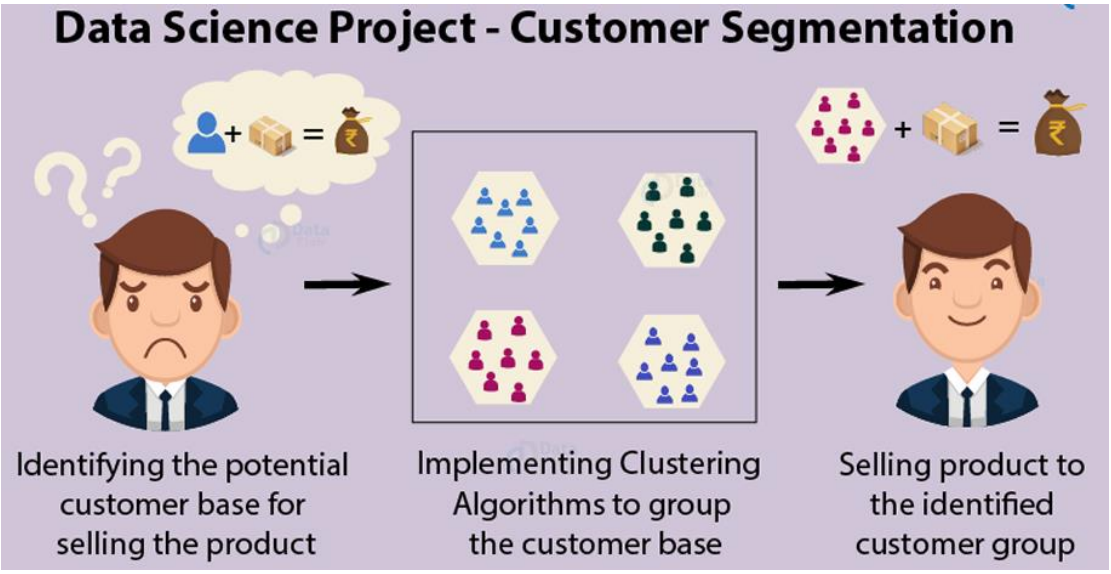
章一

什么是客户细分：

客户细分是将客户群划分为几组个人的过程，这些个人组以不同的方式与营销相关，例如性别、年龄、兴趣和其他消费习惯。

部署客户细分的公司认为，每个客户都有不同的要求，需要特定的营销工作来适当地满足这些要求。公司的目标是更深入地了解他们的目标客户。因此，他们的目标必须是具体的，并且应该根据每个客户的要求进行定制。此外，通过收集的数据，公司可以更深入地了解客户偏好以及发现有价值的细分市场的要求，从而获得最大的利润。这样，他们可以更有效地制定营销技术战略，并将投资风险的可能性降至最低。

客户细分技术取决于几个关键差异化因素，这些差异化因素将客户划分为要定位的组。与人口统计、地理、经济状况以及行为模式相关的数据在确定公司解决各个细分市场的方向方面起着至关重要的作用。



实施：

在这个数据科学项目的第一步，我们将执行数据探索。我们将导入此角色所需的基本包，然后读取我们的数据。最后，我们将遍历输入数据以获得有关它的必要见解。

数据读取：

CSV：- 在进行客户细分分析之前，第一步是读取数据以执行分析。数据保存在名为 Mall_Customers.csv 的数据集中。此数据集包含 400 条各种类型客户的记录。保存在 dataset 中的事件是非结构化的。要执行分析，使用命令“read.csv”读取数据集。

	A	B	C	D	E	F	G	H	I
1	CustomerI	Gender	Age	Annual Inc	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				

图 1 Mall_Customer.csv

数据初步认识：

通过 str()、names()、head()对数据进行初步探索，通过 summary()、sd()对数值型数据进行初步处理。

```
> str(customer_data)
'data.frame':   400 obs. of  5 variables:
 $ CustomerID            : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Gender                : chr  "Male" "Male" "Female" "Female" ...
 $ 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."
> head(customer_data)
  CustomerID Gender Age Annual.Income..k.. Spending.Score..1.100.
1          1   Male  19                 15                     39
2          2   Male  21                 15                     81
3          3 Female  20                 16                      6
4          4 Female  23                 16                     77
5          5 Female  31                 17                     40
6          6 Female  22                 17                     76
> #查看数据结构
> summary(customer_data$Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   28.75   36.00   38.85   49.00   70.00
> sd(customer_data$Age)
[1] 13.95149
> summary(customer_data$Annual.Income..k..)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  15.00   41.50   61.50   60.56   78.00  137.00
> sd(customer_data$Annual.Income..k..)
[1] 26.23179
> summary(customer_data$Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   28.75   36.00   38.85   49.00   70.00
> sd(customer_data$Spending.Score..1.100.)
[1] 25.79114
```

图 2 数据初步处理

章二、数据分析

客户性别可视化：

对非数值型数据进行可视化处理

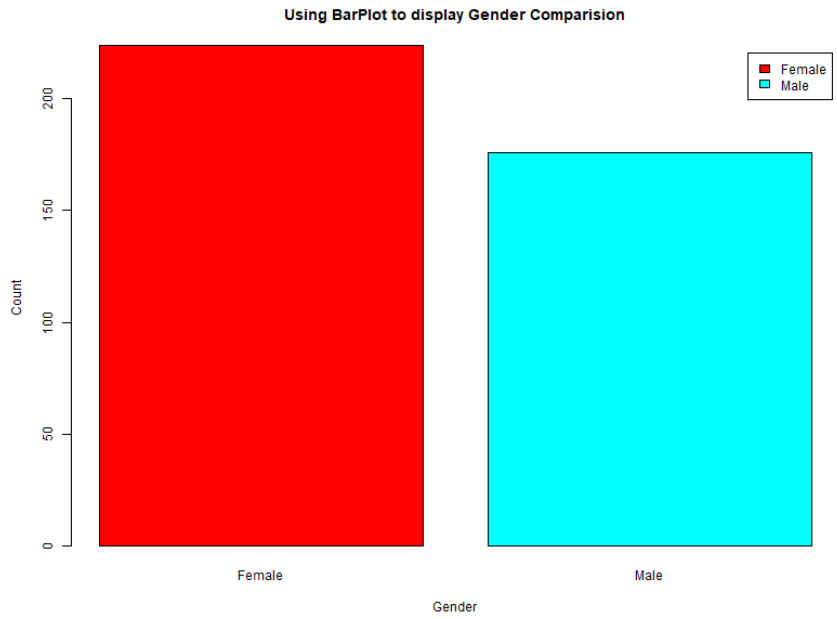


图 3、性别条形图

通过计算，并使用饼图体现具体性别占比，女性的百分比为 56%，而男性的百分比为 44%。

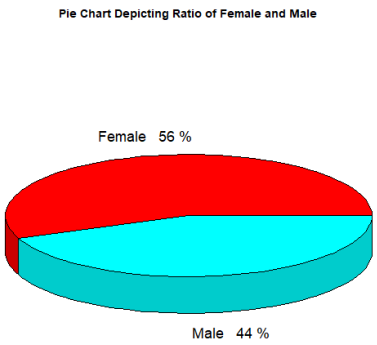
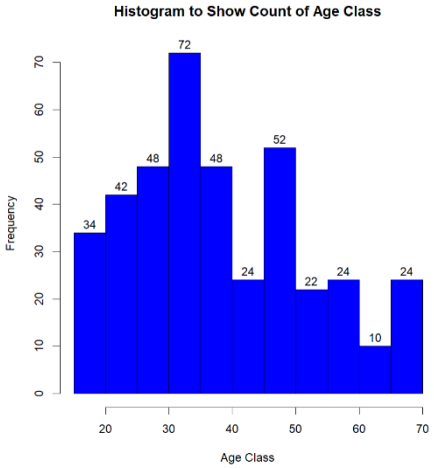
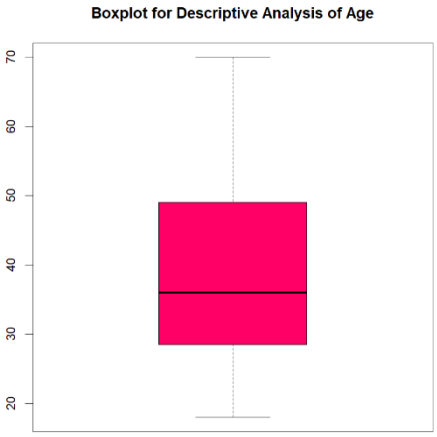


图 4、性别饼图

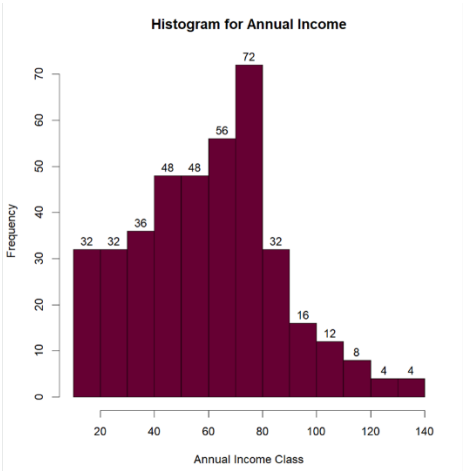
年龄可视化：

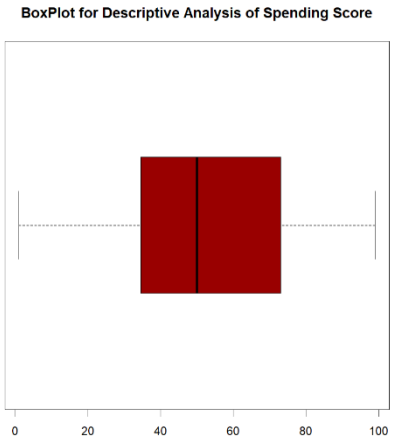
通过可视化使得数据更好说明且易懂



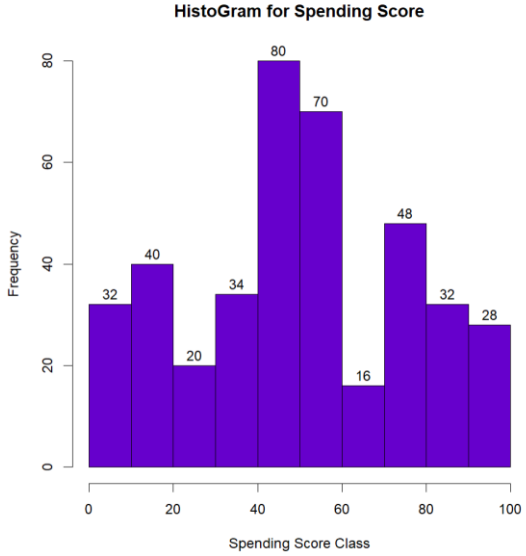


收入可视化：





购物能力得分可视化：



章三

k 均值算法

在使用 k-means 聚类算法时，第一步是指出我们希望在最终输出中产生的聚类数量 (k)。算法首先从数据集中随机选择 k 个对象，这些对象将作为我们聚类的初始中心。 这些被选择的对象就是聚类均值，也被称为质心。然后，剩下的对象有一个最近的 质心的分配。该质心由对象和簇均值之间的欧几里得距离定义。我们将这一步称为 “聚类分配”。当分配完成后，算法继续计算数据中存在的每个聚类的新平均值。 重新计算中心后，检查观察值是否更接近不同的聚类。使用更新后的聚类均值，对 对象进行重新分配。这个过程反复进行几次迭代，直到集群分配停止变化。当前迭 代中出现的聚类与前一次迭代中获得的聚类相同。

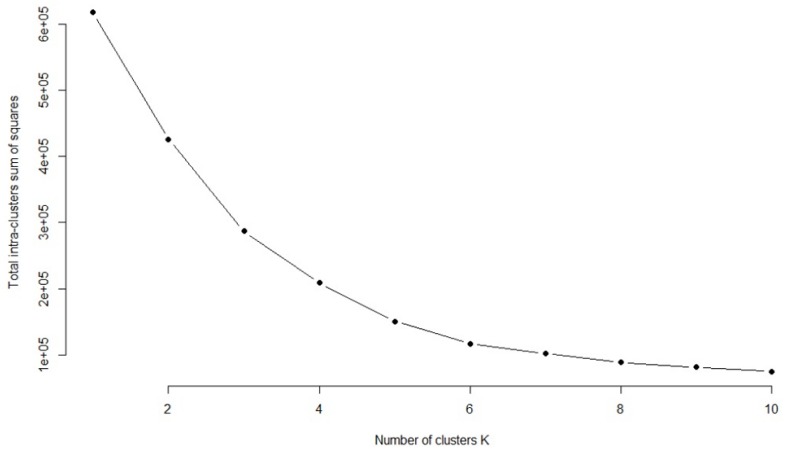
对 k 均值聚类求和

·我们指定需要创建的聚类数量，该算法从数据集中随机选择 k 个对象。该对象为初始聚类或均值。最接近的质心获得新观测值的分配。我们基于物体和质心之间的欧几里得距离来分配这个任务。

数据点中的 k 个聚类通过计算聚类中所有数据点中存在的新平均值来更新质心。第 k 个聚类的质心长度为 p，包含第 k 个聚类中观测值的所有变量的均值。我们用 p 来表示变量的个数。 ·

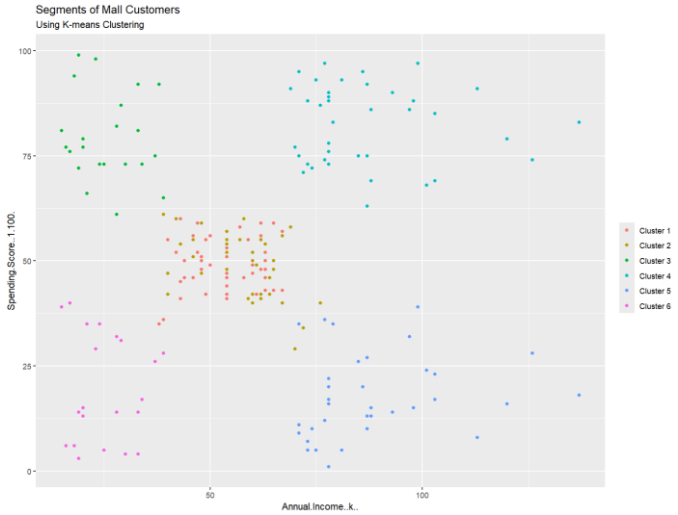
平方和内总数的迭代最小化。然后通过对总平方和的迭代最小化，当我们达到最大迭代时，赋值停止摇摆。R 软件在最大迭代时使用的默认值是 10。

我们计算几个 k 值的聚类算法，这可以通过在 k 内创建从 1 到 10 个簇的变化来完成。然后我 们计算总簇内平方和(iss)。然后，我们根据 k 个聚类的数量继续绘制 iss。该图表示我们模型中所需的适当数量的聚类。



章四

对聚类结果进行可视化：



从上面的可视化中，我们观察到有 6 个簇的分布如下

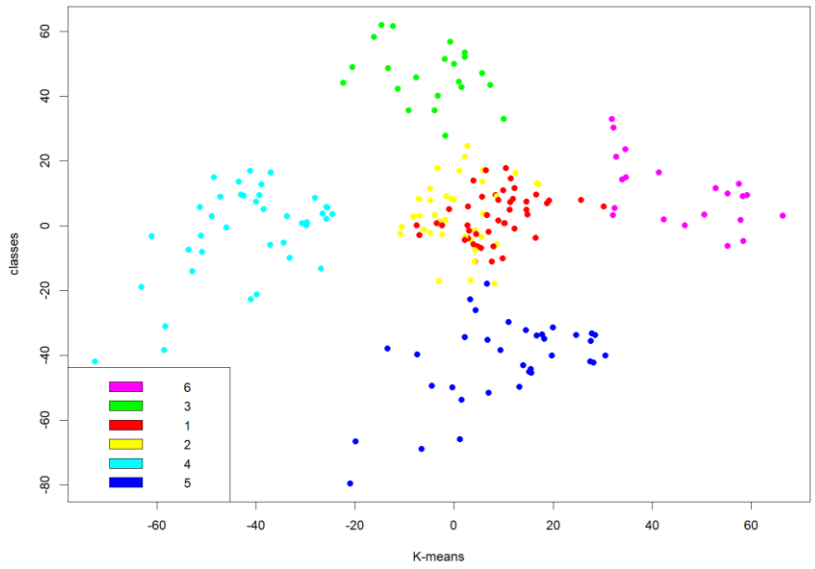
集群 6 和 4——这些集群表示具有中等收入工资和中等年度工资支出 customer_data。

集群 1——该集群表示具有高年收入和高年支出的 customer_data。

集群 3 -该集群表示年收入较低的 customer_data 以及年收入较低的年支出。

集群 2 -该集群表示高年收入和低年支出。

集群 5 -该集群年收入低，但年支出高。



集群 4 和 1——这两个集群由 PCA1 和 PCA2 评分中等的客户组成。

集群 6 -此集群表示具有高 PCA2 和低 PCA1 的客户。

集群 5 在这个集群中，有 PCA1 得分中等而 PCA2 得分较低的客户。

集群 3 -该集群由具有高 PCA1 收入和高 PCA2 的客户组成。

集群 2 -这包括具有高 PCA2 和中等年支出收入的客户。

在聚类的帮助下，我们可以更好地理解变量，促使我们做出谨慎的决策。有了对客 户的识别，公司就可以根据收入、年龄、消费模式等几个参数，发布针对客户的产 品和服务。此外，更复杂的模式，如产品评论，也会被考虑在内，以更好地进行细分。