Data Mining Project

PART 1: EDA ON GIVEN DATASET

Group Member Roll Number	Group Member Name
19F-0916	Muhammad Abdullah
19F-0917	Zainab Ijaz
19F-0965	Mahnoor
19F-0908	Maria Fatima

Section= SE(8A)

MA'AM NASREEN AKHTAR

Data Mining Project: Part 1 -> EDA Analysis

Table of Contents

Data Mining Project: Part 1 -> EDA Analysis

Introduction:

This document contains the pre-processing phase (part 1) of the Data Mining Project, which contains Python code and data analysis. We've applied different data mining techniques to normalize the data and make it ready for phase-2. We have performed EDA on the given dataset and tried to extract important information from it.

Problem:

To predict students' grades as "pass" or "fail" before: (a) Mid-II, and, (b) Final exams. For Mid-II grade prediction, use the following features: first four assignments, first four quizzes, and Mid-I score; and, for grade prediction before the final exam, use all the features (take the best 5 assignments and quizzes).

Summary of Dataset:

The dataset contains students' assessment scores including <Assignments, Quizzes, Mid- I, Mid-II>, and a predictor variable <Grade>. The data has been anonymized to hide the identities of the students and course(s). The data is shared on seven sheets **(D1 to D7),** where each sheet contains a different number of assignments and quizzes. However, only the best 5 assignments and quizzes are included for each student before calculating their grades. Also, note that total marks for assignments and quizzes are given on the top along with their corresponding weights.

Tools Used:

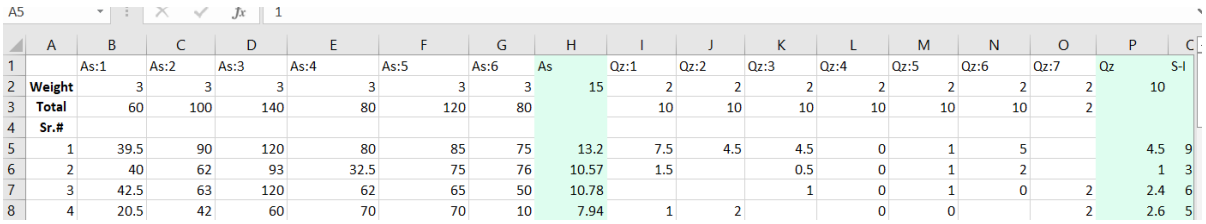
Tools	Purpose
Python	Data normalization using Code.
Orange	Data normalization using Software
MS Word	For generating report

Working Using Python:

We've tried to perform the following tasks to achieve phase 1:

Importing File while considering useful attributes only:

The dataset file starts from 1st row, which shows the name of the columns. 2nd row which shows the weightage of tasks. 3rd column which shows the total marks of the tasks. 4th column is null/free. And main data starts from the 5th column. So, due to this, we've started taking data while neglecting the 2nd, 3rd, and 4th row which contains nothing which can contribute to our analysis.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	C
1		As:1	As:2	As:3	As:4	As:5	As:6	As	Qz:1	Qz:2	Qz:3	Qz:4	Qz:5	Qz:6	Qz:7	Qz	S-I
2	Weight	3	3	3	3	3	3	15	2	2	2	2	2	2	2	10	
3	Total	60	100	140	80	120	80		10	10	10	10	10	10	2		
4	Sr.#																
5	1	39.5	90	120	80	85	75	13.2	7.5	4.5	4.5	0	1	5		4.5	9
6	2	40	62	93	32.5	75	76	10.57	1.5		0.5	0	1	2		1	3
7	3	42.5	63	120	62	65	50	10.78			1	0	1	0	2	2.4	6
8	4	20.5	42	60	70	70	10	7.94	1	2		0	0		2	2.6	5

Data Mining Project: Part 1 -> EDA Analysis

Data Duplication:

We have also performed a data duplication analysis to check the redundant data. Because the dataset is on the student's record hence, we don't need to extra check any duplicity. Yet, we need to check for duplicate serial numbers. There were no duplications in the serial numbers and our data was perfectly passed through this phase.

Null Values:

The given dataset has some missing values, some missing assignments marks, and the same for the quizzes too. There were also some missing values available under the Grade column that also needs to be handled.

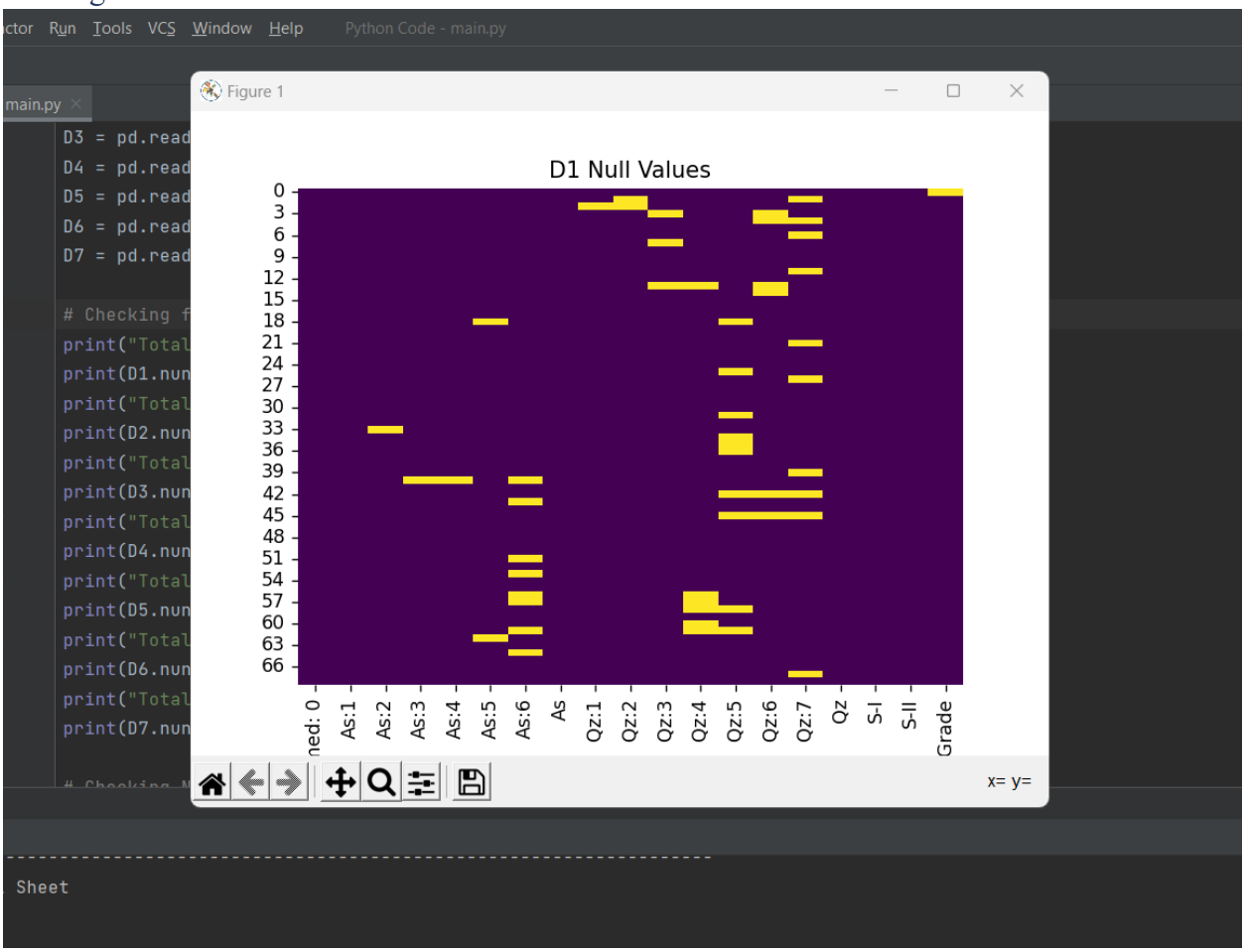
For missing numerical values, we've opted them to be replaced with the median of that column which also handles the outliers' issues.

For missing categorical values, we've opted them to be replaced with the mode of that column.

Displaying Null Values using Heatmap:

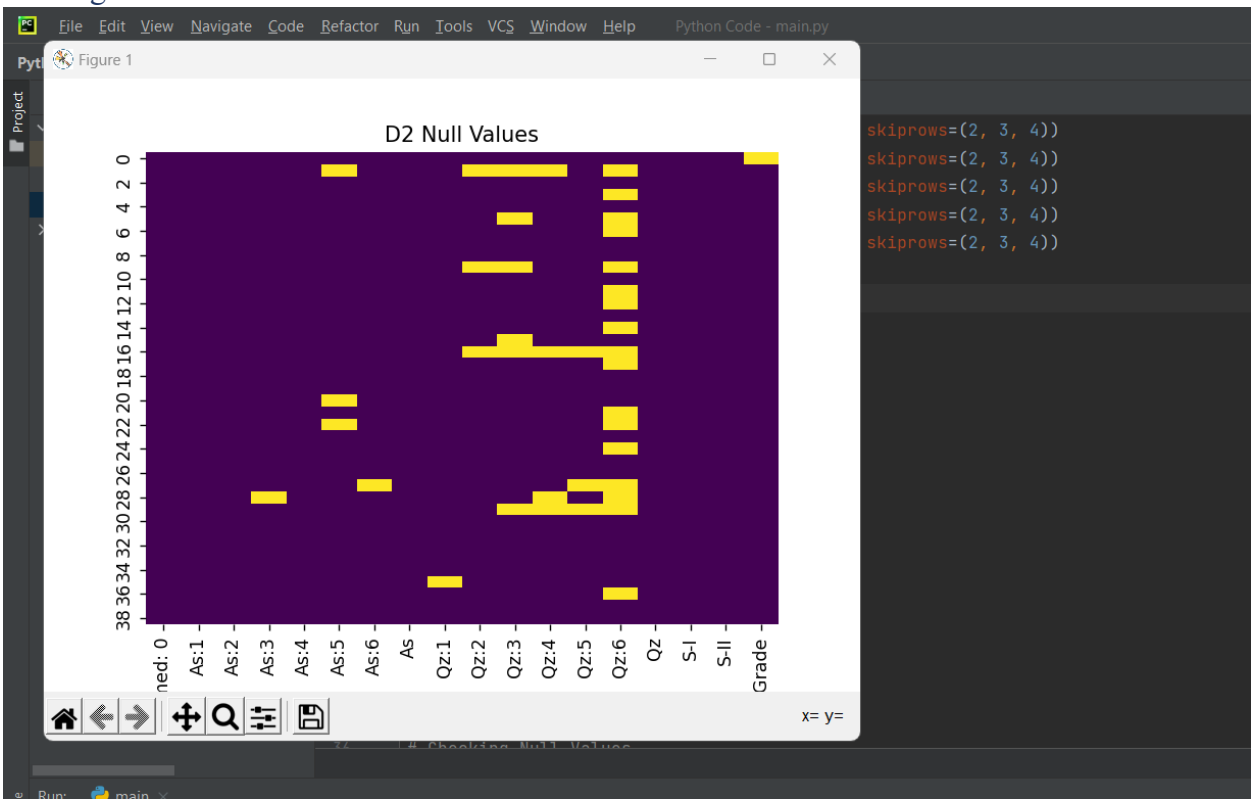
We've used Heatmap to illustrate the missing values inside the dataset and those missing values are as follows:

Missing Values in Sheet D1:

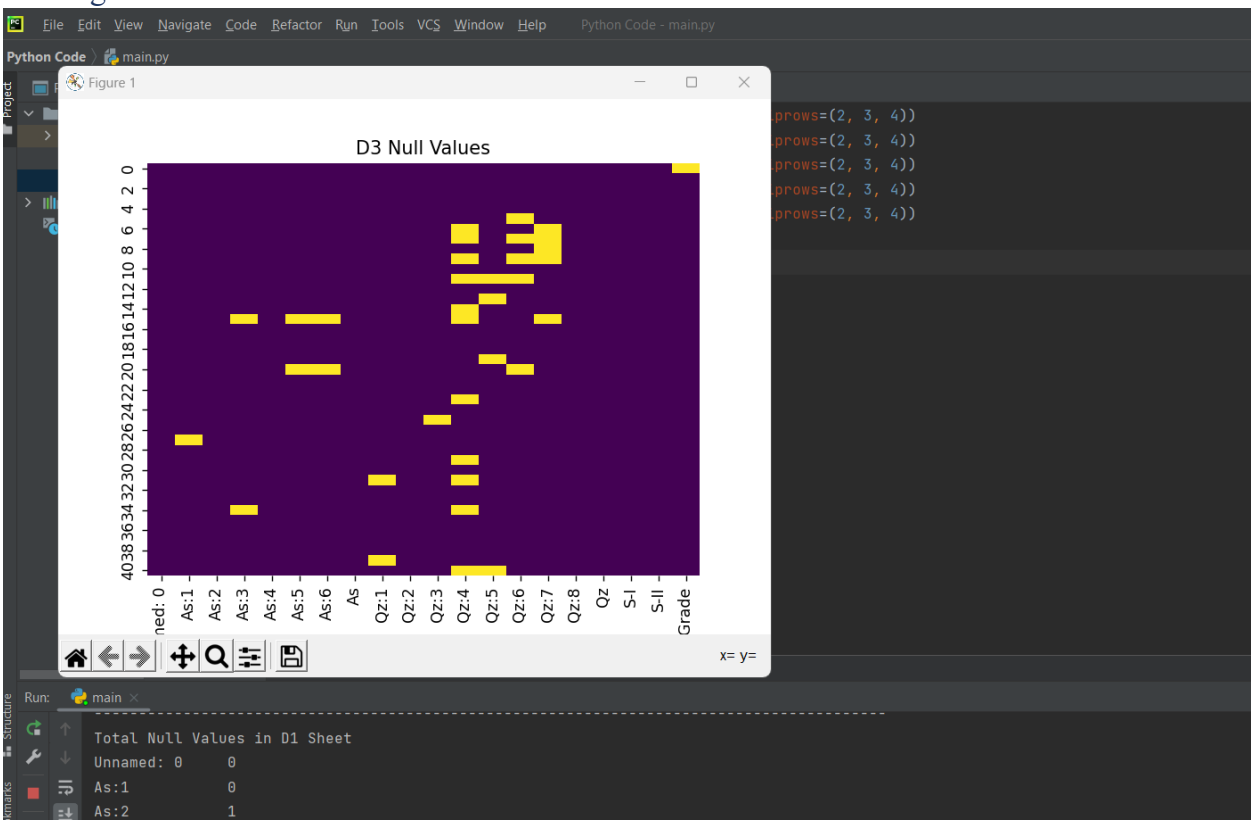


Data Mining Project: Part 1 -> EDA Analysis

Missing Values in Sheet D2:

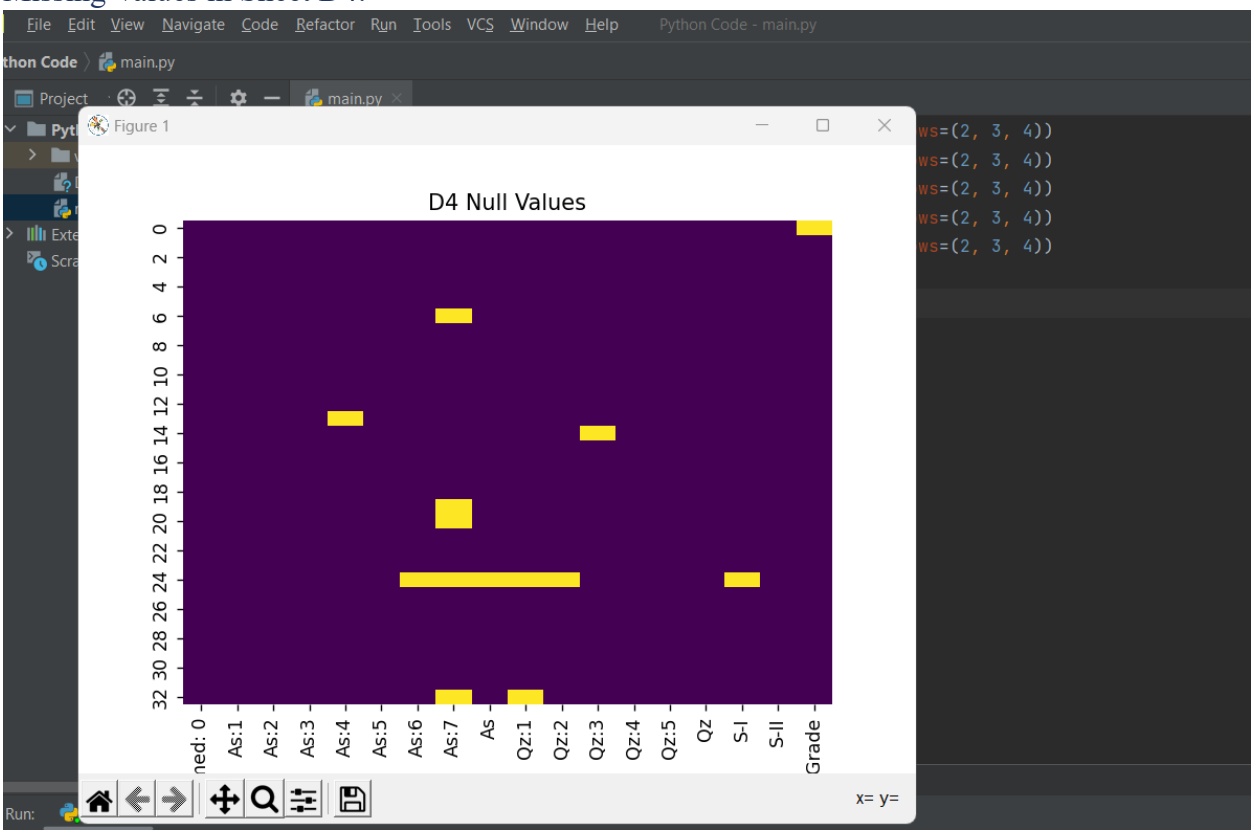


Missing Values in Sheet D3:

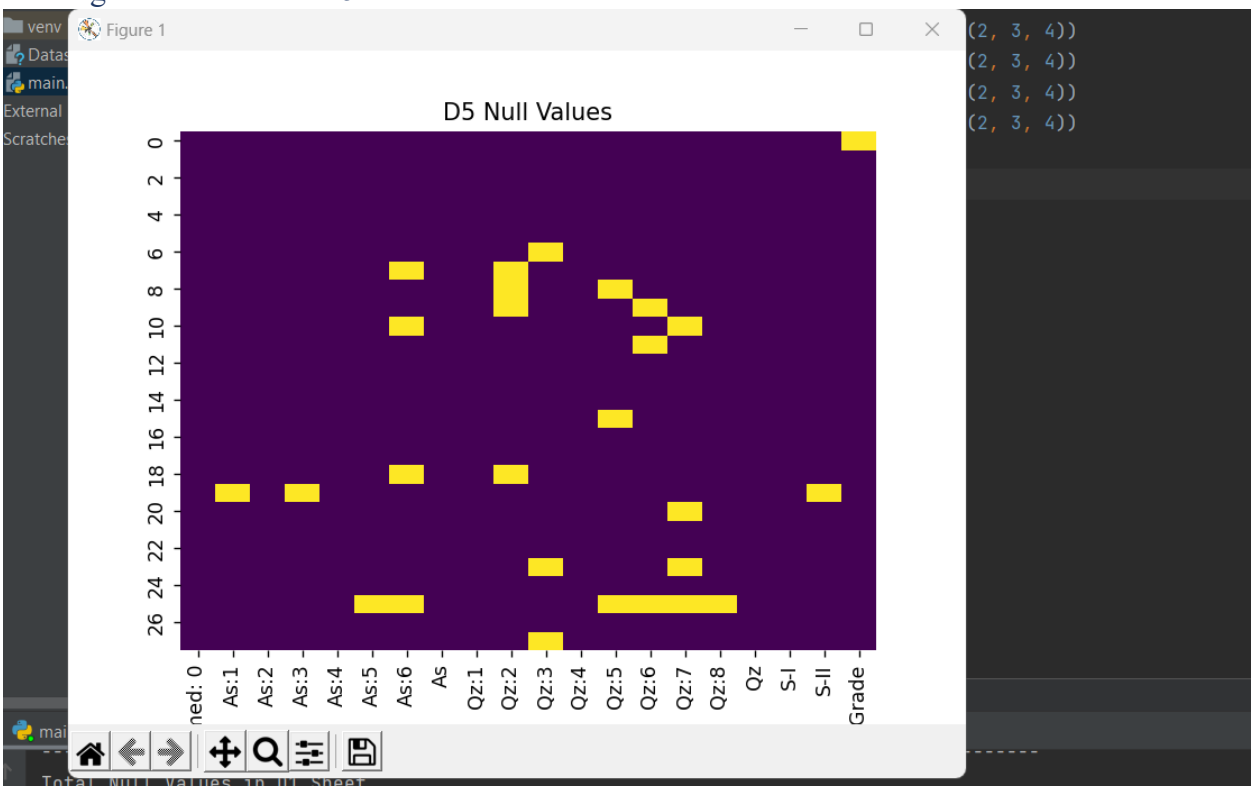


Data Mining Project: Part 1 -> EDA Analysis

Missing Values in Sheet D4:

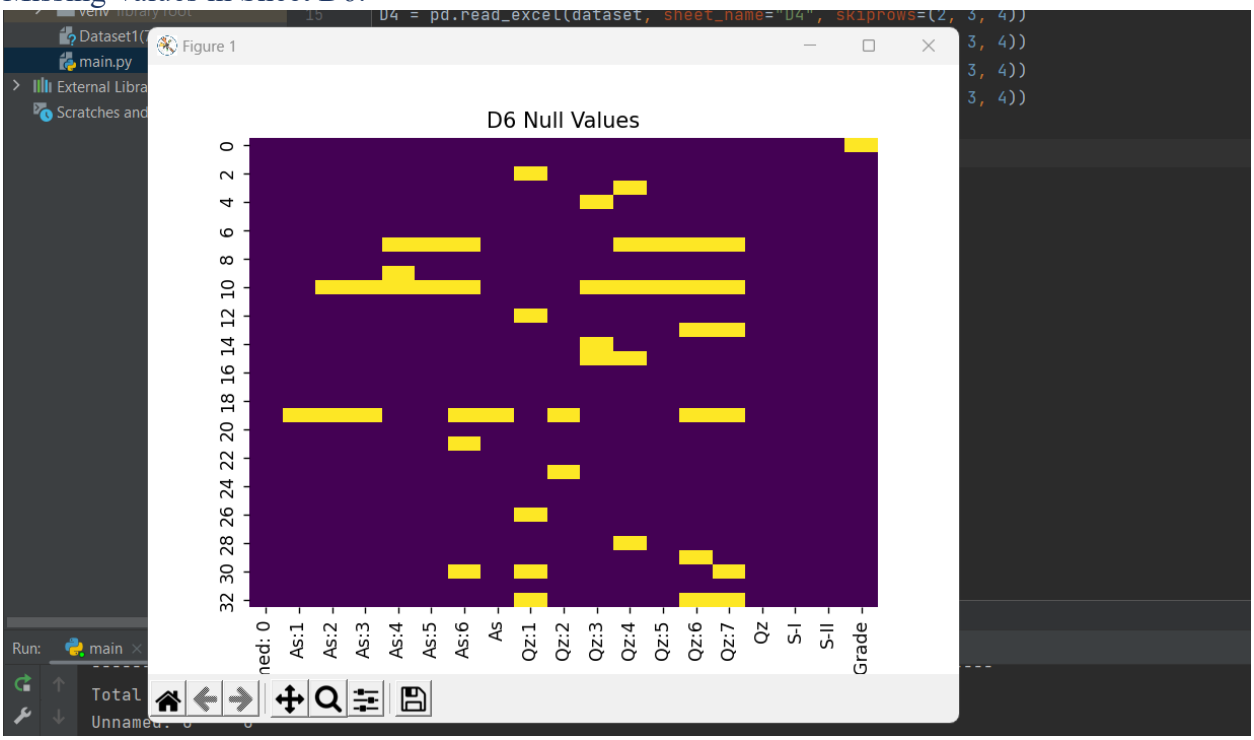


Missing Values in Sheet D5:

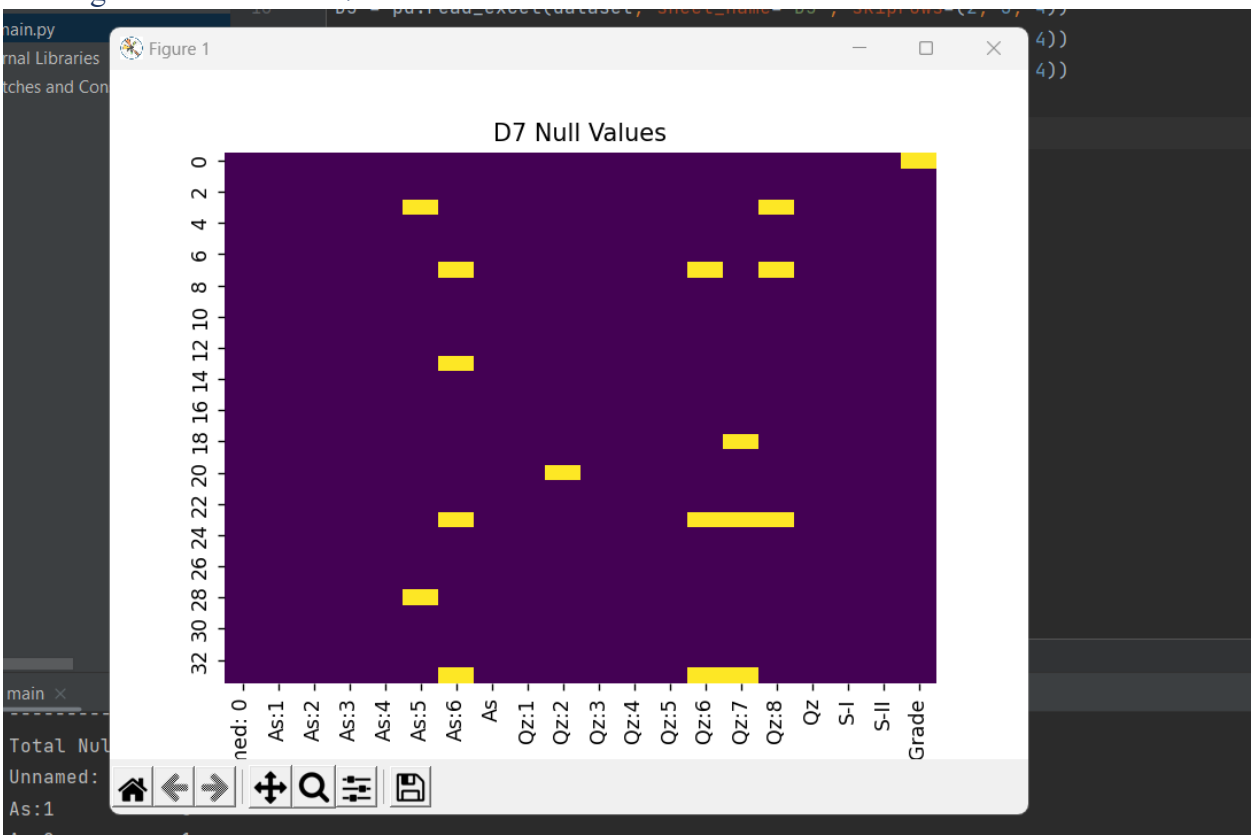


Data Mining Project: Part 1 -> EDA Analysis

Missing Values in Sheet D6:



Missing Values in Sheet D7:



Data Mining Project: Part 1 -> EDA Analysis

Handling Null Values:

As mentioned earlier, we've handled these null values in the following way:

For missing numerical values, we've opted them to be replaced with the median of that column which also handles the outliers' issues.

For missing categorical values, we've opted them to be replaced with the mode of that column.

Data Reduction:

In this phase, we performed data reduction on 1st column of the data set as this was the only column that was not playing any role in our dataset. This was just portraying serial numbers of students.

Apart from 1st column, all of the remaining columns were playing important role in the dataset and they were required for further processing of the data.

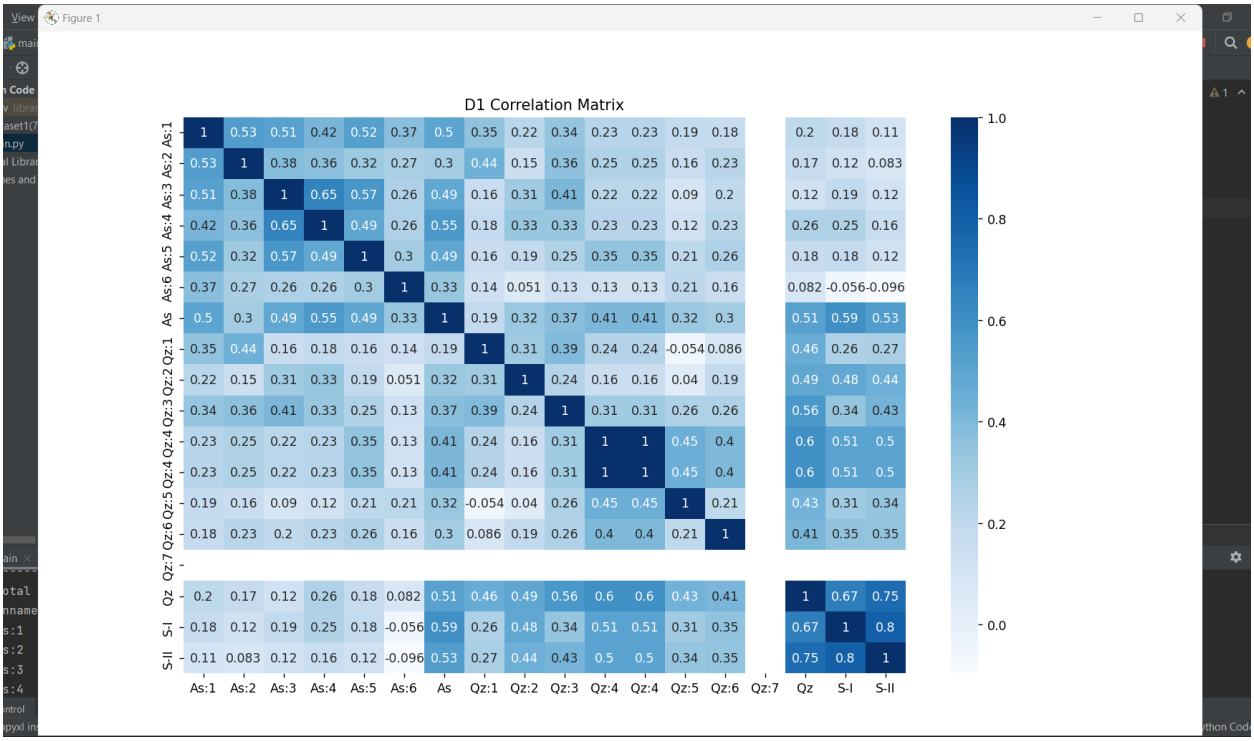
Correlation among attributes:

After data reduction, our data was ready to step on next level which represents the correlation phase. In this phase, we've tried to check out the correlation of attributes (assignments, quizzes, and sessional) to take note of how they are related to each other.

We've used the Seaborn library to illustrate the correlation of attributes in all sheets.

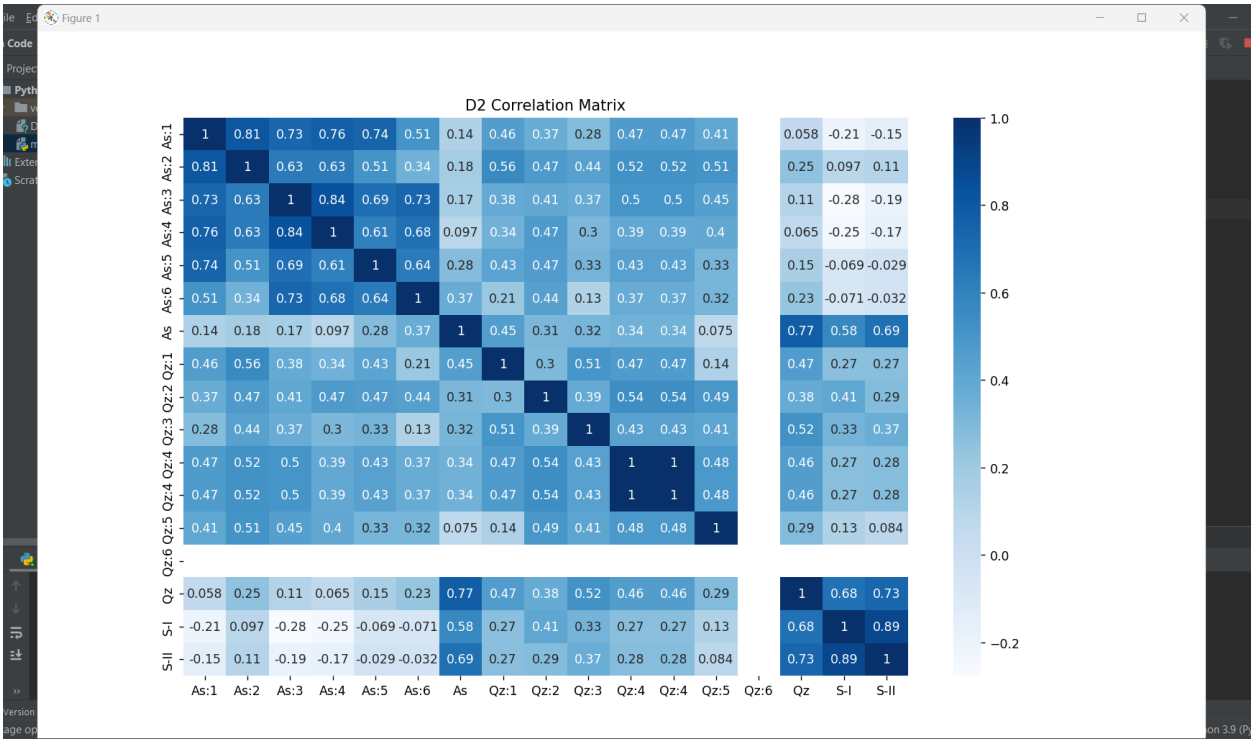
Following are the results of the correlation of attributes:

Correlation in Sheet D1:

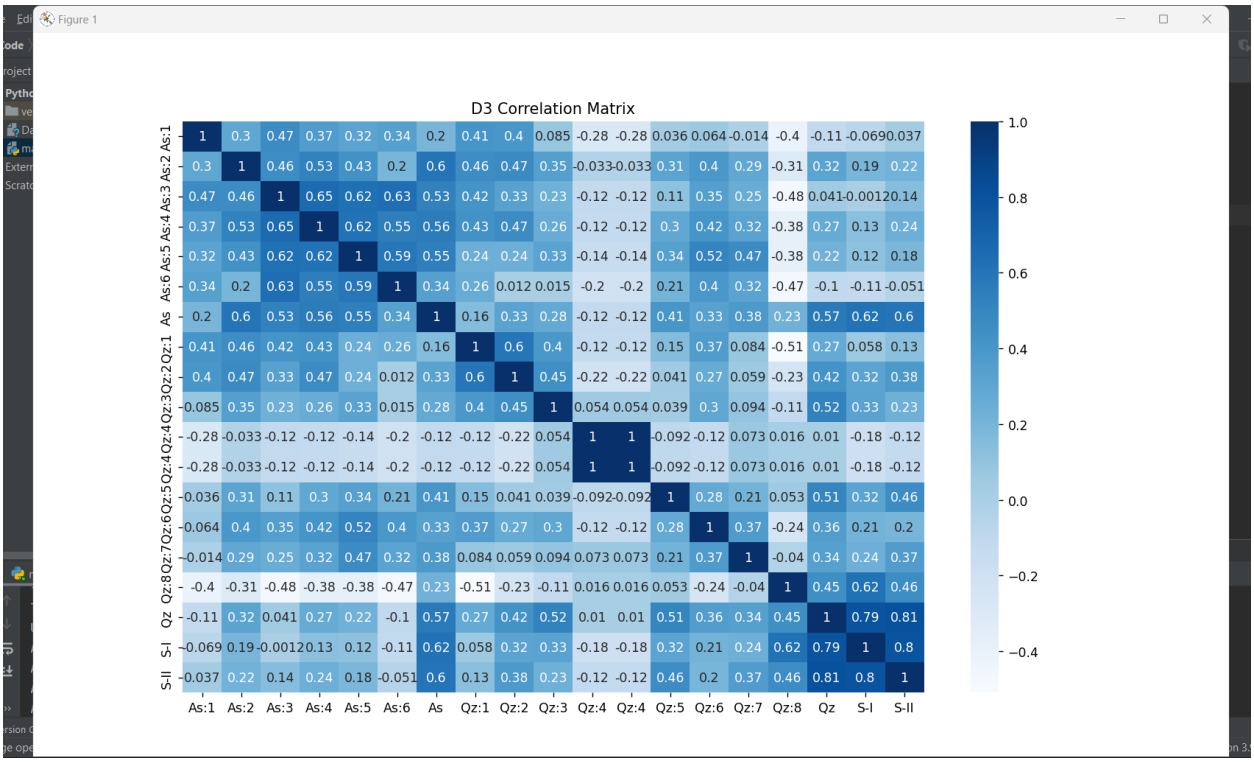


Data Mining Project: Part 1 -> EDA Analysis

Correlation in Sheet D2:

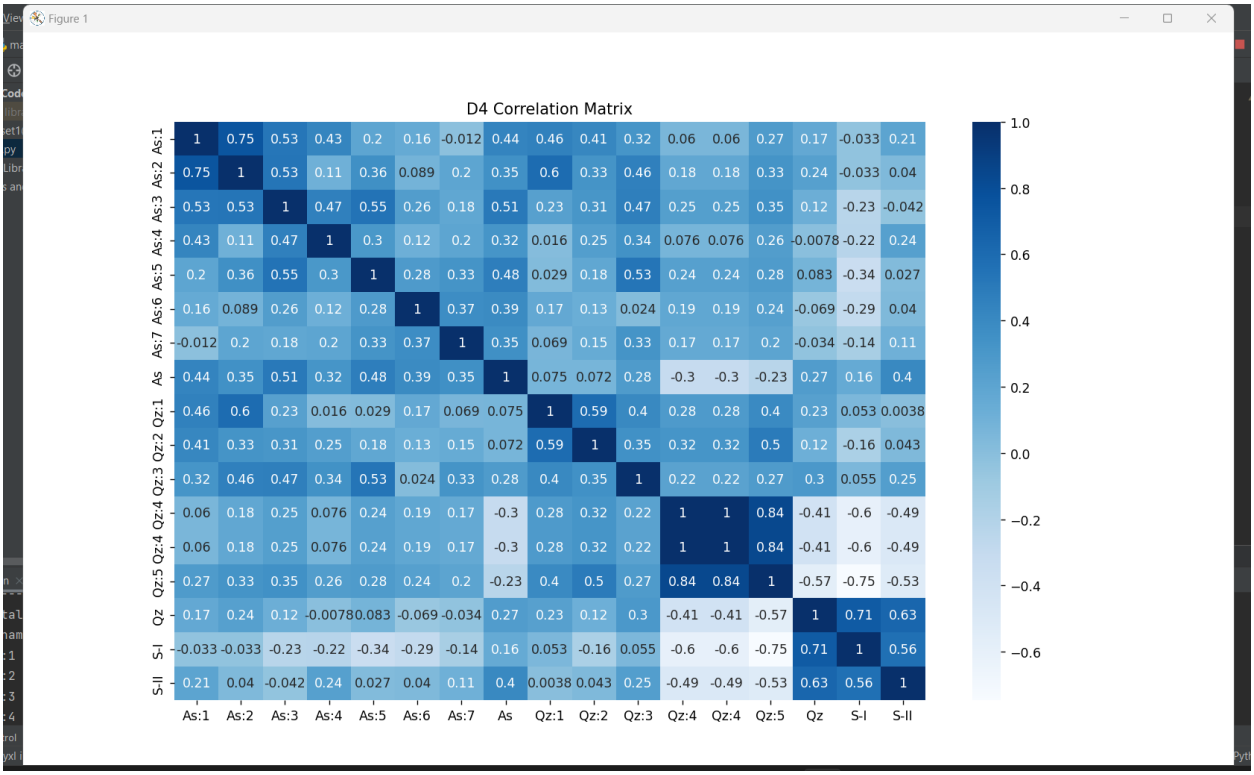


Correlation in Sheet D3:

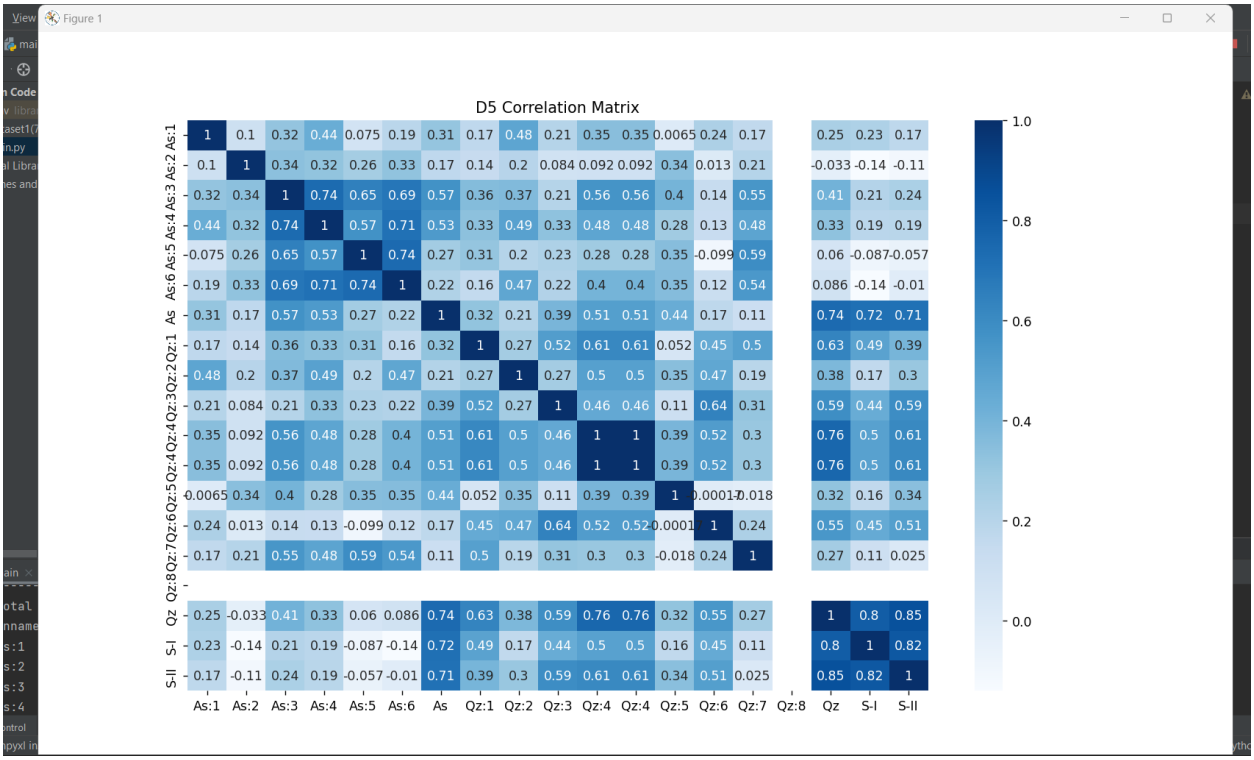


Data Mining Project: Part 1 -> EDA Analysis

Correlation in Sheet D4:

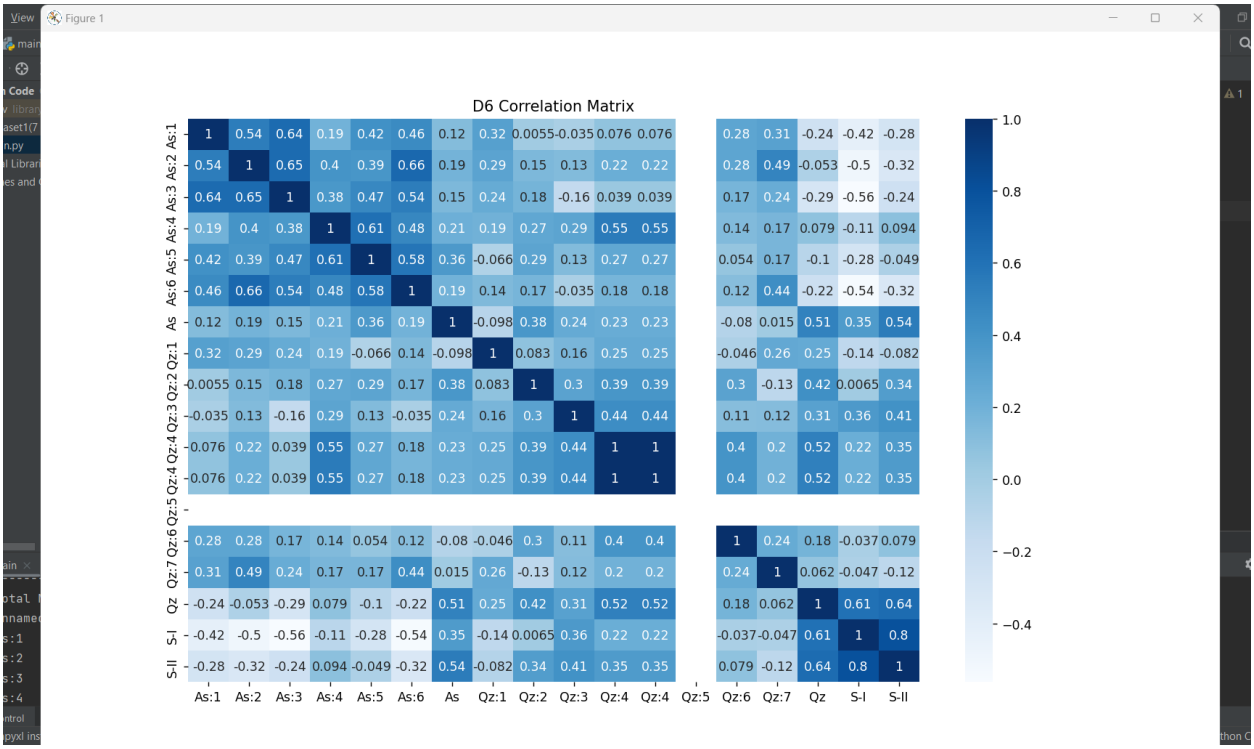


Correlation in Sheet D5:

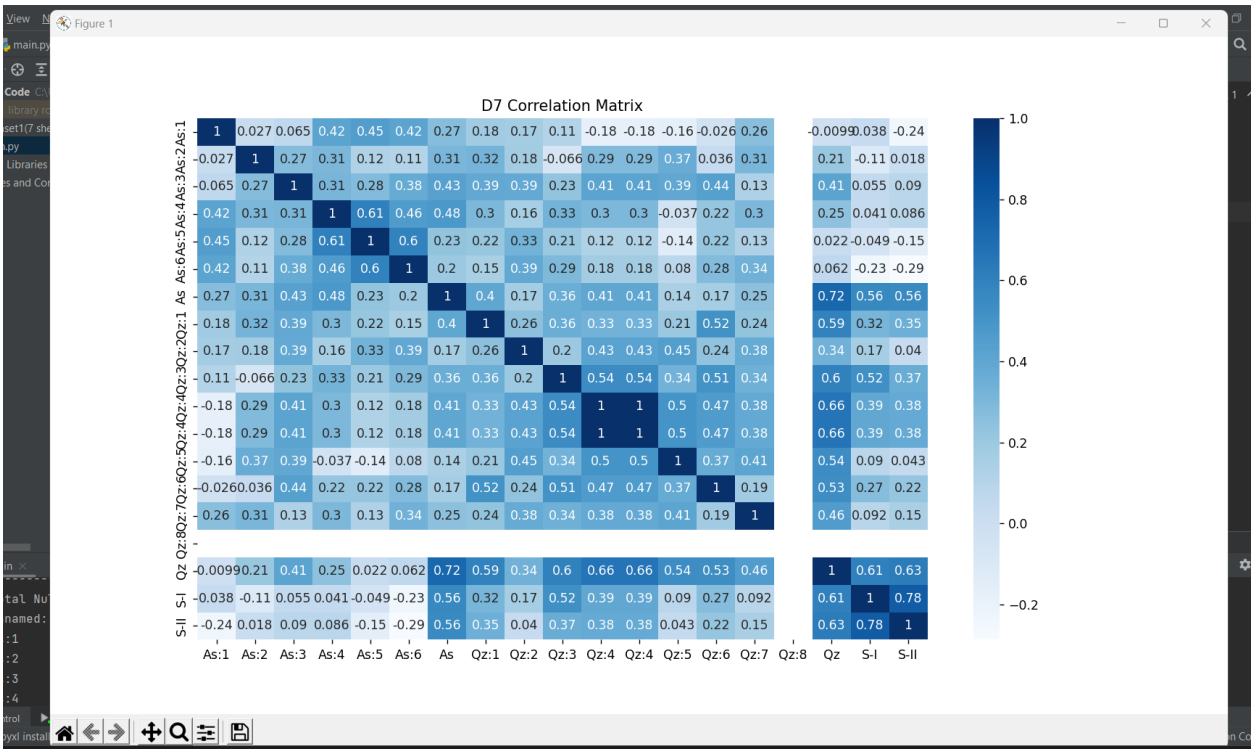


Data Mining Project: Part 1 -> EDA Analysis

Correlation in Sheet D6:



Correlation in Sheet D7:



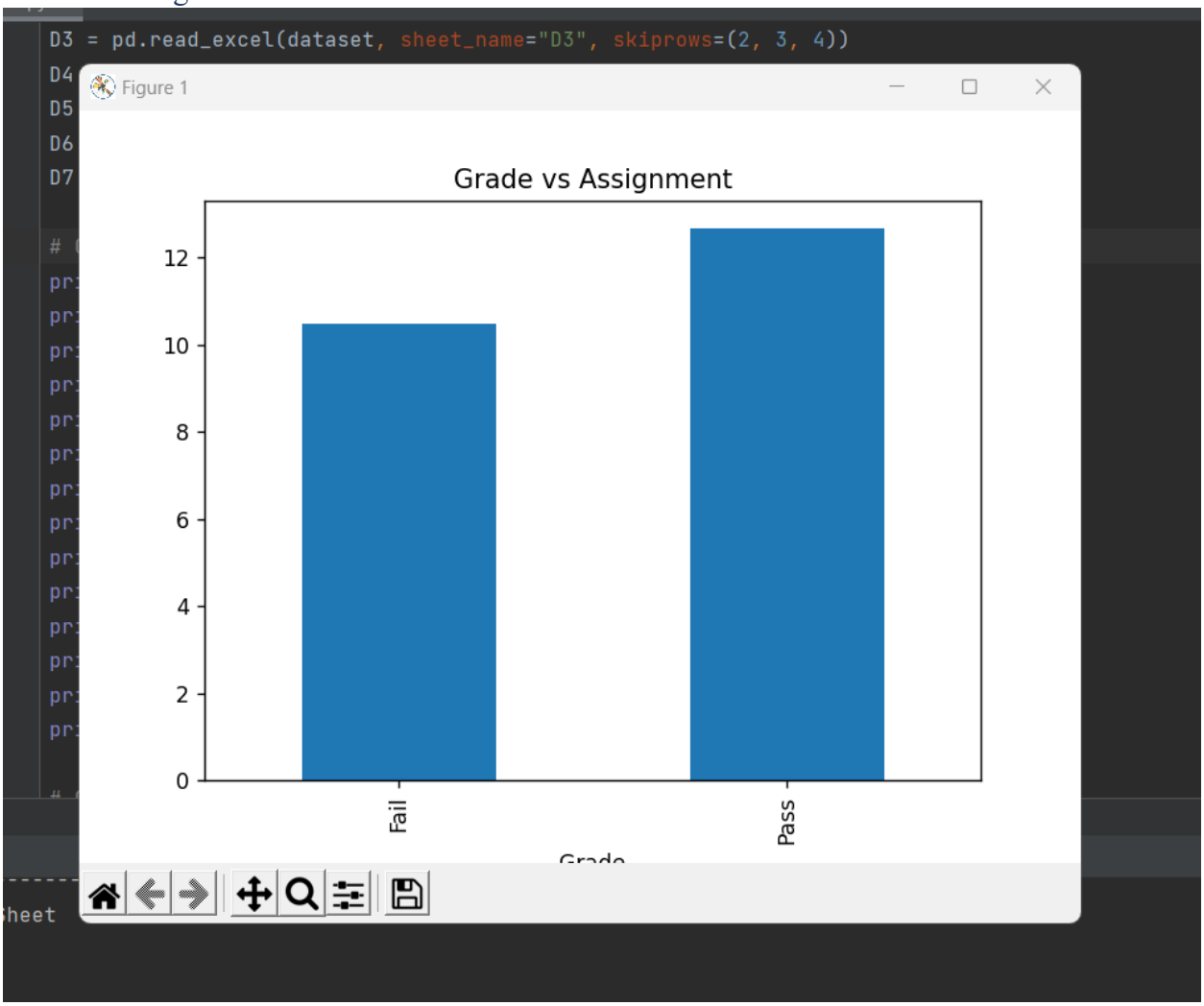
Data Mining Project: Part 1 -> EDA Analysis

Relation of attributes with main class:

After performing correlation among attributes, now it was time to relate attributes with the main class. As our dataset is on the student's record, our main class shows binary values, Fail and Pass. We've tried to map our Grade column with assignments, quizzes, and sessional exams to have an abstract view of their relation.

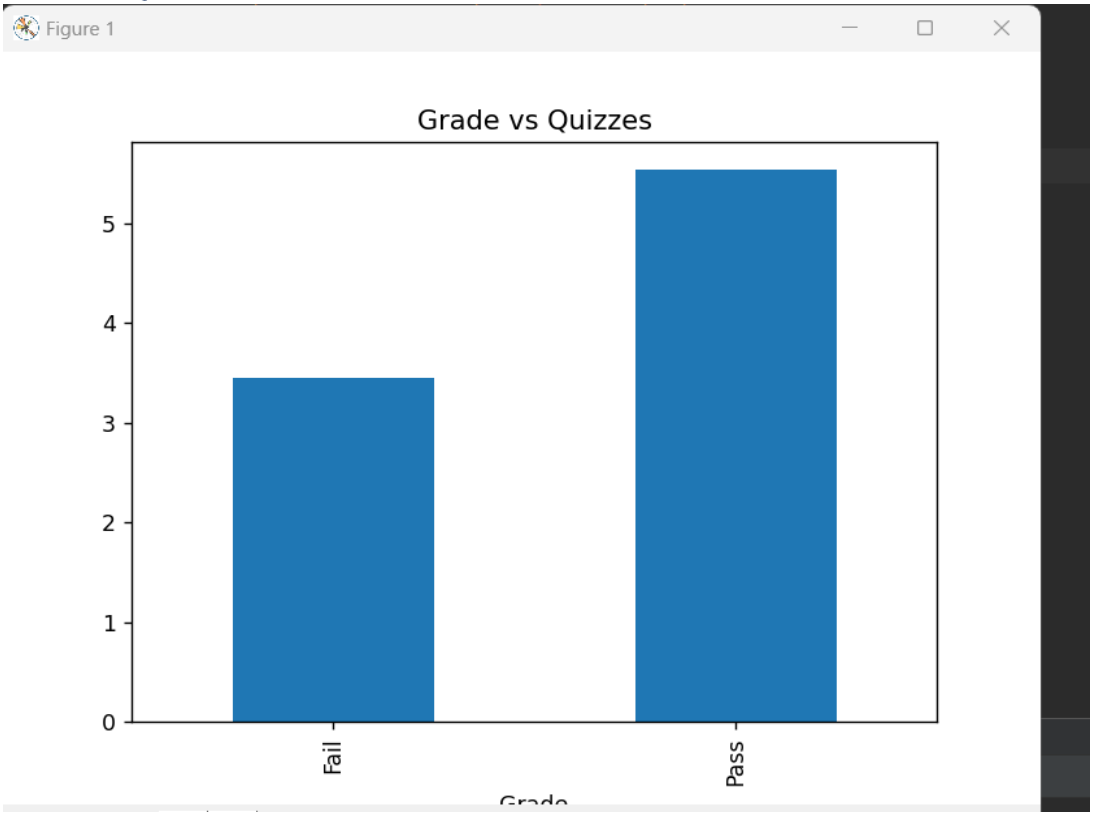
The following images show the relation of Grade with the remaining attributes:

Grade vs Assignments:

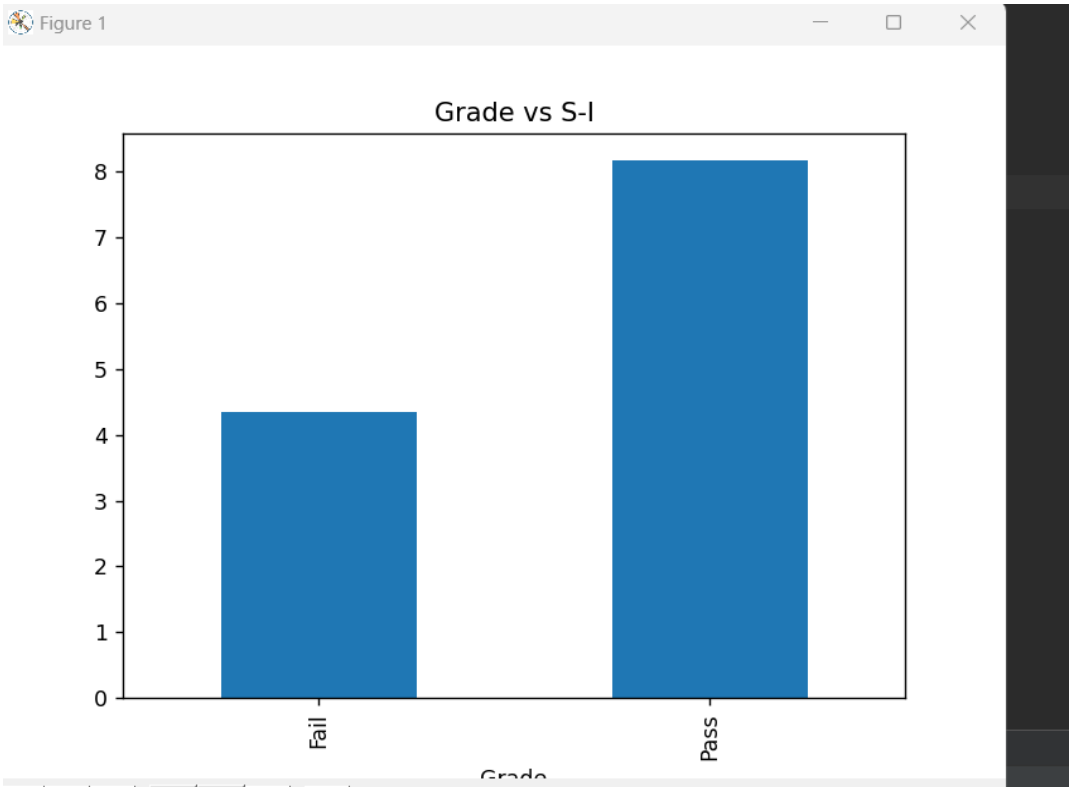


Data Mining Project: Part 1 -> EDA Analysis

Grade vs Quizzes:

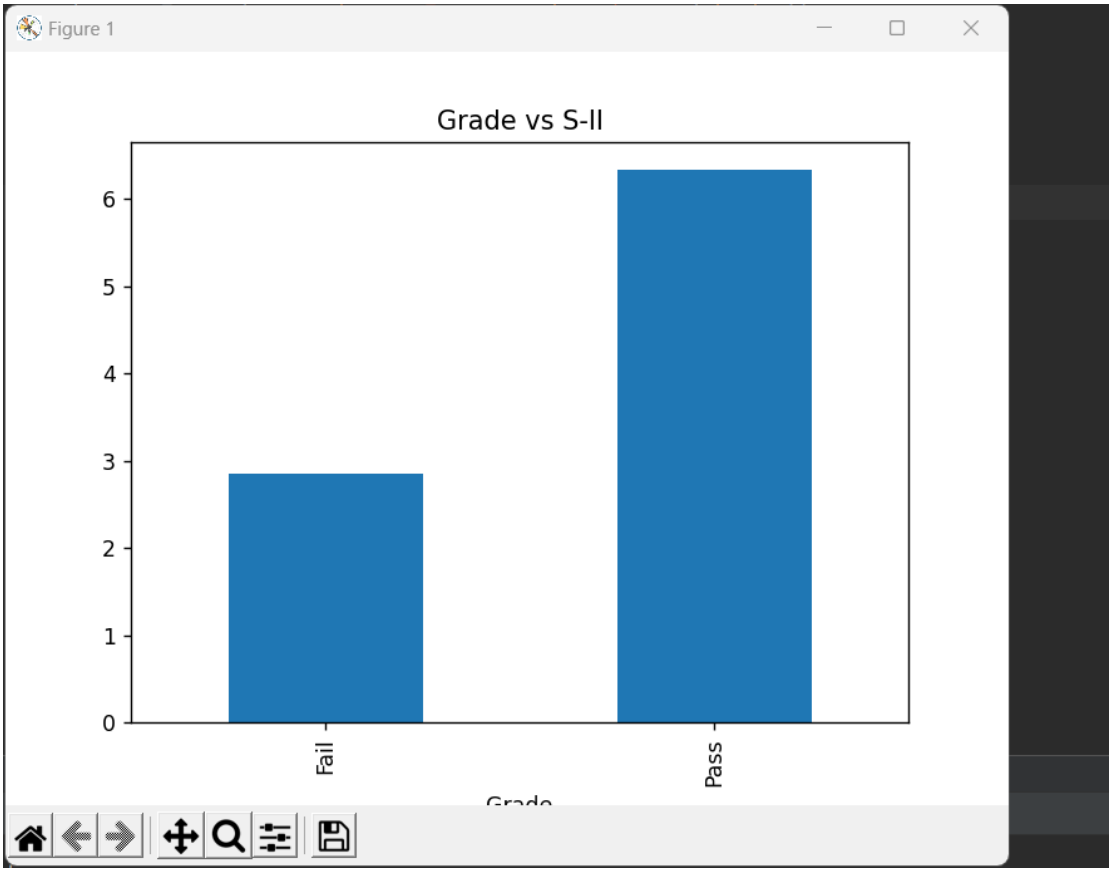


Grade vs Sessional I:



Data Mining Project: Part 1 -> EDA Analysis

Grade vs Sessional II:

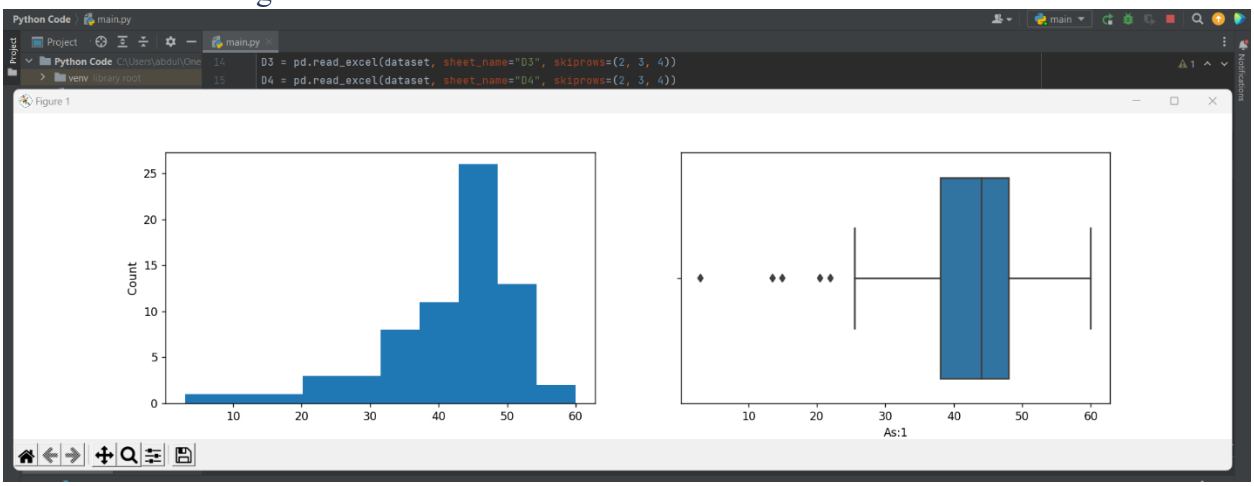


Performing EDA:

After performing all the previous steps, now it was time to move on to EDA. We've performed Univariate EDA on our dataset to extract important information from it, which tells us about the features and their count.

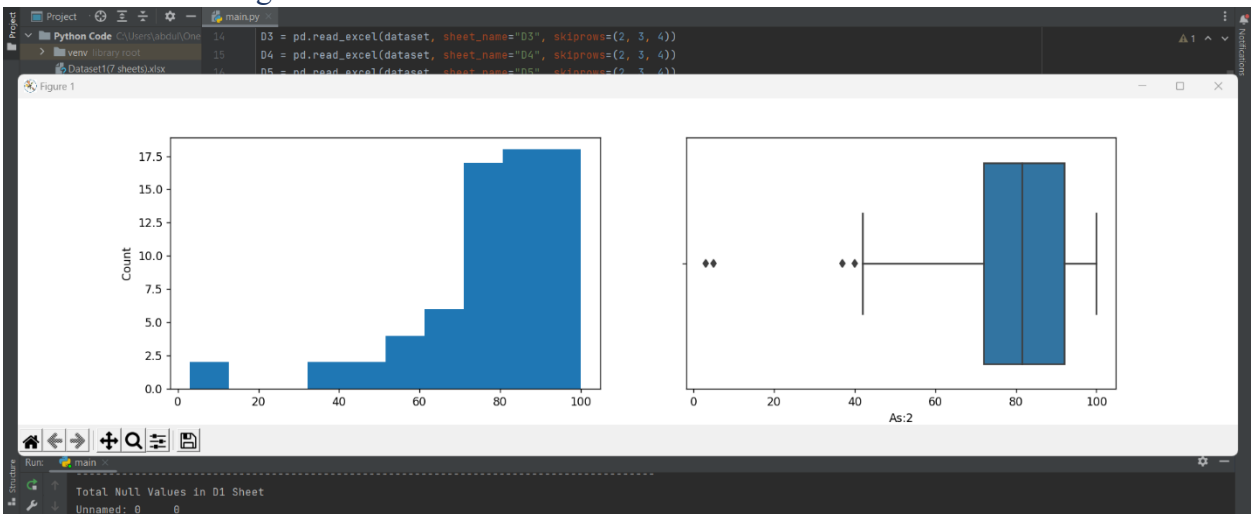
We've only performed it for sheet D1 because the process for the other is the same. It only differs in the results. The following images show the EDA of different attributes:

Sheet D1 EDA Assignment 1:

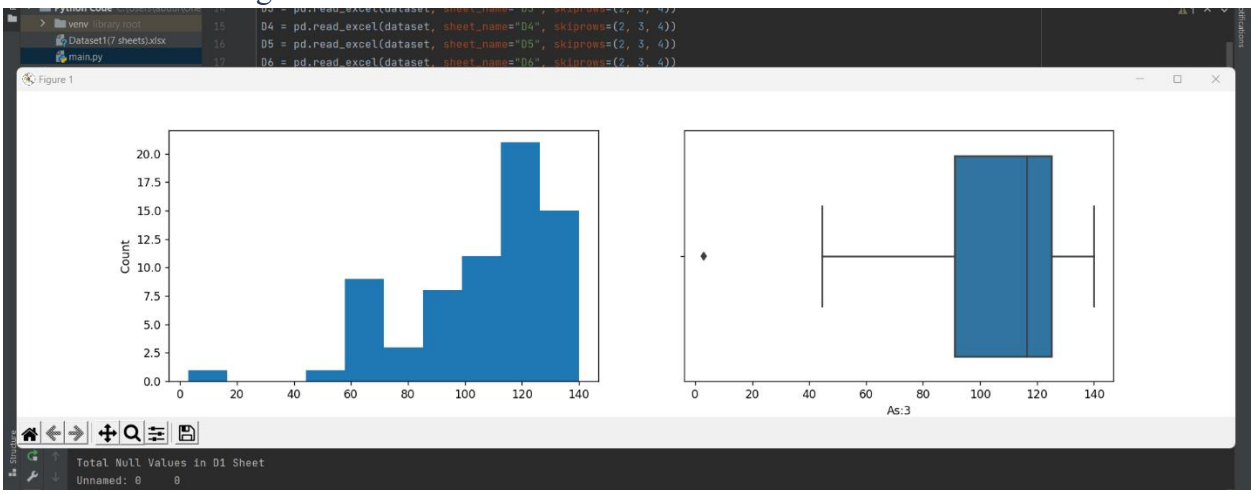


Data Mining Project: Part 1 -> EDA Analysis

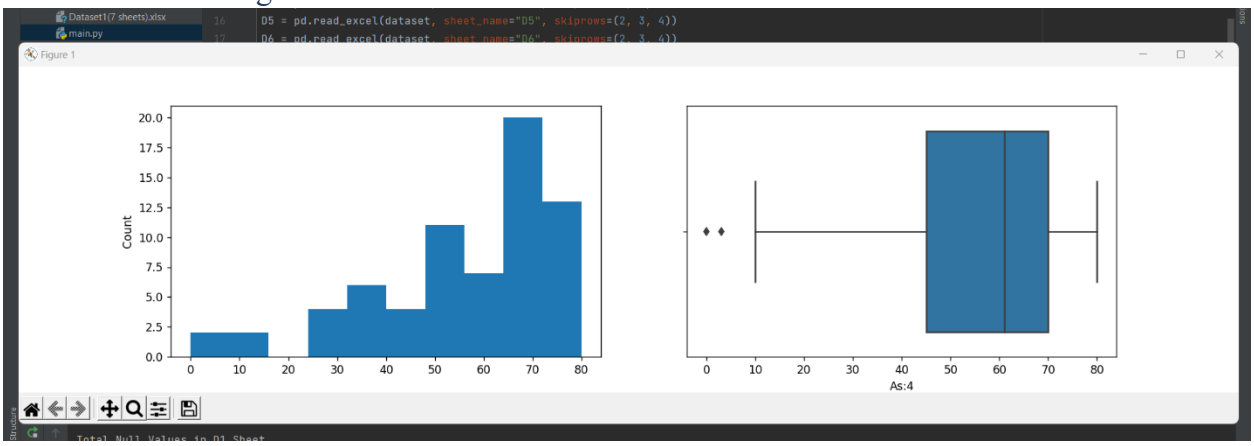
Sheet D1 EDA Assignment 2:



Sheet D1 EDA Assignment 3:

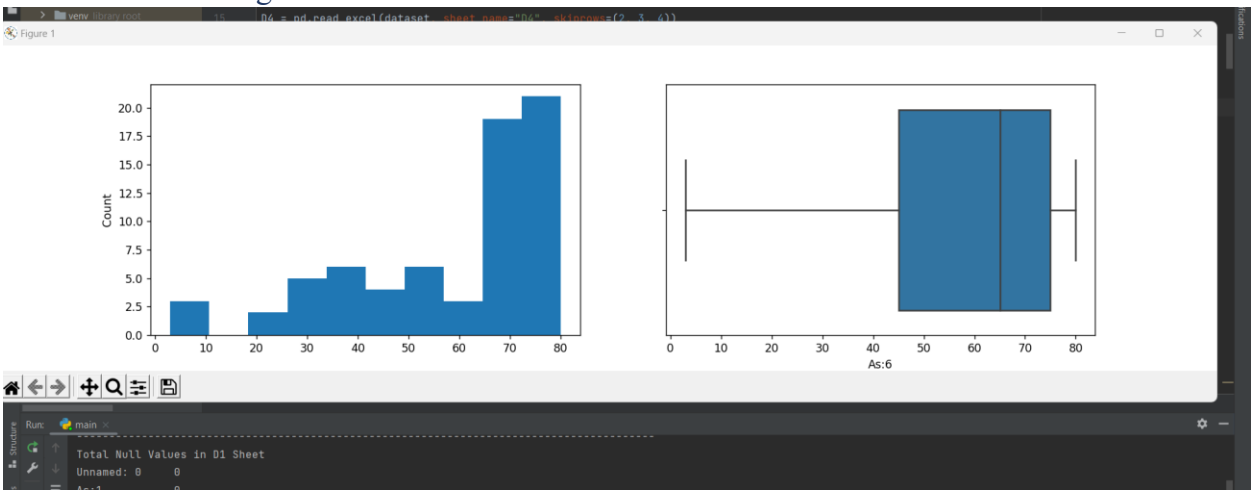


Sheet D1 EDA Assignment 4:

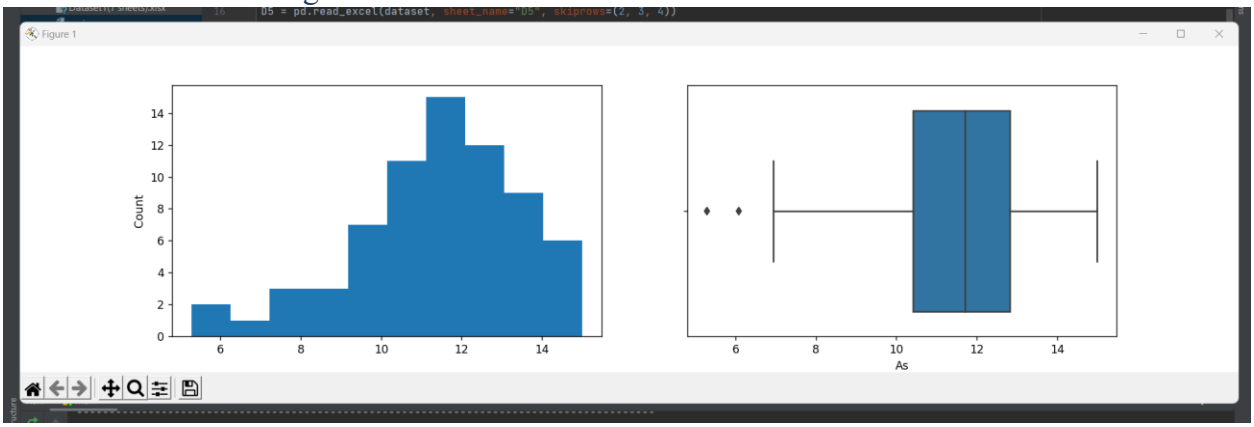


Data Mining Project: Part 1 -> EDA Analysis

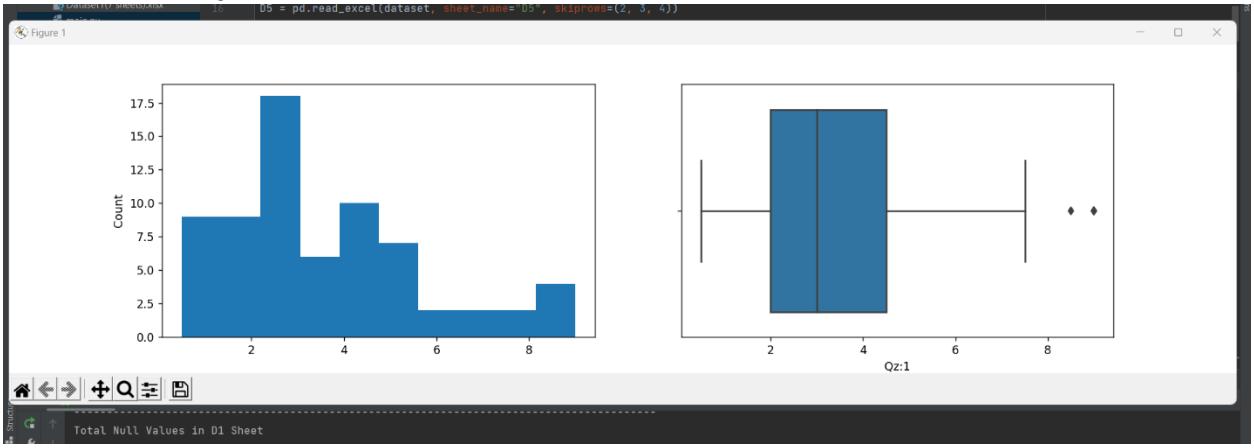
Sheet D1 EDA Assignment 5:



Sheet D1 EDA All Assignments:

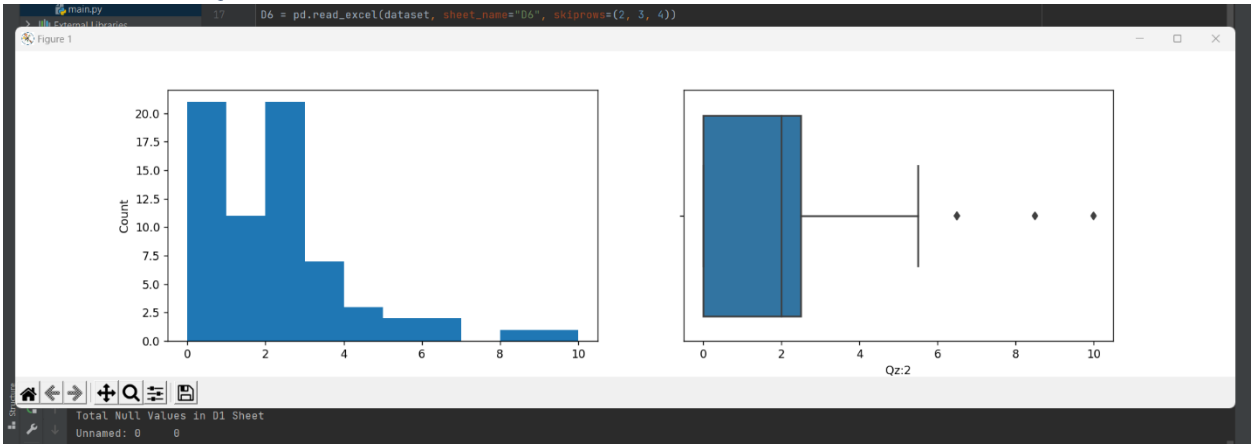


Sheet D1 EDA Quiz 1:

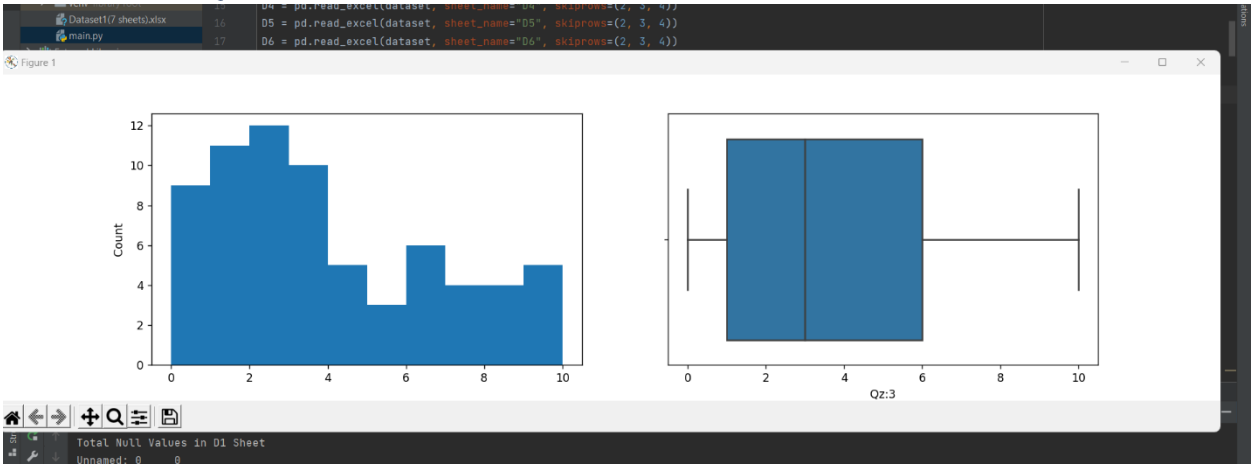


Data Mining Project: Part 1 -> EDA Analysis

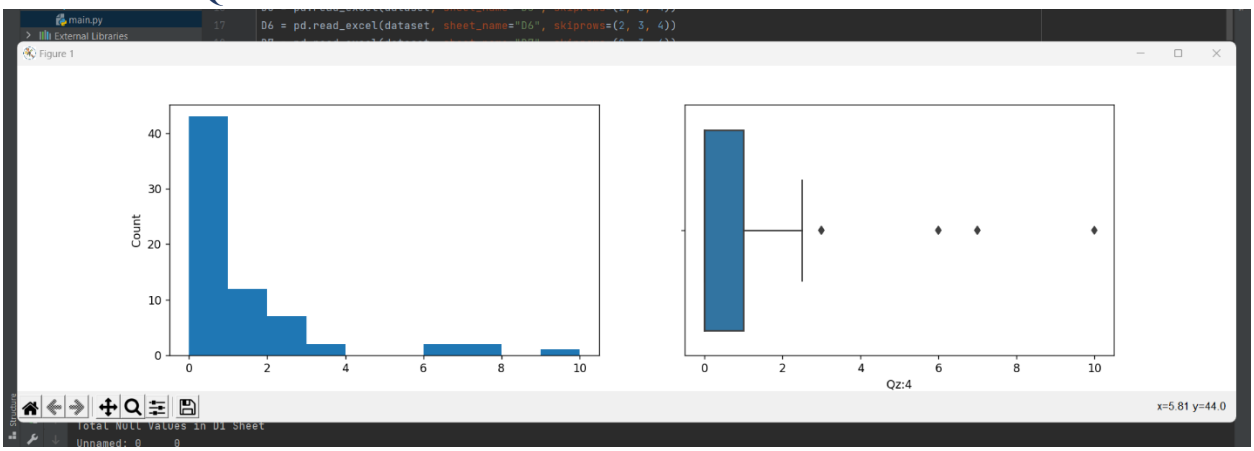
Sheet D1 EDA Quiz 2:



Sheet D1 EDA Quiz 3:

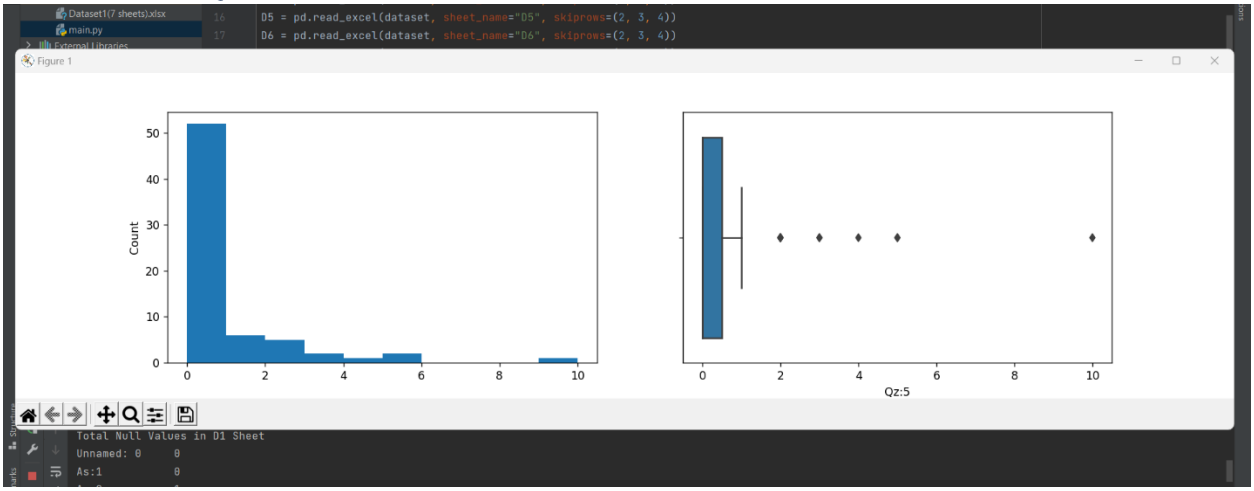


Sheet D1 EDA Quiz 4:

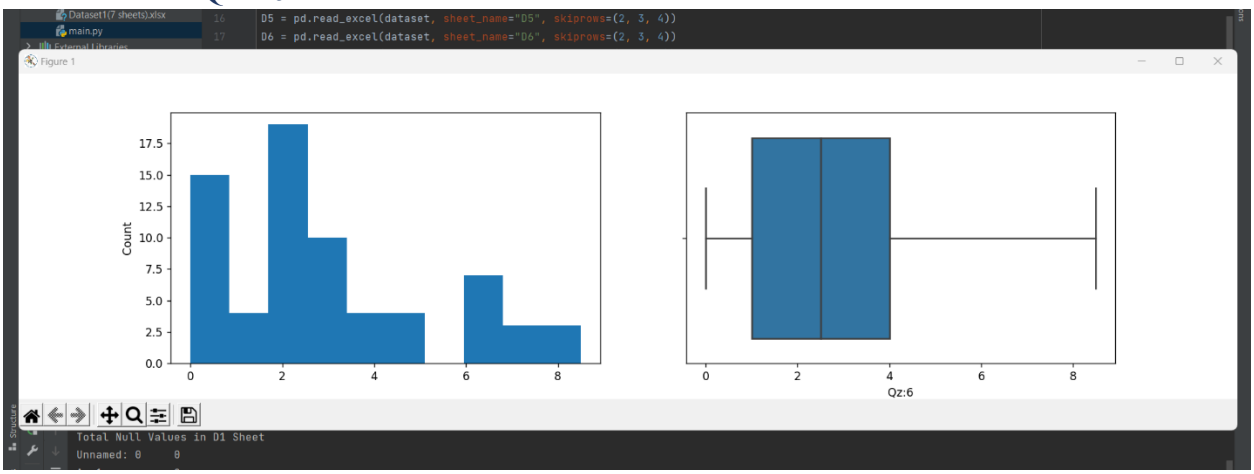


Data Mining Project: Part 1 -> EDA Analysis

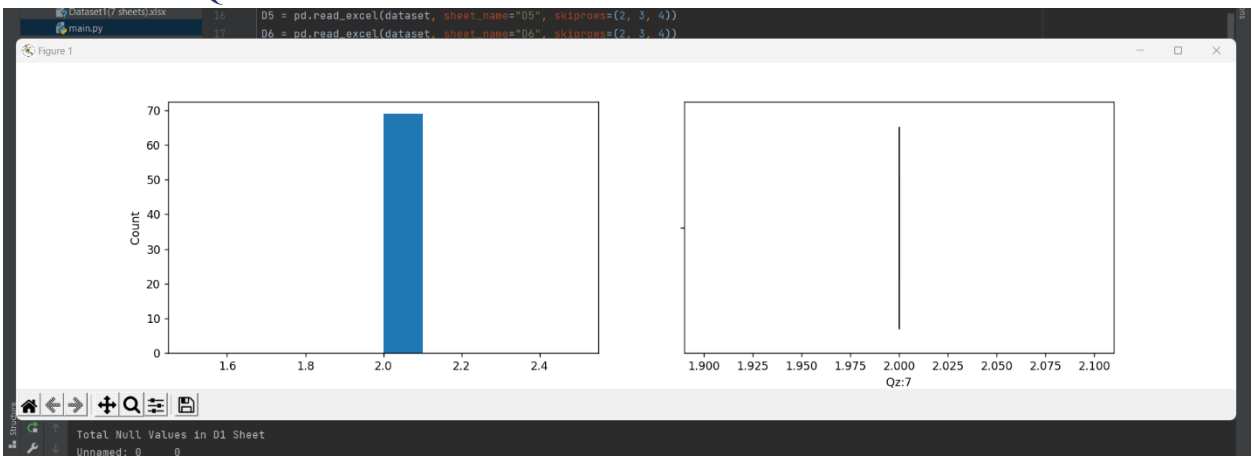
Sheet D1 EDA Quiz 5:



Sheet D1 EDA Quiz 6:

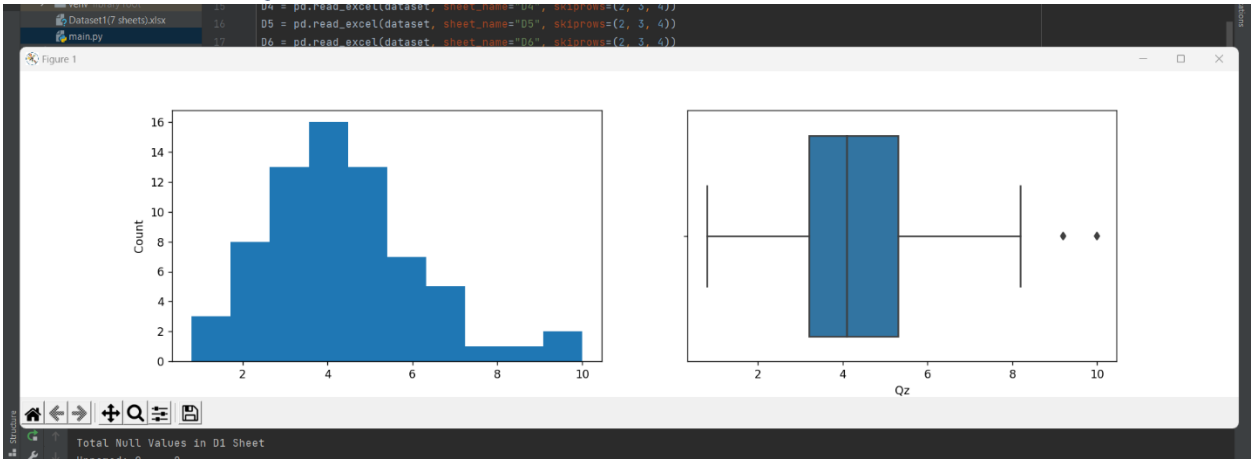


Sheet D1 EDA Quiz 7:

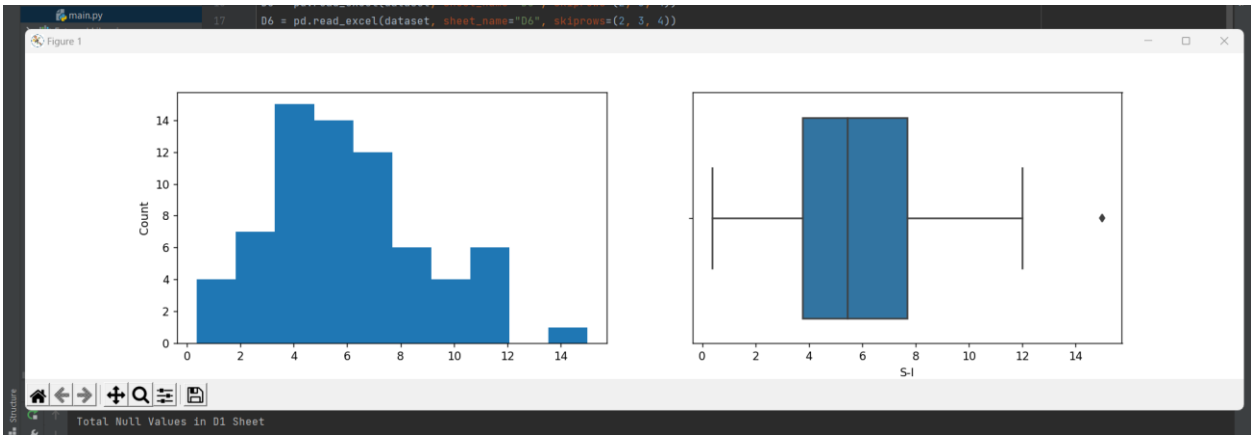


Data Mining Project: Part 1 -> EDA Analysis

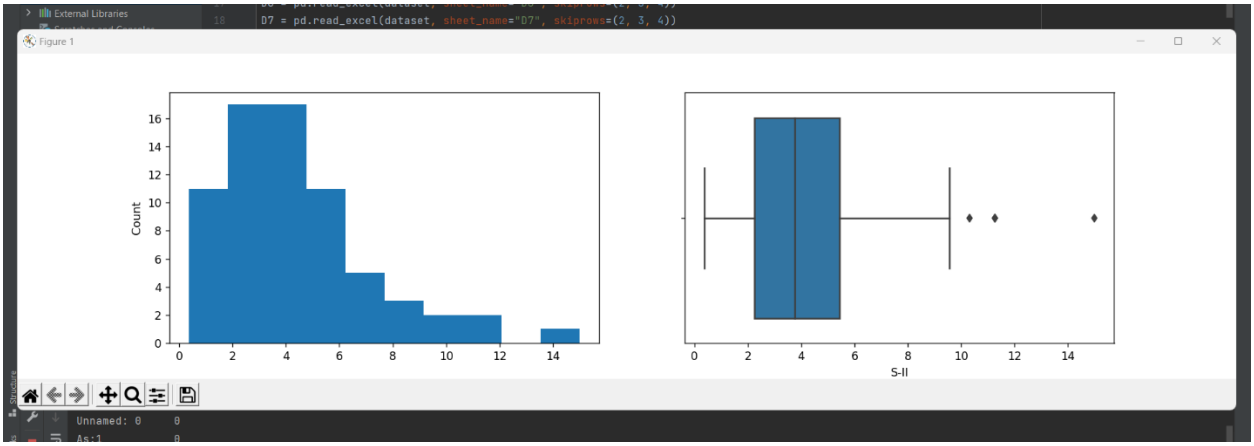
Sheet D1 EDA All Quizzes:



Sheet D1 EDA Sessional I:



Sheet D1 EDA Sessional II:



Data Mining Project: Part 1 -> EDA Analysis

Sheet D1 EDA Results:

```
As:1
Skew : -1.39
As:2
Skew : -1.75
As:3
Skew : -1.2
As:4
Skew : -1.02
As:5
Skew : -0.77
As:6
Skew : -0.99
As
Skew : -0.73
Qz:1
Skew : 0.91
Qz:2
Skew : 1.72
Qz:3
Skew : 0.69
Qz:4
Skew : 2.8
Qz:5
Skew : 3.66
Qz:6
Skew : 0.63
Qz:7
Skew : 0
Qz
Skew : 0.71
```

This analysis shows us the skewness of the attributes depending upon their count and allows us to take steps to normalize the values using data transformation and make it eligible to perform Bivariate EDA on it.

But as we know our data only consists of numerical values hence there is no need to perform Bivariate EDA for further analysis. At this time, our data is clean and ready to proceed in the next phase.

Data Mining Project: Part 1 -> EDA Analysis

Python Code:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings

warnings.filterwarnings('ignore')

# Importing File and Sheets from the given Dataset
dataset = pd.ExcelFile('Dataset1(7 sheets).xlsx', engine='openpyxl')
# We don't need Row 2,3 and 4 because it adds nothing to our data.
D1 = pd.read_excel(dataset, sheet_name="D1", skiprows=(2, 3, 4))
D2 = pd.read_excel(dataset, sheet_name="D2", skiprows=(2, 3, 4))
D3 = pd.read_excel(dataset, sheet_name="D3", skiprows=(2, 3, 4))
D4 = pd.read_excel(dataset, sheet_name="D4", skiprows=(2, 3, 4))
D5 = pd.read_excel(dataset, sheet_name="D5", skiprows=(2, 3, 4))
D6 = pd.read_excel(dataset, sheet_name="D6", skiprows=(2, 3, 4))
D7 = pd.read_excel(dataset, sheet_name="D7", skiprows=(2, 3, 4))

# Checking for Duplication
print("Total Duplications in D1 Sheet")
print(D1.nunique())
print("Total Duplications in D2 Sheet")
print(D2.nunique())
print("Total Duplications in D3 Sheet")
print(D3.nunique())
print("Total Duplications in D4 Sheet")
print(D4.nunique())
print("Total Duplications in D5 Sheet")
print(D5.nunique())
print("Total Duplications in D6 Sheet")
print(D6.nunique())
print("Total Duplications in D7 Sheet")
print(D7.nunique())

# Checking Null Values
print("--------------------------------------------------------------------
------------------")
print("Total Null Values in D1 Sheet")
print(D1.isnull().sum())
print("Total Null Values in D2 Sheet")
print(D2.isnull().sum())
print("Total Null Values in D3 Sheet")
print(D3.isnull().sum())
print("Total Null Values in D4 Sheet")
print(D4.isnull().sum())
print("Total Null Values in D5 Sheet")
print(D5.isnull().sum())
print("Total Null Values in D6 Sheet")
print(D6.isnull().sum())
print("Total Null Values in D7 Sheet")
print(D7.isnull().sum())

# Displaying Null Values using Heatmap
print("--------------------------------------------------------------------
```

Data Mining Project: Part 1 -> EDA Analysis

```
--------------------")
sns.heatmap(D1.isnull(), cbar=False, cmap='viridis')
plt.title("D1 Null Values")
plt.show()
sns.heatmap(D2.isnull(), cbar=False, cmap='viridis')
plt.title("D2 Null Values")
plt.show()
sns.heatmap(D3.isnull(), cbar=False, cmap='viridis')
plt.title("D3 Null Values")
plt.show()
sns.heatmap(D4.isnull(), cbar=False, cmap='viridis')
plt.title("D4 Null Values")
plt.show()
sns.heatmap(D5.isnull(), cbar=False, cmap='viridis')
plt.title("D5 Null Values")
plt.show()
sns.heatmap(D6.isnull(), cbar=False, cmap='viridis')
plt.title("D6 Null Values")
plt.show()
sns.heatmap(D7.isnull(), cbar=False, cmap='viridis')
plt.title("D7 Null Values")
plt.show()

# Replacing missing values with Median of the corresponding column
print("--------------------------------------------------------------------------
--------------------")
num_col_D1 = ['As:1', 'As:2', 'As:3', 'As:4', 'As:5', 'As:6', 'As', 'Qz:1',
'Qz:2', 'Qz:3', 'Qz:4', 'Qz:4', 'Qz:5',
              'Qz:6', 'Qz:7', 'Qz', 'S-I', 'S-II']
for col in num_col_D1:
    D1[col] = pd.to_numeric(D1[col])
    D1[col].fillna(D1[col].median(), inplace=True)
grades_mode = D1.Grade.mode()
D1.Grade.fillna(grades_mode, inplace=True)

num_col_D2 = ['As:1', 'As:2', 'As:3', 'As:4', 'As:5', 'As:6', 'As', 'Qz:1',
'Qz:2', 'Qz:3', 'Qz:4', 'Qz:4', 'Qz:5',
              'Qz:6', 'Qz', 'S-I', 'S-II']
for col in num_col_D2:
    D2[col] = pd.to_numeric(D2[col])
    D2[col].fillna(D2[col].median(), inplace=True)
grades_mode = D2.Grade.mode()
D2.Grade.fillna(grades_mode, inplace=True)

num_col_D3 = ['As:1', 'As:2', 'As:3', 'As:4', 'As:5', 'As:6', 'As', 'Qz:1',
'Qz:2', 'Qz:3', 'Qz:4', 'Qz:4', 'Qz:5',
              'Qz:6', 'Qz:7', 'Qz:8', 'Qz', 'S-I', 'S-II']
for col in num_col_D3:
    D3[col] = pd.to_numeric(D3[col])
    D3[col].fillna(D3[col].median(), inplace=True)
grades_mode = D3.Grade.mode()
D3.Grade.fillna(grades_mode, inplace=True)

num_col_D4 = ['As:1', 'As:2', 'As:3', 'As:4', 'As:5', 'As:6', 'As:7', 'As',
'Qz:1', 'Qz:2', 'Qz:3', 'Qz:4', 'Qz:4',
              'Qz:5', 'Qz', 'S-I', 'S-II']
for col in num_col_D4:
```

Data Mining Project: Part 1 -> EDA Analysis

```
    D4[col] = pd.to_numeric(D4[col])
    D4[col].fillna(D4[col].median(), inplace=True)
grades_mode = D4.Grade.mode()
D4.Grade.fillna(grades_mode, inplace=True)

num_col_D5 = ['As:1', 'As:2', 'As:3', 'As:4', 'As:5', 'As:6', 'As', 'Qz:1',
'Qz:2', 'Qz:3', 'Qz:4', 'Qz:4', 'Qz:5',
              'Qz:6', 'Qz:7', 'Qz:8', 'Qz', 'S-I', 'S-II']
for col in num_col_D5:
    D5[col] = pd.to_numeric(D5[col])
    D5[col].fillna(D5[col].median(), inplace=True)
grades_mode = D5.Grade.mode()
D5.Grade.fillna(grades_mode, inplace=True)

num_col_D6 = ['As:1', 'As:2', 'As:3', 'As:4', 'As:5', 'As:6', 'As', 'Qz:1',
'Qz:2', 'Qz:3', 'Qz:4', 'Qz:4', 'Qz:5',
              'Qz:6', 'Qz:7', 'Qz', 'S-I', 'S-II']
for col in num_col_D6:
    D6[col] = pd.to_numeric(D6[col])
    D6[col].fillna(D6[col].median(), inplace=True)
grades_mode = D6.Grade.mode()
D6.Grade.fillna(grades_mode, inplace=True)

num_col_D7 = ['As:1', 'As:2', 'As:3', 'As:4', 'As:5', 'As:6', 'As', 'Qz:1',
'Qz:2', 'Qz:3', 'Qz:4', 'Qz:4', 'Qz:5',
              'Qz:6', 'Qz:7', 'Qz:8', 'Qz', 'S-I', 'S-II']
for col in num_col_D7:
    D7[col] = pd.to_numeric(D7[col])
    D7[col].fillna(D7[col].median(), inplace=True)
grades_mode = D7.Grade.mode()
D7.Grade.fillna(grades_mode, inplace=True)

# Performing Data-Reduction Techniques
#   Removing 1st Column from data as it only contains serial number that is
not helpful for us
print("-------------------------------------------------------------------------
-------------------")
D1 = D1.drop(columns=D1.columns[0])
D2 = D2.drop(columns=D2.columns[0])
D3 = D3.drop(columns=D3.columns[0])
D4 = D4.drop(columns=D4.columns[0])
D5 = D5.drop(columns=D5.columns[0])
D6 = D6.drop(columns=D6.columns[0])
D7 = D7.drop(columns=D7.columns[0])

# Data Cleared. Now proceeding to Checking Correlation of attributes.
print("-------------------------------------------------------------------------
-------------------")
plt.figure(figsize=(13, 13))
sns.heatmap(D1[['As:1', 'As:2', 'As:3', 'As:4', 'As:5', 'As:6', 'As', 'Qz:1',
'Qz:2', 'Qz:3', 'Qz:4', 'Qz:4', 'Qz:5',
                'Qz:6', 'Qz:7', 'Qz', 'S-I', 'S-II']].corr(), cbar=True,
annot=True, cmap='Blues')
plt.title("D1 Correlation Matrix")
plt.show()

plt.figure(figsize=(13, 13))
```

Data Mining Project: Part 1 -> EDA Analysis

```
sns.heatmap(D2[['As:1', 'As:2', 'As:3', 'As:4', 'As:5', 'As:6', 'As', 'Qz:1',
'Qz:2', 'Qz:3', 'Qz:4', 'Qz:4', 'Qz:5',
                'Qz:6', 'Qz', 'S-I', 'S-II']].corr(), cbar=True, annot=True,
cmap='Blues')
plt.title("D2 Correlation Matrix")
plt.show()

plt.figure(figsize=(13, 13))
sns.heatmap(D3[['As:1', 'As:2', 'As:3', 'As:4', 'As:5', 'As:6', 'As', 'Qz:1',
'Qz:2', 'Qz:3', 'Qz:4', 'Qz:4', 'Qz:5',
                'Qz:6', 'Qz:7', 'Qz:8', 'Qz', 'S-I', 'S-II']].corr(),
cbar=True, annot=True, cmap='Blues')
plt.title("D3 Correlation Matrix")
plt.show()

plt.figure(figsize=(13, 13))
sns.heatmap(D4[['As:1', 'As:2', 'As:3', 'As:4', 'As:5', 'As:6', 'As:7', 'As',
'Qz:1', 'Qz:2', 'Qz:3', 'Qz:4', 'Qz:4',
                'Qz:5', 'Qz', 'S-I', 'S-II']].corr(), cbar=True, annot=True,
cmap='Blues')
plt.title("D4 Correlation Matrix")
plt.show()

plt.figure(figsize=(13, 13))
sns.heatmap(D5[['As:1', 'As:2', 'As:3', 'As:4', 'As:5', 'As:6', 'As', 'Qz:1',
'Qz:2', 'Qz:3', 'Qz:4', 'Qz:4', 'Qz:5',
                'Qz:6', 'Qz:7', 'Qz:8', 'Qz', 'S-I', 'S-II']].corr(),
cbar=True, annot=True, cmap='Blues')
plt.title("D5 Correlation Matrix")
plt.show()

plt.figure(figsize=(13, 13))
sns.heatmap(D6[['As:1', 'As:2', 'As:3', 'As:4', 'As:5', 'As:6', 'As', 'Qz:1',
'Qz:2', 'Qz:3', 'Qz:4', 'Qz:4', 'Qz:5',
                'Qz:6', 'Qz:7', 'Qz', 'S-I', 'S-II']].corr(), cbar=True,
annot=True, cmap='Blues')
plt.title("D6 Correlation Matrix")
plt.show()

plt.figure(figsize=(13, 13))
sns.heatmap(D7[['As:1', 'As:2', 'As:3', 'As:4', 'As:5', 'As:6', 'As', 'Qz:1',
'Qz:2', 'Qz:3', 'Qz:4', 'Qz:4', 'Qz:5',
                'Qz:6', 'Qz:7', 'Qz:8', 'Qz', 'S-I', 'S-II']].corr(),
cbar=True, annot=True, cmap='Blues')
plt.title("D7 Correlation Matrix")
plt.show()

# Showing the relation between assignments weightage and Grades
print("----------------------------------------------------------------------
-------------------")
D1.groupby('Grade')['As'].mean().plot.bar()
plt.title("Grade vs Assignment")
plt.show()

# Showing the relation between quizzes weightage and Grades
D1.groupby('Grade')['Qz'].mean().plot.bar()
plt.title("Grade vs Quizzes")
```

Data Mining Project: Part 1 -> EDA Analysis

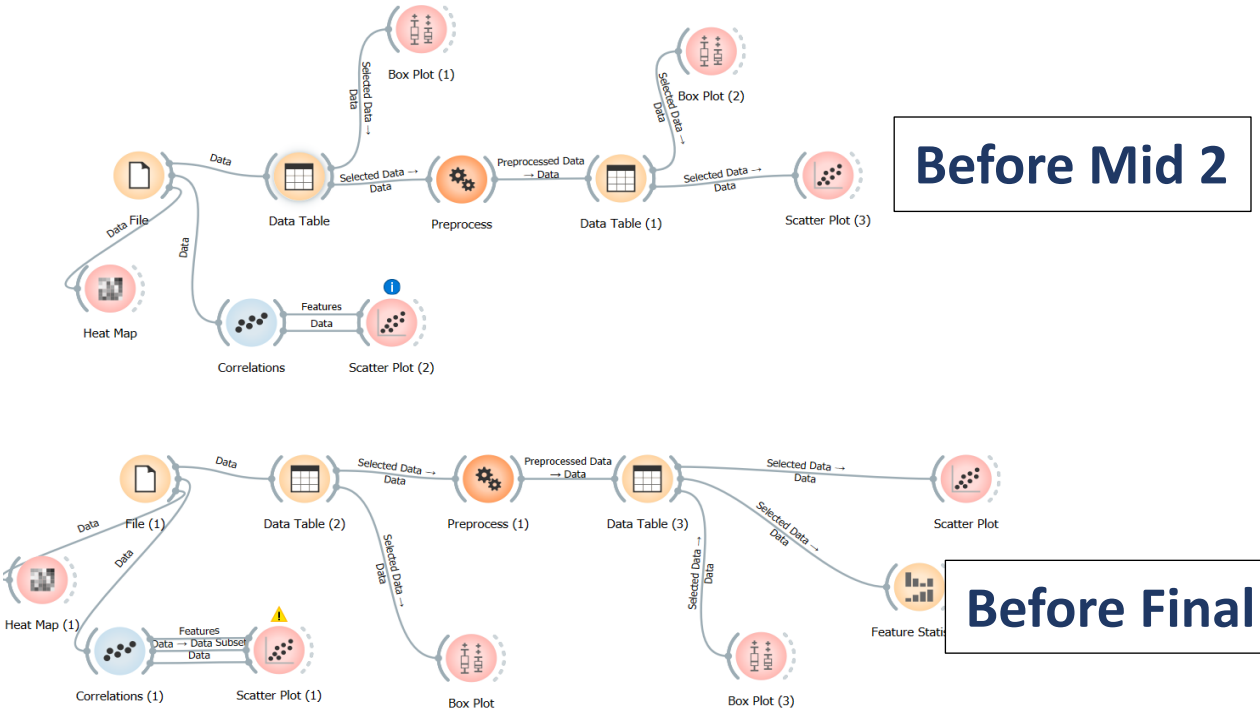
```
plt.show()

# Showing the relation between S-I weightage and Grades
D1.groupby('Grade')['S-I'].mean().plot.bar()
plt.title("Grade vs S-I")
plt.show()

# Showing the relation between S-II weightage and Grades
D1.groupby('Grade')['S-II'].mean().plot.bar()
plt.title("Grade vs S-II")
plt.show()

# Same can be done for D2 to D7 for further analysis

# Performing  EDA on D1
# Selecting only Numerical Values
print("------------------------------------------------------------------------
-------------------")
num_cols = D1.select_dtypes(include=np.number).columns.tolist()
for col in num_cols:
    print(col)
    print('Skew :', round(D1[col].skew(), 2))
    plt.figure(figsize=(15, 4))
    plt.subplot(1, 2, 1)
    D1[col].hist(grid=False)
    plt.ylabel('Count')
    plt.subplot(1, 2, 2)
    sns.boxplot(x=D1[col])
    plt.show()
# Same can be done for D2 to D7 for further analysis
```

Data Mining Project: Part 1 -> EDA Analysis

Working Using Orange:

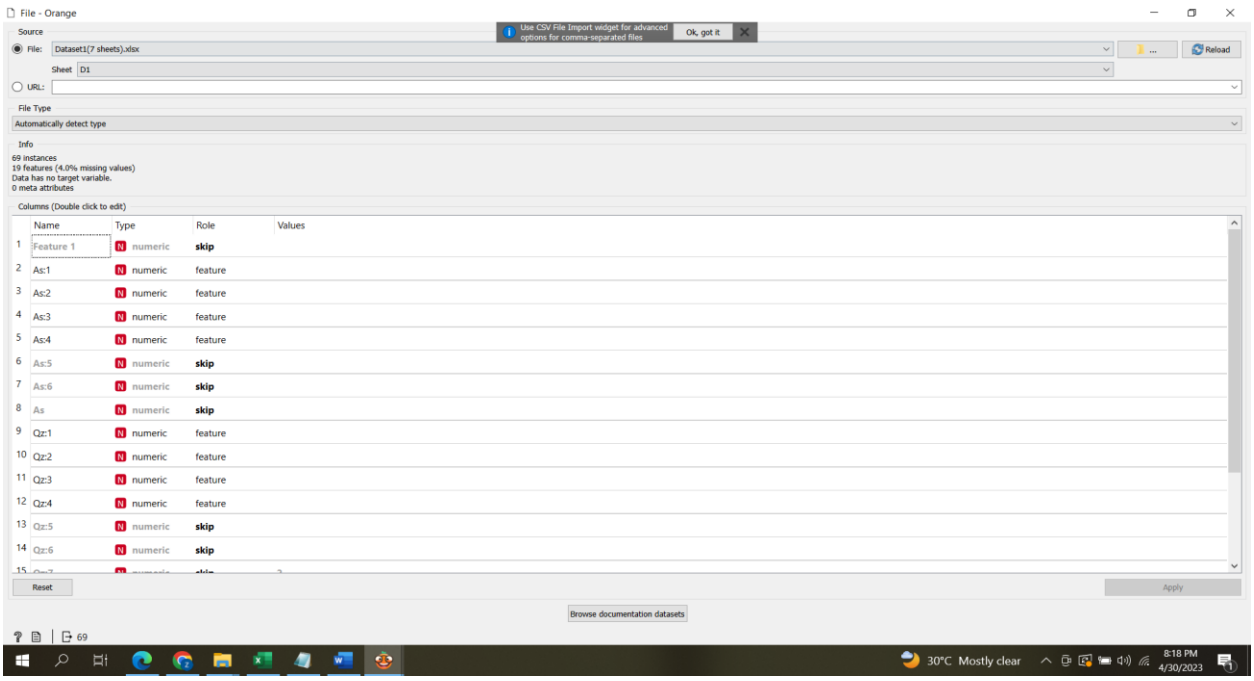
D1 to D7 process is same just showing the D1:

Workflow Demonstration:

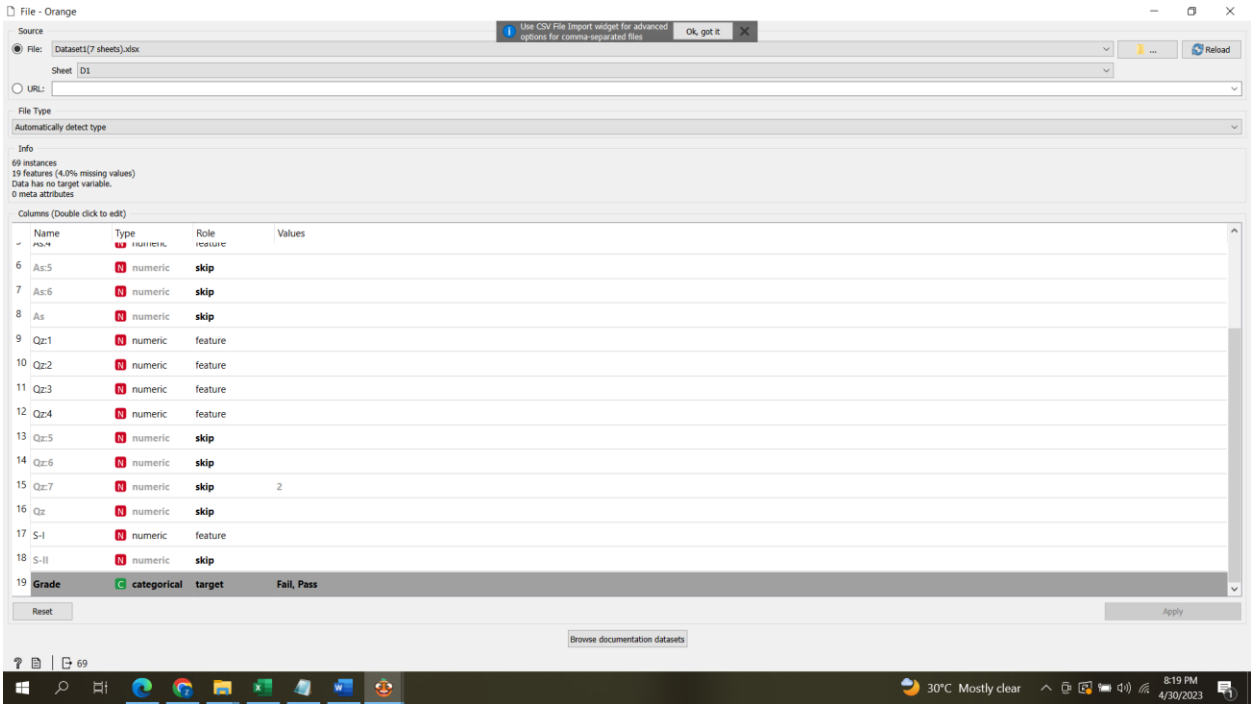


Data Importing:

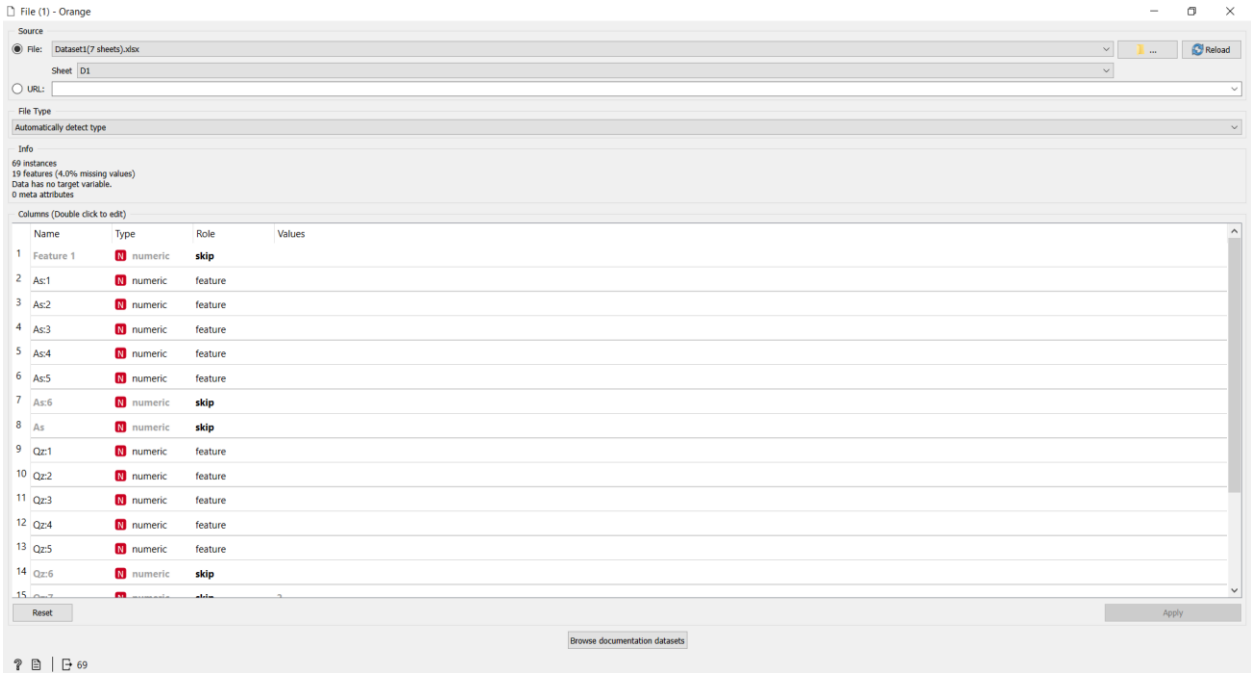
Importing the data from D1 of Assign 1 to 4, Quiz 1 to 4 & S-1 for before mid-2 grade suggestion and set their type as all are feature and grade is target.



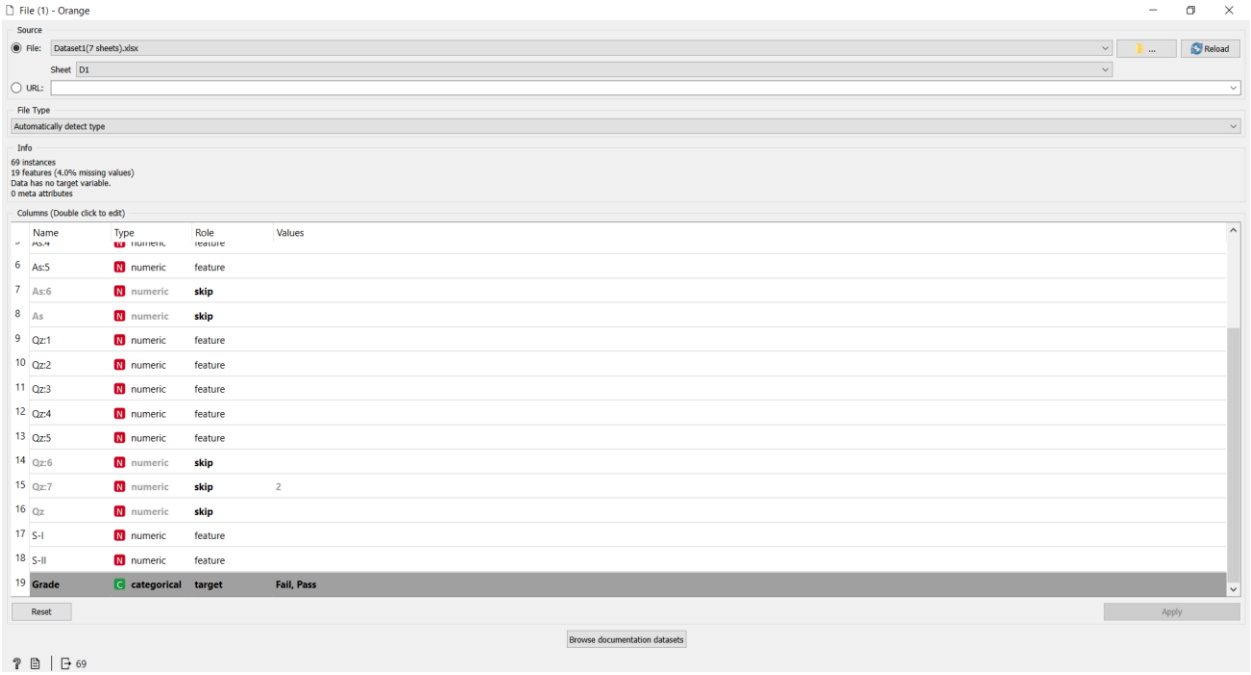
Data Mining Project: Part 1 -> EDA Analysis



Importing the data from D1 of Assign 1 to 5, Quiz 1 to 5 & S-1, S-2 for before final grade suggestion and set their type as all are feature and grade is target.



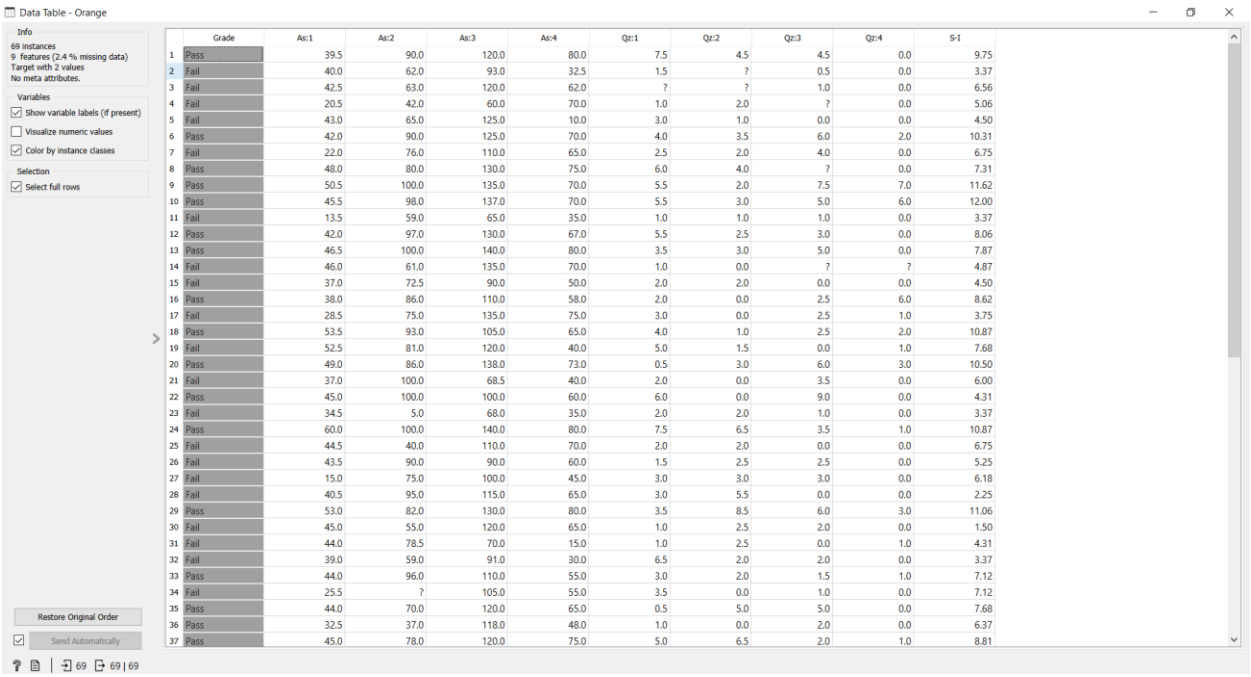
Data Mining Project: Part 1 -> EDA Analysis



Importing process is same from as D1 from D2 to D7.

Covert data into Table from File:

Data imported from file in table format is shown in table as below:

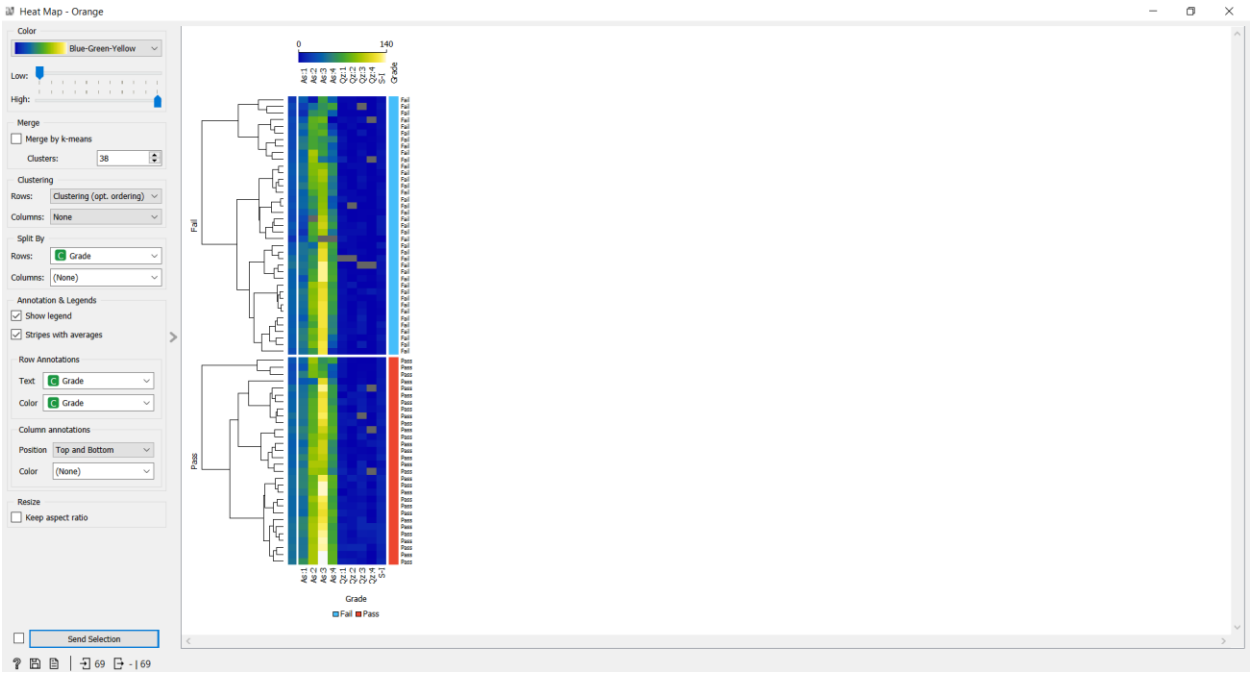


	Grade	As:1	As:2	As:3	As:4	Qz:1	Qz:2	Qz:3	Qz:4	S-I
1	Pass	39.5	90.0	120.0	80.0	7.5	4.5	4.5	0.0	9.75
2	Fail	40.0	62.0	93.0	32.5	1.5	?	0.5	0.0	3.37
3	Fail	42.5	63.0	120.0	62.0	?	?	1.0	0.0	6.56
4	Fail	20.5	42.0	60.0	70.0	1.0	2.0	?	0.0	5.06
5	Fail	43.0	65.0	125.0	10.0	3.0	1.0	0.0	0.0	4.50
6	Pass	42.0	90.0	125.0	70.0	4.0	3.5	6.0	2.0	10.31
7	Fail	22.0	76.0	110.0	65.0	2.5	2.0	4.0	0.0	6.75
8	Pass	48.0	80.0	130.0	75.0	6.0	4.0	?	0.0	7.31
9	Pass	50.5	100.0	135.0	70.0	5.5	2.0	7.5	7.0	11.62
10	Pass	45.5	98.0	137.0	70.0	5.5	3.0	5.0	6.0	12.00
11	Fail	13.5	59.0	65.0	35.0	1.0	1.0	1.0	0.0	3.37
12	Pass	42.0	97.0	130.0	67.0	5.5	2.5	3.0	0.0	8.06
13	Pass	46.5	100.0	140.0	80.0	3.5	3.0	5.0	0.0	7.87
14	Fail	46.0	61.0	135.0	70.0	1.0	0.0	?	?	4.87
15	Fail	37.0	72.5	90.0	50.0	2.0	2.0	0.0	0.0	4.50
16	Pass	38.0	86.0	110.0	58.0	2.0	0.0	2.5	6.0	8.62
17	Fail	28.5	75.0	135.0	75.0	3.0	0.0	2.5	1.0	3.75
18	Pass	53.5	93.0	105.0	65.0	4.0	1.0	2.5	2.0	10.87
19	Fail	52.5	81.0	120.0	40.0	5.0	1.5	0.0	1.0	7.68
20	Pass	49.0	86.0	138.0	73.0	0.5	3.0	6.0	3.0	10.50
21	Fail	37.0	100.0	68.5	40.0	2.0	0.0	3.5	0.0	6.00
22	Pass	45.0	100.0	100.0	60.0	6.0	0.0	9.0	0.0	4.31
23	Fail	34.5	5.0	68.0	35.0	2.0	2.0	1.0	0.0	3.37
24	Pass	60.0	100.0	140.0	80.0	7.5	6.5	3.5	1.0	10.87
25	Fail	44.5	40.0	110.0	70.0	2.0	2.0	0.0	0.0	6.75
26	Fail	43.5	90.0	90.0	60.0	1.5	2.5	2.5	0.0	5.25
27	Fail	15.0	75.0	100.0	45.0	3.0	3.0	3.0	0.0	6.18
28	Fail	40.5	95.0	115.0	65.0	3.0	5.5	0.0	0.0	2.25
29	Pass	53.0	82.0	130.0	80.0	3.5	8.5	6.0	3.0	11.06
30	Fail	45.0	55.0	120.0	65.0	1.0	2.5	2.0	0.0	1.50
31	Fail	44.0	78.5	70.0	15.0	1.0	2.5	0.0	1.0	4.31
32	Fail	39.0	59.0	91.0	30.0	6.5	2.0	2.0	0.0	3.37
33	Pass	44.0	96.0	110.0	55.0	3.0	2.0	1.5	1.0	7.12
34	Fail	25.5	?	105.0	55.0	3.5	0.0	1.0	0.0	7.12
35	Pass	44.0	70.0	120.0	65.0	0.5	5.0	5.0	0.0	7.68
36	Pass	32.5	37.0	118.0	48.0	1.0	0.0	2.0	0.0	6.37
37	Pass	45.0	78.0	120.0	75.0	5.0	6.5	2.0	1.0	8.81

Data Mining Project: Part 1 -> EDA Analysis

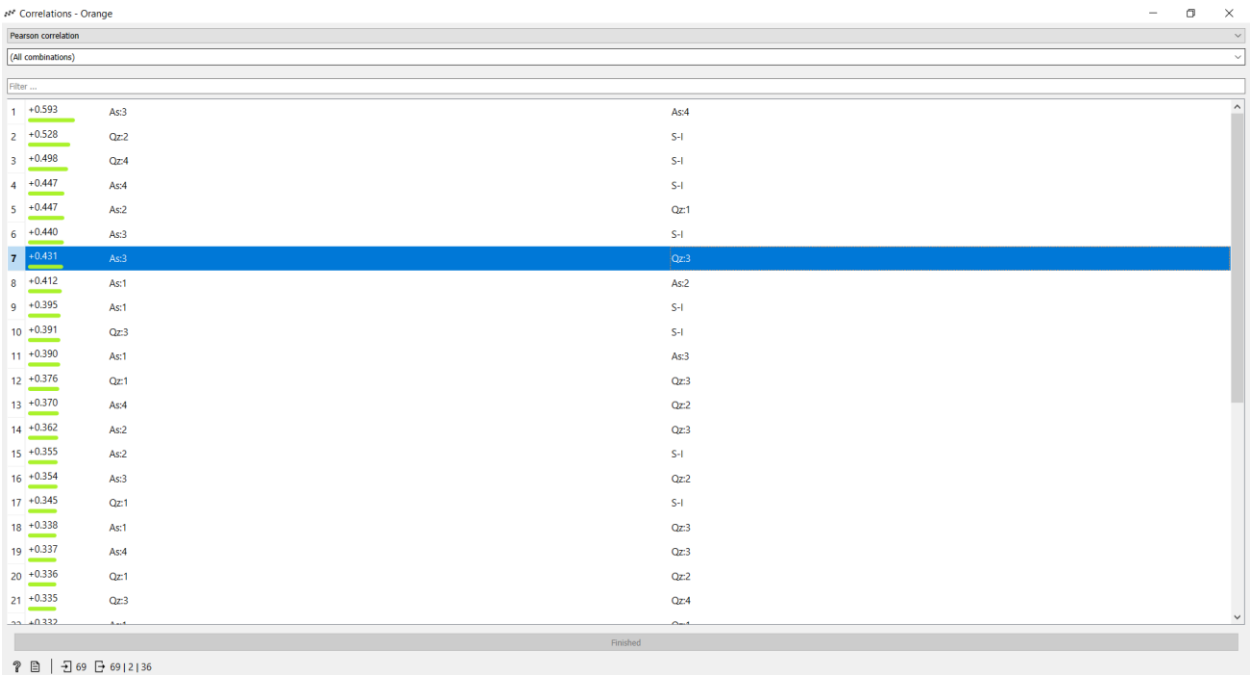
Heat Map:

Heat map is a graphical method for visualizing attribute values in a two-way matrix. It only works on datasets containing numeric variables. Heat map shows low expressions in blue and high expressions in yellow and white.

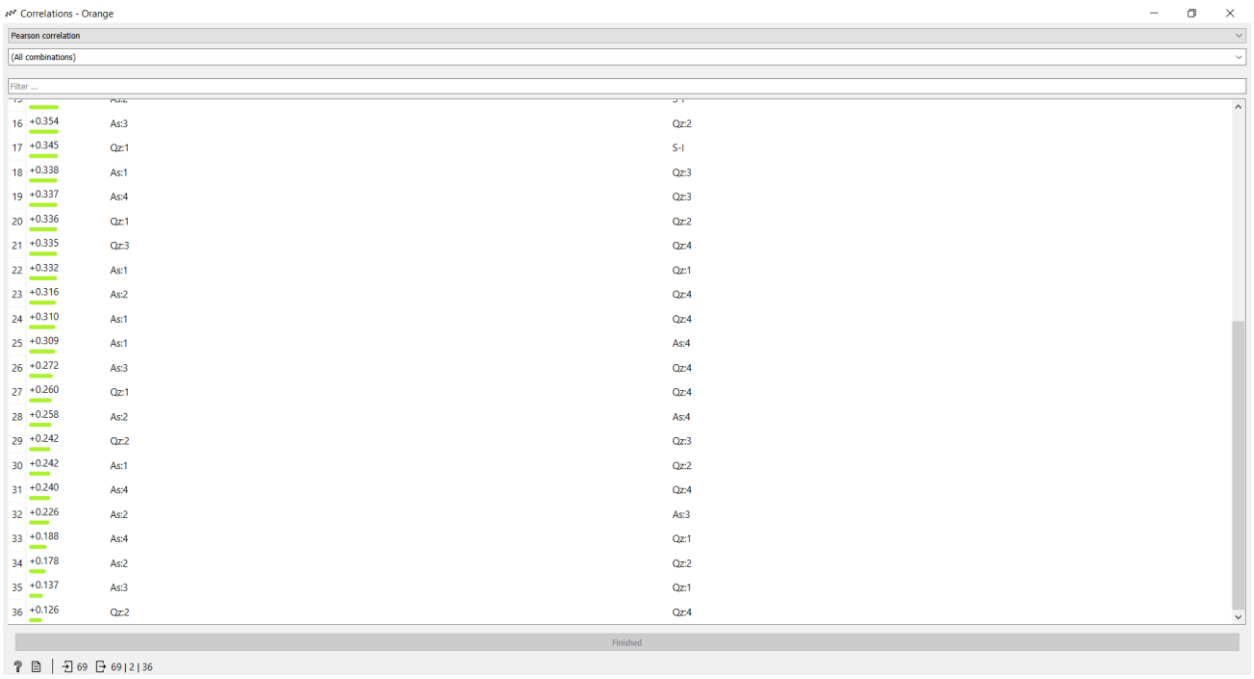


Co-relation:

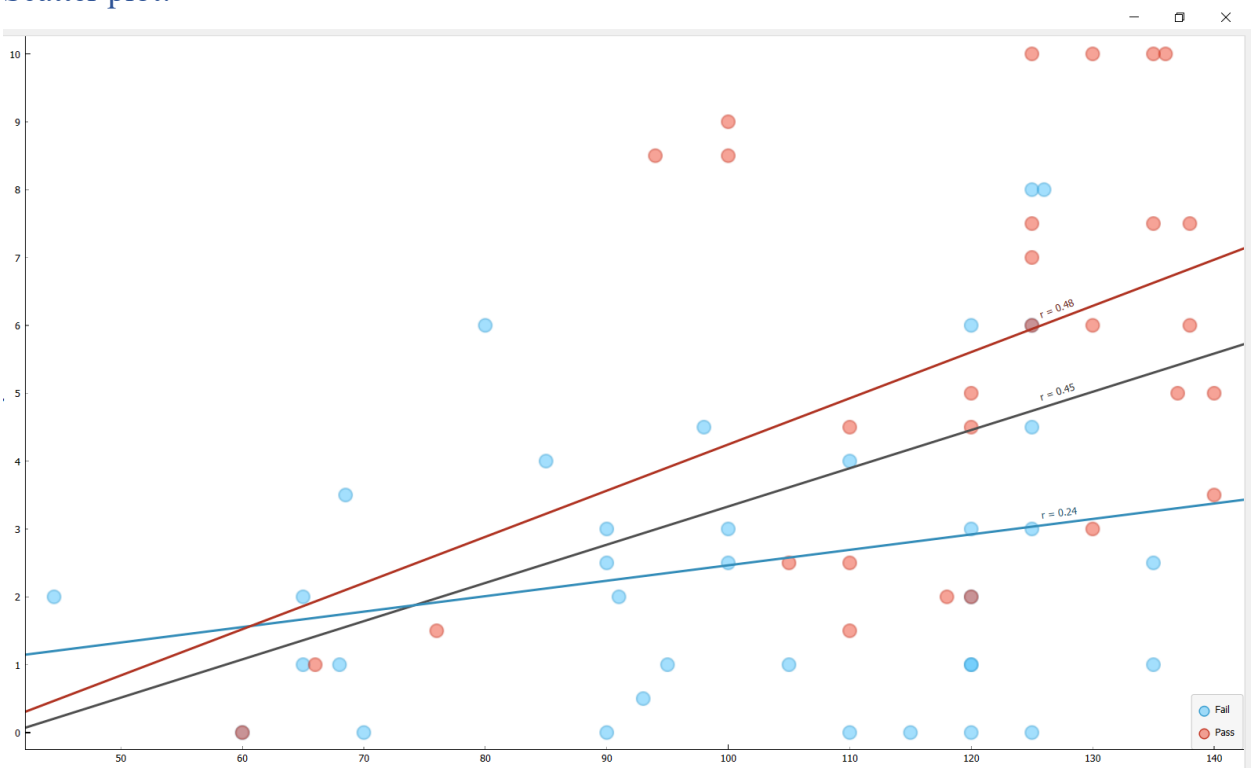
Load the File widget and connect it to Correlations (Pearson). Positively correlated feature pairs will be at the top of the list and negatively correlated will be at the bottom.



Data Mining Project: Part 1 -> EDA Analysis



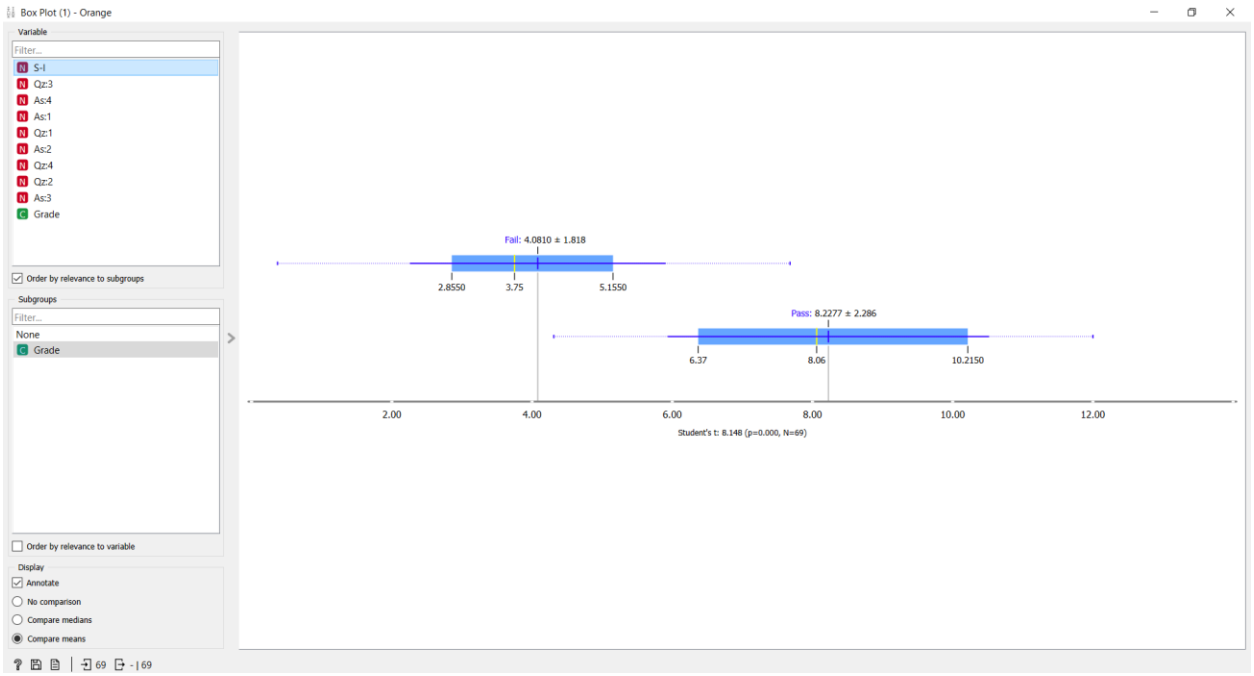
Scatter plot:



Data Mining Project: Part 1 -> EDA Analysis

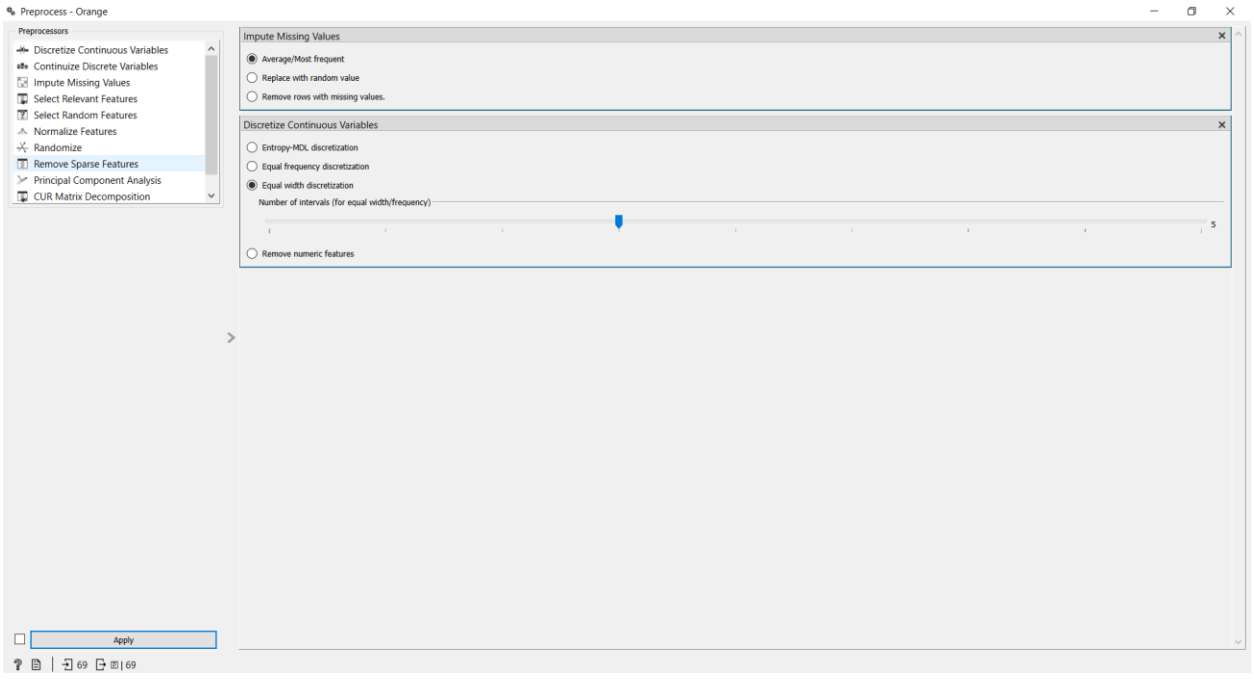
Box Plot:

S1 & grade



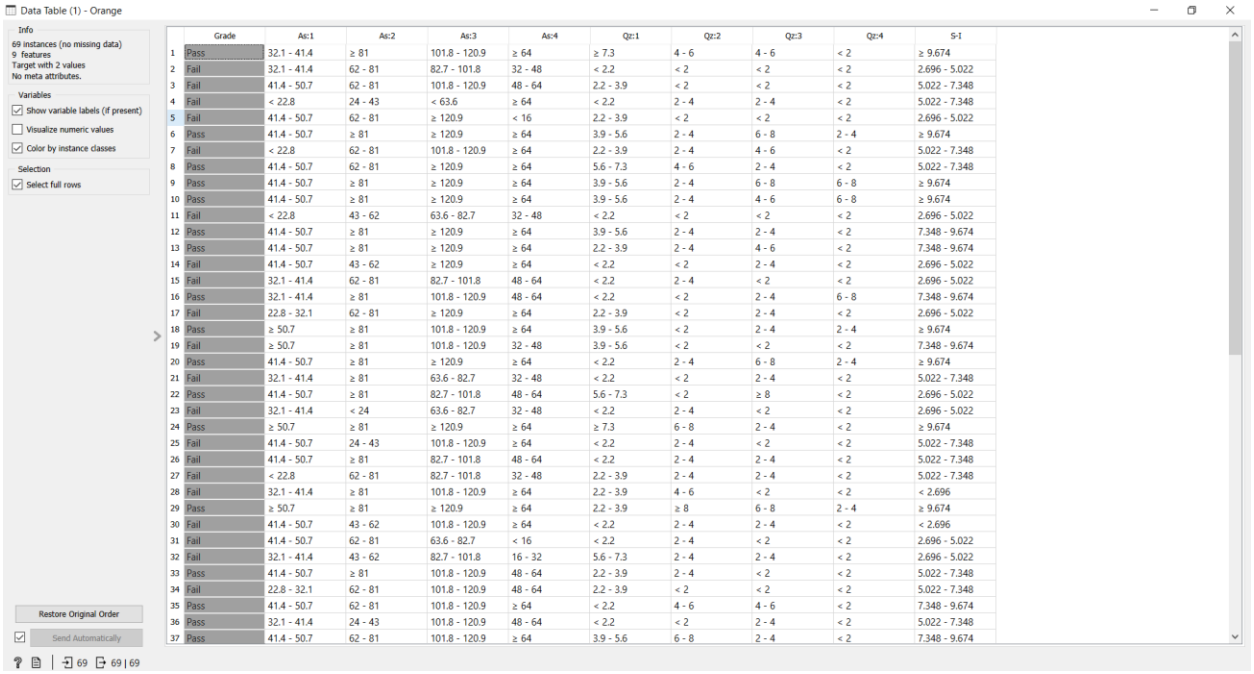
For the sample I just show the grade with S1.

Preprocessing:

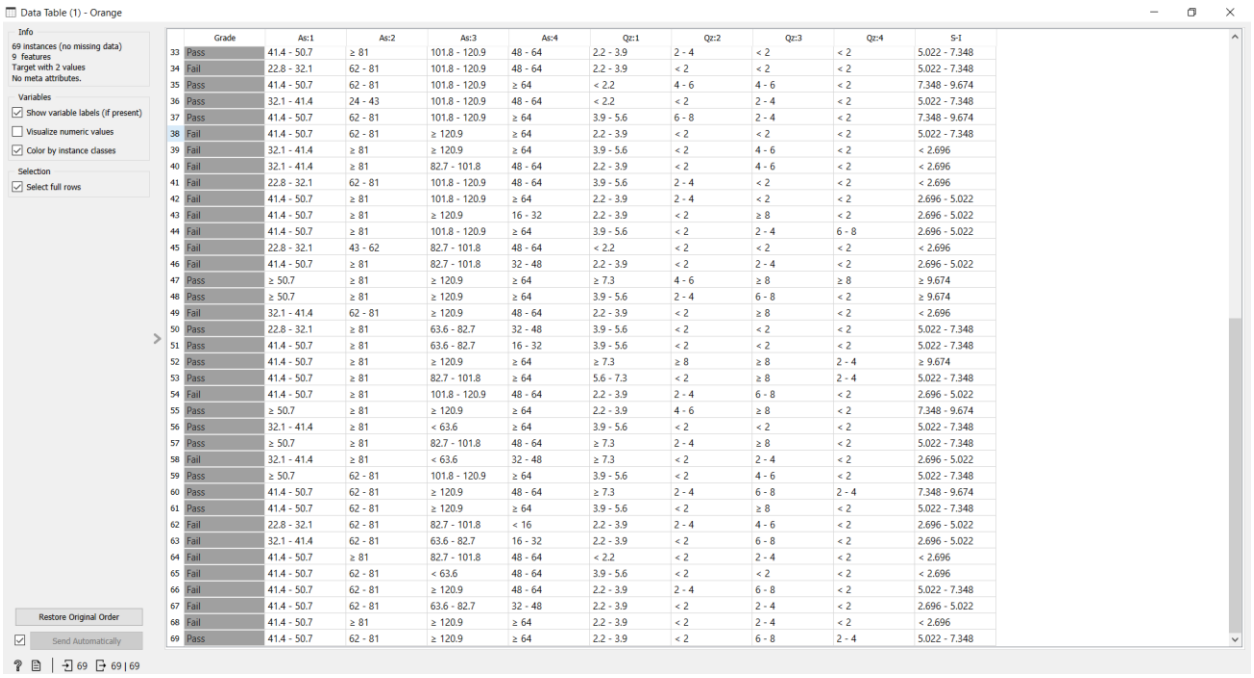


Data Mining Project: Part 1 -> EDA Analysis

Data After Preprocessing:



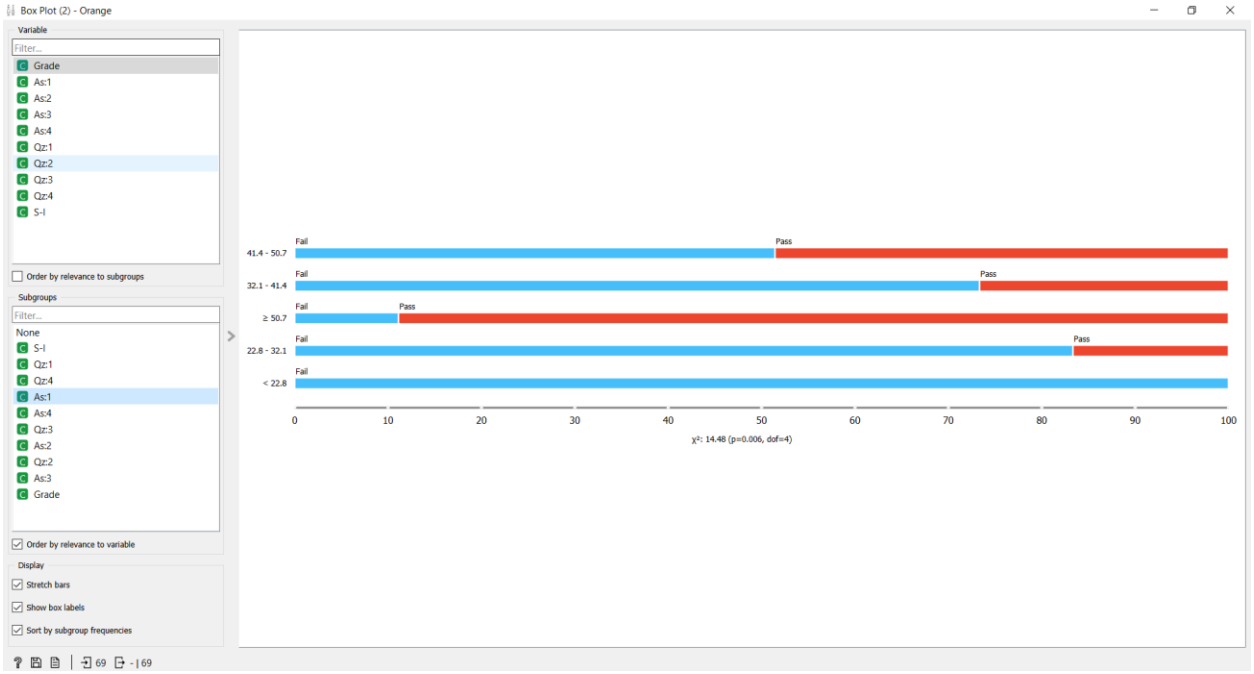
	Grade	As:1	As:2	As:3	As:4	Qz:1	Qz:2	Qz:3	Qz:4	S-I
1	Pass	32.1 - 41.4	≥ 81	101.8 - 120.9	≥ 64	≥ 7.3	4 - 6	4 - 6	< 2	≥ 9.674
2	Fail	32.1 - 41.4	62 - 81	82.7 - 101.8	32 - 48	< 2.2	< 2	< 2	< 2	2.696 - 5.022
3	Fail	41.4 - 50.7	62 - 81	101.8 - 120.9	48 - 64	2.2 - 3.9	< 2	< 2	< 2	5.022 - 7.348
4	Fail	< 22.8	24 - 43	< 63.6	≥ 64	< 2.2	2 - 4	2 - 4	< 2	5.022 - 7.348
5	Fail	41.4 - 50.7	62 - 81	≥ 120.9	< 16	2.2 - 3.9	< 2	< 2	< 2	2.696 - 5.022
6	Pass	41.4 - 50.7	≥ 81	≥ 120.9	≥ 64	3.9 - 5.6	2 - 4	6 - 8	2 - 4	≥ 9.674
7	Fail	< 22.8	62 - 81	101.8 - 120.9	≥ 64	2.2 - 3.9	2 - 4	4 - 6	< 2	5.022 - 7.348
8	Pass	41.4 - 50.7	62 - 81	≥ 120.9	≥ 64	5.6 - 7.3	4 - 6	2 - 4	< 2	5.022 - 7.348
9	Pass	41.4 - 50.7	≥ 81	≥ 120.9	≥ 64	3.9 - 5.6	2 - 4	6 - 8	6 - 8	≥ 9.674
10	Pass	41.4 - 50.7	≥ 81	≥ 120.9	≥ 64	3.9 - 5.6	2 - 4	4 - 6	6 - 8	≥ 9.674
11	Fail	< 22.8	43 - 62	63.6 - 82.7	32 - 48	< 2.2	< 2	< 2	< 2	2.696 - 5.022
12	Pass	41.4 - 50.7	≥ 81	≥ 120.9	≥ 64	3.9 - 5.6	2 - 4	2 - 4	< 2	7.348 - 9.674
13	Pass	41.4 - 50.7	≥ 81	≥ 120.9	≥ 64	2.2 - 3.9	2 - 4	4 - 6	< 2	7.348 - 9.674
14	Fail	41.4 - 50.7	43 - 62	≥ 120.9	≥ 64	< 2.2	< 2	2 - 4	< 2	2.696 - 5.022
15	Fail	32.1 - 41.4	62 - 81	82.7 - 101.8	48 - 64	< 2.2	2 - 4	< 2	< 2	2.696 - 5.022
16	Pass	32.1 - 41.4	≥ 81	101.8 - 120.9	48 - 64	< 2.2	< 2	2 - 4	6 - 8	7.348 - 9.674
17	Fail	22.8 - 32.1	62 - 81	≥ 120.9	≥ 64	2.2 - 3.9	< 2	2 - 4	< 2	2.696 - 5.022
18	Pass	≥ 50.7	≥ 81	101.8 - 120.9	≥ 64	3.9 - 5.6	< 2	2 - 4	2 - 4	≥ 9.674
19	Fail	≥ 50.7	≥ 81	101.8 - 120.9	32 - 48	3.9 - 5.6	< 2	< 2	< 2	7.348 - 9.674
20	Pass	41.4 - 50.7	≥ 81	≥ 120.9	≥ 64	< 2.2	2 - 4	6 - 8	2 - 4	≥ 9.674
21	Fail	32.1 - 41.4	≥ 81	63.6 - 82.7	32 - 48	< 2.2	< 2	2 - 4	< 2	5.022 - 7.348
22	Pass	41.4 - 50.7	≥ 81	82.7 - 101.8	48 - 64	5.6 - 7.3	< 2	≥ 8	< 2	2.696 - 5.022
23	Fail	32.1 - 41.4	< 24	63.6 - 82.7	32 - 48	< 2.2	2 - 4	< 2	< 2	2.696 - 5.022
24	Pass	≥ 50.7	≥ 81	≥ 120.9	≥ 64	≥ 7.3	6 - 8	2 - 4	< 2	≥ 9.674
25	Fail	41.4 - 50.7	24 - 43	101.8 - 120.9	≥ 64	< 2.2	2 - 4	< 2	< 2	5.022 - 7.348
26	Fail	41.4 - 50.7	≥ 81	82.7 - 101.8	48 - 64	< 2.2	2 - 4	2 - 4	< 2	5.022 - 7.348
27	Fail	< 22.8	62 - 81	82.7 - 101.8	32 - 48	2.2 - 3.9	2 - 4	2 - 4	< 2	5.022 - 7.348
28	Fail	32.1 - 41.4	≥ 81	101.8 - 120.9	≥ 64	2.2 - 3.9	4 - 6	< 2	< 2	< 2.696
29	Pass	≥ 50.7	≥ 81	≥ 120.9	≥ 64	2.2 - 3.9	≥ 8	6 - 8	2 - 4	≥ 9.674
30	Fail	41.4 - 50.7	43 - 62	101.8 - 120.9	≥ 64	< 2.2	2 - 4	2 - 4	< 2	< 2.696
31	Fail	41.4 - 50.7	62 - 81	63.6 - 82.7	< 16	< 2.2	2 - 4	< 2	< 2	2.696 - 5.022
32	Fail	32.1 - 41.4	43 - 62	82.7 - 101.8	16 - 32	5.6 - 7.3	2 - 4	2 - 4	< 2	2.696 - 5.022
33	Pass	41.4 - 50.7	≥ 81	101.8 - 120.9	48 - 64	2.2 - 3.9	2 - 4	< 2	< 2	5.022 - 7.348
34	Fail	22.8 - 32.1	62 - 81	101.8 - 120.9	48 - 64	2.2 - 3.9	< 2	< 2	< 2	5.022 - 7.348
35	Pass	41.4 - 50.7	62 - 81	101.8 - 120.9	≥ 64	< 2.2	4 - 6	4 - 6	< 2	7.348 - 9.674
36	Pass	32.1 - 41.4	24 - 43	101.8 - 120.9	48 - 64	< 2.2	< 2	2 - 4	< 2	5.022 - 7.348
37	Pass	41.4 - 50.7	62 - 81	101.8 - 120.9	≥ 64	3.9 - 5.6	6 - 8	2 - 4	< 2	7.348 - 9.674



	Grade	As:1	As:2	As:3	As:4	Qz:1	Qz:2	Qz:3	Qz:4	S-I
33	Pass	41.4 - 50.7	≥ 81	101.8 - 120.9	48 - 64	2.2 - 3.9	2 - 4	< 2	< 2	5.022 - 7.348
34	Fail	22.8 - 32.1	62 - 81	101.8 - 120.9	48 - 64	2.2 - 3.9	< 2	< 2	< 2	5.022 - 7.348
35	Pass	41.4 - 50.7	62 - 81	101.8 - 120.9	≥ 64	< 2.2	4 - 6	4 - 6	< 2	7.348 - 9.674
36	Pass	32.1 - 41.4	24 - 43	101.8 - 120.9	48 - 64	< 2.2	< 2	2 - 4	< 2	5.022 - 7.348
37	Pass	41.4 - 50.7	62 - 81	101.8 - 120.9	≥ 64	3.9 - 5.6	6 - 8	2 - 4	< 2	7.348 - 9.674
38	Fail	41.4 - 50.7	62 - 81	≥ 120.9	≥ 64	2.2 - 3.9	< 2	< 2	< 2	5.022 - 7.348
39	Fail	32.1 - 41.4	≥ 81	≥ 120.9	≥ 64	3.9 - 5.6	< 2	4 - 6	< 2	< 2.696
40	Fail	32.1 - 41.4	≥ 81	82.7 - 101.8	48 - 64	2.2 - 3.9	< 2	4 - 6	< 2	< 2.696
41	Fail	22.8 - 32.1	62 - 81	101.8 - 120.9	48 - 64	3.9 - 5.6	2 - 4	< 2	< 2	< 2.696
42	Fail	41.4 - 50.7	≥ 81	101.8 - 120.9	≥ 64	2.2 - 3.9	2 - 4	< 2	< 2	2.696 - 5.022
43	Fail	41.4 - 50.7	≥ 81	≥ 120.9	16 - 32	2.2 - 3.9	< 2	≥ 8	< 2	2.696 - 5.022
44	Fail	41.4 - 50.7	≥ 81	101.8 - 120.9	≥ 64	3.9 - 5.6	< 2	2 - 4	6 - 8	2.696 - 5.022
45	Fail	22.8 - 32.1	43 - 62	82.7 - 101.8	48 - 64	< 2.2	< 2	< 2	< 2	< 2.696
46	Fail	41.4 - 50.7	≥ 81	82.7 - 101.8	32 - 48	2.2 - 3.9	< 2	2 - 4	< 2	2.696 - 5.022
47	Pass	≥ 50.7	≥ 81	≥ 120.9	≥ 64	≥ 7.3	4 - 6	≥ 8	≥ 8	≥ 9.674
48	Pass	≥ 50.7	≥ 81	≥ 120.9	≥ 64	3.9 - 5.6	2 - 4	6 - 8	< 2	≥ 9.674
49	Fail	32.1 - 41.4	62 - 81	≥ 120.9	48 - 64	2.2 - 3.9	< 2	≥ 8	< 2	< 2.696
50	Pass	22.8 - 32.1	≥ 81	63.6 - 82.7	32 - 48	3.9 - 5.6	< 2	< 2	< 2	5.022 - 7.348
51	Pass	41.4 - 50.7	≥ 81	63.6 - 82.7	16 - 32	3.9 - 5.6	< 2	< 2	< 2	5.022 - 7.348
52	Pass	41.4 - 50.7	≥ 81	≥ 120.9	≥ 64	≥ 7.3	≥ 8	≥ 8	2 - 4	≥ 9.674
53	Pass	41.4 - 50.7	≥ 81	82.7 - 101.8	≥ 64	5.6 - 7.3	< 2	≥ 8	2 - 4	5.022 - 7.348
54	Fail	41.4 - 50.7	≥ 81	101.8 - 120.9	48 - 64	2.2 - 3.9	2 - 4	6 - 8	< 2	2.696 - 5.022
55	Pass	≥ 50.7	≥ 81	≥ 120.9	≥ 64	2.2 - 3.9	4 - 6	≥ 8	< 2	7.348 - 9.674
56	Pass	32.1 - 41.4	≥ 81	< 63.6	≥ 64	3.9 - 5.6	< 2	< 2	< 2	5.022 - 7.348
57	Pass	≥ 50.7	≥ 81	82.7 - 101.8	48 - 64	≥ 7.3	2 - 4	≥ 8	< 2	5.022 - 7.348
58	Fail	32.1 - 41.4	≥ 81	< 63.6	32 - 48	≥ 7.3	< 2	2 - 4	< 2	2.696 - 5.022
59	Pass	≥ 50.7	62 - 81	101.8 - 120.9	≥ 64	3.9 - 5.6	< 2	4 - 6	< 2	5.022 - 7.348
60	Pass	41.4 - 50.7	62 - 81	≥ 120.9	48 - 64	≥ 7.3	2 - 4	6 - 8	2 - 4	7.348 - 9.674
61	Pass	41.4 - 50.7	62 - 81	≥ 120.9	≥ 64	3.9 - 5.6	< 2	≥ 8	< 2	5.022 - 7.348
62	Fail	22.8 - 32.1	62 - 81	82.7 - 101.8	< 16	2.2 - 3.9	2 - 4	4 - 6	< 2	2.696 - 5.022
63	Fail	32.1 - 41.4	62 - 81	63.6 - 82.7	16 - 32	2.2 - 3.9	< 2	6 - 8	< 2	2.696 - 5.022
64	Fail	41.4 - 50.7	≥ 81	82.7 - 101.8	48 - 64	< 2.2	< 2	2 - 4	< 2	< 2.696
65	Fail	41.4 - 50.7	62 - 81	< 63.6	48 - 64	3.9 - 5.6	< 2	< 2	< 2	< 2.696
66	Fail	41.4 - 50.7	62 - 81	≥ 120.9	48 - 64	2.2 - 3.9	2 - 4	6 - 8	< 2	5.022 - 7.348
67	Fail	41.4 - 50.7	62 - 81	63.6 - 82.7	32 - 48	2.2 - 3.9	< 2	2 - 4	< 2	2.696 - 5.022
68	Fail	41.4 - 50.7	≥ 81	≥ 120.9	≥ 64	2.2 - 3.9	< 2	2 - 4	< 2	< 2.696
69	Pass	41.4 - 50.7	62 - 81	≥ 120.9	≥ 64	2.2 - 3.9	< 2	6 - 8	2 - 4	5.022 - 7.348

Data Mining Project: Part 1 -> EDA Analysis

Boxplot After preprocessing:



Scatterplot After preprocessing: