**Laporan Analisis Data Mall Customers**

**Tujuan**

Kode ini bertujuan untuk melakukan analisis data pelanggan dari pusat perbelanjaan (mall) menggunakan metode eksplorasi data dan pengelompokan (clustering) dengan algoritma K-means. Analisis ini mencakup distribusi demografis pelanggan, distribusi pendapatan tahunan, skor pengeluaran, dan segmentasi pelanggan berdasarkan karakteristik tertentu.

**Penjelasan Setiap Bagian Kode**

**1. Import Data**

customers <- read.csv("~/Desktop/Mall\_Customers.csv.xls", stringsAsFactors=TRUE)

View(customers)

* **Tujuan**: Membaca dataset *Mall\_Customers* yang tersimpan di direktori lokal.
* **Penjelasan**: Dataset dimuat dalam bentuk data frame dengan opsi stringsAsFactors=TRUE, yang mengubah kolom berbentuk teks menjadi tipe *factor*.

**2. Memuat Library yang Dibutuhkan**

library(ggplot2); theme\_set(theme\_bw())

library(dplyr)

library(skmeans)

library(gridExtra)

attach(customers)

* **Tujuan**: Memuat library yang dibutuhkan untuk analisis data dan visualisasi.
  + ggplot2: Untuk membuat visualisasi.
  + dplyr: Untuk manipulasi data.
  + skmeans: Untuk pengelompokan (*spherical k-means*).
  + gridExtra: Untuk menyusun beberapa grafik dalam satu tampilan.
* theme\_set(theme\_bw()): Mengatur tema bawaan pada semua visualisasi menjadi tema *black and white*.
* attach(customers): Mempermudah akses langsung ke kolom dalam dataset tanpa harus menggunakan tanda $.

**3. Eksplorasi Data Awal**

dim(customers)

str(customers)

n\_distinct(CustomerID)

* **Tujuan**: Mengeksplorasi dimensi, struktur data, dan jumlah pelanggan unik.
  + dim(customers): Memberikan dimensi data (jumlah baris dan kolom).
  + str(customers): Menampilkan struktur data, termasuk tipe data dari setiap kolom.
  + n\_distinct(CustomerID): Menghitung jumlah pelanggan unik berdasarkan kolom CustomerID.

**4. Distribusi Gender**

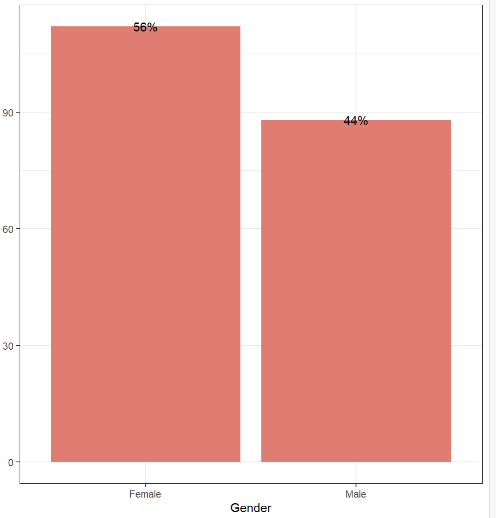
as.data.frame(table(Gender)) %>%

ggplot(aes(x = Gender, y = Freq)) +

geom\_bar(stat = "identity", fill = "#F8766D") +

geom\_text(y = as.vector(table(Gender)), label = paste0((as.vector(table(Gender))/sum(as.vector(table(Gender))))\*100, "%"))

* **Tujuan**: Menampilkan distribusi gender pelanggan dalam bentuk diagram batang.
  + table(Gender): Menghitung jumlah pelanggan berdasarkan jenis kelamin.
  + geom\_bar: Membuat diagram batang.
  + geom\_text: Menambahkan label persentase pada setiap batang.

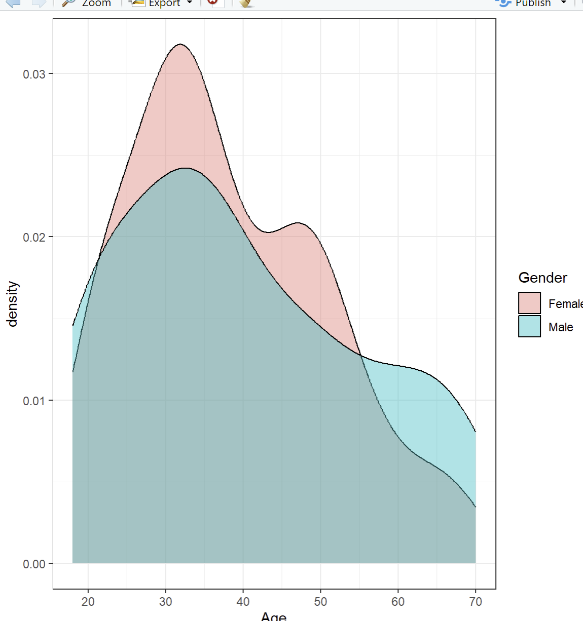


**5. Analisis Usia**

summary(Age)

ggplot(as.data.frame(Age), aes(y = Age)) + geom\_boxplot(fill='#F8766D')

ggplot(customers, aes( x = Age, fill = Gender)) + geom\_density(alpha = 0.4)

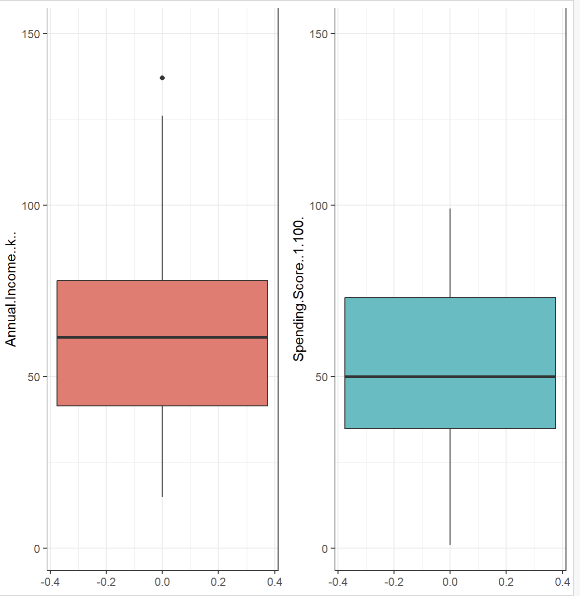
* **Tujuan**: Menganalisis distribusi usia pelanggan.
  + summary(Age): Menampilkan statistik deskriptif (rata-rata, median, min, max, dll.) untuk usia.
  + geom\_boxplot: Membuat boxplot untuk mendeteksi outlier.
  + geom\_density: Membuat grafik distribusi kepadatan usia, dipisahkan berdasarkan jenis kelamin.

**6. Analisis Pendapatan dan Skor Pengeluaran**

p1 <- ggplot(as.data.frame(Annual.Income..k..), aes(y = Annual.Income..k..)) + geom\_boxplot(fill='#F8766D') + ylim(c(1,150))

p2 <- ggplot(as.data.frame(Spending.Score..1.100.), aes(y = Spending.Score..1.100.)) + geom\_boxplot(fill='#00BFC4') + ylim(c(1,150))

grid.arrange(p1, p2, ncol = 2)

* **Tujuan**: Menampilkan distribusi pendapatan tahunan dan skor pengeluaran pelanggan.
  + geom\_boxplot: Membuat boxplot untuk kedua variabel (pendapatan tahunan dan skor pengeluaran).
  + grid.arrange: Menampilkan kedua grafik secara berdampingan dalam satu tampilan.

**7. Memeriksa Data yang Hilang**

cat("There are", sum(is.na(customers)), "N/A values.")

* **Tujuan**: Menghitung jumlah nilai yang hilang (*missing values*) dalam dataset.

**8. Penentuan Jumlah Kluster dengan Elbow Method**

Kdata <- customers[,c(4,5)]

tot.withinss <- vector("numeric", length = 10)

for (i in 1:10){

kDet <- kmeans(Kdata, i)

tot.withinss[i] <- kDet$tot.withinss

}

ggplot(as.data.frame(tot.withinss), aes(x = seq(1,10), y = tot.withinss)) +

geom\_point(col = "#F8766D") +

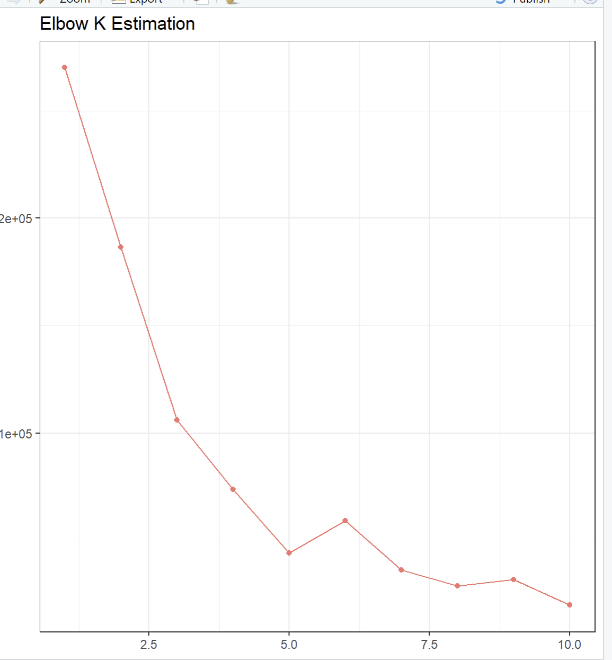
geom\_line(col = "#F8766D") +

theme(axis.title.x.bottom = element\_blank()) +

ylab("Within-cluster Sum of Squares") +

xlab("Number of Clusters") +

ggtitle("Elbow K Estimation")

* **Tujuan**: Menentukan jumlah kluster optimal menggunakan metode Elbow.
  + Kdata: Memilih kolom untuk analisis clustering (Pendapatan Tahunan dan Skor Pengeluaran).
  + kmeans: Melakukan K-means clustering untuk berbagai jumlah kluster (1 hingga 10).
  + tot.withinss: Menyimpan total *within-cluster sum of squares* untuk setiap nilai k.
  + geom\_point dan geom\_line: Membuat grafik Elbow.

**9. Membuat dan Visualisasi Kluster**

customerClusters <- kmeans(Kdata, 5)

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

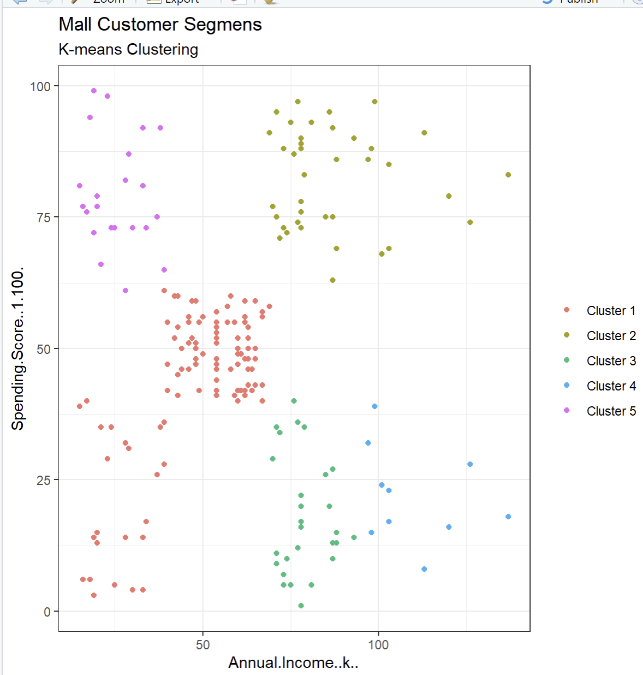
geom\_point(stat = "identity", aes(color = as.factor(customerClusters$cluster))) +

scale\_color\_discrete(name=" ",

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

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

ggtitle("Mall Customer Segmens", subtitle = "K-means Clustering")

* **Tujuan**: Membagi pelanggan menjadi 5 kluster dan menampilkan hasilnya dalam grafik.
  + kmeans(Kdata, 5): Melakukan K-means clustering dengan 5 kluster.
  + geom\_point: Memetakan pelanggan berdasarkan pendapatan tahunan dan skor pengeluaran, dengan warna yang menunjukkan kluster.

**10. Menyimpan Riwayat R**

savehistory("~/DAP.Rhistory")

* **Tujuan**: Menyimpan riwayat perintah R ke file lokal agar dapat diakses kembali di kemudian hari.

**Kesimpulan**

Kode ini berhasil mengelompokkan pelanggan mall ke dalam 5 segmen berdasarkan pendapatan tahunan dan skor pengeluaran. Hasil analisis ini dapat digunakan untuk strategi pemasaran yang lebih terarah dan optimal.