UNIVERSITI TUNKU ABDUL RAHMAN

Assignment 2

Course Code:	UECS3213 / UECS3453
Course Name:	Data Mining
Lecturer:	Dr. Simon Lau Boung Yew
Academic Session:	2019/01
Title:	Exploratory Data Analysis and Visualization

Name	I.D. No	Course	Practical Group
Tan Ying Yao	1703648	SE	P1
			Mark: /10

Marking Scheme

No.		Poor	Adequate	Proficient	Subtotal
Section 1: Summary Statistics					
1	Correctness	not reasonable, not precise, not understandable	not reasonable, precise, understandable	reasonable, precise, understandable	
		0 - 5	6 - 15	16 - 20	
2	Comprehensiveness / Completeness of analysis	incomplete, lack of information	partially complete, with some information	complete, informative	
		0 - 5	6 - 15	16 - 20	
Section 2: Visualization					
3	Visual	not tidily presented, visually not readable	moderately tidy, readable	visually appealing, rich with information	
		0 - 10	11 - 20	21 - 30	
4	Critical analysis	illogical, incorrect assumption	partially logical, correct assumption	logical, critical, correct assumption	
		0 - 10	11 - 20	21 - 30	
				Total	/100
				Assessment	/10%

Summary Statistic:

I) Iris Dataset

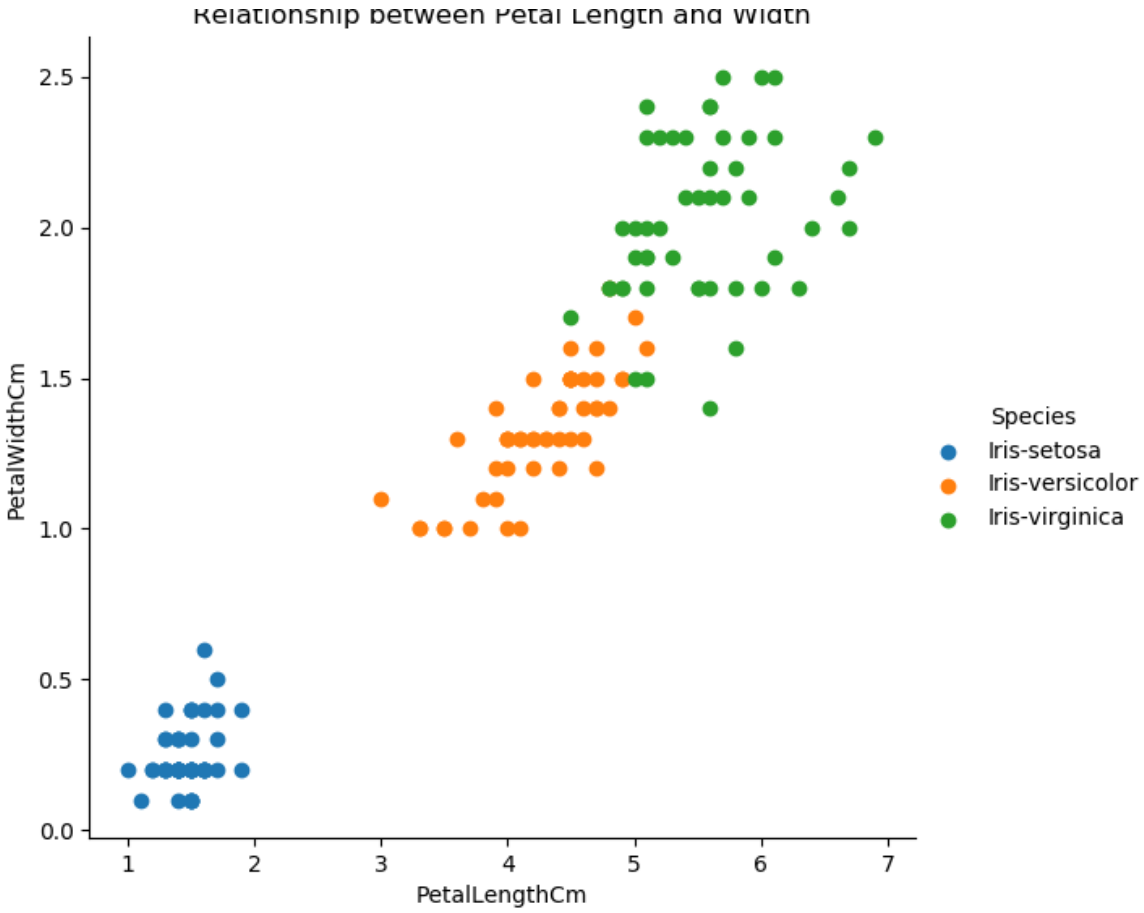
Calculation/Category	Sepal Length(cm)	Sepal Width (cm)	Petal Length (cm)	Petal Width (cm)
Mean	5.843	3.054	3.759	1.199
Standard deviation	0.828	0.434	1.764	0.7632
Minimum value	4.300	2.000	1.000	0.100
Maximum value	7.900	4.400	6.900	2.500
Median	5.80	3.00	4.35	1.30
Interquartile Range	1.10	0.50	3.50	1.50
Mean Absolute Deviation (MAD)	0.687	0.333	1.562	0.658
Mode	5.0	3.0	1.5	0.2

Brief Analysis:

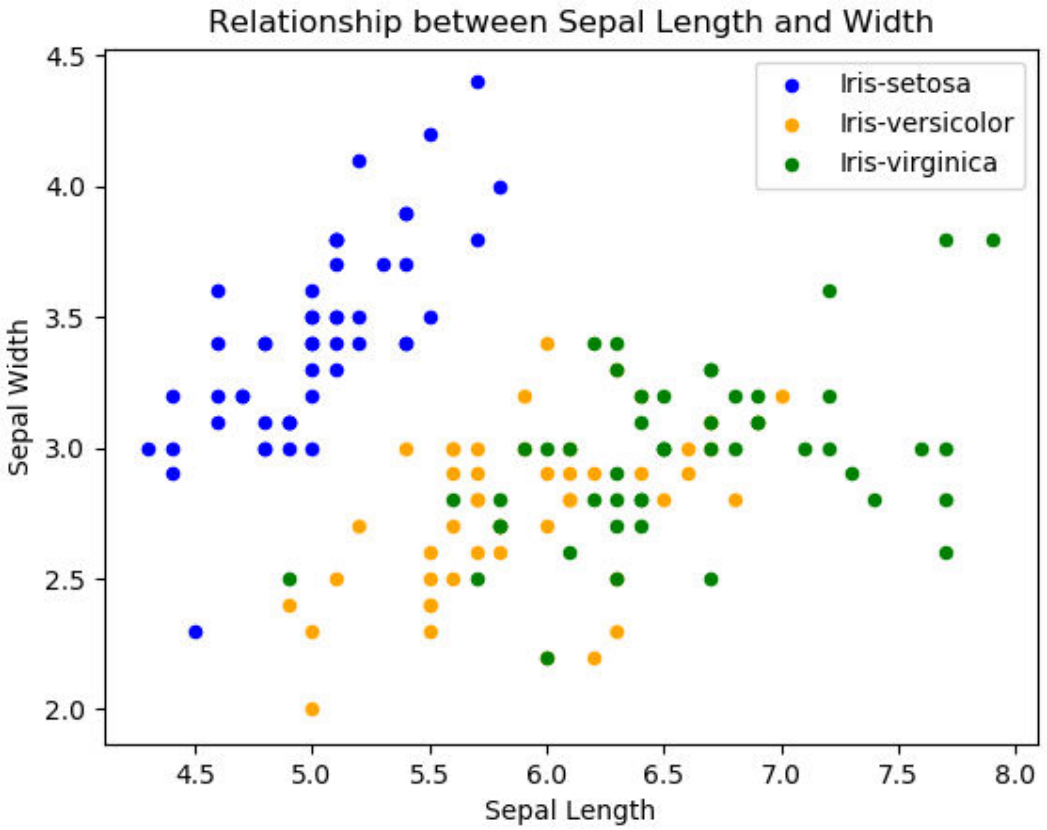
The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2. The 3 classes is as below:
-- Iris Setosa
-- Iris Versicolour
-- Iris Virginica

Each class of iris has their own sepal length, sepal width, petal length, petal width.

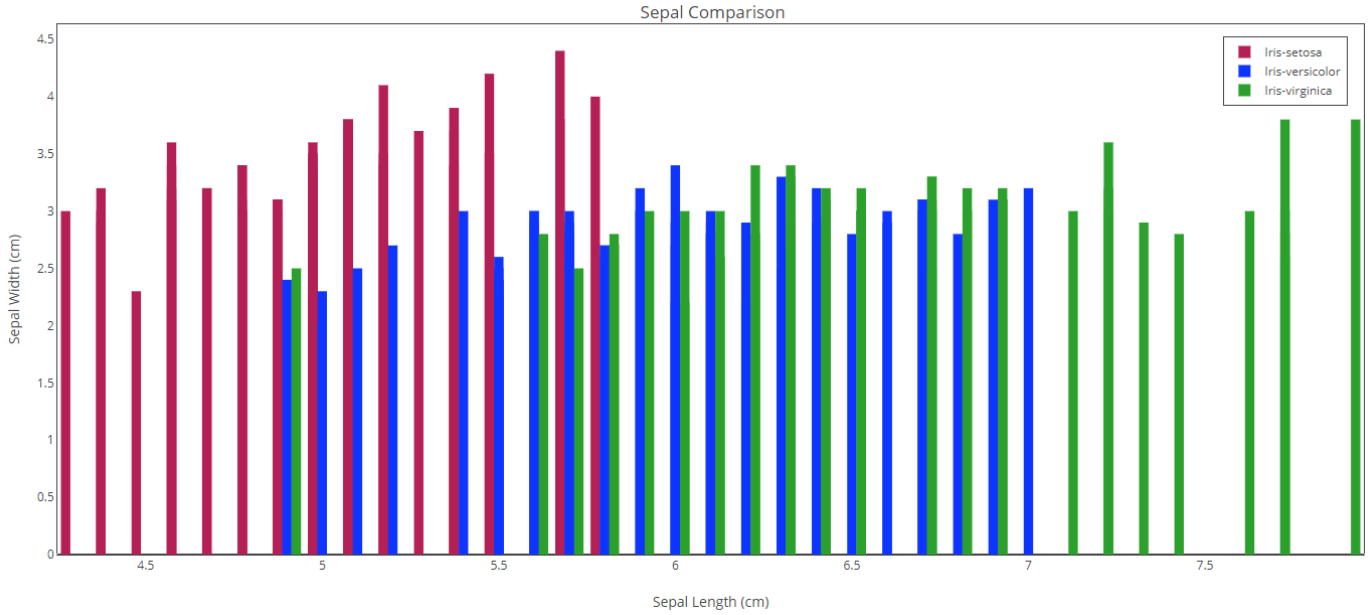
Graph 1: Relationship between Petal Length & Width



Graph 2: Relationship between Sepal Length & Sepal Width



Graph 3: Comparison between Sepal Width & Sepal Length



Critical Analysis:

Based on the first scatter plot, it can be observed that Iris-virginica has the longest petal width and petal length when compared to the other two classes. Iris-versicolor has the second longest petal width and petal length while Iris-setosa has the shortest width and length.

Based on the second scatter plot, Iris-virginica has the longest sepal length while Iris-setosa has the widest sepal width. Iris-versicolor is in the middle and has the shortest sepal length and a small sepal width.

Based on the bar chart, it can be seen that Iris-virginica has the longest sepal length while Iris-setosa has the widest sepal. Iris-versicolor is somewhere between the two with no standout characteristic except for a short sepal length and narrow sepal width.

II) Student Performance Dataset

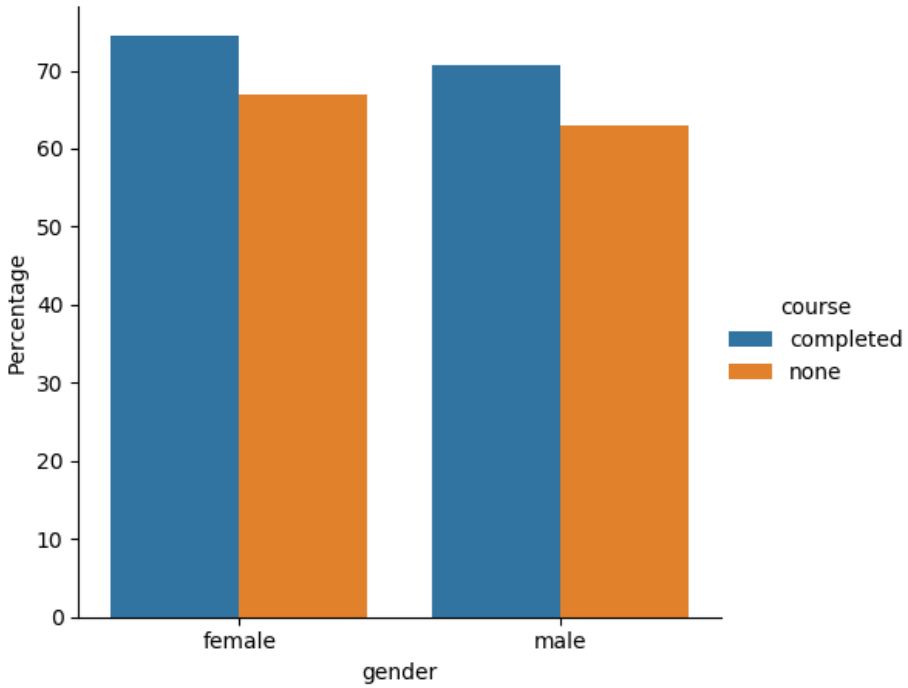
Calculation/Category	Mathematics Score	Reading Score	Writing Score
Mean	66	69	68
Standard deviation	15.16	14.60	15.19
Minimum value	0	17	10
Maximum value	100	100	100
Median	66.0	70.0	69.0
Interquartile Range	48	50	49
Mean Absolute Deviation (MAD)	12.02	11.77	12.20
Mode	65	62	74

Brief Analysis:

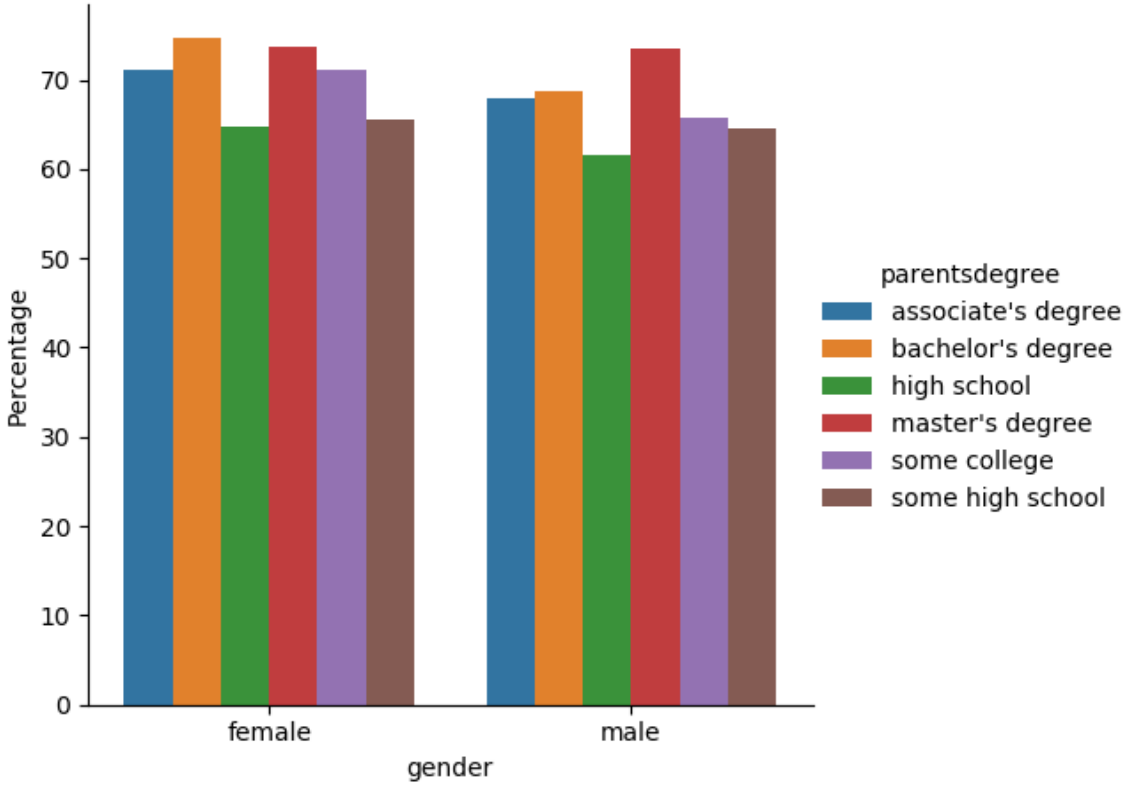
This data set consists of the marks obtained by the students in various subjects. It is further classified into several category:
Gender, race/ethnicity, parental level of education and participation in test preparation course.

All of these factors may affect the score of mathematics, reading and writing obtained by the students. A total of 100 student is recorded to analyse their correlation between score gained and factors involved.

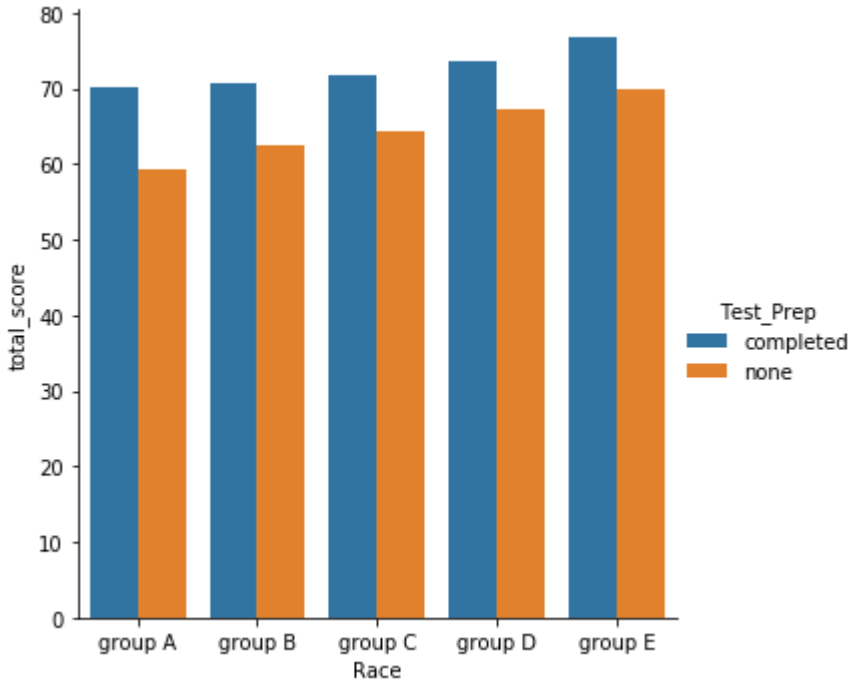
Graph 1: Relationship between percentage of gender and participation in test preparation course



Graph 2: Correlation between parental level of education & Student score



Graph 3: Score obtained by group of races according to the test preparation course



Critical Analysis:

Based on the first chart, the relationship between percentage of gender and participation in test preparation course shows a distinct difference in male and female participation. There are more females that had completed the course in comparison to it's male counterpart. Females also topped male's when comparing which gender did not participate the course. The scores are affected by whether the participants finished the course or not.

For the second chart, there is a strong correlation between parental level of education and the student's score. The female gender scores higher if the parents own at least a bachelor's degree while the male counterpart scores higher if their parents have at least a master's degree.

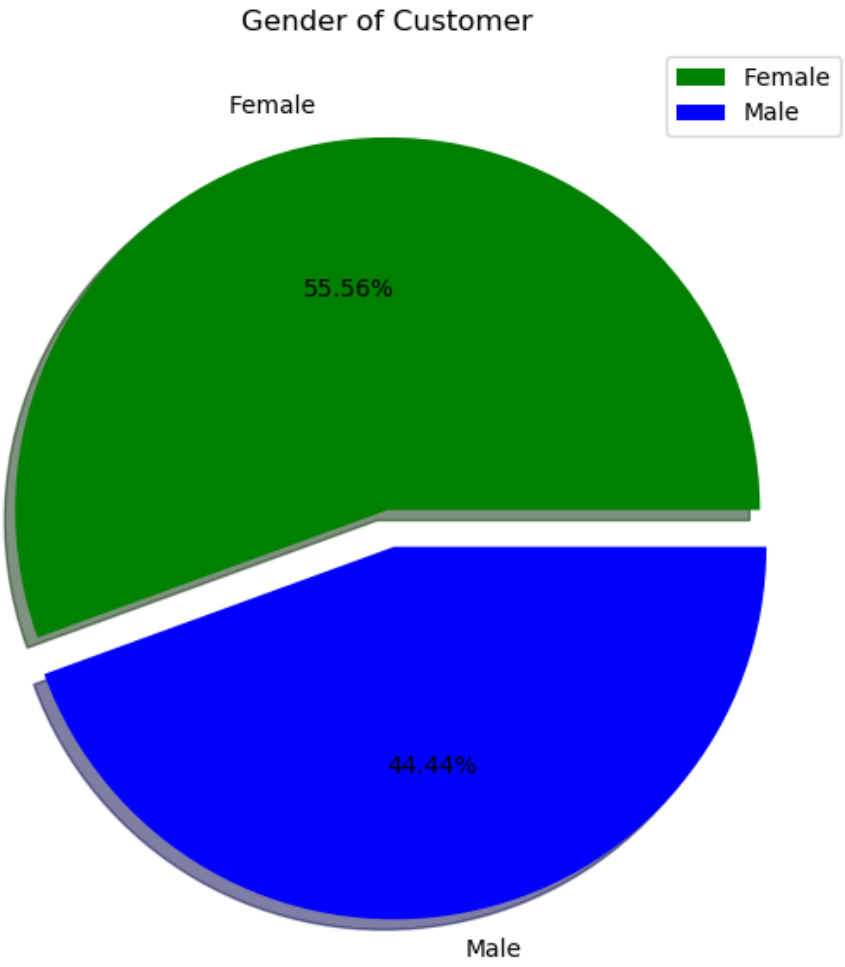
Regarding the third chart, race E seem to have the higher score in the test prepation score when compared to others. On the other hand, group A and group B both have the same total score. Those that finished the test preparation course score way better than those that didn't.

III) Shopping Mall Dataset

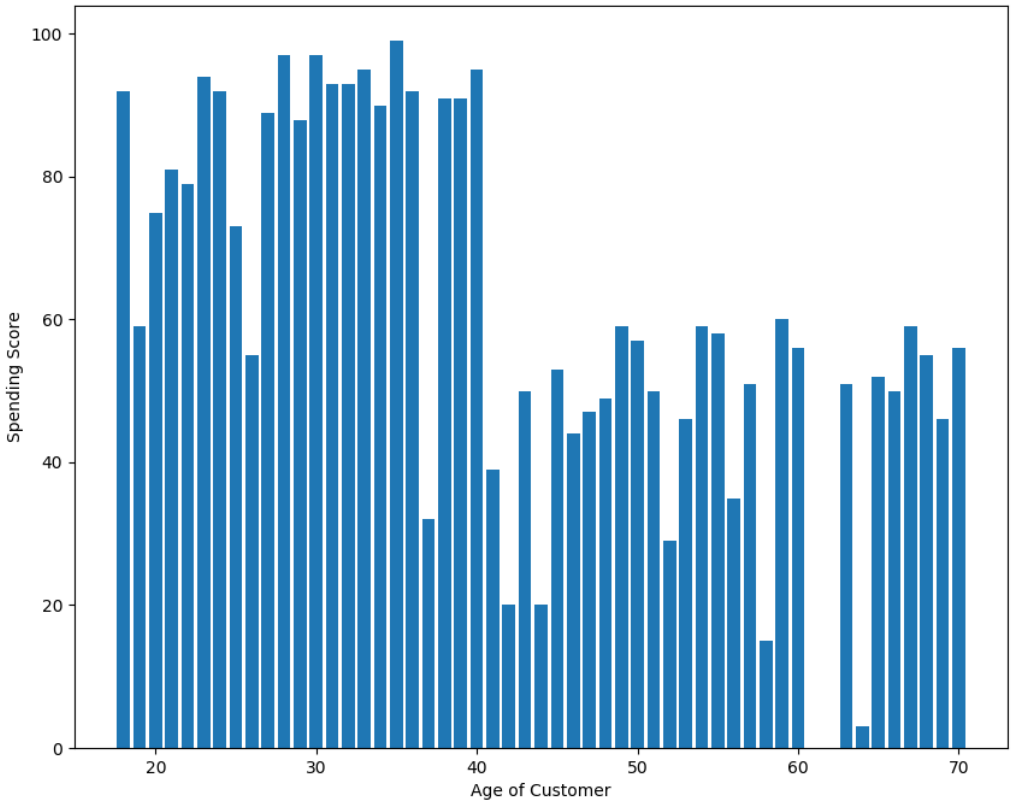
Calculation/ Category	Age	Annual Income ($)	Spending Score (1-100)
Mean	38.85	60.56	50.2
Standard Deviation	38.85	26.26	25.82
Minimum Value	18.00	15.00	1.00
Maximum Value	70.00	137.00	99.00
Mean Absolute Deviation (MAD)	11.66	21.00	20.82
Median	36.0	61.5	50.0
Mode	32.0	54.0	42.0

Brief Analysis:
The data set shows a supermarket mall's detail regarding it's membership. Customers share their data which includes Customer ID, age, gender, annual income and spending score. Spending Score is based on parameters such as customer behaviour and purchasing habits. The age and annual income of a customer may greatly affect their spending score.

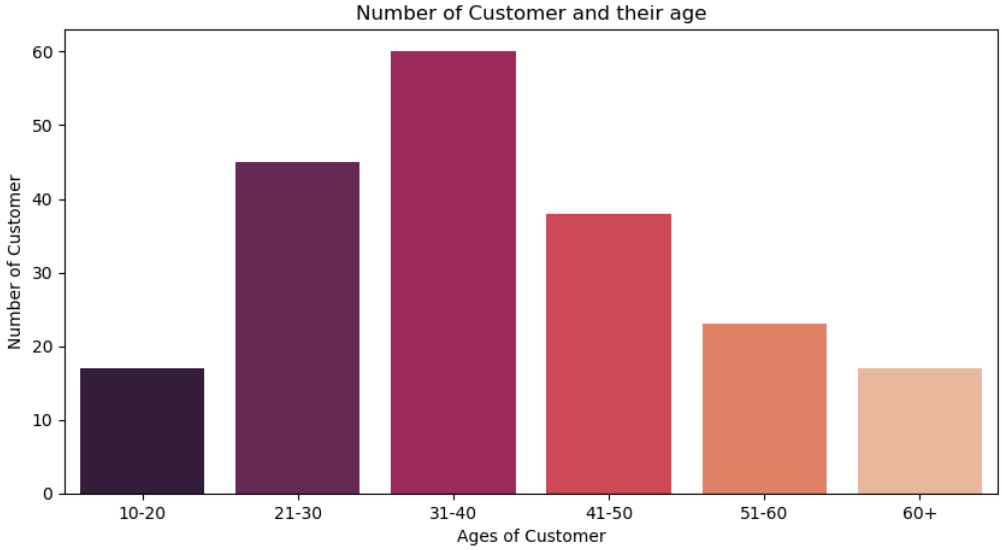
Graph 1: Percentage of gender of customer



Graph 2: Spending score and the age of customer



Graph 3: Number of customer and their age



Critical Analysis:

Based on the first pie chart, there are a total of 55.56% of female customer while there are only 44.44% of male customer. This can be attributed to the fact that it is the norm for the females to shop while males prefer not to.

Based on the second chart, the age group that spends the most seem to be in the range of age twenty to fourty. There is a sharp decline in those that shop after age fourty and above. Ages between 35 and 40 seem to have the highest spending score due to their higher financial capabilities as opposed to their younger demographic and the older demographic.

Based on the third chart, it can be inferred that the ages of customer heavily affect whether they shop or not. It can be seent hat customers of age 31 to 40 has the highest amount of shoppers. The least amount of shoppers belong the to age group 10 – 20. This can be explained as the young don't possess the ability to spend money recklessly while the older group may have money saved up and capability to do so.