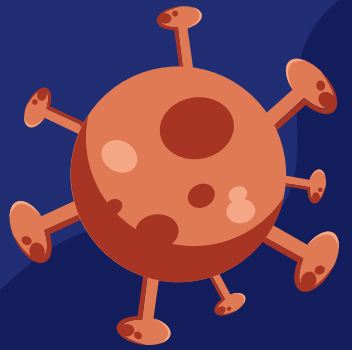
Assignment Study Case: Data Analysis C

A multivariate data analysis approach for investigating daily statistics of countries affected with COVID-19 pandemic in Europe

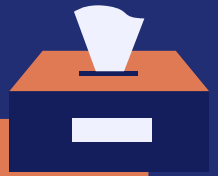
Presented by:
Haiva Qurrota A 06211840000045
Susilowati Gusinta 06211840000060

SUMBER DATA

Data tentang statistik Coronavirus dari website https://www.worldometers.info/coronavirus/ pada 14 November 2021 di benua Eropa. 1 Negara (Vatican City) dengan data yang hilang dihilangkan dari analisis untuk menjaga keterwakilan yang baik dari setiap variabel. Jumlah negara yang masuk dalam analisis adalah 47 negara.

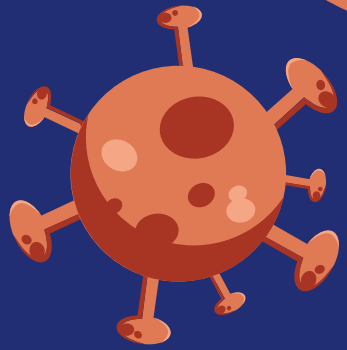


Variabel Penelitian



Variabel	Deskripsi
Total Cases	Total kasus terkonfirmasi COVID-19
Total Deaths	Total kasus kematian karena COVID-19
Total Recovered	Total kasus sembuh dari COVID-19
Active Cases	Total kasus terbuka (ringan, serius, kritis)
Mortality Recovery Ratio	Rasio antara total kematian dengan total pasien sembuh

ANALISIS PEMBAHASAN



Statistika Deskriptif

	Total Cases	Total Deaths	Total Recovery	Active Cases	Mortality Recovery Ratio
Min.	2.903	1.216	2.365	54	0,007186
1st Qu.	159.424	8.223	84.762	6.916	0,016134
Median	494.643	26.191	405.598	42.279	0,025434
Mean	1.448.410	90.200	1.286.064	133.574	2,642025
3rd Qu.	1.346.549	89.837	1.194.808	114.476	1,271811
Max	9.524.971	950.000	7.792.578	1.589.558	17,096413

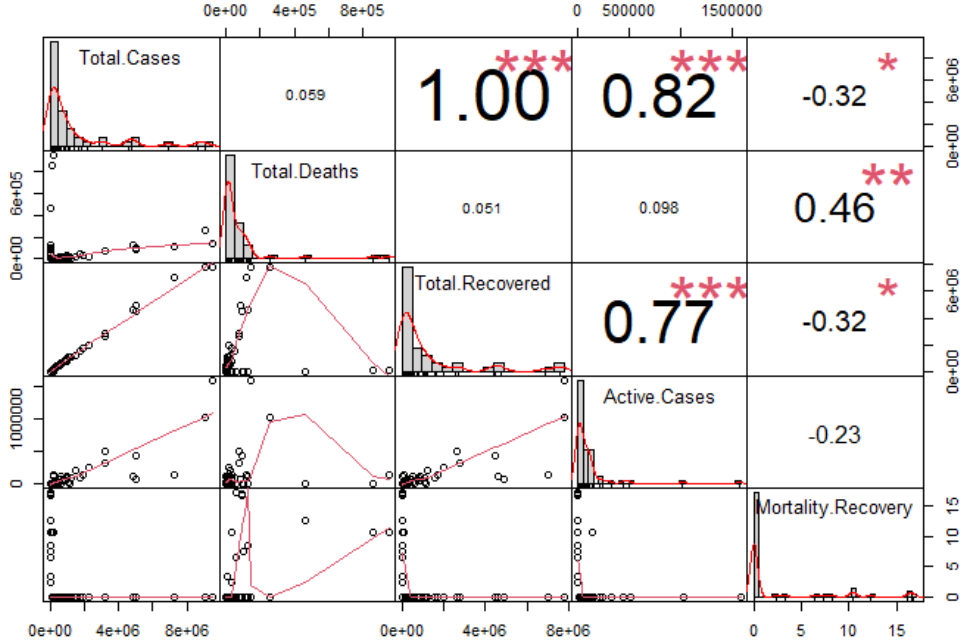
Uji Asumsi KMO

Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = r)
Overall MSA = 0.5
MSA for each item =

Total.Cases	Total.Deaths
0.5	0.5
Total.Recovered	Active.Cases
0.5	0.5
Mortality.Recovery	
0.65	

Pada tingkat kepercayaan 95%, data tersebut dikatakan tidak layak untuk analisis faktor karena memiliki nilai KMO=0,5. Akan tetapi, hampir seluruh nilai MSA (*Measure of Sampling Adequacy*) 0,5 sehingga masih bisa dianalisis lebih lanjut.

Uji Multikolinieritas



Terdapat nilai korelasi antar variabel yang melebihi 0.05, maka dapat dikatakan bahwa terjadi kasus multikolinearitas.

Uji Asumsi Homoskedastisitas

```
> bartlett.test(df_numeric)

        Bartlett test of homogeneity of variances

data:  df_numeric
Bartlett's K-squared = 1386.3, df = 4, p-value < 2.2e-16
```

Nilai p-value yang didapat adalah < 0.05, maka dapat diketahui bahwa sampel berasal dari varians populasi yang sama atau memenuhi asumsi homoskedastisitas

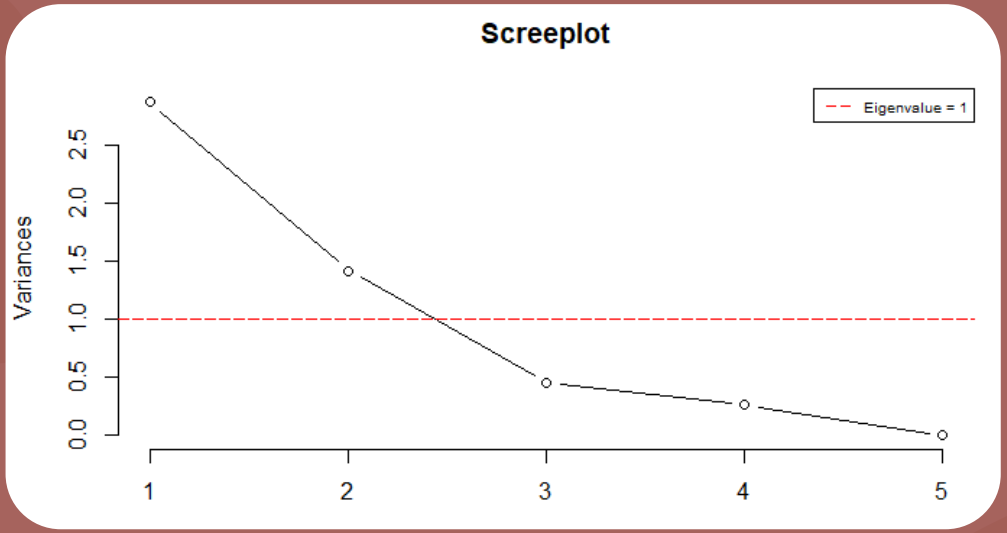
Principal Components Analysis (PCA)

```
> df_pca
Standard deviations (1, .., p=5):
[1] 1.693178470 1.191060951 0.671691483 0.513150035 0.005298365

Rotation (n x k) = (5 x 5):
                          PC1         PC2         PC3         PC4          PC5
Total.Cases       -0.578273079 0.08615684  0.08786671 -0.32196154  0.7394575380
Total.Deaths       0.003788657 0.75318988 -0.65764080  0.01410552 -0.0005076914
Total.Recovered   -0.570113211 0.07557541  0.07375311 -0.46778063 -0.6670839826
Active.Cases      -0.521840767 0.14360903  0.17905217  0.81657685 -0.0905606426
Mortality.Recovery 0.261226019 0.63162236  0.72269552 -0.10259302  0.0001480876
> summary(df_pca)
Importance of components:
                          PC1    PC2     PC3     PC4      PC5
Standard deviation     1.6932 1.1911 0.67169 0.51315 0.005298
Proportion of Variance 0.5734 0.2837 0.09023 0.05266 0.000010
Cumulative Proportion  0.5734 0.8571 0.94733 0.99999 1.000000
```

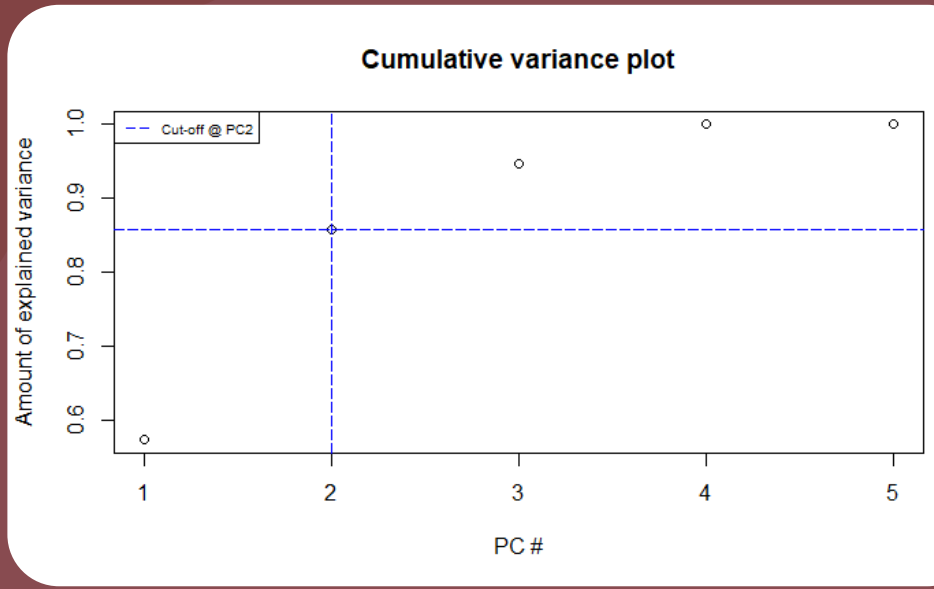
Dengan melihat nilai standar deviasi, maka jumlah faktor PC yang dapat digunakan adalah 2, yakni PC_1 dan PC_2 dikarenakan nilai standar deviasi lebih dari 1.

Scree Plot



Apabila ditinjau menggunakan screeplot, titik yang berada diatas 1 ada 2. Maka jumlah principal components yang terbentuk adalah 2, yakni PC_1 dan PC_2.

Cumulative Proportion



Maka, PC_1 dan PC_2 dapat menjelaskan kelima variabel sebesar 85,71 %

PCA

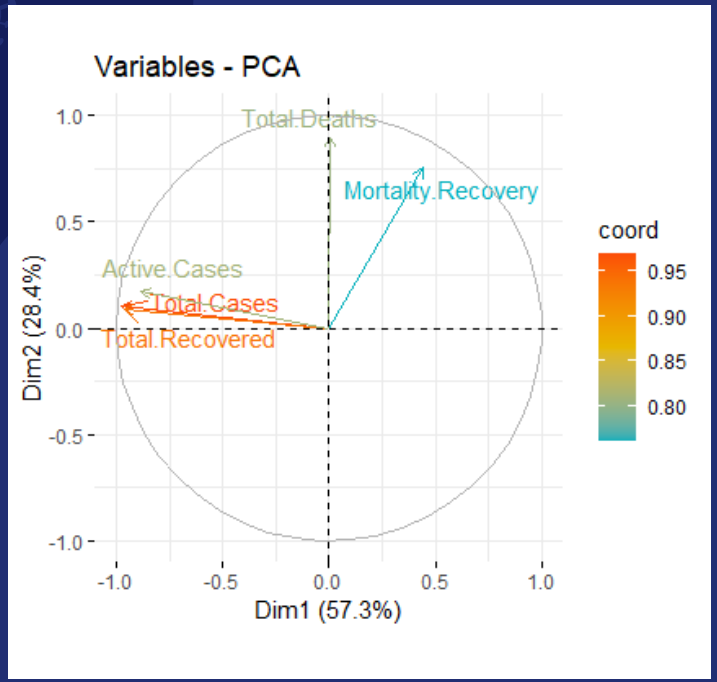
	PC1	PC2
Total Cases	-0.578273079	0.08615684
Total Deaths	0.003788657	0.75318988
Total Recovered	-0.570113211	0.07557541
Active Cases	-0.521840767	0.14360903
Mortality Recovery	0.261226019	0.63162236

Sehingga persamaan yang terbentuk yakni:

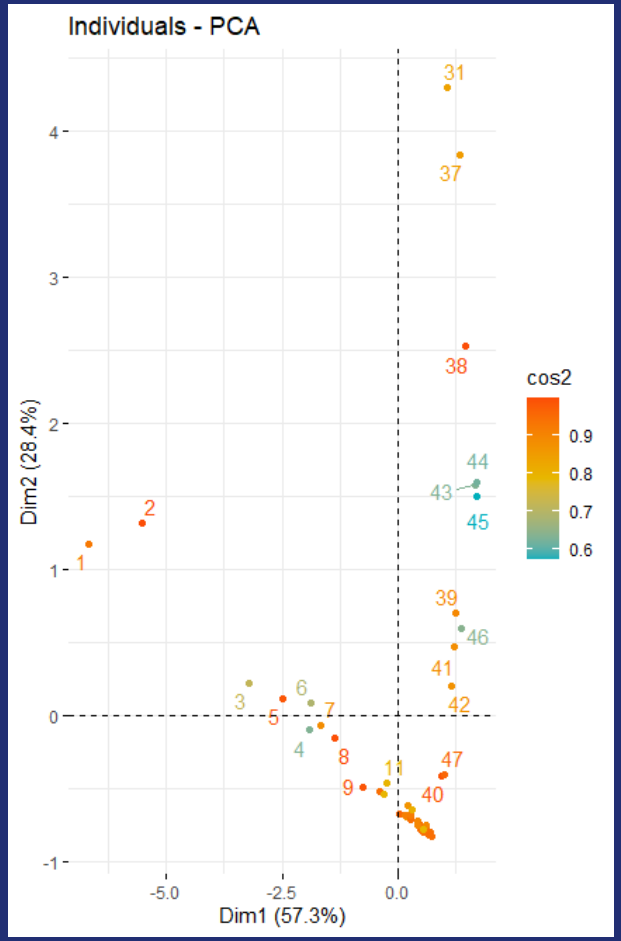
PC1 = – 0,58(Total Cases) – 0,57(Total Recovered) – 0,52(Active Cases)

PC2 = 0,75(Total Deaths) + 0,63(Mortality Recovery)

Heatmap Korelasi untuk Varibel Numerik



Gambar disamping menunjukkan korelasi variabel asli dengan faktor yang terbentuk. PC-1 menjelaskan sekitar 95% dari varians Total Cases dan Total Recovered, serta diantara 80-85% dari varians Total Deaths dan Active Cases. Sementara PC-2 menjelaskan dibawah 80% dari varians rasio Mortality Recovery.



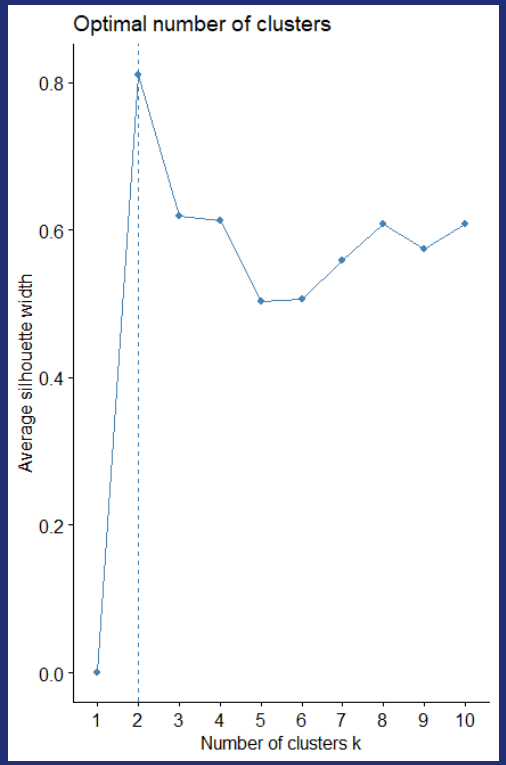
Heatmap Korelasi untuk Varibel Kategorik

Dimensi 1 dan 2 cukup untuk mempertahankan 86% dari total inersia (variasi) yang terkandung dalam data. Cosinus kuadrat (cos2) merupakan kualitas representasi yang mengukur derajat hubungan antara variabel kategori (Country) dan sumbu tertentu.

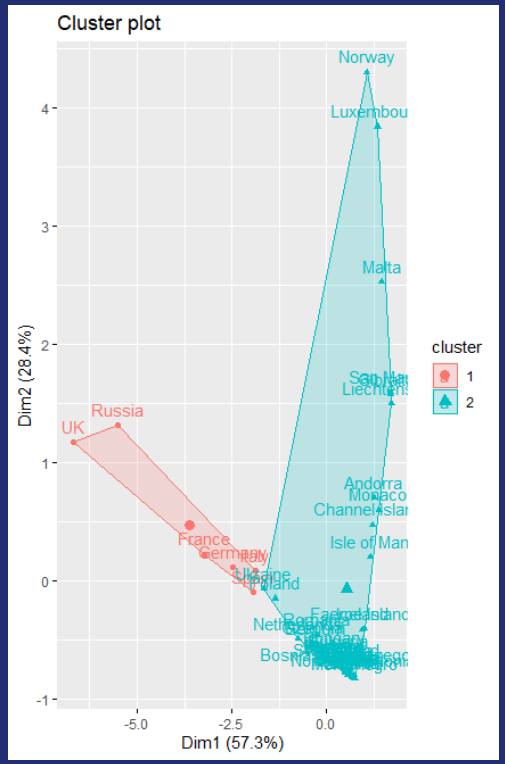
- Negara dengan cos2 rendah berwarna biru
- Negara dengan cos2 sedang berwarna kuning
- Negara dengan cos2 tinggi berwarna oranye.

Plot di samping memberikan gambaran pada dimensi yang mana variabel kategori memberikan kontribusi.

KMEANS CLUSTERING



Dengan metode *sillouette* didapatkan jumlah cluster optimum adalah 2.



Disajikan gambar clustering untuk negara di Eropa berdasarkan skor PC. Cluster 1 terdiri atas 6 negara yaitu UK, Russia, France, Spain, Germany, dan Italy. Sementara cluster 2 merupakan 41 negara lainnya di benua Eropa.

Syntax R

```
setwd("C:/Users/Iva/Documents/STATISTIK'18/THN 4/Semester 7/Andat C/Tugas/Week 10 - Kelompok")
df= read.csv("bismillah.csv", header=TRUE, sep=";")
head(df)
summary(df)
str(df)
df$Total.Cases<- as.numeric(df$Total.Cases)
df$Total.Deaths<- as.numeric(df$Total.Deaths)
df$Total.Recovered<- as.numeric(df$Total.Recovered)
df$Active.Cases<- as.numeric(df$Active.Cases)
df_numeric<-df[,2:6]
df_numeric

#KMO
library(psych)
r=cor(df_numeric)
KMO(r)

#uji Barlett
bartlett.test(df_numeric)

#Uji Multiko
library("PerformanceAnalytics")
chart.Correlation(df_numeric, histogram=TRUE, pch=19)

#PCA
df_pca <- prcomp(x = df_numeric, scale. = TRUE, center = TRUE)
names(df_pca)
summary(df_pca)

#screeplot
screeplot(df_pca, type = "l", npcs = 5, main = "Screeplot")
abline(h = 1, col="red", lty=5)
legend("topright", legend=c("Eigenvalue = 1"), col=c("red"), lty=5, cex=0.6)

#cumulative Proportion
cumpro <- cumsum(df_pca$sdev^2 / sum(df_pca$sdev^2))
plot(cumpro[0:5], xlab = "PC #", ylab = "Amount of explained variance", main = "Cumulative variance plot")
abline(v = 2, col="blue", lty=5)
abline(h = 0.8571, col="blue", lty=5)
legend("topleft", legend=c("Cut-off @ PC2"), col=c("blue"), lty=5, cex=0.6)
```

Syntax R

```
#Proportion of variance
plot(df_pca$x[,1],df_pca$x[,2],
xlab="PC1 (57.34%)", ylab = "PC2
(28.37%)",
    main = "PC1 / PC2 - plot")
ncomp=2
df_pca$rotation[,1:2]
rawLoadings     <-
df_pca$rotation[,1:ncomp] %*%
diag(df_pca$sdev, ncomp, ncomp)
rotatedLoadings <-
varimax(rawLoadings)$loadings

library("factoextra")
fviz_pca_var(df_pca,
col.var="coord",
          gradient.cols = c("#00AFBB",
"#E7B800", "#FC4E07"),
repel = TRUE, # Avoid text
overlapping
axes = c(1, 2)) # choose PCs to plot

#FAMD
library("FactoMineR")
library("factoextra")
FAMD (df, ncp = 5, sup.var = NULL,
ind.sup = NULL, graph = TRUE)
res.famd <- FAMD(df, graph =
FALSE)
print(res.famd)

#Biplot Country
ind <- get_famd_ind(res.famd)
ind
head(ind$coord)
fviz_famd_ind(res.famd, col.ind =
"cos2", repel = TRUE) +
ggtitle("Clustering according to
PC") +  theme(plot.title =
element_text(hjust = 0.5))

#K-Means Cluster
rownames(df_numeric) <-
df$Country
fviz_nbclust(df_numeric, kmeans,
method = "silhouette") # metode
silhouette
km <- kmeans(df_numeric, centers
= 3, nstart = 25)
str(km)
fviz_cluster(km, data =
df_numeric)
```

Terimakasih

Analisis Data C
Tahun Ajaran 2021/2022
S1 – Statistika
FSAD - ITS