STUDENT PERFORMANCE FACTORS

Visual Analytics Techniques Homework

Merve Pakcan Tufenk

INTRODUCTION

This project seeks to analyze and understand the factors that influence student performance.

USER: Parents, administrators, educators, psychologists, counsellors

DATA: A comprehensive dataset of student performance metrics

TASK: Uncovering patterns, relationships, and key drivers of academic success to inform effective interventions

SDG 4
Quality
Education

SDG 10
Reduced
Inequalities

6600 ROWS

20 COLUMNS

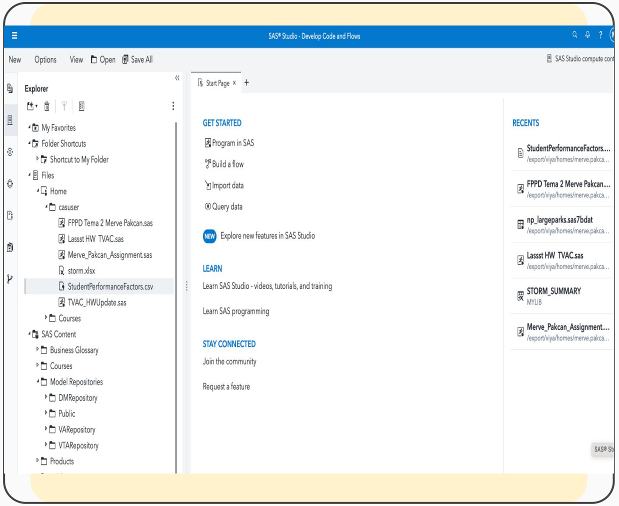
Data Sourcing & Acquisition

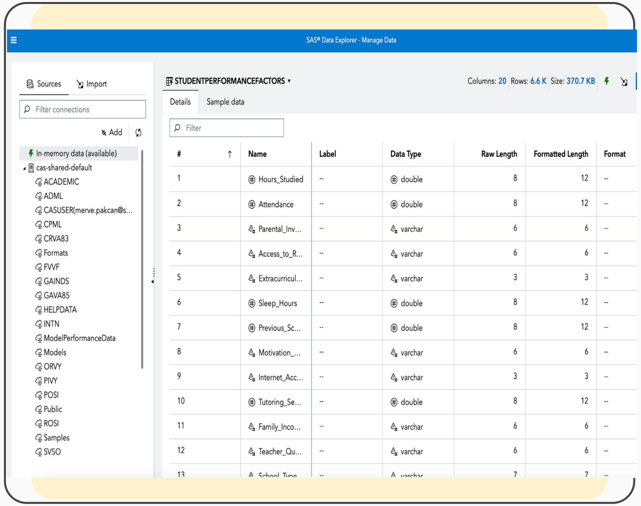
Student Performance Factors dataset from Kaggle

Uploaded CAS using *SAS Studio - Develop Code and Flows* menu.

Loaded file into the CAS environment, within the CASUSER library, using *Manage Data* menu, **indicated by the green 'In-memory data' status.**

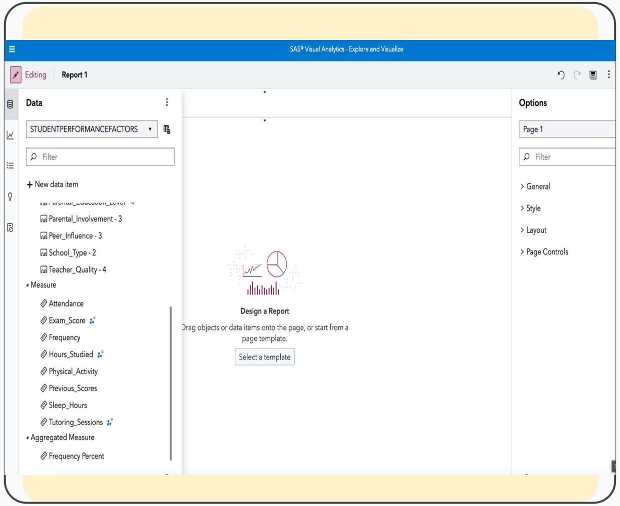
13 categorical variables, 7 numerical variables in *Explore and Visualize* menu.



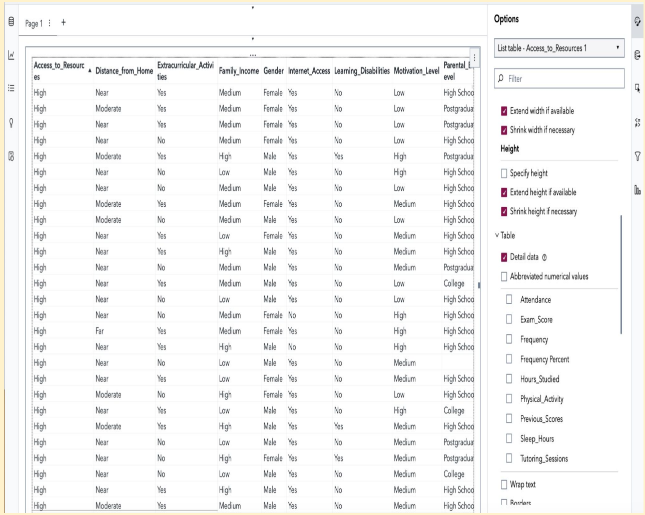


STUDENTPERFORMANCEFACTORS Columns: 20 Rows: 6.6 K Size: 370.7 KB

#	Name	Label	Data Type	Raw Length	Formatted Length	Format
1	Hours_Studied	--	double	8	12	--
2	Attendance	--	double	8	12	--
3	Parental_Inv...	--	varchar	6	6	--
4	Access_to_R...	--	varchar	6	6	--
5	Extracurricul...	--	varchar	3	3	--
6	Sleep_Hours	--	double	8	12	--
7	Previous_Sc...	--	double	8	12	--
8	Motivation_...	--	varchar	6	6	--
9	Internet_Acc...	--	varchar	3	3	--
10	Tutoring_Se...	--	double	8	12	--
11	Family_Inco...	--	varchar	6	6	--
12	Teacher_Qu...	--	varchar	6	6	--
13	School_Type	--	varchar	7	7	--



DATA WRANGLING



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

Options

List table - Access_to_Resources 1 ▾

Filter

☑ Extend width if available

☑ Shrink width if necessary

Height

☐ Specify height

☑ Extend height if available

☑ Shrink height if necessary

∨ Table

☑ Detail data

☐ Abbreviated numerical values

☐ Attendance

☐ Exam_Score

☐ Frequency

☐ Frequency Percent

☐ Hours_Studied

☐ Physical_Activity

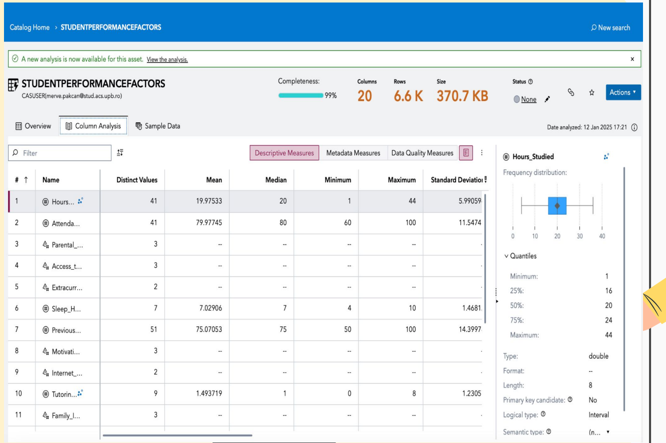
☐ Previous_Scores

☐ Sleep_Hours

☐ Tutoring_Sessions

☐ Wrap text

☐ Borders



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

Date analyzed: 12 Jan 2025 17:21

# ↑	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:

∨ Quantiles

Minimum: 1

25%: 16

50%: 20

75%: 24

Maximum: 44

Type: double

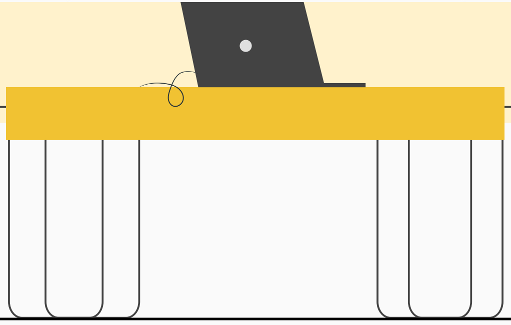
Format: --

Length: 8

Primary key candidate: No

Logical type: Interval

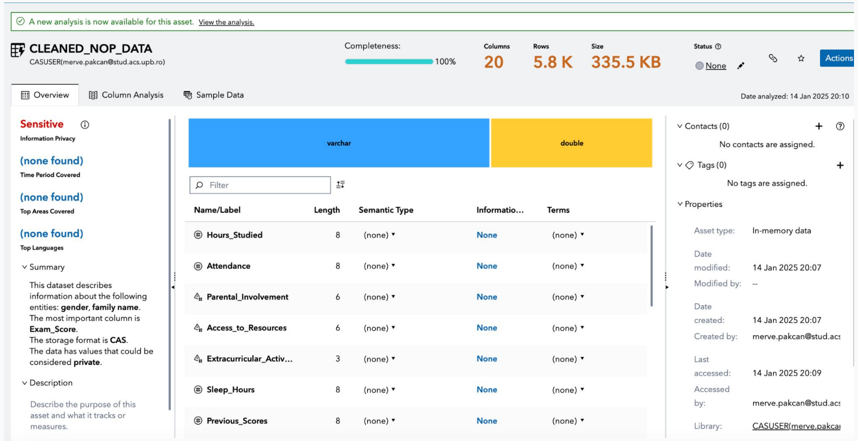
Semantic type: (n... ▾



List table created, **no new data items** are **needed**.Data analyzed in Discover Information Assets part. Cross Tabulation is not proper due to high number of categorical variables

5836 ROWS

Uploaded clean data to discover information assets part, shows that data is clean to analyze



A new analysis is now available for this asset. View the analysis.

CLEANED_NOP_DATA
CASUSER(merve.pakcan@stud.acs.upb.ro)

Completeness:	Columns	Rows	Size	Status
100%	20	5.8 K	335.5 KB	None

Date analyzed: 14 Jan 2025 20:10

Sensitive
Information Privacy

(none found)
Time Period Covered

(none found)
Top Areas Covered

(none found)
Top Languages

Summary

This dataset describes information about the following entities: **gender**, **family name**. The most important column is **Exam_Score**. The storage format is **CAS**. The data has values that could be considered **private**.

Description

Describe the purpose of this asset and what it tracks or measures.

Name/Label	Length	Semantic Type	Informatio...	Terms
Hours_Studied	8	(none)	None	(none)
Attendance	8	(none)	None	(none)
Parental_Involvement	6	(none)	None	(none)
Access_to_Resources	6	(none)	None	(none)
Extracurricular_Activ...	3	(none)	None	(none)
Sleep_Hours	8	(none)	None	(none)
Previous_Scores	8	(none)	None	(none)

Contacts (0)

No contacts are assigned.

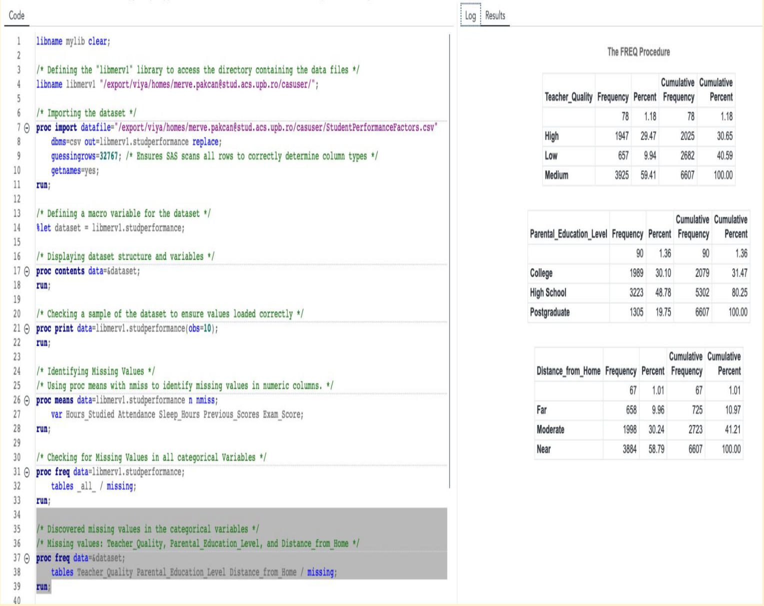
Tags (0)

No tags are assigned.

Properties

Asset type:	In-memory data
Date modified:	14 Jan 2025 20:07
Modified by:	--
Date created:	14 Jan 2025 20:07
Created by:	merve.pakcan@stud.acs
Last accessed:	14 Jan 2025 20:09
Accessed by:	merve.pakcan@stud.acs
Library:	CASUSER(merve.pakca

DATA PREPARATION



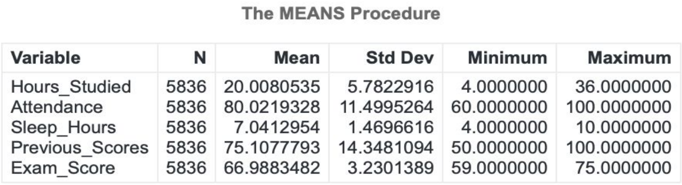
```
libname mylib clear;

/* 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;

```

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

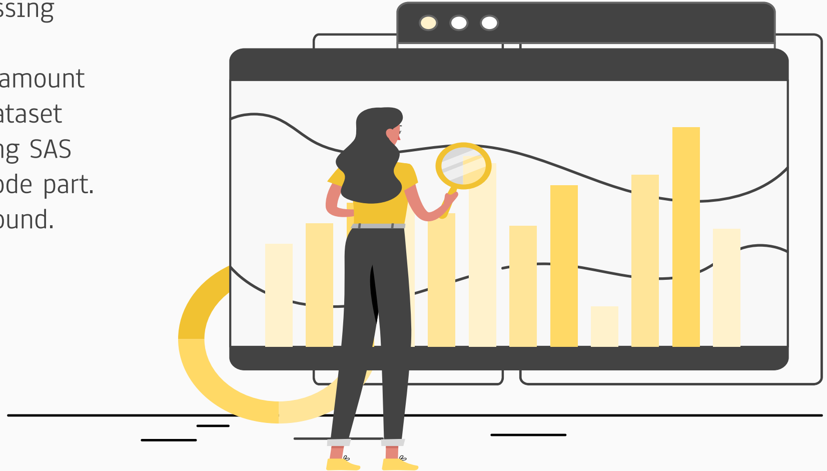
Distance_from_Home	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	67	1.01	67	1.01
Far	658	9.96	725	10.97
Moderate	1998	30.24	2723	41.21
Near	3884	58.79	6607	100.00



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

Eliminated missing values and outliers(small amount compared to dataset size), with using SAS Develop and Code part. No duplicate found.

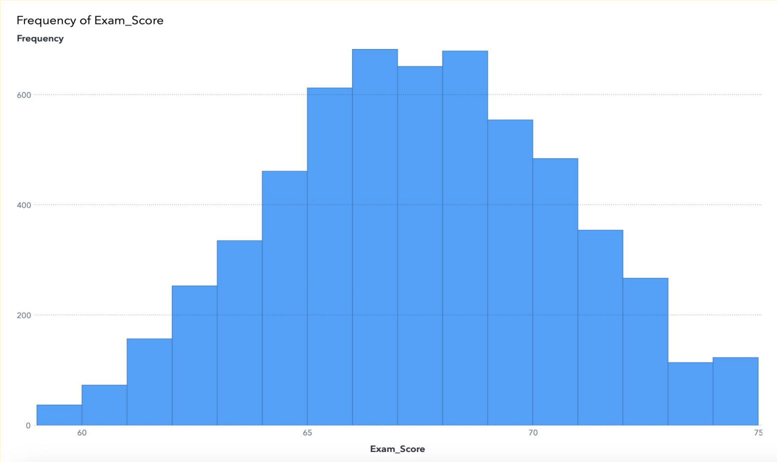


RESEARCH METHODOLOGY

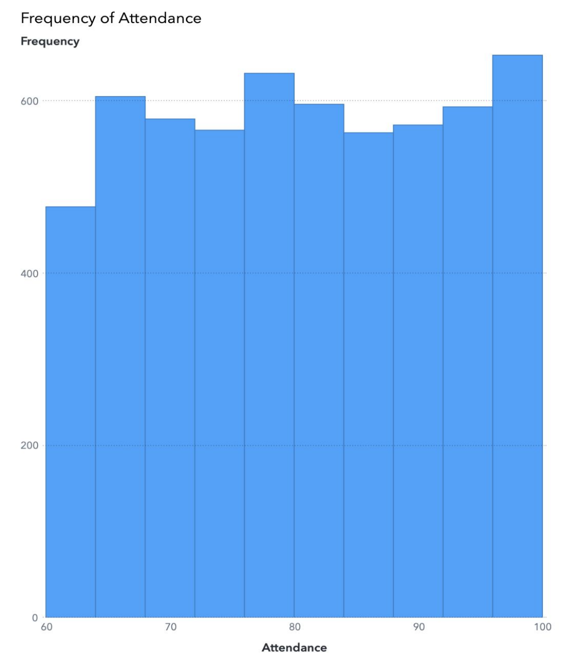
QUESTION	SUB-QUESTION	VARIABLES	NOTES
What factors most influence student performance?	Which variables have the strongest correlation with Exam_Score?	Dependent Variable: Exam_Score Independent Variable: Attendance, Hours_Studied, Previous_Scores, Parental_Involvement, Access_to_Resources, Extracurricular_Activities, 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.

FIRST EXPLORATORY DATA ANALYSIS

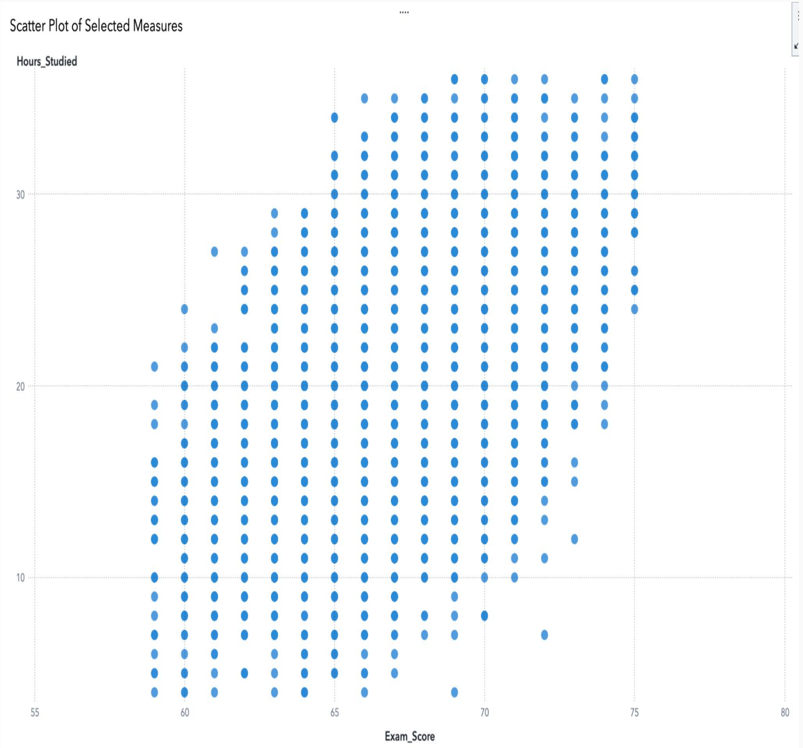
Target Variable



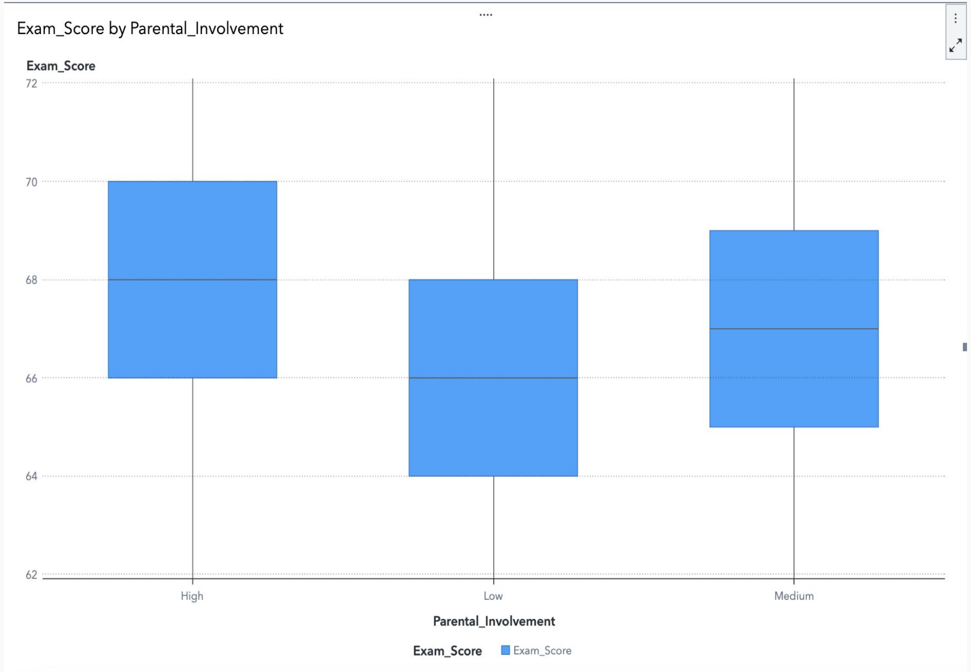
Shape: Bell-shaped distribution, **Center:**Around 67, **Spread:**Between 59 and 75



Uniform distribution



Positive correlation between hours studied and exam scores: as study hours increase, exam scores tend to increase

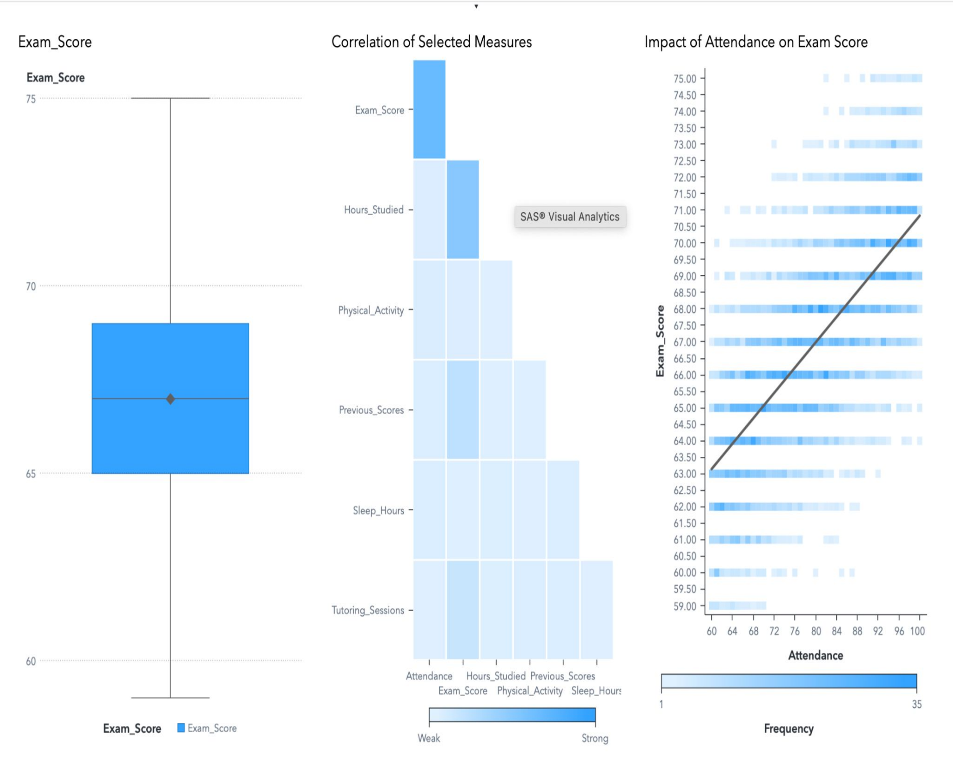


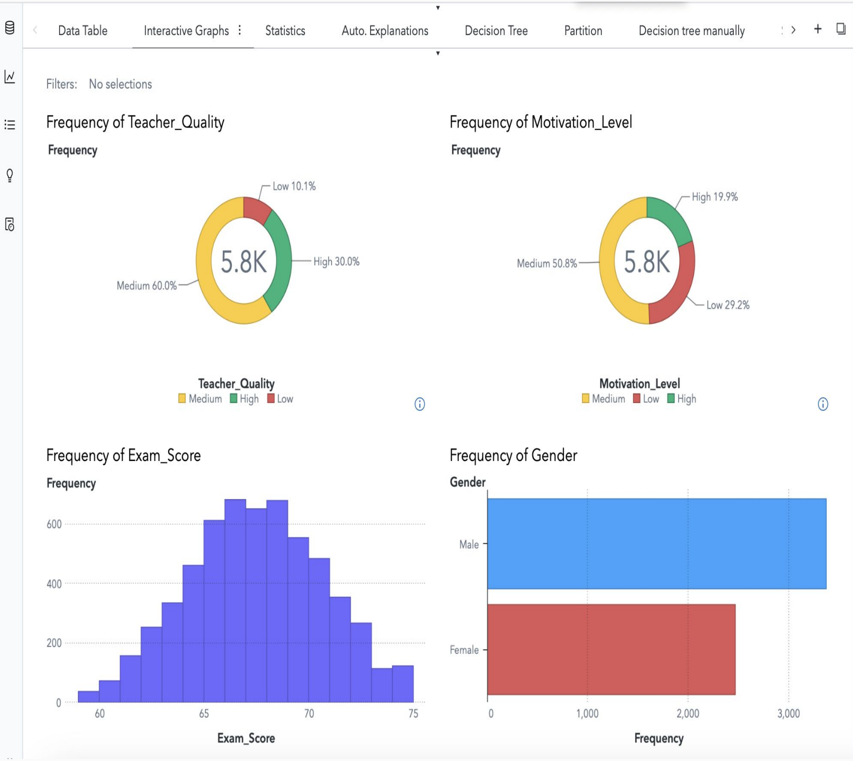
Positive correlation between parental engagement and academic success

INVESTIGATE RELATIONSHIPS IN DATA

The distribution of **Exam_Score** shows a median around 67. The range is relatively narrow. In correlation matrix and heatmap, a strong linear relationship exists between **Attendance** and **Exam_Score**

Additional interactive graphs created to understand variables more.



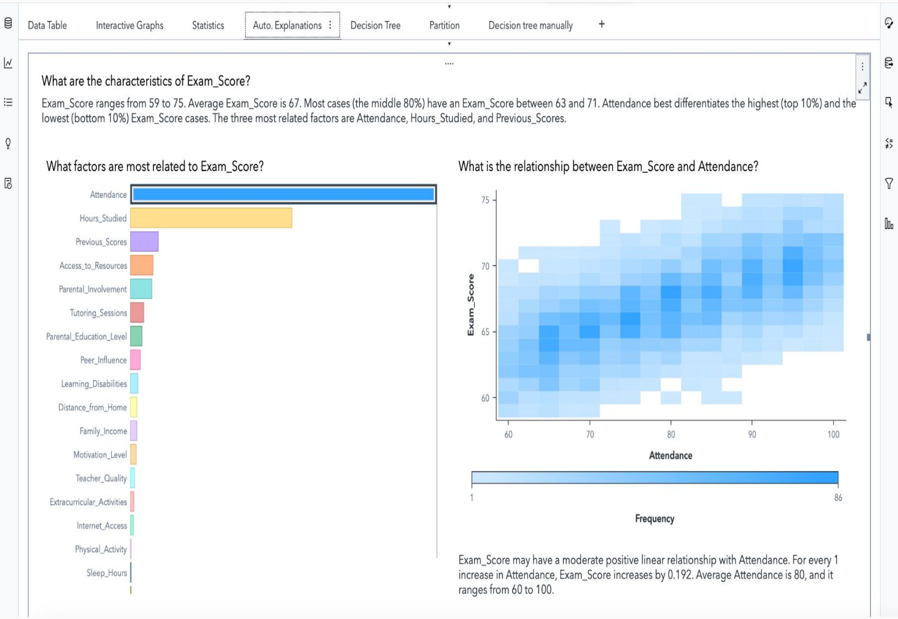


Filters: No selections

FOR QUESTION I

What factors most influence student performance?

FIRST STEP TO BUILD A DECISION



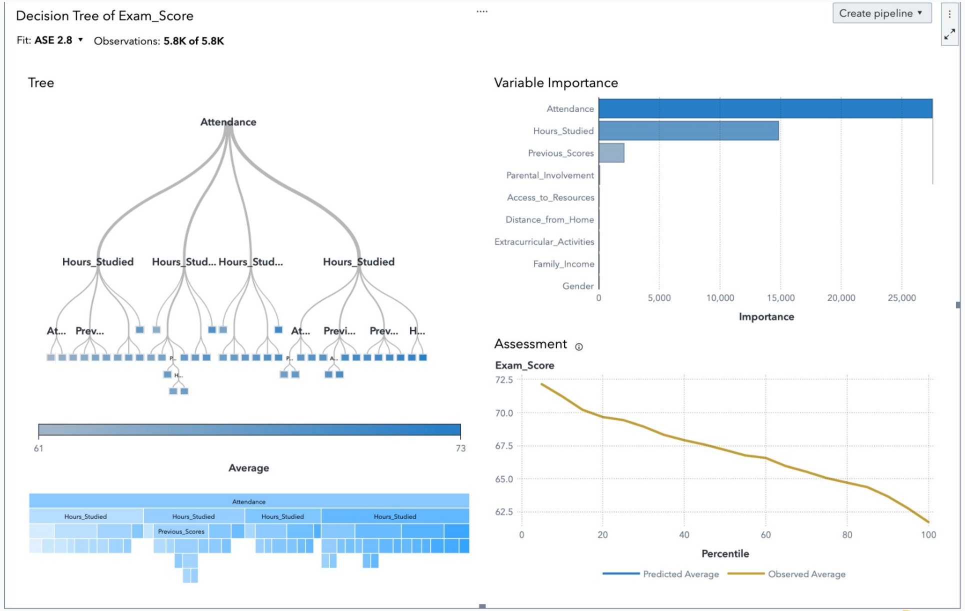
What are the characteristics of Exam_Score?

Exam_Score ranges from 59 to 75. Average Exam_Score is 67. Most cases (the middle 80%) have an Exam_Score between 63 and 71. Attendance best differentiates the highest (top 10%) and the lowest (bottom 10%) Exam_Score cases. The three most related factors are Attendance, Hours_Studied, and Previous_Scores.

What factors are most related to Exam_Score?

What is the relationship between Exam_Score and Attendance?

Exam_Score may have a moderate positive linear relationship with Attendance. For every 1 increase in Attendance, Exam_Score increases by 0.192. Average Attendance is 80, and it ranges from 60 to 100.



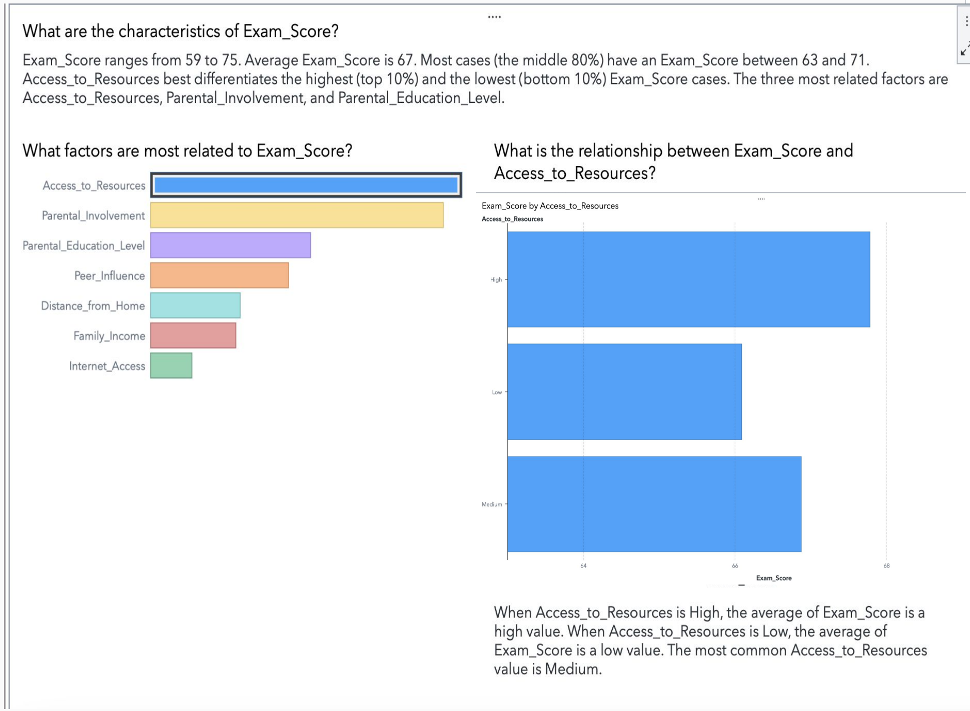
Decision Tree of Exam_Score

Fit: ASE 2.8 Observations: 5.8K of 5.8K

As it seen from automatic explanations and decision tree, the three most related factors are **Attendance**, Hours_Studied, Previous_Score

FOR QUESTION 2

What socioeconomic factors most impact exam results?



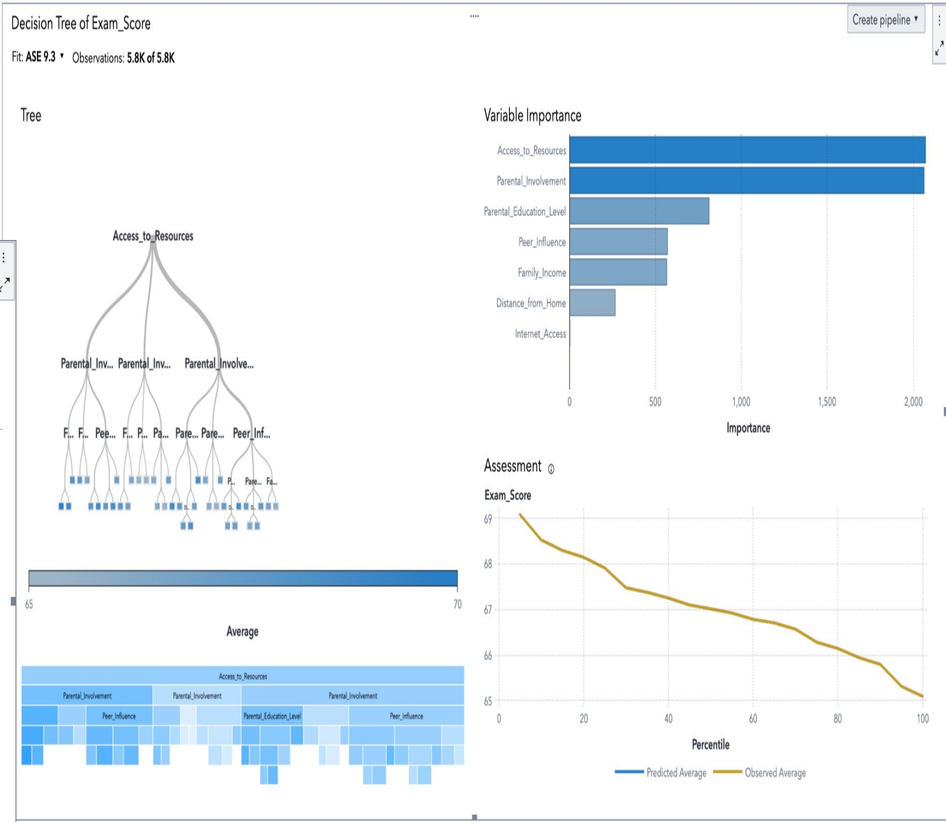
What are the characteristics of Exam_Score?

Exam_Score ranges from 59 to 75. Average Exam_Score is 67. Most cases (the middle 80%) have an Exam_Score between 63 and 71. Access_to_Resources best differentiates the highest (top 10%) and the lowest (bottom 10%) Exam_Score cases. The three most related factors are Access_to_Resources, Parental_Involvement, and Parental_Education_Level.

What factors are most related to Exam_Score?

What is the relationship between Exam_Score and Access_to_Resources?

When Access_to_Resources is High, the average of Exam_Score is a high value. When Access_to_Resources is Low, the average of Exam_Score is a low value. The most common Access_to_Resources value is Medium.



Decision Tree of Exam_Score

Fit: ASE 9.3 ▾ Observations: 5.8K of 5.8K

As it seen from automatic explanations and decision tree, the three most related socioeconomic factors are **Access_to_Resources**, Parental_Involvement, Parental_Education_Level.

THANK YOU!

FINAL CONCLUSIONS

What was easy: Exploring patterns with visualization part
What was challenging:Some technical issues at Sas

4 QUALITY EDUCATION

10 REDUCED INEQUALITIES

QUESTION	FINDINGS	DECISONS
What factors most influence student performance?	**Attendance** and **Hours_Studied** are most related factors	The findings can support policymakers, educators, and institutions in designing targeted programs to improve student performance and reduce educational inequalities.
What socioeconomic factors most impact exam results?	**Access_to_Resources**, **Parental_Involvement** are most related factors	● Empower families through community outreach to enhance parental engagement. ● Promoting attendance and study habits while offering support for students that has low score