



UTM
UNIVERSITI TEKNOLOGI MALAYSIA

MCSD1123 BIG DATA MANAGEMENT

Assignment 1 :

Data Analysis Using Google Sheet

Case Study 1a :

Examination Results

Lecturer's Name :

PM Dr. Mohd Shahizan Othman

Group Name :

F4

Group Members	Matrix Number
LYE KAH HOOI	MCS231010
LEE SEOW MING THERESA	MCS231013
THONG YEE MOON	MCS231001
SITI NORAFIZAH BINTI AB AZIZ	MCS231018

1.0 INTRODUCTION

A dataset named “dataset1.txt” was given with the purpose of examination result analysis. The dataset contained 111,519 records of *Id_No* starting from HP313300000 to HP313411518. Each *Id_No* had five examination results such as *Academic*, *Sports*, *Co-Curriculum*, *Test_1* and *Test_2*, with full marks as shown in Table 1 below.

Table 1:

Type of Examination	Full Marks
Academic	61
Sport	10
Co-Curriculum	15
Test_1	10
Test_2	10

Data analysis and data visualization process were expected to be completed using google sheet as a platform. With the help of google sheet, simple calculations and dashboard could be created using formula functions and charts.

First, examination result analysis was carried out by normalizing the full marks of each examination to a maximum value as 3.33 as shown in Table 2.

Table 2:

Type of Examination	Full Marks	Full Marks _{New}
Academic	61	3.33
Sport	10	3.33
Co-Curriculum	15	3.33
Test_1	10	3.33
Test_2	10	3.33

After that, the top three highest normalized marks of each *Id_No* were identified and added up to obtain a *Total Mark (TM)*. Percentage of *TM* was calculated. Grade and Status of each *Id-No* were assigned based on Percentage of *TM* as shown in Table 3.

Table 3:

Percentage of TM	Grade	Grade
90 - 100	A+	Pass
80 - 89	A	Pass
75 - 79	A-	Pass
70 - 74	B+	Pass
65 - 69	B	Pass
60 - 64	B-	Fail
55 - 59	C+	Fail
50 - 54	C	Fail
45 - 49	C-	Fail
40 - 44	D+	Fail
35 - 39	D	Fail
30 - 34	D-	Fail
0 - 29	E	Fail

Lastly, data visualization was done by creating a dashboard as shown in Figure 1 below:

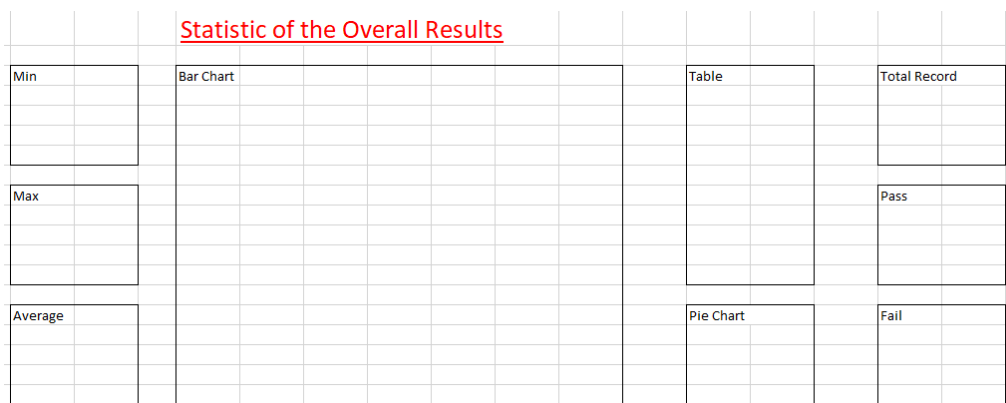


Figure 1: Dashboard design layout

2.0 METHODOLOGY - DATA PREPROCESSING

1. A new blank spreadsheet was created.
2. The “File” and “Import” buttons were clicked.

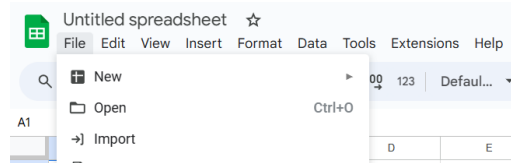


Figure 2: Google Sheet interface in order to import data

3. The dataset was then uploaded into google drive.

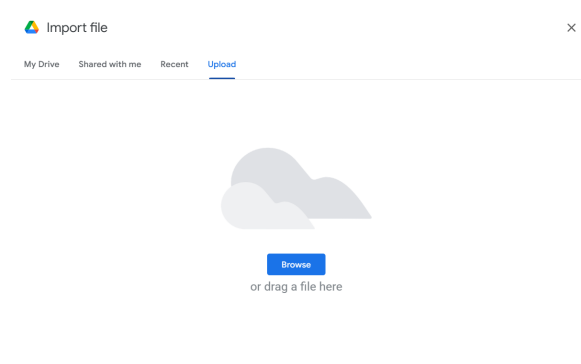


Figure 3 : Upload and import dataset file in Google Drive.

4. After uploading the dataset to google drive, the “Import data” button was clicked to import the dataset to google sheet. Then clicked “open now” to redirect to google sheet with imported data.

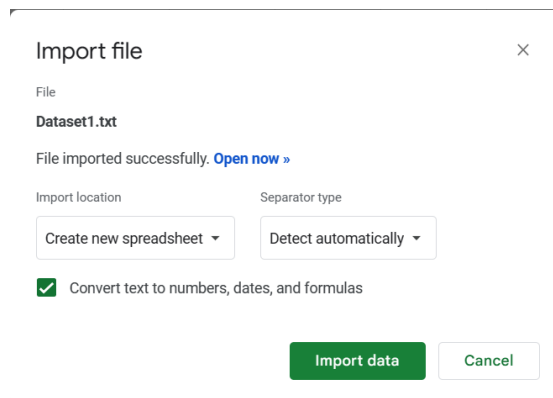

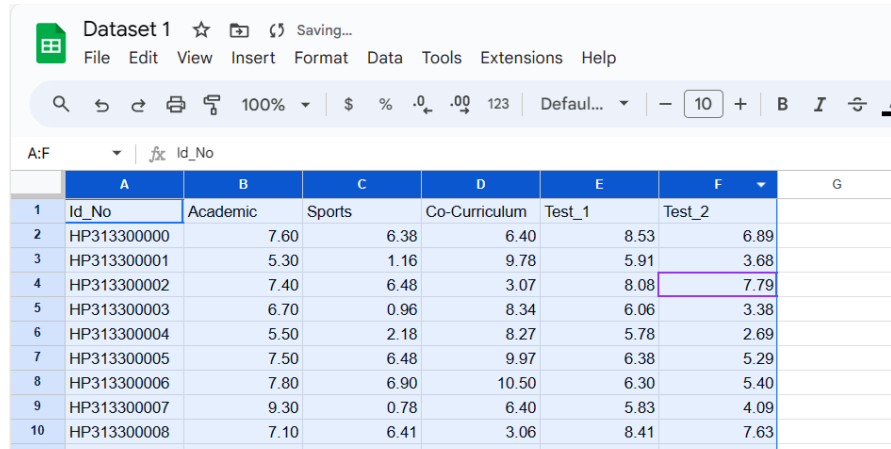


Figure 4: Import File

5. In order to convert data to two decimal places, Column B to Column F were highlighted for decimal place adjustment and the function “” on the toolbar was used to reduce the value to two decimal places.



	A	B	C	D	E	F	G
1	Id_No	Academic	Sports	Co-Curriculum	Test_1	Test_2	
2	HP313300000	7.60	6.38	6.40	8.53	6.89	
3	HP313300001	5.30	1.16	9.78	5.91	3.68	
4	HP313300002	7.40	6.48	3.07	8.08	7.79	
5	HP313300003	6.70	0.96	8.34	6.06	3.38	
6	HP313300004	5.50	2.18	8.27	5.78	2.69	
7	HP313300005	7.50	6.48	9.97	6.38	5.29	
8	HP313300006	7.80	6.90	10.50	6.30	5.40	
9	HP313300007	9.30	0.78	6.40	5.83	4.09	
10	HP313300008	7.10	6.41	3.06	8.41	7.63	

Figure 5: Column B to Column F with two decimal places

6. In order to generate the values which are more comparable in between each column of values, **value normalization** with calculation was carried out by assigning a new maximum value of 3.33 for each column of value. A general formula was used as below:

$$X_{new} = (X_{old} / \text{Full Marks}) * \text{Full Marks}_{New}$$

The calculations were slightly different from each other due to **different original Full Mark (refer to Table 2)** . The new labeling and calculation were shown in Table 4 below:

Table 4:

New Column	Column Name	Description	Formula
G	P1	Normalized mark for Academic with new maximum value of 3.33	=(‘Academic Value’/ 61)*3.33 =(B/61)*3.33
H	P2	Normalized mark for Sport with new maximum value of 3.33	=(‘Sports Value’/ 10) *3.33 =(C/10)*3.33
I	P3	Normalized mark for Co-Curriculum with new maximum value of 3.33	=(‘Co-curriculum Value’/ 15) *3.33 =(D/15)*3.33
J	P4	Normalized mark for Test_1 with new maximum value of 3.33	=(‘Test_1 Value’/ 10) *3.33 =(E/10)*3.33

K	P5	Normalized mark for Test_2 with new maximum value of 3.33	$(\text{'Test_2 Value'}/10)*3.33$ $= (F/10)*3.33$
---	----	---	---

	A	B	C	D	E	F	G	H	I	J
	Id_No	Academic	Sports	Co-Curriculum	Test_1	Test_2	P1	P2	P3	P4
2	HP313300000	7.60	6.38	6.40	8.53	6.89	0.4148852459			
3	HP313300001	5.30	1.16	9.78	5.91	3.68	0.2893278689			
4	HP313300002	7.40	6.48	3.07	8.08	7.79	0.4039672131			
5	HP313300003	6.70	0.96	8.34	6.06	3.38	0.3657540984			
6	HP313300004	5.50	2.18	8.27	5.78	2.69	0.3002459016			
7	HP313300005	7.50	6.48	9.97	6.38	5.29	0.4094262295			
8	HP313300006	7.80	6.90	10.50	6.30	5.40	0.4258032787			
9	HP313300007	9.30	0.78	6.40	5.83	4.09	0.5076885246			
10	HP313300008	7.10	6.41	3.06	8.41	7.63	0.3875901639			
11	HP313300009	5.40	5.86	1.54	8.21	7.48	0.2947868852			
12	HP313300010	6.70	7.16	11.34	6.76	4.68	0.3657540984			
13	HP313300011	8.40	6.56	8.64	7.66	4.38	0.458557377			
14	HP313300012	6.40	5.06	5.30	8.91	7.28	0.3493770492			
15	HP313300013	8.50	6.93	6.20	9.58	7.74	0.4640163934			

Figure 6: Creation of Column G (P1) using formula

- After completing the value normalization from P1 to P5 by assigning a new maximum value, the values from these columns were then converted to two decimal places using the same method in step 5.

	G	H	I	J	K
	P1	P2	P3	P4	P5
6.89	0.41	2.12	1.42	2.84	2.29
3.68	0.29	0.39	2.17	1.97	1.23
7.79	0.40	2.16	0.68	2.69	2.59
3.38	0.37	0.32	1.85	2.02	1.13
2.69	0.30	0.73	1.84	1.92	0.90
5.29	0.41	2.16	2.21	2.12	1.76
5.40	0.43	2.30	2.33	2.10	1.80
4.09	0.51	0.26	1.42	1.94	1.36
7.63	0.39	2.13	0.68	2.80	2.54
7.48	0.29	1.95	0.34	2.73	2.49
4.68	0.37	2.38	2.52	2.25	1.56
4.38	0.46	2.18	1.92	2.55	1.46
7.28	0.35	1.68	1.18	2.97	2.42

Figure 7: Column G (P1) to Column K (P5) with two decimal places

- Next, the top 3 highest mark values among Column G (P1) to Column K (P5) were determined and recorded from Column L (B1) to Column N (B3) separately. With the top 3 highest mark values, a total mark value was calculated labeled as TM at Column O.

Lastly, a percentage value of total mark value was then calculated and labeled as “Percent” at column P and had been made sure the value had been converted to two decimal places.

The formula was shown in Table 5 below.

Table 5:

Column	Column Name	Description	Formula
L	B1	Highest value among P1 to P5	=MAX (G : K)
M	B2	Second highest value among P1 to P5	=LARGE (G : K, 2)
N	B3	Third highest value among P1 to P5	=LARGE (G : K, 3)
O	TM	Total Mark Value from B1 to B3	=SUM (L : N)
P	Percent	Percentage value for Total Mark Value (TM)	=O/(3.33*3)*100

L	M	N	O	P
B1	B2	B3	TM	Percent
2.61	2.04	1.61	6.26	62.63
2.53	2.04	1.75	6.33	63.36
2.41	2.00	1.80	6.21	62.17
3.13	2.08	0.95	6.16	61.67
2.45	2.10	1.91	6.45	64.61
2.56	1.89	1.67	6.12	61.23
3.06	2.00	1.03	6.09	61.00
2.87	1.74	1.71	6.33	63.33
2.60	2.07	1.68	6.36	63.62
2.91	1.68	1.44	6.03	60.40
2.97	1.67	1.64	6.28	62.83

Figure 8: Column L (B1) to Column P (Percent) with two decimal places

- With the reference of the grading system shown in Table 3, a grade and status were able to be assigned to a particular *Id_No* based on Column P (Percent) that had been calculated.

The formula was shown in Table 6 below.

Table 6:

Column	Column Name	Formula
Q	Grade	=IFS(P>=90,"A+",P>=80,"A",P>=75,"A-",P>=70,"B+",P>=65,"B",P>=60,"B-",P>=55,"C+",P>=50,"C",P>=45,"C-",P>=40,"D+",P>=35,"D",P>=30,"D-",P>=0,"E")

R	Status	=IF(P>=65,"Pass","Fail")
---	--------	--------------------------

	P	Q	R
	Percent	Grade	Status
4	24.39	E	Fail
4	26.43	E	Fail
8	23.78	E	Fail
9	28.93	E	Fail
1	29.18	E	Fail
0	30.00	E	Fail
0	29.00	E	Fail
4	28.42	E	Fail
0	28.98	E	Fail
9	27.88	E	Fail

Figure 9: Column Q (Grade) to Column R (Status)

10. After grading, a conditional formatting rule was set at column P with Percent value in order to have better visibility. The percent value which fulfilled the pass requirement or referring to the status with “Pass” will then automatically turn to green. The setting of conditional format rules was shown as below.

Conditional format rules
×

Single color

Color scale

Apply to range

P2:P111520

Format rules

Format cells if...

Custom formula is

=R2="Pass"

Formatting style

Default

B
I
U
A

Cancel

Done

Figure 10: Conditional formatting rule for Column P (Percent) to turn green if Column R (Status) is “Pass”

11. Beside that, a conditional formatting rule was also set at the whole dataset when the status was “Pass”, the whole respective row will be converted to the red line automatically. The setting of conditional format rules was shown as below.

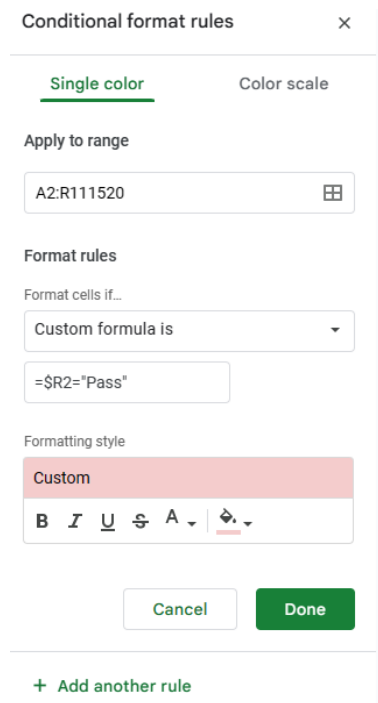


Figure 11: Conditional formatting rule for whole row to turn light red if Column R (Status) is “Pass”


12. Finally, an overview of the dataset after data preprocessing was shown as in Figure 12 below.

Dataset 1																			
File Edit View Insert Format Data Tools Extensions Help																			
Q Menus 100% \$ % 123 Default...																			
112 fx = (D12/15)*3.33																			
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	
1	Id.No	Academic	Sports	Co-Curricul	Test_1	Test_2	P1	P2	P3	P4	P5	B1	B2	B3	TM	Percent	Grade	Status	
2	HP313300000	7.60	6.38	6.40	8.53	6.89	0.41	2.12	1.42	2.84	2.29	2.84	2.29	2.12	7.26	72.67	B+	Pass	
3	HP313300001	5.30	1.16	9.78	5.91	3.68	0.29	0.39	2.17	1.97	1.23	2.17	1.97	1.23	5.36	53.70	C	Fail	
4	HP313300002	7.40	6.48	3.07	8.08	7.79	0.40	2.16	0.68	2.69	2.59	2.69	2.59	2.16	7.44	74.50	B+	Pass	
5	HP313300003	6.70	0.96	8.34	6.06	3.38	0.37	0.32	1.85	2.02	1.13	2.02	1.85	1.13	5.00	50.00	C	Fail	
6	HP313300004	5.50	2.18	8.27	5.78	2.69	0.30	0.73	1.84	1.92	0.90	1.92	1.84	0.90	4.66	46.61	C-	Fail	
7	HP313300005	7.50	6.48	9.97	6.38	5.29	0.41	2.16	2.21	2.12	1.76	2.21	2.16	2.12	6.50	65.02	B	Pass	
8	HP313300006	7.80	6.90	10.50	6.30	5.40	0.43	2.30	2.33	2.10	1.80	2.33	2.30	2.10	6.73	67.33	B	Pass	
9	HP313300007	9.30	0.78	6.40	5.83	4.09	0.51	0.26	1.42	1.94	1.36	1.94	1.42	1.36	4.72	47.29	C-	Fail	
10	HP313300008	7.10	6.41	3.06	8.41	7.63	0.39	2.13	0.68	2.80	2.54	2.80	2.54	2.13	7.48	74.83	B+	Pass	
11	HP313300009	5.40	5.86	1.54	8.21	7.48	0.29	1.95	0.34	2.73	2.49	2.73	2.49	1.95	7.18	71.83	B+	Pass	
12	HP313300010	6.70	7.16	11.34	6.76	4.68	0.37	2.38	2.52	2.25	1.56	2.52	2.38	2.25	7.15	71.60	B+	Pass	
13	HP313300011	8.40	6.56	8.64	7.66	4.38	0.46	2.18	1.93	2.66	1.46	2.66	2.18	1.93	6.66	66.60	B	Pass	

Figure 12: Full dataset from Column A to Column R

3.0 METHODOLOGY - DATA VISUALIZATION

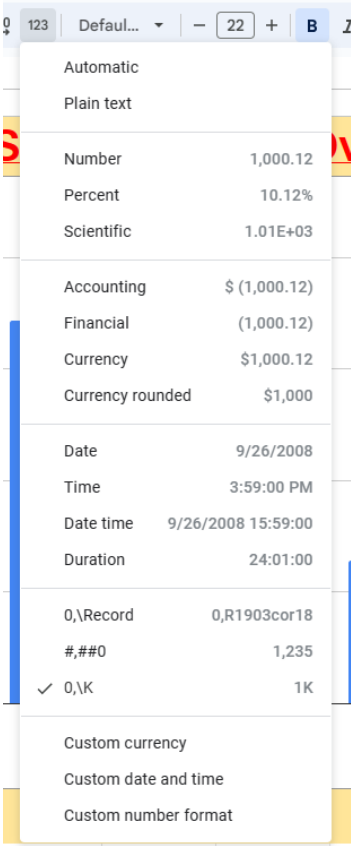
1. In order to have a simple visualization of the statistical value, few formulas were used to demonstrate the minimum Percent value, maximum Percent value, average Percent value, total number of records, percentage value of pass record and percentage of fail record.

Lastly the cells were then merged to have better visibility using “” button in the toolbar.

The formulas used and results were tabulated as follows.

Table 7 :

Statistical Value	Formula	Outcome
Minimum Percent Value	=MIN(Dataset1!\$P:\$P)	<div>Min</div> <div>14.83</div>
Maximum Percent Value	=MAX(Dataset1!\$P:\$P)	<div>Max</div> <div>98.67</div>
Average Percent Value	=AVERAGE(Dataset1!\$P:\$P)	<div>Average</div> <div>69.91</div>

<p>Total Number of Record</p>	<p>=COUNTA(Dataset1!A2:A)</p> <p>To format the data to from 111,519 to 112K, a number format was customized as “0,\K” as figure below.</p> 	<div data-bbox="1079 210 1429 451"> <p>Total Record</p> <p>112K</p> <p>111,519</p> </div>
<p>“Pass” Percentage Value & “Pass” Records</p>	<p>“Pass” Percentage Value =COUNTIF(Dataset1!R:R,"Pass")/Q4 = 65.16%</p> <p>“Pass” Records =text(COUNTIF(Dataset1!R:R,"Pass"), "0,000") & " Records" = 72,664 Records</p>	<div data-bbox="1079 1333 1429 1564"> <p>Pass</p> <p>65.16%</p> <p>72,664 Records</p> </div>

“Fail” Percentage Value & “Fail” Records	“Fail” Percentage Value =COUNTIF(Dataset1!R:R,"Fail")/Q4 = 34.84% “Fail” Records =text(COUNTIF(Dataset1!R:R,"Fail"), "0,000") & " Records" = 38,855 Records	<div> <div>Fail</div> <div>34.84%</div> <div>38,855 Records</div> </div>
--	---	--

2. Next, to summarize the data, a pivot table was inserted with the given setting as below.

<i>Grade</i>	<i>Data</i>
A+	8,588
A	24,264
A-	12,464
B+	13,956
B	13,392
B-	10,677
C+	7,803
C	7,566
C-	5,982
D+	3,436
D	1,664
D-	1,019
E	708
	111,519

Figure 13 : Number of records by Grade

3. A bar chart of the number of *Id_No* (Student) versus Grade was created as shown in Figure 14, and the “setup” in the chart editor which represented the setting of the chart was shown in Figure 15.

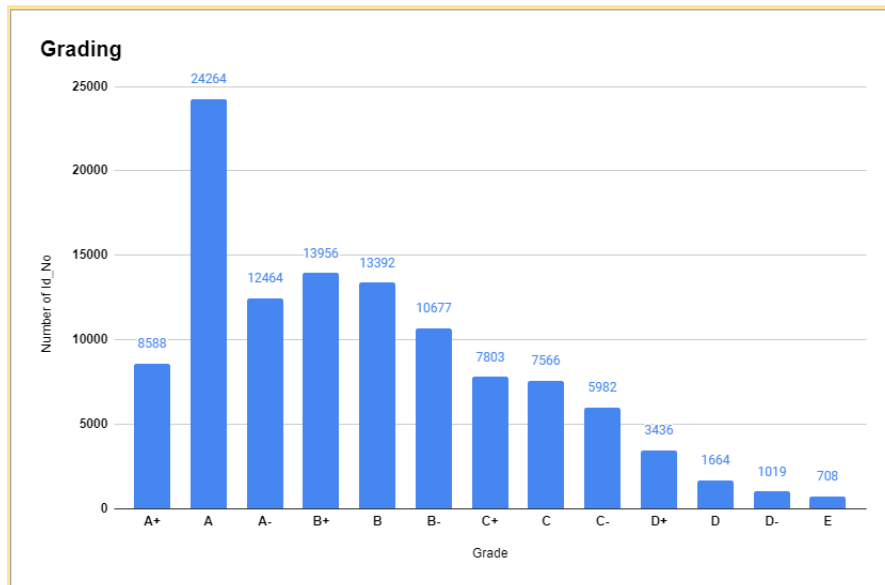


Figure 14 : Bar chart of Number of *Id_No* versus Grade

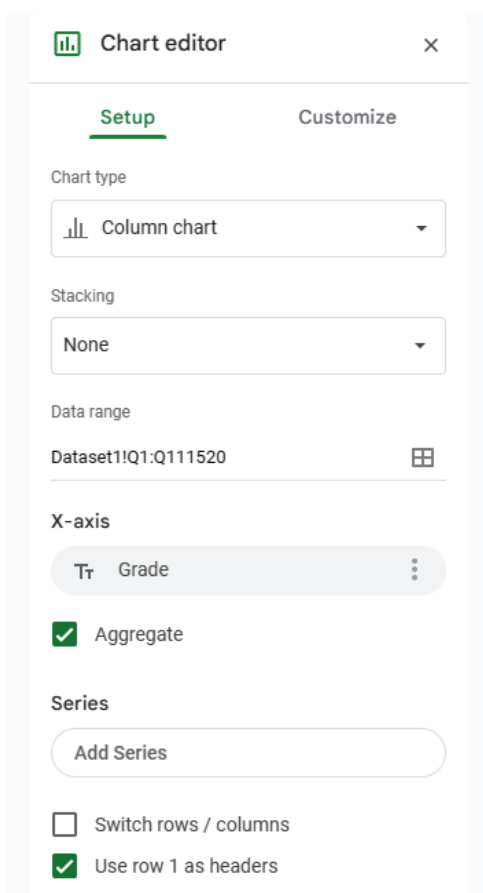


Figure 15: “Setup” in the chart editor for Figure 14

4. Lastly, a pie chart of pass and fail percentage was created as shown in Figure 16, and the “setup” in the chart editor which represented the setting of the chart was shown in Figure 17.

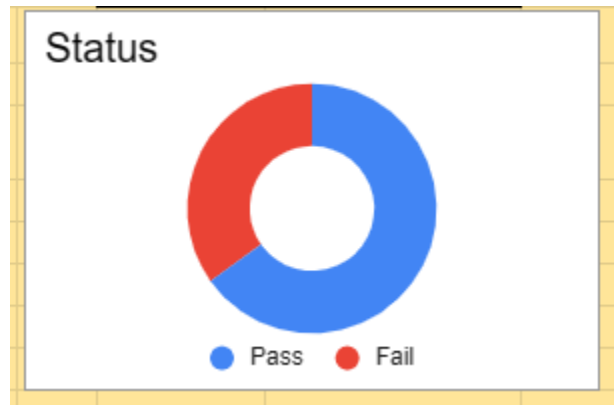


Figure 16 : Pie Chart for Pass and Fail Percentage

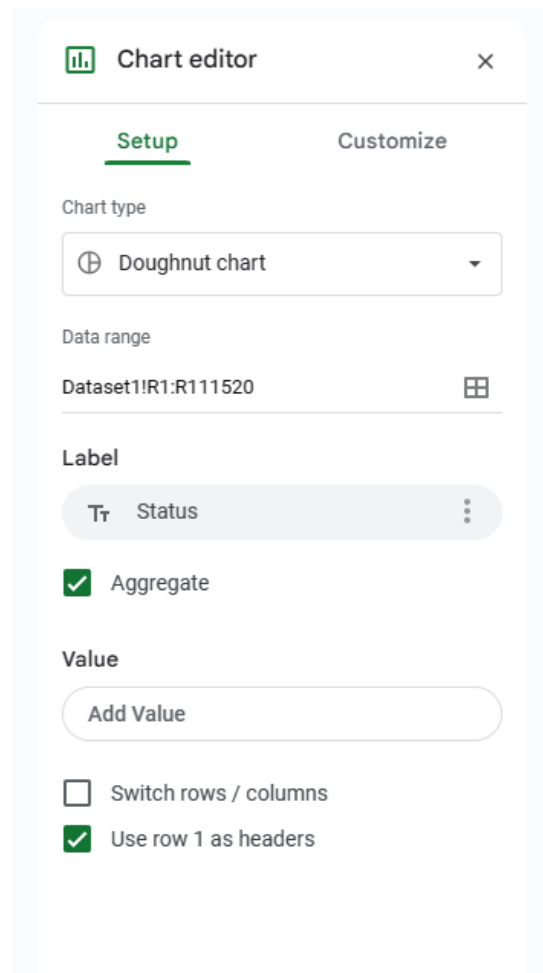


Figure 17 : “Setup” in the chart editor for Figure 16

5. A dashboard was successfully created as shown in Figure 18 below.

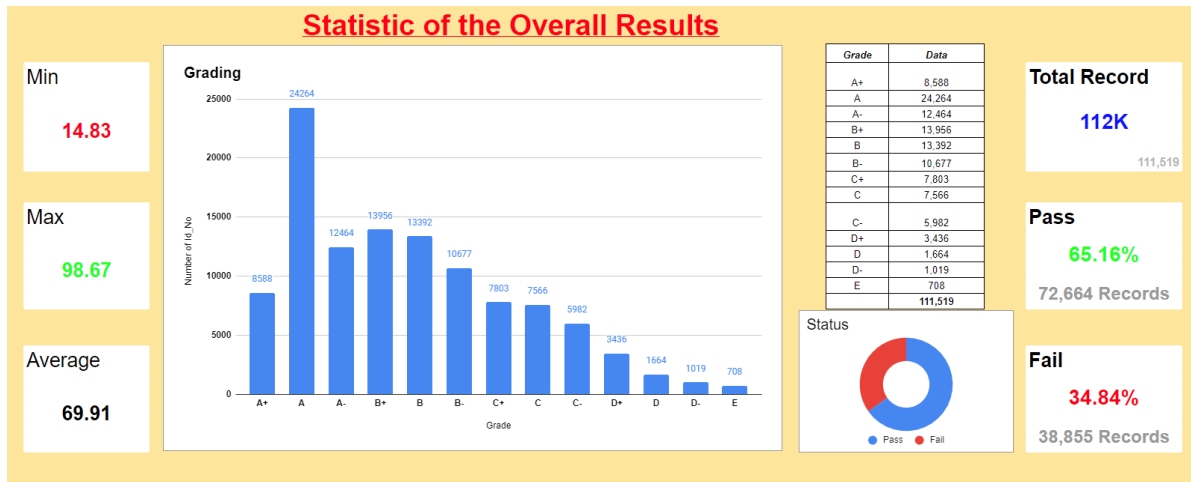


Figure 18: Dashboard showing overall results

4.0 CONCLUSION

Through the above dashboard shown in Figure 18, we can see that approximately 65% examinees passed the examinations and we have the most number of the examinees scored A and least number of the examinees scored E. Average score is 69.91, minimum score is 14.83 and maximum score is 98.67.