National University of Science and Technology

POLITEHNICA Bucharest

Faculty of Automation and Computers

Master Program

Advanced Analytics for Business

Visual Analytics Techniques

First semester, 2024-2025

Homework

Student Performance Factors

Teachers: Monica Drăgoicea

Fadi Hosni

Student: Merve Pakcan Tufenk

E-mail: merve.pakcan@stud.acs.upb.ro

Contents

1. Introduction

This project seeks to analyze and understand the factors that influence student performance, using the *Student Performance Factors* dataset from Kaggle. This study aims to explore how various factors, including socioeconomic background, school resources, and parental involvement, influence academic outcomes, as measured by Exam Scores. By analyzing variables such as Hours Studied, Sleep Hours, Extracurricular Activities, and Internet Access, the study seeks to identify patterns and correlations. By applying the SAS Visual Analytics methodology—Access, Investigate, Prepare, Analyze, and Report—this project will provide data-driven insights into the disparities in academic achievement and their underlying causes across diverse student demographics. The findings will guide strategies to improve educational equity and resilience within school systems.

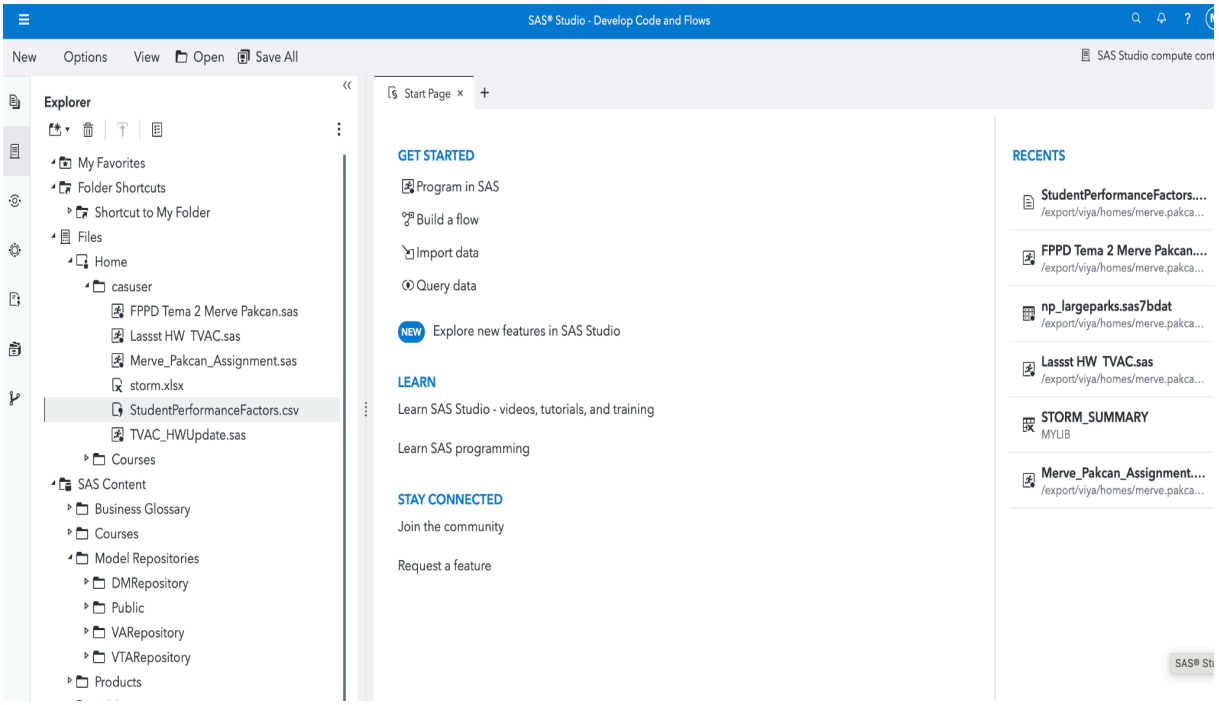
Aligned with the United Nations Sustainable Development Goals (SDGs), particularly SDG 4: Quality Education and SDG 10: Reduced Inequalities, this project aims to address issues of educational fairness and inclusivity.Specifically, it examines how disparities in socioeconomic status and educational resources shape academic resilience and performance outcomes. By highlighting the impact of these inequities, the project aspires to offer actionable insights for educators, policymakers, and communities, ultimately fostering more inclusive and supportive learning environments.As a result, promoting educational quality, inclusion, and fairness in and through education are crucial components of the SDG4 aim, and monitoring progress until 2030 is essential (Friedman et al., 2020).

The long-term goal is to support efforts that reduce performance gaps and promote equal access to quality education, creating a more resilient and equitable educational system.To guarantee that no one falls behind, all students, regardless of background or handicap status, require appropriate physical infrastructure and safe, inclusive learning settings (Rad et al., 2022).The insights generated can assist in developing targeted interventions to empower all students, especially those from underserved backgrounds, in reaching their academic potential. This project ultimately contributes to building a fairer, more sustainable educational landscape that supports all students in achieving success and resilience in their learning journeys.

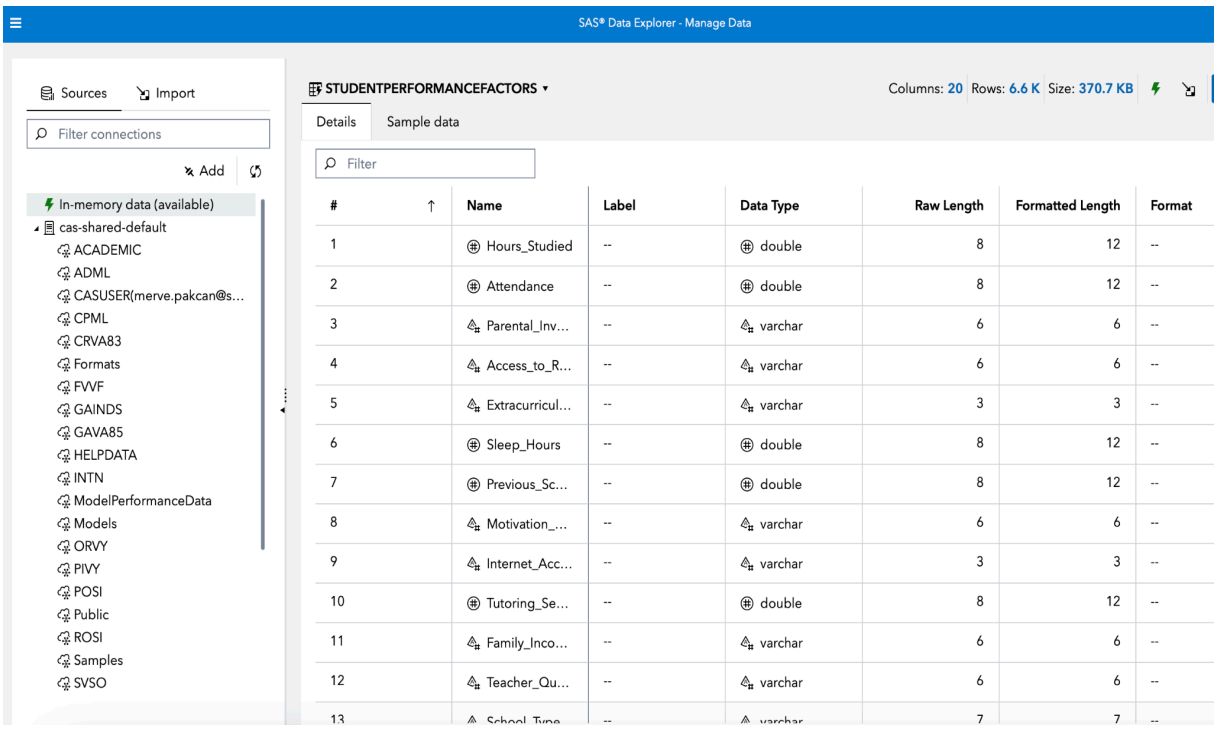
The analysis of factors influencing student performance can be framed within the User–Data–Task Triangle model. The Users in this context include educators, policymakers, and other stakeholders who aim to enhance academic outcomes and equity in education.Poor academic performance in students has been a subject of concern to many people, including parents, administrators, educators, psychologists, and counsellors (Oldayo & Fakai, 2020). The Data comprises a comprehensive dataset of student performance metrics. The Task focuses on uncovering patterns, relationships, and key drivers of academic success to inform effective interventions. By aligning user needs, data exploration, and analytical objectives, this project strives to generate actionable insights that foster equitable and data-driven improvements in the education system.

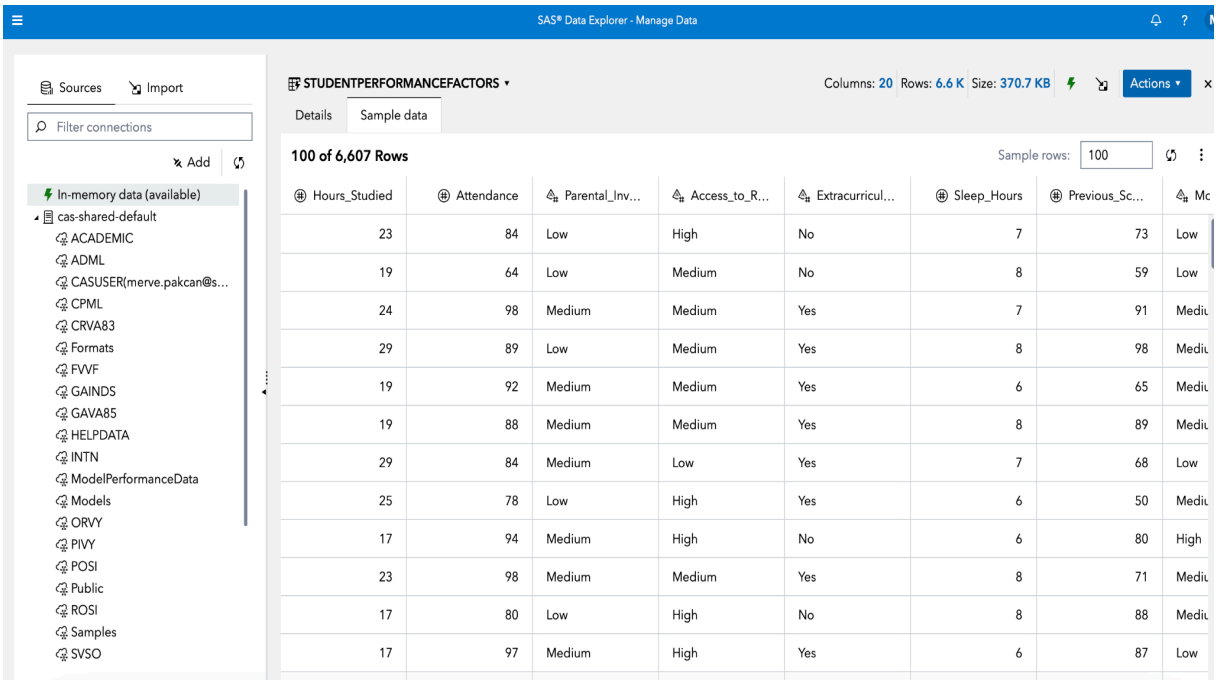
2. Data Sourcing & Acquisition

I uploaded the *StudentPerformanceFactors.csv* file into the CAS using *SAS Studio - Develop Code and Flows* menu.



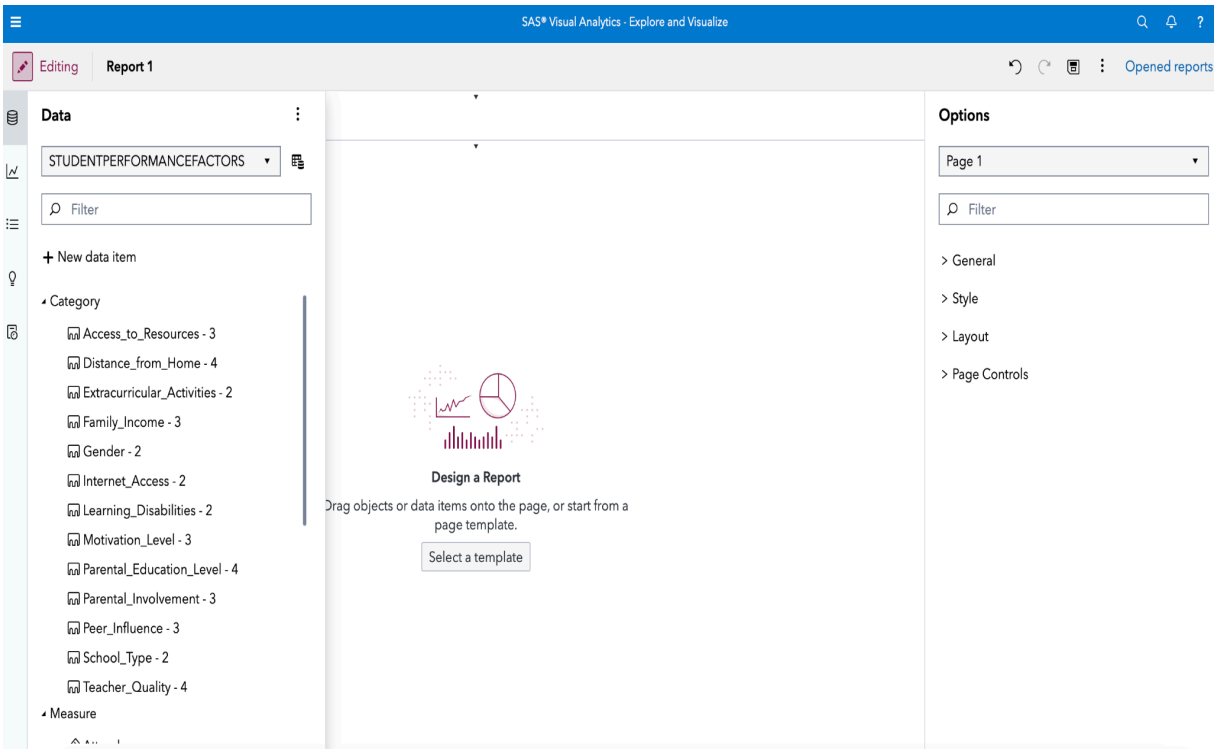
I uploaded the StudentPerformanceFactors.csv file into the CAS (Cloud Analytics Services) environment, within the CASUSER library, using the Manage Data option in SAS Data Explorer. The successful in-memory load, indicated by the green 'In-memory data' status, confirms that the file is ready for analysis and visualization in SAS Visual Analytics.

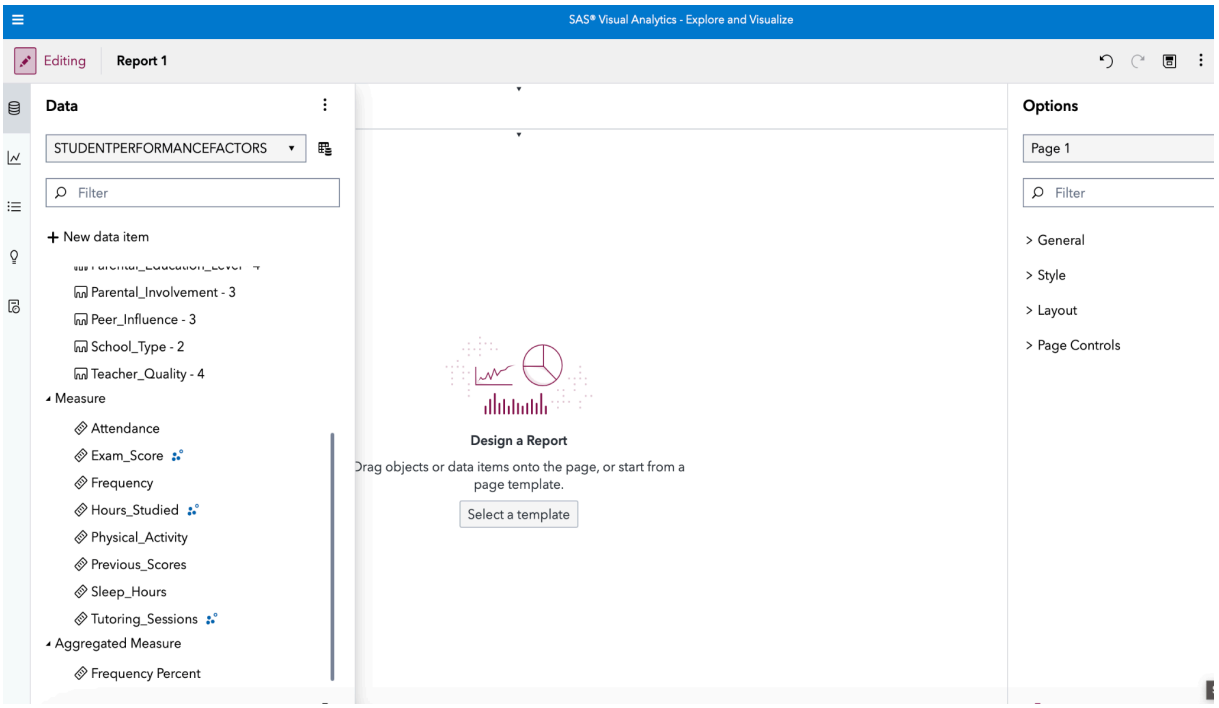




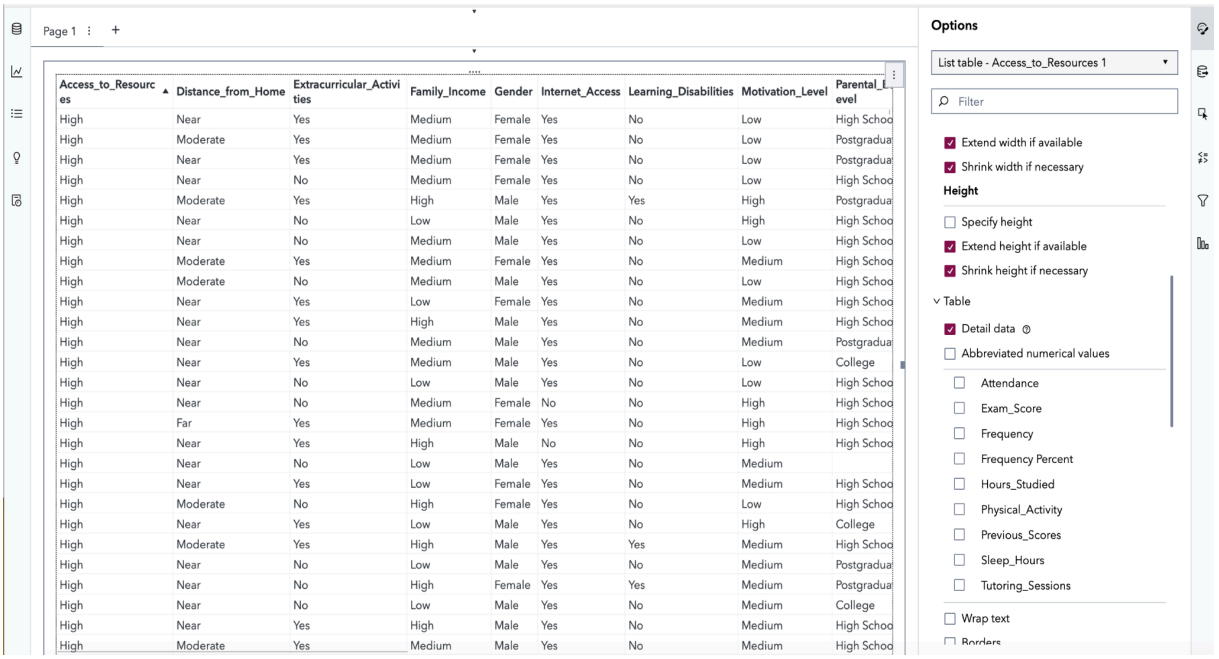
The dataset comprises 6,600 rows and 20 columns, with a total size of 370.7 KB. Out of these 20 columns, 7 are numerical (DataType: double) and 13 are character data (DataType: varchar), offering a robust data structure for analysis.

3. Data Wrangling



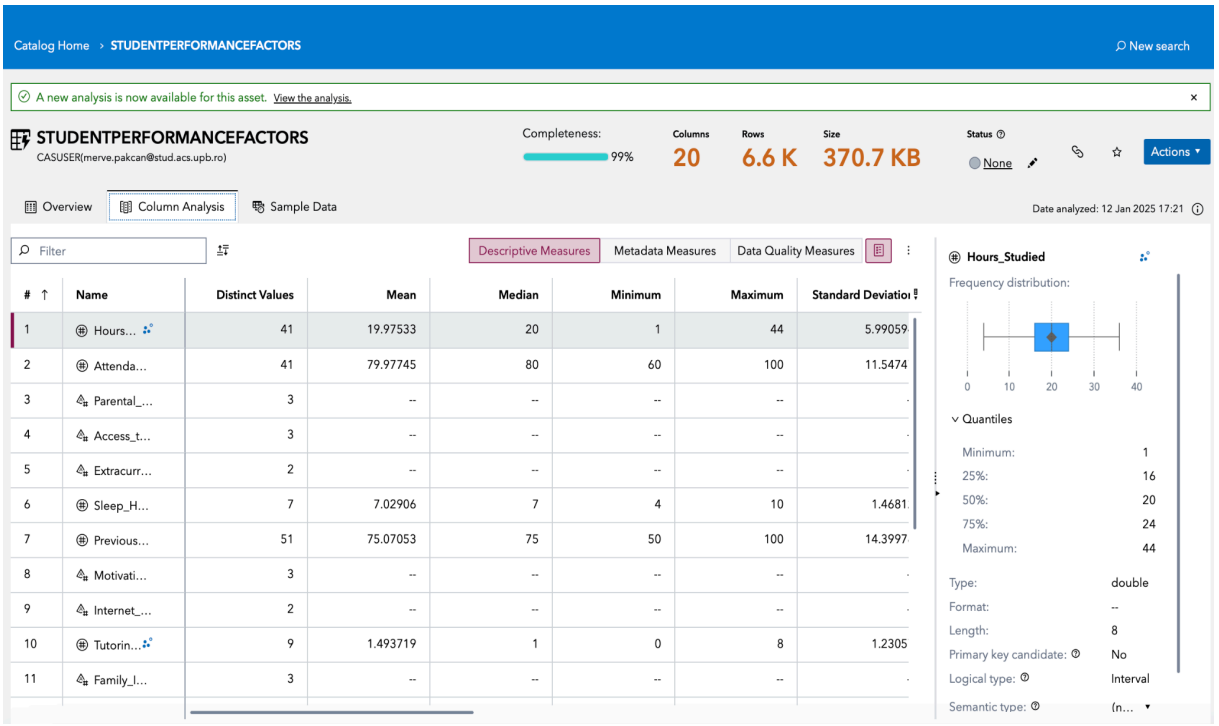


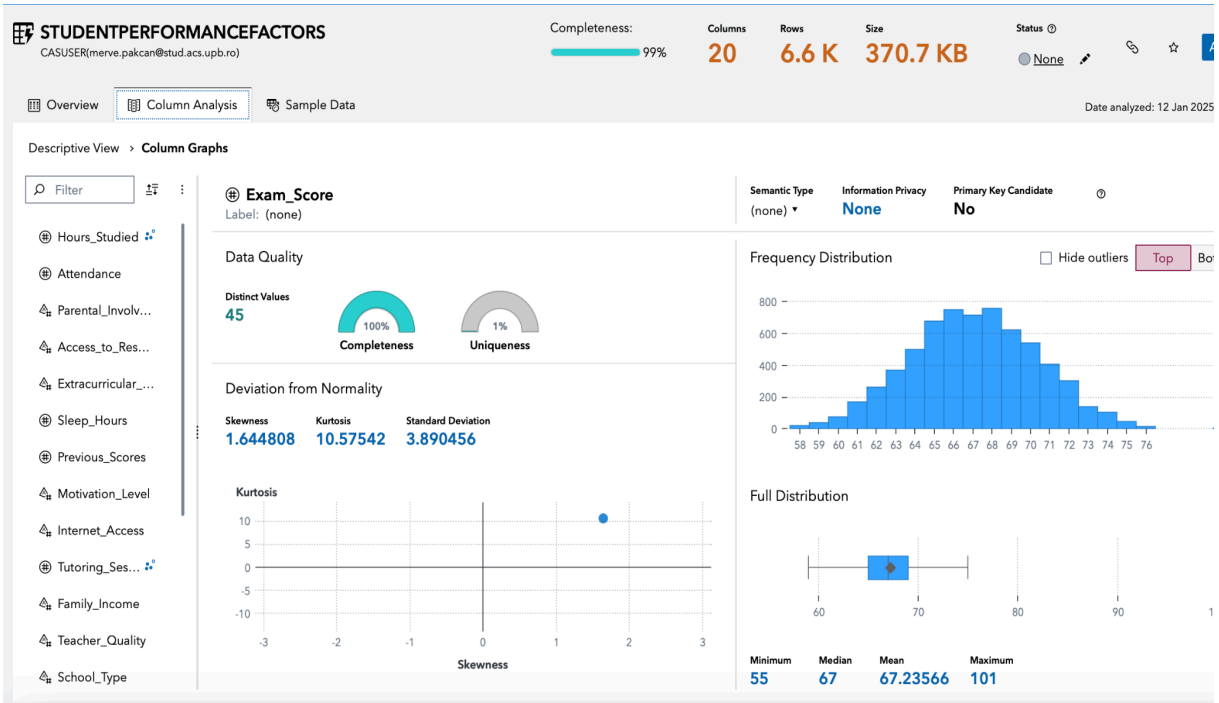
I utilized the Explore and Visualize menu in SAS Visual Analytics to load the dataset as an in-memory table, making it accessible for analysis within the report.The dataset consists of 13 categorical variables (categories), 7 numerical variables (measurements), and a single aggregated measure.

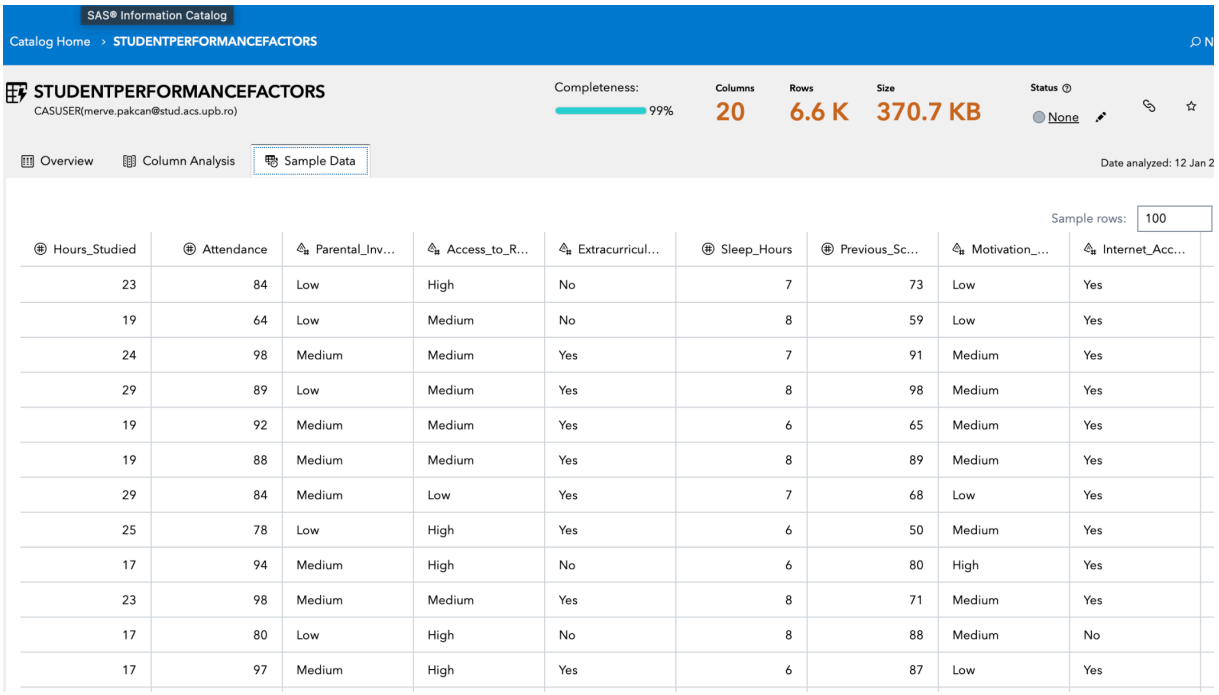


Access_to_Resources	Distance_from_Home	Extracurricular_Activities	Family_Income	Gender	Internet_Access	Learning_Disabilities	Motivation_Level	Parental_Level
High	Near	Yes	Medium	Female	Yes	No	Low	High Schoo
High	Moderate	Yes	Medium	Female	Yes	No	Low	Postgradua
High	Near	Yes	Medium	Female	Yes	No	Low	Postgradua
High	Near	No	Medium	Female	Yes	No	Low	High Schoo
High	Moderate	Yes	High	Male	Yes	Yes	High	Postgradua
High	Near	No	Low	Male	Yes	No	High	High Schoo
High	Near	No	Medium	Male	Yes	No	Low	High Schoo
High	Moderate	Yes	Medium	Female	Yes	No	Medium	High Schoo
High	Moderate	No	Medium	Male	Yes	No	Low	High Schoo
High	Near	Yes	Low	Female	Yes	No	Medium	High Schoo
High	Near	Yes	High	Male	Yes	No	Medium	High Schoo
High	Near	No	Medium	Male	Yes	No	Medium	Postgradua
High	Near	Yes	Medium	Male	Yes	No	Low	College
High	Near	No	Low	Male	Yes	No	Low	High Schoo
High	Near	No	Medium	Female	No	No	High	High Schoo
High	Far	Yes	Medium	Female	Yes	No	High	High Schoo
High	Near	Yes	High	Male	No	No	High	High Schoo
High	Near	No	Low	Male	Yes	No	Medium	
High	Near	Yes	Low	Female	Yes	No	Medium	High Schoo
High	Moderate	No	High	Female	Yes	No	Low	High Schoo
High	Near	Yes	Low	Male	Yes	No	High	College
High	Moderate	Yes	High	Male	Yes	Yes	Medium	High Schoo
High	Near	No	Low	Male	Yes	No	Medium	Postgradua
High	Near	No	High	Female	Yes	Yes	Medium	Postgradua
High	Near	No	Low	Male	Yes	No	Medium	College
High	Near	Yes	High	Male	Yes	No	Medium	High Schoo
High	Moderate	Yes	Medium	Male	Yes	No	Medium	High Schoo

I created a *List Table* in SAS Visual Analytics, allowing me to view a summary of the dataset as well as the detailed values for each row and column. This provided a clear overview of the data structure.After reviewing the data, it was concluded that **no new data items** are needed, and the existing data is sufficient for analysis.


Catalog Home › STUDENTPERFORMANCEFACTORS
New search
A new analysis is now available for this asset. View the analysis.
STUDENTPERFORMANCEFACTORS
CASUSER(merve.pakcan@stud.acs.upb.ro)
Completeness: 99%
Columns 20
Rows 6.6 K
Size 370.7 KB
Status None
Actions
Overview
Column Analysis
Sample Data
Date analyzed: 12 Jan 2025 17:21
Filter
Descriptive Measures
Metadata Measures
Data Quality Measures
Name Distinct Values Mean Median Minimum Maximum Standard Deviation
1 Hours... 41 19.97533 20 1 44 5.99059
2 Attenda... 41 79.97745 80 60 100 11.5474
3 Parental_... 3 -- -- -- -- .
4 Access_t... 3 -- -- -- -- .
5 Extracurr... 2 -- -- -- -- .
6 Sleep_H... 7 7.02906 7 4 10 1.4681
7 Previous... 51 75.07053 75 50 100 14.3997
8 Motivati... 3 -- -- -- -- .
9 Internet_... 2 -- -- -- -- .
10 Tutorin... 9 1.493719 1 0 8 1.2305
11 Family_I... 3 -- -- -- -- .
Hours_Studied
Frequency distribution:
0 10 20 30 40
Quantiles
Minimum: 1
25%: 16
50%: 20
75%: 24
Maximum: 44
Type: double
Format: --
Length: 8
Primary key candidate: No
Logical type: Interval
Semantic type: (n...


STUDENTPERFORMANCEFACTORS
CASUSER(merve.pakcan@stud.acs.upb.ro)
Completeness: 99%
Columns 20
Rows 6.6 K
Size 370.7 KB
Status None
Overview
Column Analysis
Sample Data
Date analyzed: 12 Jan 2025
Descriptive View › Column Graphs
Filter
Hours_Studied
Attendance
Parental_Involv...
Access_to_Res...
Extracurricular_...
Sleep_Hours
Previous_Scores
Motivation_Level
Internet_Access
Tutoring_Ses...
Family_Income
Teacher_Quality
School_Type
Exam_Score
Label: (none)
Semantic Type (none)
Information Privacy None
Primary Key Candidate No
Data Quality
Distinct Values 45
100% Completeness
1% Uniqueness
Deviation from Normality
Skewness 1.644808
Kurtosis 10.57542
Standard Deviation 3.890456
Kurtosis
Skewness
Frequency Distribution
Hide outliers
Top
Full Distribution
Minimum 55
Median 67
Mean 67.23566
Maximum 101



Hours_Studied	Attendance	Parental_Inv...	Access_to_R...	Extracurricul...	Sleep_Hours	Previous_Sc...	Motivation_...	Internet_Acc...
23	84	Low	High	No	7	73	Low	Yes
19	64	Low	Medium	No	8	59	Low	Yes
24	98	Medium	Medium	Yes	7	91	Medium	Yes
29	89	Low	Medium	Yes	8	98	Medium	Yes
19	92	Medium	Medium	Yes	6	65	Medium	Yes
19	88	Medium	Medium	Yes	8	89	Medium	Yes
29	84	Medium	Low	Yes	7	68	Low	Yes
25	78	Low	High	Yes	6	50	Medium	Yes
17	94	Medium	High	No	6	80	High	Yes
23	98	Medium	Medium	Yes	8	71	Medium	Yes
17	80	Low	High	No	8	88	Medium	No
17	97	Medium	High	Yes	6	87	Low	Yes

Data is selected and followed by lectures for Discover Information Assets parts and received column analysis and descriptive statistics.

Due to the presence of 13 categorical variables, applying a crosstab results in an overly complex and difficult-to-interpret table. The large number of categories creates excessive rows and columns, making the table impractical for meaningful analysis or presentation. As a result, the crosstab lacks clarity and fails to provide actionable insights in its current form.

Gender ▲	Female	Male
Internet_Access ▲	Exam_Score	Exam_Score
No	66.540178571	66.530909091
Yes	67.306344881	67.283130828

To address this, a focused analysis was conducted using two categorical variables: "Internet Access" and "Gender," alongside the continuous variable "Average Exam Score." The results show that internet access positively impacts the average exam score for both male and female students, with slight improvements observed in both groups.

4. Data Preparation

In the data preparation step, I ensured data quality by addressing missing values, outliers and removing duplicates to create a clean, analysis-ready dataset.I utilized SAS Studio's Develop Code and Flows menu to write and execute the code necessary for data preparation tasks.

Code

```
libname mylib clear;

/* Defining the "libmerv1" library to access the directory containing the data files */
libname libmerv1 "/export/viya/homes/merve.pakcan@stud.acs.upb.ro/casuser/";

/* Importing the dataset */
proc import datafile="/export/viya/homes/merve.pakcan@stud.acs.upb.ro/casuser/StudentPerformanceFactors.csv"
    dbms=csv out=libmerv1.studperformance replace;
    guessingrows=32767; /* Ensures SAS scans all rows to correctly determine column types */
    getnames=yes;
run;

/* Defining a macro variable for the dataset */
%let dataset = libmerv1.studperformance;

/* Displaying dataset structure and variables */
proc contents data=&dataset;
run;

/* Checking a sample of the dataset to ensure values loaded correctly */
proc print data=libmerv1.studperformance(obs=10);
run;

/* Identifying Missing Values */
/* Using proc means with nmiss to identify missing values in numeric columns. */
proc means data=libmerv1.studperformance n nmiss;
    var Hours_Studied Attendance Sleep_Hours Previous_Scores Exam_Score;
run;
```

Log Results

The MEANS Procedure

Variable	N	N Miss
Hours_Studied	6607	0
Attendance	6607	0
Sleep_Hours	6607	0
Previous_Scores	6607	0
Exam_Score	6607	0

As shown in the results, there were no missing values in the numeric fields.

Code

```
libname mylib clear;

/* Defining the "libmerv1" library to access the directory containing the data files */
libname libmerv1 "/export/viya/homes/merve.pakcan@stud.acs.upb.ro/casuser/";

/* Importing the dataset */
proc import datafile="/export/viya/homes/merve.pakcan@stud.acs.upb.ro/casuser/StudentPerformanceFactors.csv"
    dbms=csv out=libmerv1.studperformance replace;
    guessingrows=32767; /* Ensures SAS scans all rows to correctly determine column types */
    getnames=yes;
run;

/* Defining a macro variable for the dataset */
%let dataset = libmerv1.studperformance;

/* Displaying dataset structure and variables */
proc contents data=&dataset;
run;

/* Checking a sample of the dataset to ensure values loaded correctly */
proc print data=libmerv1.studperformance(obs=10);
run;

/* Identifying Missing Values */
/* Using proc means with nmiss to identify missing values in numeric columns. */
proc means data=libmerv1.studperformance n nmiss;
    var Hours_Studied Attendance Sleep_Hours Previous_Scores Exam_Score;
run;

/* Checking for Missing Values in all categorical Variables */
proc freq data=libmerv1.studperformance;
    tables _all_ / missing;
run;

/* Discovered missing values in the categorical variables */
/* Missing values: Teacher_Quality, Parental_Education_Level, and Distance_from_Home */
proc freq data=&dataset;
    tables Teacher_Quality Parental_Education_Level Distance_from_Home / missing;
run;
```

Log Results

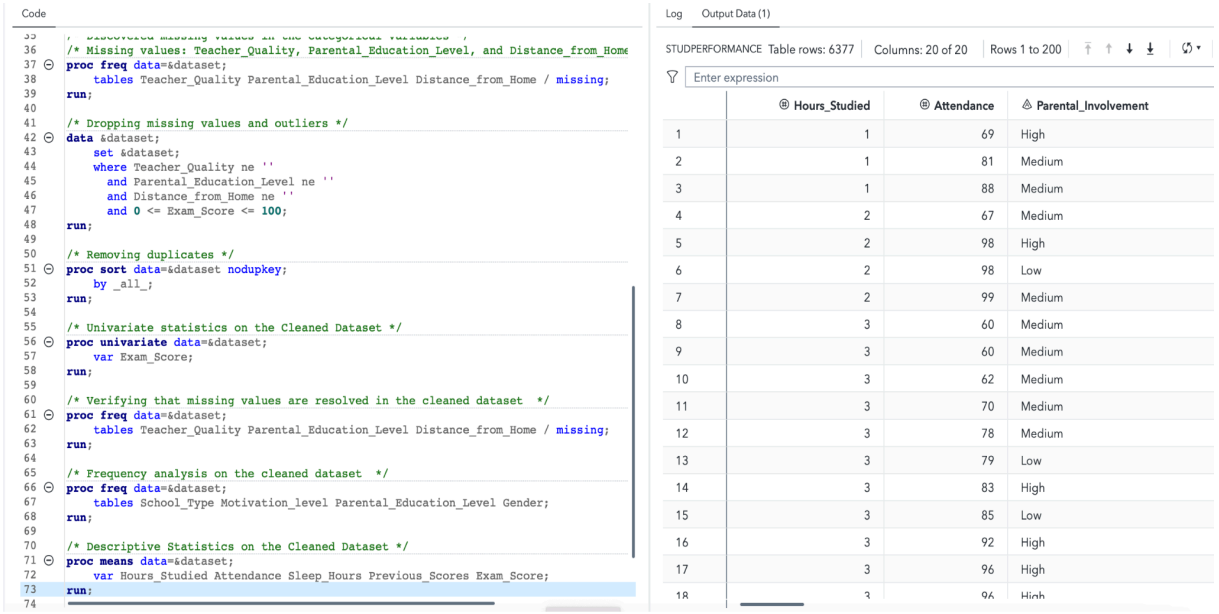
The FREQ Procedure

Teacher_Quality	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	78	1.18	78	1.18
High	1947	29.47	2025	30.65
Low	657	9.94	2682	40.59
Medium	3925	59.41	6607	100.00

Parental_Education_Level	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	90	1.36	90	1.36
College	1989	30.10	2079	31.47
High School	3223	48.78	5302	80.25
Postgraduate	1305	19.75	6607	100.00

Distance_from_Home	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	67	1.01	67	1.01
Far	658	9.96	725	10.97
Moderate	1998	30.24	2723	41.21
Near	3884	58.79	6607	100.00

All categorical variables were checked for missing values.In the data preparation step, a deeper analysis revealed missing values in specific categorical variables, which were not initially visible in the preliminary List Table review. Using proc frec with the missing option, we identified 78 missing values in Teacher_Quality, 90 in Parental_Education_Level, and 67 in Distance_from_Home.



```
Code

/* Missing values: Teacher_Quality, Parental_Education_Level, and Distance_from_Home
proc freq data=&dataset;
    tables Teacher_Quality Parental_Education_Level Distance_from_Home / missing;
run;

/* Dropping missing values and outliers */
data &dataset;
    set &dataset;
    where Teacher_Quality ne ''
      and Parental_Education_Level ne ''
      and Distance_from_Home ne ''
      and 0 <= Exam_Score <= 100;
run;

/* Removing duplicates */
proc sort data=&dataset nodupkey;
    by _all_;
run;

/* Univariate statistics on the Cleaned Dataset */
proc univariate data=&dataset;
    var Exam_Score;
run;

/* Verifying that missing values are resolved in the cleaned dataset  */
proc freq data=&dataset;
    tables Teacher_Quality Parental_Education_Level Distance_from_Home / missing;
run;

/* Frequency analysis on the cleaned dataset  */
proc freq data=&dataset;
    tables School_Type Motivation_level Parental_Education_Level Gender;
run;

/* Descriptive Statistics on the Cleaned Dataset */
proc means data=&dataset;
    var Hours_Studied Attendance Sleep_Hours Previous_Scores Exam_Score;
run;
```

Log | Output Data (1)

STUDPERFORMANCE Table rows: 6377 | Columns: 20 of 20 | Rows 1 to 200

Enter expression

	Hours_Studied	Attendance	Parental_Involvement
1	1	69	High
2	1	81	Medium
3	1	88	Medium
4	2	67	Medium
5	2	98	High
6	2	98	Low
7	2	99	Medium
8	3	60	Medium
9	3	60	Medium
10	3	62	Medium
11	3	70	Medium
12	3	78	Medium
13	3	79	Low
14	3	83	High
15	3	85	Low
16	3	92	High
17	3	96	High
18	3	96	High

Out of the all dataset, only a small number of rows contained missing values,so dropped missing values.Also ensured that exam_score values were within the valid range of 0 to 100.No duplicates were found in the dataset. (Needed duplicate steps done.)Outliers found in Hours_studied, exam_score and tutoring session that was also small considering the all dataset, so eliminated outliers using Python(43 in Hours_ Studied, 430 in Tutoring Session, 104 in exam score).After all data cleaning, dataset has 5836 rows.

Descriptive Statistics (Mean Analysis): The mean and standard deviation were calculated for study hours, attendance, sleep hours, previous scores, and exam scores. The average Exam_Score is 66.98, while Previous_Scores is 75.10, suggesting a possible decline in students' current performance compared to past achievements.

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
Hours_Studied	5836	20.0080535	5.7822916	4.0000000	36.0000000
Attendance	5836	80.0219328	11.4995264	60.0000000	100.0000000
Sleep_Hours	5836	7.0412954	1.4696616	4.0000000	10.0000000
Previous_Scores	5836	75.1077793	14.3481094	50.0000000	100.0000000
Exam_Score	5836	66.9883482	3.2301389	59.0000000	75.0000000

Frequency Analysis: The distribution of categorical variables shows that the majority of students (69.59%) attend public schools, and over half (50.84%) have medium motivation levels, indicating that motivation could impact student performance.

The FREQ Procedure

School_Type	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Private	1775	30.41	1775	30.41
Public	4061	69.59	5836	100.00

Motivation_Level	Frequency	Percent	Cumulative Frequency	Cumulative Percent
High	1162	19.91	1162	19.91
Low	1707	29.25	2869	49.16
Medium	2967	50.84	5836	100.00

Parental_Education_Level	Frequency	Percent	Cumulative Frequency	Cumulative Percent
College	1766	30.26	1766	30.26
High School	2892	49.55	4658	79.81
Postgraduate	1178	20.19	5836	100.00

Gender	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Female	2466	42.25	2466	42.25
Male	3370	57.75	5836	100.00

Univariate Analysis: The univariate analysis of Exam_Score shows no outliers, with values falling within expected ranges. For Exam_Score, scores range from 59 to 75, with an average of 66.98 and a narrow spread (standard deviation: 3.23).

The UNIVARIATE Procedure
Variable: Exam_Score

Moments			
N	5836	**Sum Weights**	5836
Mean	66.9883482	**Sum Observations**	390944
Std Deviation	3.23013891	**Variance**	10.4337974
Skewness	0.00471108	**Kurtosis**	-0.4185257
Uncorrected SS	26249574	**Corrected SS**	60881.2077
Coeff Variation	4.82194142	**Std Error Mean**	0.04228278

Basic Statistical Measures			
Location		**Variability**	
Mean	66.98835	**Std Deviation**	3.23014
Median	67.00000	**Variance**	10.43380
Mode	66.00000	**Range**	16.00000
		Interquartile Range	4.00000

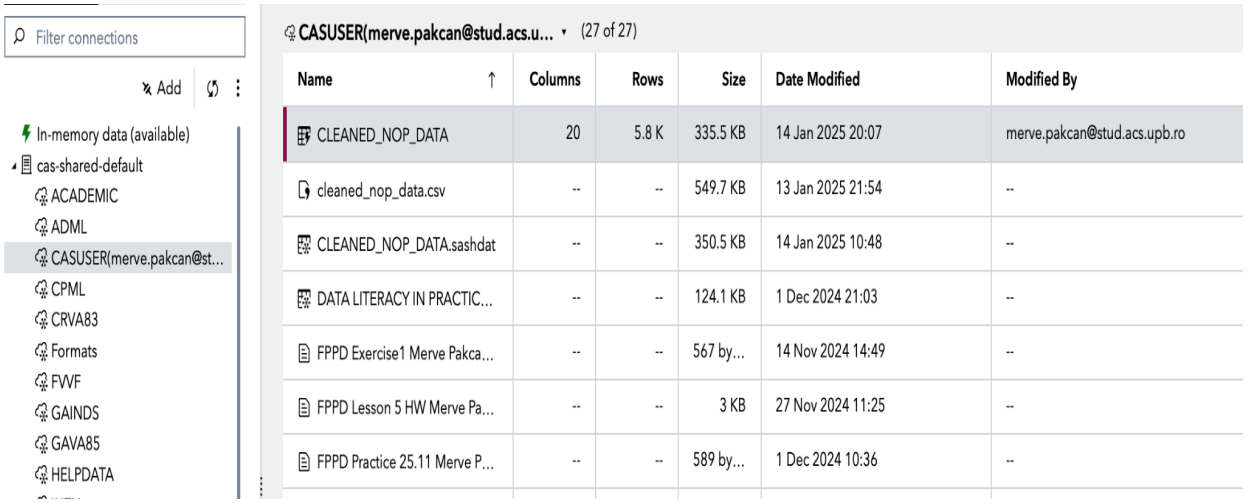
Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	1584.294	**Pr > |t|**	<.0001
Sign	**M**	2918	**Pr >= |M|**	<.0001
Signed Rank	**S**	8516183	**Pr >= |S|**	<.0001

Quantiles (Definition 5)	
Level	**Quantile**
100% Max	75
99%	74
95%	72
90%	71
75% Q3	69
50% Median	67
25% Q1	65
10%	63
5%	62
1%	60
0% Min	59

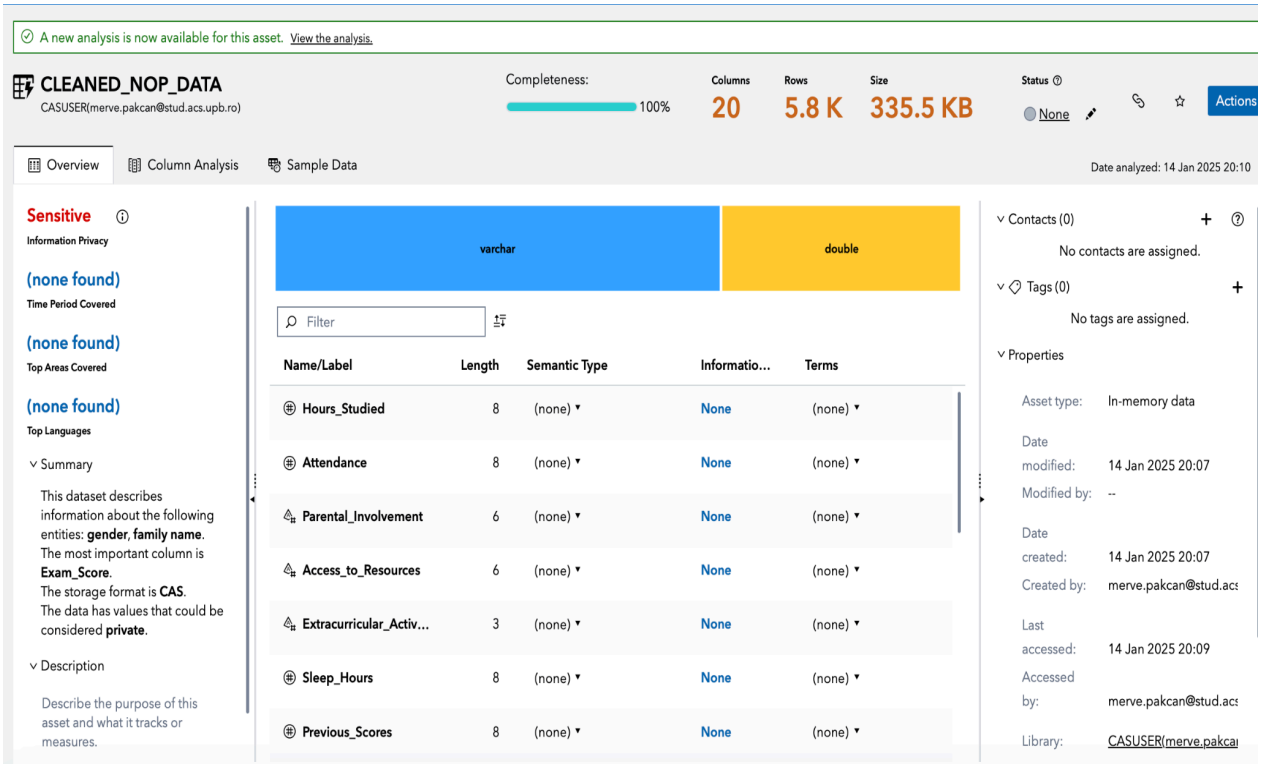
Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
59	5421	75	5365
59	5402	75	5382
59	5394	75	5587
59	5164	75	5627
59	5151	75	5773

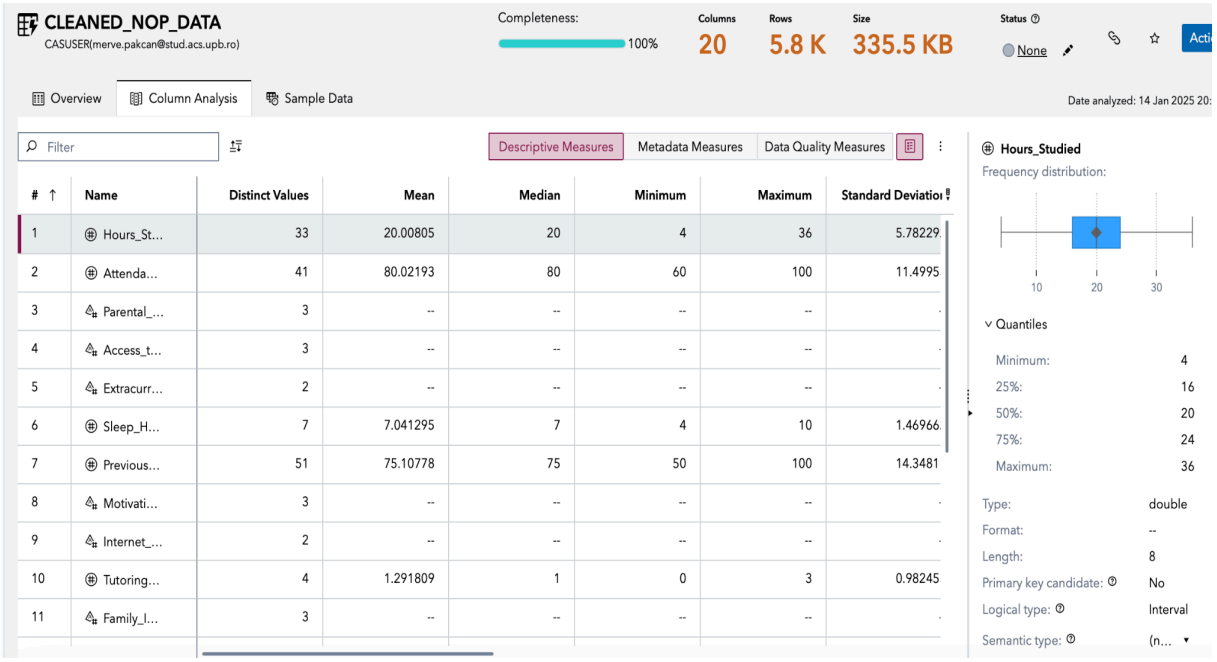
SAS

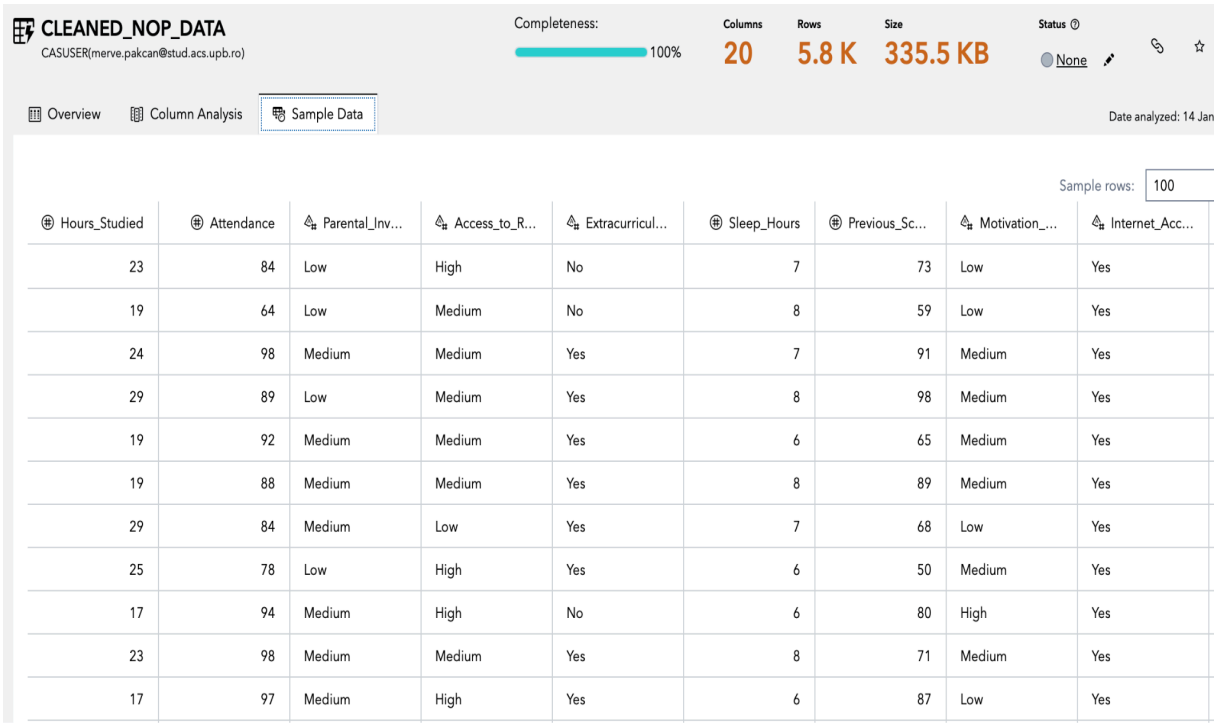
I exported my dataset and imported my file in SAS Data Explorer Manage Data part.



The name of the cleaned data is CLEANED_NOP_DATA that is already in-memory.







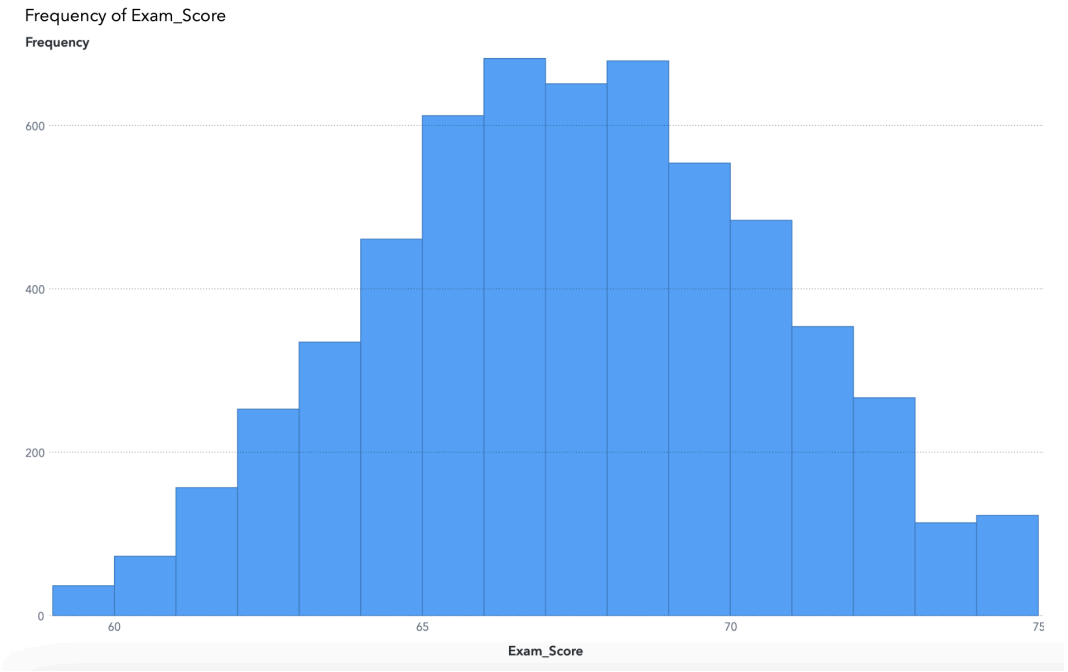
After uploading clean data, in discover information assets part, no missing value and outliers found in data, data is clean and ready to analyze.

5. Research Methodology

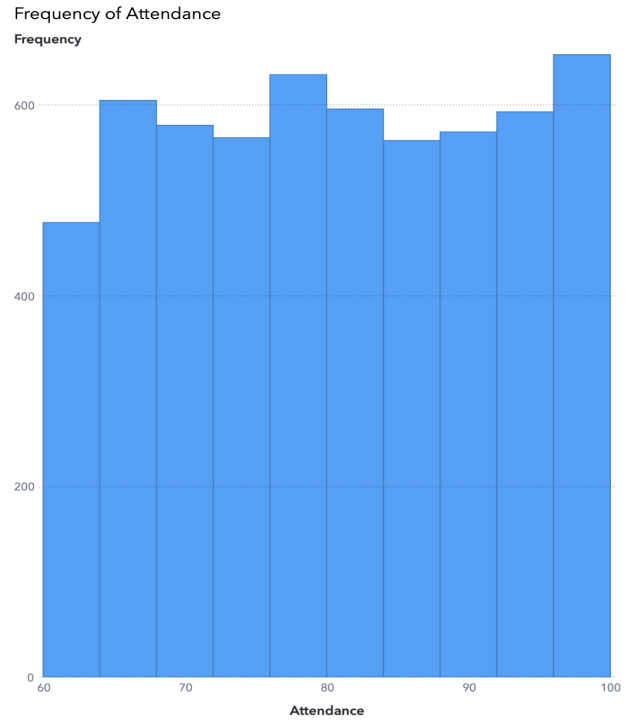
Research Questions

Question	Sub-Question	Variables	Notes
What factors most influence student performance?	Which variables have the strongest correlation with Exam_Score?	Dependent Variable: Exam_Score Independent Variable: Attendance, Hours_Studied, Previous_Scores, Parental_Involvement, Access_to_Resources, Extracurricular_Activities, Sleep_Hours, Previous_Scores, Motivation_Level, Internet_Access, Tutoring_Sessions, Family_Income, Teacher_Quality, School_Type, Peer_Influence, Physical_Activity, Learning_Disabilities, Parental_Education_Level, Distance_from_Home,Gender	Needed to identify key variables of success.
What socioeconomic factors most impact exam results?	Which socioeconomic variables have the strongest correlation with Exam_Score?	Dependent Variable: Exam_Score Independent Variable: Parental_Education_Level, Family_Income, Access_to_Resources, Internet_Access, Distance_from_Home, Peer_Influence, Parental_Involvement	Exploring educational equity and disparities.

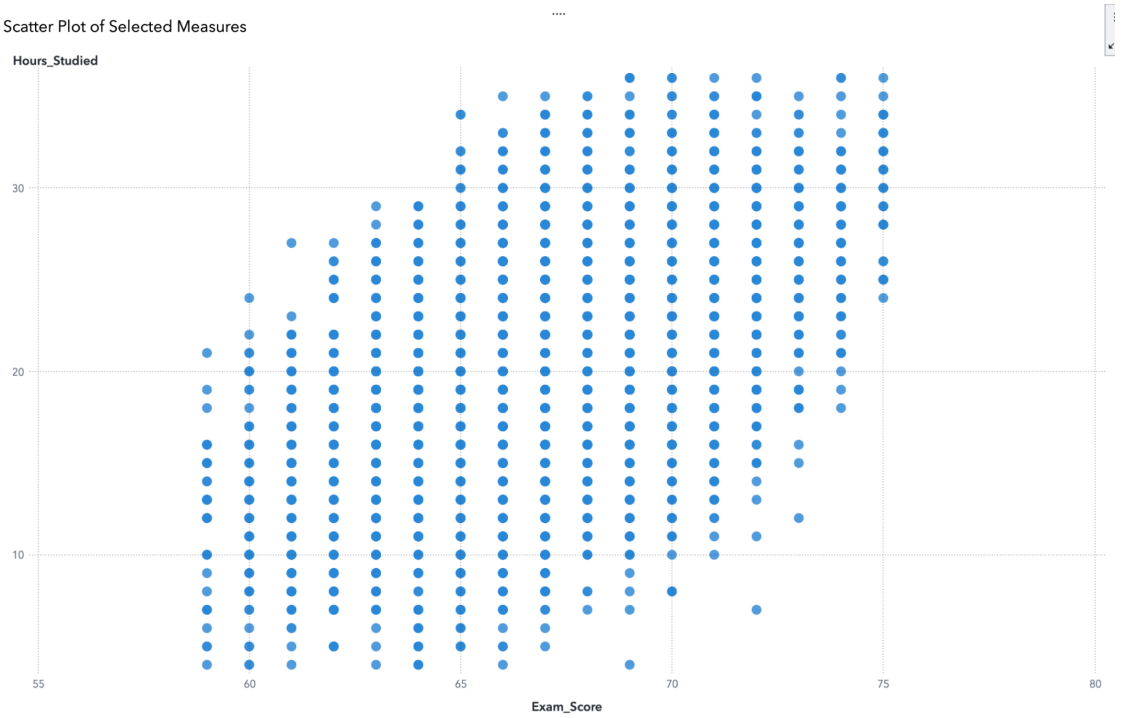
6. First Exploratory Data Analysis



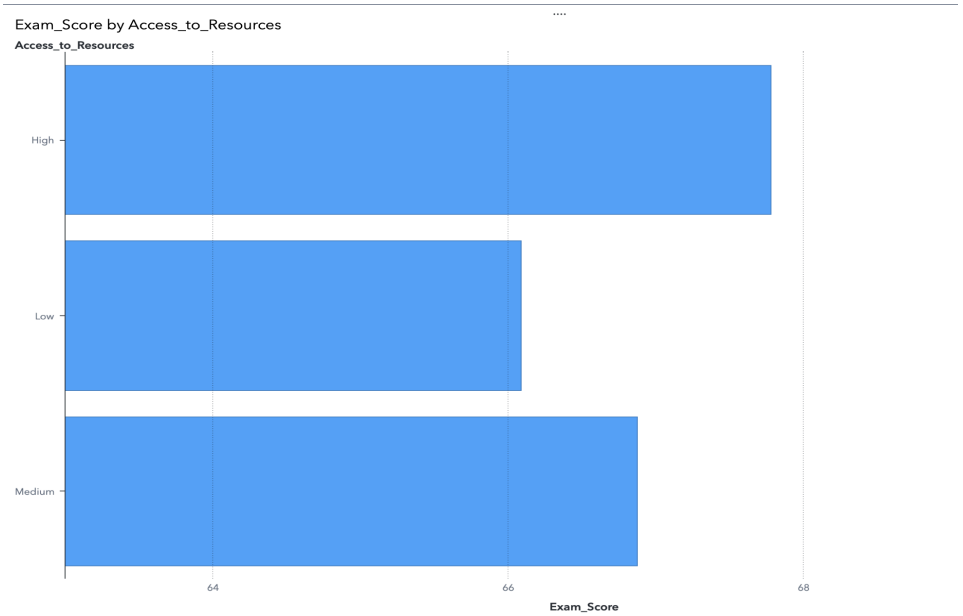
The Exam_Score histogram exhibits a roughly bell-shaped distribution. The center of the distribution is around 67, while the scores are spread densely between 59 and 75.



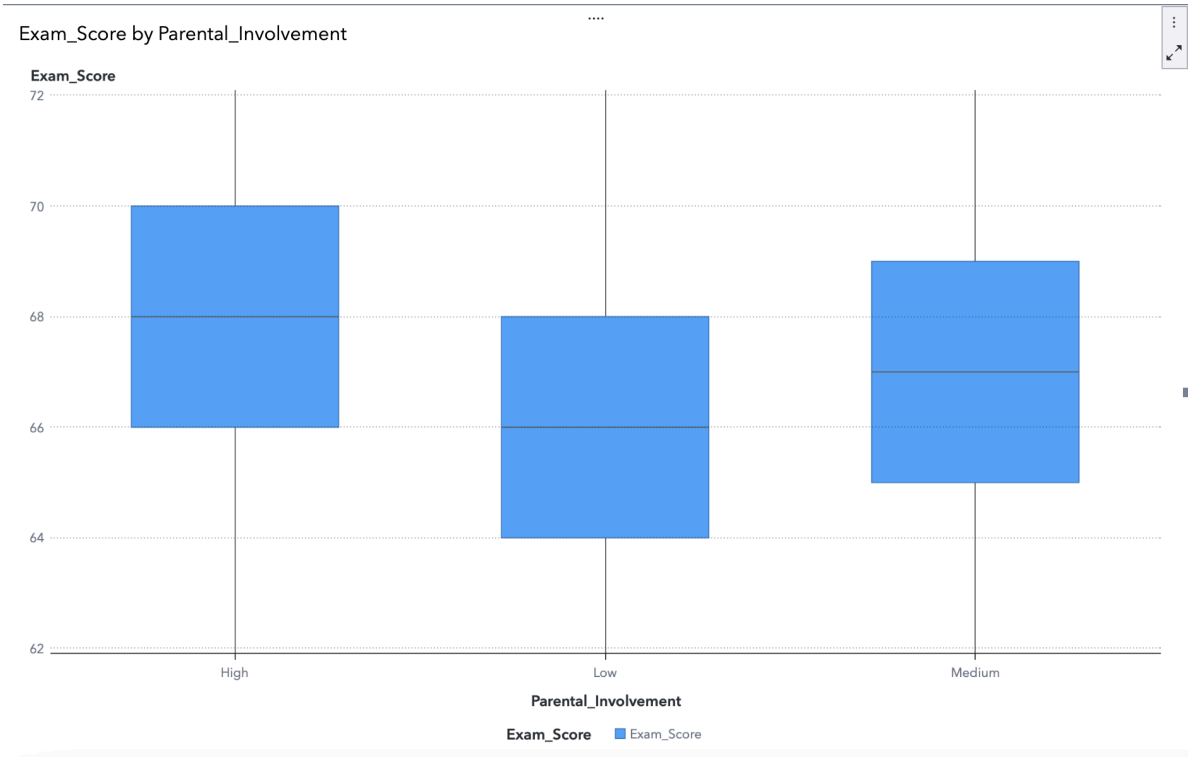
The Attendance histogram exhibits a roughly uniform distribution. The frequency of attendance is distributed fairly evenly across the range of values, with no significant peaks or dips. The attendance values are spread consistently between 60 and 100, suggesting a balanced distribution of attendance rates among the students.



The scatter plot shows a positive correlation between hours studied and exam scores. As study hours increase, exam scores generally rise, indicating that dedicated study time contributes to better performance. While some variability exists, the overall trend highlights the importance of consistent preparation.



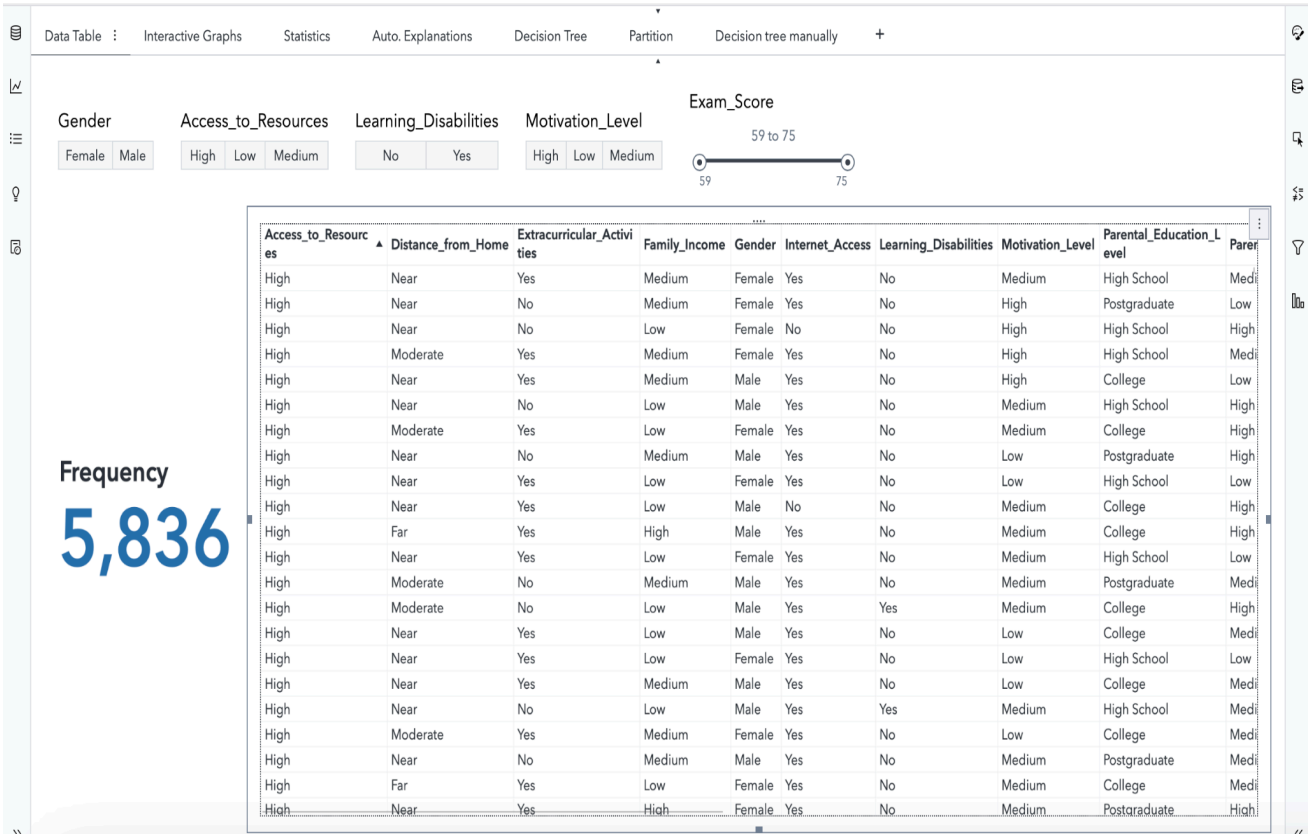
The bar chart shows the relationship between access to resources and exam scores. While students with higher access to resources tend to score slightly better, the narrow range of exam scores results in relatively close values across all groups.



The box plot shows that higher parental involvement leads to higher median exam scores. Low involvement results in the lowest scores, while medium involvement shows greater

variability. This highlights the positive impact of strong parental support on academic performance.

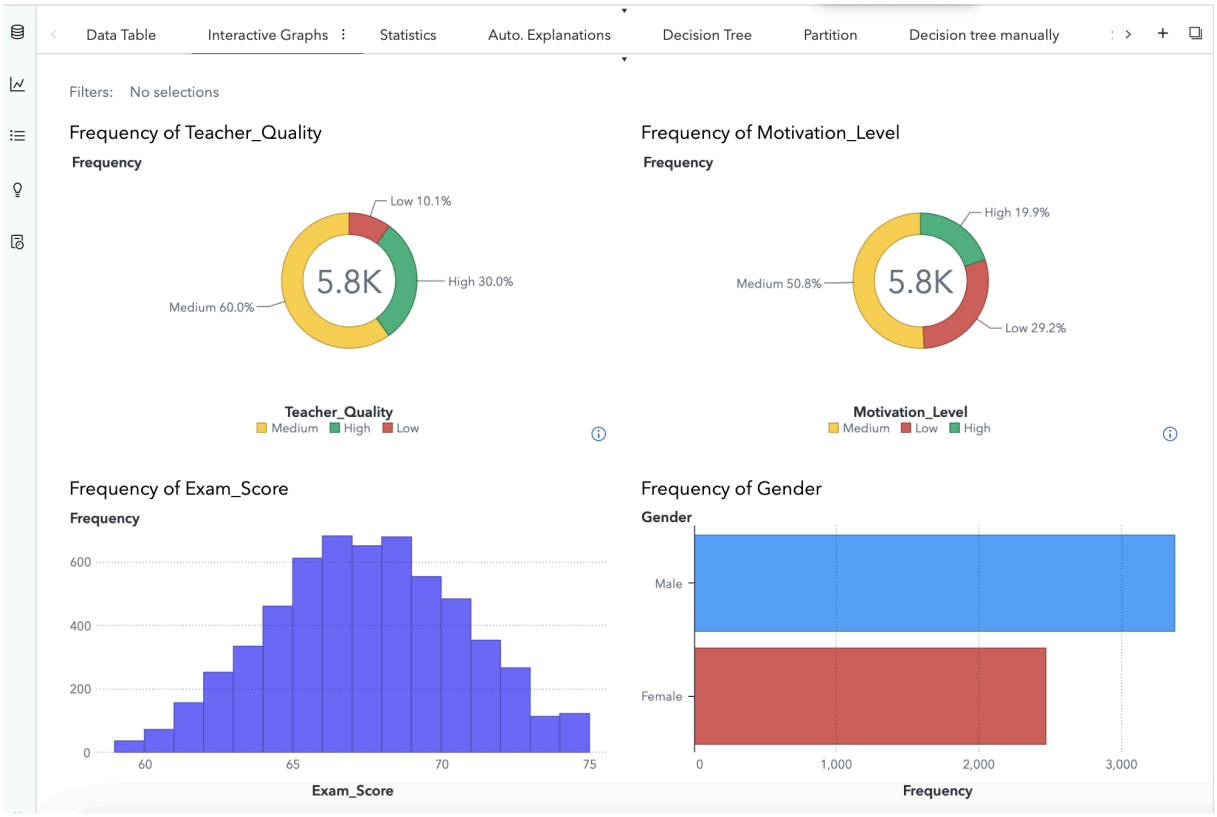
7. Investigate relationships in data



Access_to_Resources	Distance_from_Home	Extracurricular_Activities	Family_Income	Gender	Internet_Access	Learning_Disabilities	Motivation_Level	Parental_Education_Level	Parer
High	Near	Yes	Medium	Female	Yes	No	Medium	High School	Medi
High	Near	No	Medium	Female	Yes	No	High	Postgraduate	Low
High	Near	No	Low	Female	No	No	High	High School	High
High	Moderate	Yes	Medium	Female	Yes	No	High	High School	Medi
High	Near	Yes	Medium	Male	Yes	No	High	College	Low
High	Near	No	Low	Male	Yes	No	Medium	High School	High
High	Moderate	Yes	Low	Female	Yes	No	Medium	College	High
High	Near	No	Medium	Male	Yes	No	Low	Postgraduate	High
High	Near	Yes	Low	Female	Yes	No	Low	High School	Low
High	Near	Yes	Low	Male	No	No	Medium	College	High
High	Far	Yes	High	Male	Yes	No	Medium	College	High
High	Near	Yes	Low	Female	Yes	No	Medium	High School	Low
High	Moderate	No	Medium	Male	Yes	No	Medium	Postgraduate	Medi
High	Moderate	No	Low	Male	Yes	Yes	Medium	College	High
High	Near	Yes	Low	Male	Yes	No	Low	College	Medi
High	Near	Yes	Low	Female	Yes	No	Low	High School	Low
High	Near	Yes	Medium	Male	Yes	No	Low	College	Medi
High	Near	No	Low	Male	Yes	Yes	Medium	High School	Medi
High	Moderate	Yes	Medium	Female	Yes	No	Low	College	Medi
High	Near	No	Medium	Male	Yes	No	Medium	Postgraduate	Medi
High	Far	Yes	Low	Female	Yes	No	Medium	College	Medi
High	Near	Yes	High	Female	Yes	No	Medium	Postgraduate	High

To further explore and visualize the data, I started by creating a list table to display key variables such as Gender, Access to Resources, Learning Disabilities, Motivation Level, and Exam Score. To enhance interactivity, I added page controls through the expand page controls option. I selected Gender as the first filter and enabled the show object title feature. Next, I added additional controls. There are 4 control objects above the list table.3 button bars are created for Gender, Access to Resources and Motivation Level variables.A slider control is created for Exam Score variable.

To refine the analysis, I dragged the Frequency measure to the left of the table using the key value object in objects, enabling a count of records matching the filtering criteria. When filters are applied, the table dynamically updates to reflect the selected criteria. For example, selecting "Male" as the gender automatically filters the table to show only the relevant records that meet the applied conditions. This interactive approach enables a more targeted and flexible analysis of the dataset.

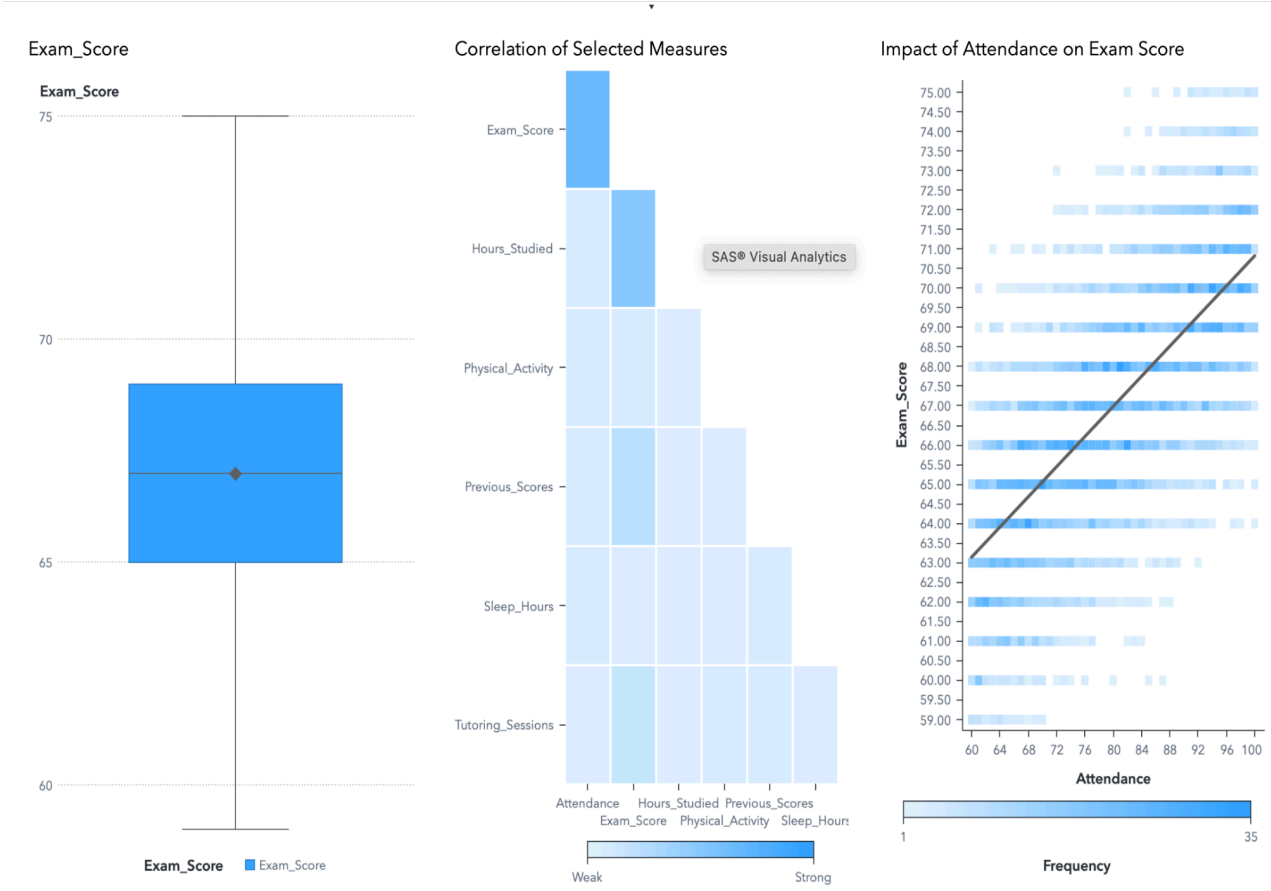


To perform a simple statistical analysis in SAS Visual Analytics,automatic actions are utilized to link all objects on the report page, ensuring seamless interactivity across visualizations. **Histograms** are created to examine the distribution of single measures, such as the **Exam_Score**, which helps identify the concentration and spread of scores. The histogram in this analysis reveals a relatively symmetric distribution, with most scores clustering between 65 and 70.

In addition to histograms, **frequency visualizations** like the donut charts for **Teacher_Quality** and **Motivation_Level** allow us to understand categorical distributions. For instance, **Teacher_Quality** is predominantly rated as Medium (60%) and High (30%), with only a small proportion marked as Low (10.1%). Similarly, **Motivation_Level** shows Medium as the most frequent level (50.8%), followed by Low (29.2%) and High (19.9%).

The bar chart for **Gender Frequency** highlights a nearly equal distribution between male and female participants, while the inclusion of a frequency total (5.8K) provides an overview of the dataset size.

Exploratory Data Analysis (EDA) plays a pivotal role in this process, helping to uncover general patterns and make informed decisions before addressing specific research questions. By understanding data distributions and group differences, we build a foundation for deeper and more targeted analyses.



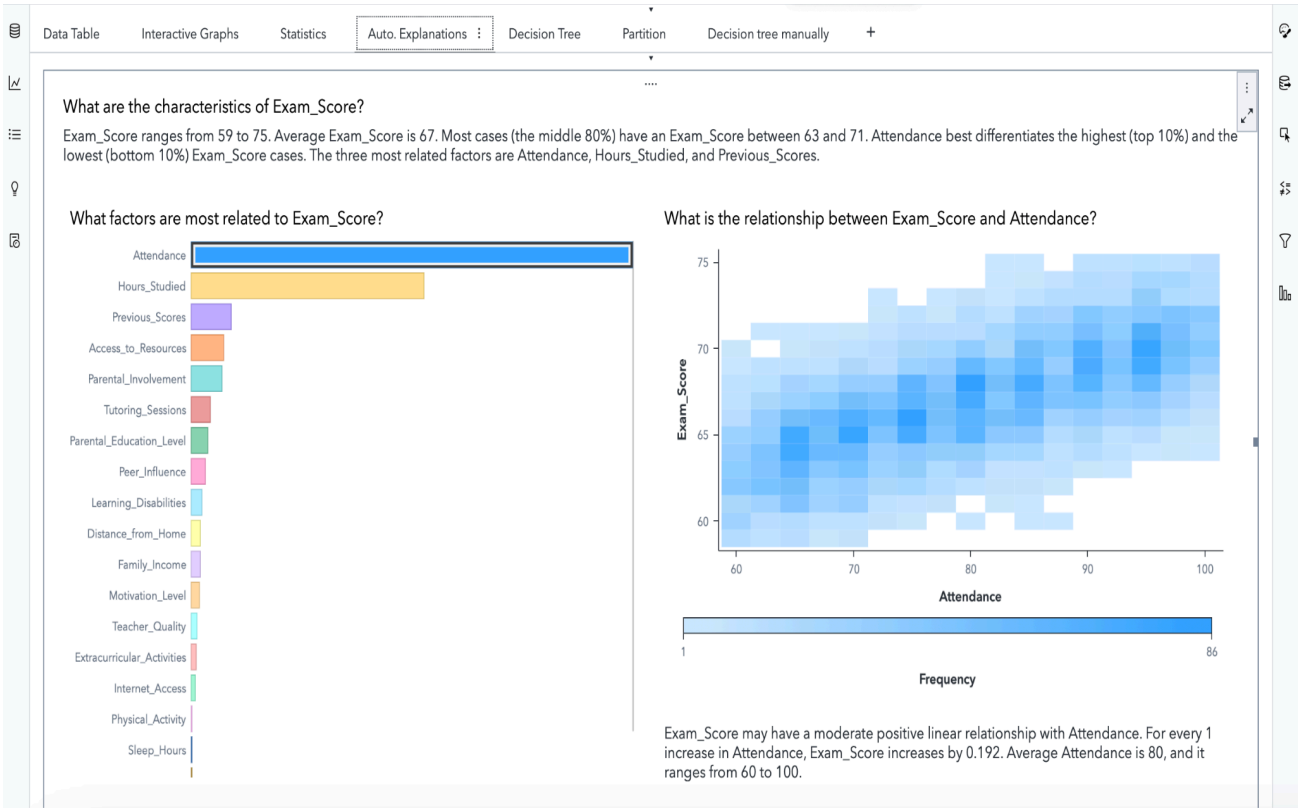
The box plot of exam scores reveals that most scores fall between 65 and 69, as indicated by the interquartile range (IQR), which encompasses the middle 50% of the data. The median score is 67, closely aligning with the mean of 66.99, suggesting a relatively symmetric distribution without significant skewness. The scores range from a minimum of 59 to a maximum of 75, with no extreme outliers observed. This indicates a consistent performance trend among participants, with the majority clustering within a narrow and well-defined range.

The correlation matrix illustrates the strength of relationships between variables, with darker blocks indicating stronger correlations. "Attendance" shows a clear positive correlation with "Exam_Score," underscoring its crucial role in academic success. Additionally, "Hours_Studied" and "Previous_Scores" also have strong positive correlations, indicating their significant influence on performance. This highlights the importance of consistent participation, prior preparation, and study habits in improving exam outcomes.

The heatmap shows a positive linear trend between "Attendance" and "Exam_Score," with higher attendance linked to better performance. Students with high attendance achieve significantly higher scores, highlighting attendance as a key factor for academic success.

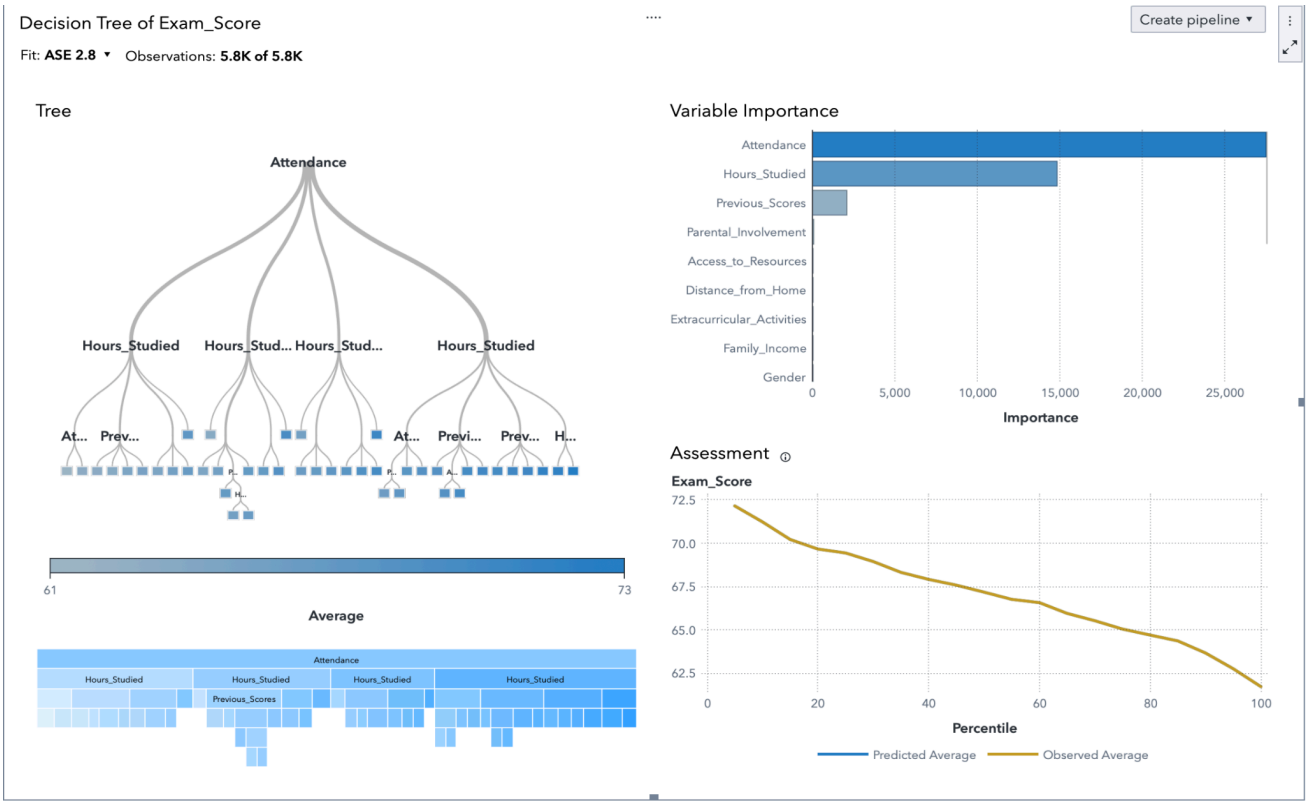
8. First step to build a decision

First Question

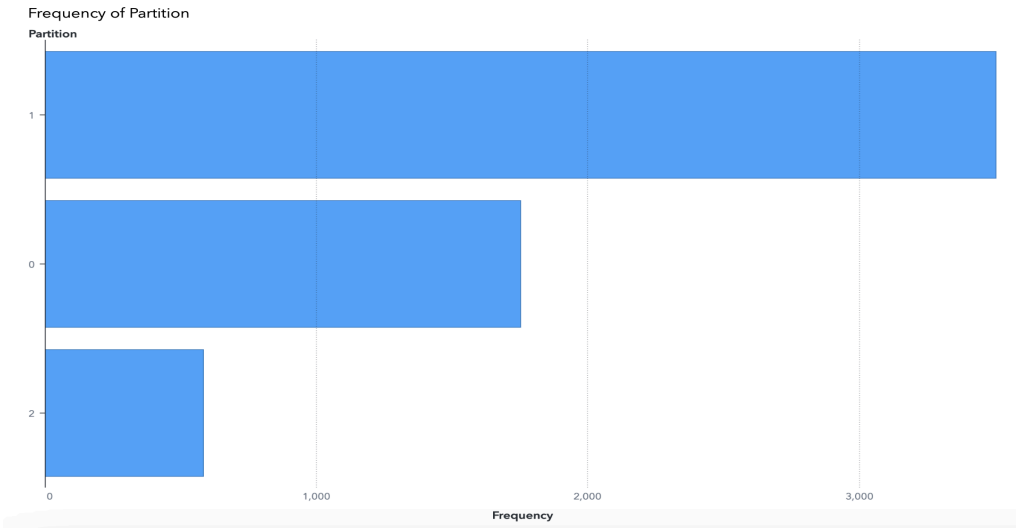


To answer the first question, *"What factors most influence student performance?"*, **automated explanations** and a **decision tree** were utilized. This analysis identified the most significant variables affecting Exam Scores, namely **Attendance**, **Hours Studied**, and **Previous Scores**. Automated explanations provided a clear and prioritized understanding of these relationships, demonstrating that enhancing attendance and study habits can lead to meaningful improvements in academic performance.

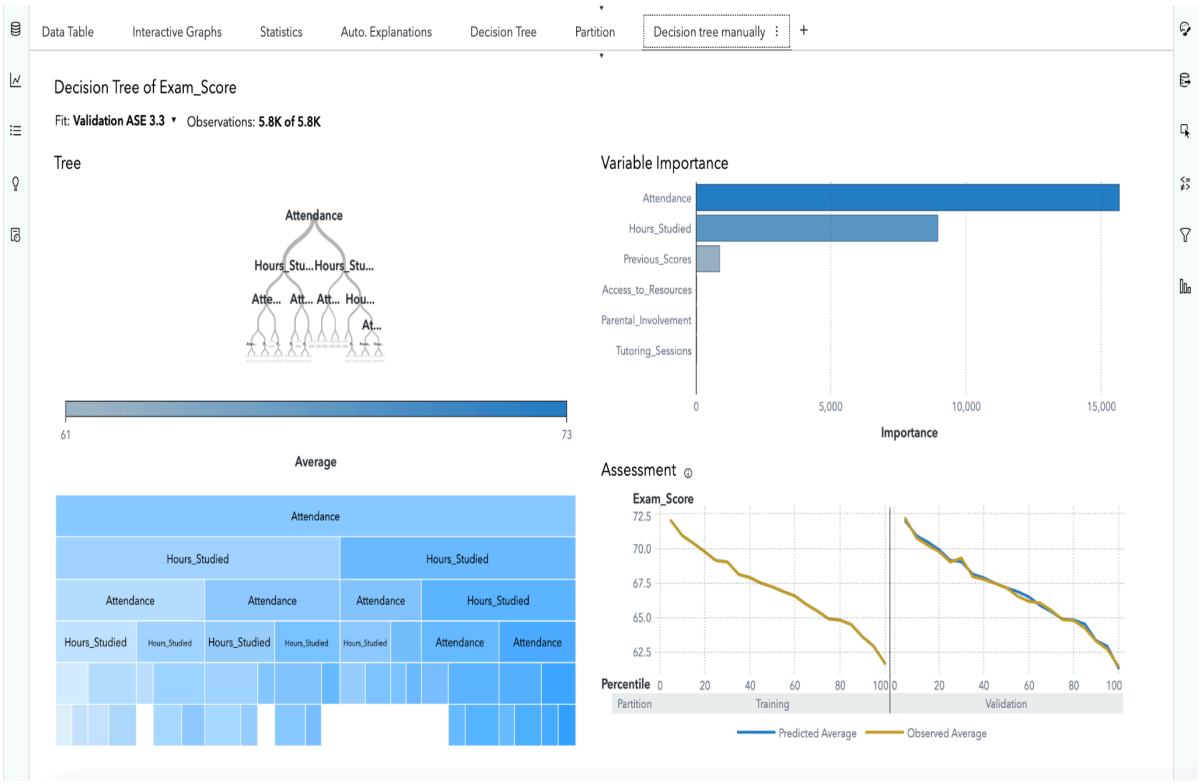
The graph on the right illustrates a **moderate positive linear relationship** between **Attendance** and **Exam Score**. As attendance increases, Exam Scores show a consistent upward trend, with the analysis indicating that for every 1 increase in Attendance, Exam Score increases by approximately **0.192 points**.

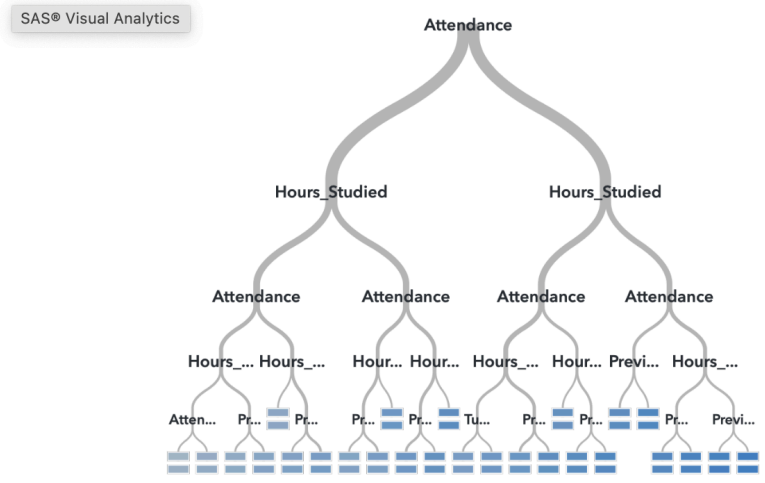


The decision tree analysis identifies **Attendance** as the most critical variable in predicting **Exam Score**, which is the root node.Decision node based on **Hours Studied** and **Previous Scores** further emphasize their importance as secondary factors. This is supported by the Variable Importance chart, which highlights Attendance and Hours Studied as the dominant predictors. The bottom-left chart shows the average Exam Score distribution across different combinations of Attendance, Hours Studied, and Previous Scores. Darker shades indicate higher scores, highlighting the strong impact of Attendance and Hours Studied on performance.



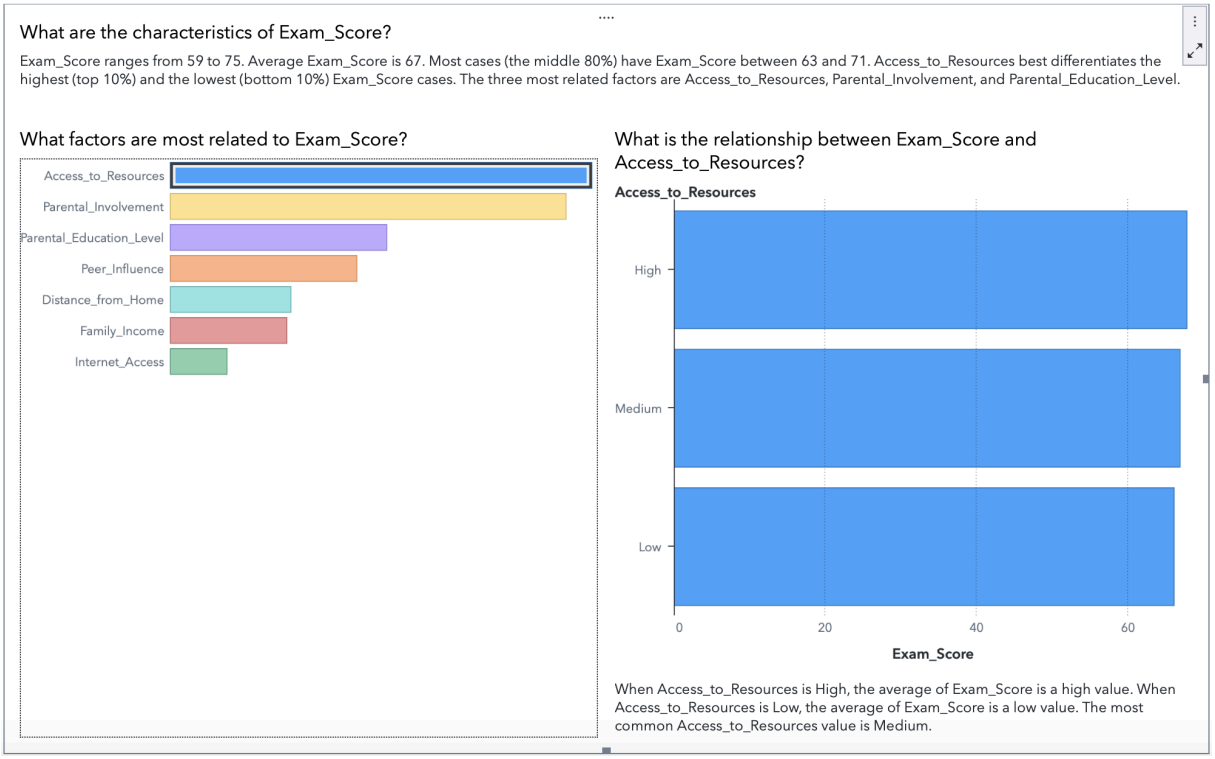
To be sure the model is accurate, partition variable is created with adding new data items.The dataset is divided into three partitions: **Training** (60%, 3501), **Validation** (30%,1751), and **Test** (10%, 584), resulting in a more reliable and effective model.



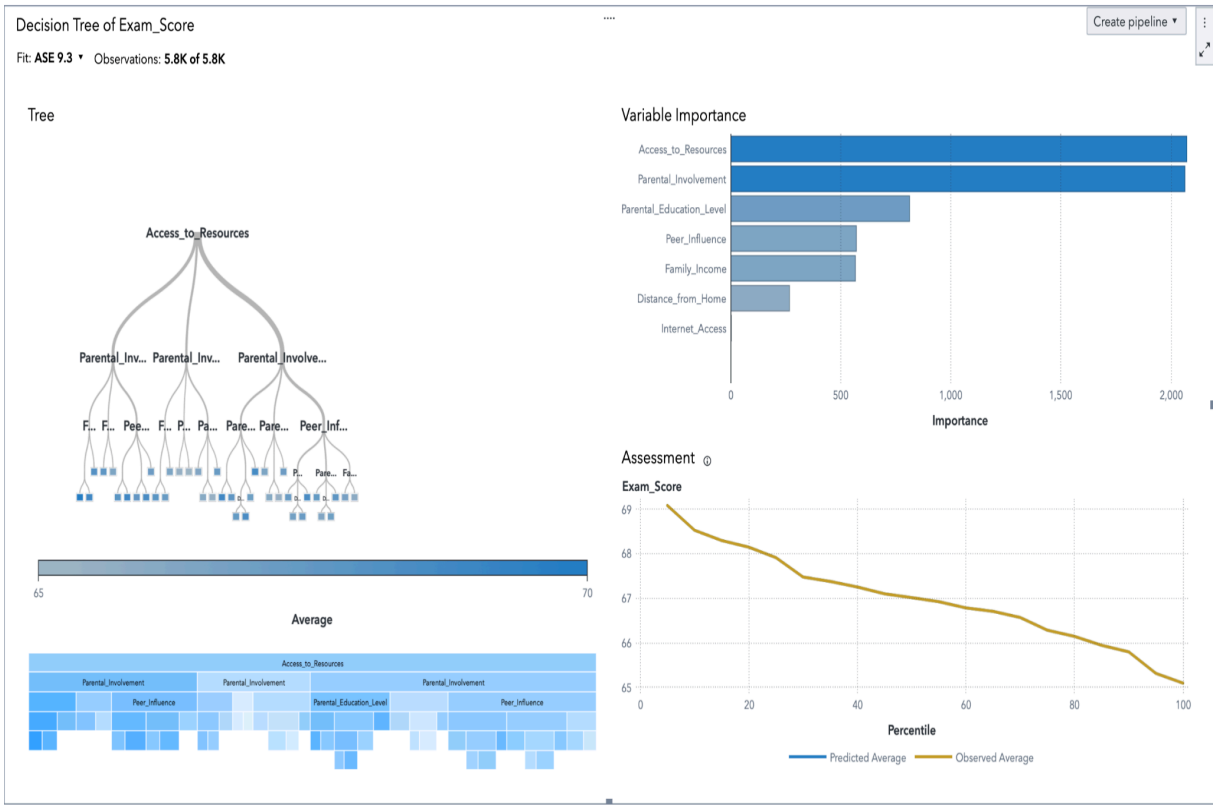


After partitioning, the analysis confirmed that **Attendance** is the most critical predictor of **Exam Score**, followed by **Hours Studied** and **Previous Scores**. The decision tree maintained a consistent structure across training and validation sets, demonstrating robustness. The predicted averages closely matched the observed averages, validating the model's accuracy and generalization. Partitioning preserved balanced distributions, ensuring reliable evaluation while highlighting that higher **Attendance** and **Hours Studied** lead to better outcomes.

Second Question



To answer the second question, "*What socioeconomic factors most impact exam results?*", both automated explanation and decision tree analysis were used. The automated explanation revealed that Access to Resources, Parental Involvement are the most significant factors influencing Exam Scores**.**



The decision tree analysis confirmed these findings, with **Access to Resources** identified as the root node, followed by splits based on **Parental Involvement** and **Parental Education Level**. These splits illustrate how these factors interact to influence scores. The model's assessment showed strong alignment between predicted and observed averages, validating its accuracy and emphasizing the critical role of socioeconomic factors in shaping academic outcomes.

9. Final conclusions

As a result, this project provided clear answers to my research questions. Attendance, hours studied, and previous scores were identified as the most significant predictors of academic performance, highlighting the importance of consistent participation and preparation. Additionally, socioeconomic factors, particularly access to resources and parental involvement, were shown to play critical roles in shaping exam outcomes. These findings not only validate the robustness of the analysis process but also emphasize the potential of data-driven insights to address educational disparities and align with SDG goals.Achieving educational equality requires a shared responsibility between families, educational institutions, and governments to address systemic barriers (Boeren, 2019).

The easiest part of the project was creating visualizations and decision trees using clean data in SAS Visual Analytics, which made exploring patterns simple and effective. Sas Viya Explore and Visualize part is so useful to create visualisations easily, I found it so effective. The most challenging part was handling some technical issues during data cleaning.Eventually I handled those problems, which also helped me to improve my coding skills in the Sas.

This project was highly valuable for developing skills in SAS, particularly in understanding each step of the process. From cleaning the data and conducting first exploratory analysis to gain deeper insights in data, to formulating correct key questions and answering them using tools like automated explanations and decision trees, it provided a structured learning experience. Moreover, the project aligns with certain SDG principles, making it socially significant. It also revealed valuable insights, highlighting the potential for more comprehensive analyses with larger datasets in the future.

The findings of this project have broadened my understanding of how data-driven insights can be leveraged to create tangible societal impacts. By identifying key factors such as attendance, study habits, and parental involvement, the analysis offers actionable recommendations for reducing educational inequalities. These insights empower policymakers, educators, and institutions to design targeted programs, such as community outreach initiatives to enhance parental engagement and strategies to promote attendance and study habits while supporting low-performing students.Parental involvement not only enhances academic outcomes but also builds children’s emotional resilience and self-regulatory abilities, key factors for achieving sustainable educational success (Sanders et al., 2022). The knowledge gained through this project highlights the potential of advanced analytics to drive meaningful change and contribute to a more equitable and effective education system.Addressing structural inequalities in education requires both inclusive policies and transformative approaches(Unterhalter, 2019).

Tackling the environmental, economic, and social dimensions of sustainability requires a concerted effort, high quality data, and the integration of innovative tools, including robust big data analysis activities(WEF, 2024).Advanced data analysis is a powerful tool for driving progress toward the United Nations Sustainable Development Goals (SDGs). By identifying patterns, correlations, and disparities within large datasets, it enables informed decision-making and targeted interventions. For example, in alignment with **SDG 4 (Quality Education)**, data analysis can reveal the factors influencing student performance, helping to design programs that reduce educational inequalities and improve learning outcomes. Similarly, for **SDG 10 (Reduced Inequalities)**, it uncovers systemic disparities, enabling inclusive strategies to support underserved populations. Overall, advanced data analysis fosters evidence-based solutions, ensuring resources are allocated efficiently and equitably, ultimately accelerating global progress toward a sustainable future.

10. References

AH, K., Oldayo, A. A., & Fakai, A. A. (2020). Factors and Effects of Poor Background on the Students Academic Performance in Physics at Senior Secondary School in Birnin Kebbi Metropolis.

Boeren, E. (2019). Understanding Sustainable Development Goal (SDG) 4 on "quality education" from micro, meso and macro perspectives. *International review of education*, *65*, 277-294.

Friedman, J., York, H., Graetz, N., Woyczynski, L., Whisnant, J., Hay, S. I., & Gakidou, E. *(2020).* Measuring and forecasting progress towards the education-related SDG targets. *Nature, 580(7805), 636-639.*

Kaggle. *Student Performance Factors Data.* Retrieved from https://www.kaggle.com/datasets/lainguyn123/student-performance-factors

Rad, D., Redeş, A., Roman, A., Ignat, S., Lile, R., Demeter, E., ... & Rad, G. *(2022). Pathways to inclusive and equitable quality early childhood education for achieving SDG4 goal—a scoping review. Frontiers in psychology, 13, 955833.*

Sanders, M. R., Divan, G., Singhal, M., Turner, K. M., Velleman, R., Michelson, D., & Patel, V. (2022). Scaling up parenting interventions is critical for attaining the sustainable development goals. *Child Psychiatry & Human Development, 53(5), 941-952.*

Unterhalter, E. (2019). The many meanings of quality education: Politics of targets and indicators in SDG 4. *Global Policy, 10, 39-51.*

WEF, World Economic Forum (2024). The digital dividend: How to harness data for sustainability wins. Retrieved from https://www.weforum.org/stories/2024/08/how-to-harness-data-for-sustainability-wins/