1 Q2 – Problem definition and data analysis.

1.1 Data exploration.

The data set is representing information collected from different participants and it contains information about data analytics students. The given data consist of 19 variables and 109 observations and it comes in CSV format.

The data collected through Microsoft survey forms which allow input of different values and types of data. A closer look into the data we will figure out that the data might be collected from two different sources or followed two different formats and that is clear by checking the different timestamp formats later in our analysis.

The records show that the period of the survey was 27 days starting from 17th of December to 12th January, the data represent the current students who are taking data analytics module from different background, cities, ages, and even physical characteristics such as their height and gender, a glimpse of the data (figure3).

ID	Start time	Completion ti...	Email	Gender	Age	Closest bigge...	Number of kid...	Birth Year	Number of mo...
2	17-Dec-20 9:09:3...	17-Dec-20 9:10:5...	anonymous	Woman	40-49	Lagos	0	1975	5
3	17-Dec-20 9:09:2...	17-Dec-20 9:11:0...	anonymous	Man	30-39	Dhaka	0	1011986	5
4	17-Dec-20 9:09:3...	17-Dec-20 9:14:3...	anonymous	Man	30-39	Hanoi	0	1989	5
5	17-Dec-20 9:19:5...	17-Dec-20 9:27:0...	anonymous	Woman	50-59	Stockholm	0	1970	5
6	17-Dec-20 9:28:1...	17-Dec-20 9:35:4...	anonymous	Man	40-49	Accra	3	1979	5

Average mark ...	Prediction for ...	I enjoy workin...	I am excited a...	I am afraid of ...	I am intereste...	I plan to work ...	Height (absol...	Height (rel
62.000	62	Agree	Agree	Completely Disa...	Agree	Agree	175.00	very tall
57.000	65	Agree	Agree	Completely Disa...	Agree	Agree	172.00	average
61.000	64	Agree	Meh	Disagree	Meh	Meh	178.00	average
60.000	100	Agree	Agree	Completely Disa...	Agree	Completely Agree	164.00	average
60.000	70	Completely Agre	Agree	Completely Disa...	Completely Agree	Meh	170.00	average

Figure 1 Overview of the dataset

1.2 Business projects and Audience

1.2.1 Project 1:

Analyze the data to understand the motivations and characteristics of the current data analytics program students and create a segmentation to help the marketing team enhancing their targeting techniques and increase their admission rate.

Questions:

Q1. Does age impact people's interest to study data analytics?
Q2. Is gender important in segmentation targeting?
Q3. can the geographical location of current students help to enhance our marketing segmentation?

Audience:

Robert Kennedy College marketing team.

1.2.2 Project 2

The second business problem will be directed to the learning board of the college, to understand if the new students understand the marking grade system in United Kingdom college's higher education and ensure they address this issue probably in their induction module.

Questions:

1. How many marks new students expect to get?
2. Do their expectations change after completing some modules?
3. After how many modules their expectations changed?

Audience

Robert Kennedy College induction team leaders.

1.3 Project 1 analysis:

Matching to data: The Age, Gender, and location are ideal variable to the problem, in addition to that we will be creating new variables for better analysis and hypothesis, which are calculated age variable that contains the current age of each participant using their birth year column, generate the country, continent, and population for better geological analysis (figure4)

ID	Country/region	Continent	Population	calculated Age
1	Nigeria	Africa	200,963,599	46
2	Bangladesh	Asia	163,046,161	35
3	Vietnam	Asia	96,462,106	32
4	Sweden	Europe	10,285,453	51
5	Ghana	Africa	30,417,856	42
6	South Africa	Africa	58,558,270	44
7	Singapore	Asia	5,703,569	38
8	Switzerland	Europe	8,574,832	34
9	South Africa	Africa	58,558,270	44
10	United States	North America	328,239,523	37

Figure 2 The new variables

.

1.3.1 Exploration and hypothesis:

1.3.1.1 Age

Age group	Count of students
20-29	9
30-39	32
40-49	42
50-59	25
70+	1
Grand Total	**109**

The table shows that out of 109 records of our dataset there is a noticeable difference between groups, the 40-49 group is the highest followed by 30-39. Let's use our calculated age variable to check the distribution of the age. (figure5)

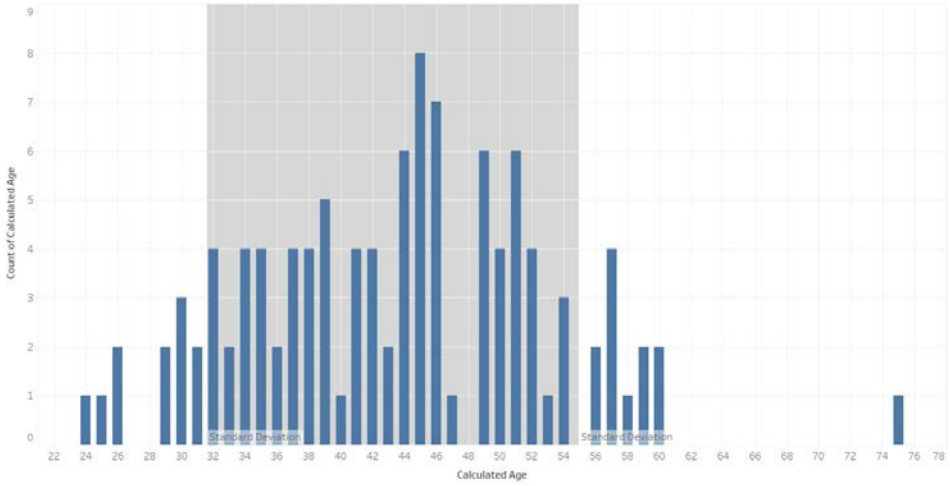


Figure 3 calculated age distribution

The calculated age explained the age that falls within the grey area between 32 and 55 are representing most of the population in the course. Now let's analyze how students feel toward studying data against age to check if there is any difference between them.

	I am excited about working with data				
Age	Agree	Completely Agree	Completely Disagree	Disagree	Meh
20-29	33.33%	44.44%			22.22%
30-39	50.00%	31.25%		3.13%	15.63%
40-49	52.38%	33.33%	2.38%	2.38%	9.52%
50-59	24.00%	40.00%			36.00%
70+	100.00%				

	I plan to work hard in this module			
Age	Agree	Completely Agree	Disagree	Meh
20-29	22.22%	77.78%		
30-39	43.75%	46.88%		9.38%
40-49	33.33%	61.90%		4.76%
50-59	32.00%	60.00%	4.00%	4.00%
70+		100.00%		

	I am interested in understanding Data Analytics			
Age	Agree	Completely Agree	Disagree	Meh
20-29	11.11%	66.67%		22.22%
30-39	50.00%	46.88%		3.13%
40-49	26.19%	66.67%	2.38%	4.76%
50-59	44.00%	48.00%		8.00%
70+		100.00%		

From the tables explain how students feel toward study data analytics within their respective groups, we notice group 40-49 has the most agree and completely agree percentage followed by group 30-39. Therefore, we will put age in our marketing project consideration.

1.3.1.2 Gender

Our main concern here is to find out if there is any difference between men and women who are studying data analytics.

Gender	Count of Gender	Percentage of total
Man	64	58.72%
Woman	45	41.28%
Total	109	100.00%

From the above graph, we can find that men have a greater number than women in our dataset, about 17% of the difference.

Despite men are more than women, that can only be our sample and it might differ in other samples, so let's see if men are more successful than women by running a descriptive statistic against average marks so far.

	women	*men*
Mean	62.26190476	63
Standard Error	0.845881959	0.768309
Median	62	63
Mode	60	60
Standard Deviation	5.481941635	5.90149
Sample Variance	30.05168409	34.82759
Kurtosis	-0.318999631	0.444323
Skewness	-0.11439061	0.5887
Range	25	27
Minimum	50	51
Maximum	75	78
Sum	2615	3717
Count	42	59

Figure 4 descriptive statistics between gender

The mean of 62.26 is very similar to the men's mean of 63, also the median and standard are very close, the data range of 25 and 27 explained that marks are equally distributed among the two samples.

	women
Mean	62.26190476
Variance	30.05168409
Observations	42
Pooled Variance	32.84968735
df	99
t Stat	-0.637876713
P(T<=t) one-tail	0.262513299
t Critical one-tail	1.660391156

t-Test: Assuming Equal Variances

Figure 5 t-test of gender

To be certain, a t-test conducted and gave a *p-value* of *0.26,* it is greater than the alpha of 0.05 which means there is no significant difference. Therefore, we will not take gender into consideration for our marketing campaign.

1.3.1.3 Location

The geographical location could be vital in targeting new joiners thus we will analyze it through our biggest city near you.

Cities	Count of cities
Zurich	6
Dubai	5
singapore	4
Johannesburg	4
Lagos	4
New York	3
Nairobi	3
Accra	3
Toronto	3
Oman	2
Ho Chi Minh City	2
Hanoi	2
Prague	2
Brussels	2
Riyadh	2
Harare	2
Shanghai	2

Figure 6 city count

Distinct count of Closest biggest city near you	69
Distinct count of Country/region	45

Zurich comes at the top of the table followed by Dubai, but since we do not have an accurate location, analysis of cities could be misleading. Hence, we will use countries in our analysis for the wider picture.

Countries	Count of Countries
Switzerland	9
United States	8
Nigeria	6
South Africa	5
United Arab Emirates	5
Ghana	4
United Kingdom	4
Canada	4
Singapore	4
Vietnam	4
Italy	3
Iraq	3
Democratic Republic of the Congo	3
Zambia	3
Kenya	3
Belgium	2
Israel	2

	Country	Count
1	Switzerland	9
2	United States	8
3	Nigeria	6
4	South Africa	5
5	United Arab Emirates	5
6	Canada	4
7	Ghana	4
8	Singapore	4
9	United Kingdom	4
10	Vietnam	4
11	Democratic Republic of the Congo	3
	Total	**56**
	Percentage	**51.38%**

Figure 7 country count

Another interesting finding that out of 45 countries only the top 11 countries represent over 50% of the total students which indicates that some countries contribute more than others.

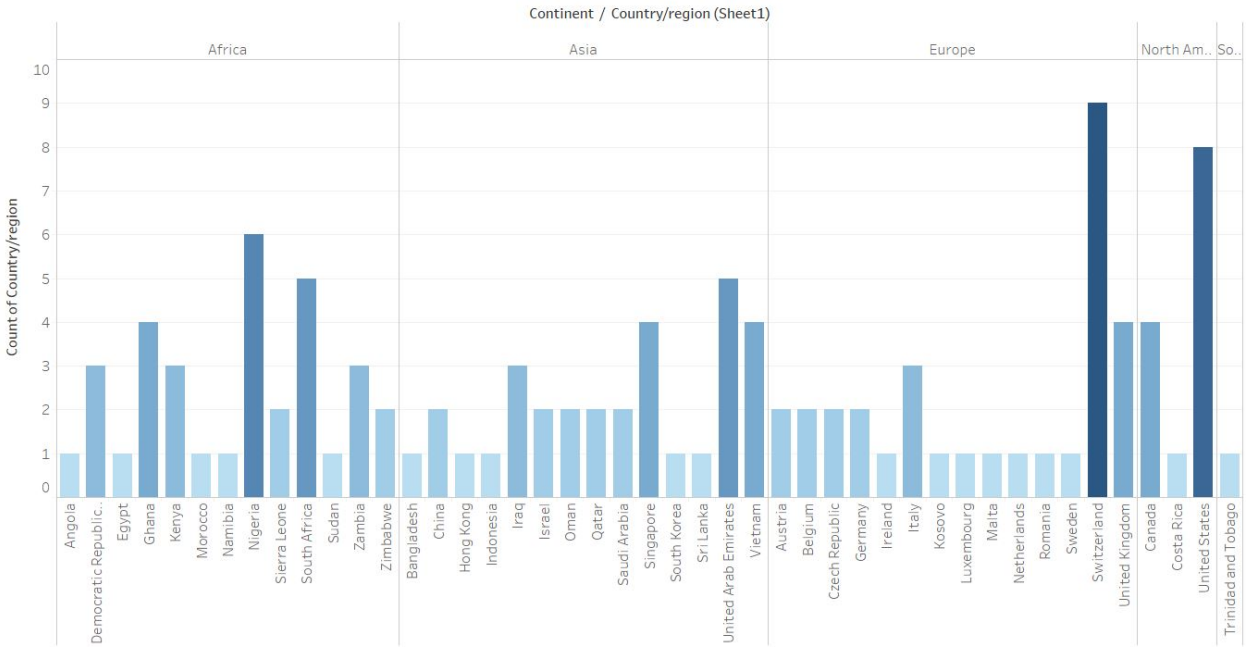


Figure 8 country and continent distribution

When we look at the continent view, Surprisingly, Africa came first in our analysis close to Asia and Europe however north and south America they are very low, thus we can conclude that there are overlooked countries that need more attention of our marketing team.

Africa	33
Asia	31
Europe	31
North America	13
South America	1

1.3.2 Conclusion of the analysis

From our previous analysis, we can conclude that age and location are important variables to be considered when conducting a marketing campaign, gender did not show a significant difference so it will not be considered.

1.4 Project2 analysis:

Matching to data: prediction for this module and average mark so far along with the number of modules completed will be used to check the average of how much students expect to score grouped by the number of their completed modules.

1.4.1 Data exploration

Number of modules completed so far	Count of Number of modules completed so far	Avg. Average marks so far	Avg. Prediction for this module's mark
1	7.00	72.80	75.50
2	9.00	63.50	72.44
3	12.00	64.58	64.25
4	4.00	61.50	63.25
5	74.00	61.65	66.26

The number of students who completed 5 modules is greater than the rest of the groups which might indicate the preferences of students to study data analytics after finishing their other modules.
This table shows an interesting point also which is the group who completed fewer modules score more than the ones who completed more, that can have many reasons that we do not have the answer in our dataset, such as the first modules might be easier, or the students might be studying harder at the beginning rather than towards the end.

However, when we look at the prediction of the module's marks, we can notice a difference between the student who completed one or two modules and the rest of the groups, which means there is a difference when we increase the module number. To ensure that there is a significant difference we will run an ANOVA analysis between the groups.

1.4.2 Hypothesis testing

The test of prediction for this module and the number of modules shows *p-values* less than 0.05 our alpha which means there is a significant difference between groups.

Analysis of Variance, Predictionforthismodulesmark Numberofmodulescompletedsofar:

	Sum of squares	df	Mean square
Treatment	940.51	4	235.127
Residual	6038.6	98	61.6183
Total	6979.11	102	68.4226

Grand mean = 66.6602

Level	n	mean	std. dev
1	6	75.5	9.9750
2	9	72.4444	6.6165
3	12	64.25	9.5644
4	4	63.25	6.9940
5	72	65.7917	7.5413

F(4, 98) = 235.127 / 61.6183 = 3.81587 [p-value 0.0063]

Figure 9 ANOVA for preditction_of_this_module

Likewise, the test of average marks so far shows *p-values* less than our alpha which means there is a significant difference between groups.

Analysis of Variance, Averagemarkssofar Numberofmodulescompletedsofar:

	Sum of squares	df	Mean square
Treatment	642.449	4	160.612
Residual	2623.04	96	27.3233
Total	3265.49	100	32.6549

Grand mean = 62.6931

Level	n	mean	std. dev
1	5	72.8	5.8907
2	8	63.5	7.3872
3	12	64.5833	6.1120
4	4	61.5	4.7958
5	72	61.6528	4.7801

F(4, 96) = 160.612 / 27.3233 = 5.87822 [p-value 0.0003]

Figure 10 ANOVA for Average_mark_so_far

1.4.3 Conclusion of the analysis.

We can conclude from the previous hypothesis that there is a change of behavior of predicting the score of this module after the second module, average marks test also supports our analysis but will not be taken into considerations because of the reasons mentioned earlier.

2 Q3 – Presentation of the analysis results.

2.1 Project 1 Mission:

Finding valuable information for Robert Kenndey college marketing team to increase enrollment rate.

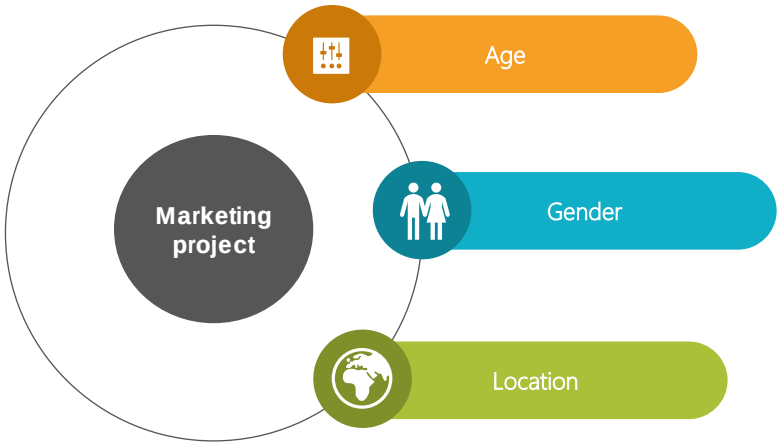
2.2 Approach:

Analyze the most relevant data variables in our dataset and find the most useful element to adopt (age, gender, and location).

Project 1 Analysis

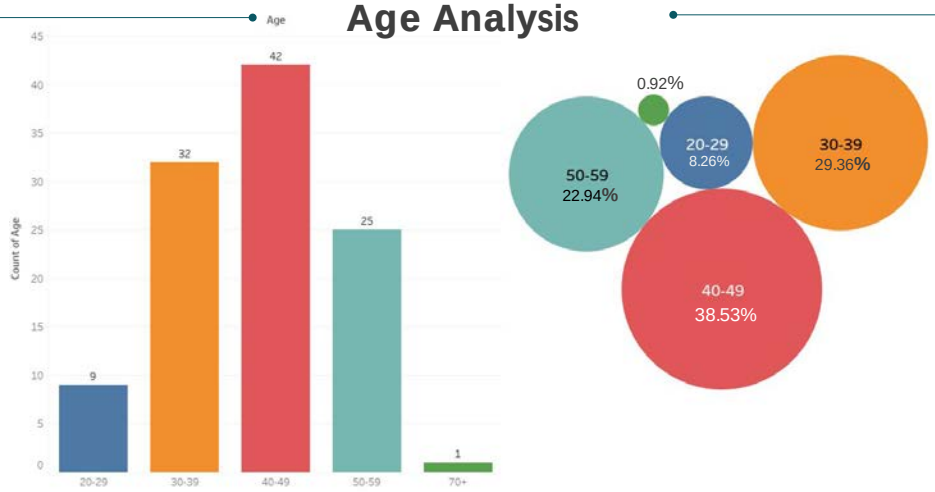
The goal of this project is to investigate the three factors that may be beneficial for our marketing team.

Exploring the points that need more attention and ensure targeting is directed to the correct potential students.



Let start by exploring the Age variable to decide whether if it has any significance or not.

Age Analysis



From 40 to 49

38.53%

40-49 age has the highest percentage of our students currently enrolled in data analytics course

From 30 to 39

29.36%

30-39 comes the second with almost 30% of total

From 50-59

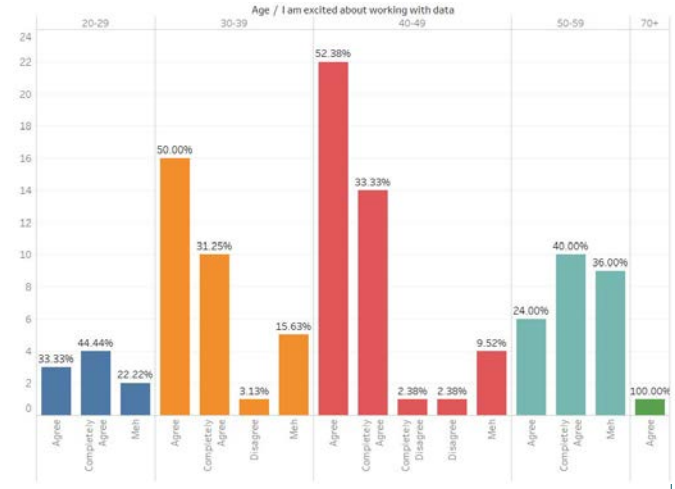
22.94%

50-59 comes as third but with high percentage of almost the quarter of the total.

From the above visualization we recognize that most of those who currently studying data analytics age between 40 to 49 followed by 30 to 39, both categories have almost 68% of total students. +70 group has

only one record so we cannot consider it in our analysis. Now after we know that let's see students' feelings of working with data according to their respective ages.

Age/ excited about working with data



40-49 85.71% Excited

30-39 81.25% Excited

50-59 64.00% Excited

20-29 77.77% Excited

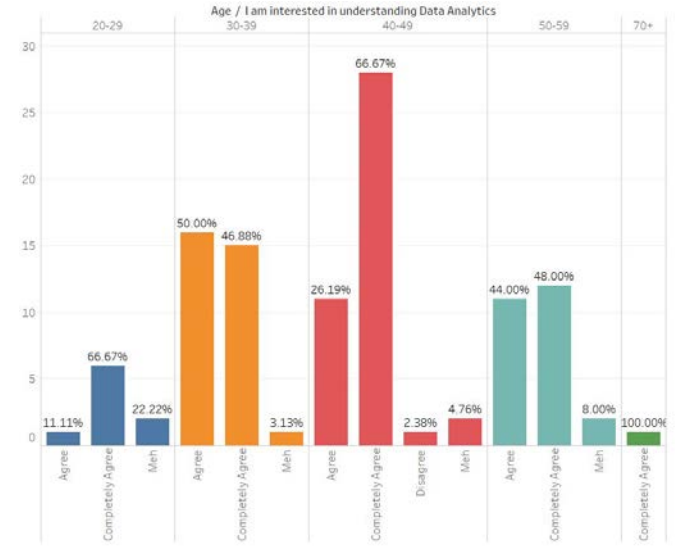
Exploring how student excited about working with data based on their age?

Agree - completely agree (Positive)

Disagree – completely disagree – Meh (Negative)

According to how excited students are working with data we see that the same groups of 40-49 and 30-39 have great positive excitement among the rest of the groups.

Age/ interested in understanding data



40-49 92.86% Interested

30-39 96.88% Interested

50-59 92.00% Interested

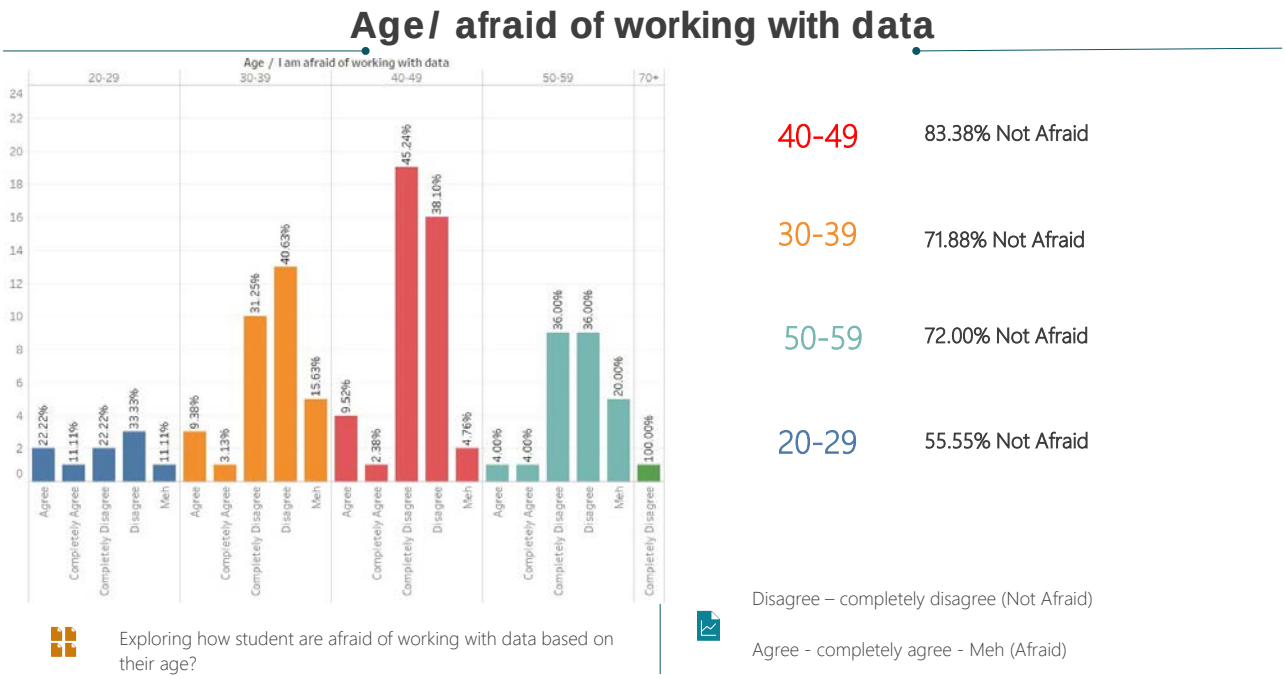
20-29 77.78% Interested

Exploring how student interested to understand data analytics based on their age?

Agree - completely agree (Positive)

Disagree – completely disagree – Meh (Negative)

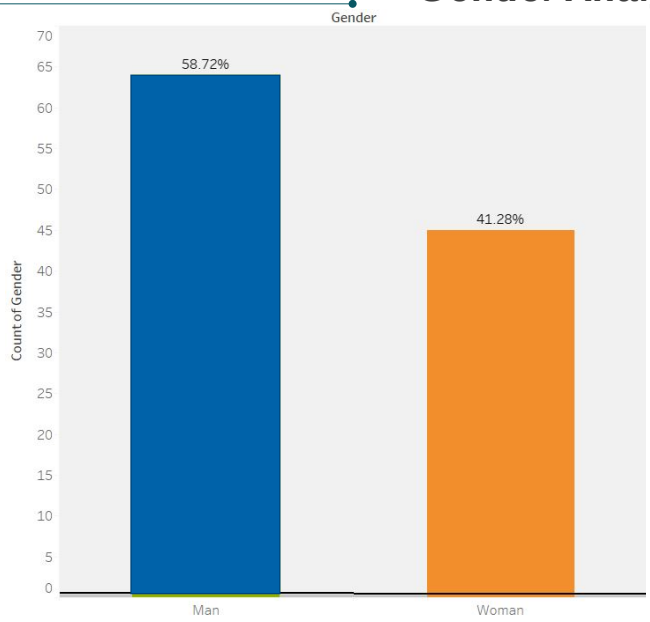
Likewise, the previous variable, but this time is how interested students are. The most noticeable difference here is group 30-39 has the highest percentage and they seem to have more interested than others.



Another factor that could help us to understand our students depend on their age is how are they afraid of working with data, our findings confirmed the previous results as group 40-49 are the most confident to work with data followed by 30-39 and 50-59 then the least confident group is 20-20 with almost 30% difference of group 40-49.

Now let's explore our second variable which is gender.

Gender Analysis



Men

58.72%

Men are more than women, 64 out of 109 are Men.

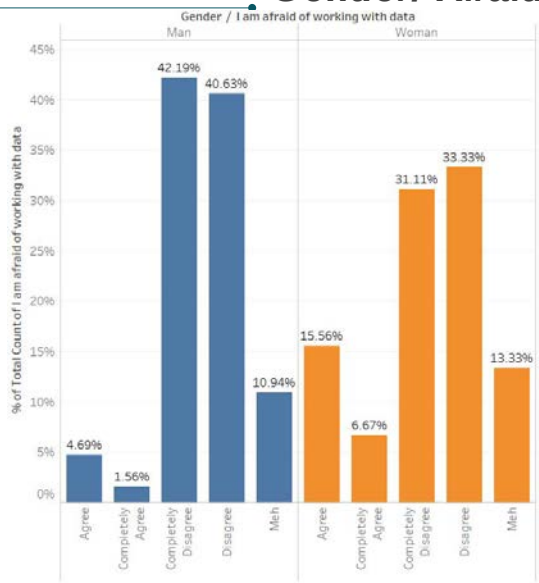
Women

41.28%

Women are less with 46 out of 109.

Men appear to be more than women in RKC the current data analytics course almost 17% greater than women. But are men also more confident working with data?

Gender/ Afraid of working with data



Exploring how student are afraid of working with data based on their gender?

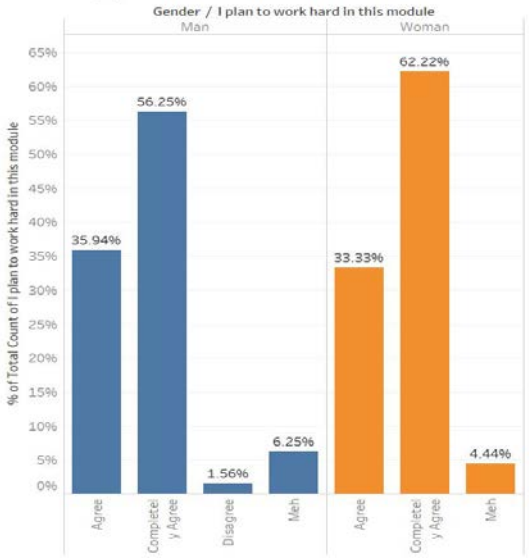
6.25% of Men Are afraid of working with data

22.23% of women Are afraid of working with data

Agree - completely agree (Afraid)

This representation shows that women are more likely to be scared working with data more than men, women who are not scared to work with data are 64.44% of total women, but in men, it's 82.82% of all men feel more confident working with data.

Gender/ Planning to work hard



Exploring if students are planning to work hard with data based on their gender?

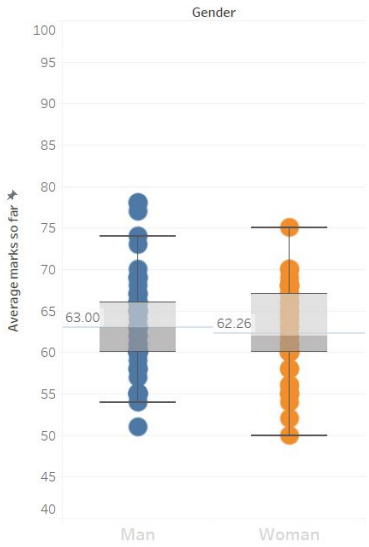
92.19% of Men Are planning to work hard in this module

95.55% of women Are planning to work hard in this module

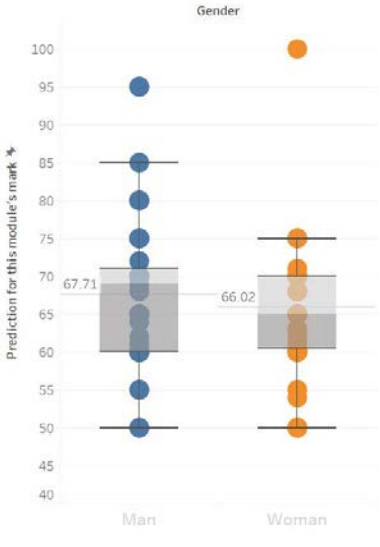
Agree - completely agree (work harder)

Thus, this will take us to consider another factor showing the student's intention to work hard on this module based on their gender, women are planning to work 95.55% harder while men 92.19%. Hence, women are planning to work harder in this module than men, this finding might explain the previous outcome that showed the ratio of women is less confident working with the data. That takes us to conclude that however, women are more afraid dealing with data they are planning to work harder than men.

Gender marks' Analysis



Boxplot for average marks so far between men and women

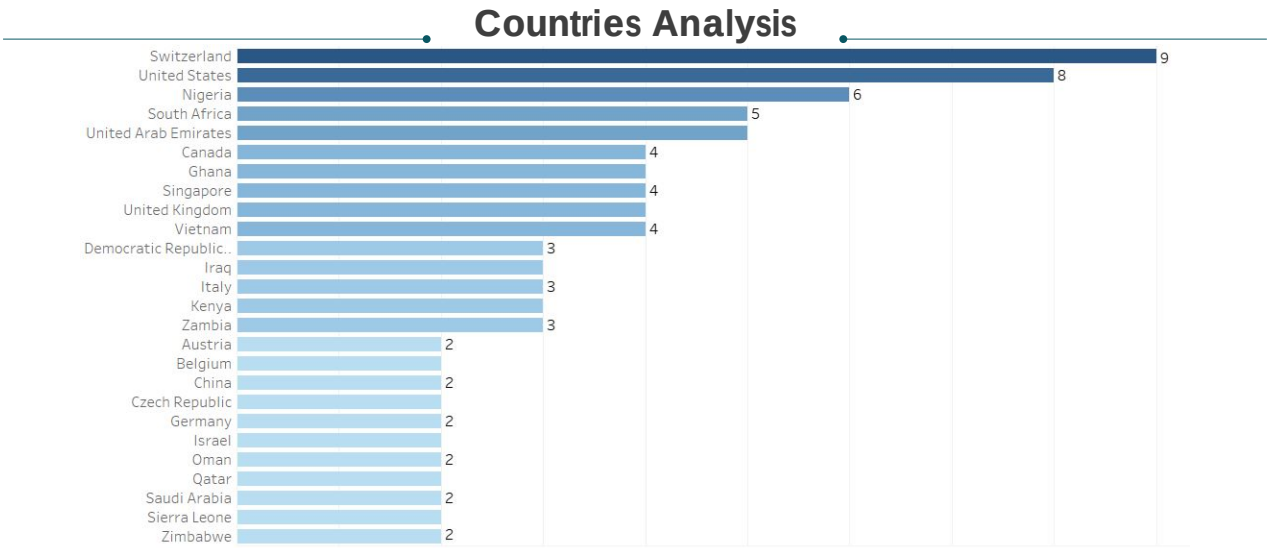


Boxplot for prediction for this module's mark between men and women

The average marks so far explain how each gender did in their previous modules the boxplot shows similarities between them with an average of 63.00 for men and 62.26 women, likewise, the boxplot for prediction of this module's marks shows a similar average of 67.71 for men and 66.02 for women despite some influential points which means mean did not differ than women on their previous marks or even their prediction of data analytics module.

Now let's explore the location variable.

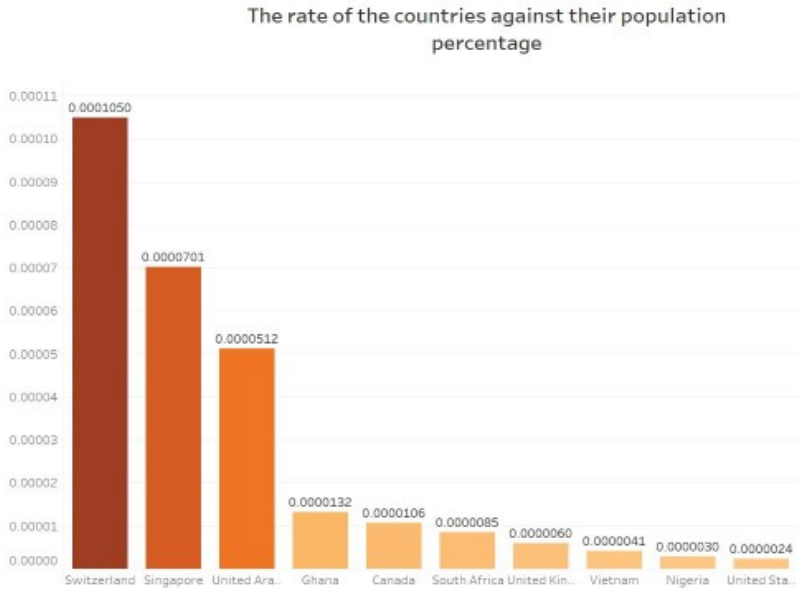
The top cities with at least 2 students currently studying data analytics, Zurich at the top with 6 followed by Dubai 5 student, then Johannesburg, Lagos and Singapore by 4 students each.



The top countries with at least 2 students currently studying data analytics, Switzerland 9, followed by United States 8, and Nigeria 6.

Countries may vary in their population, since we will use our analysis for geographical segmentation, we will consider the most successful countries comparing the number of students to the population of the country.

Countries/population Analysis

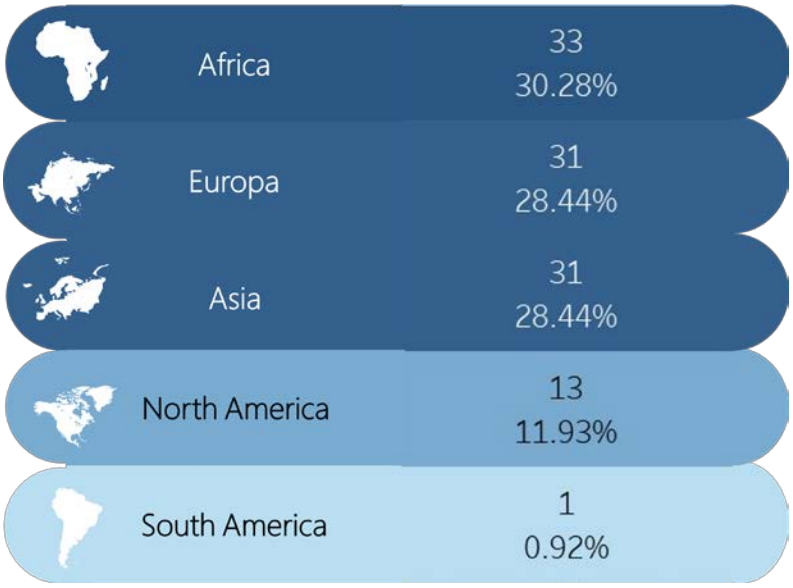


The analysis of the top 10 countries confirms that Switzerland has the highest rate, but surprisingly Singapore, and UAE came after. USA and Nigeria came last because of their high population against their number of students.

Location (map) Analysis

The map shows the location of our students relative to their countries, the map explained most students live in African, Europe, and Asia. The following table has detailed numbers for each continent.

Location Analysis



Therefore, we conclude that location may affect targeting students, some countries, even continents need more attention from the marketing team. And the overall conclusion will be as followed:

2.3 Analysis conclusion and outcome.

Project 1 Outcome

Significant

Age

› Age is considered a significant factor on targeting potential data analytics students, as analysis explained age can determine if the students will be interested on studying data analytics more than others.

Location

› The student's location also plays an important role of figuring out where our most of student live and what are the overlooked places to direct our marketing to, such as the whole continent of Australia is missed and very low enrolment from South America.

Not Significant

Gender

› Despite of men are greater than women, analysis indicated that there is no different between women and men on their potential of studying data analytics.

2.4 Project 2 Mission:

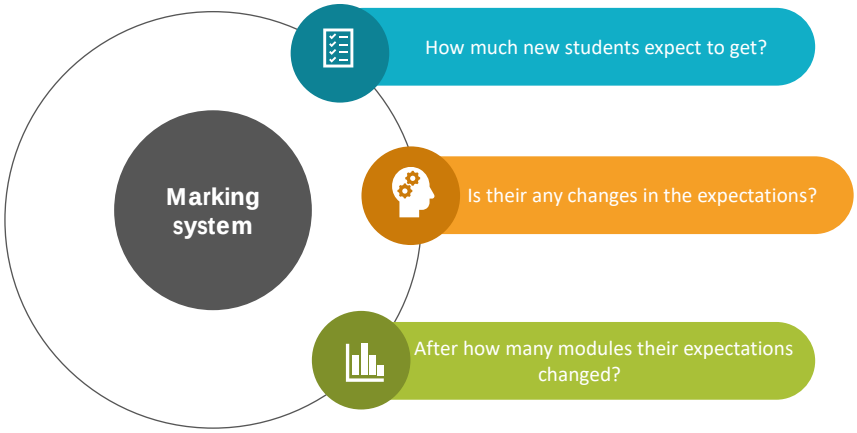
To find out how familiar newer students with the grading system.

2.5 Approach:

Explore and analyze how students predict their marks and if there any difference between students who took more modules than others.

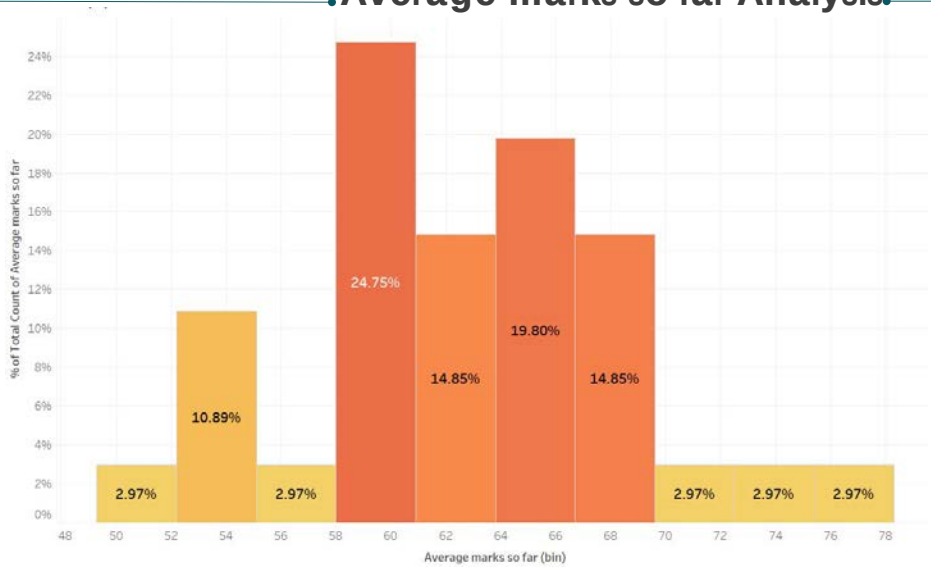
Project 2 Analysis

The goal of this project is to analyze the prediction marks of the data analytics module and the previous modules score against the number of modules completed to understand the knowledge of the students about marking system.



Before we answer these questions, we will try to understand how the students did in their previous module's grades.

Average marks so far Analysis

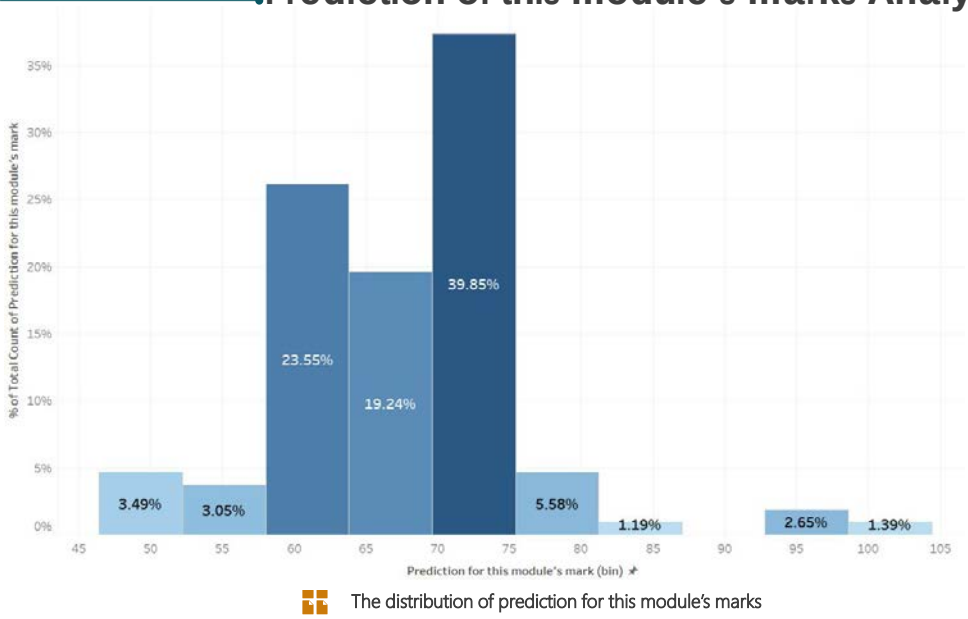


The distribution of average marks so far

74.25%
of the population score between 58 to 69

24.75% of students scored from 58 to 60 but most of the students did not exceed 70, only 8.91% manage to score over 70. Now let's explore what are the expectation of our data analytics course.

Prediction of this module's marks Analysis



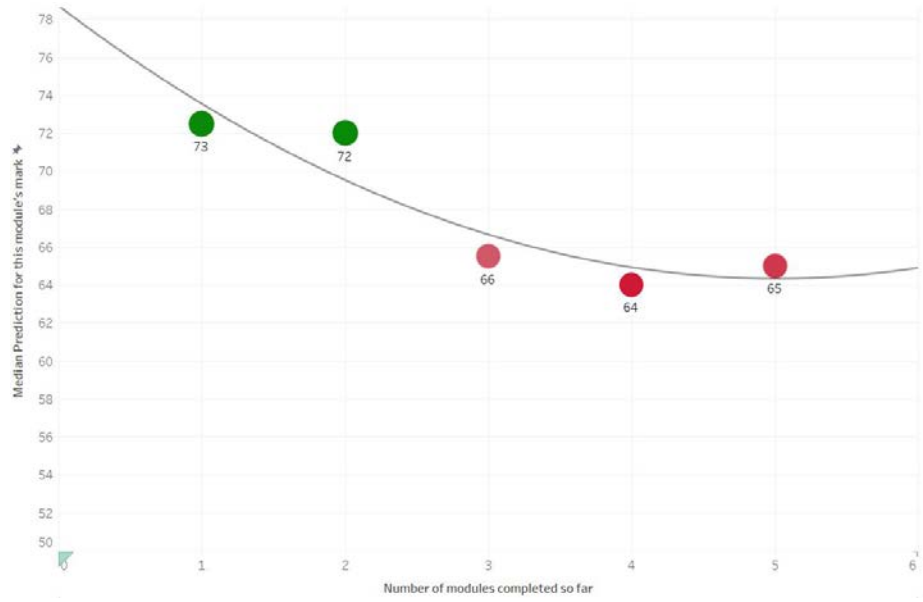
The distribution of prediction for this module's marks

82.64%
of the population predict to score between 60 to 75

The prediction marks showed higher expectation than the previous modules marks, 39.85% expect to get from 70 to 75 comparing to 8.91% who only achieved it in their past average marks.

Now let's see the average of marks that students expect to achieve in this course depending on how many modules they already completed.

Prediction of this module's marks Analysis



Completed one or two modules

72 to 73

The average expected marks of new students is greater than 70.

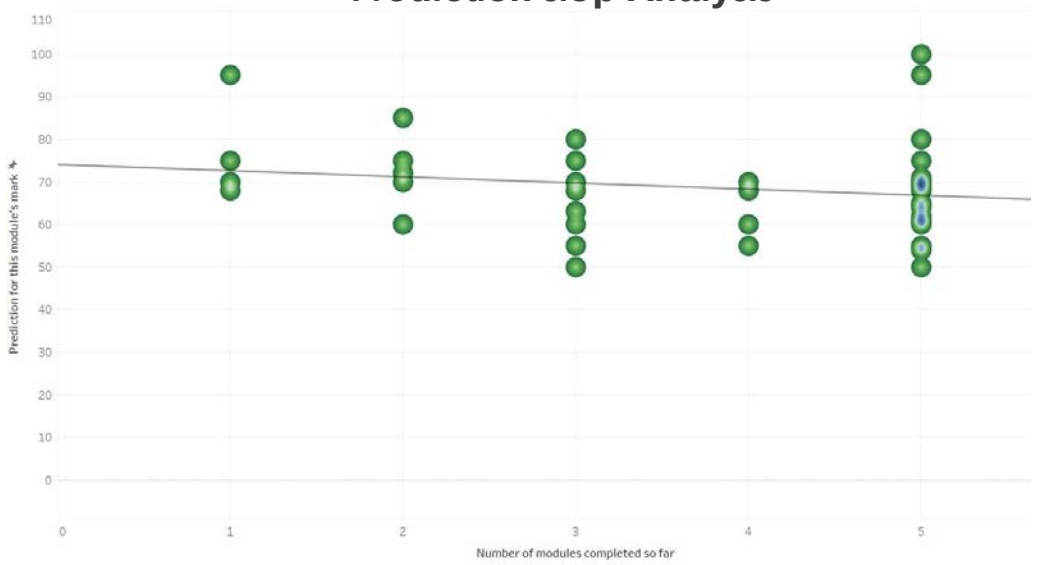
Completed three, four or five.

64 to 65

The average of expected marks of students completed more than two modules dropped down to 64.

This chart explains the drop of the average expected mark after finishing the second module and remains down.

Prediction slop Analysis



Drop of marks by -1.44284 when number of modules increase by one

The slope of the line indicates a negative correlation between the number of modules and the prediction of marks -1.44 of each module.

2.6 Analysis conclusion and outcome:

Project 2 Outcome

Business problem	Outcome
How much new students expect to get?	Above 70%
Is their any changes in the expectations?	Yes, students' expectations dropped down
After how many modules their expectations changed?	After the second module

The newer student of RKC appear to be not familiar with the UK grading system, it appears they expect higher marks, however, student who finished at least two modules are more familiar with the grading system, therefore, raising the awareness of the new students in the induction module is required.