Data Science Project – Customer Segmentation using Machine Learning in R

- **Overview:** Customer Segmentation is a key application of unsupervised learning that allows companies to identify different customer segments for targeted marketing.
- **Project Goal:** To implement Customer Segmentation using K-means clustering in R.
- **Dataset:** Mall_Customers.csv: Contains demographic data of mall visitors.

Photo by Digitawise Agency on Unsplash

Data Exploration

- **Importing Packages:** Use the necessary R packages for data analysis and visualization.
- **Reading Data:** Load the 'Mall_Customers.csv' dataset and inspect its structure using str() and head() functions.
- **Data Summary:** Generate summary statistics and visualize data distribution using summary() and other descriptive statistics functions.

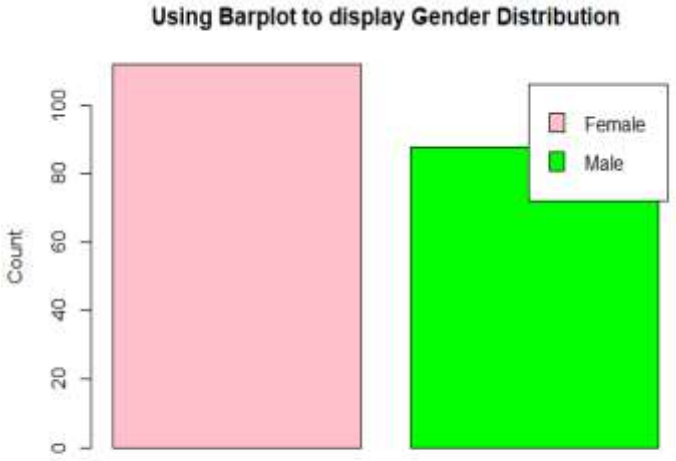
Photo by Ilija Boshkov on Unsplash

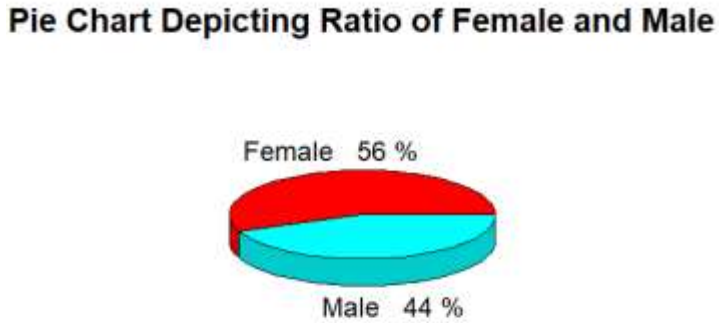
Customer Gender Visualization

- **Bar Plot:** Code: a=table(customer_data$Gender) barplot(a, main='Using BarPlot to display Gender Comparison', ylab='Count', xlab='Gender', col=rainbow(2), legend=rownames(a))

- **Pie Chart:** Code: pct=round(a/sum(a)*100) lbs=paste(c('Female','Male'), ' ', pct, '%', sep=' ') library(plotrix) pie3D(a, labels=lbs, main='Pie Chart Depicting Ratio of Female and Male')

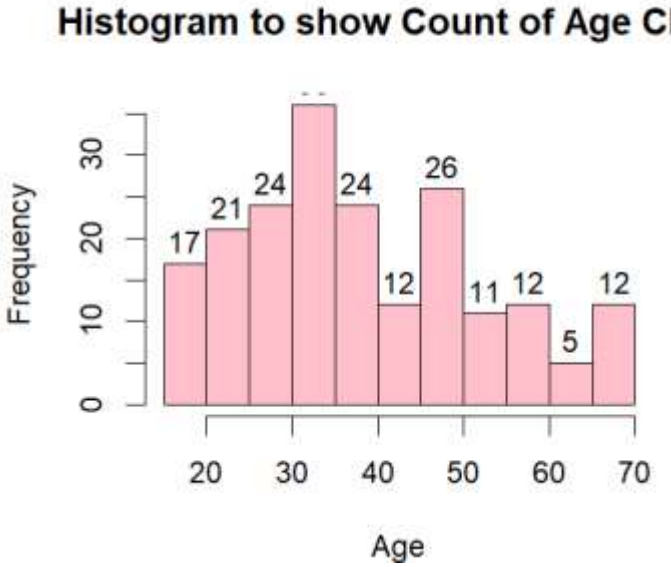
- **Insights:** The bar plot and pie chart show that the percentage of females is 56%, whereas the percentage of males is 44%.

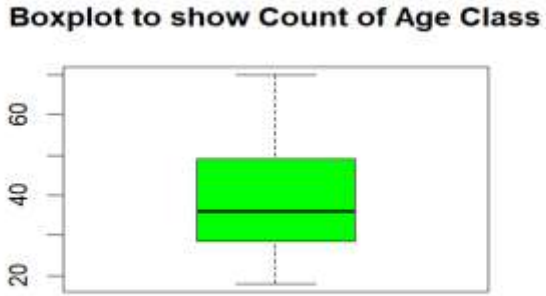




Age Distribution Visualization

- **Histogram:** Code: hist(customer_data$Age, col='blue', main='Histogram to Show Count of Age Class', xlab='Age Class', ylab='Frequency', labels=TRUE)
- **Box Plot:** Code: boxplot(customer_data$Age, col='ff0066', main='Boxplot for Descriptive Analysis of Age')
- **Insights:** Most customers are between 30 and 35 years old. The minimum age is 18, and the maximum age is 70.



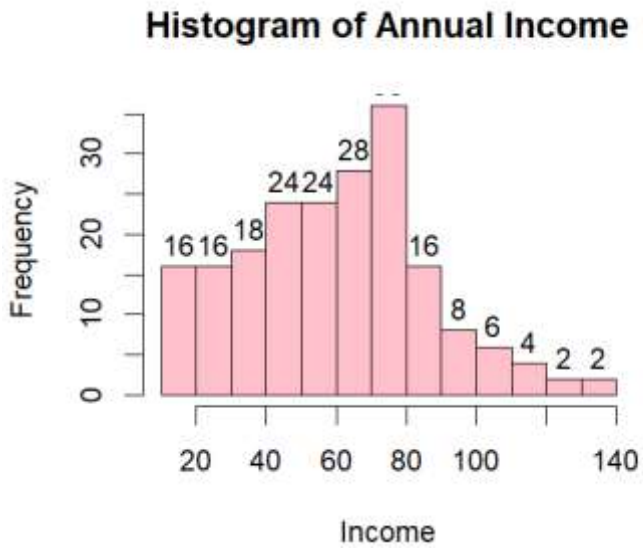


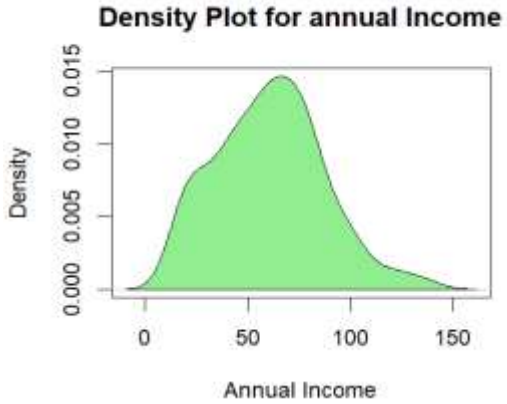
Annual Income Analysis

- **Histogram:** Code:
  hist(customer_data$Annual.Income..k.., col='#660033', main='Histogram for Annual Income', xlab='Annual Income Class', ylab='Frequency', labels=TRUE)

- **Density Plot:** Code:
  plot(density(customer_data$Annual.Income..k..), col='yellow', main='Density Plot for Annual Income', xlab='Annual Income Class', ylab='Density')
  polygon(density(customer_data$Annual.Income..k..), col='#ccff66')

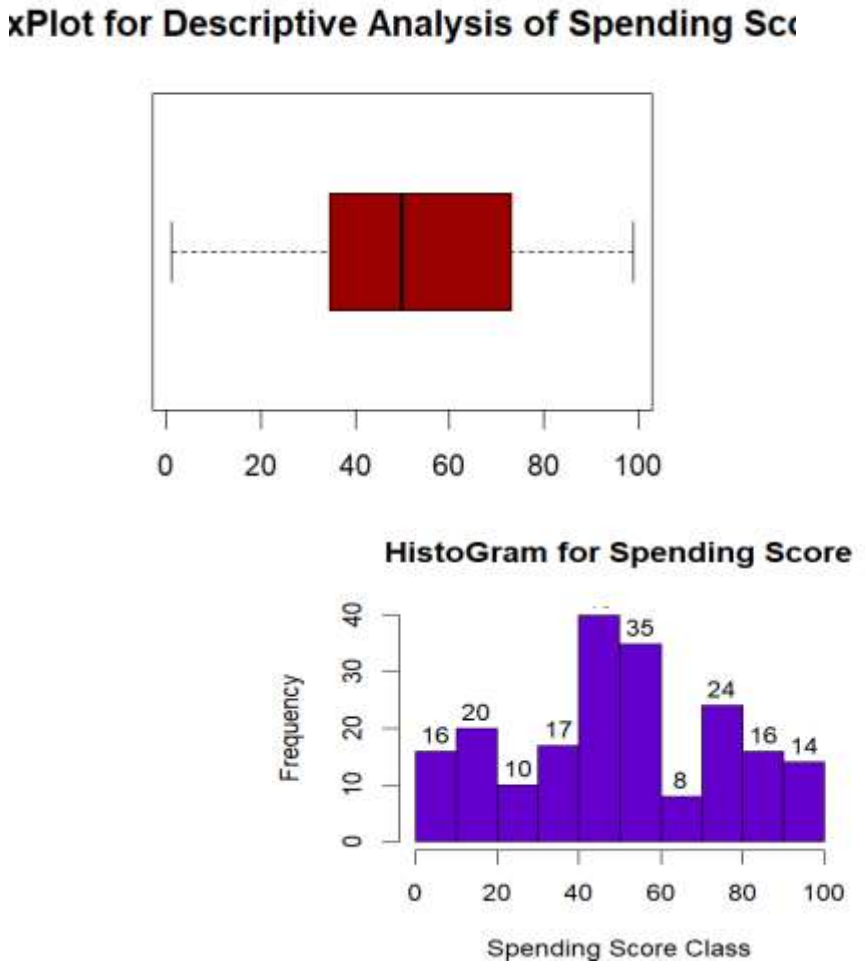
- **Insights:** The minimum annual income is 15k, and the maximum is 137k. The average income is 60.56k, with a normal distribution observed in the density plot.





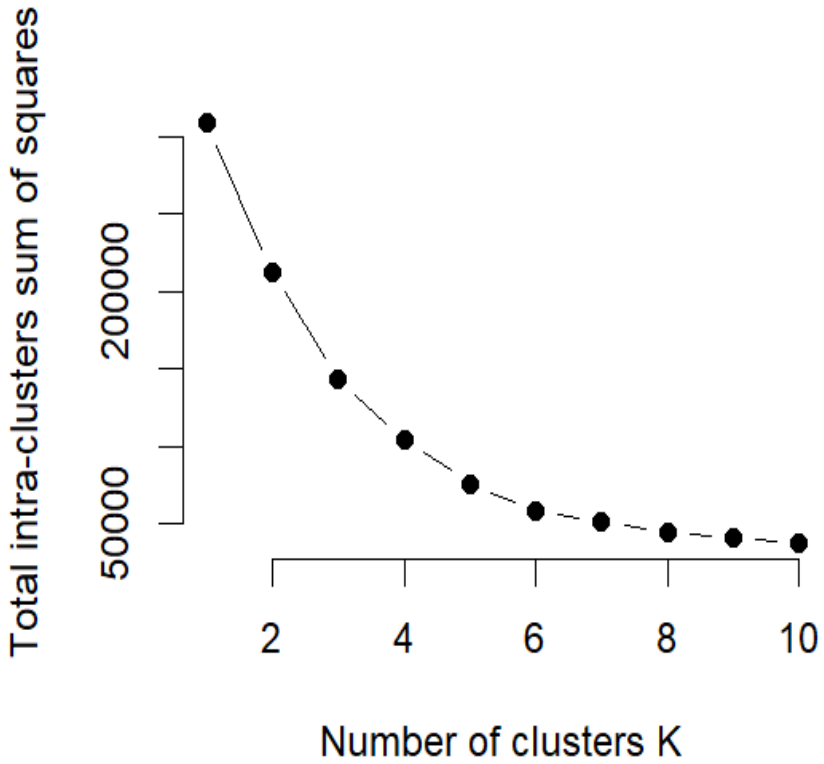
Spending Score Analysis

- **Box Plot:** Code: boxplot(customer_data$Spending.Score..1.100., horizontal=TRUE, col='#990000', main='BoxPlot for Descriptive Analysis of Spending Score')
- **Histogram:** Code: hist(customer_data$Spending.Score..1.100., main='Histogram for Spending Score', xlab='Spending Score Class', ylab='Frequency', col='#6600cc', labels=TRUE)
- **Insights:** The minimum spending score is 1, the maximum is 99. Most customers have a median spending score of 50.



K-means Algorithm

- **Optimal Clusters:** Use the elbow method to determine the optimal number of clusters by plotting the total intra-cluster sum of squares for different values of k.
- **Code Implementation:** Code: library(purrr) set.seed(123) iss <- function(k) {kmeans(customer_data[,3:5],k,iter.max=100,nstart=100,algorithm='Lloyd')$tot.withinss} k.values <- 1:10 iss_values <- map_dbl(k.values, iss) plot(k.values, iss_values, type='b', pch=19, frame=FALSE, xlab='Number of clusters K', ylab='Total intra-clusters sum of squares')
- **Insights:** The elbow plot shows that 4 is the optimal number of clusters as it appears at the bend in the graph.

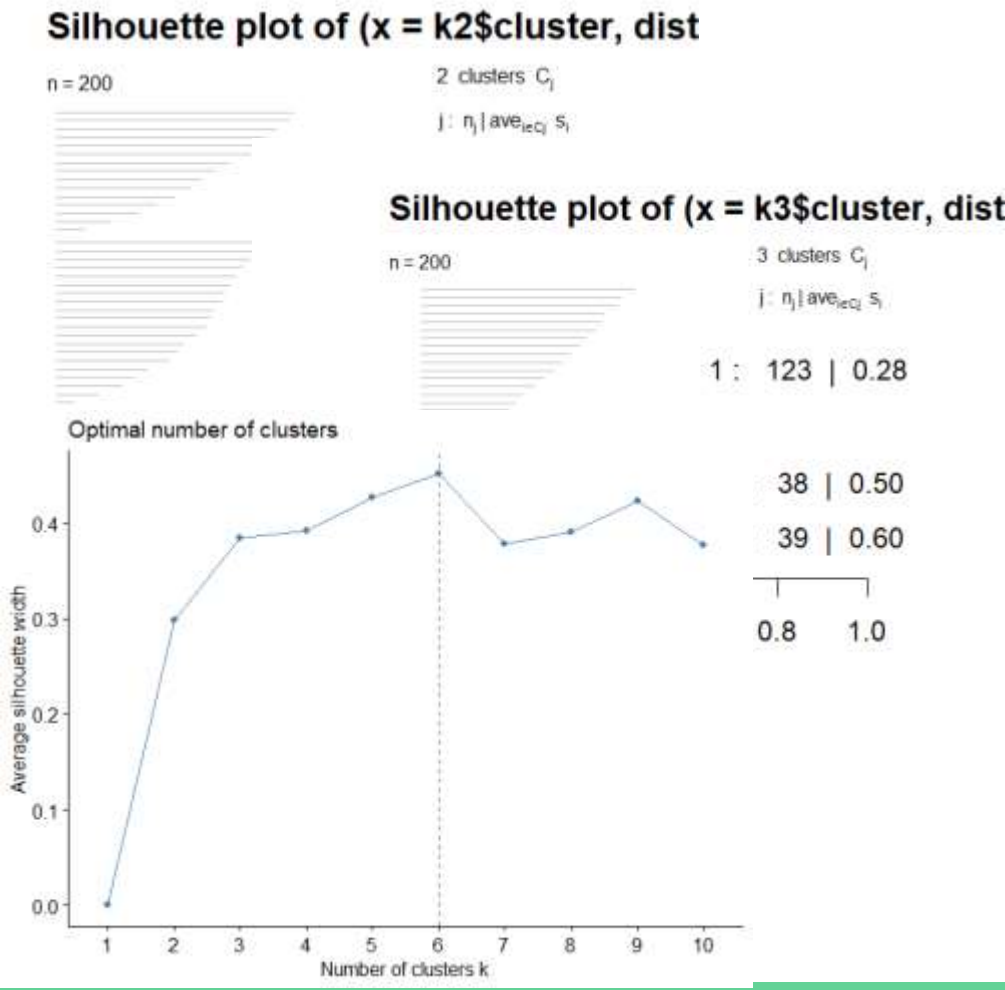


Average Silhouette Method

- **Introduction:** The silhouette method is used to determine the optimal number of clusters by computing the average silhouette width for different numbers of clusters.

- **Code Implementation:** Code: library(cluster) k2<- kmeans(customer_data[,3:5],2,iter.max=100,nstart=50,algorithm='Lloyd') s2<- plot(silhouette(k2$cluster,dist(customer_data[,3:5],'euclidean'))) k3<- kmeans(customer_data[,3:5],3,iter.max=100,nstart=50,algorithm='Lloyd') s3<- plot(silhouette(k3$cluster,dist(customer_data[,3:5],'euclidean')))

- **Insights:** The optimal number of clusters is determined by the highest average silhouette width.

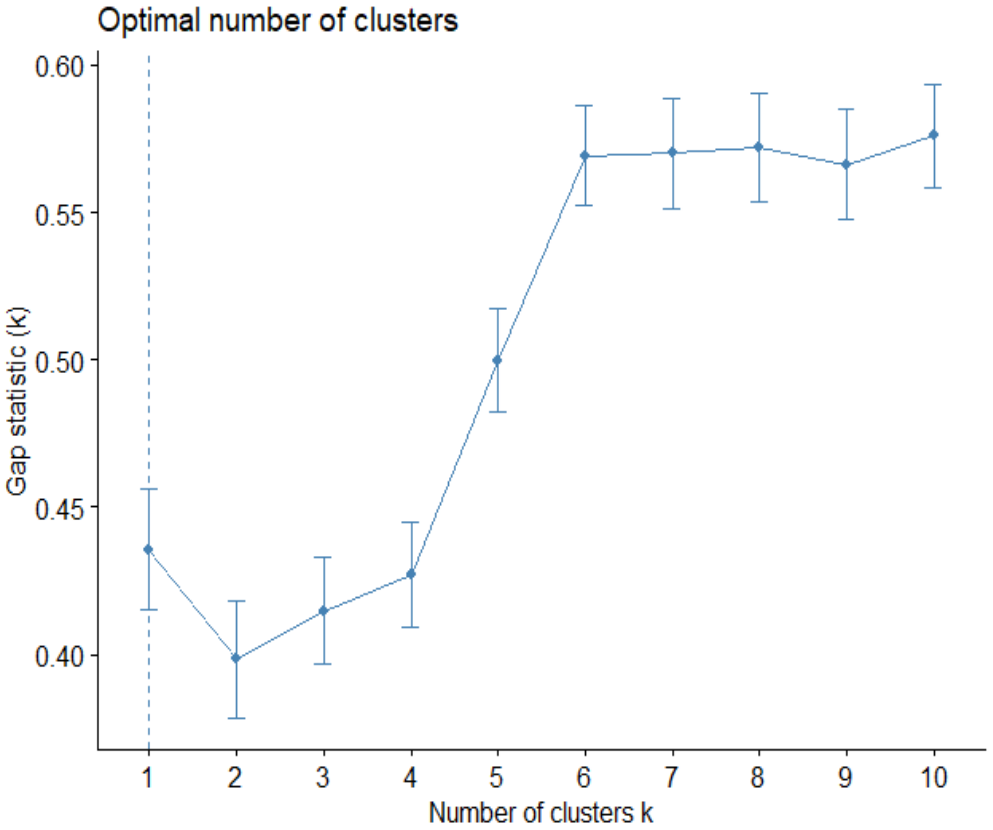


Gap Statistic Method

- **Introduction:** The gap statistic method is used to estimate the optimal number of clusters by comparing the total within intra-cluster variation for different numbers of clusters with their expected values under null reference distribution.

- **Code Implementation:** Code: library(cluster) stat_gap <- clusGap(customer_data[,3:5], FUN=kmeans, nstart=25, K.max=10, B=50) plot(stat_gap)

- **Insights:** The optimal number of clusters is determined by the value of k that maximizes the gap statistic.

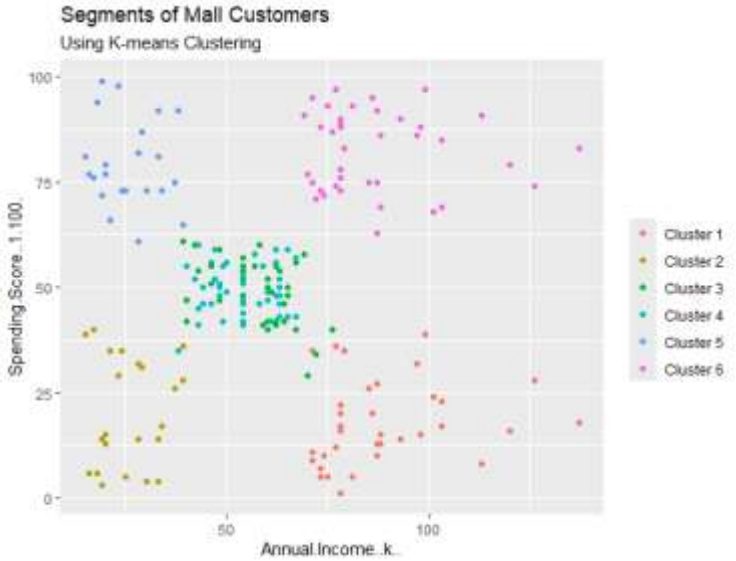


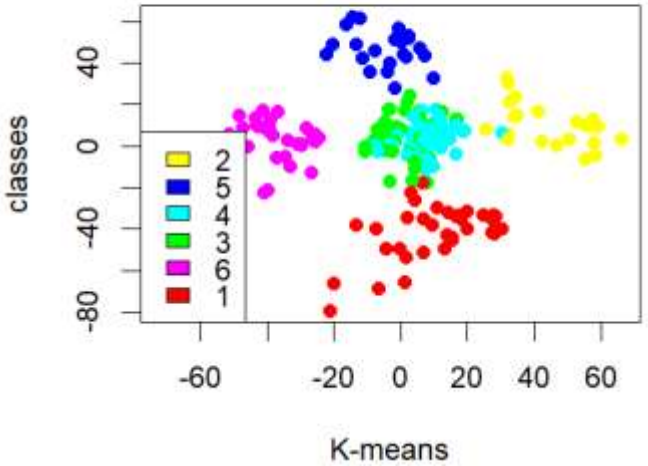
Visualizing Clustering Results

- **Principal Component Analysis (PCA):** Perform PCA to reduce dimensionality and visualize the clustering results.

- **Cluster Visualization:** Code:
  pcclust=prcomp(customer_data[,3:5],scale=FALSE)
  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')

- **Insights:** The visualization shows a clear separation of the six clusters based on their annual income and spending score.
  - Cluster 6 and 4 – These clusters represent the customer_data with the medium income salary as well as the medium annual spend of salary.
  - Cluster 1 – This cluster represents the customer_data having a high annual income as well as a high annual spend.
  - Cluster 3 – This cluster denotes the customer_data with low annual income as well as low yearly spend of income.
  - Cluster 2 – 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.





Summary and Conclusion

- **Project Recap:** Recapped the key steps: data exploration, visualizations, and K-means clustering to segment customers.
- **Key Insights:** Identified key insights: different customer segments based on demographic and spending patterns.
- **Future Work:** Potential future improvements: incorporating more features, exploring other clustering algorithms, and applying the model to other datasets.