4-Motion • Control4-MotionArea • Control4-Temperature
Activity	Human activities (the activities were annotated manually) • Cook_Breakfast • Watch_TV • Personal_Hygiene • Dress • Wash_Breakfast_Dishes • Wash_Dishes • Drink • Work_At_Table • Phone • Cook_Lunch • Toilet • Enter_Home • Sleep • Leave_Home • Step_Out • Cook_Dinner • Eat_Breakfast • Read • Eat_Dinner • Morning_Meds • Bed_Toilet_Transition • Eat_Lunch • Groom • Wash_Dinner_Dishes • Relax • Take_Medicine • Bathe

Create a new workflow, named as "*SensorData_setA*"

(show me how)



Use **File Reader**, configure and execute.

Data Cleaning

Check for any missing values.

How many observations are there? _____

Perform data cleaning.

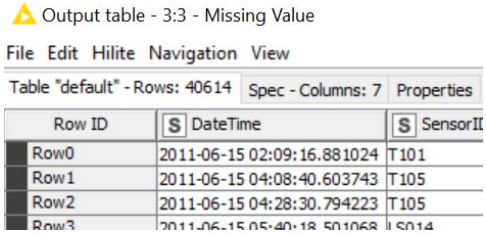
(Show me how)

Show screen shot of your workflow. Explain what is removed and how many observations left.

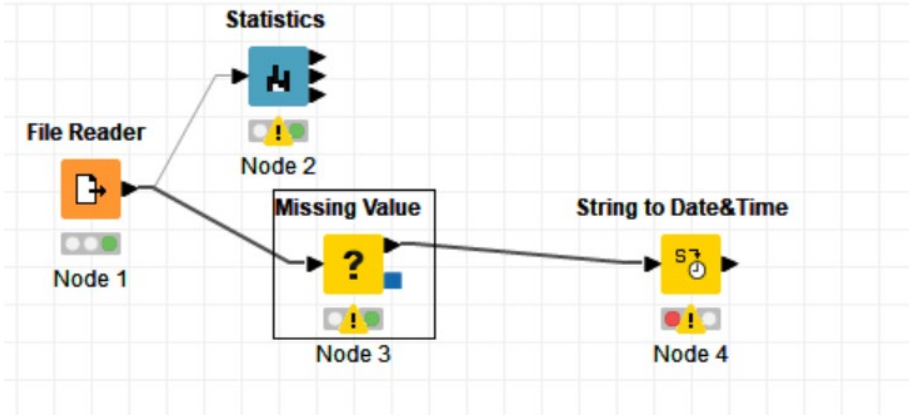
String to Date&Time Transformation

Are all data types correct?

DateTime variable is not in correct data type.

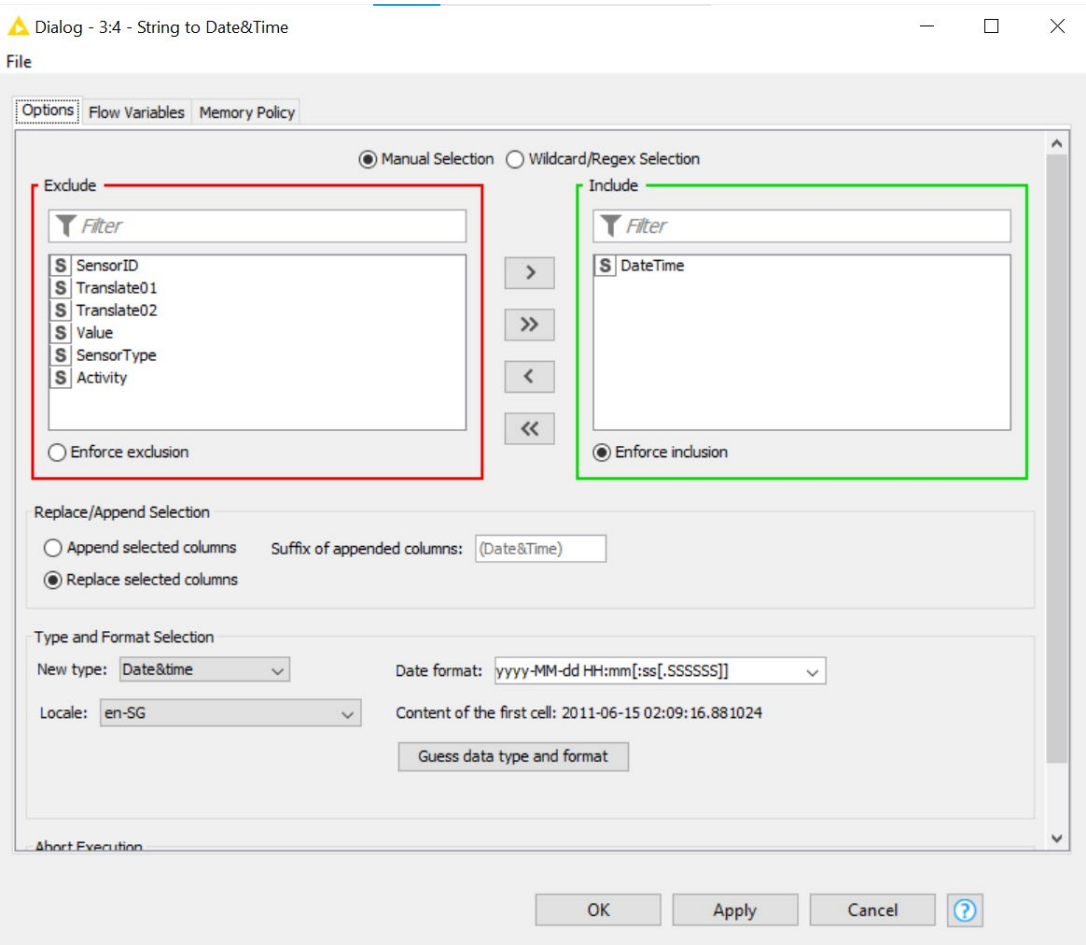


Use String to Date&Time node.



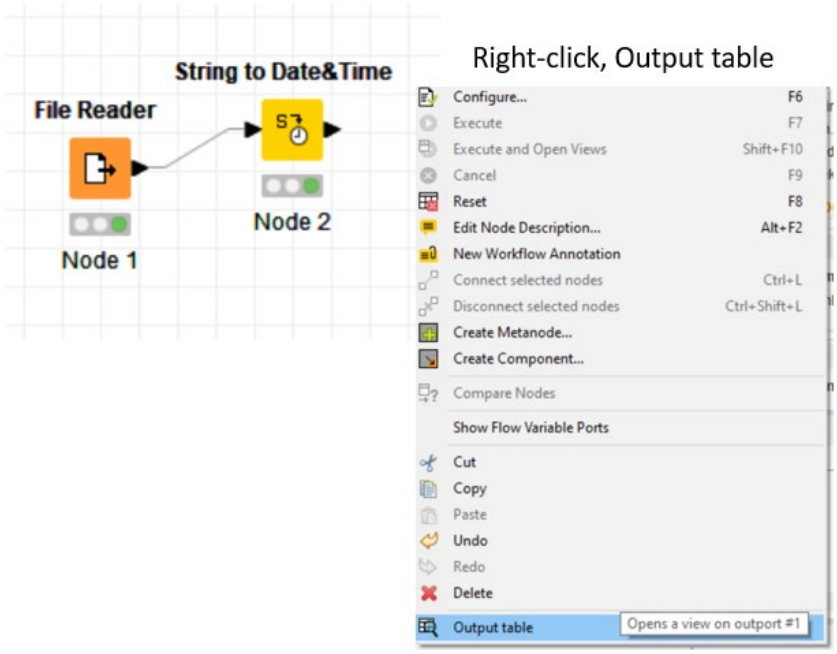
Configure "String to Date&Time" node to include *DateTime* variable.

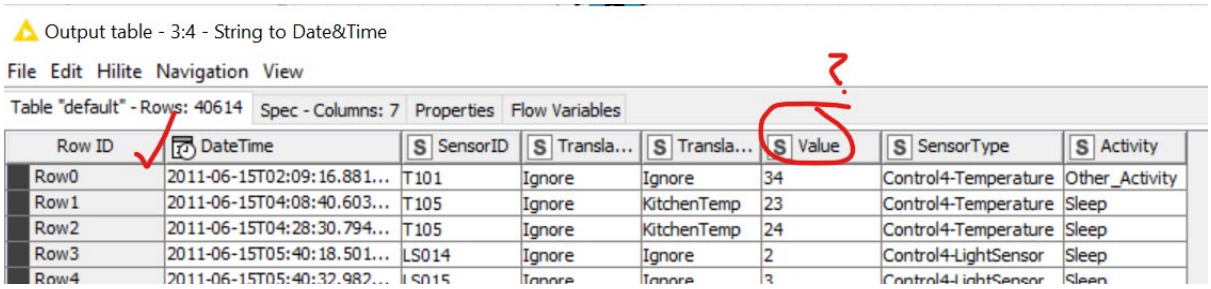
(Show me how)



Execute the **String to Date&Time** node.

Check the output file:




Output table - 3:4 - String to Date&Time

File Edit Hilite Navigation View

Table "default" - Rows: 40614 | Spec - Columns: 7 | Properties | Flow Variables

Row ID	DateTime	SensorID	Transla...	Transla...	Value	SensorType	Activity
Row0	2011-06-15T02:09:16.881...	T101	Ignore	Ignore	34	Control4-Temperature	Other_Activity
Row1	2011-06-15T04:08:40.603...	T105	Ignore	KitchenTemp	23	Control4-Temperature	Sleep
Row2	2011-06-15T04:28:30.794...	T105	Ignore	KitchenTemp	24	Control4-Temperature	Sleep
Row3	2011-06-15T05:40:18.501...	LS014	Ignore	Ignore	2	Control4-LightSensor	Sleep
Row4	2011-06-15T05:40:32.982...	LS015	Ignore	Ignore	3	Control4-LightSensor	Sleep

The data type for *DateTime* is correct now.

Why is the variable "*Value*" read as STRING?

There are supposed to be numerical values.

Refer to the Appendix to understand what does "Value" measure.

Value	Value from the sensor depending on the type: • Control4-Door : OPEN or CLOSE • Control4-LightSensor : Integer values ranging from 0 to 100 (pitch black to very bright) • Control4-Motion : ON or OFF (The sensor will instantly send an ON message when detecting motion. 1.25 seconds after it no longer observes motion the sensor will send OFF.) • Control4-MotionArea : ON or OFF • Control4-Temperature : a decimal in Celsius with 0.5 degrees Celsius accuracy

"*Value*" can be ON or OFF when Control4-Motion sensor type is triggered. It can also be OPEN or CLOSE when Control4-Door sensor is triggered.

What does the numbers in the "*Value*" measure?

They could be the reading from Temperature and light sensors.

So, depend on which sensor is triggered, the "*Value*" will capture the respective value.

If temperature sensor is triggered, the "*Value*" will show numerical number.

If Door sensor is triggered, the "*Value*" will capture ON and after 1.25seconds, capture as OFF.

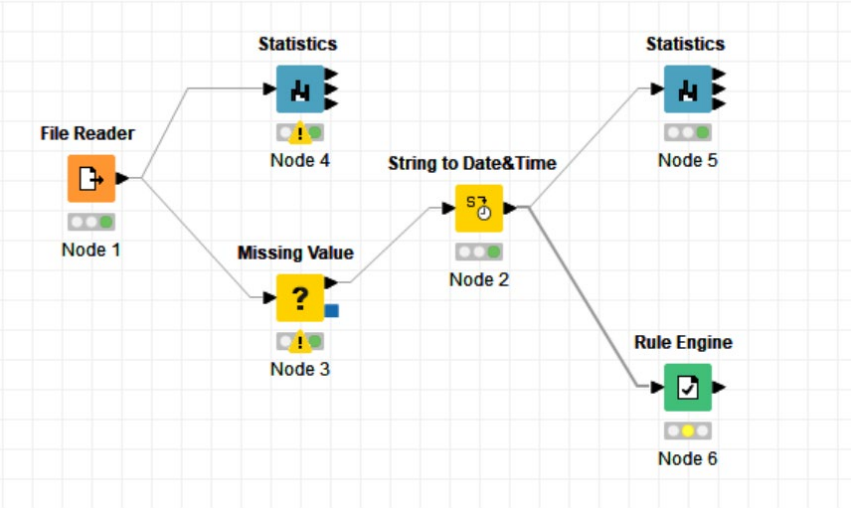
To analyse the numerical readings (eg. Temperature and lighting level), we can assign a number to ON, OFF, OPEN and CLOSE.

For example, we will assign 800 for OPEN, and 888 for CLOSE, 900 for ON and 999 for OFF.

This can be done using **Rule Engine node.**

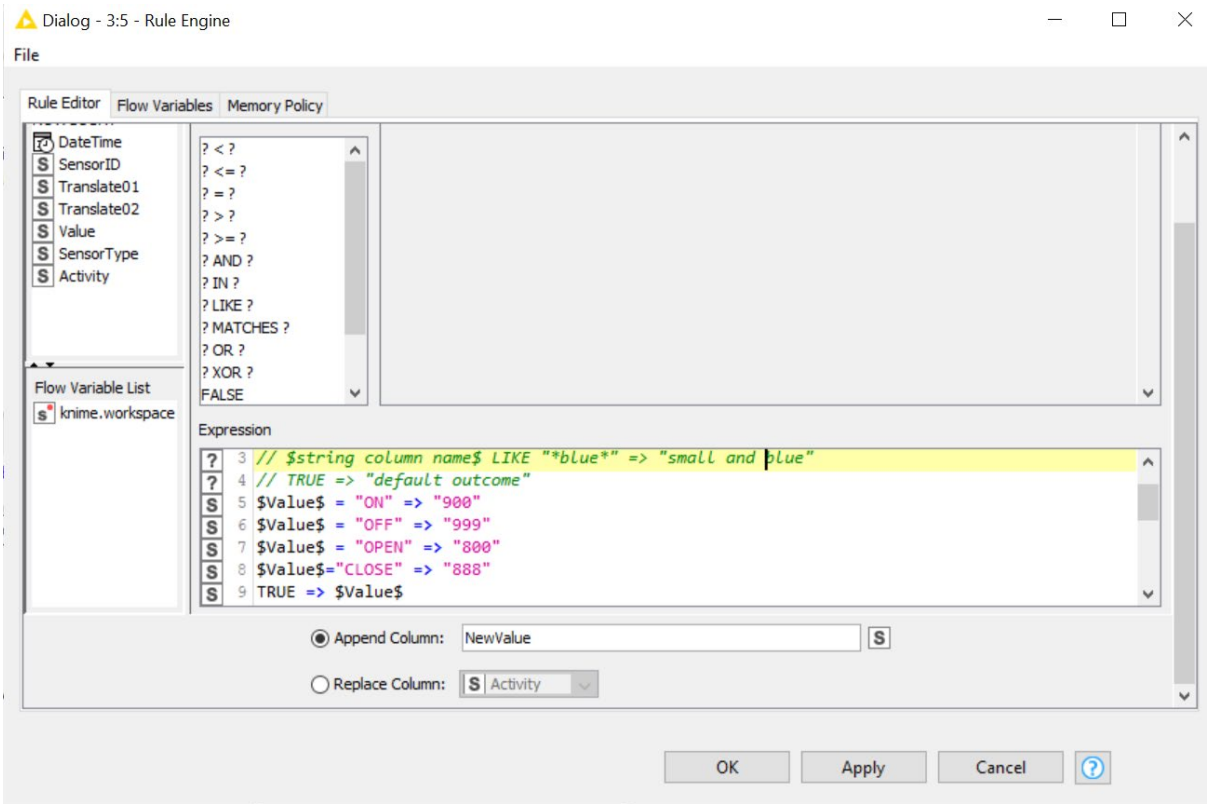
Data Transformation using Rule Engine

Add **Rule Engine** node to the workflow.



Configure the **Rule Engine** node:

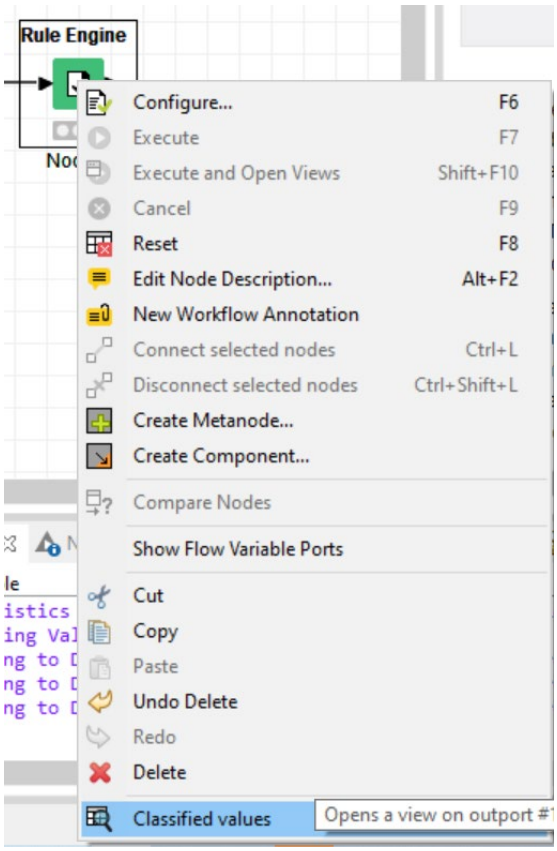
(Show me how)

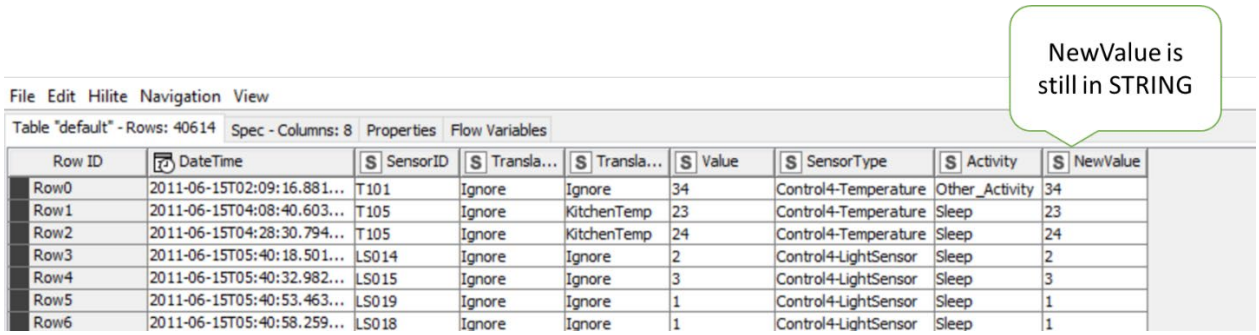


Give a new variable name "*NewValue*".

Execute the node.

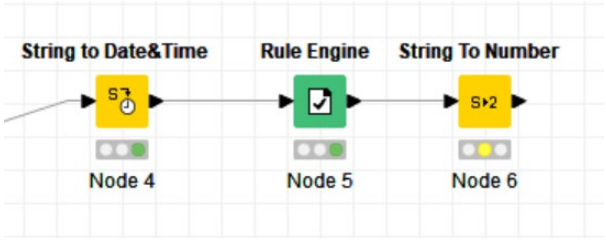
You can check the result from *Classified value*:



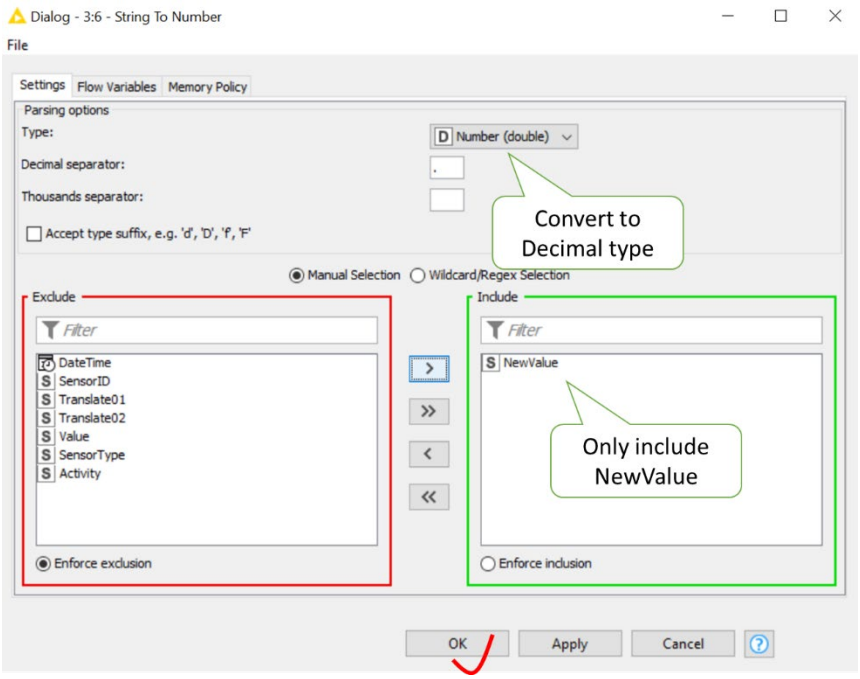


Row ID	DateTime	SensorID	Transla...	Transla...	Value	SensorType	Activity	NewValue
Row0	2011-06-15T02:09:16.881...	T101	Ignore	Ignore	34	Control4-Temperature	Other_Activity	34
Row1	2011-06-15T04:08:40.603...	T105	Ignore	KitchenTemp	23	Control4-Temperature	Sleep	23
Row2	2011-06-15T04:28:30.794...	T105	Ignore	KitchenTemp	24	Control4-Temperature	Sleep	24
Row3	2011-06-15T05:40:18.501...	LS014	Ignore	Ignore	2	Control4-LightSensor	Sleep	2
Row4	2011-06-15T05:40:32.982...	LS015	Ignore	Ignore	3	Control4-LightSensor	Sleep	3
Row5	2011-06-15T05:40:53.463...	LS019	Ignore	Ignore	1	Control4-LightSensor	Sleep	1
Row6	2011-06-15T05:40:58.259...	LS018	Ignore	Ignore	1	Control4-LightSensor	Sleep	1

But the "*NewValue*" is still in STRING data type. Add a **String to Number** node:



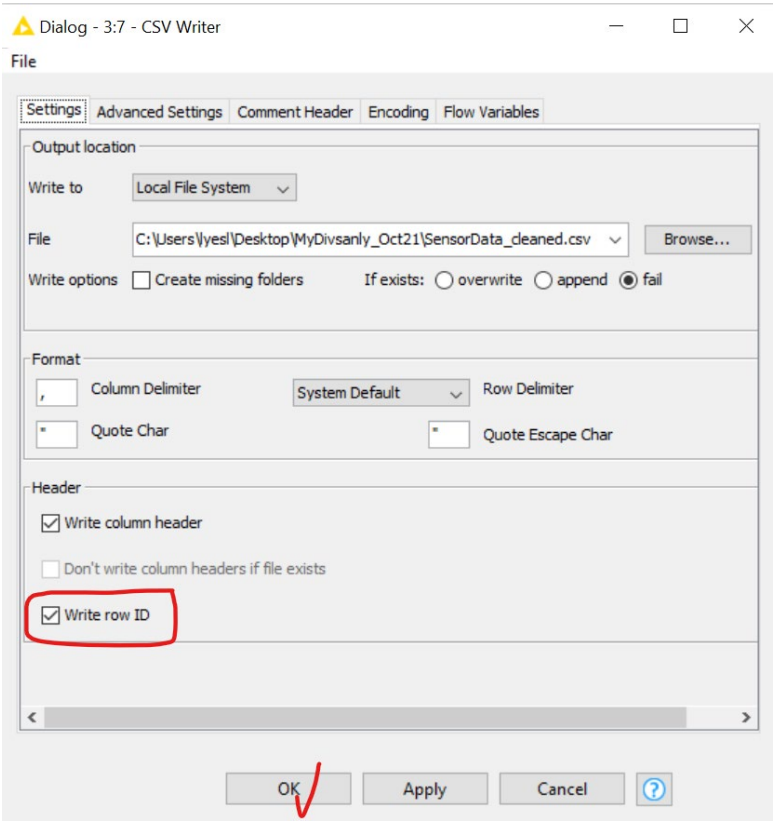
Configure **String to Number** node:



Execute the node.

The data is ready for analysis in TABLEAU.

Add a **CSV Writer** and save the clean and transformed data to "*SensorData_cleaned*".



Remember to **execute** the node.

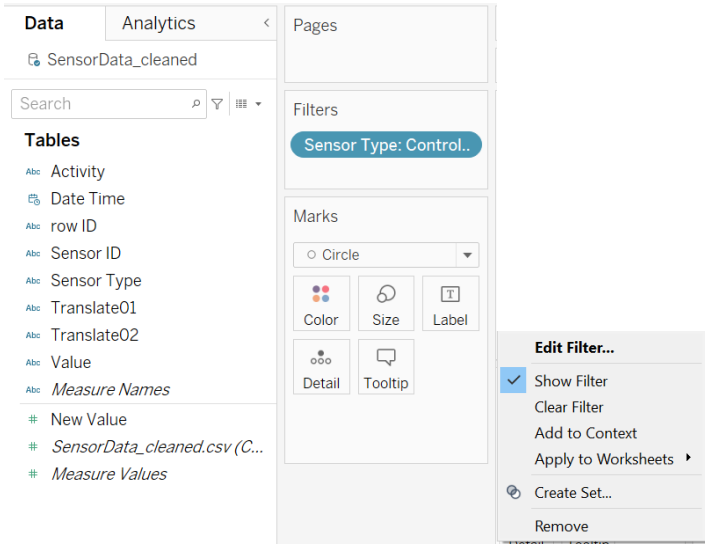
Read the cleaned SensorData file into TABLEAU.

Exploratory Data Analysis

Ex1-Question 1

There are 4 temperature sensors in the room. They are labelled as SensorID T101, T102, T103, T104. Show the box-plot distribution of temperature values for each Temperature Sensor. Which SensorID has the highest temperature reading?

Hint: Use the Filter card to filter only Control4-Temperature group.



Exercise 2

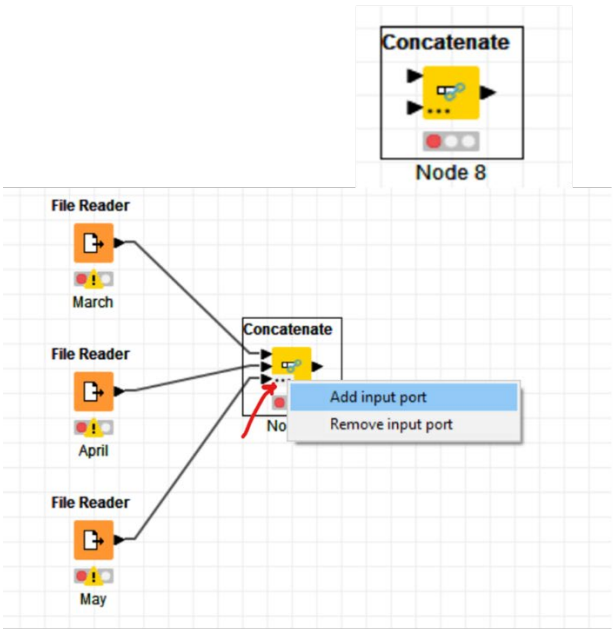
Data set: daily rainfall data for March, April and May

- DAILYDATA_S24_202103
- DAILYDATA_S24_202104
- DAILYDATA_S24_202105

Objective:

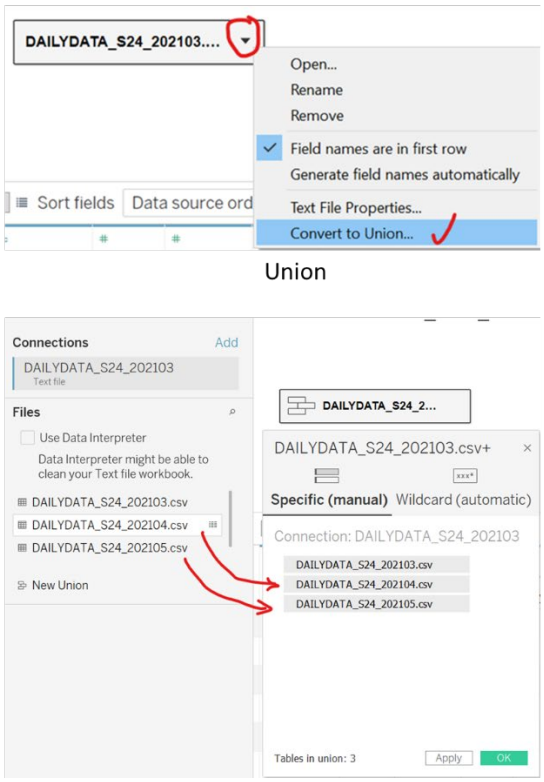
To append the 3 months into one single file

KNIME



OR

TABLEAU



Exploratory data analysis

Ex2-Question 1

Append the 3 files into one single file using the Concatenate node.

Add a Linear Correlation node in KNIME,after Concatenate node.

Show your workflow and the Correlation Matrix. Compare the "Daily Rainfall Total", "Mean Temperature" and "Mean Windspeed".

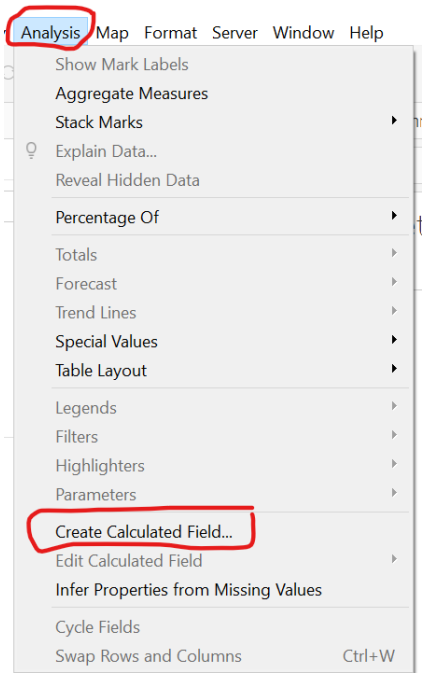
Describe the insights.

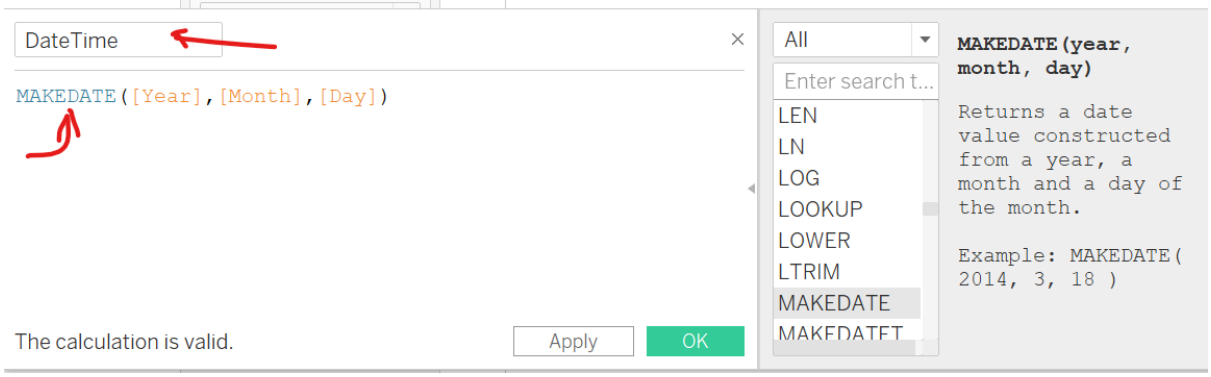
Show me how

Ex2-Question 2

Append the 3 files in TABLEAU.

Create a calculated field to make a datetime field:





(show me how)

Plot a line graph showing the average rainfall total vs the Day of the Week. Which day of the week has the lowest average rainfall total?

Show a screen shot of your graph.

(show me how)

Exercise 3 – Using Calculated Field

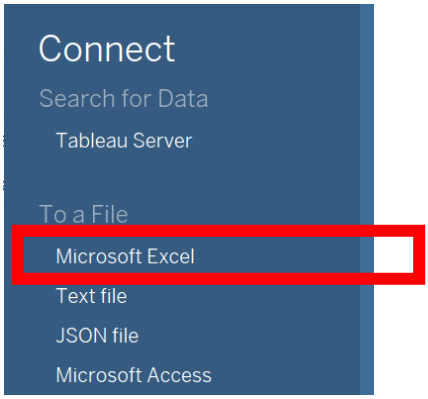
We will now read the CSV file *Purchase_History.xlsx* into Tableau Desktop.

Specific questions that we would like to answer in this session:

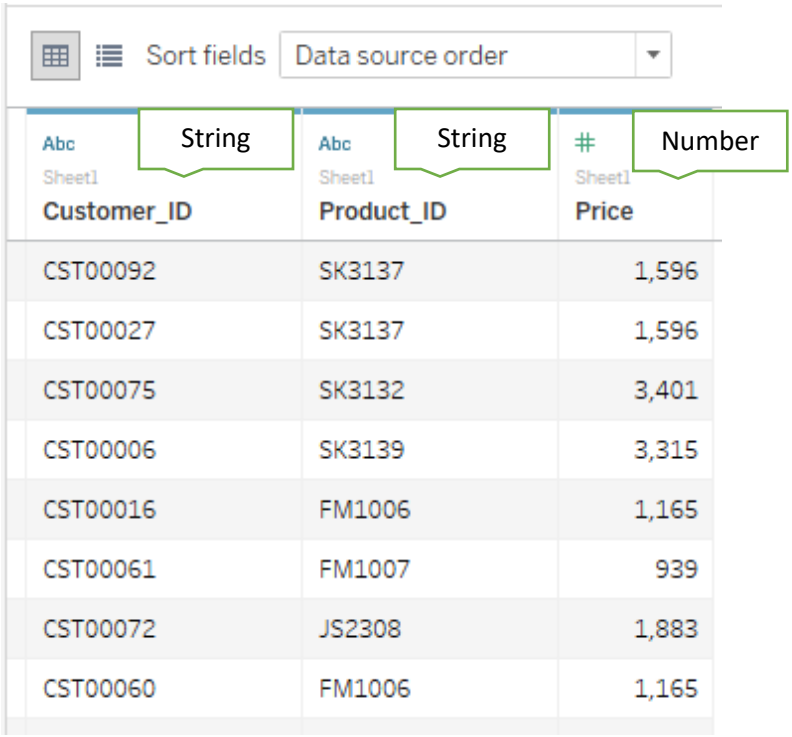
1. Who are the top customers in terms of amount spent?
2. Who are the top customers in terms of number of products purchased?
3. Who are the top Customers in terms of number of distinct products?
4. How many products are there for each category?
5. What is the average amount spent for each product category?

Launch Tableau desktop.

Click on Connect → Microsoft Excel



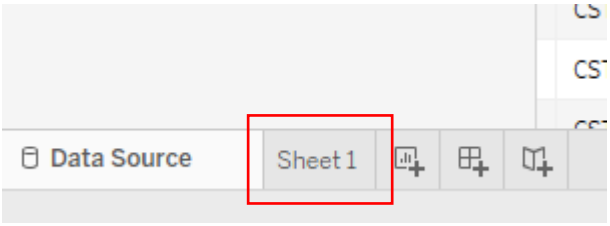
Locate the *Purchase_History.xlsx* file and select Open.
You will see a preview of the data loaded into Tableau Desktop.



Customer_ID	Product_ID	Price
CST00092	SK3137	1,596
CST00027	SK3137	1,596
CST00075	SK3132	3,401
CST00006	SK3139	3,315
CST00016	FM1006	1,165
CST00061	FM1007	939
CST00072	JS2308	1,883
CST00060	FM1006	1,165

There are about 1032 rows of observation, and 3 variables/fields.

Click on “Sheet 1” at the bottom-left of the program to begin.



Question 1:
Who are the top 3 customers in terms of amount spent?

Answer:

Question 2:

Who are the top 3 customers in terms of number of products purchased?

Answer:

Question 3:

Who are the top 3 Customers in terms of number of **distinct** products?

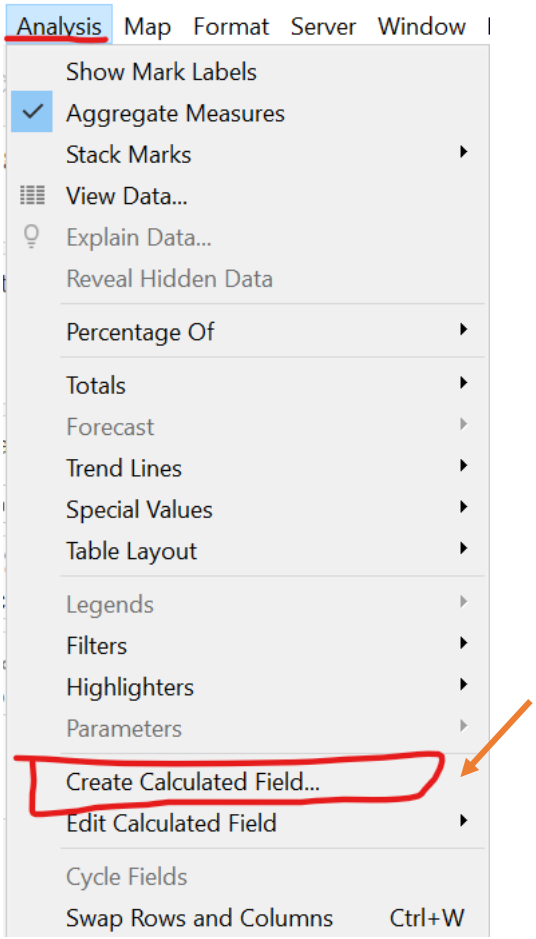
Answer:

We can group the Product_ID base on the first 2 letters.

Product_ID begins with	Product Category
FM	FM Group
SK	SK Group
TF	TF Group
JS	JS Group

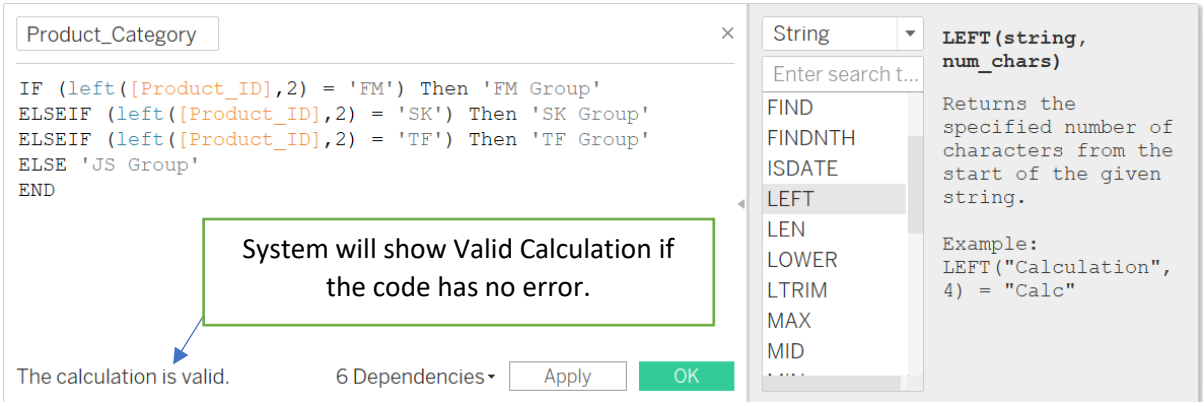
We are going to create another field(column) called Product Category, to group all products begin with the same first 2 letters together.

We will create a Calculated field:



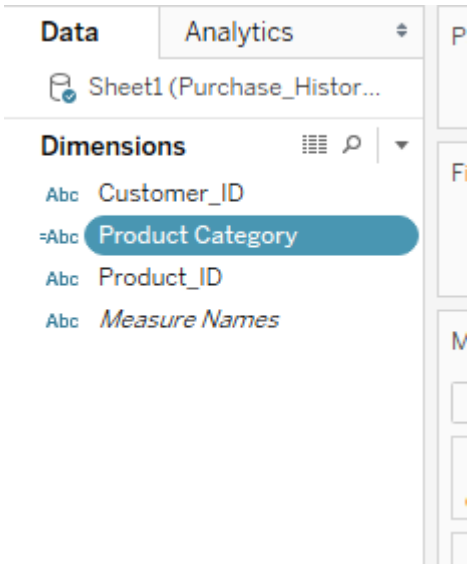
Give the name of the calculated field “Product Category”. Enter the code below:

```
IF (left([Product_ID],2) = 'FM') Then ’FM roup”
ELSEIF (left([Product_ID],2) = 'SK') Then ‘SK Group’
ELSEIF (left([Product_ID],2) = 'TF') Then 'TF Group’
ELSE 'JS Group'
END
```



Click “**OK**” to continue.

Notice a new field appear in the dimension area:



Let’s add another Sheet for Question 4.

Question 4:
How many products are there for each category?

Answer:

Discussion: What is the underlying assumption when we answered question 4 – how many products are there for each category?

Answer:

Since we are studying the product purchase data, we are assuming that there was at least one transaction for each product.

Question 5:
What is the average amount spent for each product category?

Answer:

Save your Tableau workbook.

Lab 8 Data Modelling

Learning Outcome

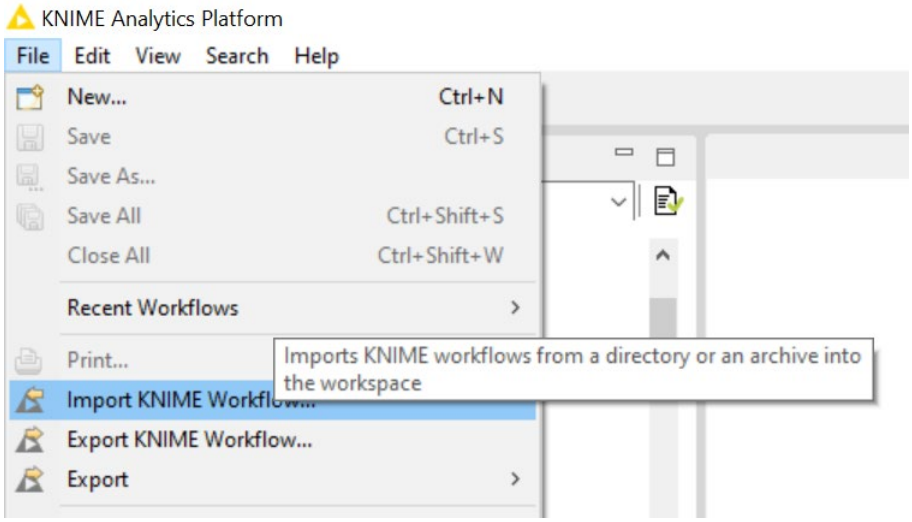
To use the Linear Regression Learner to generate the model

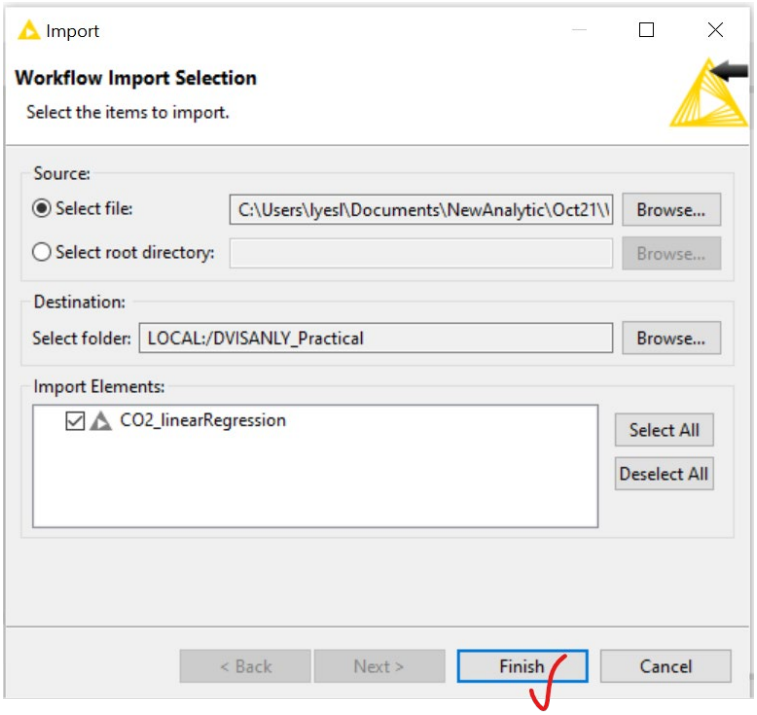
To use Regression Predictor to validate the model

Exercise 1

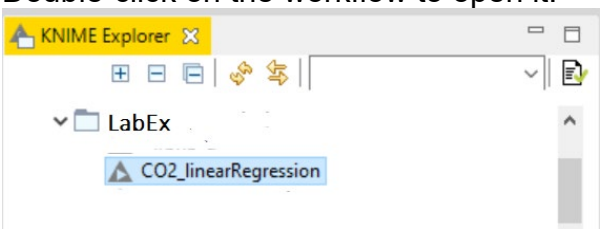
Data set: *CO2_datamodelling.csv*

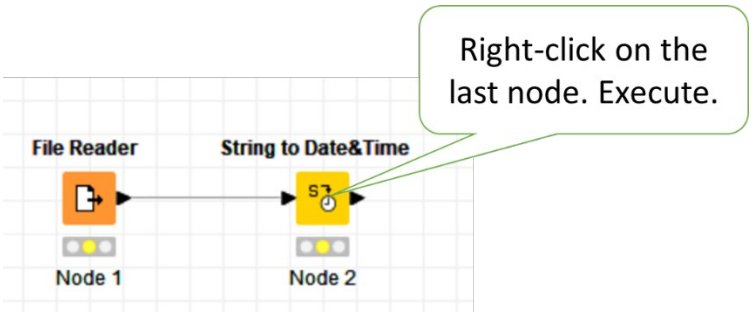
Import the KNIME workflow: *CO2_LinearRegression.knwf*





Double-click on the workflow to open it.





Open the output table:

File Edit Hilite Navigation View

Table "default" - Rows: 97 | Spec - Columns: 4 | Properties | Flow Variables

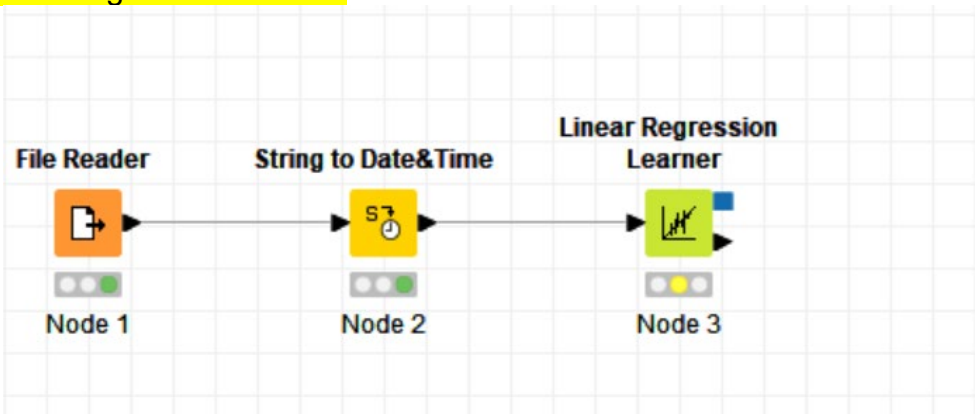
Row ID	Date and Time	CO2 (ppm)	Temperature (°C)	Relative Humidity (%)
1	2014-01-18T00:29...	458.45	-0.873	39.492
2	2014-01-18T01:29...	463.098	-1.056	39.626
3	2014-01-18T02:29...	465.712	-1.682	40.358
4	2014-01-18T03:29...	466.88	-2.459	41.797
5	2014-01-18T04:29...	468.992	-2.577	41.274

Key Questions:

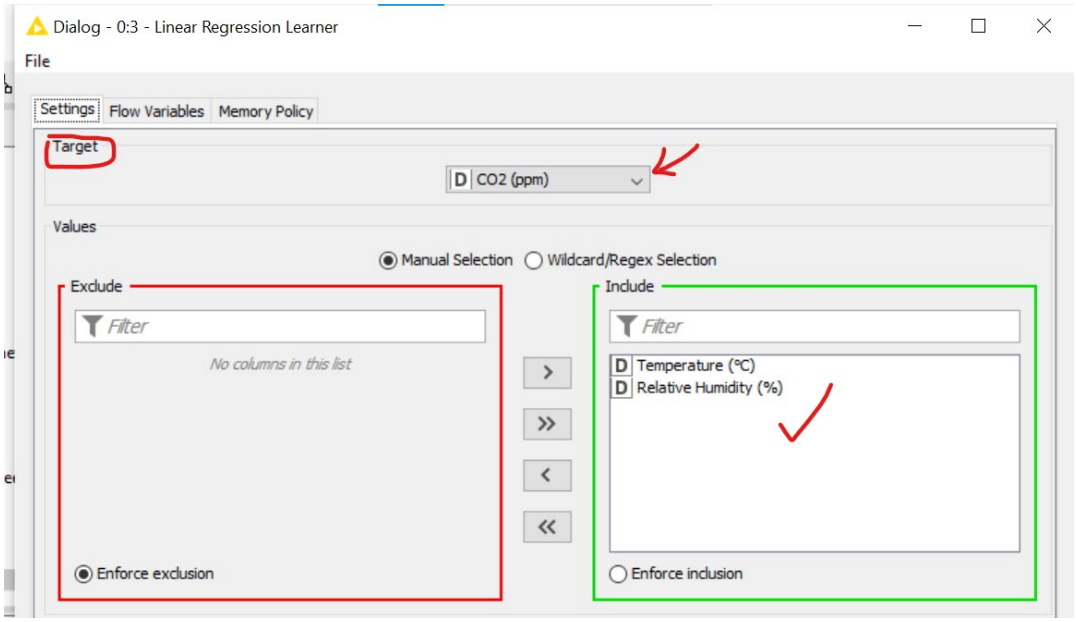
How accurate is the Linear Regression model in predicting CO2 level base on Temperature and Relative Humidity?

Target/Label: CO2
Independent variables: Temperature, Relative Humidity

Add a Linear Regression Learner.



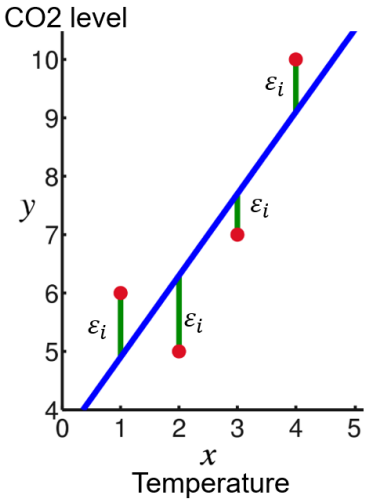
Configure the Target to be “CO2” and the independent variables are Temperature and Relative Humidity %.



How does the Learner work?

- Linear Regression Model with **multiple independent variables**:
  - $Y=\beta_1 X_1 +\beta_2 X_2 +\beta_3 X_3 +\ldots + \beta_0$

Learner will determine a trend line that passes through all/most of the observation points, and check the Root Mean Square Error.

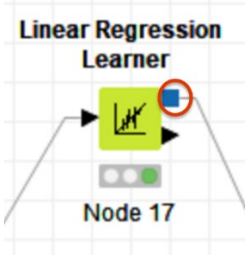


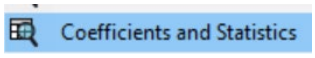
Observed value

Root mean squared error

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$$

It will converge to a line with the least RMSE. The coefficients of the trend line can be obtained from output port 1.





File Edit Hilite Navigation View

Table "Coefficients and Statistics" - Rows: 3 | Spec - Columns: 5 | Properties | Flow Variables

Row ID	S Variable	D Coeff.	D Std. Err.	D t-value	D P>|t|
Row1	Temperature (°C	-0.704	.112	-6.306	0
Row2	Relative Humidit.	0.979	.099	9.897	0
Row3	Intercept	409.778	.298	95.338	0

Y = -0.704*T + 0.979*RH + 409.778

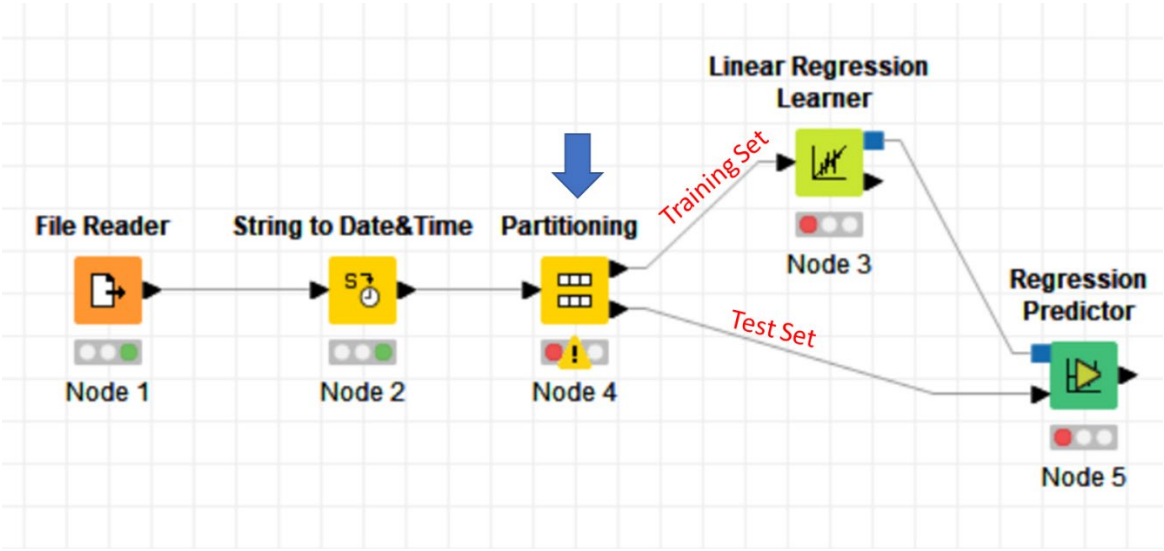
This is the linear regression model that predict the CO2 (Y-value) from temperature (T) and RH (relative humidity %) base on the given data.

How accurate is this mode?

We need to test the model by trying out with new data. But, we don't have any more data.

We can divide the data that we have, into Training set and Test set.

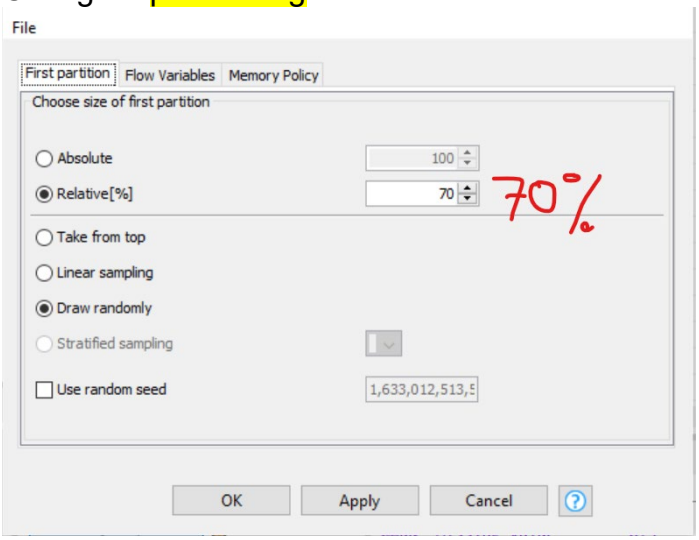
Use Partitioning node.



Divide the data into 70% and 30%.
70% will be the training set used by Linear Regression

30% will be the test set for Regression Predictor.

Configure partitioning node:



There is nothing to configure for predictor node.

Execute the Regression Predictor node.

File Edit Hilite Navigation View

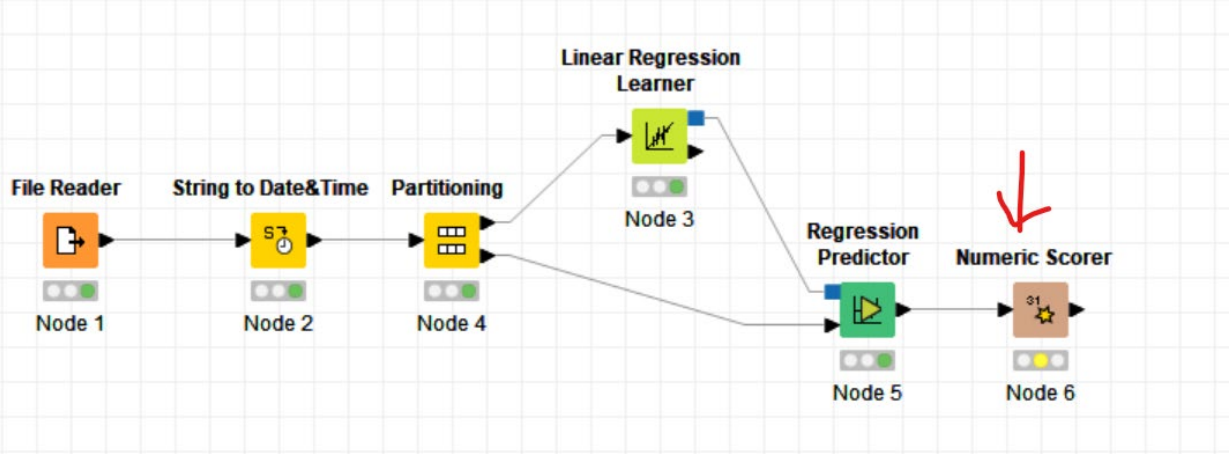
Table "default" - Rows: 30 | Spec - Columns: 5 | Properties | Flow Variables

Row ID	Date and Time	CO2 (ppm)	Temperatur...	Relative Hu...	Prediction (CO2 (pp...
6	2014-01-18T05:29...	463.085	-2.048	39.289	460.897
11	2014-01-18T10:29...	403.111	14.07	18.427	401.416
13	2014-01-18T12:29...	400.343	16.24	18.274	393.525
15	2014-01-18T14:29...	411.916	12.369	21.713	407.741

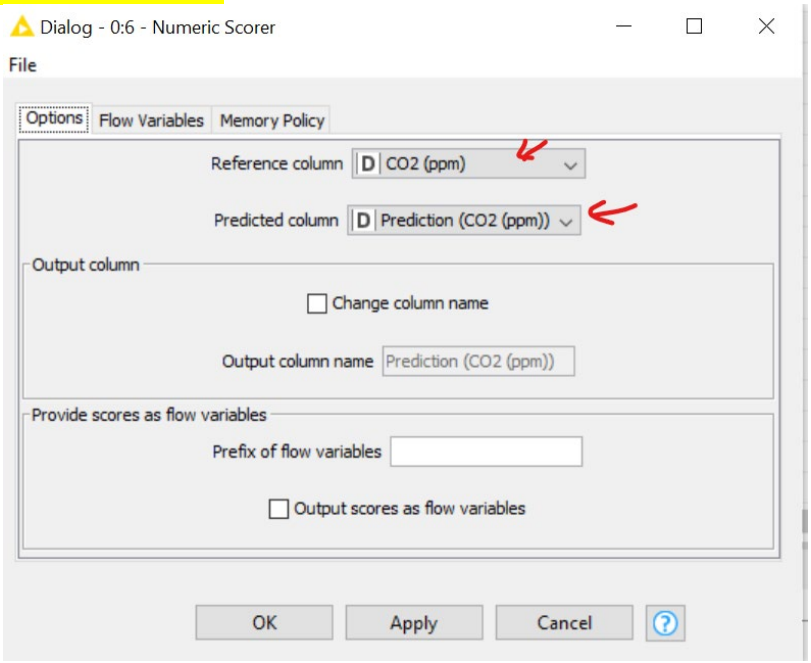
New column added. It is the predicted value using the regression model.

Next, we need to calculate the Root Mean Square Error (RMSE) to check the accuracy of linear regression model.

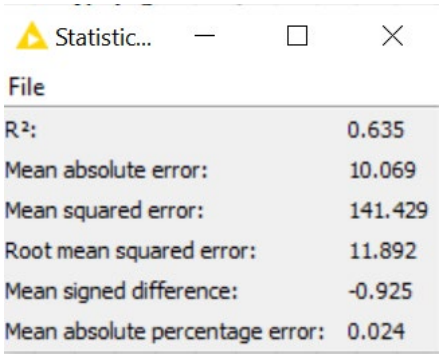
Add a Numeric Scorer node.



Configure the Numeric Scorer:



Execute the Numeric Scorer node.



The Root Mean Squared Error is 11.892.
The R-squared value shows 0.635. It means 63.5% of the data fit well to the regression line.

Conclusion:
Linear Regression model may be able to predict the CO2 level.

Exercise 2

Data set: *laptop_prices.csv*

For our task this week, we make use of a dataset containing a list of laptop prices for close to 1000 laptops with their specs such as CPU, GPU, RAM, screen size, weight, hard disk capacity and prices. For more details, you can visit this link (https://www.kaggle.com/ionaskel/laptop-prices). Take note that some modifications were made to the dataset.

Specific questions that we would like to answer in this session:

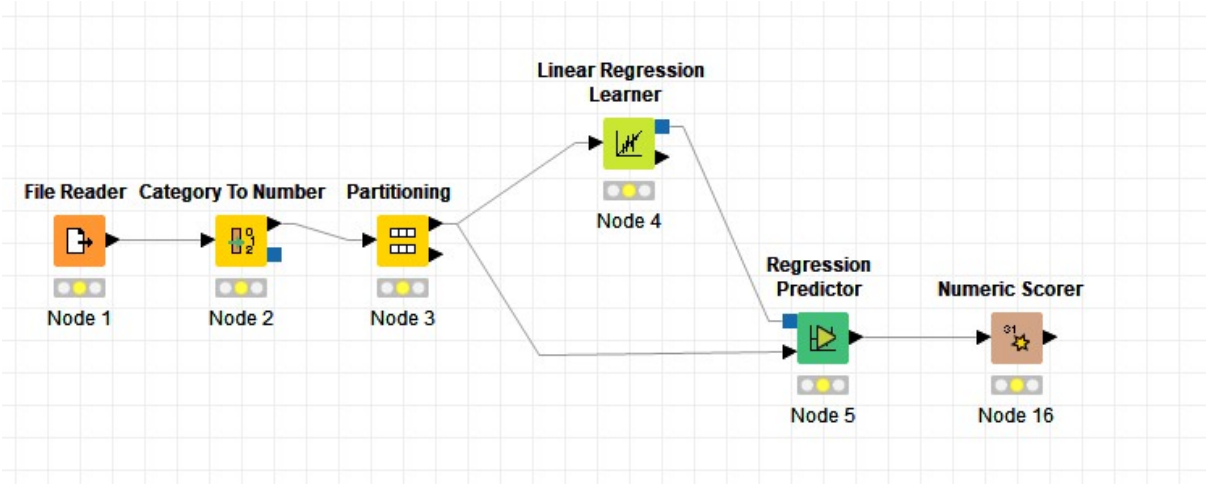
a) Which are the important features (key predictor) that determine a laptop's price?
b) How accurate is the model to predict the laptop's price?

1. Download the workflow *DataModeling_labex_start.knwf* from TP-LMS (or the data set folder)
2. Launch KNIME. Select **File > Import KNIME Workflow**.

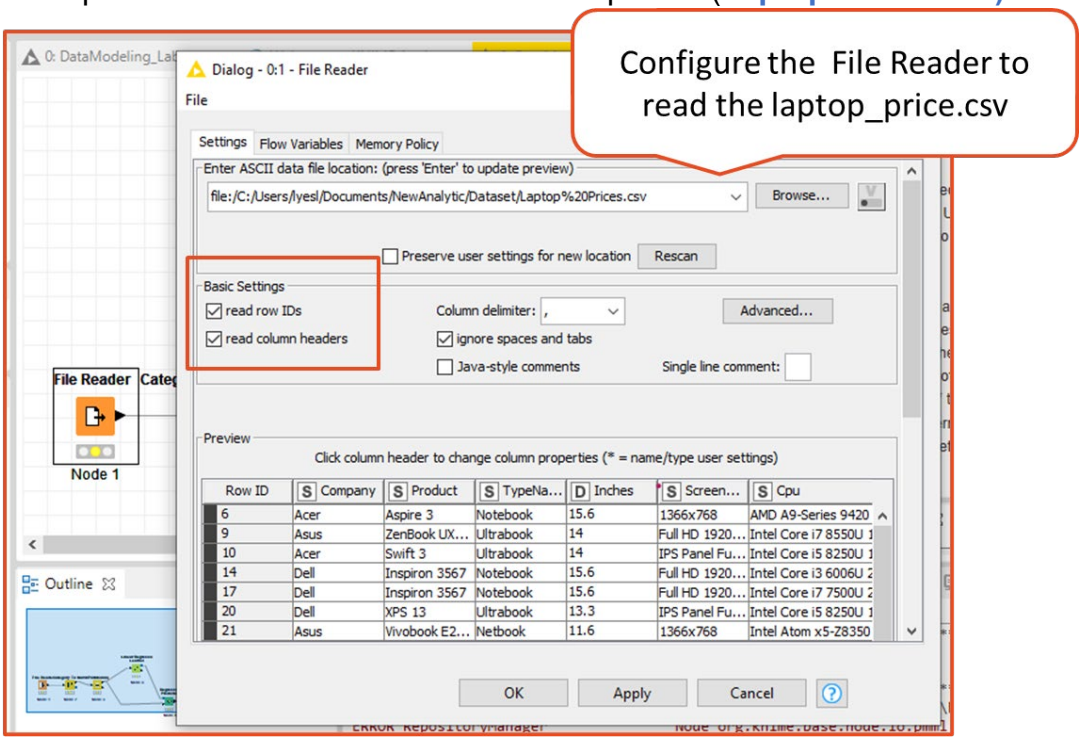
   Double-click on the workflow in the explorer pane to bring up the workflow in the main canvas.
   Your workflow should resemble the one shown below.

SCAN ME

https://videotp.ap.panopto.com/Panopto/Pages/Viewer.aspx?id=2e2688af-f0c4-4917-91a1-abd9003bf22f



If the File Reader node has a warning sign, you need to make sure that File Reader node points to the correct location of the Input file (**Laptop Prices.csv)**

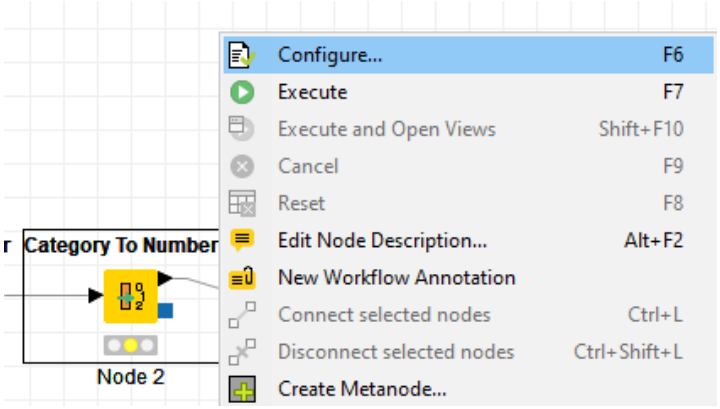


Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is the dependent (target) variable while one or more variable(s) is/are considered to be the explanatory (input) variable(s). For example, a modeller might want to relate the weights of individuals to their heights using a linear regression model.

Before attempting to fit a linear model to observed data, a modeller should first determine whether or not there is a relationship between the variables of interest. This does not necessarily imply that one variable causes the other (for example, higher SAT scores do not cause higher college grades), but that there is some significant association between the two variables.

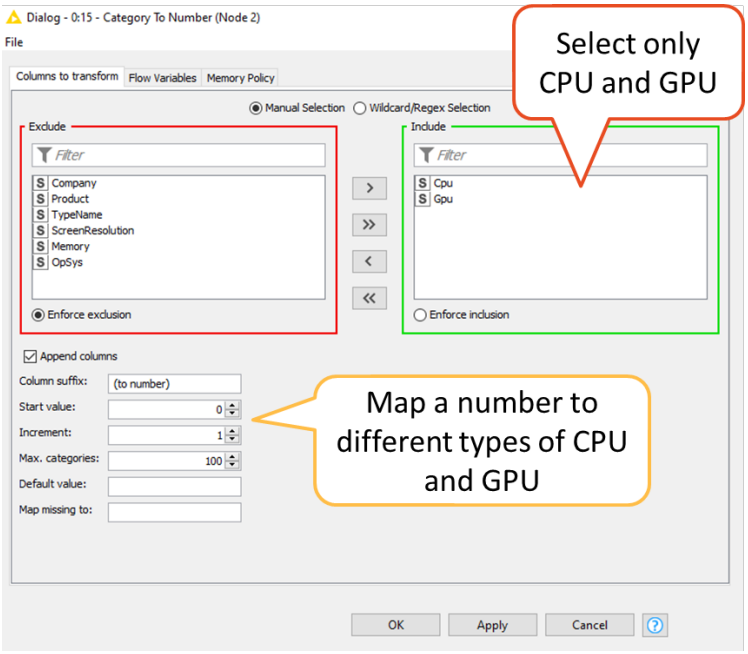
A linear regression line has an equation of the form **Y = a + bX**, where **X** is the explanatory variable and **Y** is the dependent variable. The slope of the line is **b**, and **a** is the intercept (the value of y when x = 0).

3. Double click on the Category to Number node to bring out its configuration window.

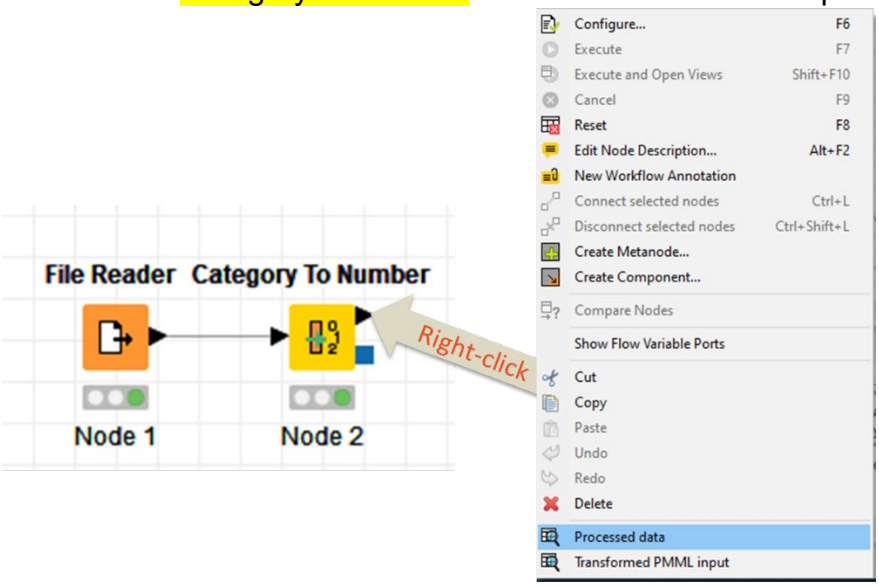


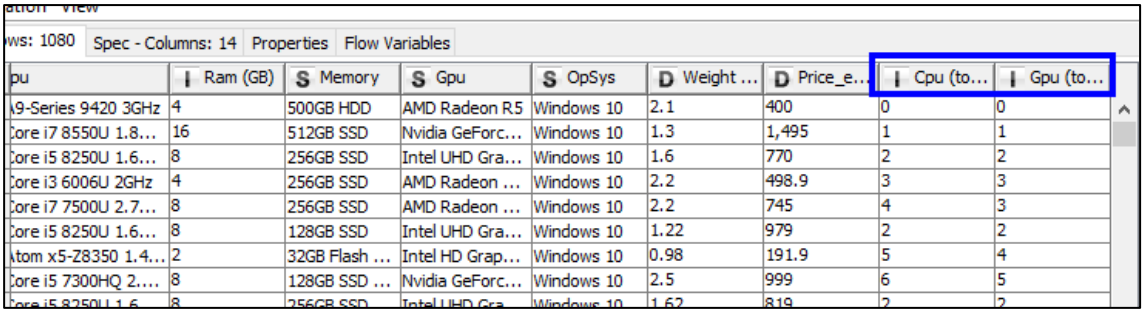
Since linear regression only works on numeric fields. For our exercise, we use the Category to Number node to covert two fields (*Cpu* and *Gpu* type) into numbers.

	S Cpu	I Ram (GB)	S Memory	S Gpu	S OpSys
	AMD A9-Series 9420 3GHz	4	500GB HDD	AMD Radeon R5	Windows 10
0...	Intel Core i7 8550U 1.8...	16	512GB SSD	Nvidia GeForc...	Windows 10
u...	Intel Core i5 8250U 1.6...	8	256GB SSD	Intel UHD Gra...	Windows 10
0...	Intel Core i3 6006U 2GHz	4	256GB SSD	AMD Radeon ...	Windows 10
0...	Intel Core i7 7500U 2.7...	8	256GB SSD	AMD Radeon ...	Windows 10
u...	Intel Core i5 8250U 1.6...	8	128GB SSD	Intel UHD Gra...	Windows 10
	Intel Atom x5-Z8350 1.4...	2	32GB Flash ...	Intel HD Grap...	Windows 10
u...	Intel Core i5 7300HQ 2....	8	128GB SSD ...	Nvidia GeForc...	Windows 10
...	Intel Core i5 8250U 1.6...	8	256GB SSD	Intel UHD Gra...	Windows 10
0...	Intel Core i7 8550U 1.8...	8	256GB SSD	Intel HD Grap...	Windows 10
	Intel Core i3 6006U 2GHz	4	1TB HDD	Intel HD Grap...	Windows 10
0...	Intel Core i5 8250U 1.6...	8	256GB SSD	AMD Radeon ...	Windows 10
0...	Intel Core i7 8650U 1.9...	8	256GB SSD ...	Intel UHD Gra...	Windows 10
0...	Intel Core i5 8250U 1.6...	8	1TB HDD	Nvidia GeForc...	Windows 10
0...	Intel Atom x5-Z8300 1.4...	4	64GB Flash ...	Intel HD Grap...	Windows 10
	AMD E-Series E2-6110 1...	2	32GB Flash ...	AMD Radeon R2	Windows 10
0...	AMD A6-Series 9220 2.5...	4	500GB HDD	AMD Radeon ...	Windows 10
...	Intel Core i7 8550U 1.8...	16	512GB SSD	Intel UHD Gra...	Windows 10
	Intel Celeron Dual Core ...	4	64GB Flash ...	Intel HD Grap...	Windows 10



Execute the Category to Number node and look at the output. (Processed Data)





ws: 1080 | Spec - Columns: 14 | Properties | Flow Variables

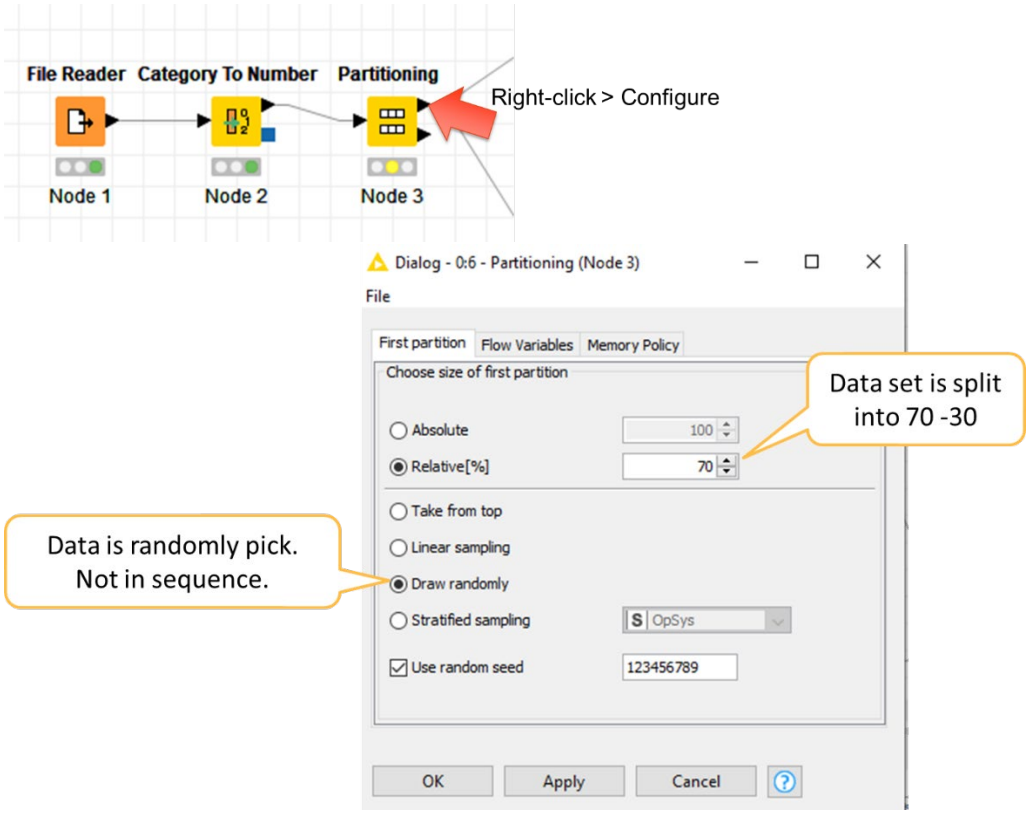
pu	Ram (GB)	Memory	Gpu	OpSys	Weight ...	Price_e...	Cpu (to...	Gpu (to...
A9-Series 9420 3GHz	4	500GB HDD	AMD Radeon R5	Windows 10	2.1	400	0	0
Core i7 8550U 1.8...	16	512GB SSD	Nvidia GeForc...	Windows 10	1.3	1,495	1	1
Core i5 8250U 1.6...	8	256GB SSD	Intel UHD Gra...	Windows 10	1.6	770	2	2
Core i3 6006U 2GHz	4	256GB SSD	AMD Radeon ...	Windows 10	2.2	498.9	3	3
Core i7 7500U 2.7...	8	256GB SSD	AMD Radeon ...	Windows 10	2.2	745	4	3
Core i5 8250U 1.6...	8	128GB SSD	Intel UHD Gra...	Windows 10	1.22	979	2	2
Atom x5-Z8350 1.4...	2	32GB Flash ...	Intel HD Grap...	Windows 10	0.98	191.9	5	4
Core i5 7300HQ 2....	8	128GB SSD ...	Nvidia GeForc...	Windows 10	2.5	999	6	5
Core i5 8250U 1.6	8	256GB SSD	Intel UHD Gra	Windows 10	1.62	819	2	2

Discussion 1:
CPU variable is now a numerical values ranging from 0 to ??? What is the highest number?

Does it mean CPU with value “6” is better than a CPU with value “1”? (Yes/No)

What assumptions are we making when we convert categorical data (like CPU) to numbers?

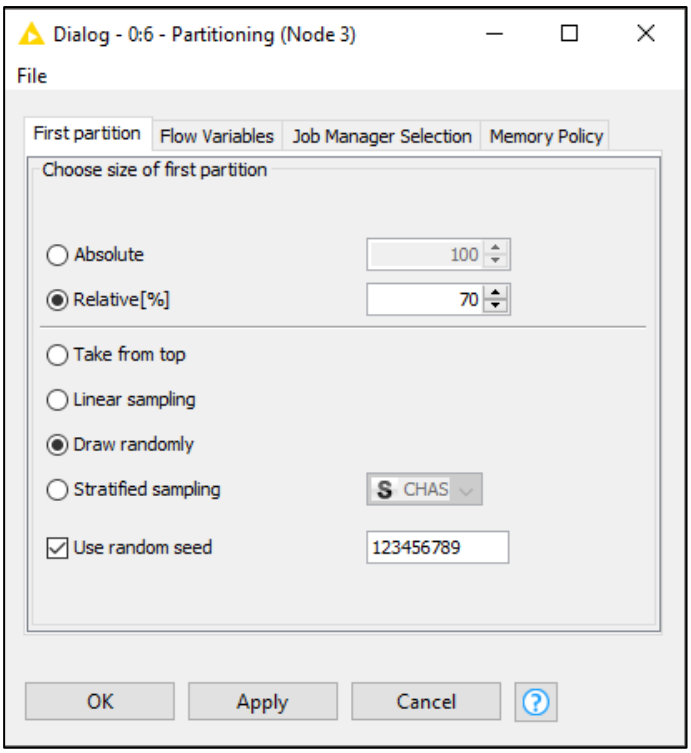
4. Double click on the Partitioning node or right-click on the node and select “configure”.



Look at the configuration of the Partitioning node. We partition our dataset into 70% - 30%. The plan is to use the 70% subset as the "training" dataset; this subset would be used to generate the regression model.

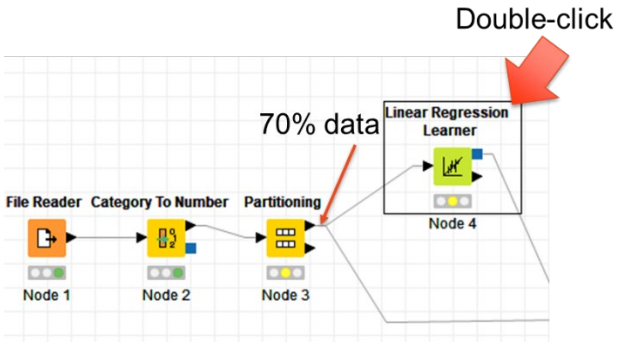
The 30% subset would be used as the "testing" dataset; this subset would act as a previously unseen dataset which would provide an unbiased assessment of the regression model.

The node is configured to use a random seed (in this case 123456789) such that the sampling can be replicated.



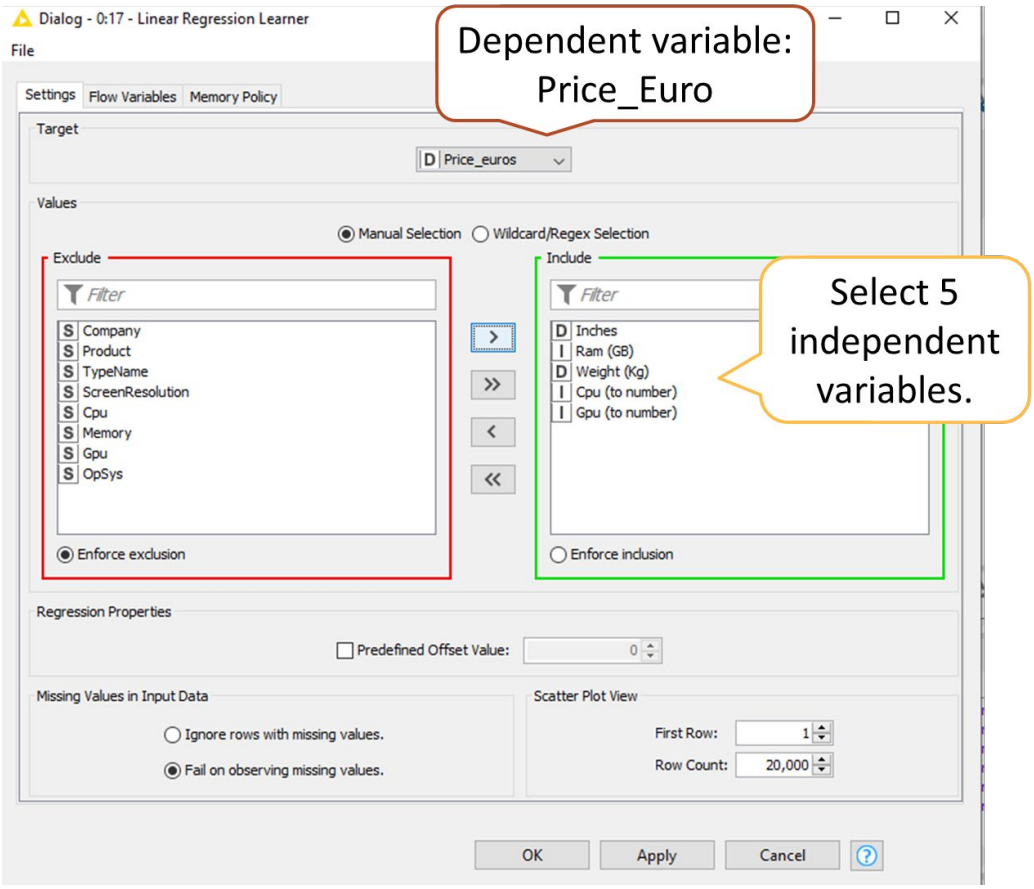
Click “OK” and Execute the node.

5. Proceed to configure the Linear Regression Learner node

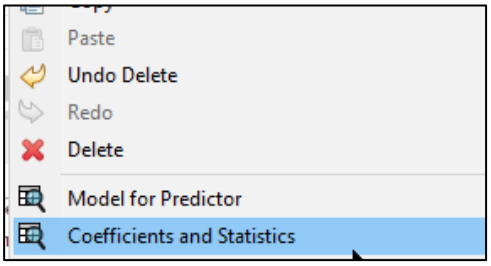


We are setting *Price_euros* as the target (dependent variables).

For the inputs (independent variables), we are using *Inches*, *Ram (GB)*, *Weight* (Kg), *Cpu (to number)* and *GPU (to number)*.



Go ahead and execute the Linear Regression Learner node.
Right-click on the node and select **Coefficients and Statistics**



Note: There might be a slight difference in the values you see and the one shown in the screenshot below.

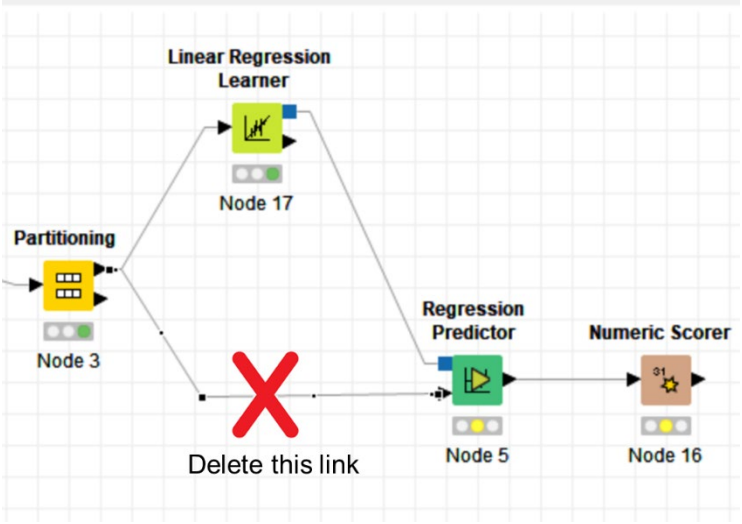
Table "Coefficients and Statistics" - Rows: 6 | Spec - Columns: 5 | Properties | Flow Variables

Row ID	Variable	Coeff.	Std. Err.	t-value	P>|t|
Row1	Inches	-76.192	20.536	-3.71	0
Row2	Ram (GB)	107.967	3.477	31.055	0
Row3	Weight (Kg)	26.191	46.217	0.567	0.571
Row4	Cpu (to num...	-0.224	0.88	-0.255	0.799
Row5	Gpu (to num...	0.97	1.007	0.963	0.336
Row6	Intercept	1,302.848	245.732	5.302	0

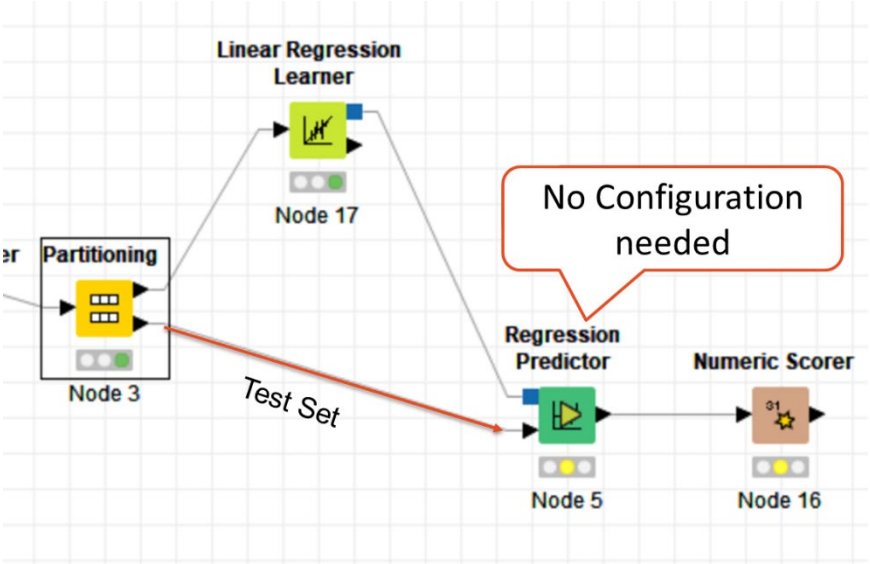
Discussion 2:
Which are the important features (key predictor) that determine a laptop's price?

How accurate is the Model?

Before we proceed to validate the model, delete the existing link:

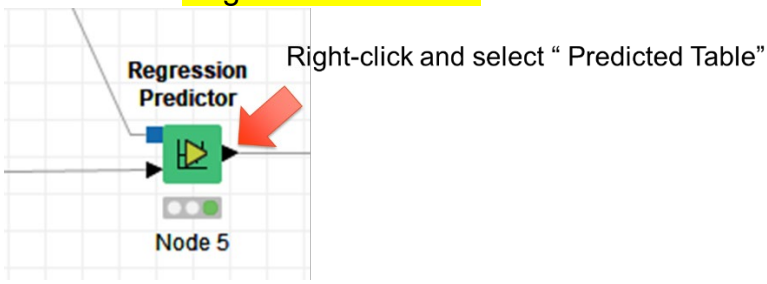


Re-link the 30% Test data to “Regression Predictor” Node.

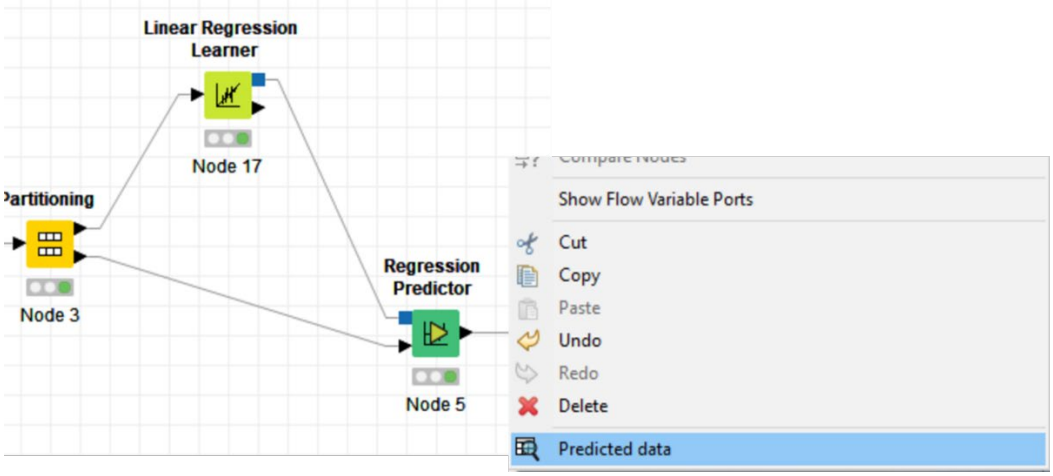


6. Next, we move on to the Regression Predictor node. This node allows you to make use of the regression model produced by the Linear Regression Learner node and carry out prediction based on input data.

   Execute the Regression Predictor node and look at the output.

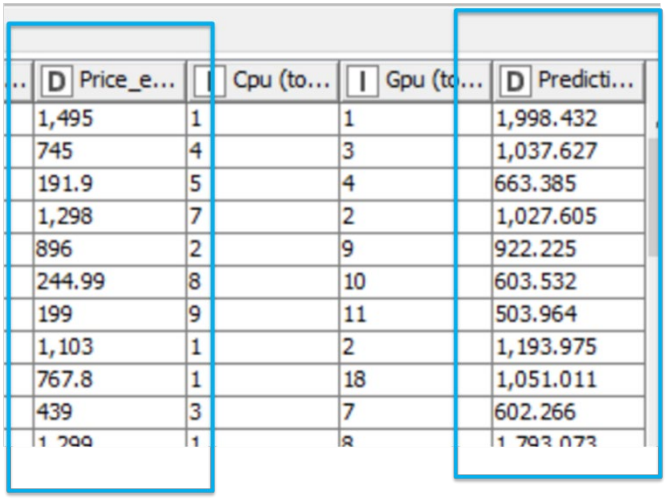


   (Right-click on the node, select Predicted data).



The two columns of interest to us are:

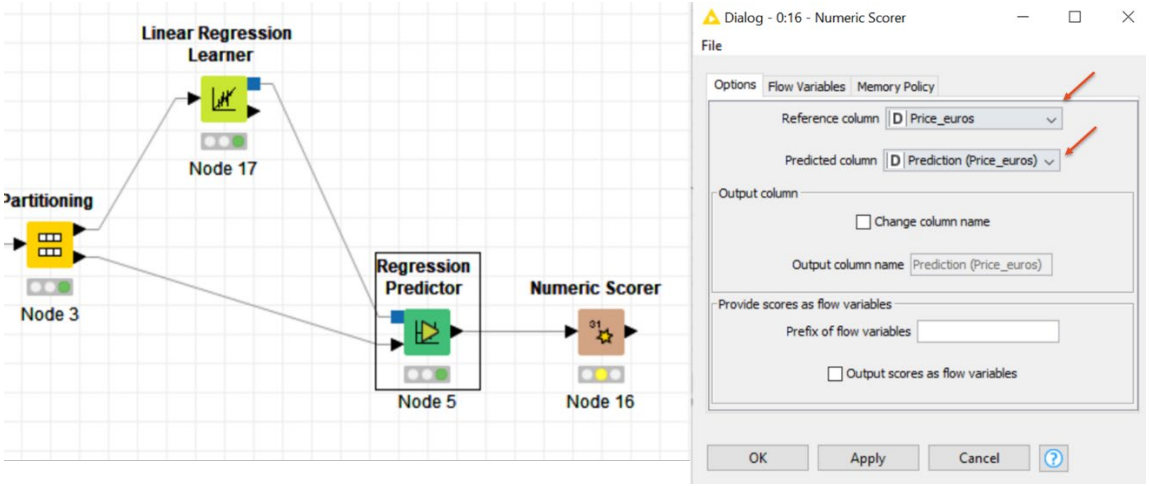
- *Price_euros* – this is the actual price of the laptop, it was in the original dataset)
- *Prediction (Price_euros)* – this is the predicted price based on your regression model

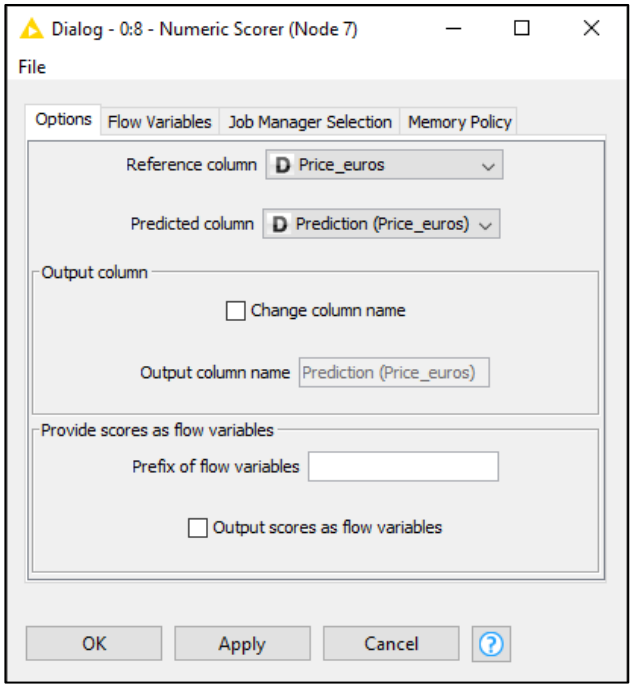


D Price_e...	I Cpu (to...	I Gpu (to...	D Predicti...
1,495	1	1	1,998.432
745	4	3	1,037.627
191.9	5	4	663.385
1,298	7	2	1,027.605
896	2	9	922.225
244.99	8	10	603.532
199	9	11	503.964
1,103	1	2	1,193.975
767.8	1	18	1,051.011
439	3	7	602.266
1,299	1	8	1,793.073

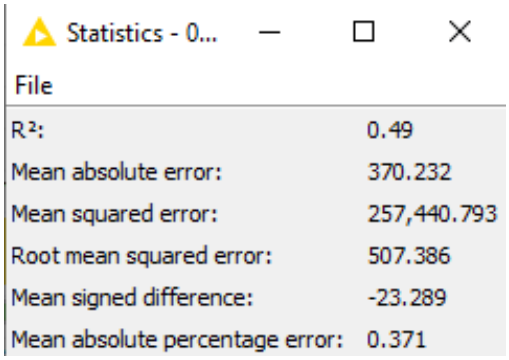
You can see that some predictions are quite close to the actual, but some there are way off target. However, rather than eyeball each pair of actual price vs predicted price, there is a better way for us to assess the quality of our regression model.

7. The Numeric Scorer node is the node used to provide a summary of the quality of the regression model. In the configuration, you select the Reference column (in our context, it is the actual laptop price *Price_euros*) and the Predicted column (in our context, it is the predicted laptop price *Prediction (Price_euros)*)





Click "OK" and "Execute and Open Views".



R^2 (R-squared) is a statistical measure of how close the data are to the fitted regression line. It is also known as the *coefficient of determination*. It is sometimes expressed as a percentage. R^2 value ranges from 0% to 100% (or 0.0 to 1.0 as in the case in KNIME)

$$R^2 = \frac{\text{Variance explained by the model}}{\text{Total variance}}$$

0% indicates that the model explains none of the variability of the response data around its mean.
100% indicates that the model explains all the variability of the response data around its mean.

Usually, the larger the R^2, the better is the model in explaining the predicted values. Along with R^2 are four measures of error. Naturally, errors should be as small as possible.

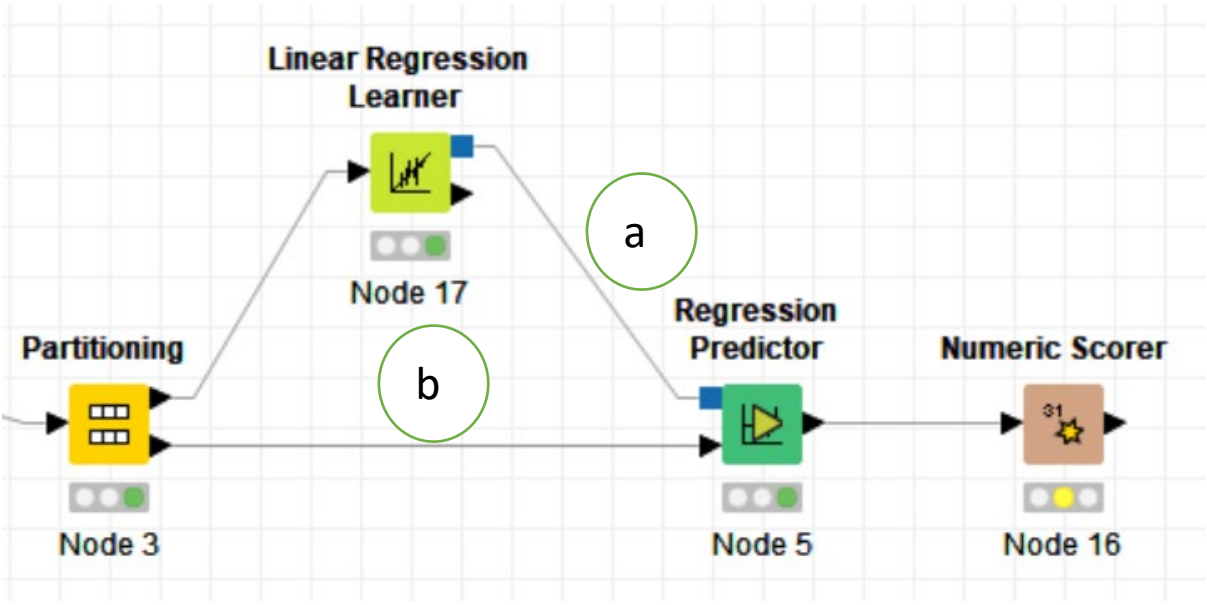
Discussion 3:

(a) A regression model that can produce the right prediction should have… (delete away "high" or "low" for each case)

High / Low R^2
High / Low mean absolute error
High / Low mean square error
High / Low root mean squared error
High / Low mean signed error

(b) What magnitude of error is considered small?

Discussion 4:



Discussion 4:

What are the data send into the “Regression Predictor”?

(a) Is _____

(b) Is _____

Save your workflow.

Exporting Predicted outcomes and Observed outcomes to TABLEAU to create an interactive dashboard (optional)

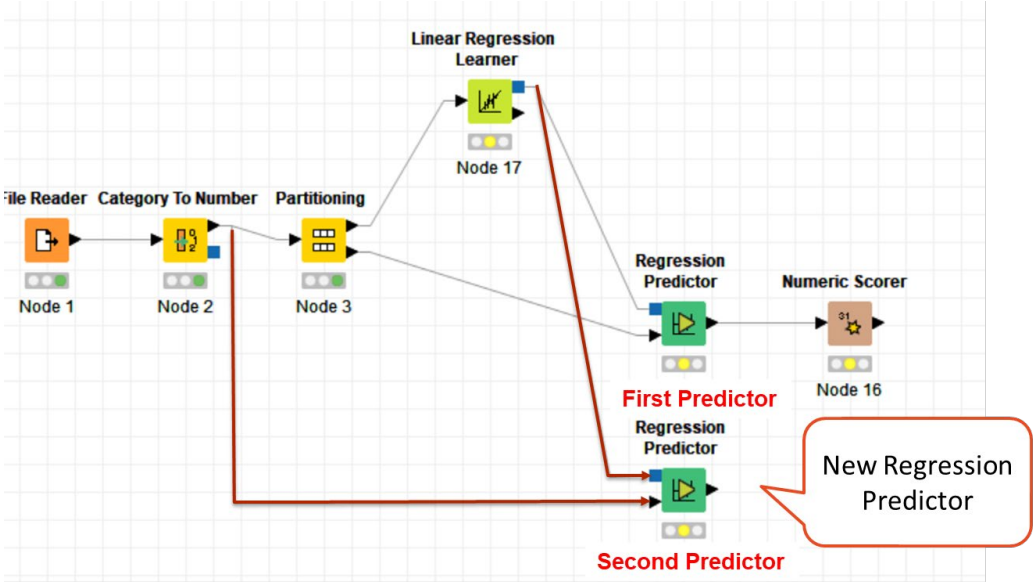
In this exercise, we will create an interactive dashboard to display the laptop price with the selection option for Cpu, Gpu, Ram Size and Screen Size.

Part 1 (Export the result to CSV)

1. Add a new Regression Predictor node (second one) to your workflow. Repeat the process of connecting the output model from the Linear Regression Learner node to this newly added Regression Predictor node.

   Feed the pre-partitioned data from the Category To Number node into this new Regression Predictor node. Execute the Regression Predictor and examine the output, verify that you have 1080 rows.

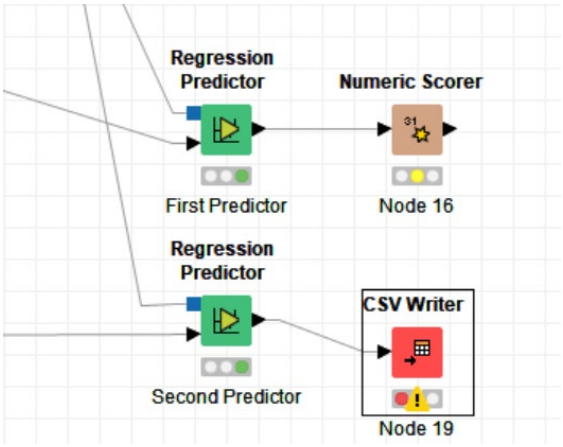
Your workflow should resemble the one shown in the next page.



	Discussion 1 What is the difference between the output of the Second Predictor and the first predictor?
The second predictor takes all the data set to predict the outcome. It will have 1080 observations. *The first predictor used 30% of the total data, so the predicted column will have 0.3x1080 = 324 observations.*	

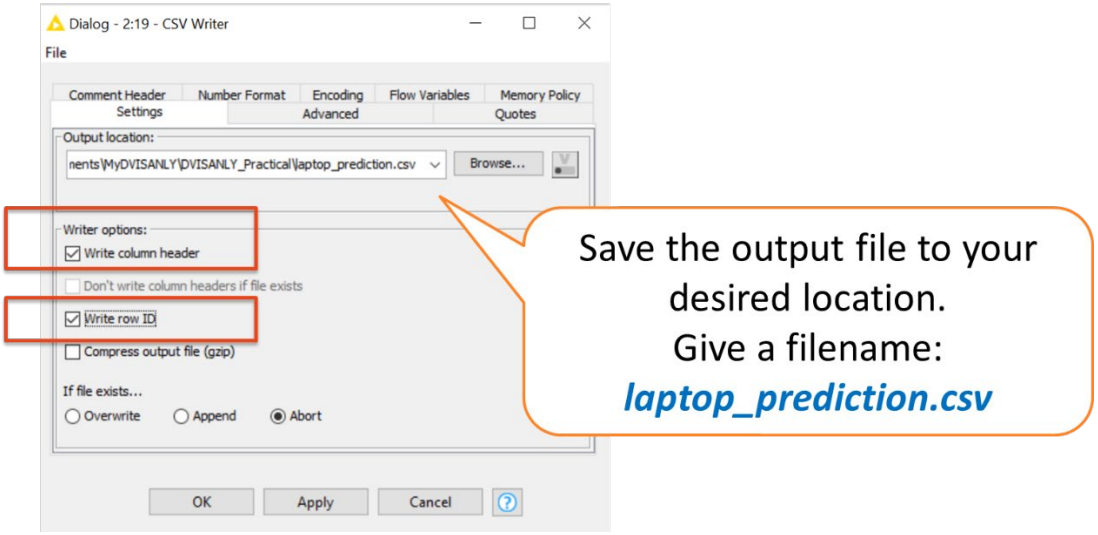
The purpose of the Second Predictor is to export all the data, with predicted result to a new file.

2. Let's export the predicted result from the Second Regression Predictor, which has all the data (1080 observations) to a CSV file. (by adding a CSV Writer)



Configure the CSV writer to include the Column Header and Row Header.
Set the output folder to your desired location.

We want to read this file into Tableau and create a dashboard with interactive selector to show the predicted laptop price.

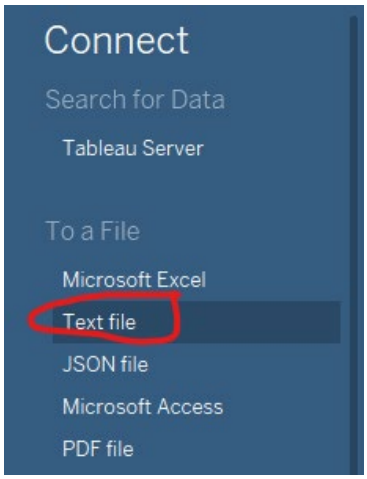


Execute the CSV Writer node and verify that the file is created in the destination that you have configured.

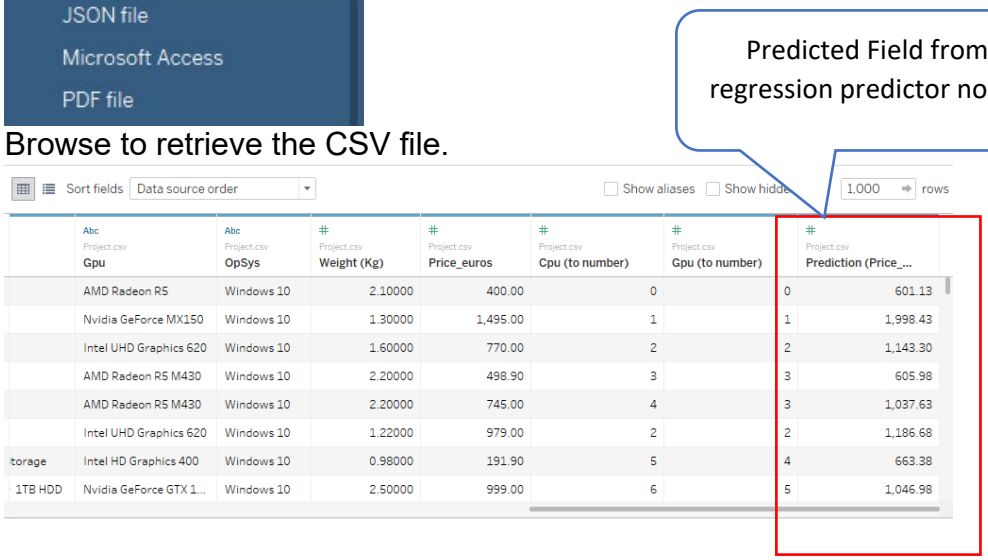
Part 2 (Visualising the data in TABLEAU)

We will now read the CSV file prepared in Question 1 into Tableau Desktop. This allows us to prepare visualisations that can help us answer our questions.

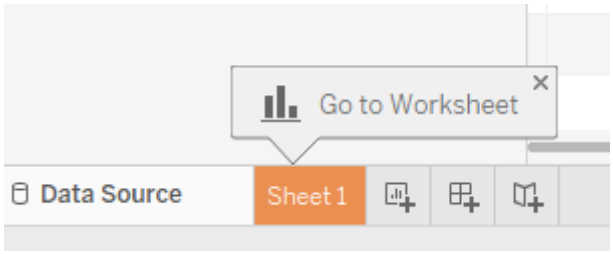
1. Launch the Tableau desktop.
   Click on Connect **Text** File.



Browse to retrieve the CSV file.



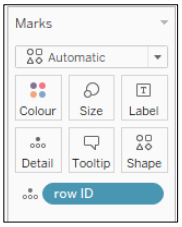
Continue to start a new Sheet.



2. We will create 2 separate scatter plots.

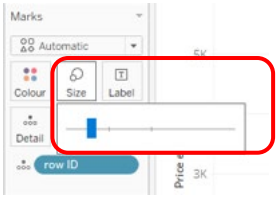
The first scatterplot – Predicted vs Actual Laptop Price.

- Column: *Price_euros*
- Row: *Prediction (Price_euros)*
- Detail: *Row ID*

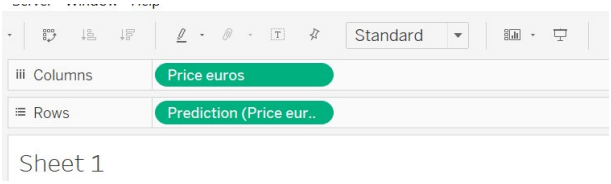


- Set to

- Reduce the bubble size for each data point.



SCAN ME https://videotp.ap.panopto.com/Panopto/Pages/Viewer.aspx?id=9493251e-7c30-4761-ae25-abdc01680dab

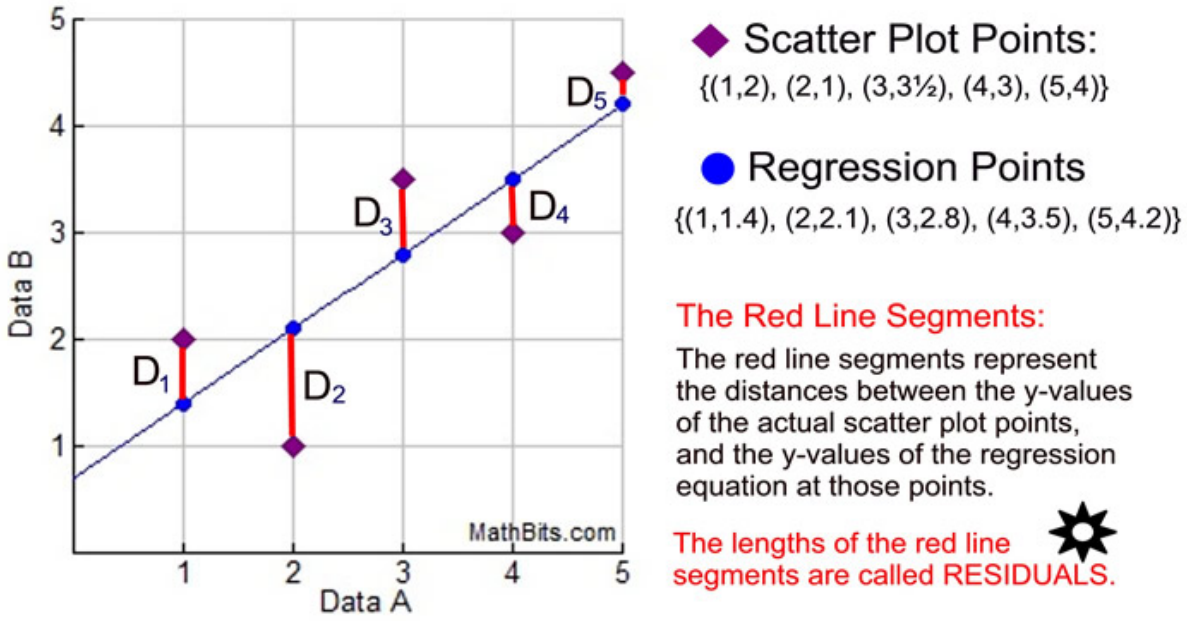


Discussion 2:
How should the Predicted vs Actual graph for a perfect regression model (100% accuracy model) look like?

The data point for all observed values = predicted value. All data points will fall on a straight line.

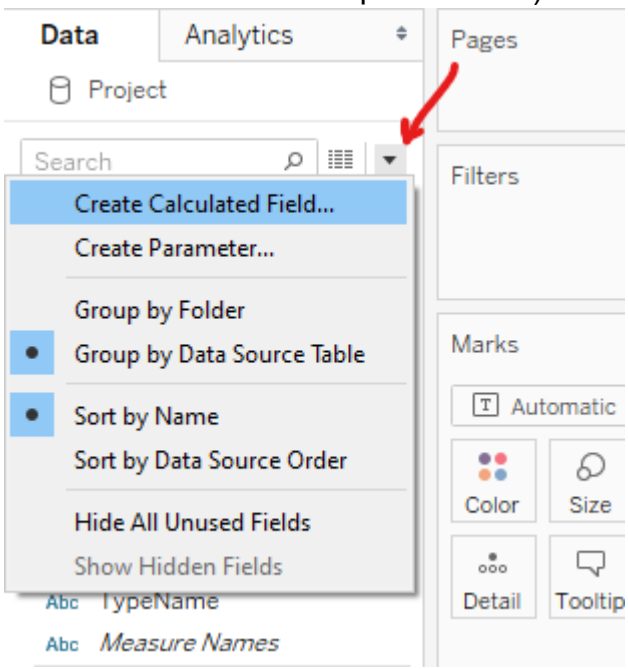
The second scatterplot – Residual Plot

In regression tasks, the difference between the observed value of the dependent variable (y) and the predicted value (ŷ) is called the **residual**. Having residual values that are close to 0 indicates accurate prediction by the regression model.



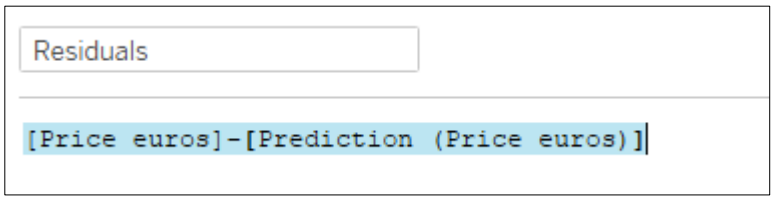
https://mathbitsnotebook.com/Algebra2/Statistics/residualgraph1aa.jpg

We start by creating a new field called *Residual*. (Recall how to create a new calculated field from the previous lab)



Enter the expression:

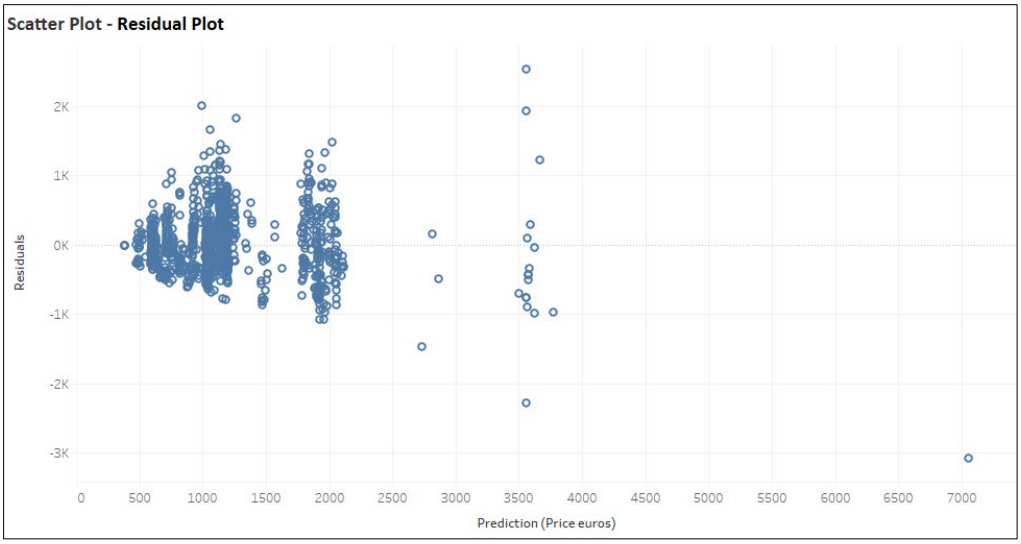
Residuals = [Price euros]-[Prediction (Price euros)]



Click [OK] to continue.

SCAN ME Show Me
https://videotp.ap.panopto.com/Panopto/Pages/Viewer.aspx?id=d8911071-d345-434d-b571-abdc0161f0a8

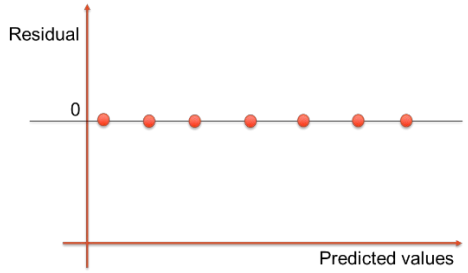
- Column: *Prediction (Price_euros)*
- Row: *Residuals*
- Details: *Row ID*



SCAN ME [Show me](https://videotp.ap.panopto.com/Panopto/Pages/Viewer.aspx?id=a4c9f022-a259-465a-9a25-abdc01674478)
https://videotp.ap.panopto.com/Panopto/Pages/Viewer.aspx?id=a4c9f022-a259-465a-9a25-abdc01674478

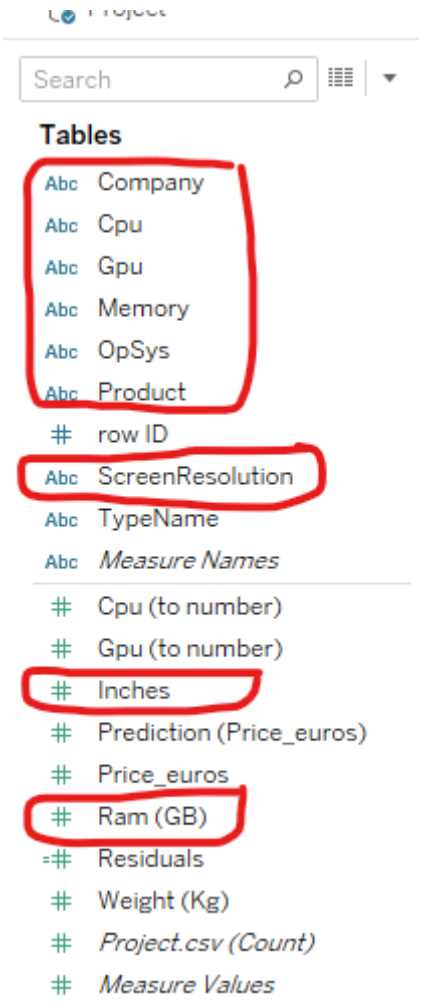
Discussion 3:
How should the residuals plot for a perfect regression model (100% accuracy model) look like?

The residual data points will fall on the "0" line.



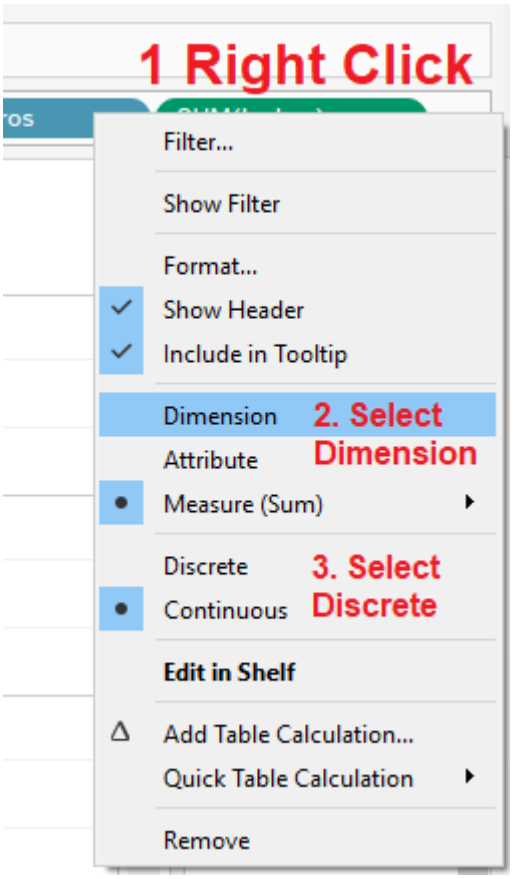
The Table should consist of the following data fields:

1) *Company*
2) *Product*
3) *OpSys*
4) *Cpu*
5) *Gpu*
6) *Ram (GB)*
7) *Memory*
8) *Inches*
9) *ScreenResolution*
10) *Price_euros*

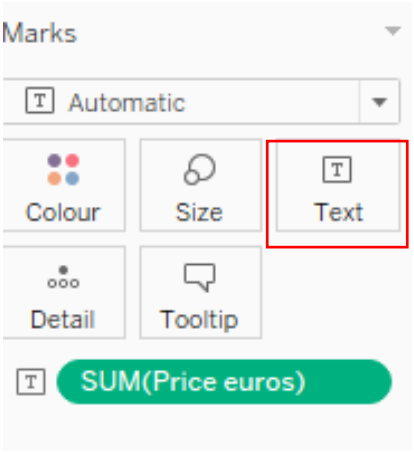


Add the above fields (1 to 9) to "Rows".

You need to convert to Dimension if the field is from Measure. The completed process appears as the blue colour pill.



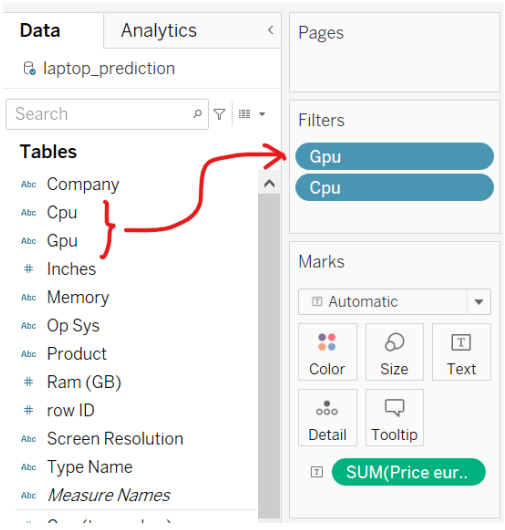
Add “Price_euros” to Text mark.



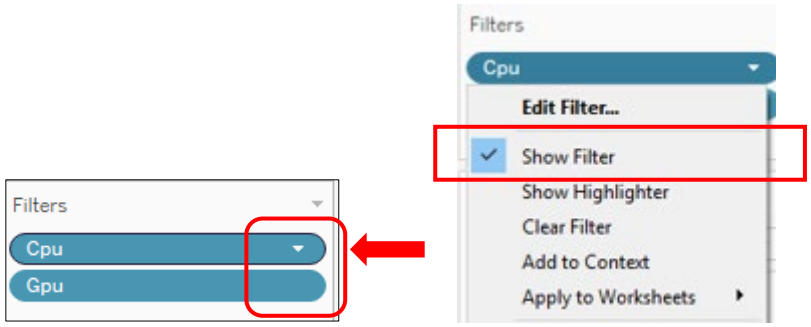
SCAN ME [Show me](https://videotp.ap.panopto.com/Panopto/Pages/Viewer.aspx?id=59556b55-876d-4e2f-838c-abdc0179b979)
https://videotp.ap.panopto.com/Panopto/Pages/Viewer.aspx?id=59556b55-876d-4e2f-838c-abdc0179b979

In the same worksheet sheet, add two Filters.

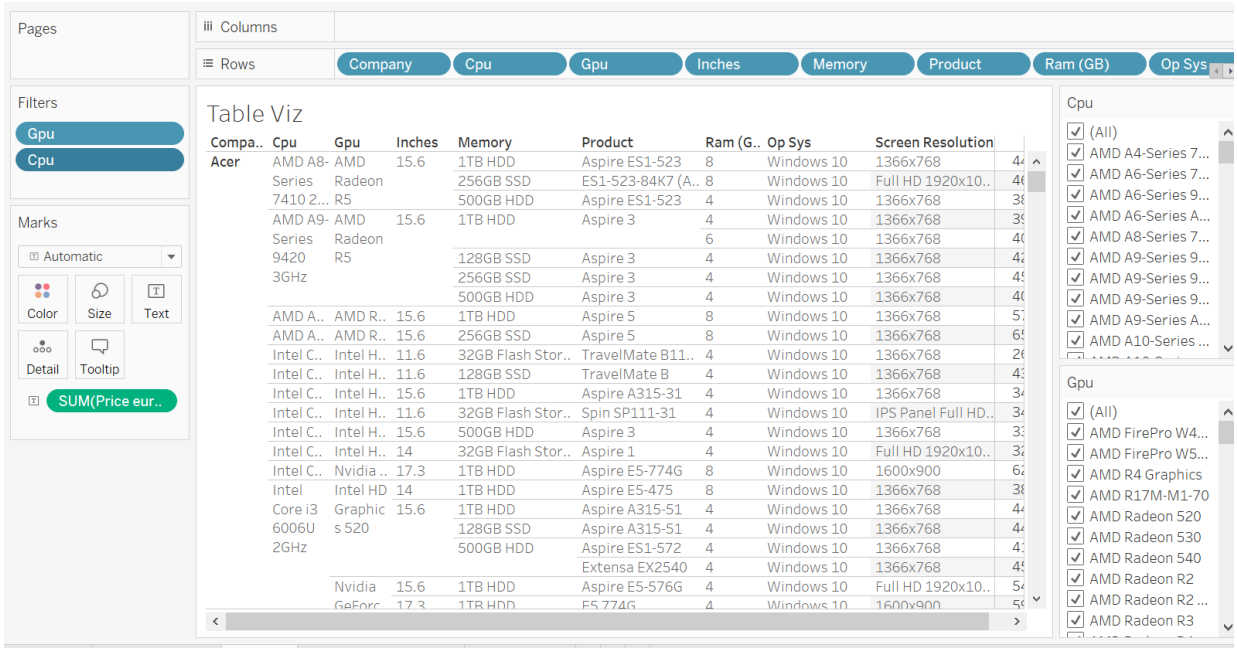
- Drag and drop *Cpu and Gpu* to Filter shelf as the fields for the filter



- *Click* from the drop-down arrow for each filter, check on "Show Filter" so that the CPU filter is shown on the right-hand side of the worksheet. This allows the user to select the CPU of their choice.



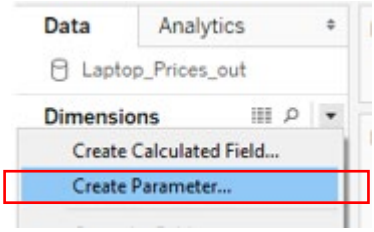
Finally, your report sheet should resemble the one shown below:



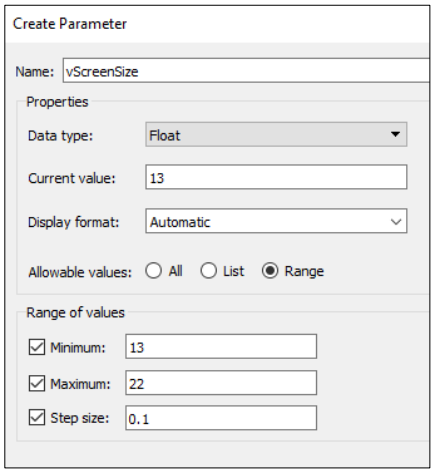
Rename the above worksheet as ***Table Viz***. This worksheet will be used for building up the dashboard.

Create parameters for interaction

Click on the *small arrow* next to **Dimensions**, select "*Create Parameter*".

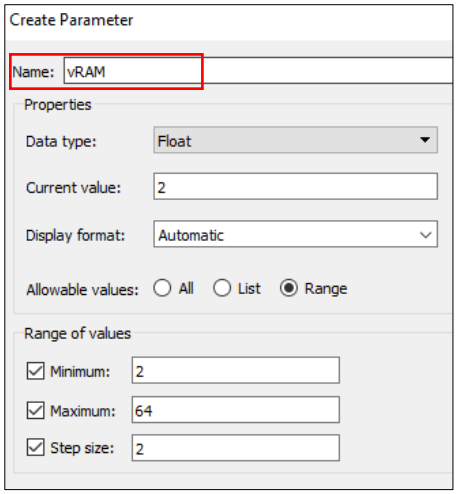


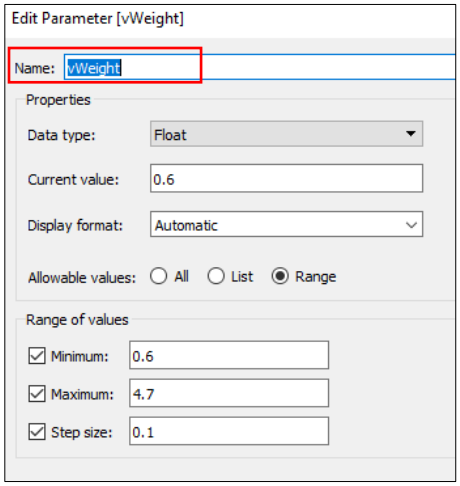
First, let's create a parameter named ***vScreenSize.*** See the configuration as follow.



Click **Ok**.

5. Repeat the steps in step 4 and create two more parameters with the following settings
   The parameters are named as ***vRAM*** and ***vWeight.***





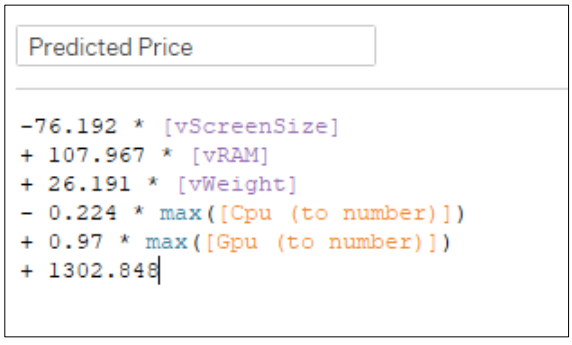
SCAN ME

https://videotp.ap.panopto.com/Panopto/Pages/Viewer.aspx?id=d52c6414-d2de-4bf0-8d33-abdd00418f10

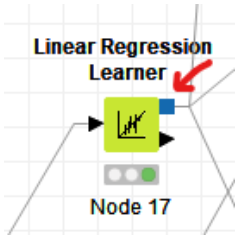
Still in the same worksheet tab, create a new calculated field (a new Measure).

We will call this measure *Predicted Price*

Predicted Price = -76.192 * vScreenSize
+ 107.967 * vRAM
+ 26.191 * vWeight
- 0.224 * max([Cpu (to number)])
+ 0.97 * max([Gpu (to number)])
+ 1302.848



The numbers (-76.192, +107,967 etc.) are based on the Linear Regression coefficients in KNIME



Linear Regression Result View - 0:17 - Linear Regression Learner

File

Statistics on Linear Regression

Variable	Coeff.	Std. Err.	t-value	P>|t|
Inches	-76.1917	20.536	-3.7102	0.0002
Ram (GB)	107.9671	3.4767	31.0548	0.0
Weight (Kg)	26.1907	46.2169	0.5667	0.5711
Cpu (to number)	-0.224	0.8796	-0.2546	0.7991
Gpu (to number)	0.9697	1.0066	0.9634	0.3357
Intercept	1,302.8484	245.7316	5.3019	1.51E-7

Multiple R-Squared: 0.6144
Adjusted R-Squared: 0.6119

(Linear Regression Statistics)

Still, in the same worksheet tab, drag the new field to the worksheet. Update the Chart Title to "Predicted Laptop Price". Your worksheet should look like below.

Predicted Laptop Price

$609.50

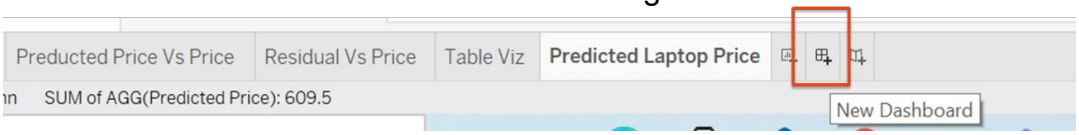
Rename the above worksheet as Predicted Price.

Build a Dashboard in Tableau

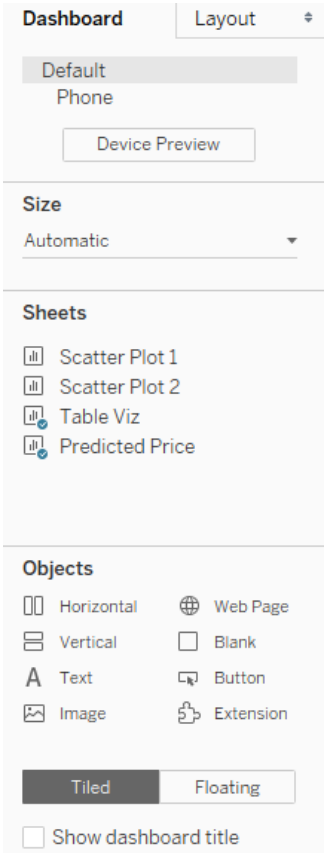
https://videotp.ap.panopto.com/Panopto/Pages/Viewer.aspx?id=8a3bba29-2f95-4ca3-b53f-abdd0062e2c4

Now, we want to create a Dashboard in Tableau.

6. Click on the **dashboard** icon at the most right-hand corner of the worksheet listing.

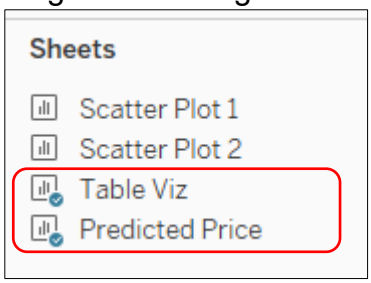


7. Below is the *Tableau Dashboard screen*. Device preview allows you to preview the dashboard on various devices. Note that device option is only applicable for Tableau Server with deployment option.

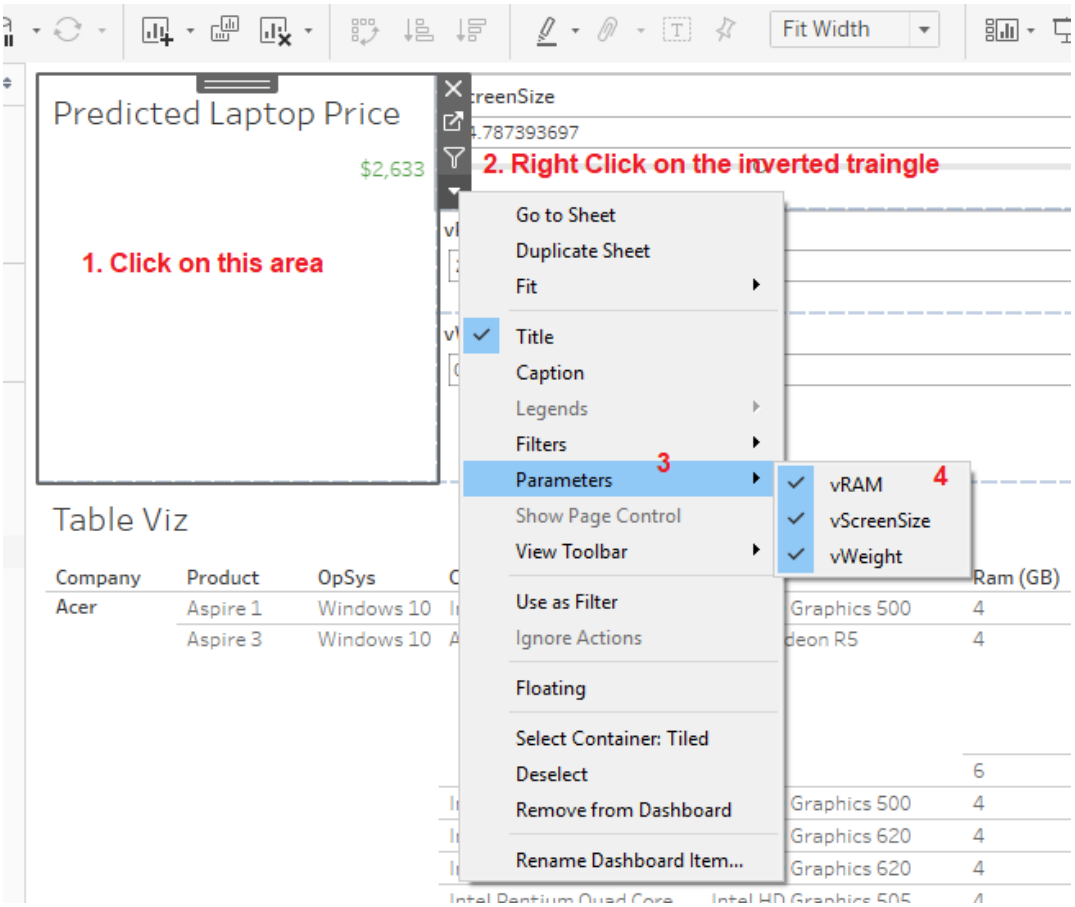


8. Next, set the size of the Dashboard. Note that you can set the size as Automatic, which fit automatically to your current screen. But when you open the file on another laptop/monitor, the Automatic does not work. It does not set automatically set to any size screen except the first monitor size that you use during the creation.

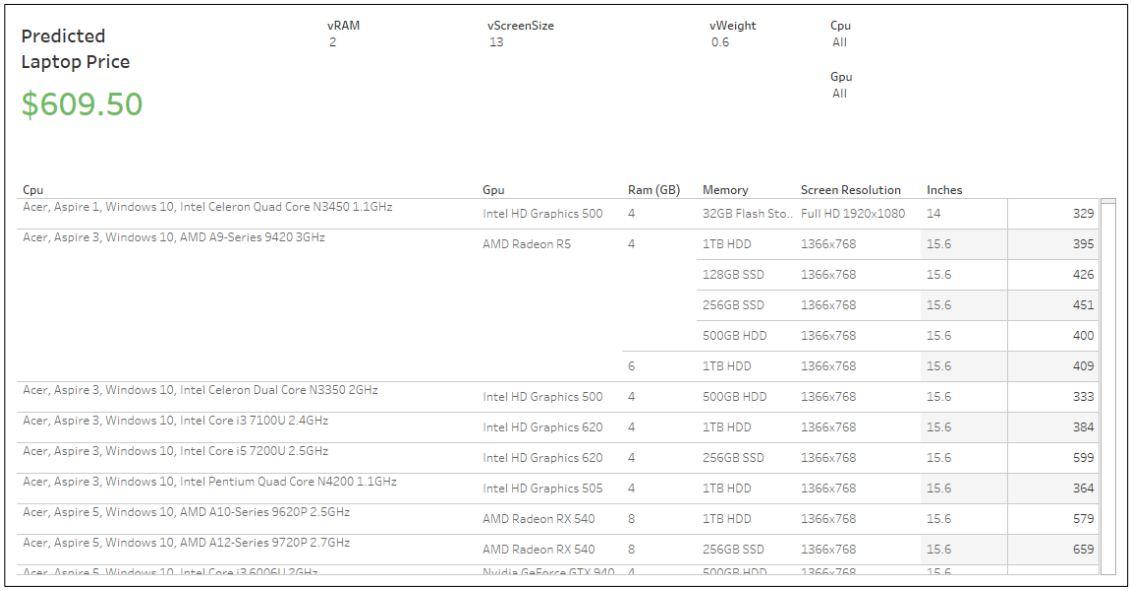
9. To create a dashboard, you can drag the chart to the workspace on the right. Try to drag the following charts to the dashboard. (1) Table Viz (2) Predicted Price.



Re-arrange your tableau objects with the Parameter Slider, Filter.
To include the parameter with the slider which you have created previously, perform the following.



10. Your Dashboard should resemble the one as shown below.



Use the Sliders to select the screen size, weight and RAM, the visualisation make use of the regression equation you have derived in Question 1 to compute the ***Predicted laptop price*** matching the specs you have entered via the Filters and Sliders.