A red text on a white background

Description automatically generated

**SCHOOL OF INFORMATICS & IT**

Data Visualisation & Analytics

(CIA1C11)

**DATA ANALYTICS ASSIGNMENT**

**AY2024/2025 April Semester**

Student Name (Matric Number) : Mulakkal Rosemol Saju (2403878H)

Practical Group : P05

Jenny LING (TP)

Tutor : PohChai Chua

Submission Date : 27/7/2024

**Declaration of Originality**

I am the originator of this work and I have appropriately acknowledged all other original sources used as my references for this work.

I understand that Plagiarism is the act of taking and using the whole or any part of another person’s work, including work generated by AI, and presenting it as my own.

I understand that Plagiarism is an academic offence

and if I am found to have committed or abetted the offence of plagiarism in relation to this submitted work, disciplinary action will be enforced.

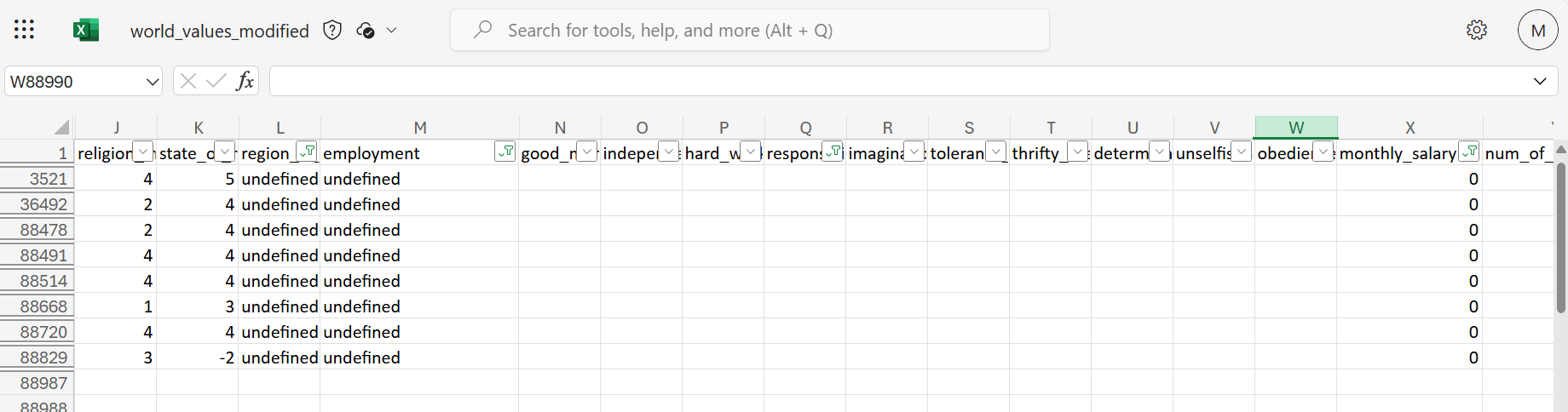
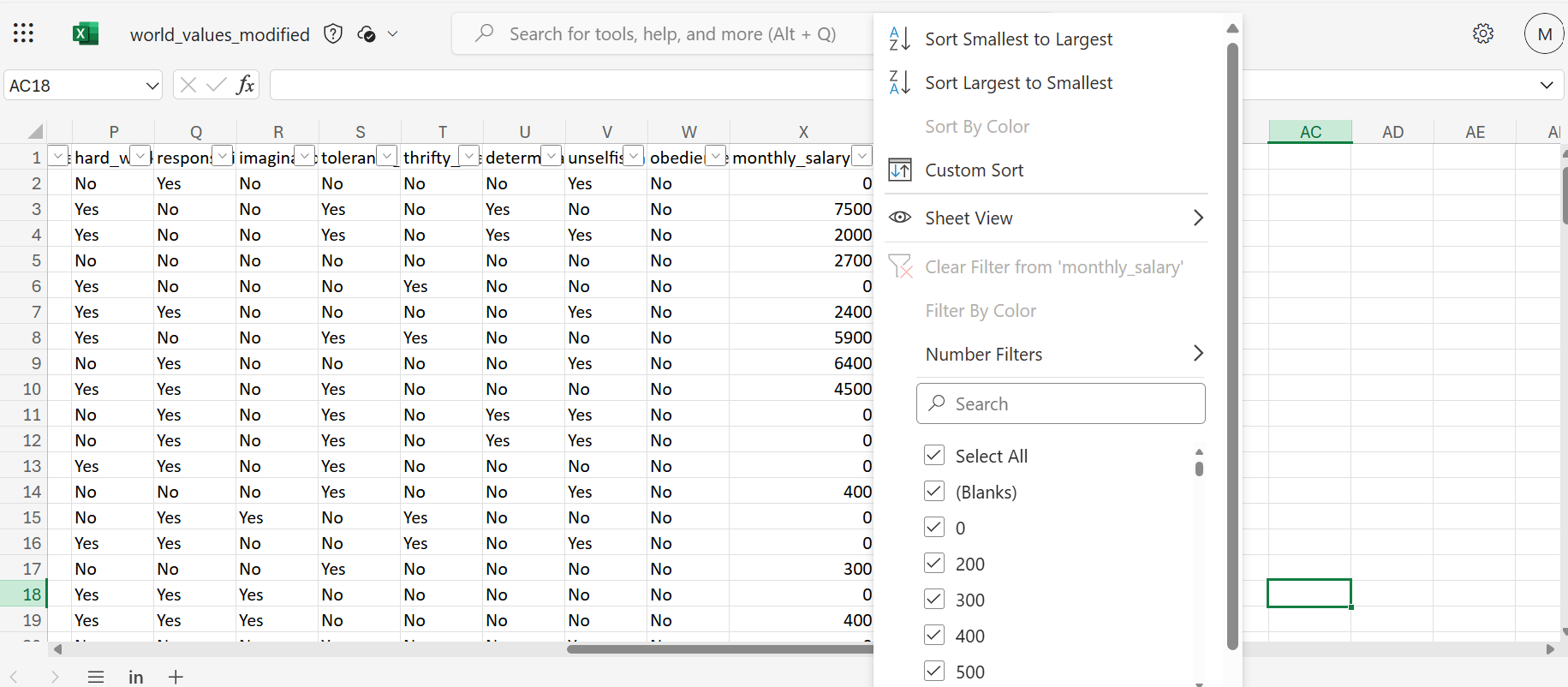
# It’s Data Prep Time

**1(a).** **Data Profiling**

To profile data, we need to check for the **missing values, inconsistency, outliers and duplicates**.

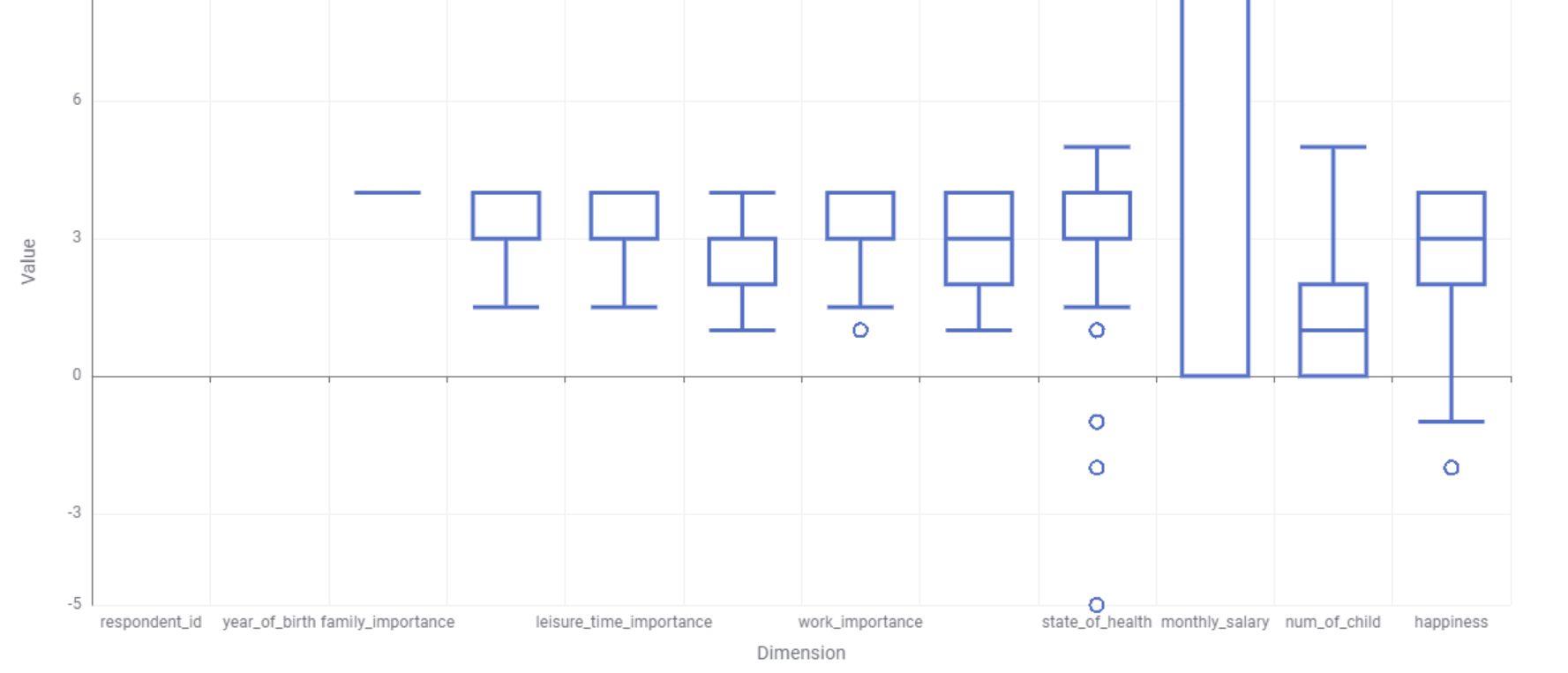
**Missing values:** In the dataset provided, there are several columns that contain missing values. These columns include monthly\_salary, good\_manners\_mentioned, independence\_mentioned, hard\_work\_mentioned, responsibility\_mentioned, imagination\_mentioned, tolerance\_and\_respect\_mentioned, thrifty\_mentioned, determination\_mentioned, unselfishness\_mentioned, and obedience\_mentioned. Missing values are problematic because they can lead to biased analyses which can lead to inaccurate conclusions.

The presence of blank indicates there are missings values in the columns.

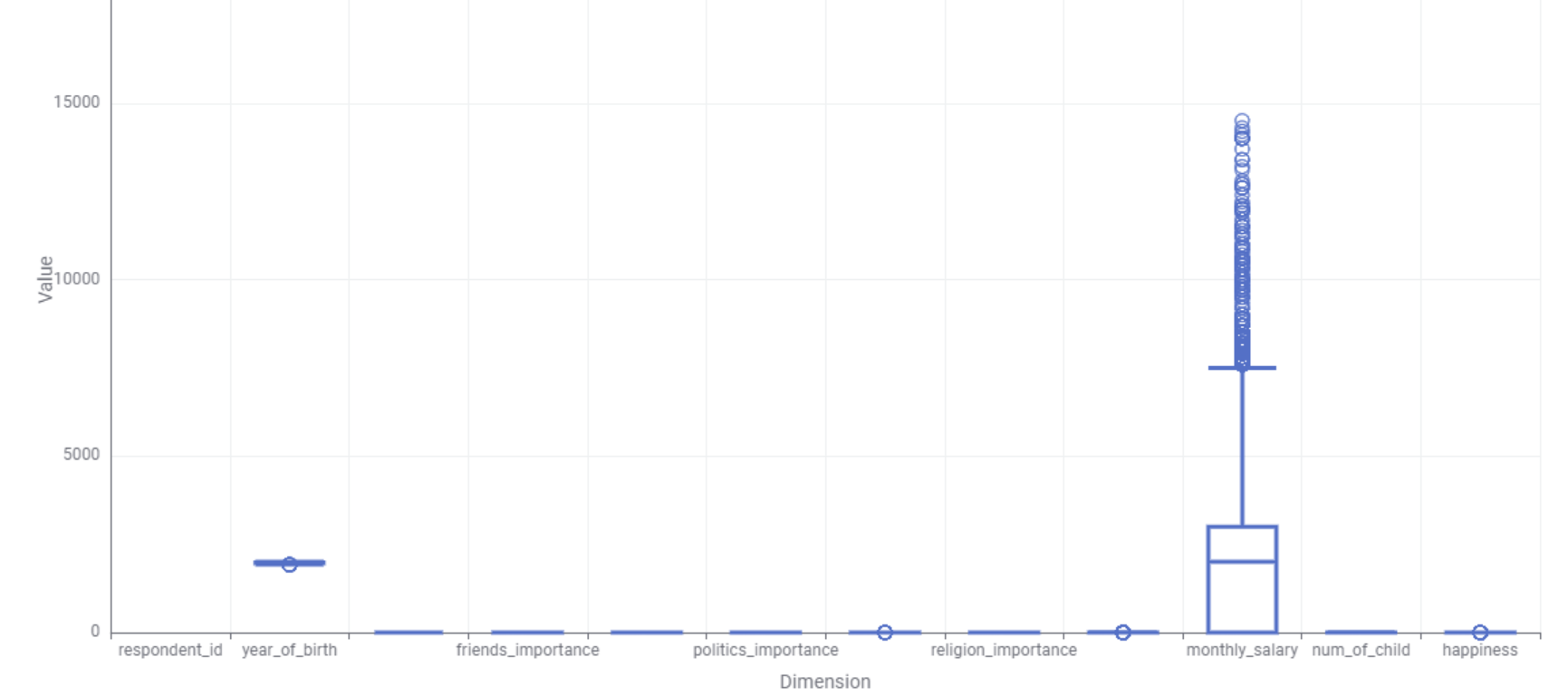


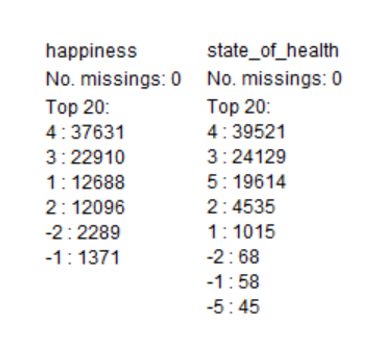
**Outliers:** Using KNIME box plot node, there were many outliers present. Specifically, outliers were observed in work\_importance, state\_of\_health, happiness, monthly\_salary, and year\_of\_birth. These outliers are visually represented by data points that extend beyond the whiskers of the box plot. Outliers can significantly impact the data by potentially skewing the results which will affect the mean. To handle the outliers we can trim it, transformation, or exclusion depending on the type of data.

The outliers below the boxplot these outliers tend to skew the mean to a lower value.



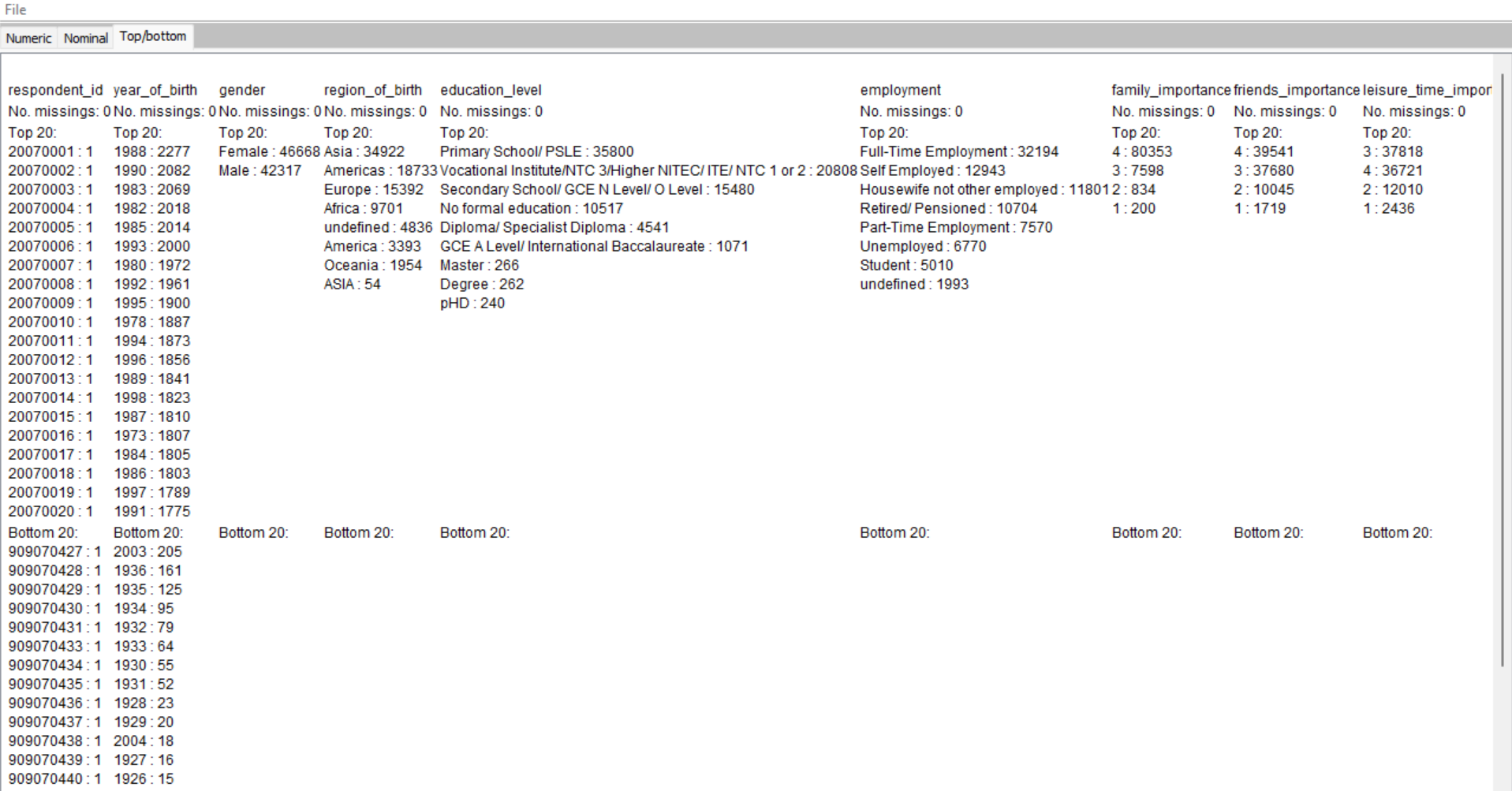
The outliers above the boxplot tend to skew the mean to a higher value



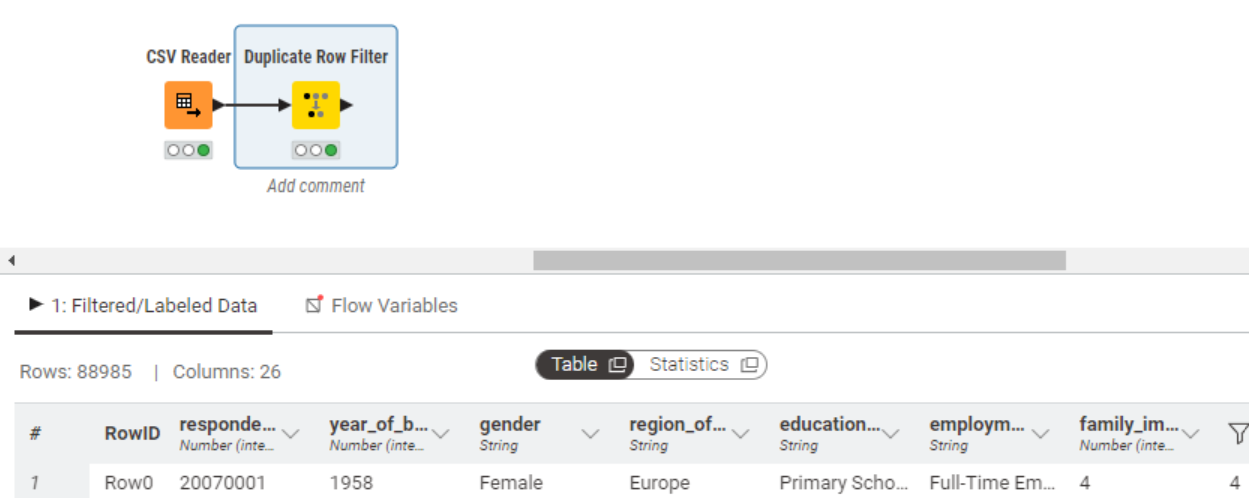
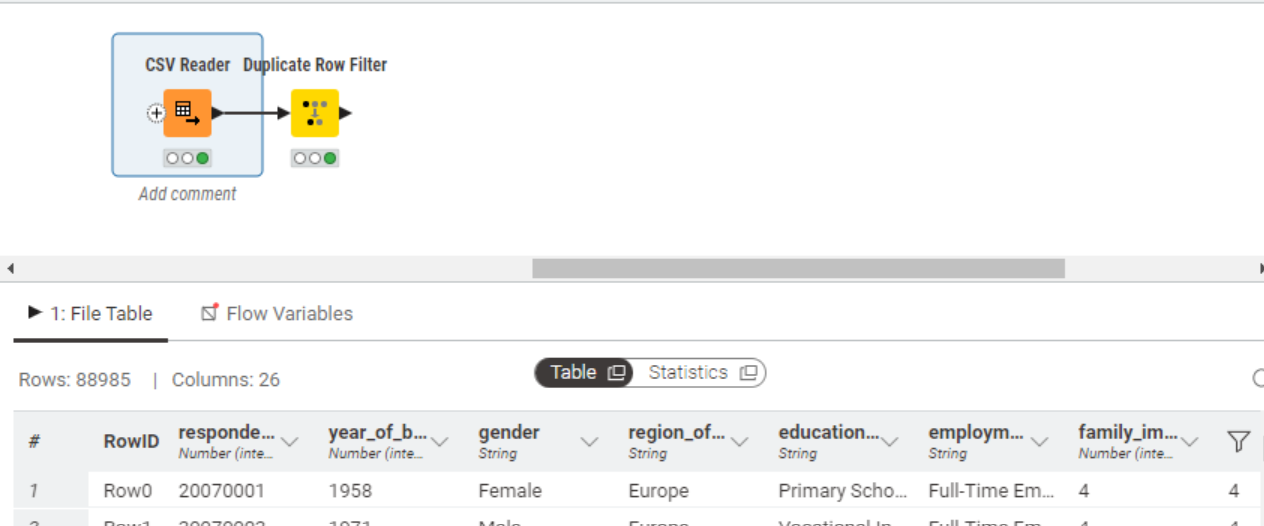


The negative values here will skew the mean to a much lower value.

**Inconsistent values:** The inconsistent values are in the region of birth and employment. Both have “undefined” as their inconsistency and region of birth has “ASIA” in all caps and “Americas” instead of America. To resolve these inconsistencies, I will utilize KNIME's Rule Engine node.

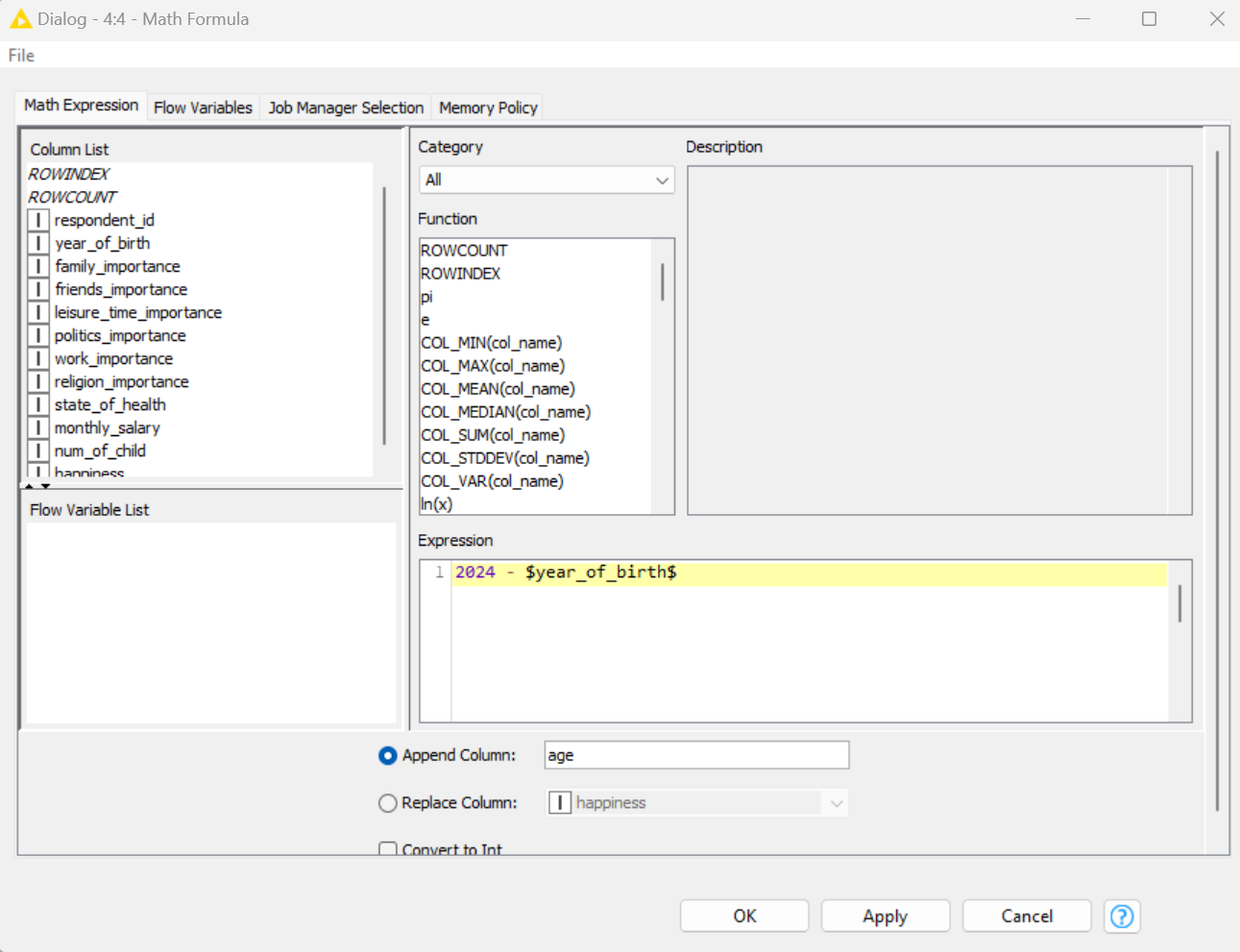


**Duplicates**: There are no duplicates as seen below. I have used the KNIME duplicate row filter to check if there are any duplicate rows. Since both have the same number of rows there are no duplicate rows. If there were duplicates, we will need to use the remove duplicate rows using Knime or aggregating duplicate entries to consolidate information.



**1(b).** **Data Cleaning**

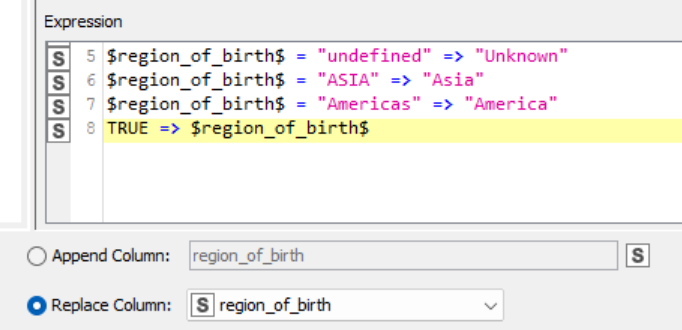
1. Derive a new column called age.

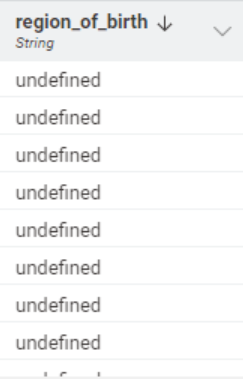


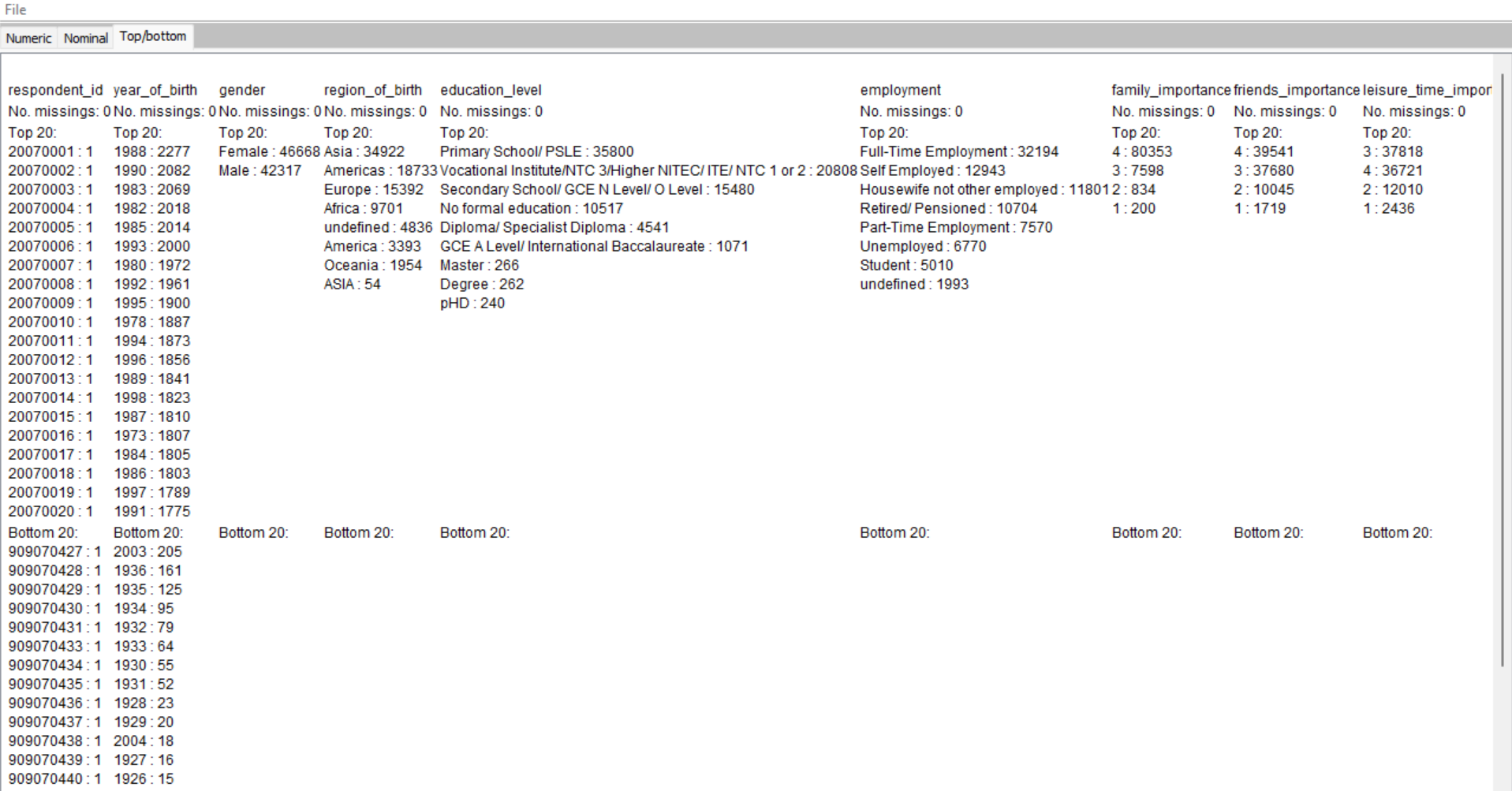
Added an age coloumn using math formula node



1. Rectify the problems found in the column called region\_of\_birth



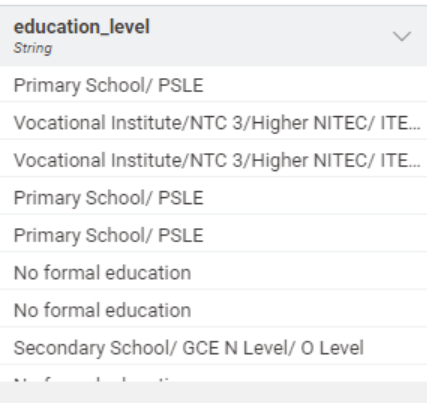
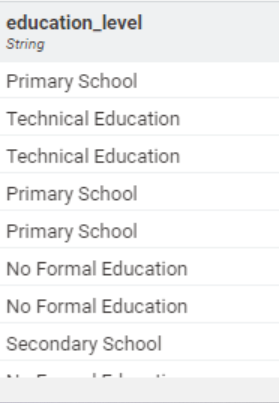
 



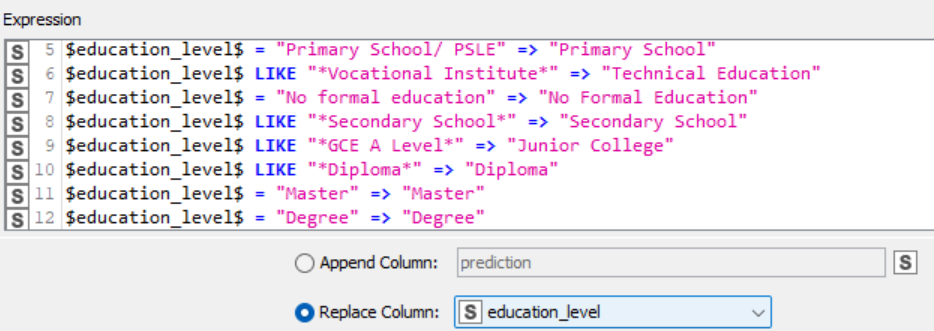


I used rule engine node and rectified the inconsistancies. I changed “ASIA” to Asia and “Americas” to America and “undefined” to Unknown. I changed it to unknown as it makes more sense to others as to where they were born.

iii. Recode the education\_level column into the following 9 categories using a Rule Engine node.

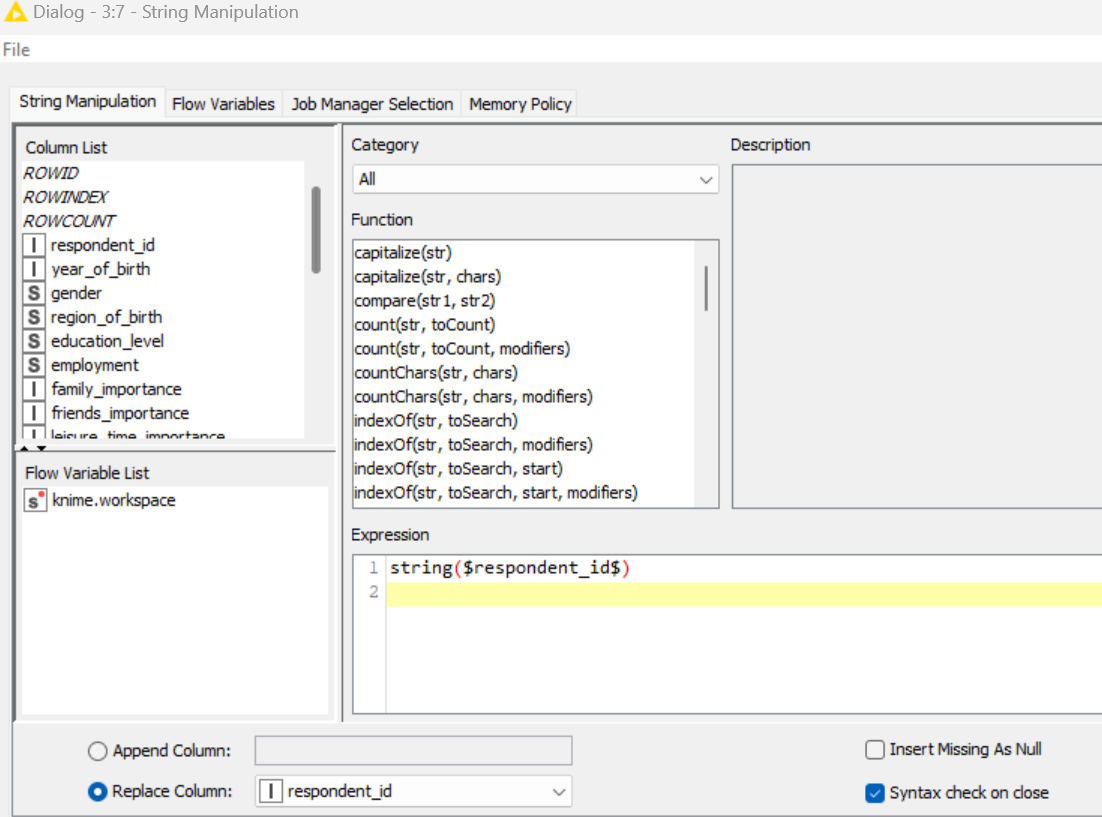
 

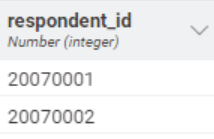
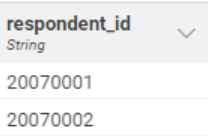
I used rule engine and recoded the nine categories.



iv. Convert the columns respondent\_id to string type as they represent IDs

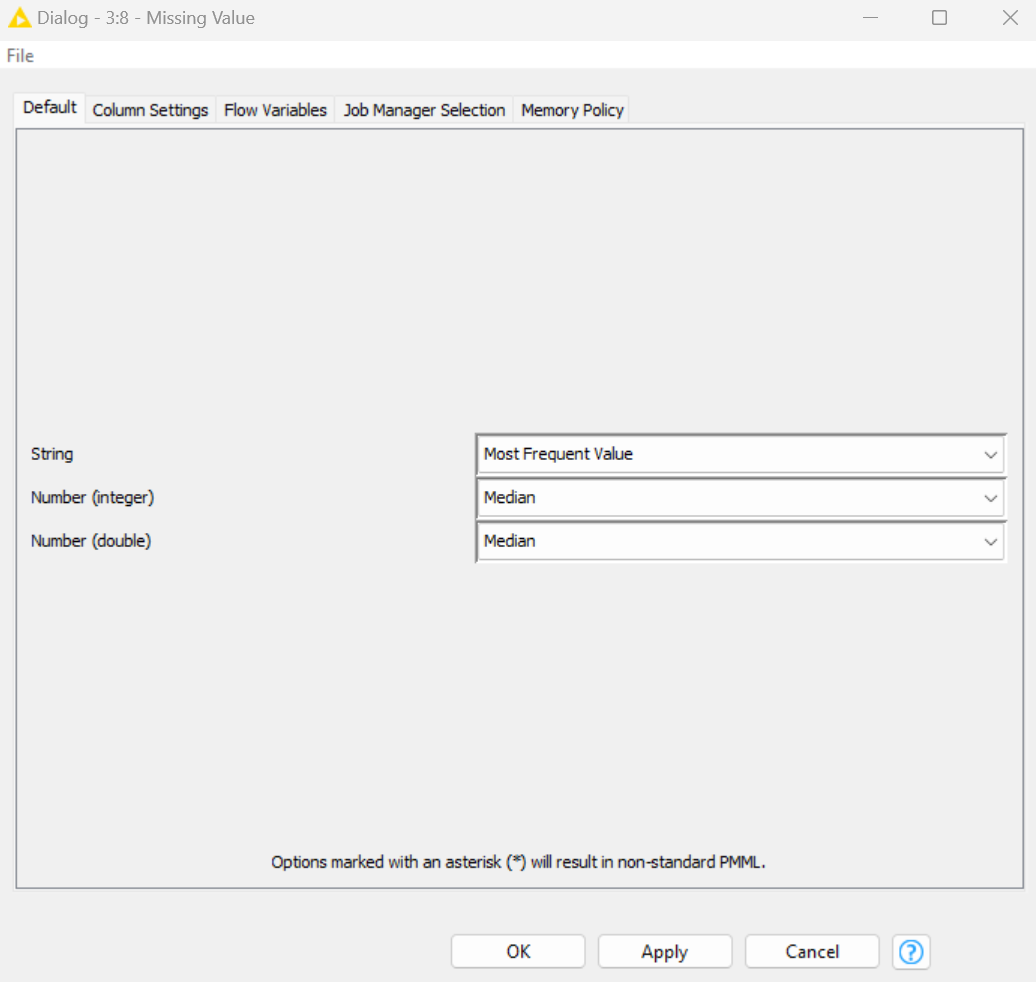
Using string manipulation node can change integer to string



v. Are there any missing values in the data? If so, decide on the treatment for each case. Execute the handling of missing values. Justify the method chosen for each case.

**Using Missing Value Node,** I have decided **for string** that I should not delete it as a large proportion of data will be lost but instead changed to **most frequent value** and **for numbers to median.**

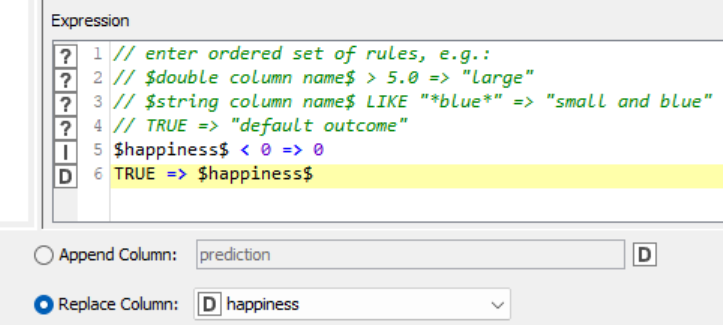


It is better to convert to median for numeric data as the median is less sensitive to outliers compared to mean.

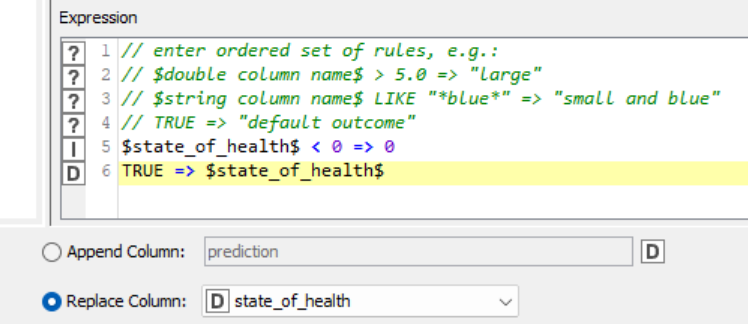
vi. Are there any negative values in the data? If so, decide on the treatment for each case. Execute the handling of negative values. Justify the method chosen for each case.

There are **2 negative values**, and these values will skew the mean results to a lower value which affects the data. The 2 negative values are **happiness and state of health**. These two are both important indicators represented by the ordinal scale where negative values will be incorrect and anomalous. To address this, I used the **rule engine node** to **replace all the negative values to the least possible value in this case 0**. This approach ensures that the data set maintains its integrity and remains accurate to use while also preserving the nature of the intended data inputed.



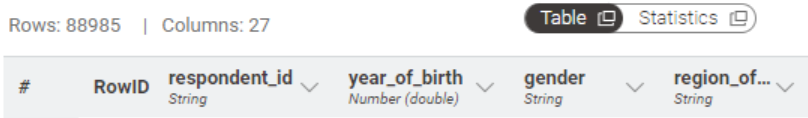
Using rule engine I changed all negative values to 0.

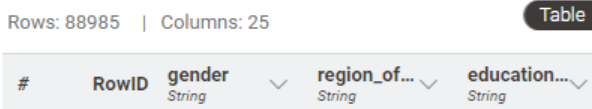


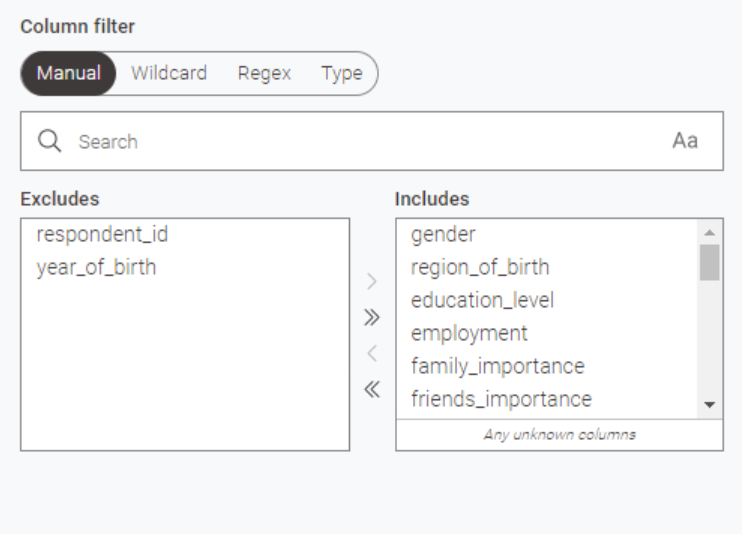
vii. Filter out redundant columns.

To me respondent\_id(Unique ID of the respondent) is often not used for analysis unless identifying specific respondents and hence not necessary. Also, year\_of\_birth since we already calculated the age.

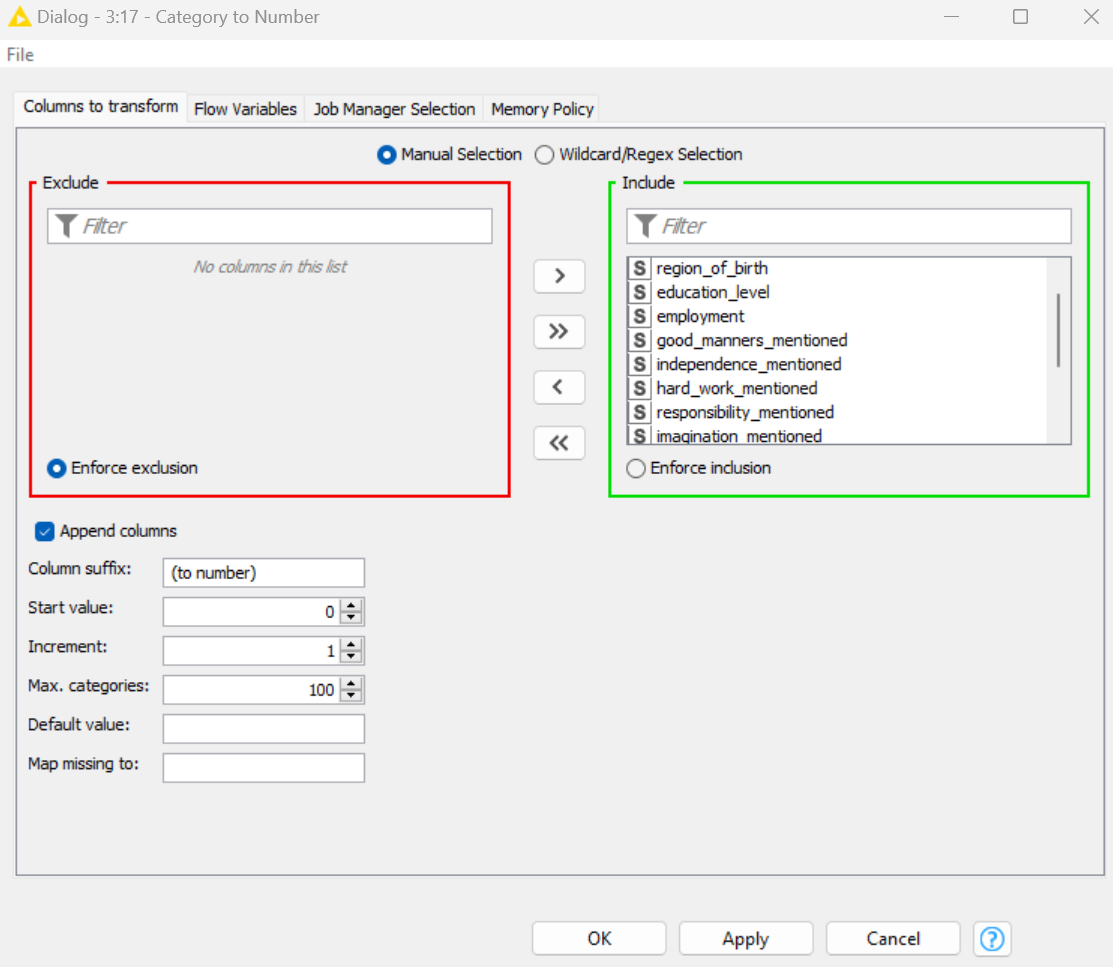




Using column filter remove responded id and year of birth.



viii. To facilitate downstream work for clustering or linear regression (which works with only numeric values), you may need to convert some of the columns to numeric data types. Decide on the best course of action and execute the conversion.



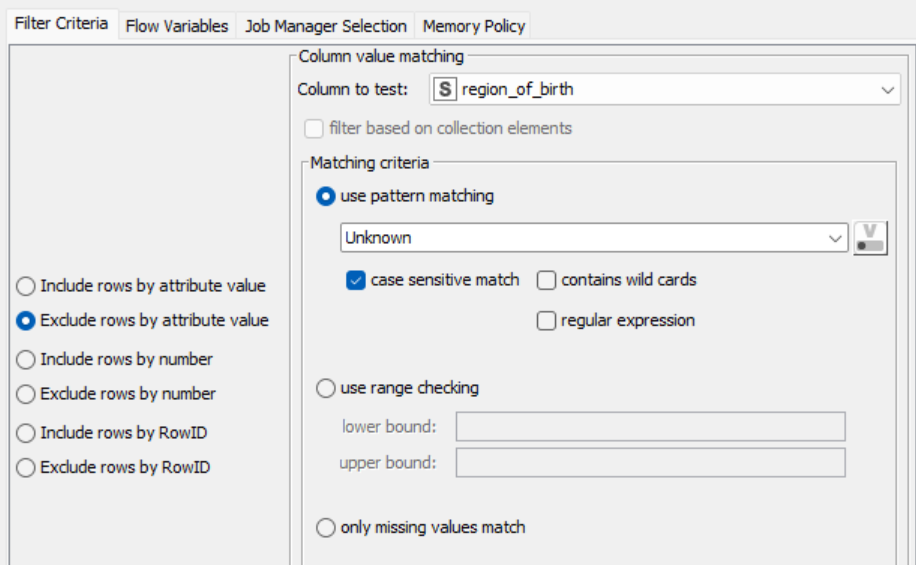
To be able to do clustering I changed state of health and all other string values to numeric data type using the Category to number node.

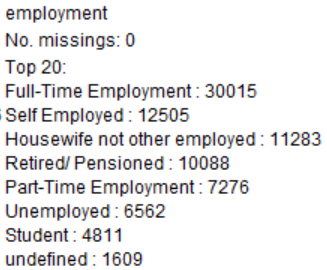
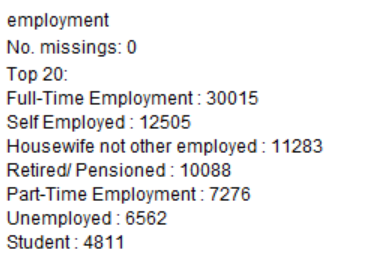
1(c). Other Possible Improvements

I have chosen to remove rows which are undefined in employment and unknown in region of birth since these values do not add value into the data set and are useless remaining inside data set.

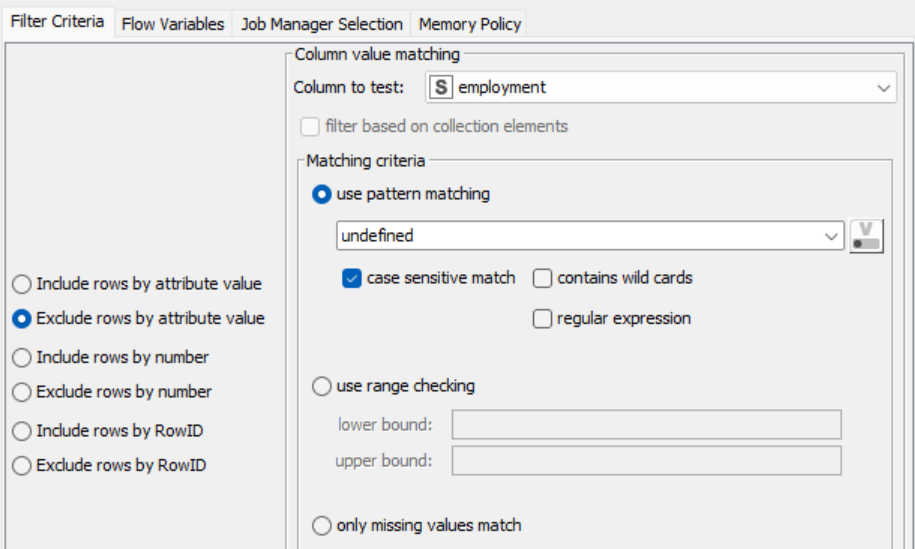
 

Using row filter I removed region of birth

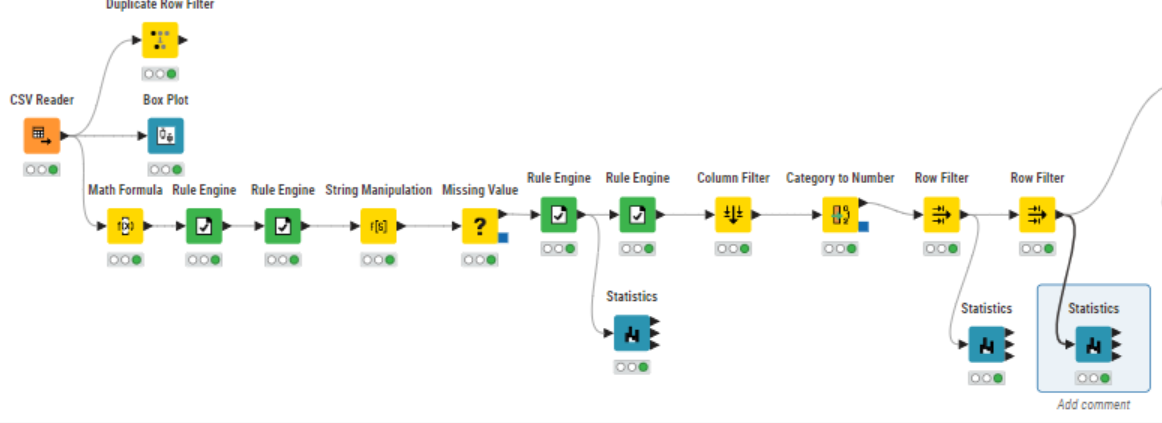


Using row filter, I removed undefined in employment.

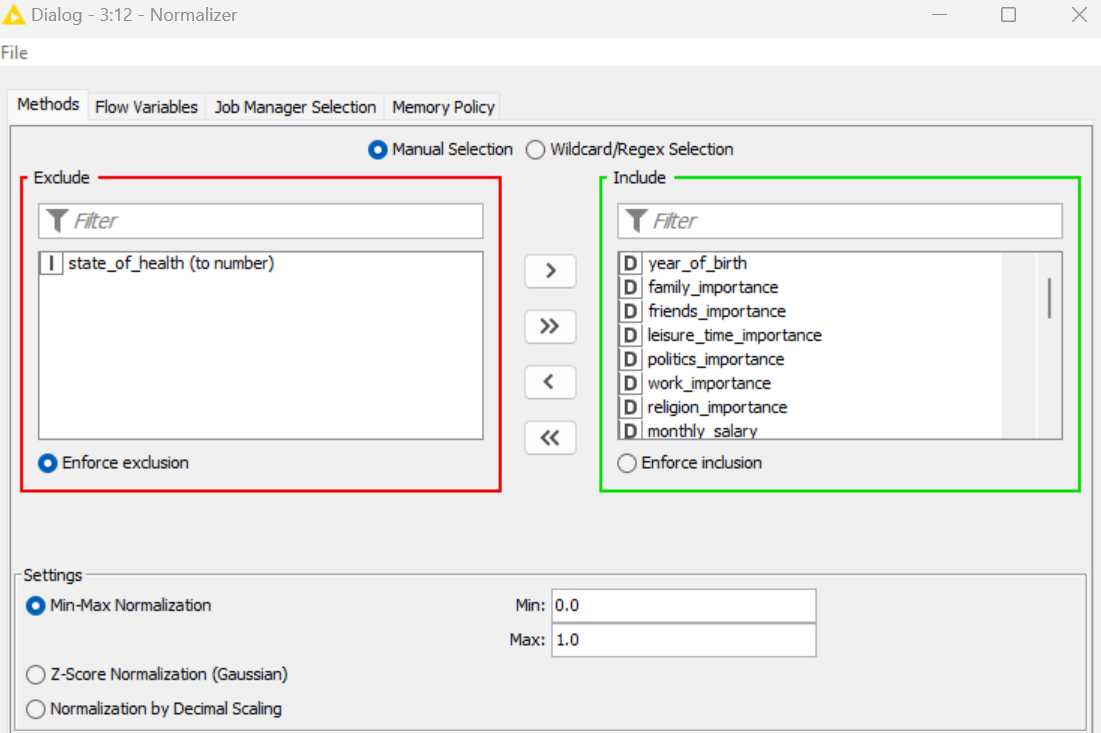


**Workflow in Knime to Clean and Profile Data.**



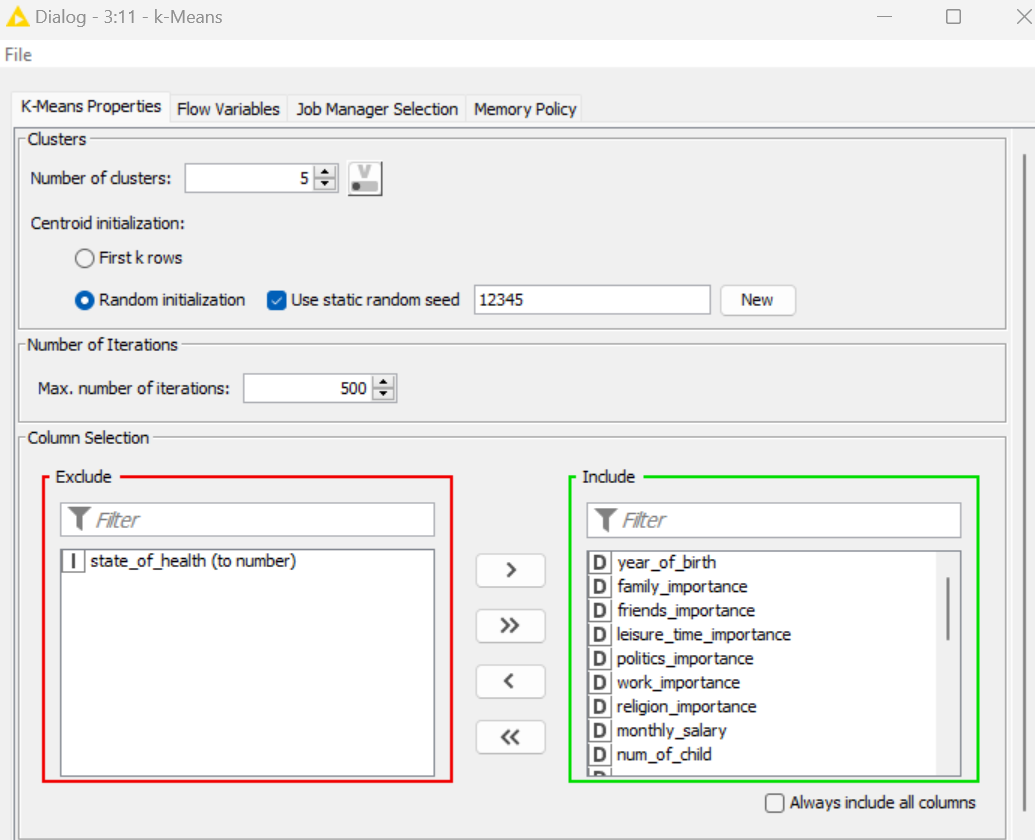
# Clustering

2(a). Steps taken:

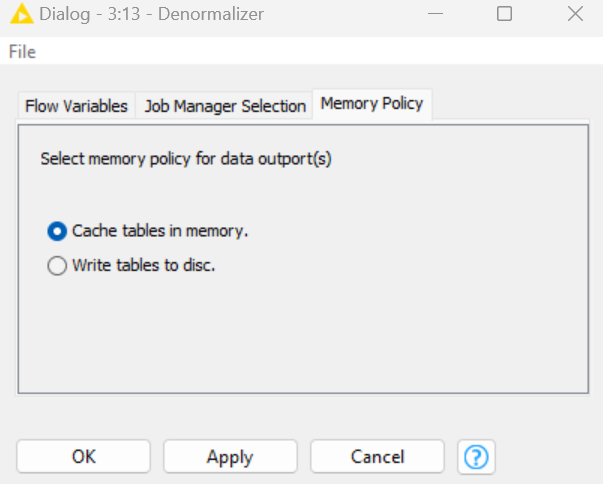
****

\*Need to **exclude state of health** to see if we can group their health by their characteristics.

**STEP 1:** We then need to use normalizer node to make sure that we normalize the fields so that it would reduce the effects of unequal ranges across the fields.

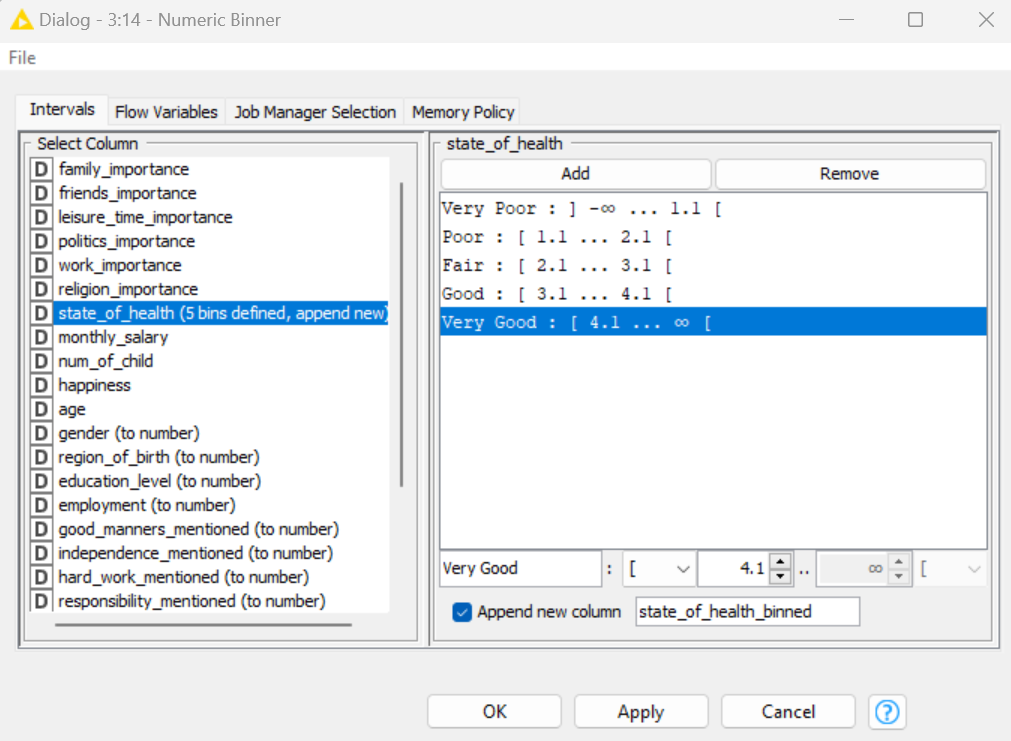
****

**STEP 2:** K means

****

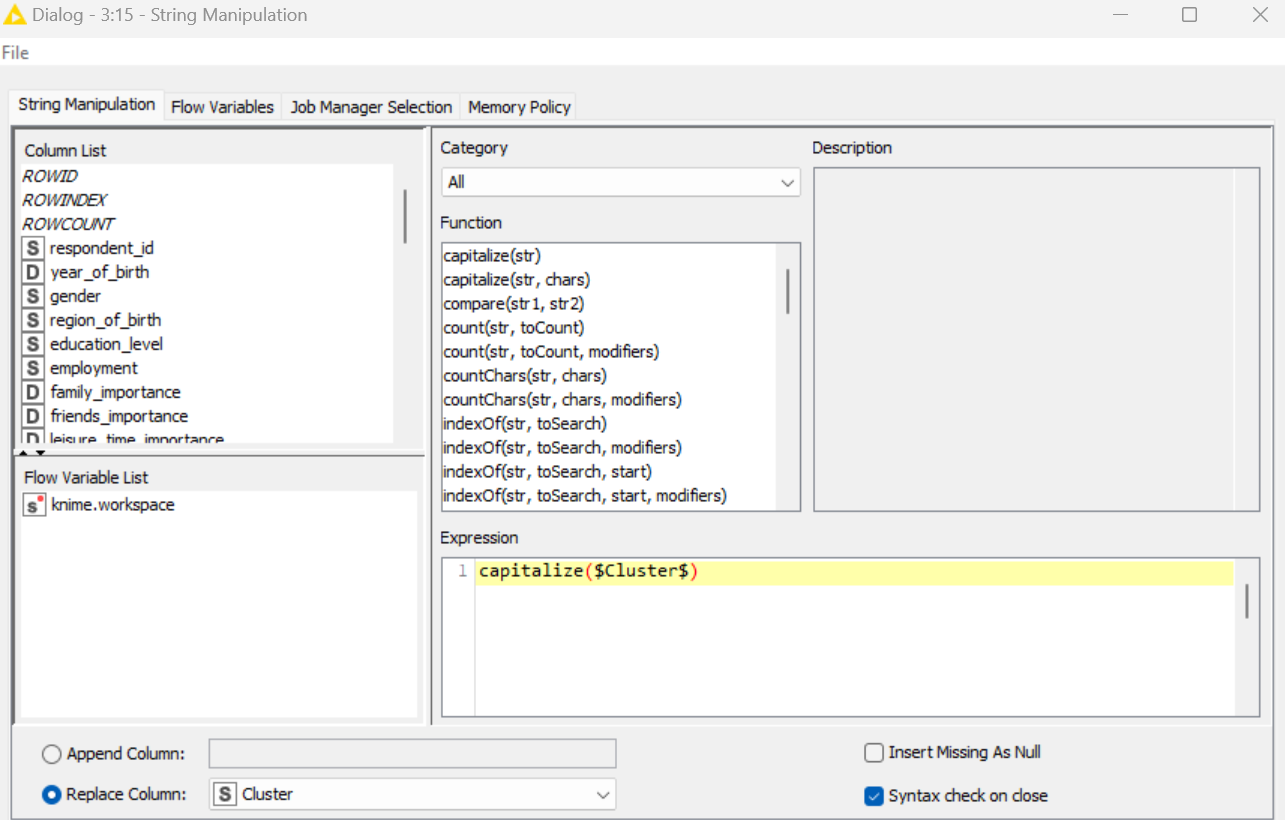
\*Use denormaliser to change the data back to original measurement of scale.

**STEP 3:** Demonrmalizer

****

\*Used 1.1 , 2.1, 3.1,4.1 as the number. This is because after checking, the numbers will tend to double count if I put the range like 1.0-2.0 and 2.0-3.0\*

**STEP 4:** Use numeric binner to categorise the state of health into 5 bins.

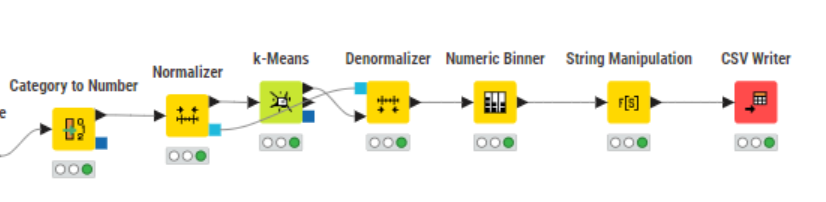
****

Results

**STEP 5:** Now we need to write this information into a file so that we can then visualize it in our dashboard tool, so we need to use the string manipulation tool.

****

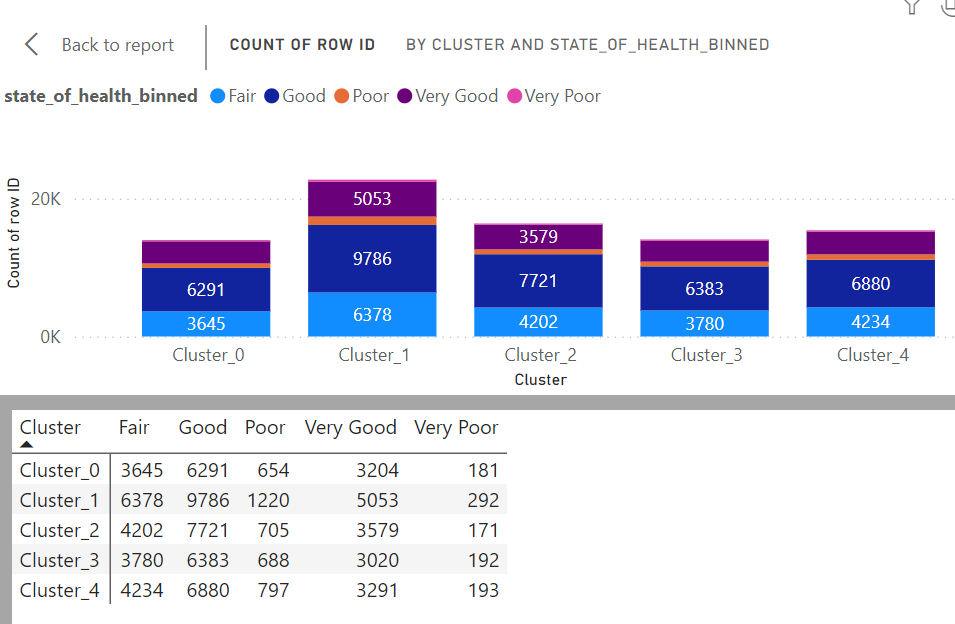
**Finally, I added a csv writer to write a csv file which I can then use in Power BI to visualise data.**

****

Workflow for Clustering

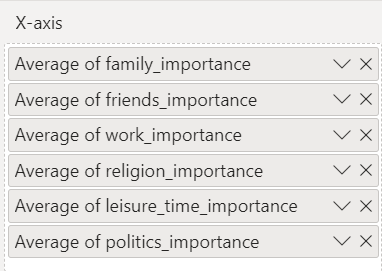
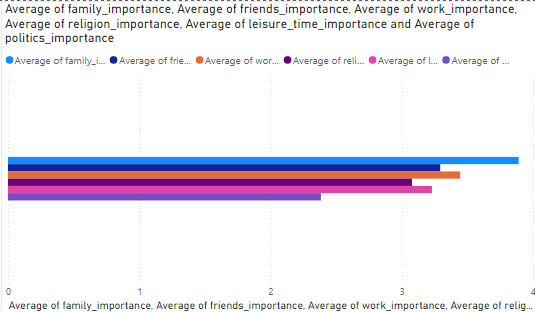
**Use Power BI to determine if there is a high degree of correspondence between the clusters and state\_of\_health. Explain your findings and provide relevant screenshots**.

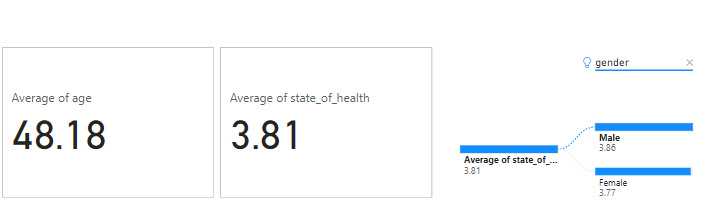
After inputing the data inside PowerBI, I used the bar chart to see the correspondence. Based on the bar chart as seen below, **there is a low correspondence**. This is because **all the clusters have the same proportions** (of fair, good, poor, very good, very poor state of health) and **they are evenly distributed across all the clusters** **which suggests it has a weak relationship** and low correspondence. Eg: If there was a high correspondence each cluster will each have high percentage of one state of health which the other clusters will not have.

****

**2(b). Findings / Explanation from Power BI:**

I am using the clustered bar chart to find the characteristics of the respondents in each cluster.

****

****

The average age of the respondants is 48.18 and their average state of health is 3.81. Furthermore the males generally have a better state of health(3.86) than females(3.77) as seen in the decomposition tree.

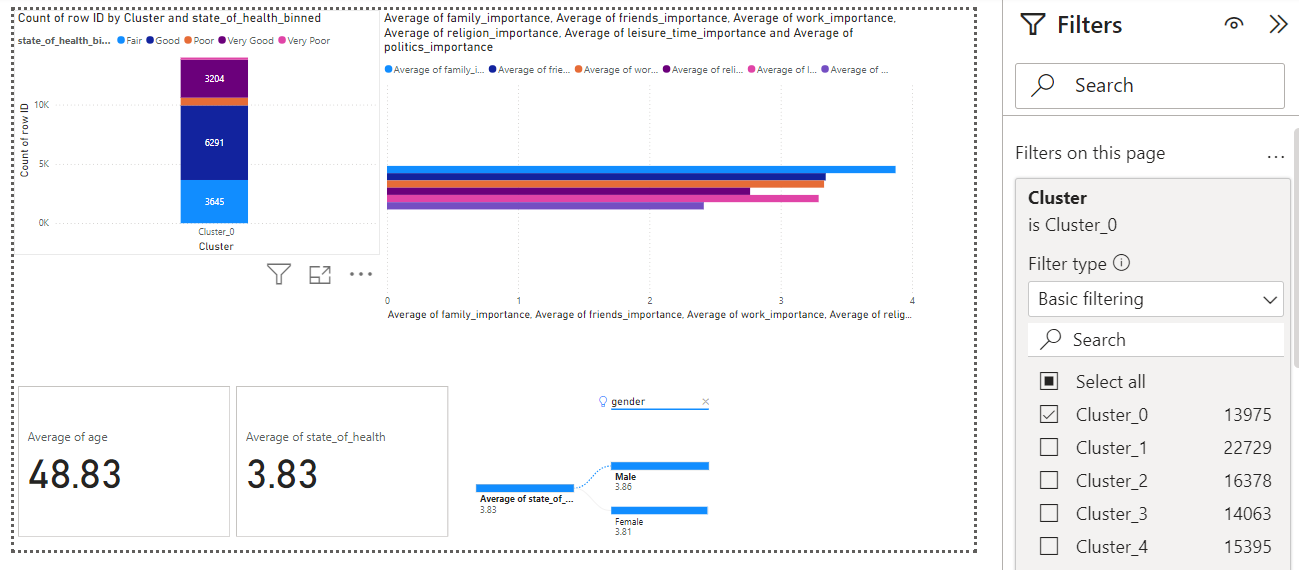
**Cluster 0**

**Characteristics**:

* High importance placed on family, friends, work, and leisure\_time.
* Low importance on religion and politics.
* Likely to have a balanced lifestyle with a focus on personal relationships and enjoyment.
* Average age 48.83
* Average state of health 3.83

**Key Insights**:

* Respondents are likely to prioritize work-life balance and personal satisfaction over religious and political involvement.

****

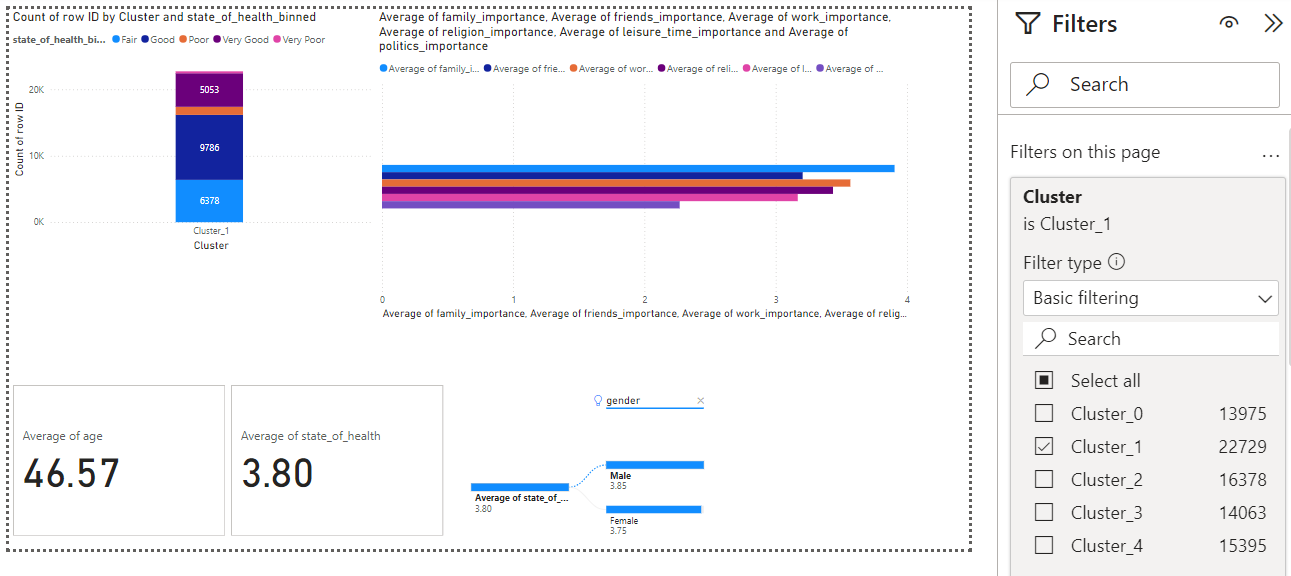
**Characteristics**:

* + High importance on work and religion.
  + Low importance on friends, leisure\_time, and politics.
  + Indicates a focus on career and religious values, potentially at the expense of social and leisure activities.
  + Average age 46.57
  + Average state of health 3.80

**Key Insights**:

* + Respondents are likely driven by career ambitions and religious commitments, with less emphasis on social and leisure aspects.

**Cluster 1**

****

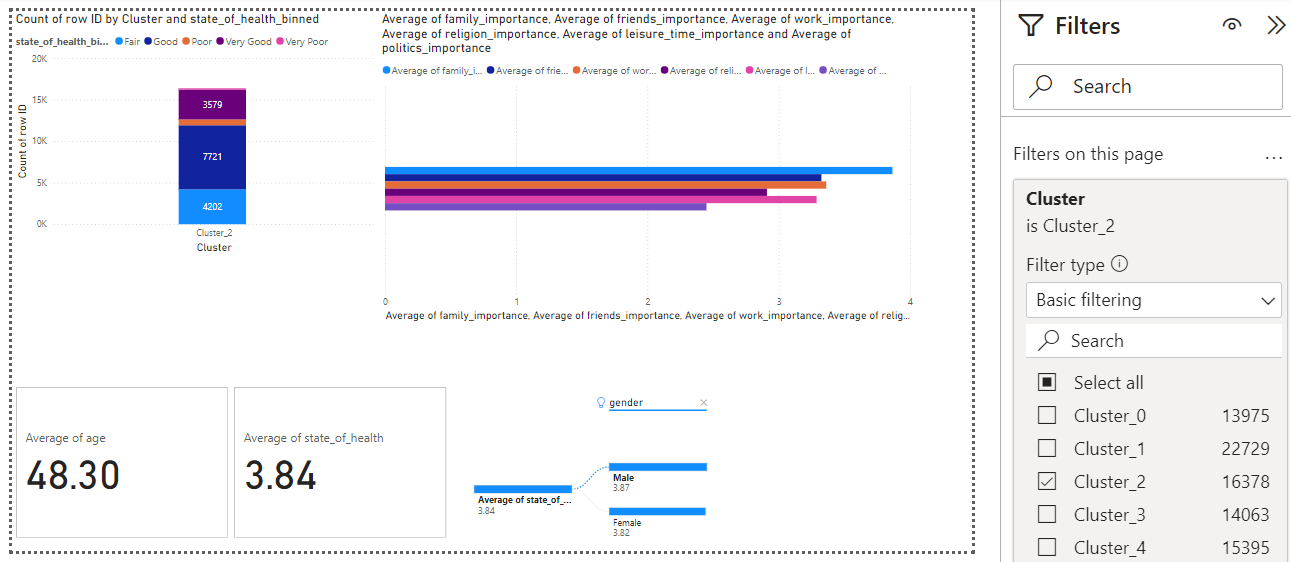
**Cluster 2**

**Characteristics**:

* + Similar to Cluster 0 with high importance on friends, work, and leisure\_time.
  + Low importance on religion and politics.
  + Average age 48.30
  + Average state of health 3.84

**Key Insights**:

* + This cluster shares a similar lifestyle focus with Cluster 0, emphasizing personal and social fulfilment.



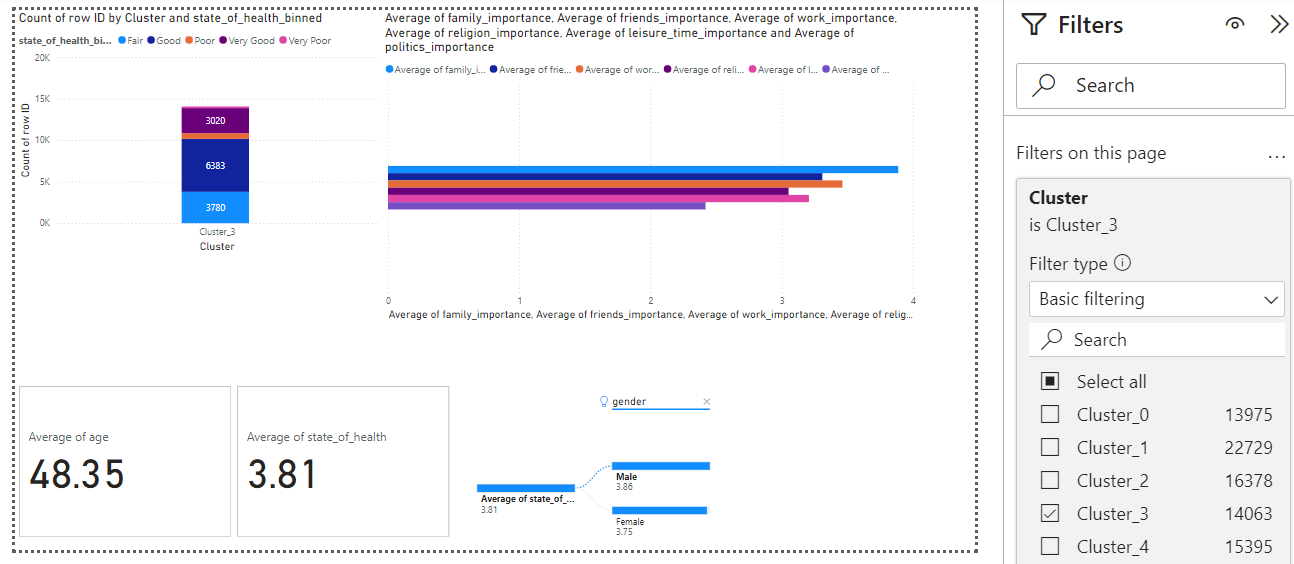
**Characteristics**:

* + Almost identical to Clusters 0 and 2 with high importance on friends, work, and leisure\_time.
  + Low importance on religion and politics.
  + Average age 48.35
  + Average state of health 3.81

**Key Insights**:

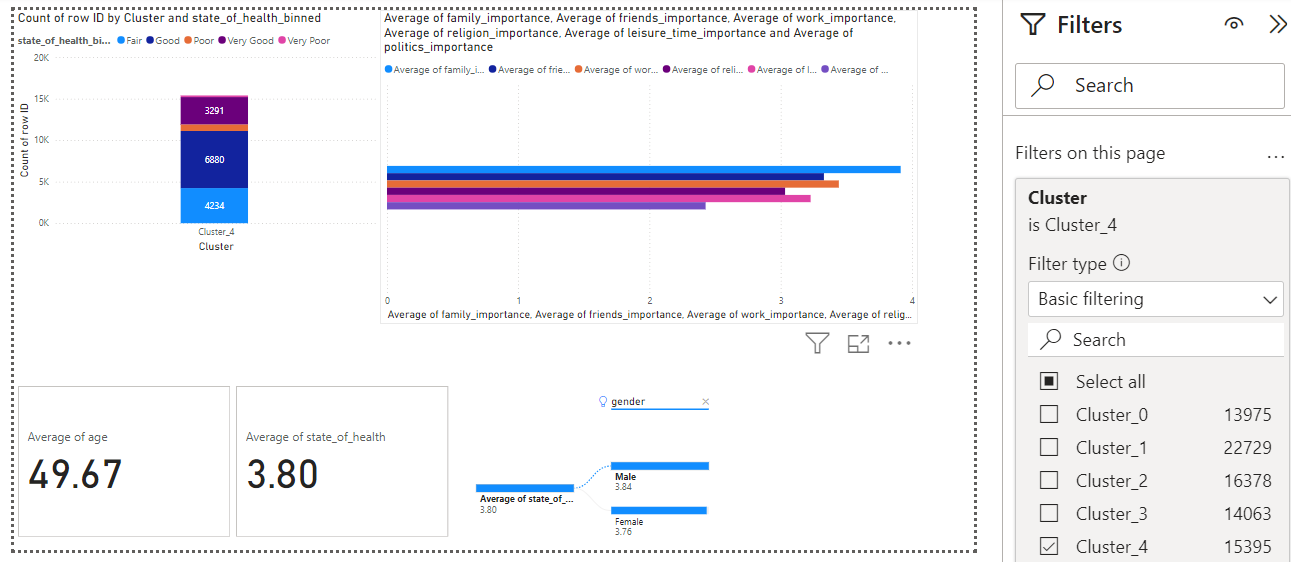
* + Respondents in this cluster also prioritize a balanced lifestyle focused on work and social life, similar to Clusters 0 and 2.

**Cluster 3**



**Cluster 4**

* **Characteristics**:
  + Similar to Clusters 0, 2, and 3 with high emphasis on friends, work, and leisure\_time.
  + Low emphasis on religion and politics.
  + Average age 49.67
  + Average state of health 3.80
* **Key Insights**:
  + This cluster exhibits the same lifestyle preferences as the other clusters that value a balanced personal and social life.



**Cluster 0, 2 , 3 4**

**Shared/Similar Characteristics Between clusters 0, 2, 3 and 4:**

1. **High Emphasis on Friends, Work, and Leisure:**
2. **Low Importance on Religion and Politics:**

**Differences: While** Clusters 0, 2, 3, and 4 exhibit a similar lifestyle pattern, there are subtle differences in how strongly each cluster prioritizes work, friends, and leisure. The average age and state of health may vary slightly among these clusters, but overall, they share a consistent approach to balancing work and personal life.

**Cluster 1**

**Distinctive Characteristics:**

1. **High Emphasis on Work and Religion:**
2. **Less Focus on Politics and Friends:**

**Differences from Other Clusters: The Lifestyle Prioritization** is different as unlike Clusters 0, 2, 3, and 4, which emphasize a balance among work, social life, and leisure, Cluster 1 has a more pronounced focus on career and religion. This results in less attention to social activities and political involvement.

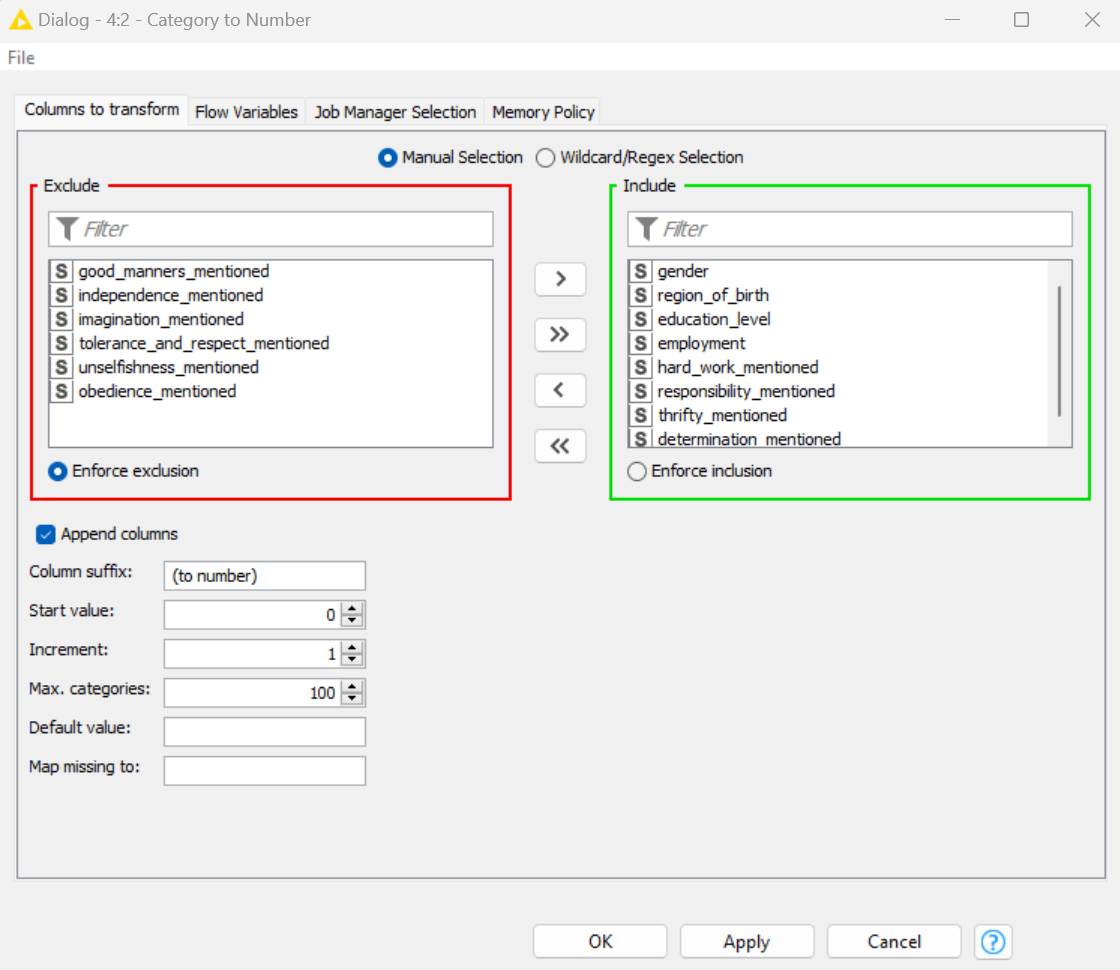
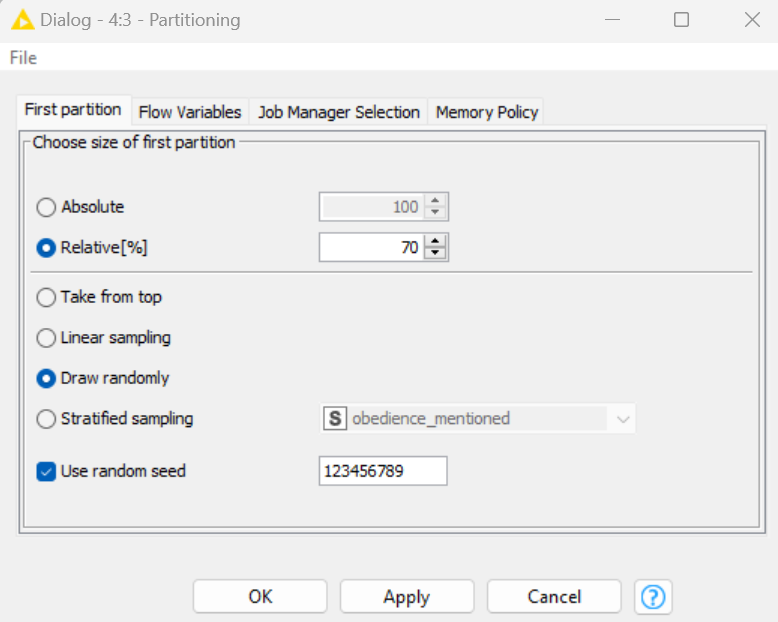
**Conclusion**

Clusters 0, 2, 3, and 4 exhibit a similar lifestyle profile, characterized by a strong emphasis on maintaining a balance between friends, work, and leisure time. These clusters consistently prioritize social interactions, career achievements, and recreational activities while placing lower importance on religion and politics. Although subtle differences exist among them, such as very minute variations in age and health status, the overall lifestyle approach remains consistent. In contrast, Cluster 1 stands out with its distinct focus on career and religious commitments. Respondents in this cluster place a high priority on work and religious values but exhibit less interest in social activities and political engagement. This prioritization results in a unique lifestyle profile where the emphasis on career and religion may impact social connections and overall health compared to the more balanced approach observed in the other clusters.

# Linear Regression

3(a). Steps taken:

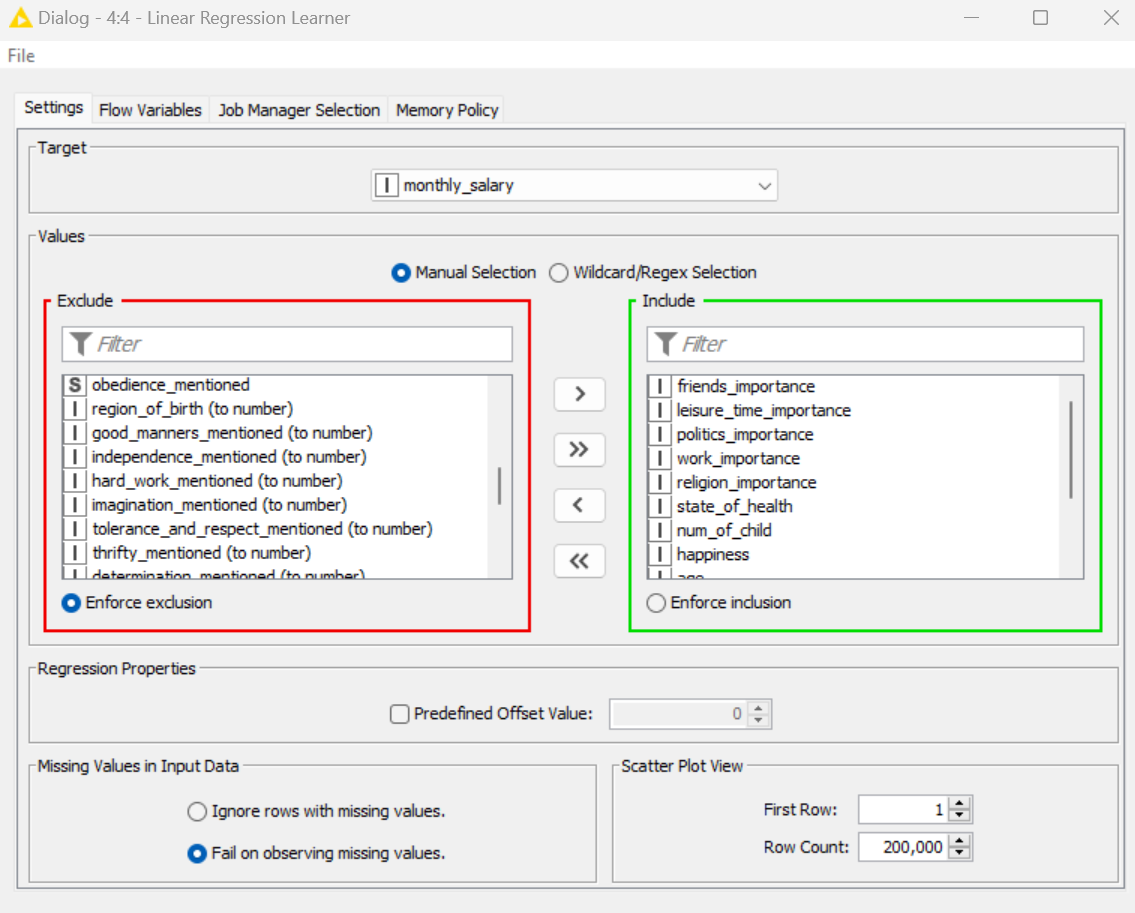
In knime I linear regressed the data.

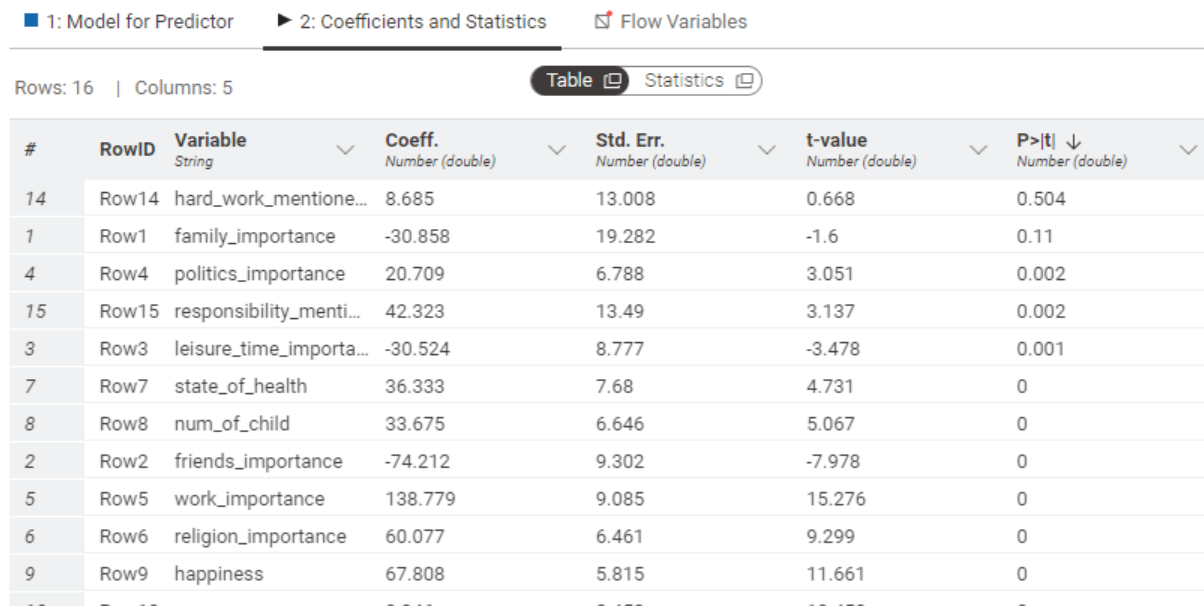
Training set (70%) is for training the regression model. Testing set (the remaining 30%) is an “unseen” set of data to check model performance.

**Step 2:** Using partioning node to split the dataset into 2 partitions: training set and testing set.

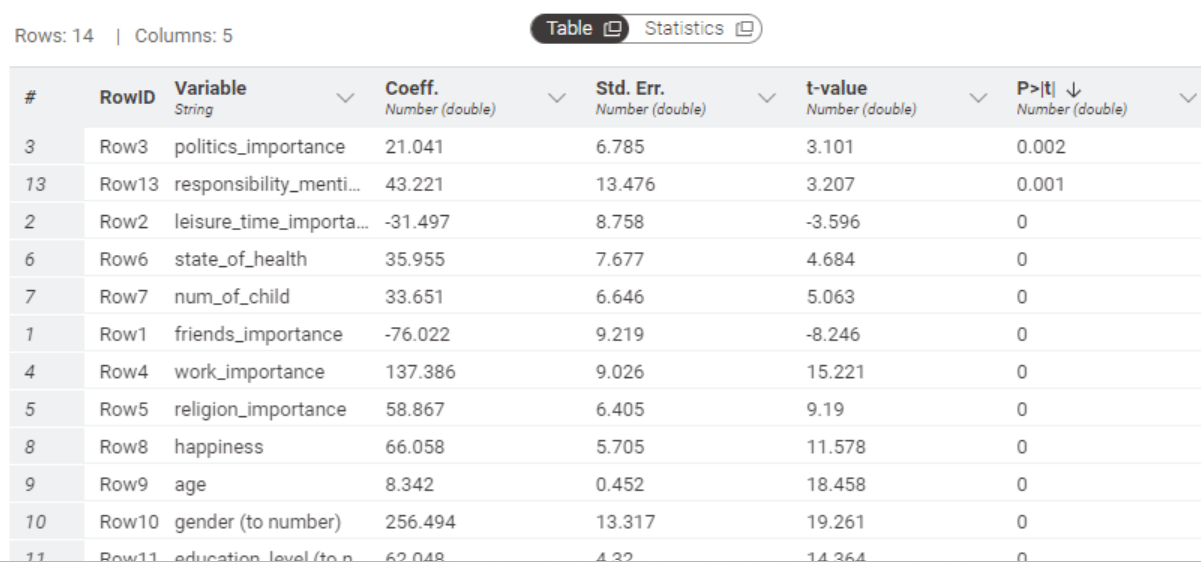
**Step 1:** Linear regression model can only take in numeric inputs. Hence the need to convert categorical fields to numeric fields using category to number node.

****

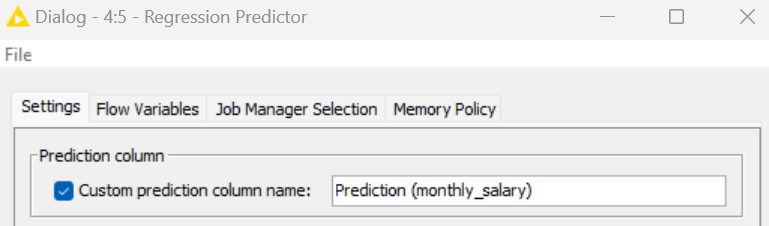
**Step 3:** To learn from the inputs and target of the training set.Then produce a linear regression model by writing down the regression equation using the coefficents.

****

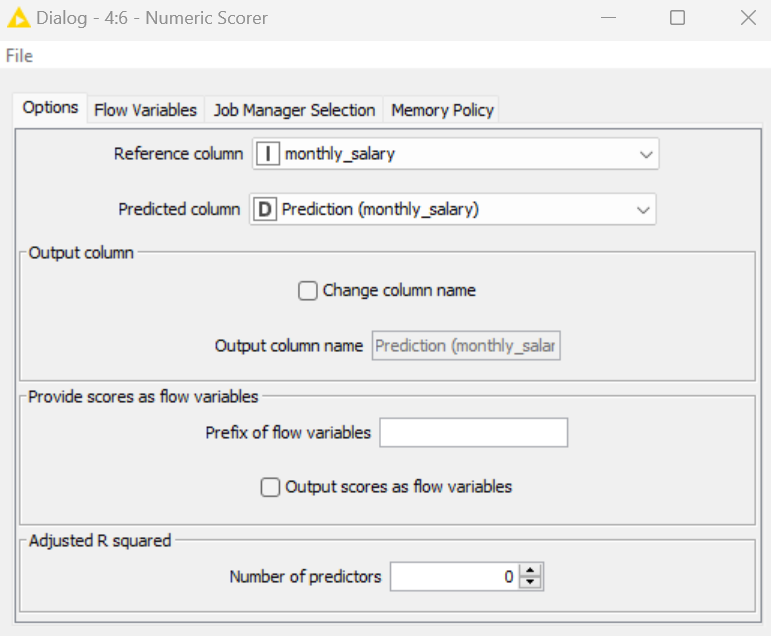
Removed all values with p-value greater than 0.05. This is because they are not good predictors and variables with p-value< 0.05 are important inputs. In this case , **hard work mentioned and family importance** are both greater than 0.05 hence I removed them.

****

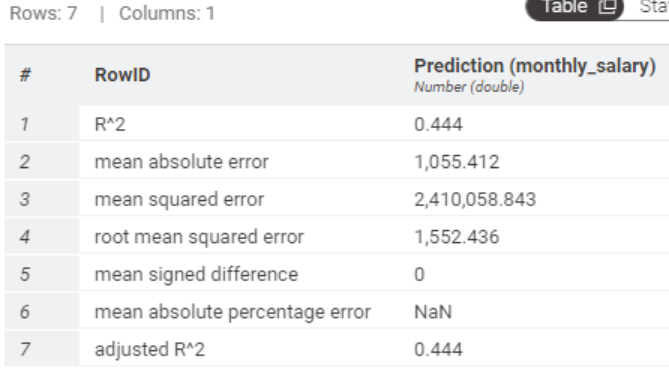
After redoing step 3 with variables only with p-value<0.05, produce a linear regression model. Use the coefficients then write down the regression equation.

****

Step 4: Using regression Predictor we can use the model produced in Learner node to predict target value for each row of data for the training set.

****

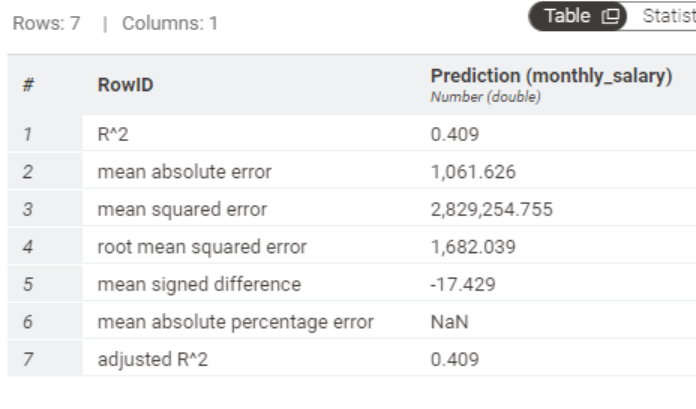
Step 5: Using numeric scorer it will compares predicted with actual target. Measures the prediction errors for the training set.

****

\*A regression model that can produce good prediction should have high R^2, low mean absolute error, Low mean square error, Low root mean squared error and Low mean signed error\*

Training set values

**Similarly do the same with the testing set using another regression predictor and numeric scorer node.**

****

The decrease in r square value from 0.444 to 0.409 when moving from training to testing data indicates that the model's predictive performance is less robust on new data compared to the data it was trained on. This drop suggests that the model might be overfitting or that there is room for improvement in the data quality. This is because the testing set is never seen before set of data and the testing set is the determining factor of how well the model will perform more than the training set.

Testing set values

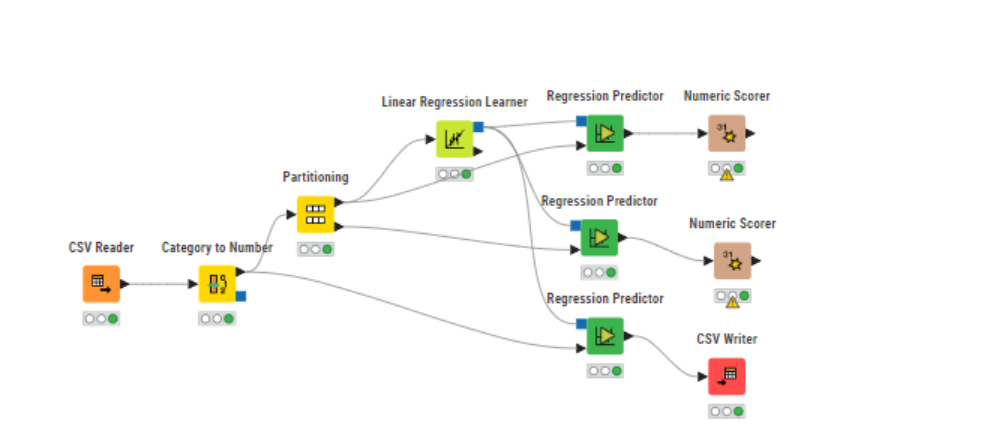
**Which are the important features that can determine the predicted values for monthly salary?**

|  |
| --- |
| friends\_importance |
| leisure\_time\_importance |
| politics\_importance |
| work\_importance |
| religion\_importance |
| state\_of\_health |
| num\_of\_child |
| happiness |
| age |
| gender |
| education\_level |
| employment |
| responsibility\_mentioned |
|  |

**The regression equation is**

Predicted monthly salary = y = friends\_importance x -76.02185358413321 + leisure\_time\_importance x -31.497129940661033 + politics\_importance x 21.041138500655276 + work\_importance x 137.38631442997146 + religion\_importance x 58.86716811153762 + state\_of\_health x 35.95543615672434 + num\_of\_child x 33.65126285518895 + happiness x 66.05802231241744

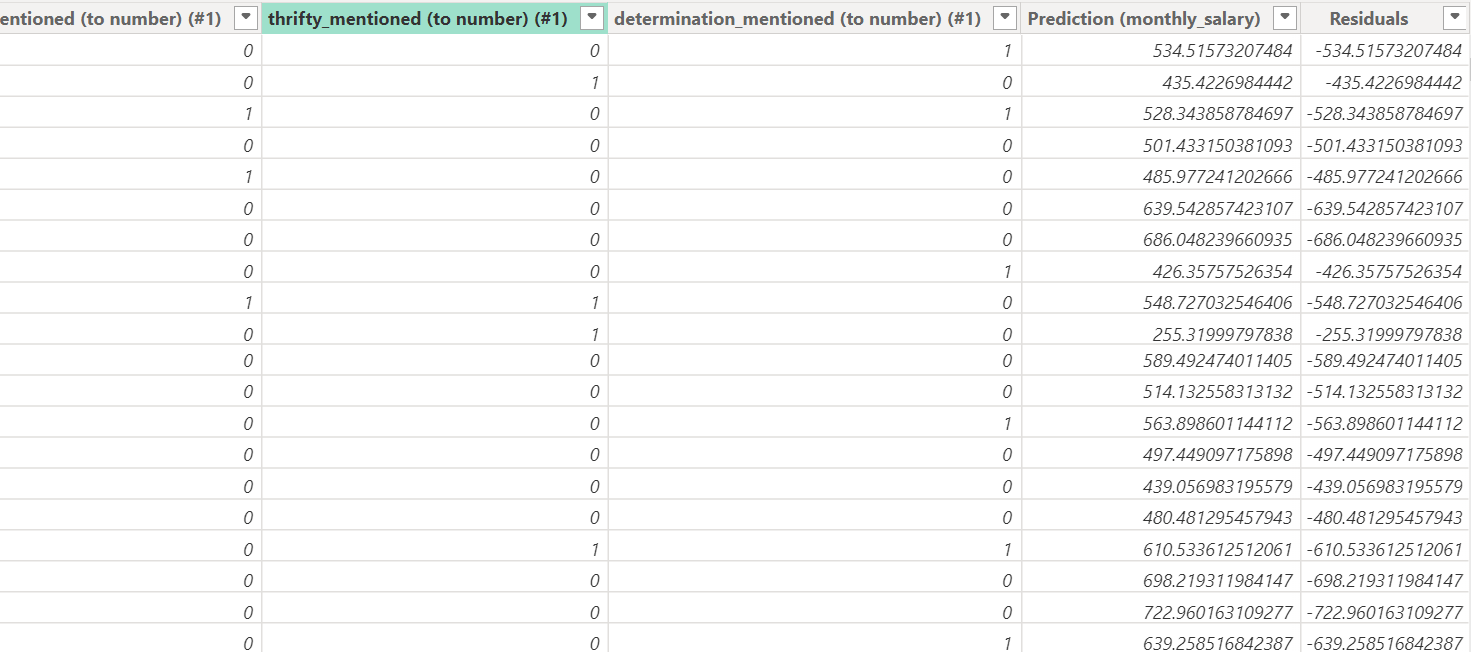
+ age x 8.341730568616994+ gender x 256.4943460482255 + education\_level x 62.04841159432335 + employment x -614.7999181629497 + responsibility\_mentioned x 43.220545382612436 + 1817.8687911882669

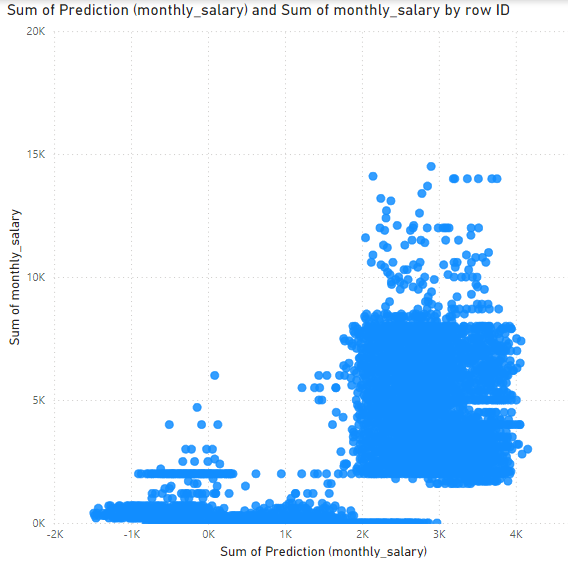
****

Final work flow

3(b). Findings / Explanation from Power BI:

****

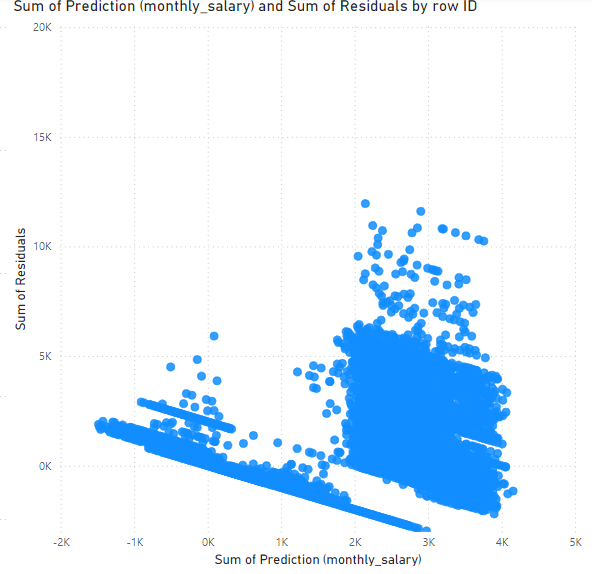
****

****

A perfect model would predict the price with 100% accuracy, this means, a salary of $1000 should have predicted salary of $1000.

A plot for the perfect model would have all the points forming a perfect diagonal line in your graph.

Since the Data points are clustering near the x-axis it suggests that this pattern is **underfitting / underestimated.**

****

A perfect model would predict the price with 100% accuracy, this means, residuals would be 0 for all monthly salary.

A residual plot for the perfect model would have all the points forming a perfect horizontal line on 0 on the Y-axis.

Since there are data points over and under the 0 line this means **the data is not accurate**.

In conclusion based on the analysis, the model is **not a good fit** for the data. The r square values of 0.444 for the training set and 0.409 for the testing set indicate that the model explains only a moderate portion of the variance in the target variable, which means that a significant amount of the variance remains unexplained. Additionally, the clustering of predicted values near the x-axis and the observed residuals above and below zero reveal that the model consistently underestimates actual salaries and displays systematic prediction errors. These issues suggest that the model is underfitting, meaning it is too simplistic to capture the underlying patterns in the data effectively.

|  |
| --- |
| SUBMISSION INSTRUCTIONS   * Save this file as “DA\_[Your Class]\_[Your Full Name]\_[Your Student ID]”, e.g. “***DA \_P01\_ALBERT EINSTEIN\_2409999A***”. * Submit your **report**, together with your **KNIME (.knwf)** and **Power BI (.pbix)** files, to DAVA LMS site > Assessment > Submission Link - DA Assignment. |

\*\*\*\*\* END OF DATA ANALYTICS ASSIGNMENT \*\*\*\*\*