Table of Contents

Task 1A Initial Data Exploration

A. Identify Attribute Types

In my given file, there are 29 feature columns where two columns are Unnamed and 27 columns are identified.

Attribute	Type	Measured Value
Unnamed	Interval	66770,28206,138297
Unnamed	Interval	57514,68503,50721
RA_ICRS	Ratio(Dimensionless)	321.24,9.4853507
DE_ICRS	Ratio(Dimensionless)	50.307,29.23
Source	Ratio(Dimensionless)	675544277730196480,188222190225
Plx	Ratio(Dimensionless)	0.5714,1.7775,1.0924
PM	Ratio(Dimensionless)	5.511,34.036,1.287
pmRA	Ratio(Dimensionless)	-8.788,-0.966,-0.926
pmDE	Ratio(Dimensionless)	-4.64,-8.077,-6.172
Gmag	Ratio(Dimensionless)	14.618119,10.567277
e_Gmag	Ratio(Dimensionless)	0.002768,0.00279,0.002776
BPmag	Ratio(Dimensionless)	15.047198,10.696631
e_BPmag	Ratio(Dimensionless)	0.003104,0.002946
RPmag	Ratio(Dimensionless)	14.009335,13.941367
e_RPmag	Ratio(Dimensionless)	0.003853,0.003817
GRVSmag	Ratio(missing)	
e_GRVSmag	Ratio(missing)	
BP-RP	Ratio(Dimensionless)	1.037864,0.555986
BP-G	Ratio(Dimensionless)	0.429079,0.479731
G-RP	Ratio(Dimensionless)	0.243632,0.778166
pscol	Ratio(missing)	
Teff	Ratio(Dimensionless)	0.778166,12224.2,10033.3
Dist	Ratio(Dimensionless)	1896.6731,956.4874
Rad	Ratio(Dimensionless)	2.1159,3.2847
Lum-Flame	Ratio(Dimensionless)	10.30679,43.061054
Mass-Flame	Ratio(Dimensionless)	43.061054,2.163

Age-Flame	Ratio(Dimensionless)	0.331,0.252
z-Flame	Ratio(Dimensionless)	
SpType-ELS	Categorical	A,B

Justifications for Choices

The given data provides information on different qualities, their respective categories, and the corresponding measured values. Allow me to provide a concise elucidation:

1. Unnamed: There are two attributes labeled as Unnamed, which appear to be IDs or indices. The values allocated to them are measured on an interval scale, such as 66770, 28206, 138297, and so on. the recorded values of an attribute are not valid.

2. RA_ICRS, DE_ICRS, Source: These are the coordinates and source of the data. The properties in question pertain to celestial coordinates, specifically Right Ascension and Declination, as well as a source identifier. They are measured on a ratio scale with values such as 321.24, 9.4853507, 675544277730196480, etc.

3. Plx, PM, pmRA, pmDE, Gmag, e_Gmag, BPmag, e_BPmag, RPmag, e_RPmag, BP-RP, BP-G, G-RP, Teff, Dist, Rad: These properties indicate numerous astronomical measures such as parallax, proper motion, magnitude, effective temperature, distance, radius, etc. They are also measured on a ratio scale with values such as 0.5714, 1.7775, 5.511, 34.036, etc.

4. GRVSmag, e_GRVSmag, pscol, z-Flame: These properties have missing values and their distributions are not stated.

5. Lum-Flame, Mass-Flame, Age-Flame: These qualities indicate properties of heavenly objects such as luminosity, mass, and age. They are also quantified on a ratio scale with numbers such as 10.30679, 43.061054, 0.331, 0.252, etc.

6. SpType-ELS: This characteristic represents spectral types of celestial objects and is categorical with values A and B.

In essence, the dataset covers several astronomical features assessed on different scales, including ratio scales and categorical scales. Some properties have missing values that need to be treated appropriately.

Handling Missing Values

In This data set two Attribute are Unnamed which is not import so that remove two attribute this dataset

Attribute	Missing Value
GRVSmag	1302
e_GRVSmag	1302
pscol	2912
Lum-Flame	52
Mass-Flame	221
Age-Flame	767
z-Flame	52

The dataset has multiple columns with considerable missing values. Specifically, the columns "GRVSmag", "e_GRvSmag", and "Pascol" have a substantial number of missing data. Therefore, these columns will be eliminated from the dataset.

Additionally, the "e_Gmag" and "e_BPmag" characteristics contain identical values, indicating redundancy. Therefore, the "e_BPmag" column will be eliminated as well.

Furthermore, the columns "Lum-Flame", "Mass-Flame", "Age-Flame", and "z-Flame" feature missing data. The missing values in these columns will be filled by using the rounded mean.

To summarize, the dataset will be processed as follows:

1. Remove columns "GRVSmag", "e_GRvSmag", "Pascol", and "e_BPmag" owing to large missing values or redundancy.

2. Fill missing values in the columns "Lum-Flame", "Mass-Flame", "Age-Flame", and "z-Flame" by using the rounded mean. Rounding the mean is a basic yet effective method for filling missing values in a dataset. It finds a balance between conserving the original data distribution and delivering interpretable imputed values, making it a regularly used technique in data preprocessing operations, including those implemented in KNIME.

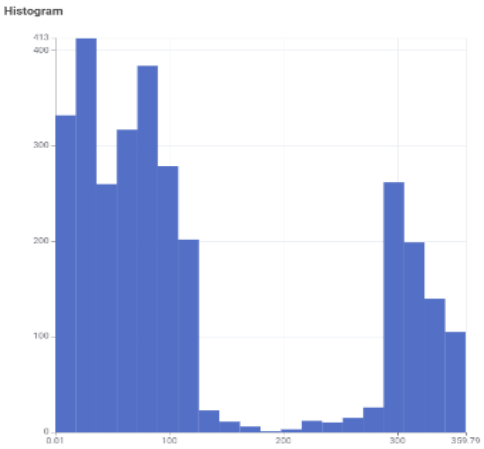
B. Statistical Summaries

Using the statistics node in KNIME, this summary table was generated.

Name	Minimum	Maximum	25% Quantile	50% Quantile (Me...	75% Quantile	Mean	Mean Absolute D...	Standard Deviation	Skewness	Kurtosis
RA_ICRS	0.012	359.791	36.518	78.275	249.716	124.946	96.991	115.443	0.912	-0.795
DE_ICRS	-29.526	86.732	32.672	47.45	59.655	41.277	18.656	23.913	-1.081	0.265
Source	35,922,973,320,48...	3,131,197,086,259...	413,475,595,360,7...	516,481,658,574,4...	2,035,279,968,188...	1,181,451,830,770...	932,628,872,535,4...	1,020,174,395,625...	0.688	-1.068
Plx	-1.912	25.979	0.3	0.607	1.218	0.938	0.686	1.138	6.097	89.851
PM	0.052	76.555	1.39	2.713	5.366	4.307	3.177	5.209	4.404	32.755
pmRA	-37.938	75.324	-2.122	-0.644	0.621	-0.574	2.548	4.632	1.837	38.068
pmDE	-55.704	43.588	-3.341	-1.189	-0.195	-2.012	2.649	4.456	-1.978	30.257
Gmag	4.911	17.643	11.541	12.901	14.744	13.039	1.834	2.17	-0.205	-0.384
e_Gmag	0.003	0.012	0.003	0.003	0.003	0.003	0	0	14.154	241.38
BPmag	4.898	18.613	11.691	13.208	15.147	13.322	1.945	2.296	-0.184	-0.459
RPmag	4.933	17.689	11.238	12.462	14.146	12.605	1.706	2.034	-0.188	-0.247
e_RPmag	0.004	0.044	0.004	0.004	0.004	0.004	0.001	0.002	7.813	88.022
GRVSmag	4.932	14.09	10.541	11.346	11.898	11.305	1.081	1.464	-0.476	0.828
e_GRVSmag	0.004	0.337	0.01	0.016	0.035	0.033	0.028	0.041	2.771	9.707
BP-RP	-0.37	2.979	0.412	0.642	0.97	0.717	0.337	0.421	0.662	0.431
BP-G	-0.429	1.73	0.143	0.238	0.384	0.283	0.153	0.198	1.146	2.216
G-RP	-0.267	1.249	0.268	0.404	0.59	0.434	0.186	0.228	0.297	-0.231
pscol	1.3	1.769	1.546	1.627	1.671	1.605	0.071	0.09	-0.926	0.753
Teff	5,394.3	30,016.6	7,798.325	9,423.7	10,443.925	9,641.139	1,694.691	2,344.835	1.71	5.082
Dist	39.056	24,289.97	817.746	1,661.013	3,318.369	2,301.889	1,515.613	2,128.229	3.048	17.538
Rad	0.935	17.119	1.912	2.416	3.289	2.824	0.994	1.418	2.473	10.09
Lum-Flame	1.886	3,452.37	16.08	35.755	86.292	107.071	116.248	266.056	6.396	51.406
Mass-Flame	1.261	7.003	1.894	2.227	2.717	2.437	0.588	0.829	1.984	5.01
Age-Flame	0.2	3.198	0.373	0.77	1	0.713	0.303	0.337	0.201	-0.038
z-Flame	0.118	1.582	0.445	0.514	0.604	0.546	0.121	0.172	1.587	4.441

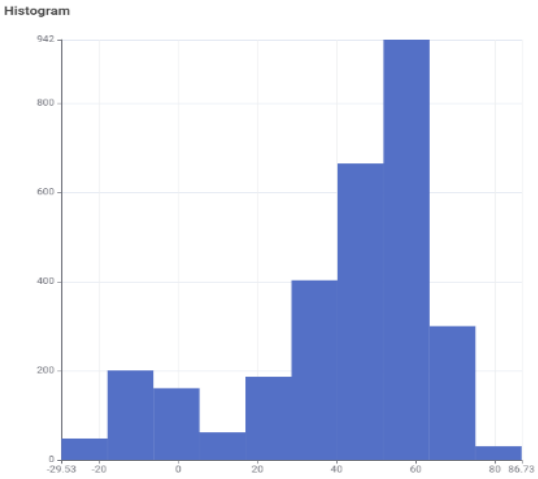
Identify Attribute Frequency With Histogram

Draw Histogram Plot for RA_ICRS Using KNIME



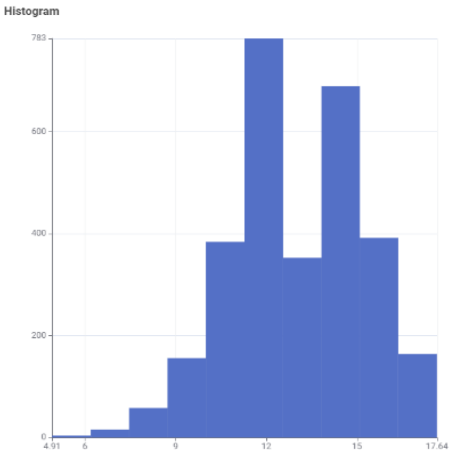
There is a considerable reduction in frequency, indicating that there are very few data points in that range. The bars show varying frequencies of data points.

Draw Histogram Plot for DE_ICRS Using KNIME



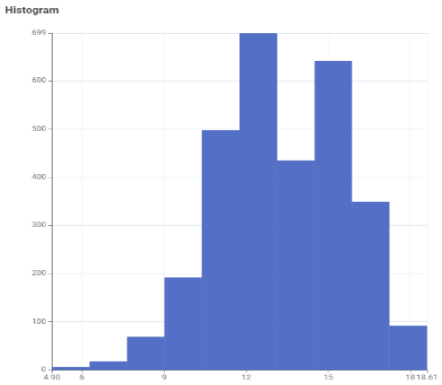
The distribution looks to be skewed, with fewer data points on the left side (negative values) and a more prolonged tail on the right side (positive values).

Draw Histogram Plot for Gmag Using KNIME



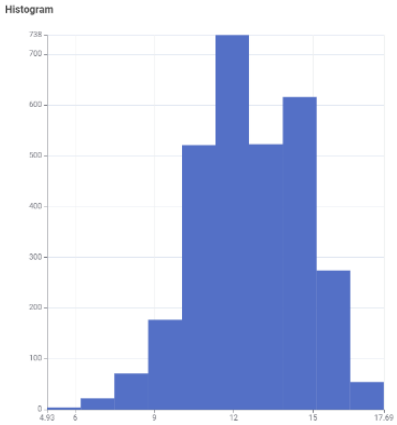
The distribution appears to be skewed, with fewer data points on the left side (lower Gmag values) and a more stretched tail on the right side (higher Gmag values).

Draw Histogram Plot for BPmag Using KNIME



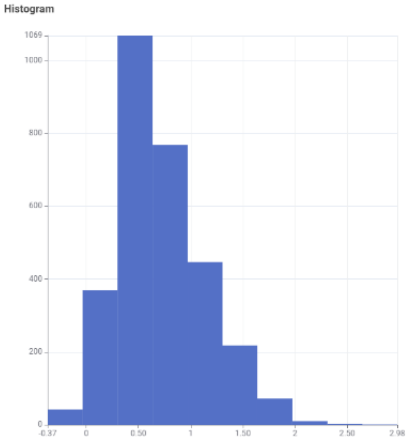
The distribution appears to be skewed, with fewer data points on the left side (lower BPmag values) and a more prolonged tail on the right side (higher BPmag values).

Draw Histogram Plot for RPmag Using KNIME



The distribution appears to be normal

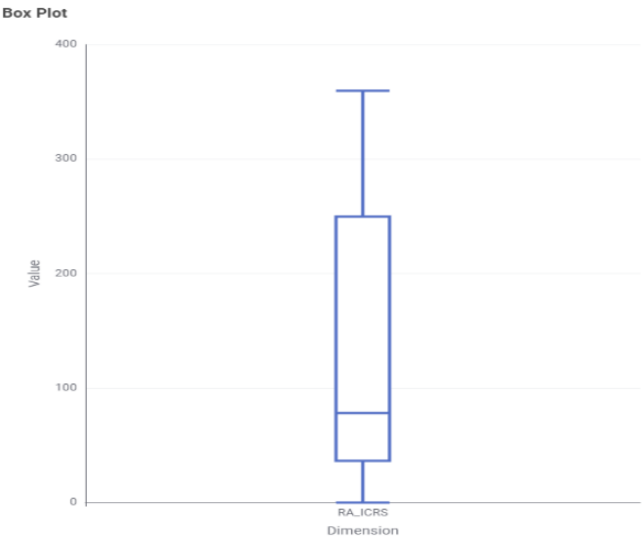
Draw Histogram Plot for BP-RP Using KNIME



The distribution appears to be Right skewed

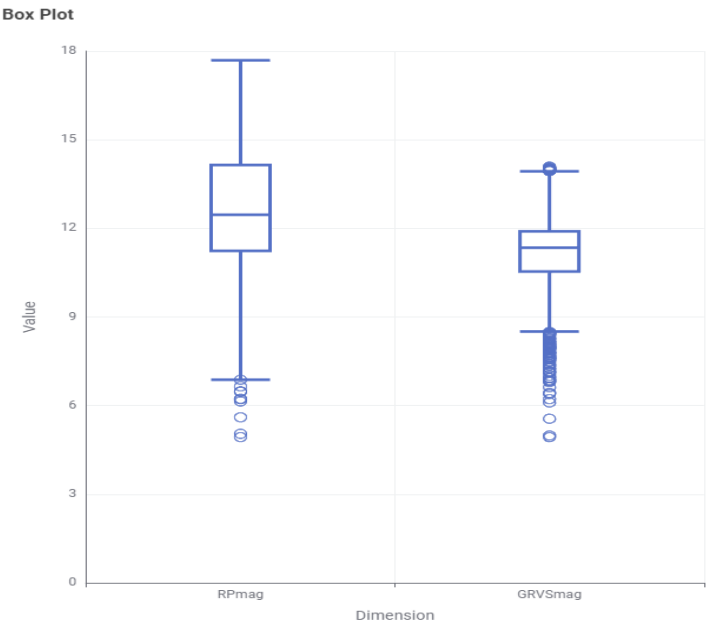
Identifying outliers with Boxplot

Box Plot for RA_ICRS



Inside the box, there's a horizontal line, which reflects the median value of RA_ICRS.An upper whisker extends from the top of the box to around 300 (showing variability beyond the highest quartile). No lower whisker is shown, implying that there are no data points below the lower quartile. There are no outliers plotted individually in this specific box plot.

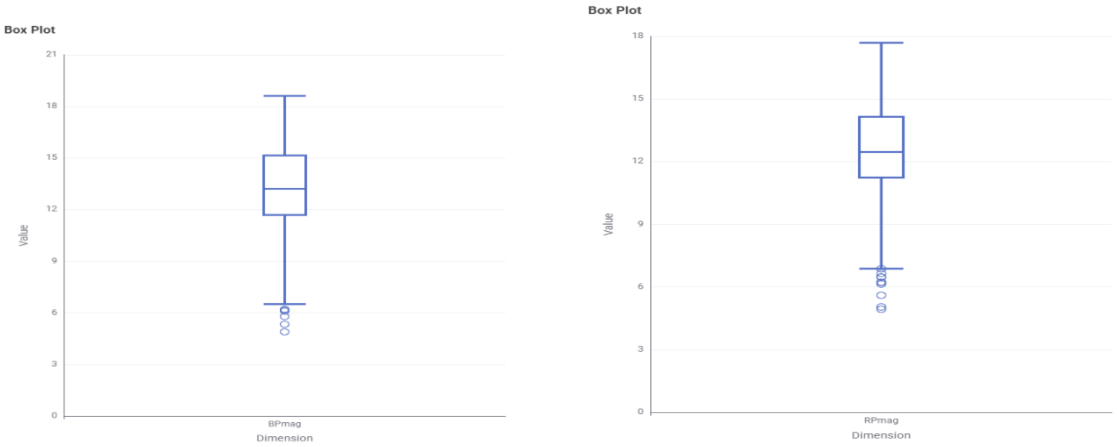
Box Plot for RPmag and GRVSmag



Inside the box, a horizontal line shows the median value of RPmag (about 12). Extending from the top of the box, the upper whisker reaches roughly 300, suggesting variability beyond the upper quartile.

Outliers: Several data points fall below the lower whisker, signifying values that are more than 1.5 times the IQR below the first quartile. The box plot for GRVSmag also has a median value around 12. The IQR for GRVSmag is narrower, extending from about 11 to 13. Outliers are present both below and above the whiskers.

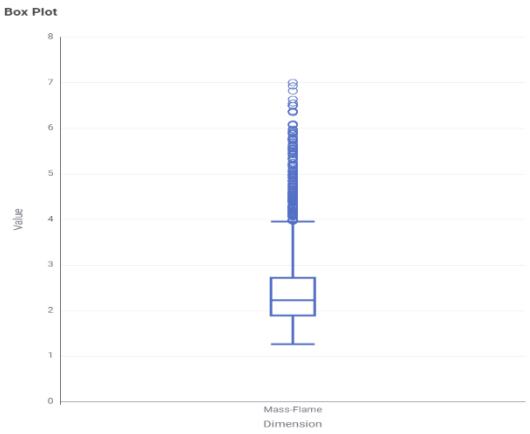
Box Plot For Bpmag and RPmag



The bottom edge of the Bpmag box plot corresponds to a number about 12, indicating the first quartile (Q1). The upper border of the box corresponds to a number about 15, denoting the third quartile (Q3). This range (from 12 to 15) shows the interquartile range (IQR) for BPmag.Inside the box, there's a horizontal line, which reflects the median value of BPmag (about 14). The upper whisker extends from the top of the box up to approximately 18, demonstrating variability beyond the upper quartile. Below the lower whisker, there are Four outliers having values about between 6 and 8.

The bottom border of the RPmag box corresponds to a number about 12, representing the first quartile (Q1). The upper border of the box corresponds to a number about 13, denoting the third quartile (Q3). This range (from 12 to 15) is the interquartile range (IQR) for RPmag. Inside the box, there's a horizontal line, which reflects the median value of RPmag (about 14). The upper whisker extends from the top of the box up to approximately 18, demonstrating variability beyond the upper quartile. Below the lower whisker, there are outliers having values about between 6 and 8.

Box Plot For Mass-Flame

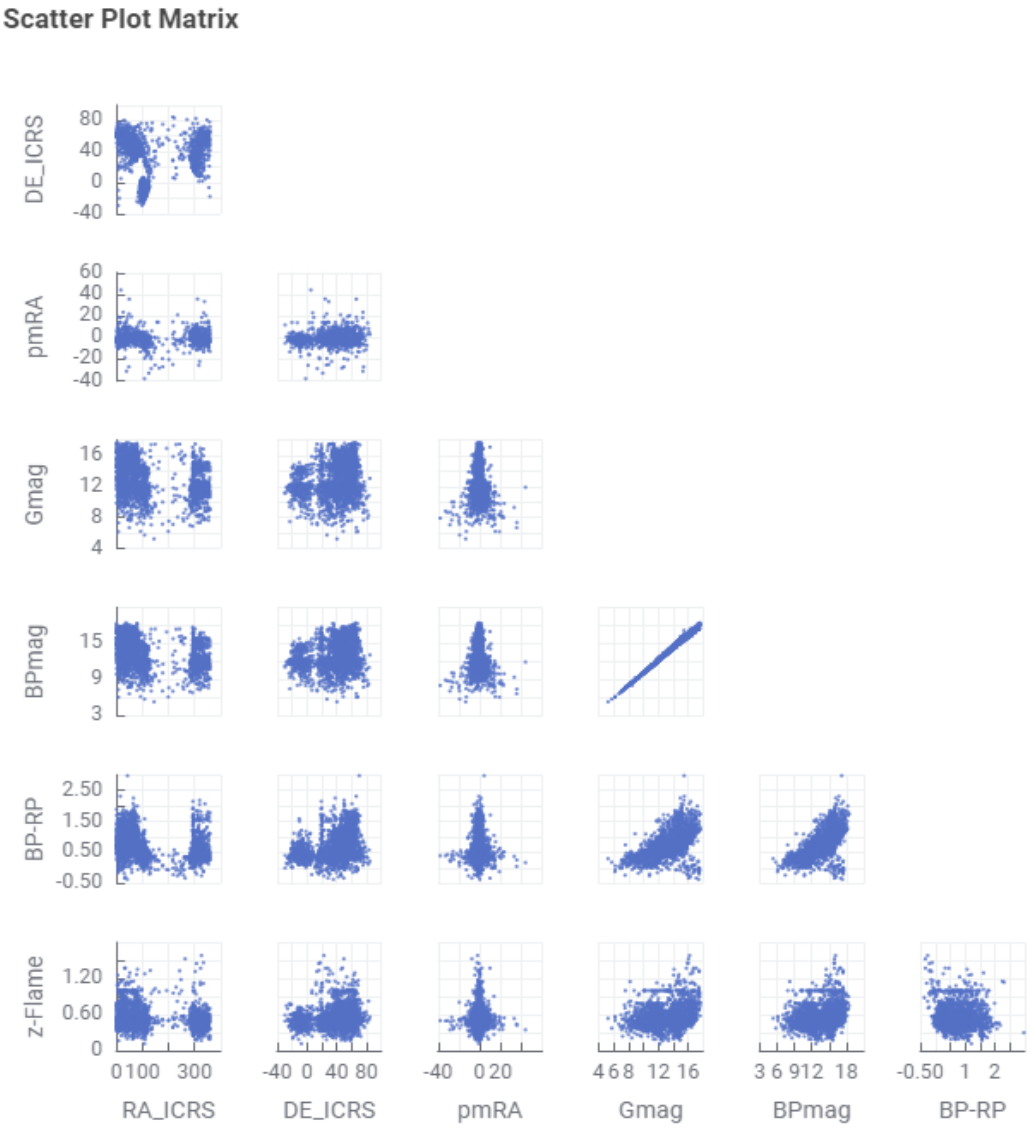


The bottom edge of the Mass_flam box plot corresponds to a number about 8, indicating the first quartile (Q1). The upper border of the box corresponds to a number about 4, denoting the third quartile (Q3). This range (from 4 to 1) shows the interquartile range (IQR) for BPmag.Inside the box, there's a horizontal line, which reflects the median value of Mass_flame

(about 2.5). The upper whisker extends from the top of the box up to approximately 4, demonstrating variability beyond the upper quartile. Below the lower whisker, there are outliers having values about between 4 to 7.

Scatter Plot Visualization

Visualize pairwise relationships between several numerical attributes with Scatter Plot Matrix with Columns DE_ICRS, pmRA, Gmag, BPmag, z-Flame, RA_ICRS



Notice a linear relationship between "BPmag" and "Gmag" in a scatter plot, it shows that changes in one variable are related with proportional changes in the other variable, revealing insights into their interdependence.

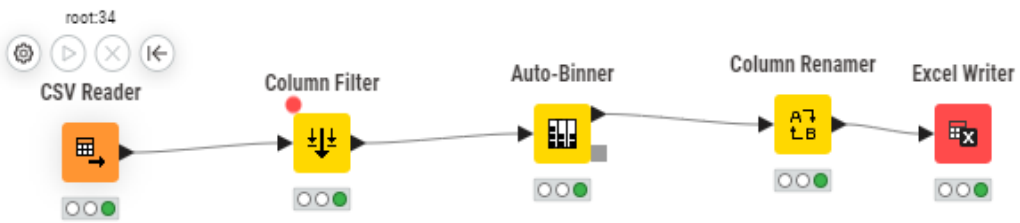
Task-1B: Data Preprocessing

A1. Binning For RA_ICRS

Steps that were followed for all the binning tasks:

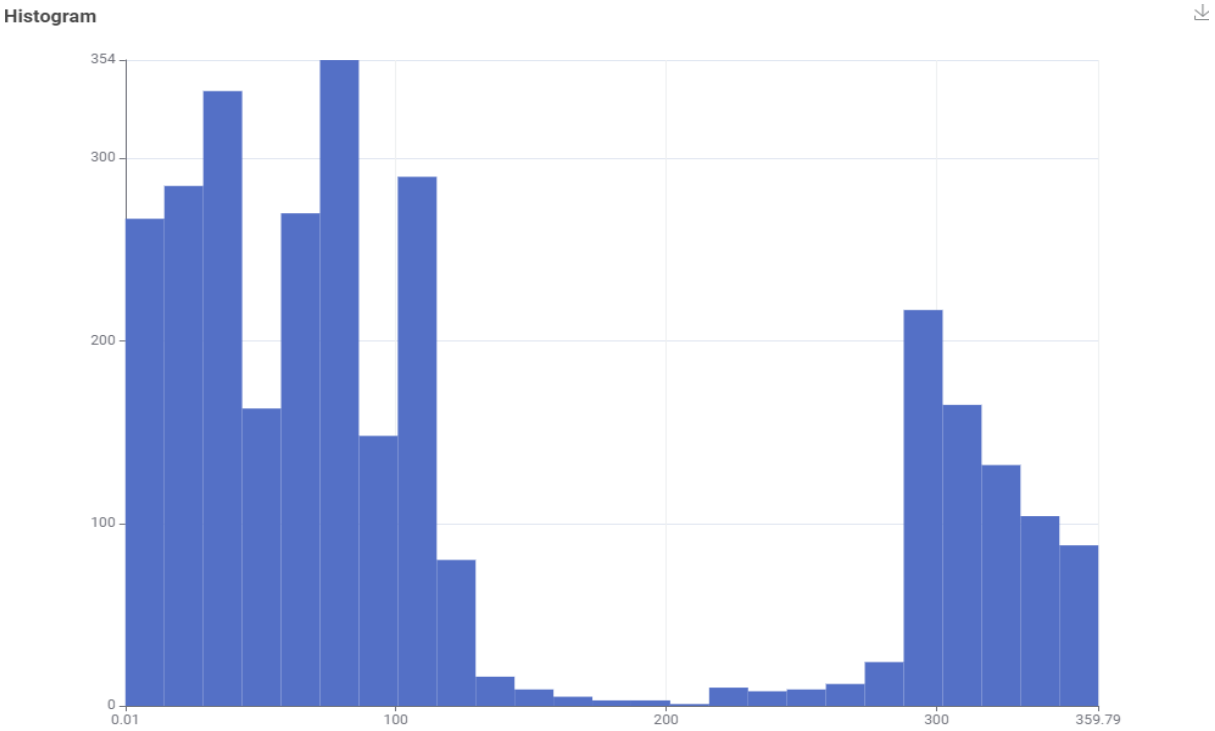
1. Load the CSV file
2. Using Column Filter, only select the relevant column
3. Using Auto Binner, first use Equi Width Binning option with number of bins 25.
4. Using Auto Binner, then use Equi Frequency Binning option with number of bins 15
5. After step 3 or 4,using column renamer, rename the column to our desired one.
6. Finally, write the output in a excel file.
7. As for histogram, it is generated after column filter, and in histogram, we can select Equi Width Binning or Equi Frequency Binning and Number of Bins for each round.

Equi Width Binning



Number of Bins: 25

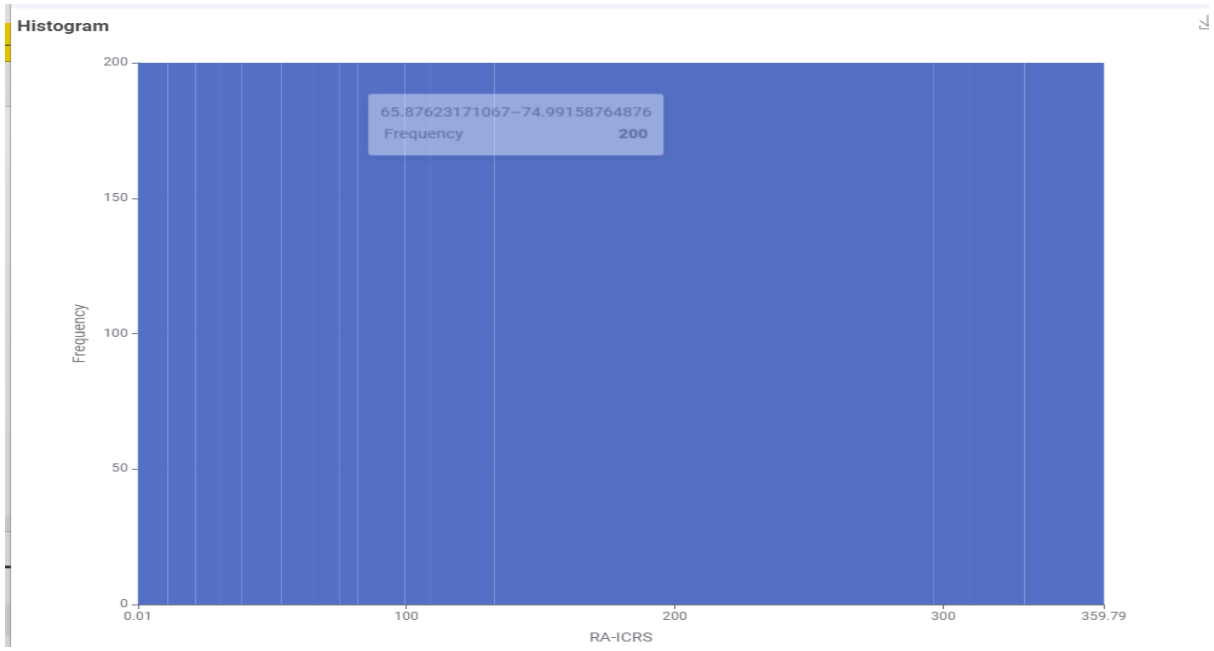
Justification: This number was chosen based on data distribution. Histogram was used for visualization.



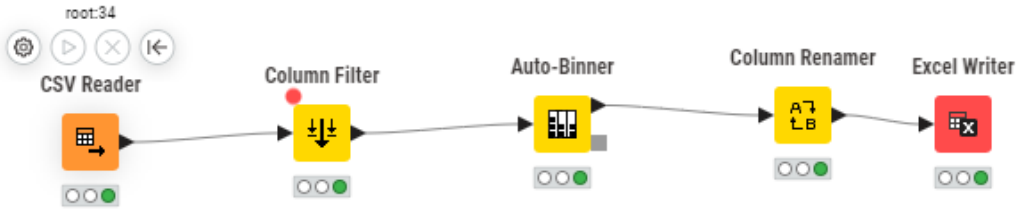
Equi Depth Binning

Number of Bins: 15

Justification: To distribute the data points in a way that every bin has the same number of instances.



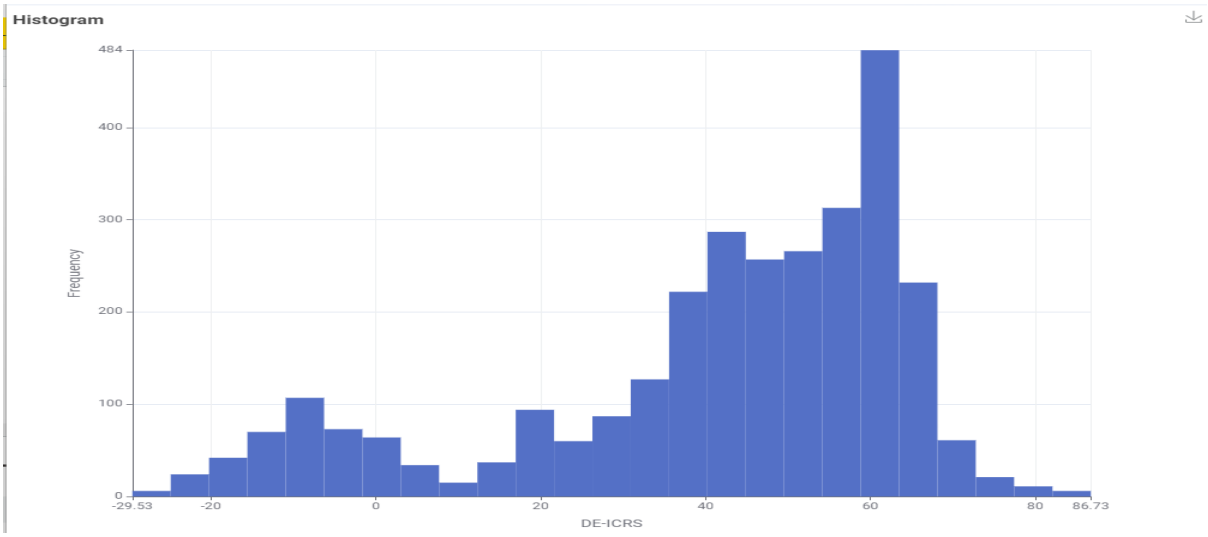
A2. Binning For DE_ICRS



Equi Width Binning

Number of Bins: 25

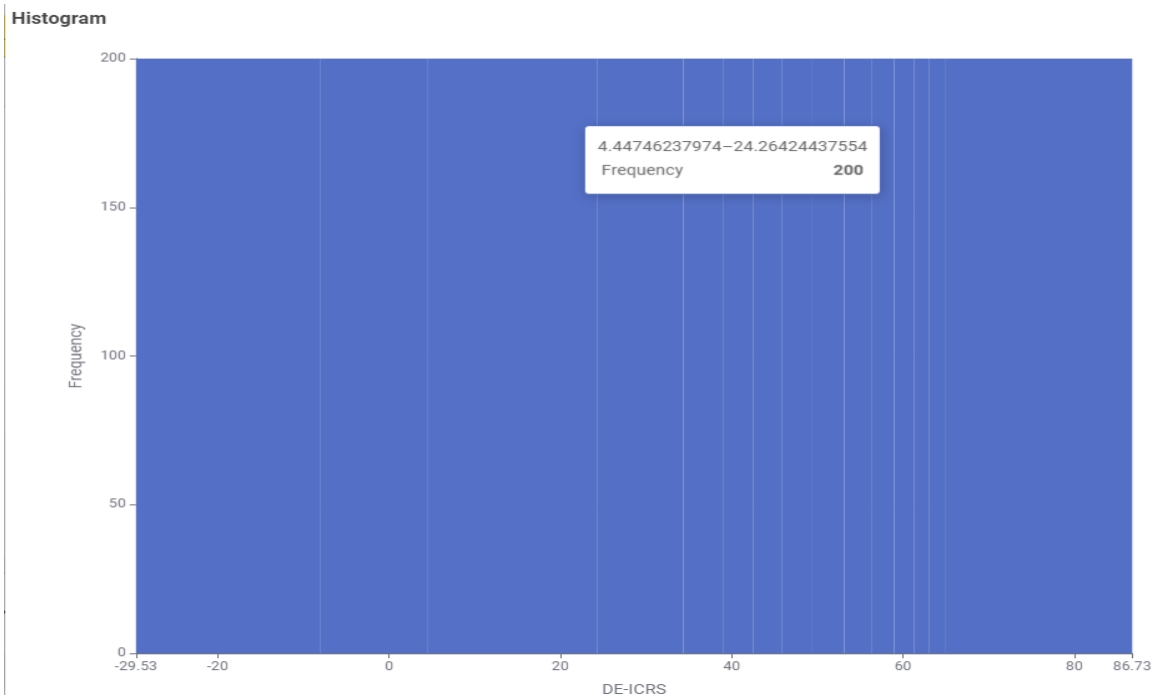
Justification: This number was chosen based on data distribution. Histogram was used for visualization.



Equi Depth Binning

Number of Bins: 15

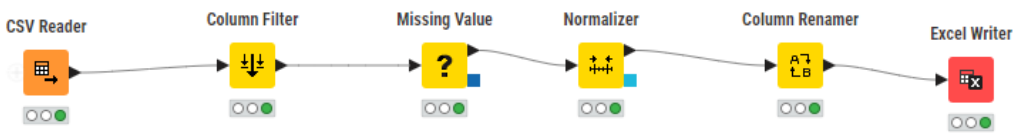
Justification: To distribute the data points in a way that every bin has the same number of instances.



B. Age-Flame Normalization

Steps:

1. Load the CSV file
2. Using Column Filter, only select the relevant column
3. Handling Missing Values, using Rounded Mean Imputation.
4. Using Normalizer node, first select Min-Max option with values [0.0-1.0] and go to step 6
5. Using Normalizer node, select z-score option and go to step 6
6. After step 4 or 5,using column renamer, rename the column to our desired one.
7. Finally, write the output in a excel file.



Min-Max Normalization

	A	B
1	Age-Flame	Age-Flame-Normalised-Min-Max
2	1.071	0.290527018
3	0.593	0.131087392
4	1	0.266844563
5	1	0.266844563
6	1.007	0.269179453
7	0.727	0.175783856
8	0.331	0.043695797
9	0.275	0.025016678
10	0.271	0.023682455

Z-Score Normalization

	A	B
1	Age-Flame	Age-Flame-Normalised-Z-score
2	1.071	1.061709265
3	0.593	-0.355579324
4	1	0.851191504
5	1	0.851191504
6	1.007	0.871946776
7	0.727	0.041735887
8	0.331	-1.132419512
9	0.275	-1.29846169
10	0.271	-1.310321846

C. Mass-Flame Categorization

Steps:

1. Load the CSV file
2. Using Column Filter, only select the relevant column
3. Handling Missing Values, using Rounded Mean Imputation.
4. Using Rule Engine Node, give the following code in its code dialog box -

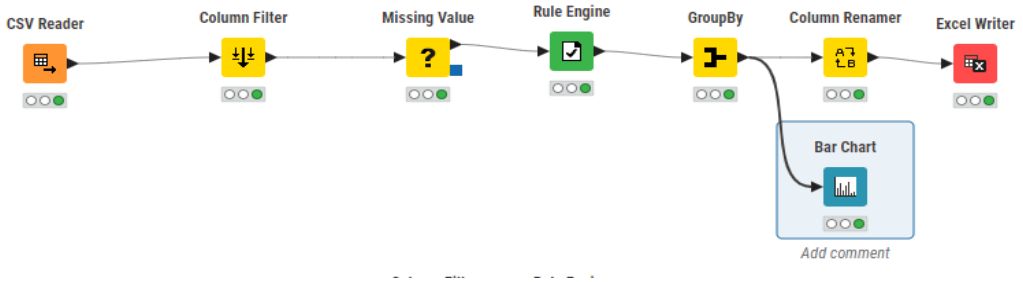
   $Mass-Flame$ <= 2 => "Small"

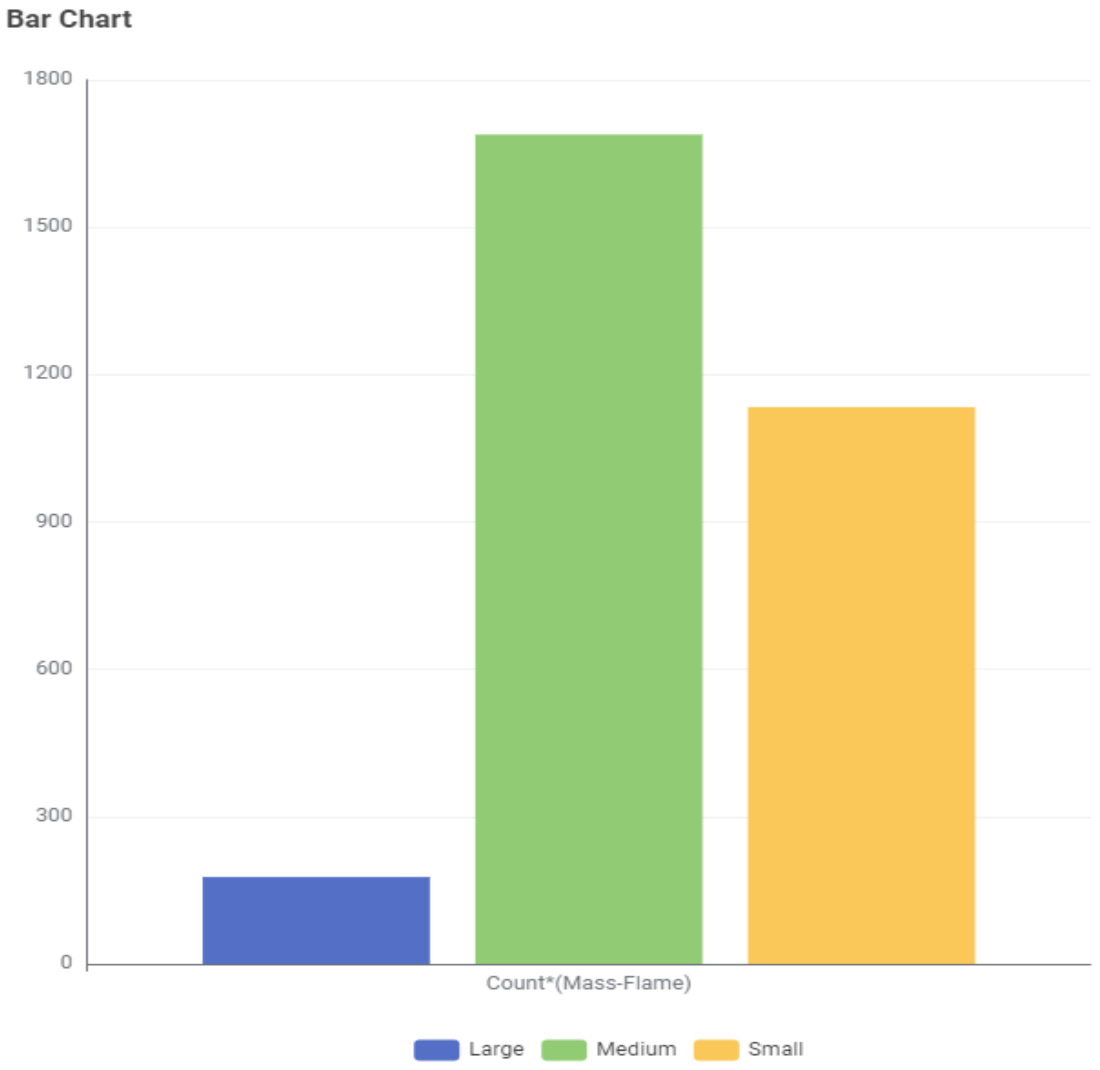
   $Mass-Flame$ > 2 AND $Mass-Flame$ <= 4 => "Medium"

$Mass-Flame$ > 4 AND $Mass-Flame$ <= 8 => "Large"

TRUE => "Out of Range"

5. Then, using GroupBy Node to count the frequencies of each group
6. After step 4 or 5,using column renamer, rename the column to our desired one.
7. Finally, write the output in a excel file.
8. As for Bar Chart, it is generated after the group by Node.





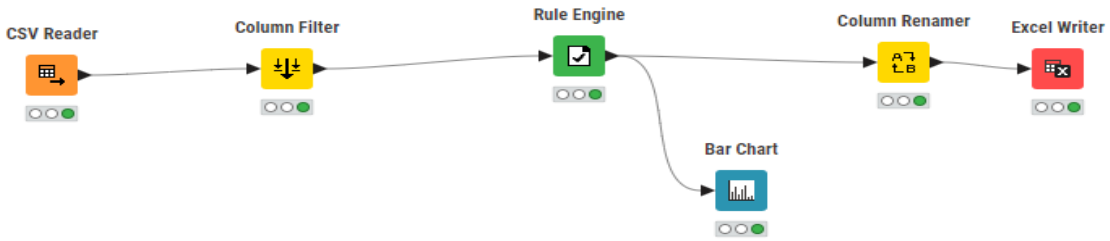
D. SP-Type-ELS Binary Transform

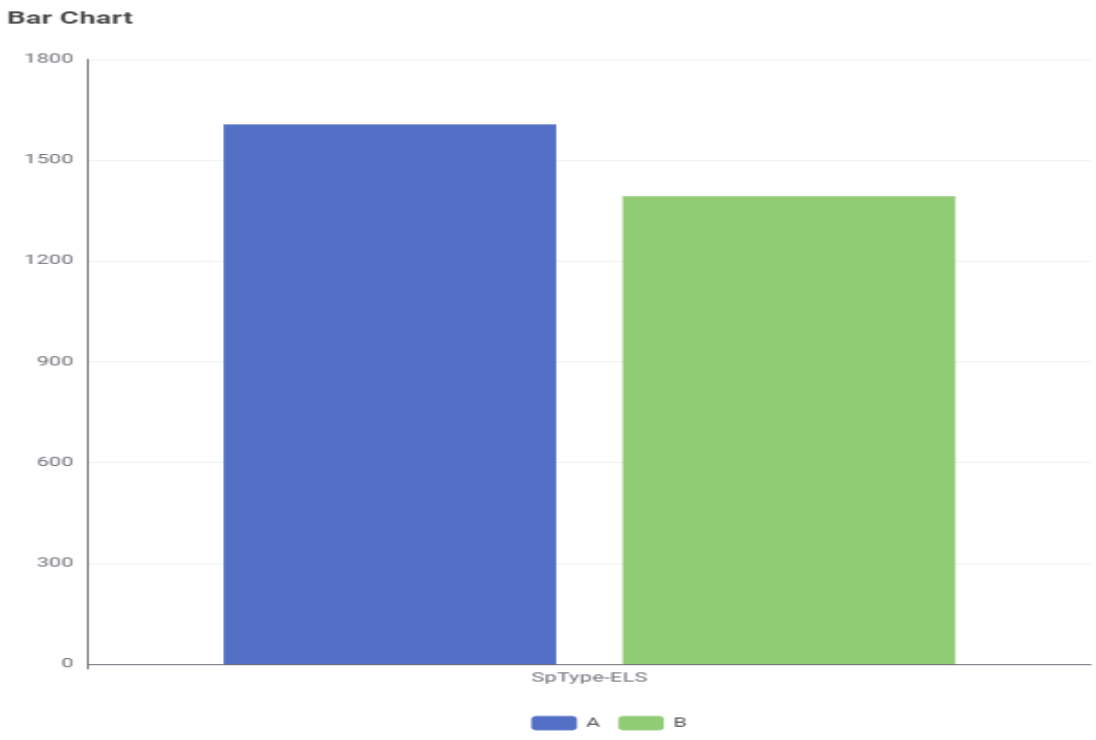
Steps:

1. Load the CSV file
2. Using Column Filter, only select the relevant column
3. Handling Missing Values, using Rounded Mean Imputation.
4. Using Rule Engine Node, give the following code in its code dialog box -

   $SpType-ELS$ = "A" => 1

   TRUE => 0
5. After step 4 or 5,using column renamer, rename the column to our desired one.
6. Finally, write the output in a excel file.
7. As for Bar Chart, it is generated after the Rule Engine Node.





1C. Executive Summary

Data Exploration and Preprocessing

- Conducted a comprehensive initial data exploration, identifying 27 key features for astronomical analysis, with special focus on celestial coordinates and identifiers, as well as vital astronomical measures such as parallax, proper motion, and magnitudes.
- Addressed missing values decisively, eliminating columns with significant data gaps and employing mean imputation for others, ensuring a robust dataset for subsequent analysis.

Binning Techniques

- Applied Equi-Width Binning to RA_ICRS and DE_ICRS attributes to capture the distribution across 25 bins, enhancing the visibility of data frequencies and distributions.
- Implemented Equi-Depth Binning for the same attributes with 15 bins, ensuring even distribution of data points and a balanced representation across bins.

Normalization

- Performed Min-Max Normalization on the Age-Flame attribute, scaling the data to a fixed range of [0.0-1.0], which standardizes the scale for model compatibility.
- Executed Z-Score Normalization to re-scale Age-Flame data to a mean of 0 and a standard deviation of 1, facilitating comparisons across different scales and mitigating the impact of outliers.

Categorization and Transformation

- Discretized the Mass-Flame attribute into 'Small', 'Medium', and 'Large' categories using logical rule sets. This categorical breakdown paves the way for more nuanced group analysis.
- Converted the SpType-ELS attribute from categorical to a binary format using a rule-based approach, enhancing the dataset's suitability for binary classification tasks.

Visualization and Statistical Summaries

- Employed histograms and boxplots to visualize the frequency distributions and identify outliers in the dataset, uncovering underlying patterns and potential data entry errors.
- Utilized scatter plots to illuminate relationships between several numerical attributes, identifying a notably linear relationship between 'BPmag' and 'Gmag'.

Findings Summary

1. Among 29 features, two features has no descriptions and contained same data instances.
2. Only one Categorical Data, which is our Target Class. (SPType-ELS)
3. Three attributes has more than 1000 missing values those are dropped to do further analysis.
4. Lum-Flame, Mass-Flame, Age-Flame, z-Flame has smaller number of missing values which is handled using Rounded Mean Imputation.
5. Statistical Summary Highlights the min and max value of each attribute, along many other statistical tools such as Mean, Standard deviation, Skewness, Kurtosis.
6. Histograms identify the followings for the attributes:
   a) RA _ ICRS have a range from (130-245) where there are very few data points.
   b) The distribution of DE_ICRS is left skewed.
7. Boxplots identify the followings for the attributes:
   a) RA _ ICRS has no outliers.
   b) RPmag and GRVSmag has several outliers below the lower whisker.
8. Scatter Plot matrix identifies a positive linear relationship between "BPmag" and "Gmag".
9. Using different binning techniques for RA _ ICRS and DE_ICRS, helps to understand the diversity of the data and their frequency properly.
10. For Age-Flame Normalization, Min-Max option seems more suitable.
11. After categorizing Mass-Flame, it is found that majority of the data points belong to Medium Range, then comes Small range and then with less than 300 data points comes Large range group.
12. In out dataset, there are more type A data points for SP-Type-ELS feature than type B.