MINI PROJECT ON

Project Entitled:

‘CUSTOMER SEGMENTATION ‘

Submitted by:

Kaustubh Patil (A-36)

Yash Surana (A-45)

CUSTOMER SEGMENTATION USING R

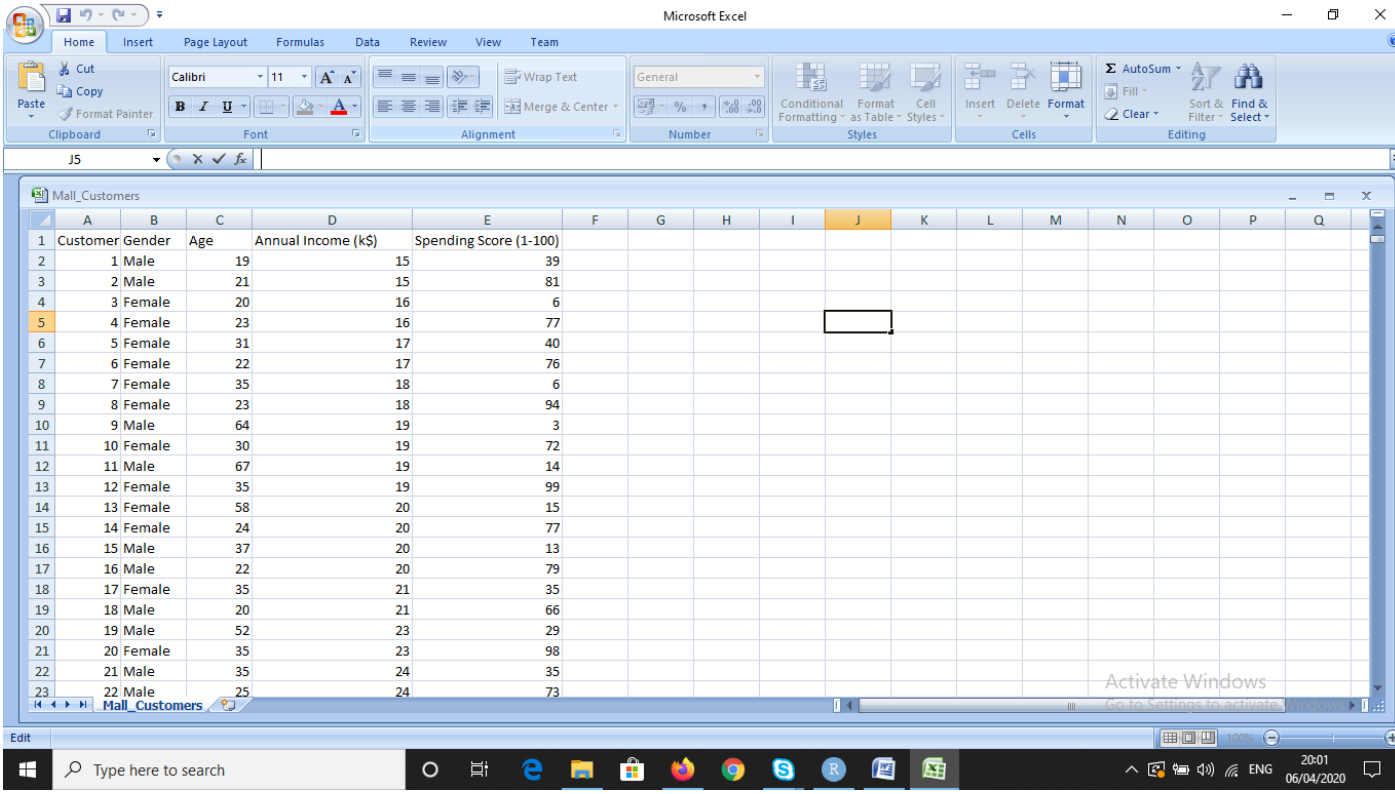
CUSTOMER SEGEMENTATION -

Customer Segmentation is one the most important applications of unsupervised learning. Using clustering techniques, companies can identify the several segments of customers allowing them to target the potential user base. In this machine learning project, we will make use of K-means Clustering which is the essential algorithm for clustering unlabeled dataset.

Customer Segmentation is the process of division of customer base into several groups of individuals that share a similarity in different ways that are relevant to marketing such as gender, age, interests, and miscellaneous spending habits.

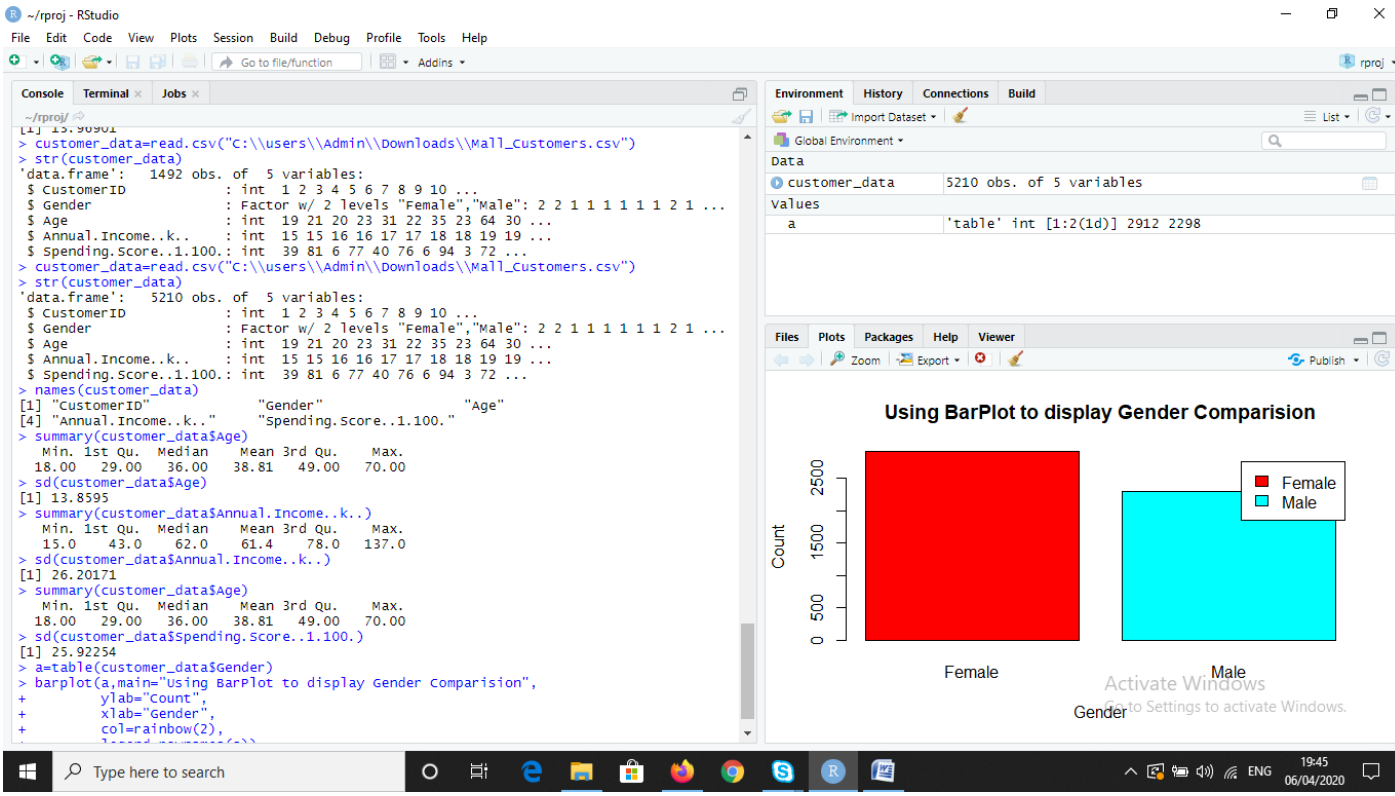
Companies that deploy customer segmentation are under the notion that every customer has different requirements and require a specific marketing effort to address them appropriately. Companies aim to gain a deeper approach of the customer they are targeting. Therefore, their aim has to be specific and should be tailored to address the requirements of each and every individual customer. Furthermore, through the data collected, companies can gain a deeper understanding of customer preferences as well as the requirements for discovering valuable segments that would reap them maximum profit. This way, they can strategize their marketing techniques more efficiently and minimize the possibility of risk to their investment.

DATASET SCREENSHOT-



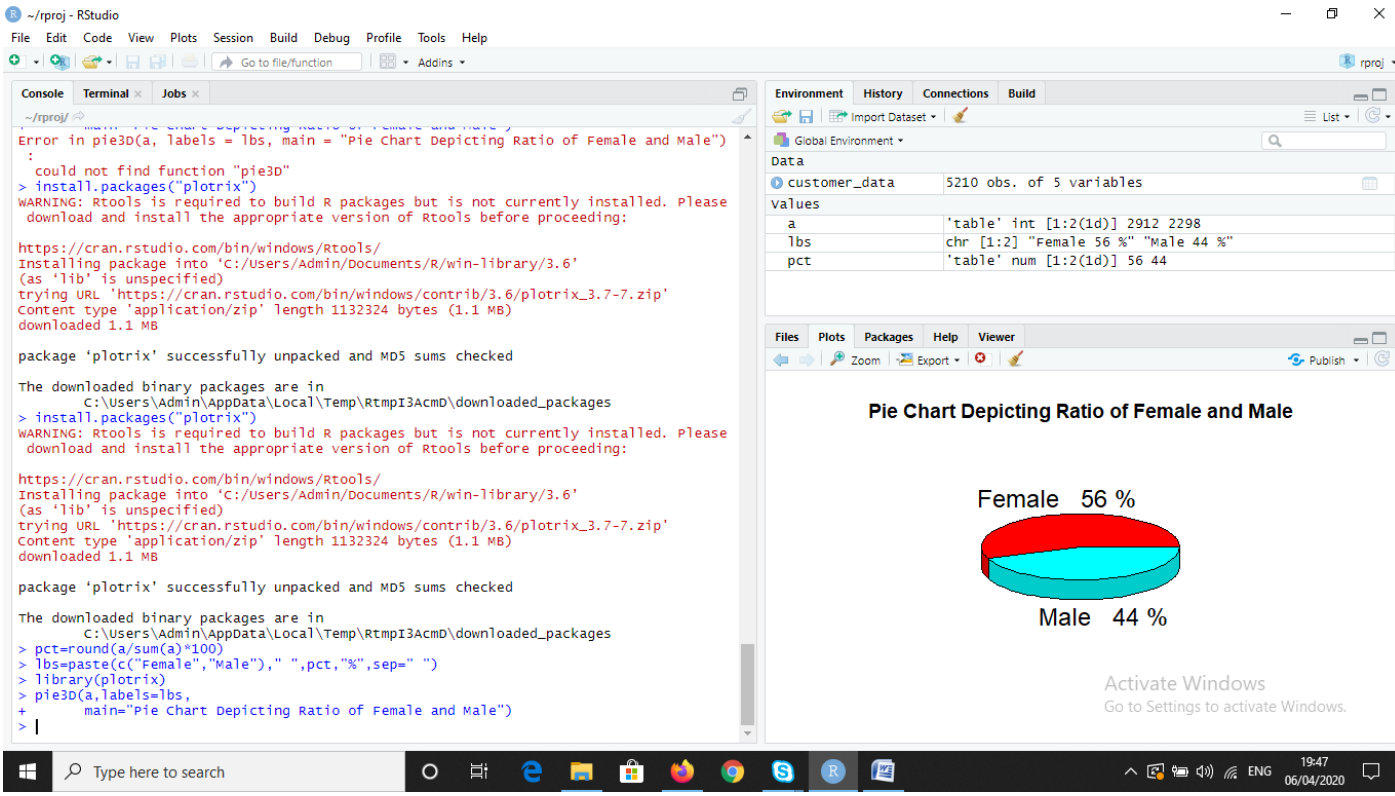
	A	B	C	D	E
1	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
2	1	Male	19	15	39
3	2	Male	21	15	81
4	3	Female	20	16	6
5	4	Female	23	16	77
6	5	Female	31	17	40
7	6	Female	22	17	76
8	7	Female	35	18	6
9	8	Female	23	18	94
10	9	Male	64	19	3
11	10	Female	30	19	72
12	11	Male	67	19	14
13	12	Female	35	19	99
14	13	Female	58	20	15
15	14	Female	24	20	77
16	15	Male	37	20	13
17	16	Male	22	20	79
18	17	Female	35	21	35
19	18	Male	20	21	66
20	19	Male	52	23	29
21	20	Female	35	23	98
22	21	Male	35	24	35
23	22	Male	25	24	73

OUTPUT-

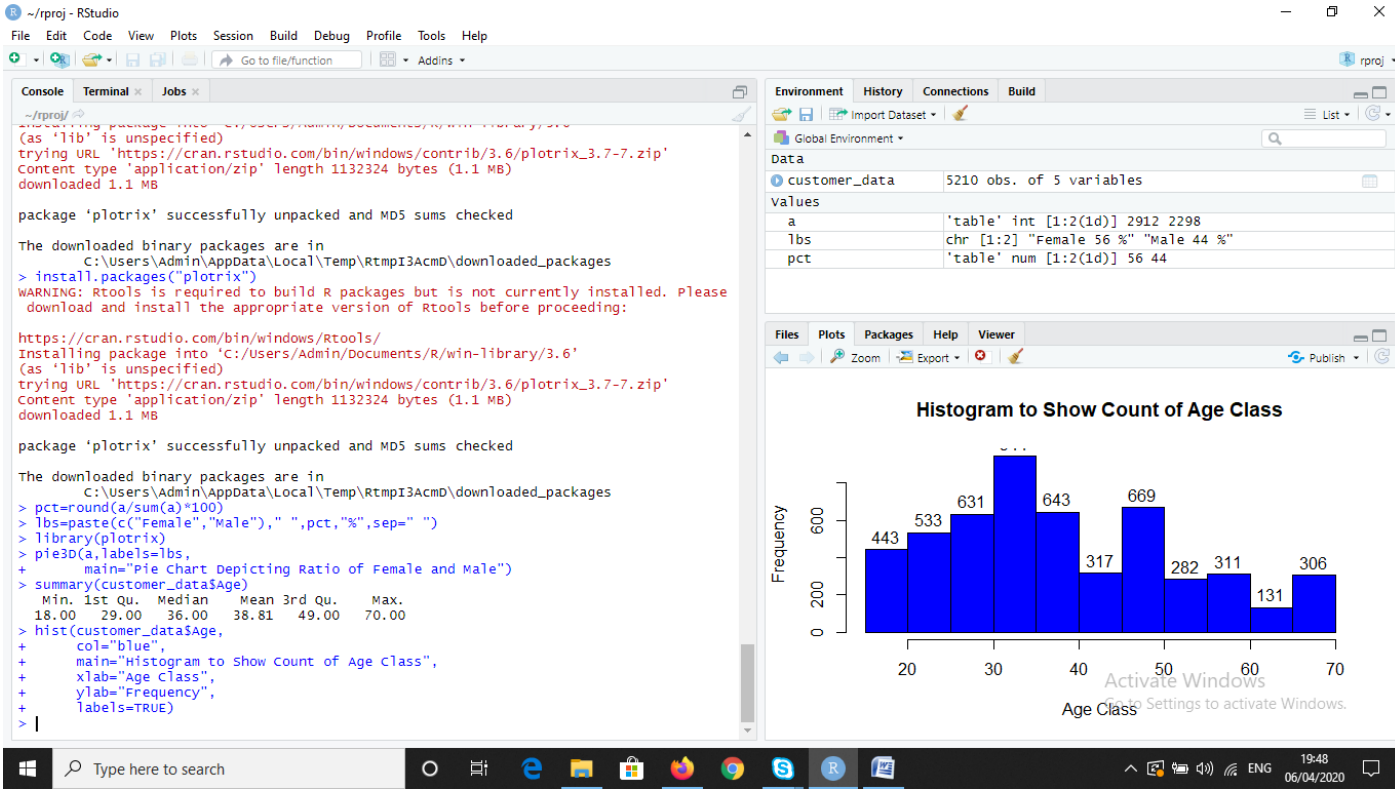


```
> customer_data=read.csv("C:\\users\\Admin\\Downloads\\Mall_Customers.csv")
> str(customer_data)
'data.frame':   1492 obs. of  5 variables:
 $ CustomerID            : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Gender                : Factor w/ 2 levels "Female","Male": 2 2 1 1 1 1 1 1 2 1 ...
 $ Age                   : int  19 21 20 23 31 22 35 23 64 30 ...
 $ Annual.Income..k..    : int  15 15 16 16 17 17 18 18 19 19 ...
 $ Spending.Score..1.100.: int  39 81 6 77 40 76 6 94 3 72 ...
> customer_data=read.csv("C:\\users\\Admin\\Downloads\\Mall_Customers.csv")
> str(customer_data)
'data.frame':   5210 obs. of  5 variables:
 $ CustomerID            : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Gender                : Factor w/ 2 levels "Female","Male": 2 2 1 1 1 1 1 1 2 1 ...
 $ Age                   : int  19 21 20 23 31 22 35 23 64 30 ...
 $ Annual.Income..k..    : int  15 15 16 16 17 17 18 18 19 19 ...
 $ Spending.Score..1.100.: int  39 81 6 77 40 76 6 94 3 72 ...
> names(customer_data)
[1] "CustomerID"             "Gender"                 "Age"
[4] "Annual.Income..k.."     "Spending.Score..1.100."
> summary(customer_data$Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   29.00   36.00   38.81   49.00   70.00
> sd(customer_data$Age)
[1] 13.8595
> summary(customer_data$Annual.Income..k..)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   15.0    43.0    62.0    61.4    78.0   137.0
> sd(customer_data$Annual.Income..k..)
[1] 26.20171
> summary(customer_data$Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   29.00   36.00   38.81   49.00   70.00
> sd(customer_data$Spending.Score..1.100.)
[1] 25.92254
> a=table(customer_data$Gender)
> barplot(a,main="Using BarPlot to display Gender Comparision",
+         ylab="Count",
+         xlab="Gender",
+         col=rainbow(2),
```

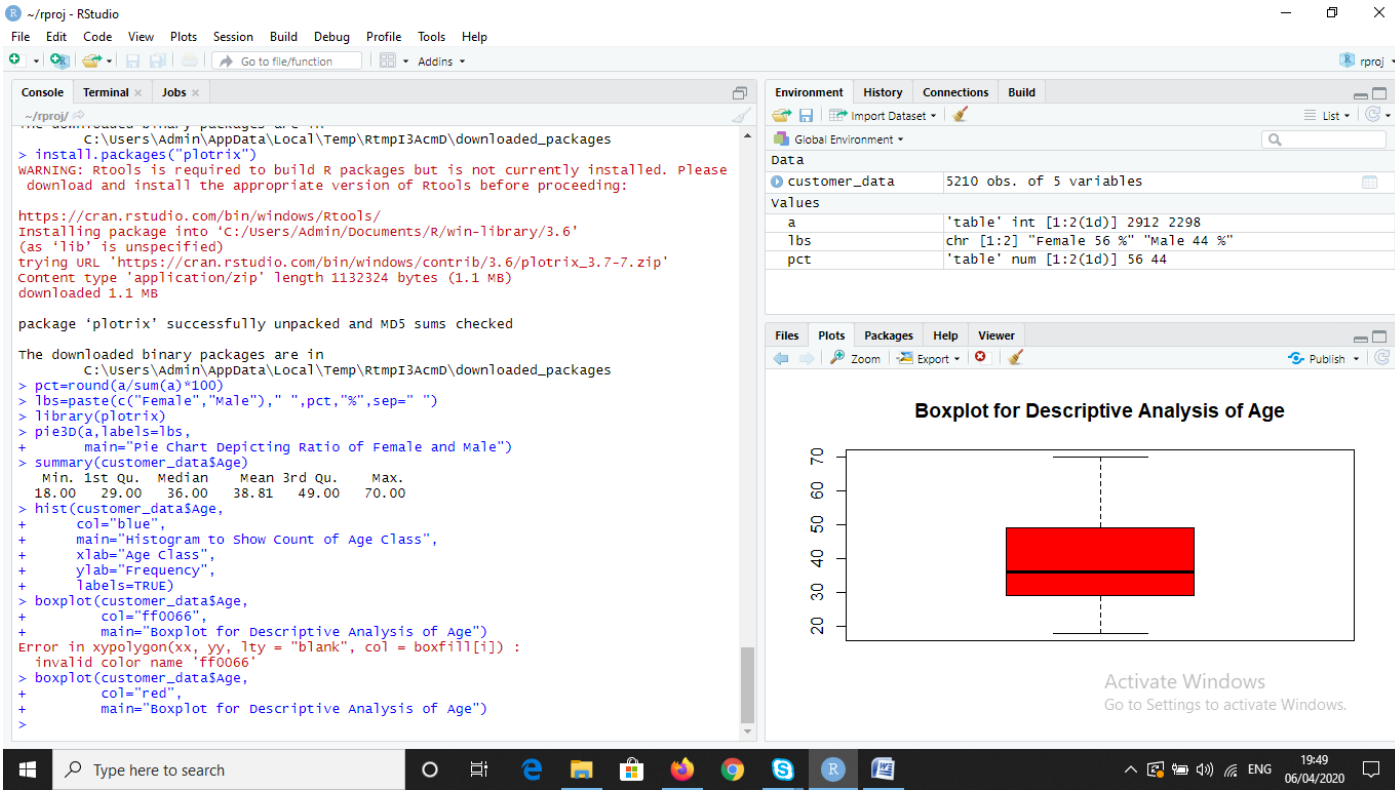
BAR PLOT



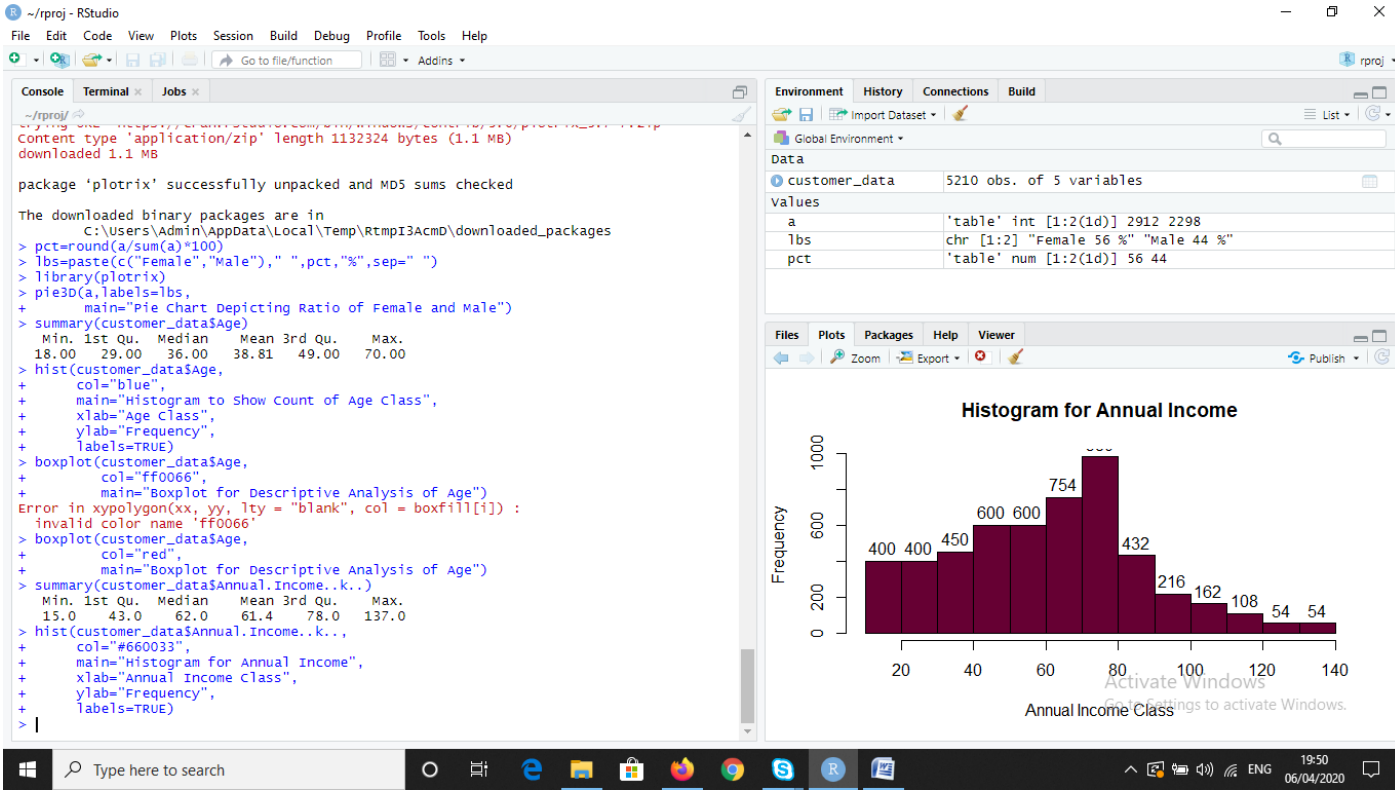
PIE CHART



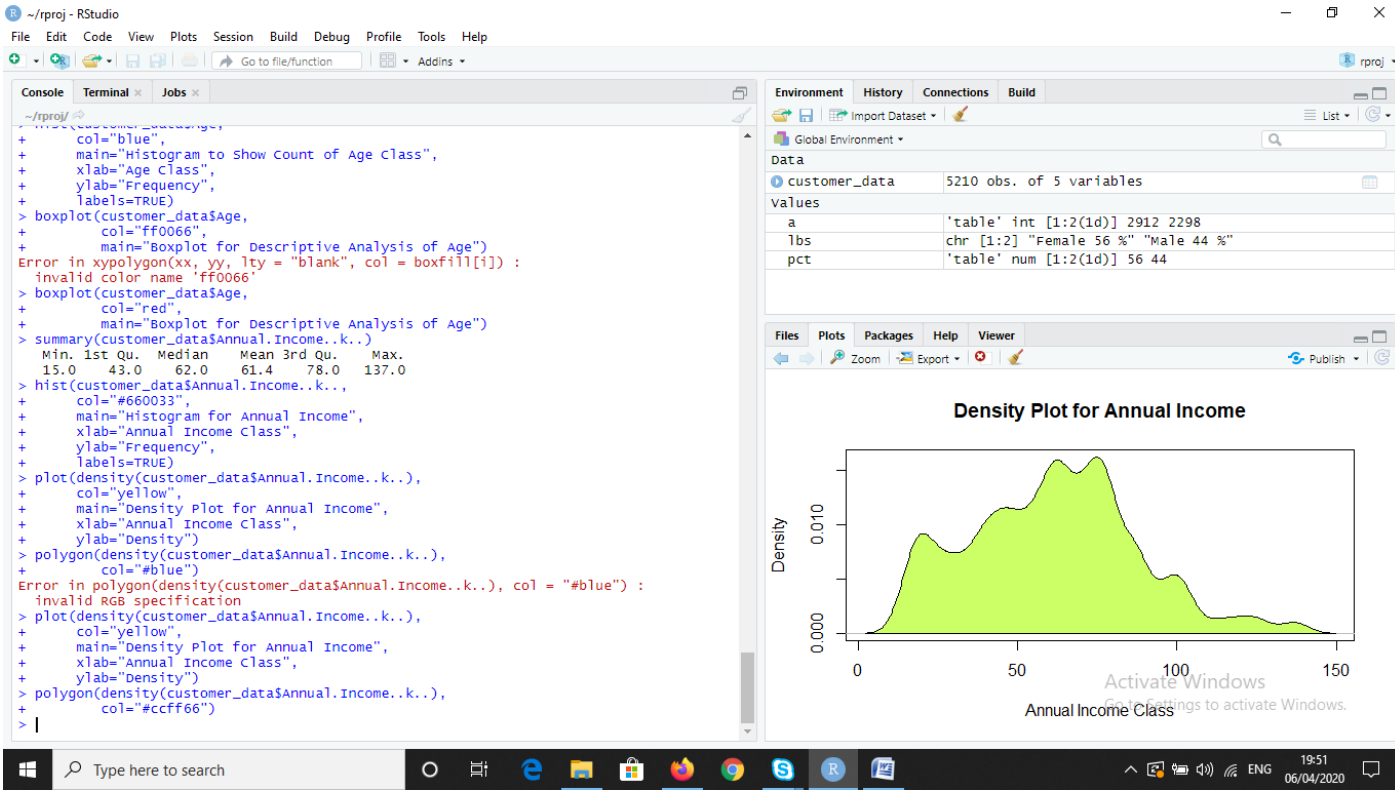
HISTROGRAM



BOX PLOT



HISTROGRAM



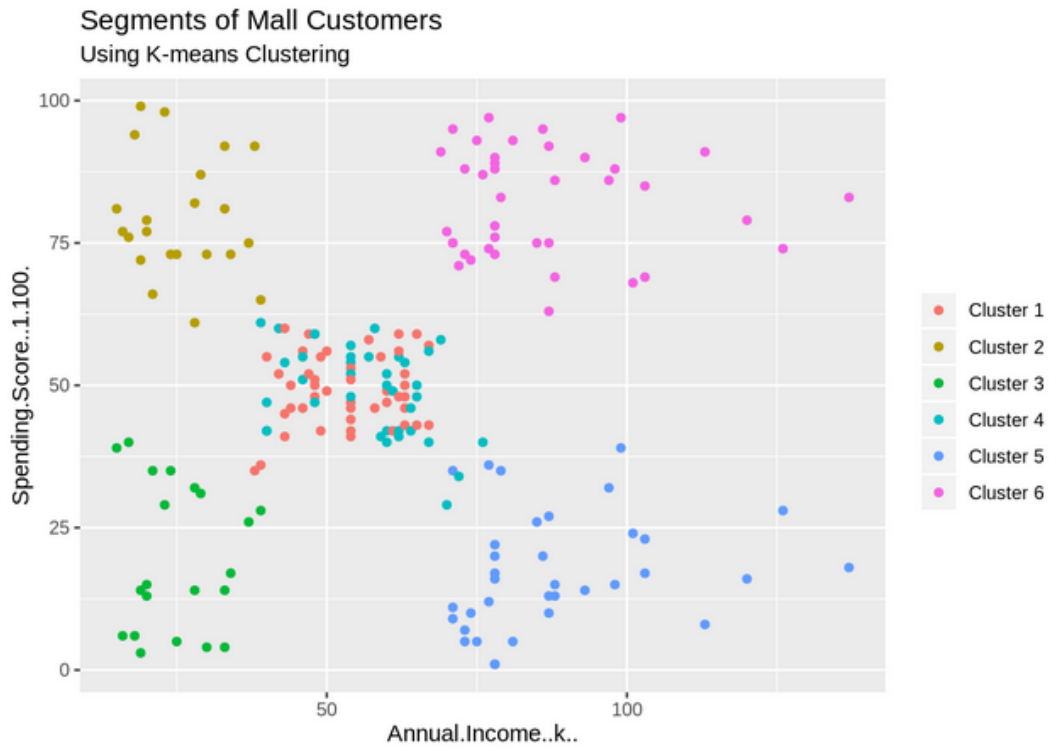
DENSITY PLOT

K-means Algorithm-

While using the k-means clustering algorithm, the first step is to indicate the number of clusters (k) that we wish to produce in the final output. The algorithm starts by selecting k objects from dataset randomly that will serve as the initial centers for our clusters. These selected objects are the cluster means, also known as centroids. Then, the remaining objects have an assignment of the closest centroid. This centroid is defined by the Euclidean Distance present between the object and the cluster mean. We refer to this step as “cluster assignment”. When the assignment is complete, the algorithm proceeds to calculate new mean value of each cluster present in the data. After the recalculation of the centers, the observations are checked if they are closer to a different cluster. Using the updated cluster mean, the objects undergo reassignment. This goes on repeatedly through several iterations until the cluster assignments stop altering. The clusters that are present in the current iteration are the same as the ones obtained in the previous iteration.

CODE-

```
set.seed(1)

ggplot(customer_data, aes(x =Annual.Income..k.., y = Spending.Score..1.100.)) +

geom_point(stat = "identity", aes(color = as.factor(k6$cluster))) +

scale_color_discrete(name=" ",

breaks=c("1", "2", "3", "4", "5","6"),

labels=c("Cluster 1", "Cluster 2", "Cluster 3", "Cluster 4", "Cluster 5","Cluster 6")) +

ggtitle("Segments of Mall Customers", subtitle = "Using K-means Clustering")
```



Cluster 1 – This cluster represents the customer_data having a high annual income as well as a high annual spend.

Cluster 2 – This cluster denotes the customer_data with low annual income as well as low yearly spend of income.

Cluster 3 – These clusters represent the customer_data with the medium income salary as well as the medium annual spend of salary.

Cluster 4 – This cluster denotes a high annual income and low yearly spend.

Cluster 5 – This cluster represents a low annual income but its high yearly expenditure

Cluster 6 – These clusters represent the customer_data with the medium income salary as well as the medium annual spend of salary.

Summary-

In this data science project, we went through the customer segmentation model. We developed this using a class of machine learning known as unsupervised learning. Specifically, we made use of a clustering algorithm called K-means clustering. We analyzed and visualized the data and then proceeded to implement our algorithm.