Name/ ID:

1. **ABIRAMI BASKARAN**	**P2241801**
2. **NAVEEN GOPALKRISHNAN**	**P2246020**
3. **TILERON LEVI JAN LACANG**	**P2241632**

Module class: **DCPE/FT/2A/03**

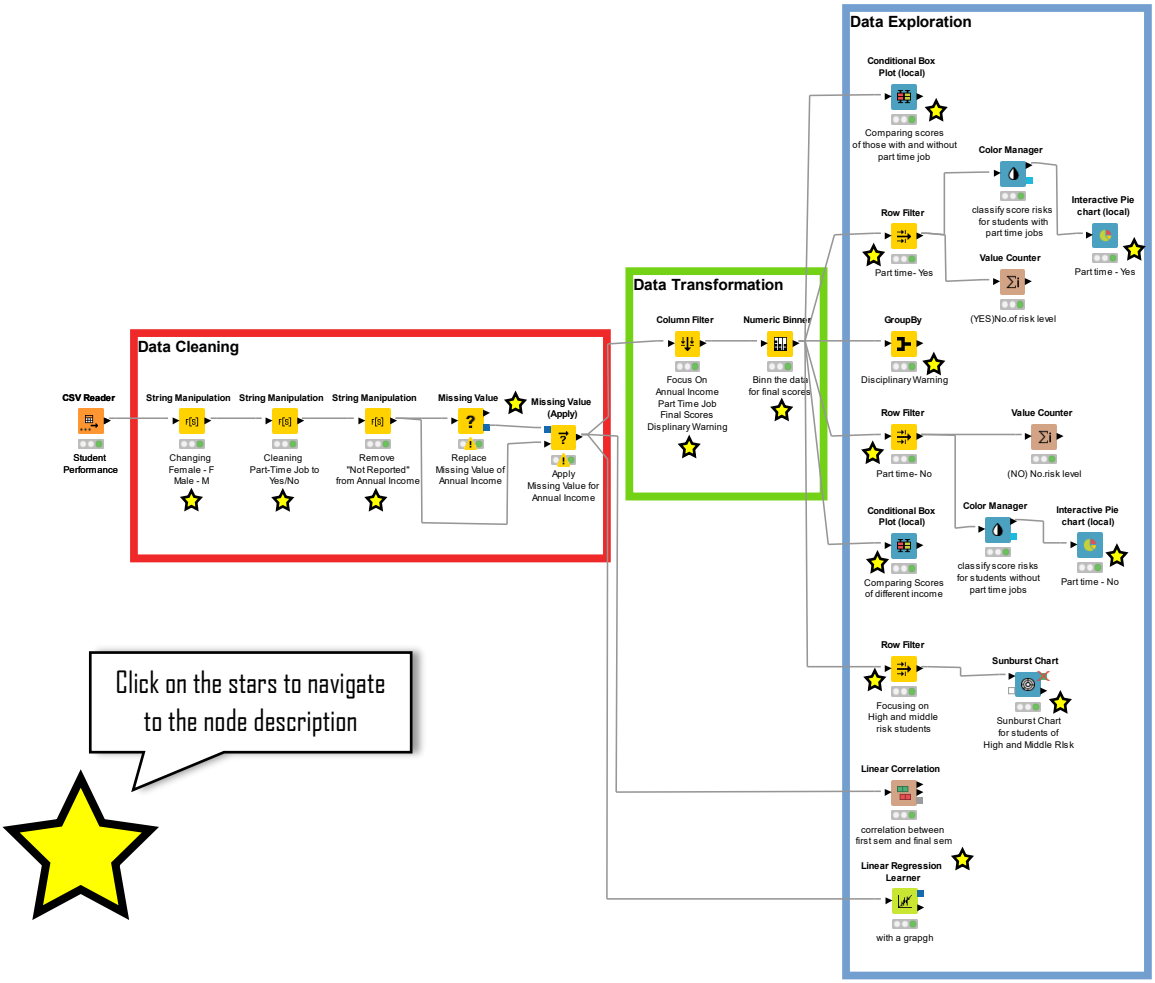
Tutor: **Mr Zhangsheng LAI**

Project choice (A or B): **B**

NOTE: Milestone 1 report **should not** exceed 5 pages (excluding figures, tables, and pictorials).

Business understanding

Every academic year, there will be a certain group of students who are more likely to either fail or underperform in academics. Every aspect of a young person's life can affect their ability to learn. It is important to identify these students earlier so that sufficient help can be provided for a progress in their education. Data mining is used to identify the possible reasons causing of students to be at risk in education. Through modelling and application of algorithms, a platform is created where we can understand the background of students at risk. This in turn helps to identify at-risk students earlier in future cohorts. By cleaning, transforming, and preparing data, we can produce satisfactory information for decision-making, improve data quality and be more consistent and accurate in data analysis.



Data understanding and preparation

Relevant data

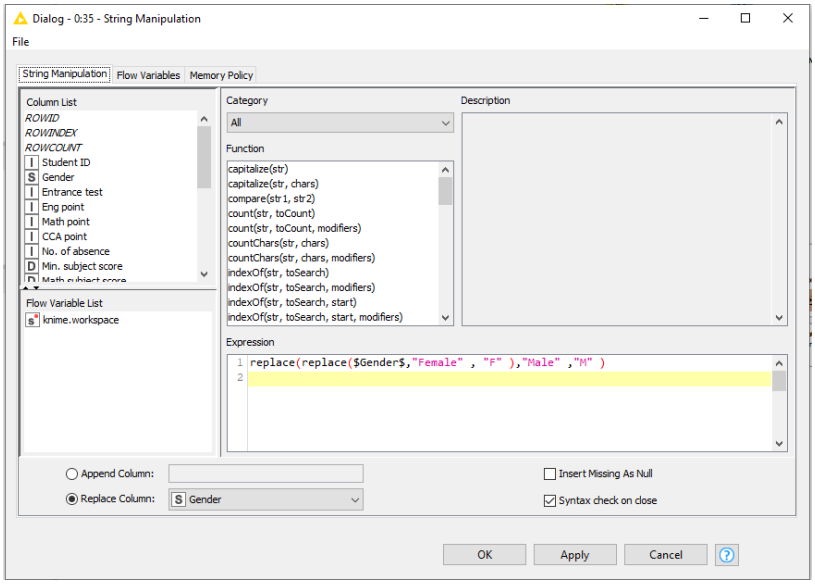
Attributes	***Reason***
Overall First Sem Score	This attribute can be used to split students by their scores. The overall scroll of a student shows his or her general performance in academics which helps to determine if a student is at risk.
Annual household income level	To study if the household income affects individuals' grades (Low, Middle, High)
Part Time Job	To study if having a part- time employment affects a student's performance in academics.
Disciplinary Warning	To determine if students with disciplinary warning fall under the risk category most likely than the rest.
Final Score	This attribute can be used to split students by their scores. The overall scroll of a student shows his or her general performance in academics which helps to determine student's risk level. (High Risk, Medium Risk, Low Risk)

Irrelevant data

Attribute	***Reason***
Entrance Test	Entrance test is too soon to determine a student's academic prowess
Gender	Gender of student (M – Male, F – Female). Gender of a student does not affect the final score
CCA Points	Student's accumulated CCA points might not be related to his or her academics and whether if the students at risk.
No. of absence	All students are generally present most days and hence it does not prove if students are at risk.
All individual Subject Scores (Math, Language & Comms., Programming, Science)	Student's skills in one individual subject does not prove his or her general academic prowess.
All individual Subject Attendance (Math, Language & Comms., Programming, Science)	Looking from the data, all students are generally present for most days. Hence, attendance for each subject does not prove if a student can be at risk.

Data Cleaning

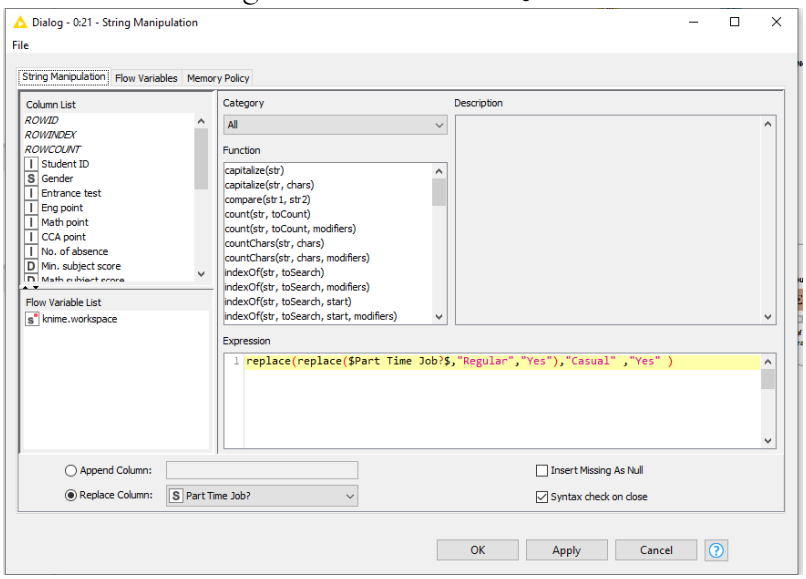
Standardising data for Gender



String Manipulation

There were irregular types of data from the original excel file. One of this irregularity error was found in the gender group. The data was inconsistent, such as “Female”, “F”, “Male” and “M”. By using the “string manipulation” node, we replaced all the “Female” to “F” and “Male” to “M”. This allows us to keep the data more uniformed and eases way of access.

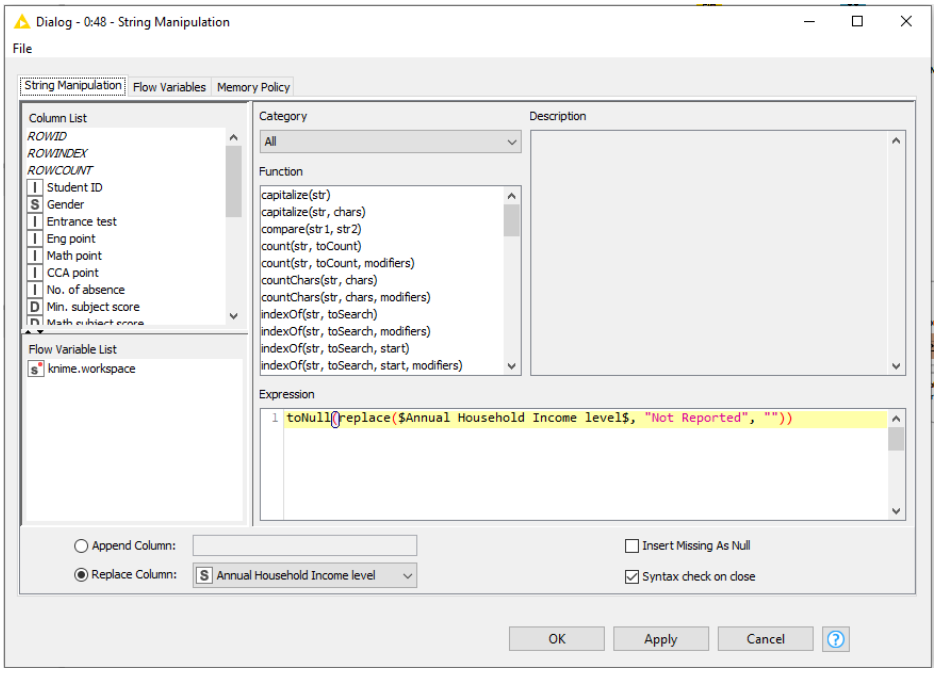
Standardizing data for Part Time Job



String Manipulation

In the original dataset, Part Time Job column contained “Regular”, “Casual” and “No”. Since “Regular” and “Casual” indicates to having a part time job, both those data were replaced with “Yes”. This would allow us to swiftly classify data and discover the common reason among the academically weak pupils for early intervention.

Replacing empty data

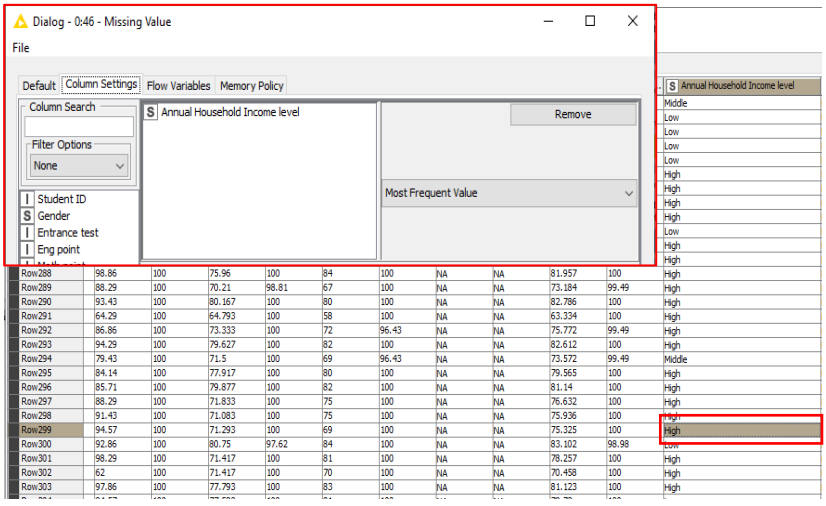


String Manipulation

In the original dataset, Row 299 of "Annual Household Income Level" was declared to be "Not Reported". Since there is only one data as such, it was first planned to be taken out, but we instead used the "String Manipulation" node to change it to a null value "?". Moreover, we will be replacing it. This will be shown in the next part.

Replacing missing data

Missing Value & Missing Value(apply)



Missing Value & Missing Value (Apply)

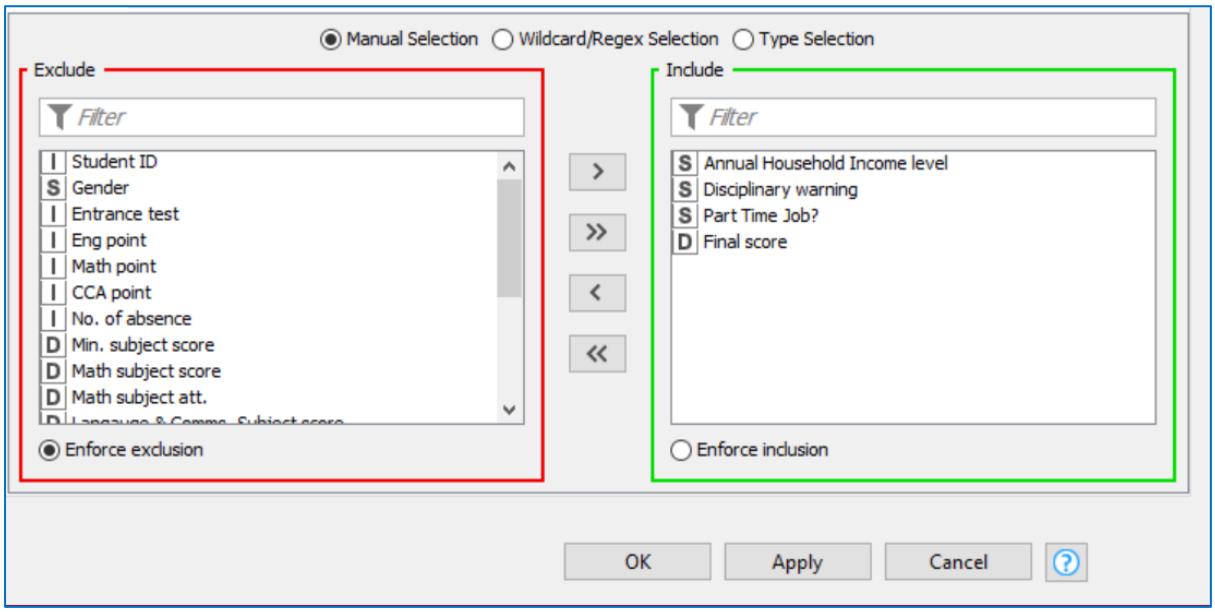
- Following that, after using "Missing Value" node on the empty string(?) it will then be replaced with the most frequent value(mode) of Annual Household Income which is "High".

- Once the data has been replaced, the new value is applied back to the file using the node "Missing Value (Apply)".

Data Transformation & Data Preparation

Focus on relevant data

Column Filter

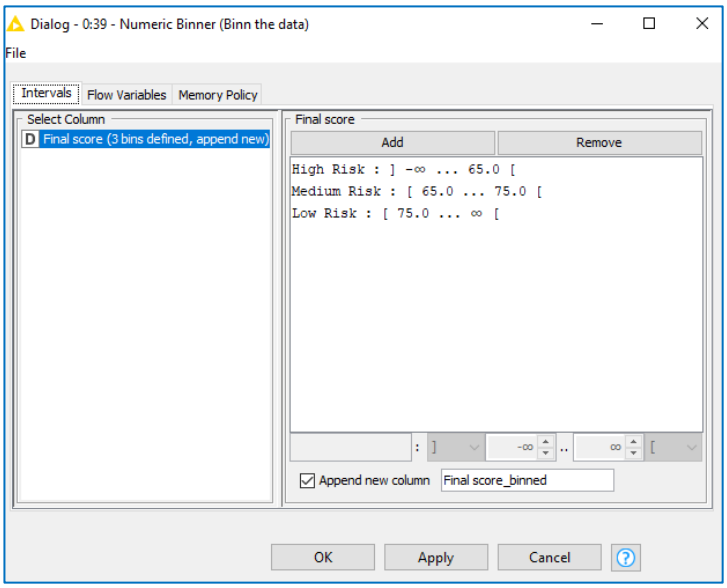


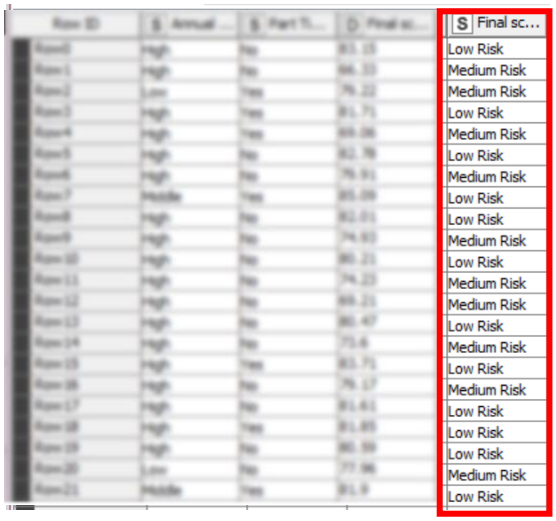
To focus more on the relevant data which provides details of students' academic prowess and their behavioural assets, we filtered the column which only focuses on data that are required. Using the "Column Filter" node, columns such as "Annual Household Income level", "Part Time Job?", "Final score" and "Disciplinary Warning" were filtered.

Row ID	S Annual ...	S Disciplin...	S Part Ti...	D Final sc...
Row0	High	N	No	83.15
Row1	High	N	No	66.33
Row2	Low	N	Yes	79.22
Row3	High	N	Yes	81.71
Row4	High	N	Yes	69.06
Row5	High	N	No	82.78
Row6	High	N	No	79.91
Row7	Middle	N	Yes	85.09
Row8	High	N	No	82.01
Row9	High	N	No	74.93
Row10	High	N	No	80.21
Row11	High	N	No	74.23
Row12	High	N	No	69.21
Row13	High	N	No	80.47
Row14	High	N	No	73.6
Row15	High	N	Yes	83.71
Row16	High	N	No	79.17
Row17	High	N	No	81.61
Row18	High	N	Yes	81.85
Row19	High	N	No	80.59
Row20	Low	N	No	77.96
Row21	Middle	N	Yes	81.9
Row22	Middle	N	No	68.68
Row23	Middle	N	Yes	72.02
Row24	High	N	Yes	73.73

Bin scores into different groups

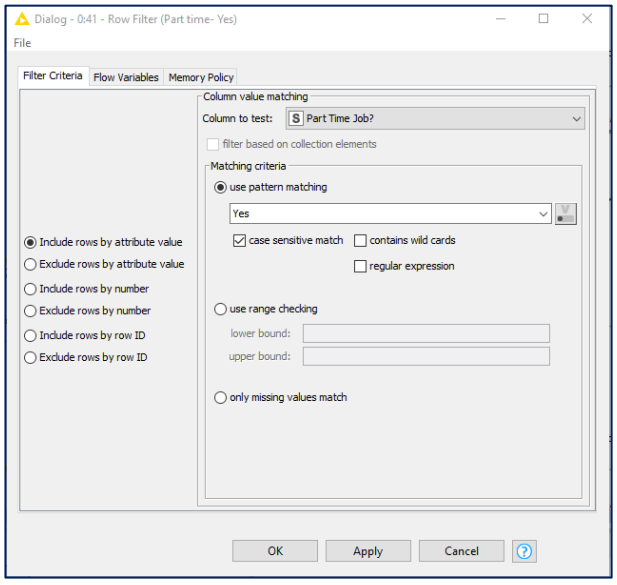
Numeric Binner





A "Numeric Binner" node is used to categorise students' scores. Scores that are lesser than 60, are categorized as 'High Risk'. Scores that are in between 60 to 75 is considered 'Medium Risk'. Lastly, scores that are above 75 are considered 'Low Risk'. A new column is appended to view this data and to not overwrite the original scores.

Students with Part Time Job

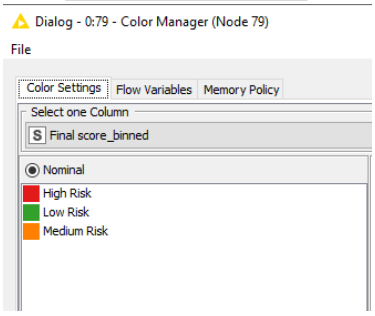


Row Filter

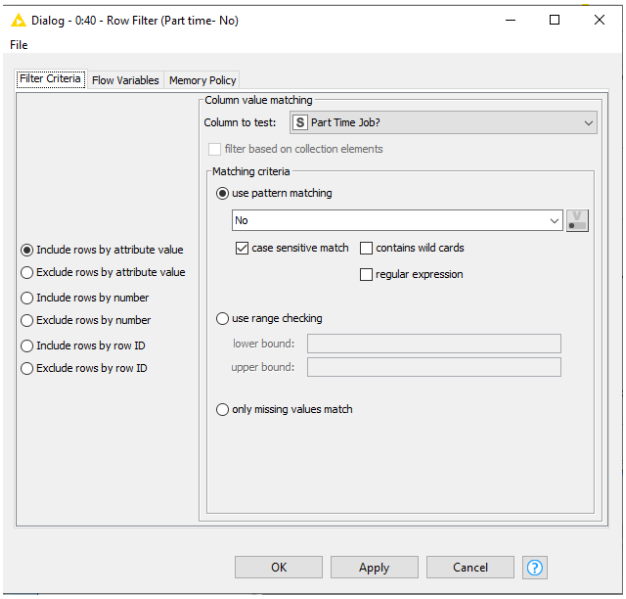
Focusing on the part time job, we kept the students who had part-time employment and filtered out those with using this Row Filter node. Doing this will help us with our upcoming parts.

Colour Manager shows the score risks based on different colours.

Colour Manager

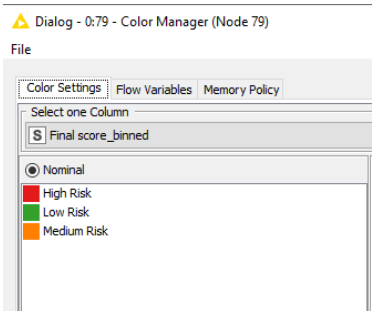


Students without Part Time Job

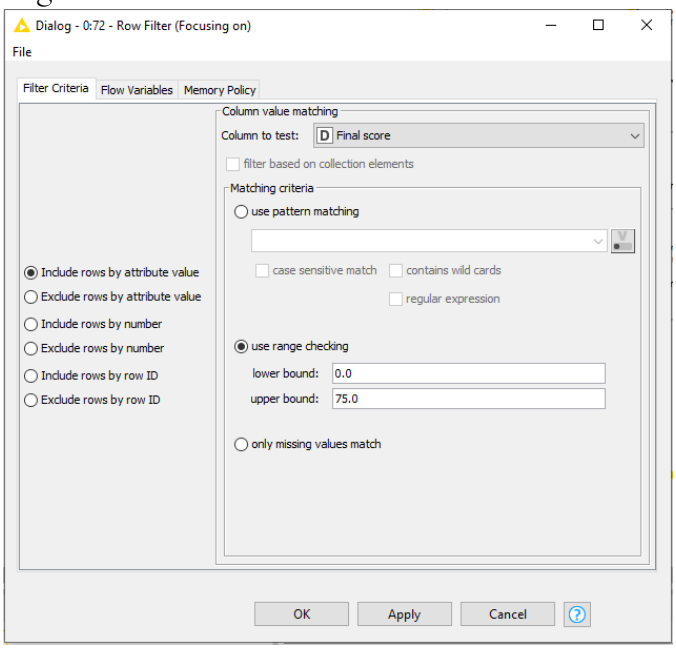


Row Filter

Focusing on the part time job, we kept the students who do not have part-time employment and filtered out those with using this Row Filter node. Doing this will help us with our upcoming parts.



High and Middle Risk Students

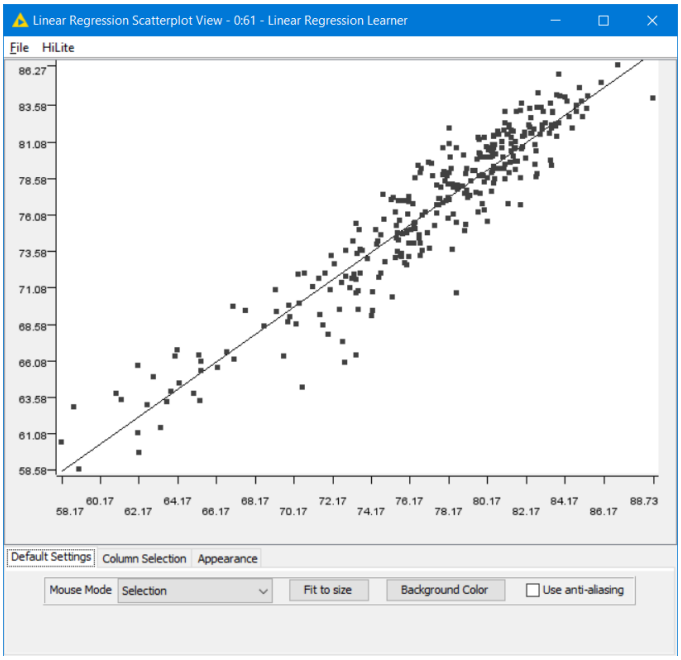


Row Filter

We are filtering to focus on students who are in the range of high and middle risk. These students will be scoring between 0 and 75 marks.

Data Exploration

First Sem. Score & Final Score



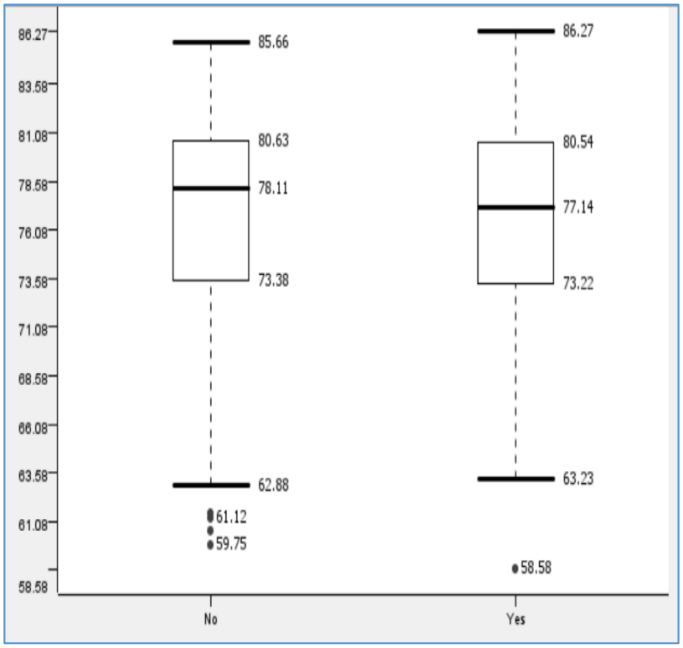
Linear Correlation &
Linear Regression Learner

We used the Linear Regression Learner node to determine the relationship between the two variables. The scatterplot indicates a strong positive correlation between First Sem. Score and Final Score with a value of 0.943.(value taken from the table below). We might then conclude that students who perform poorly in the first semester will perform poorly in the final semester. However, those who do well in the first semester will also do well on the final exam. Therefore, those who do not perform well needs to improve on their marks.

Row ID	S First column name	S Second...	D Correlation v...	D p value	I Degree...
Row0	Overall first sem. score	Final score	0.943276013597...	0.0	312

Part Time Job & Final Score

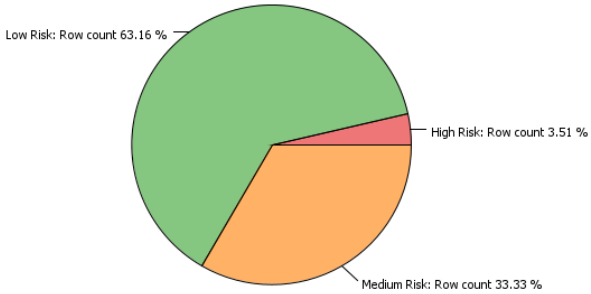
Conditional Box Plot



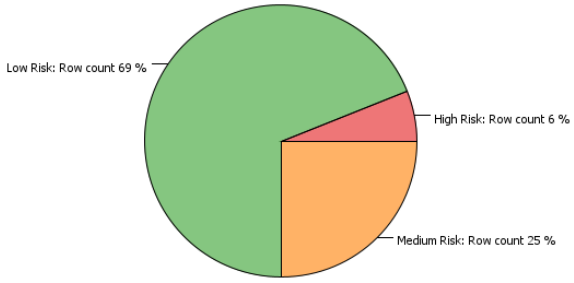
The boxplot above depicts the results of students who work part-time jobs (Yes) and those who do not (No). We believe that students who work part-time would do poorly in their tests, even if they were good at time management. We discovered that there is no significant variation in their median scores and risk percentage level between students who have (77.14) and do not have (78.11) part-time jobs based on the box plot. Moreover, those with (3.5%) and without (6%) part time jobs that are under high risks have no significant difference like the scores. As a result, we can conclude that students who work part-time are able to manage their work and it has no effect on their academic performance. As a result, we will be unable to identify "at risk students."

Interactive Pie chart (local)

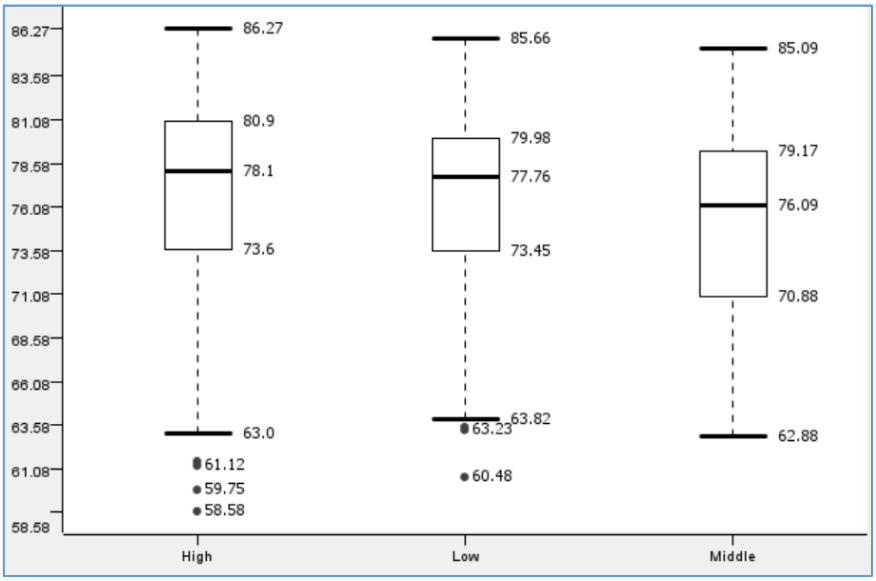
YES, part time job



NO, part time job



Students of different household incomes & Final Score | Conditional Box Plot



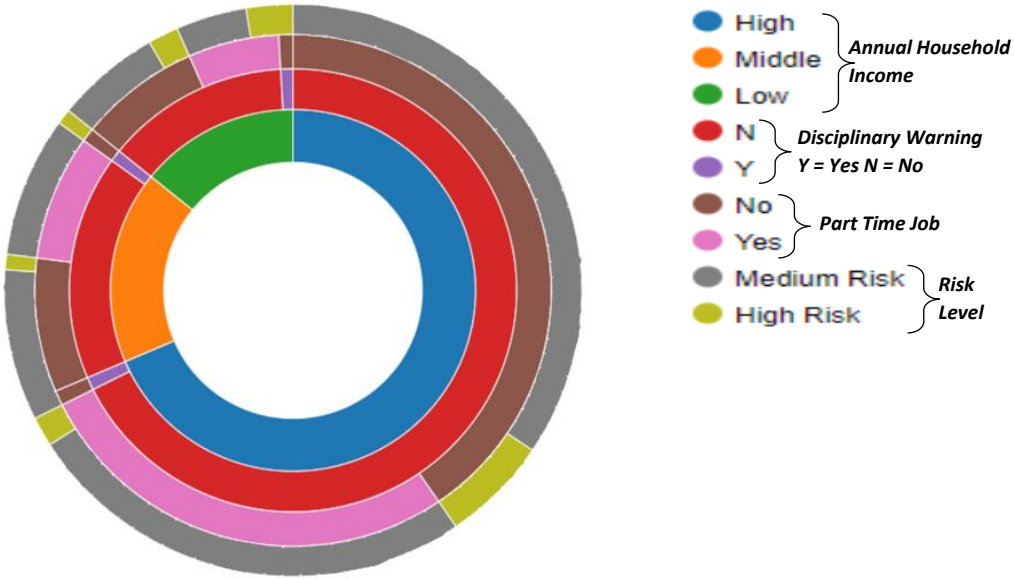
The above figure shows a comparison of grades amongst students from different household incomes. This comparison is done using box plots where we compare the inter-quartile range, and the full range of final scores. After the comparison, we were able to add on to our findings. We noticed that the grades of students from low and middle household income happen to score slightly lower than students of high household income. Though it does not prove that students of low and middle income are more likely to fall under the risk category, we can conclude that students of high income are less likely to fall under the risk category.

Students with and without Disciplinary Record | Group By

Row ID	S Disciplinary warning	I Count...	D Mean(Final score)	S Mode(Final score_binned)
Row0	N	311	76.598	Low Risk
Row1	Y	3	65.08	High Risk

The table above shows the comparison of final scores amongst students with and without disciplinary warning. Students with disciplinary warning score an average about 65.08 marks, causing them to fall under high risk. Meanwhile, students without any disciplinary warning tend to score better. An average of the final score for this group of students is 76.60 marks and most of these students fall under the "Low Risk" category. From this we can conclude that majority of students with disciplinary record fall under the "High Risk" category. Hence, when a student is found to have a disciplinary warning, he or she can be monitored from the start and provided the sufficient help needed.

Sunburst Charts



The sunburst chart above represents students of High Risk and Medium risk.
(High Risk and Medium Risk is determined from the final score and filtered using 'Row Filter' node)

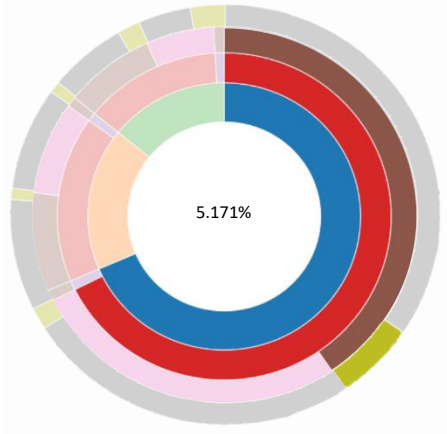
From this sunburst chart, we can see an overview of the percentage of students from different backgrounds falling under a risk category.
('different backgrounds' refer to annual household income, disciplinary warning, and part time job)

Reason for adding this node: By including a sunburst chart, we have the benefit to simply visualize the hierarchical data structure. In other words, it makes it easier to see multiple layers of data at once. The sunburst chart gives us a summary of all the focused data in one page and provides the proportion of certain variables.

<u>Examples on how we used sunburst to derive certain findings is shown below.</u>

Example 1

5.171%



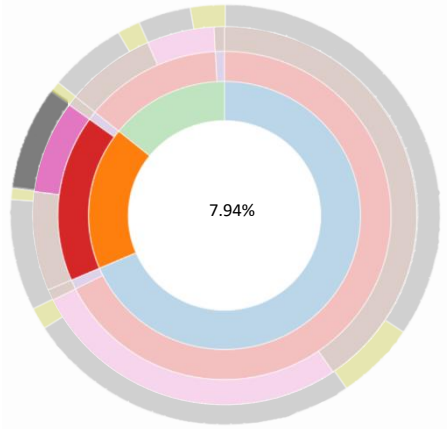
About 5.171% of the students who come from families with **high annual household income**, has no **part time job**, fall under the **High-Risk category**.

This 5.171% is also found to not have any **disciplinary record**.

This data can be later compared with similar income students without a part time job and/or a disciplinary record.

Example 2

7.94%



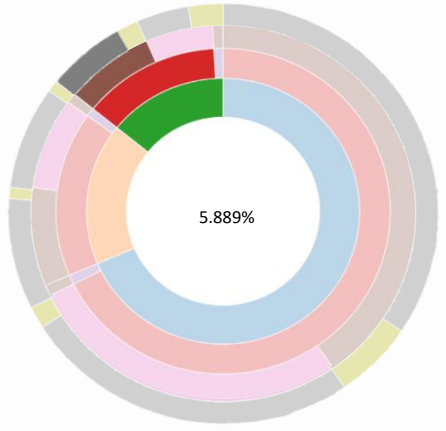
About 7.94% of the students who came from families with **medium household income**, have a **part time job** fall under the **Medium-risk category**.

This 7.94% is also found to not have any **disciplinary record**.

This data can be later compared with similar income students without a part time job and/or students with a disciplinary record.

Example 3

5.889%

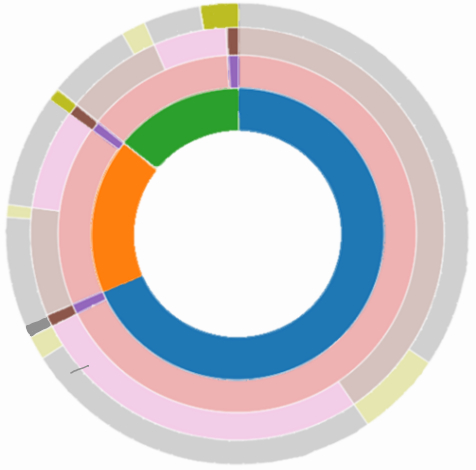


About 5.889% of the students who came from families with **low household income**, do not have a **part time job** fall under the **Medium-risk category**.

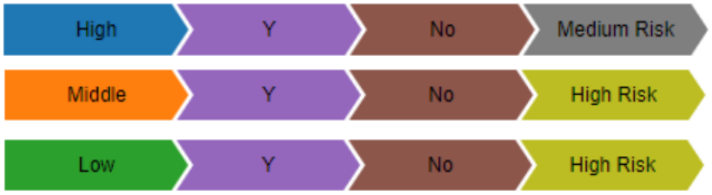
This 5.889% is also found to not have any **disciplinary warning**.

This data can be later compared with similar income students with a part time job and/or students with a disciplinary record.

Example 4



From this image, we can see that the 3 people with disciplinary warning are each from different levels of household income. The 2 students who are from **middle** and **low household income** fall under the **High-risk** category, while the student in **high household income** falls under **Medium-risk category**.



From this we can conclude that students who get disciplinary warning will have a higher chance of falling under a risk category. Most will fall under high risk, but there is still a possibility for some to fall under medium risk.