Final Project

STAT-340 Applied Regression Methods

Mina Dao & Yerim Oh

Load Packages

The following R code loads packages needed in this assignment.

Data

2023 Global Country Development & Prosperity Index https://www.kaggle.com/datasets/tarktunataalt/2023-global-country-development-and-prosperity-index

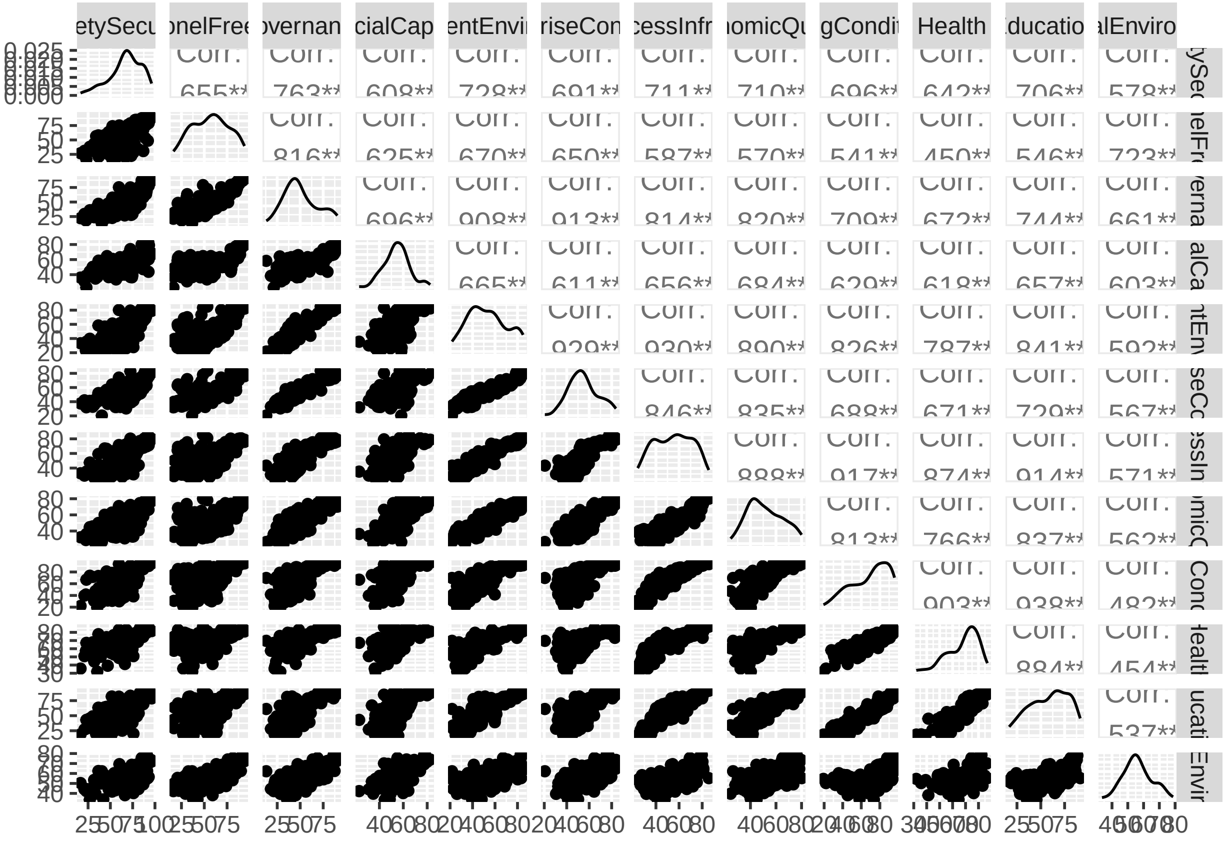
This dataset contains rankings and indicators from the 2023 Legatum Prosperity Index, assessing various dimensions of prosperity and development across countries. There are 167 observations according to 167 countries and 12 numerical variables:

- `SafetySecurity`: Freedom from conflict, terrorism, and crime
- `PersonelFreedom`: Right to speech, assembly, and individual autonomy
- `Governance`: Quality of democracy, rule of law, and government effectiveness
- `SocialCapital`: Strength of personal relationships and civic engagement
- `InvestmentEnvironment`: Conditions for private investment and credit access
- `EnterpriseConditions`: Business environment and market competition
- `MarketAccessInfrastructure`: Ease of trade and quality of infrastructure
- `EconomicQuality`: Macroeconomic stability and employment quality
- `LivingConditions`: Standard of living and access to basic services.
- `Health`: Population health and healthcare access.
- `Education`: Quality and accessibility of education.
- `NaturalEnvironment`: Environmental quality and sustainability.

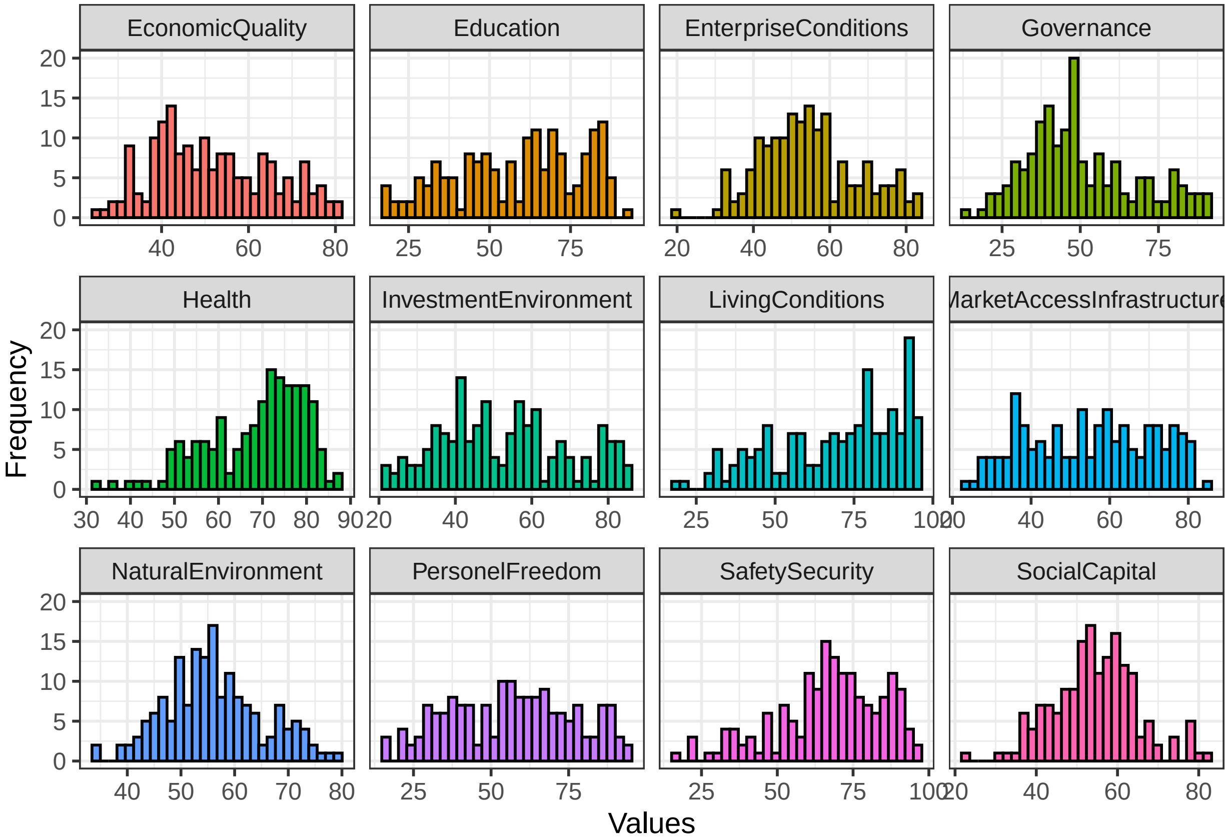
```
##               SafetySecurity PersonelFreedom Governance SocialCapital
##  Denmark               92.59           94.09      89.45         82.56
##  Sweden                90.97           91.90      86.41         78.29
##  Norway                93.30           94.10      89.66         79.03
##  Finland               89.56           91.96      90.41         77.27
##   Switzerland          95.66           87.50      87.67         69.14
##  Netherlands           91.19           90.08      87.34         74.03
##               InvestmentEnvironment EnterpriseConditions
##  Denmark                      82.42                79.64
##  Sweden                       82.81                75.54
##  Norway                       82.24                75.95
##  Finland                      84.12                77.25
##   Switzerland                 80.81                83.84
##  Netherlands                  84.11                79.09
##                MarketAccessInfrastructure EconomicQuality LivingConditions
##  Denmark                           78.79           76.81            95.77
##  Sweden                            79.67           76.18            95.33
##  Norway                            75.87           77.25            94.70
##  Finland                           78.77           70.28            94.46
##   Switzerland                      78.65           79.71            94.66
##  Netherlands                       80.82           74.34            95.86
##                Health Education NaturalEnvironment
##  Denmark        81.07     87.48              73.94
##  Sweden         82.28     85.92              78.74
##  Norway         82.98     85.68              72.37
##  Finland        81.19     88.38              77.99
##   Switzerland   82.11     87.72              73.60
##  Netherlands    82.05     86.43              62.49
```

Data Visualization

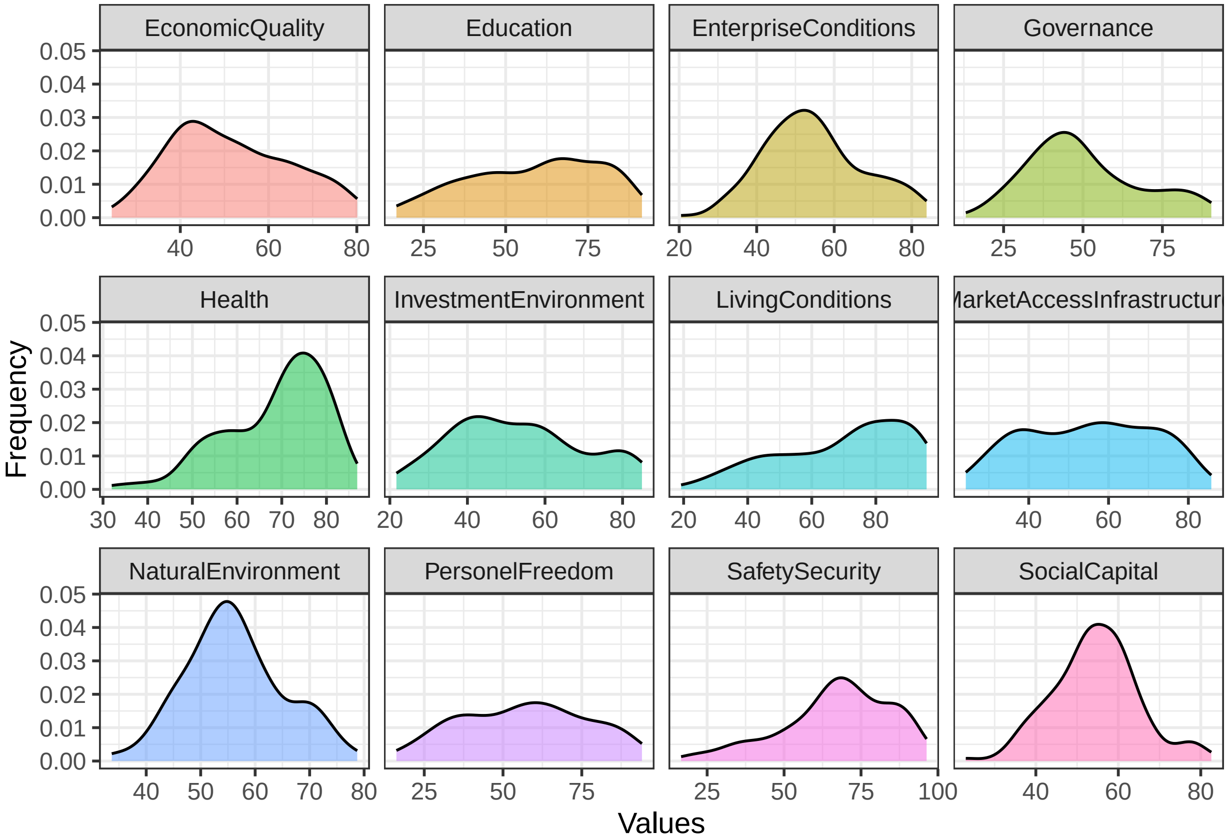
```r
ggpairs(country)
```



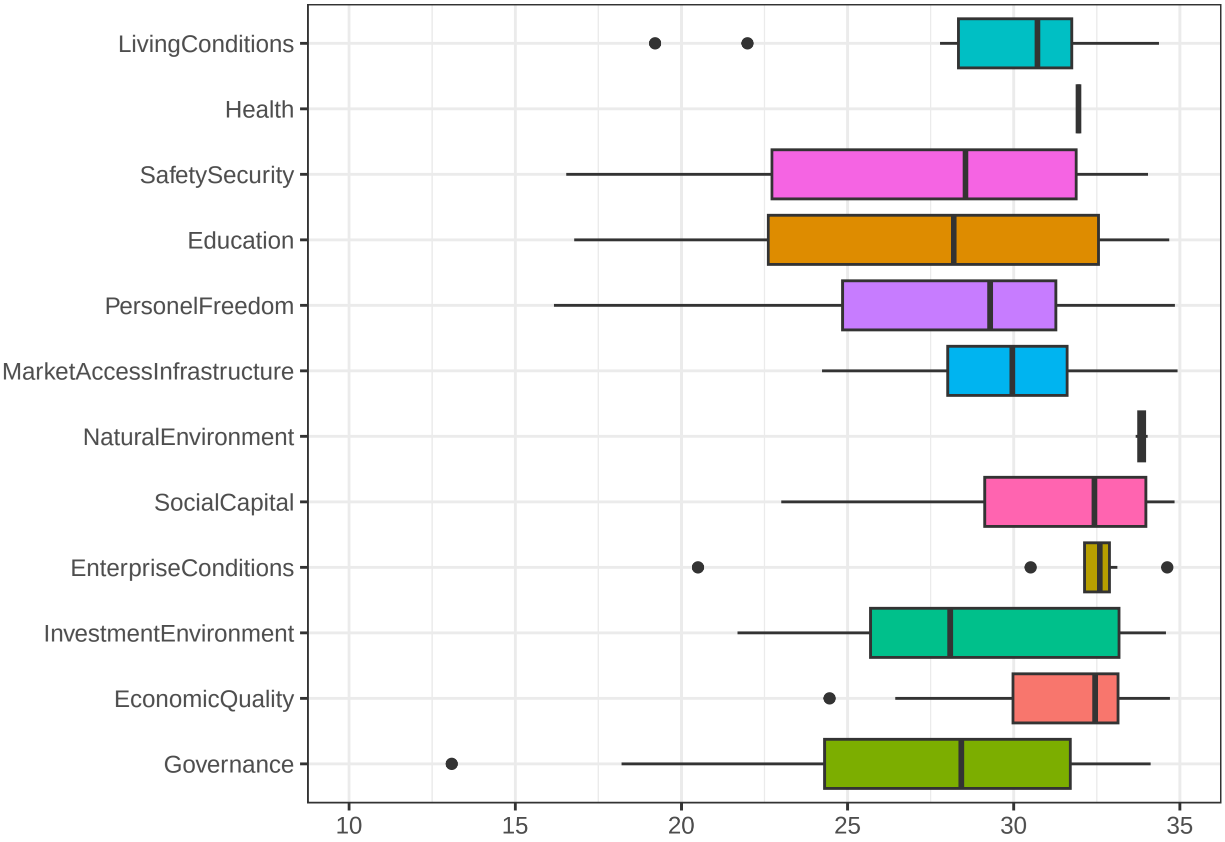
```
# Histogram for each variable
country %>%
  gather(Attributes, value, 1:12) %>%
  ggplot(aes(x=value, fill=Attributes)) +
  geom_histogram(colour="black", show.legend=FALSE) +
  facet_wrap(~Attributes, scales="free_x") +
  labs(x="Values", y="Frequency") +
  theme_bw()
```



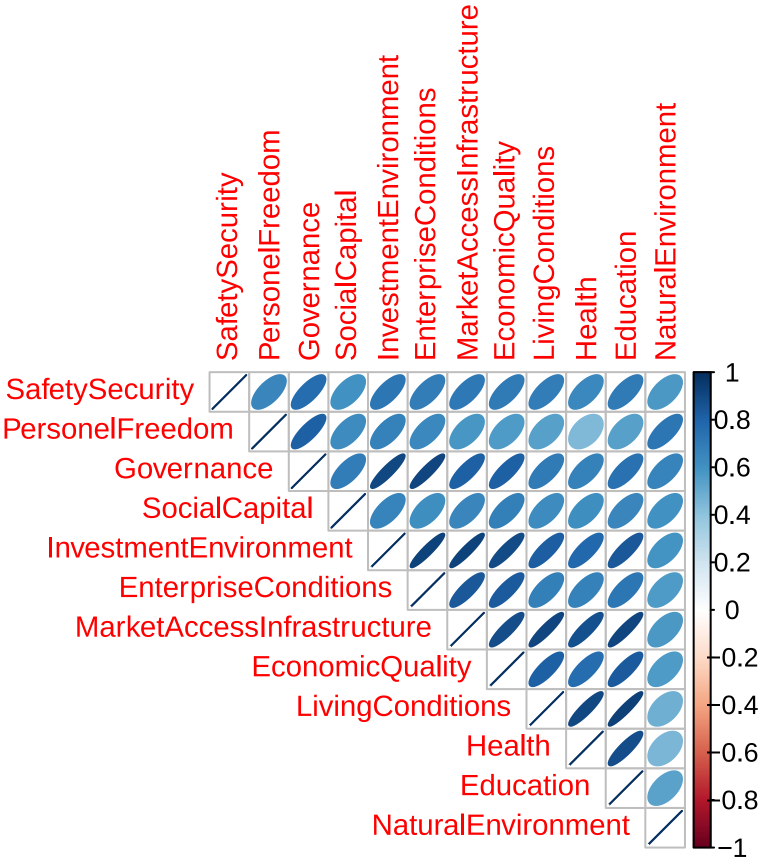
```r
# Density plot for each variable
country %>%
  gather(Attributes, value, 1:12) %>%
  ggplot(aes(x=value, fill=Attributes)) +
  geom_density(colour="black", alpha=0.5, show.legend=FALSE) +
  facet_wrap(~Attributes, scales="free_x") +
  labs(x="Values", y="Frequency") +
  theme_bw()
```



```
# Boxplot for each variable
country %>%
  gather(Attributes, values, c(1:12)) %>%
  ggplot(aes(x=reorder(Attributes, values, FUN=median), y=values, fill=Attributes)) +
  geom_boxplot(show.legend=FALSE) +
  theme_bw() +
  theme(axis.title.y=element_blank(),
        axis.title.x=element_blank()) +
  ylim(10, 35) +
  coord_flip()
```



```
# Correlation matrix
corrplot(cor(country), type="upper", method="ellipse", tl.cex=0.9)
```

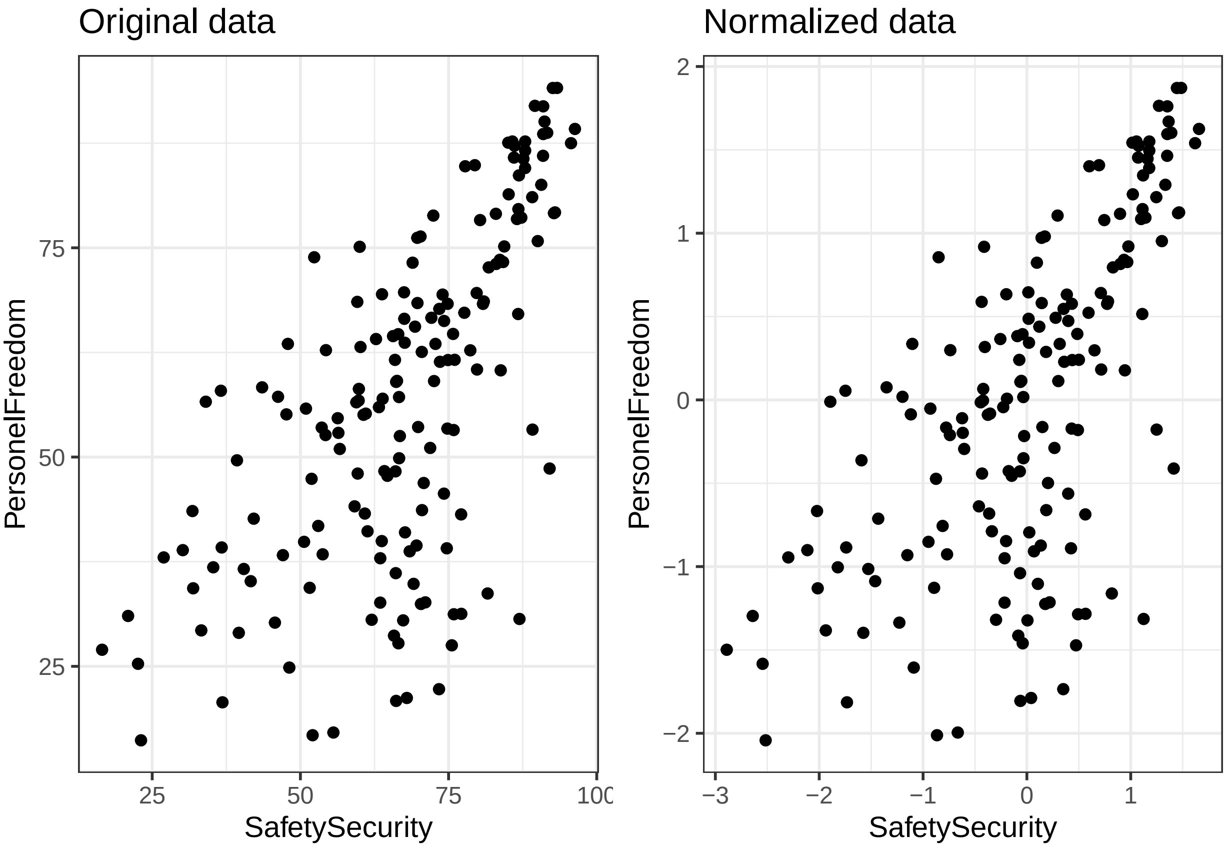


We could see that most of the variables, except for `PersonalFreedom` and `NaturalEnvironment`, have a strong linear correlation.

Data Preparation

We have to normalize the variables to express them in the same range of values. In other words, normalization means adjusting values measured on different scales to a common scale.

```r
# Normalization
country_norm <- as.data.frame(scale(country))

# Original data example
p1 <- ggplot(country, aes(x=SafetySecurity, y=PersonelFreedom)) +
  geom_point() +
  labs(title="Original data") +
  theme_bw()

# Normalized data
p2 <- ggplot(country_norm, aes(x=SafetySecurity, y=PersonelFreedom)) +
  geom_point() +
  labs(title="Normalized data") +
  theme_bw()

# Subplot
grid.arrange(p1, p2, ncol=2)
```



The points in the normalized data are the same as the original one. The only thing that changes is the scale of the axis.

K-means Execution

`cluster`: A vector of integers indicating the cluster to which each point is allocated.

`centers`: A matrix of cluster centers.

`size`: The number of points in each cluster.

`betweenss`: The between-cluster sum of squares. In an optimal segmentation, one expects this ratio to be as higher as possible, since we would like to have heterogeneous clusters.

`withinss`: Vector of within-cluster sum of squares, one component per cluster. In an optimal segmentation, one expects this ratio to be as lower as possible for each cluster, since we would like to have homogeneity within the clusters.

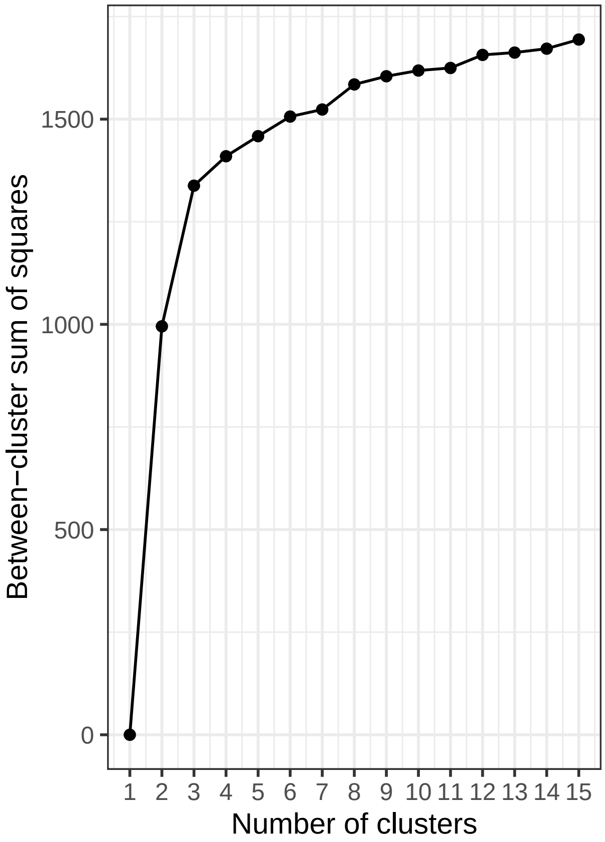
`tot.withinss`: Total within-cluster sum of squares.

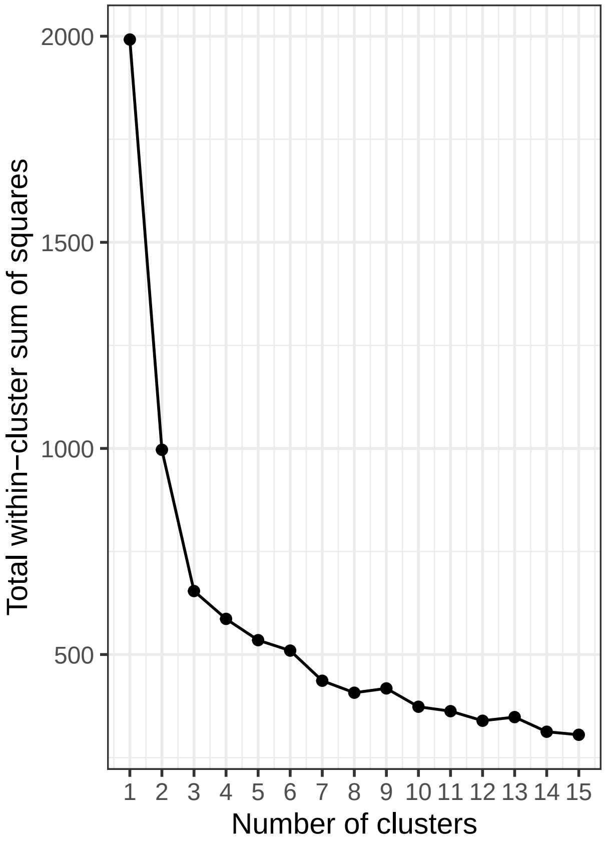
`totss`: The total sum of squares.

```
bss <- numeric() # Between-cluster sum of squares
wss <- numeric() # Total within-cluster sum of squares

# Run the algorithm for different values of k
set.seed(1234)

for(i in 1:15){
  # For each k, calculate betweenss and tot.withinss
  bss[i] <- kmeans(country_norm, centers=i)$betweenss
  wss[i] <- kmeans(country_norm, centers=i)$tot.withinss
}

# Between-cluster sum of squares vs Choice of k
k1 <- qplot(1:15, bss, geom=c("point", "line"),
            xlab="Number of clusters", ylab="Between-cluster sum of squares") +
  scale_x_continuous(breaks=seq(0, 15, 1)) +
  theme_bw()
```

```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
# Total within-cluster sum of squares vs Choice of k
k2 <- qplot(1:15, wss, geom=c("point", "line"),
            xlab="Number of clusters", ylab="Total within-cluster sum of squares") +
  scale_x_continuous(breaks=seq(0, 15, 1)) +
  theme_bw()

# Subplot
grid.arrange(k1, k2, ncol=2)
```



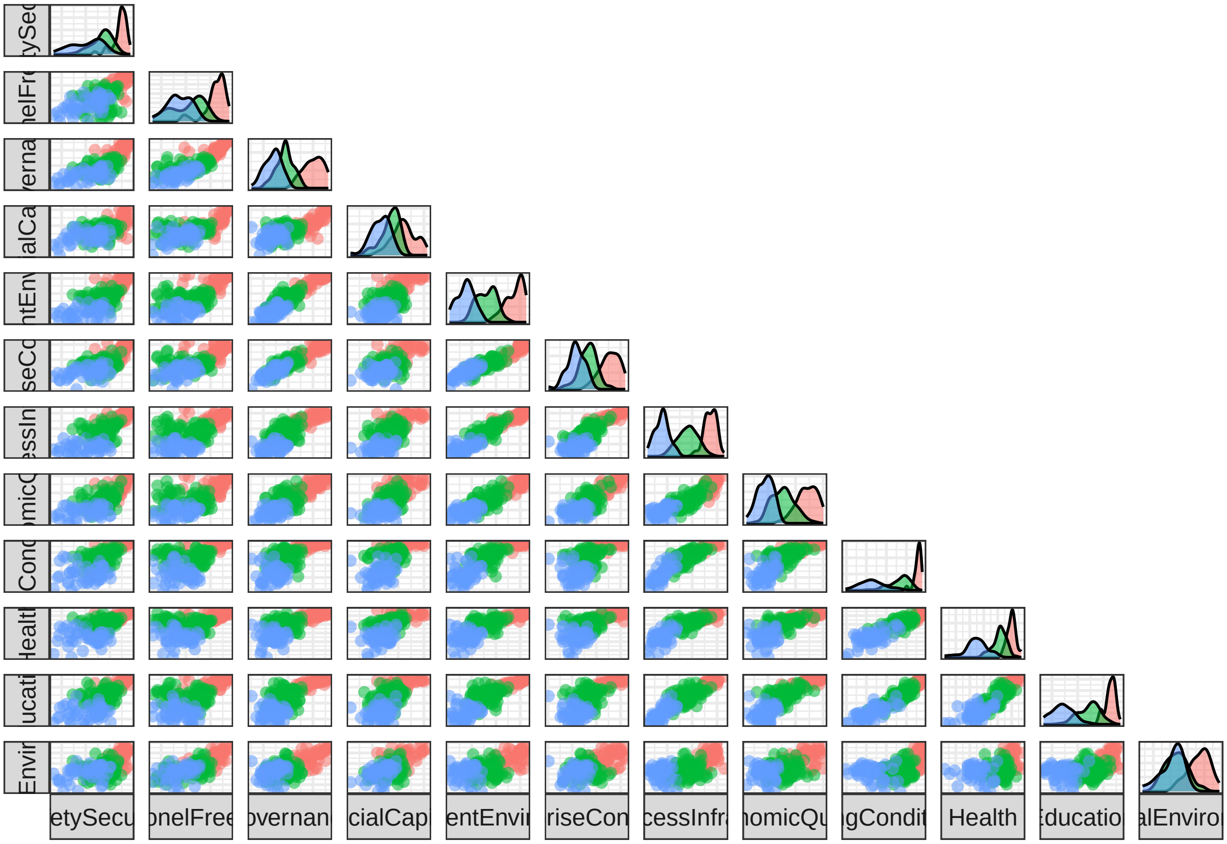


How to choose the optimal value for k:

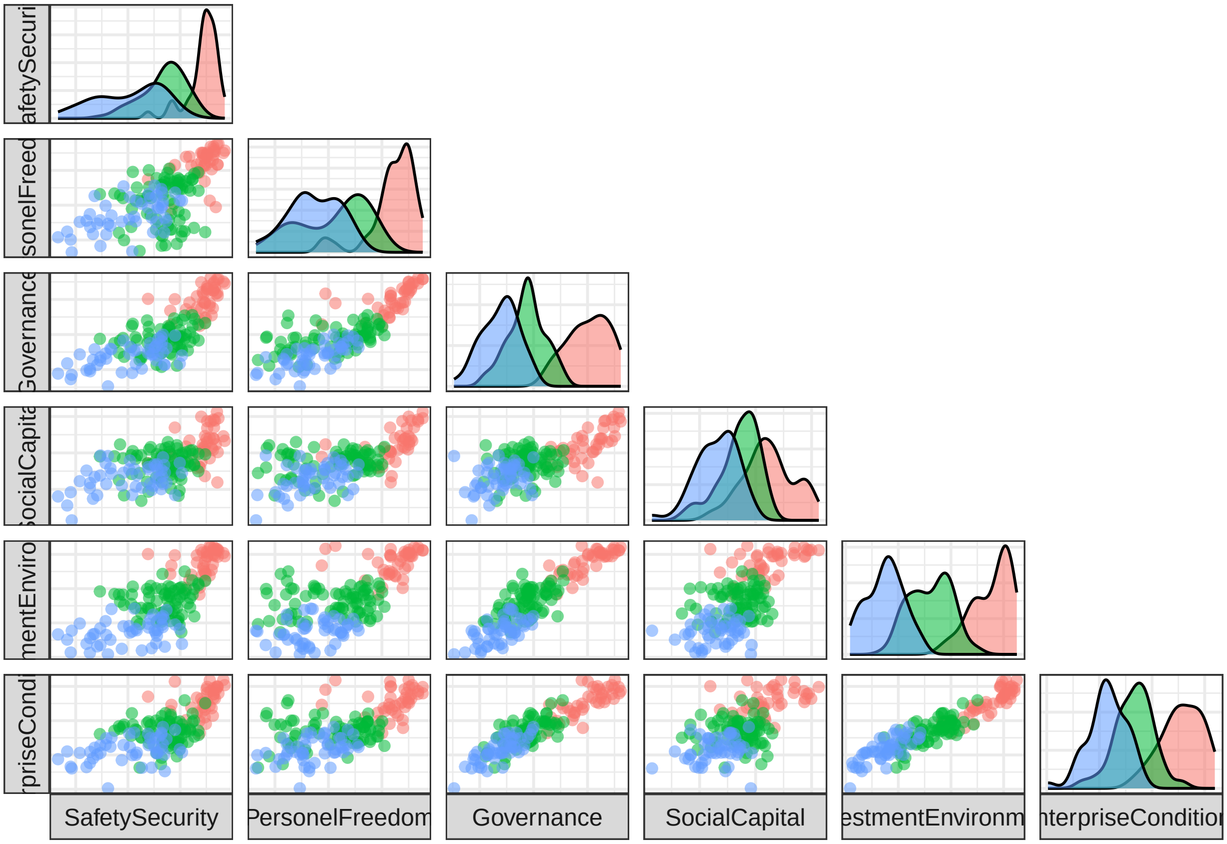
We should choose a number of clusters so that adding another cluster does not give much better partition of the data. The number of clusters, K, is chosen at the point where the gain will drop giving an angle in the graph. Therefore, our optimal value for k is 3. The total within-cluster sum of squares drops sharply at this point and then begins to level off, suggesting that adding more clusters beyond this point does not significantly improve the variance explained.

Result

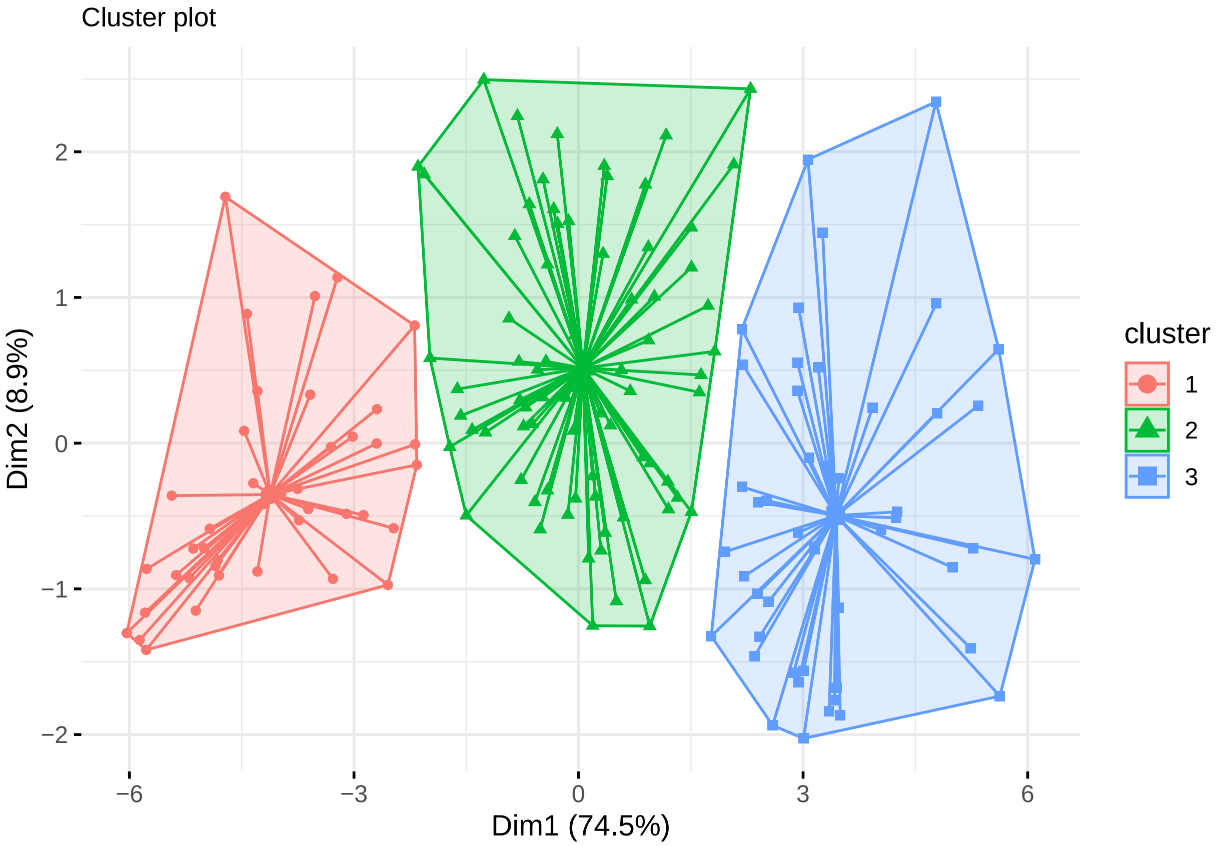
```
# Execution of k-means with k=3
set.seed(1234)

country_k3 <- kmeans(country_norm, centers=3)

# Clustering
ggpairs(cbind(country, Cluster=as.factor(country_k3$cluster)),
        columns=1:12, aes(colour=Cluster, alpha=0.5),
        lower=list(continuous="points"),
        upper=list(continuous="blank"),
        axisLabels="none", switch="both") +
        theme_bw()
```



```
# parts of the ggpairs
ggpairs(cbind(country, Cluster=as.factor(country_k3$cluster)),
        columns=1:6, aes(colour=Cluster, alpha=0.5),
        lower=list(continuous="points"),
        upper=list(continuous="blank"),
        axisLabels="none", switch="both") +
        theme_bw()
```



```
fviz_cluster(country_k3, data = country,
             ellipse.type = "convex",
             star.plot = TRUE,
             repel = TRUE,
             labelsize = 0,
             ggtheme = theme_minimal()) +
  theme(plot.title = element_text(size = 10))
```



```
# Mean values and SD of each cluster
mean <- aggregate(country, by=list(country_k3$cluster), mean)
mean
```

```
##   Group.1 SafetySecurity PersonelFreedom Governance SocialCapital
## 1       1       86.16310        80.53524   75.37071      64.74524
## 2       2       67.32882        52.48566   46.73789      54.16145
## 3       3       50.89102        43.24735   34.54122      46.13449
##   InvestmentEnvironment EnterpriseConditions MarketAccessInfrastructure
## 1              75.46429             71.36333                   74.66738
## 2              52.26105             53.22724                   56.55829
## 3              34.93755             43.01388                   35.79245
##   EconomicQuality LivingConditions   Health Education NaturalEnvironment
## 1        68.47429         91.48500 79.90167  81.83714           66.99119
## 2        50.99724         73.45066 71.43145  61.17355           53.61500
## 3        37.96449         45.56184 54.60612  35.11265           51.06796
```

```
# SD of each cluster
sd <- aggregate(country, by=list(country_k3$cluster), sd)
sd
```

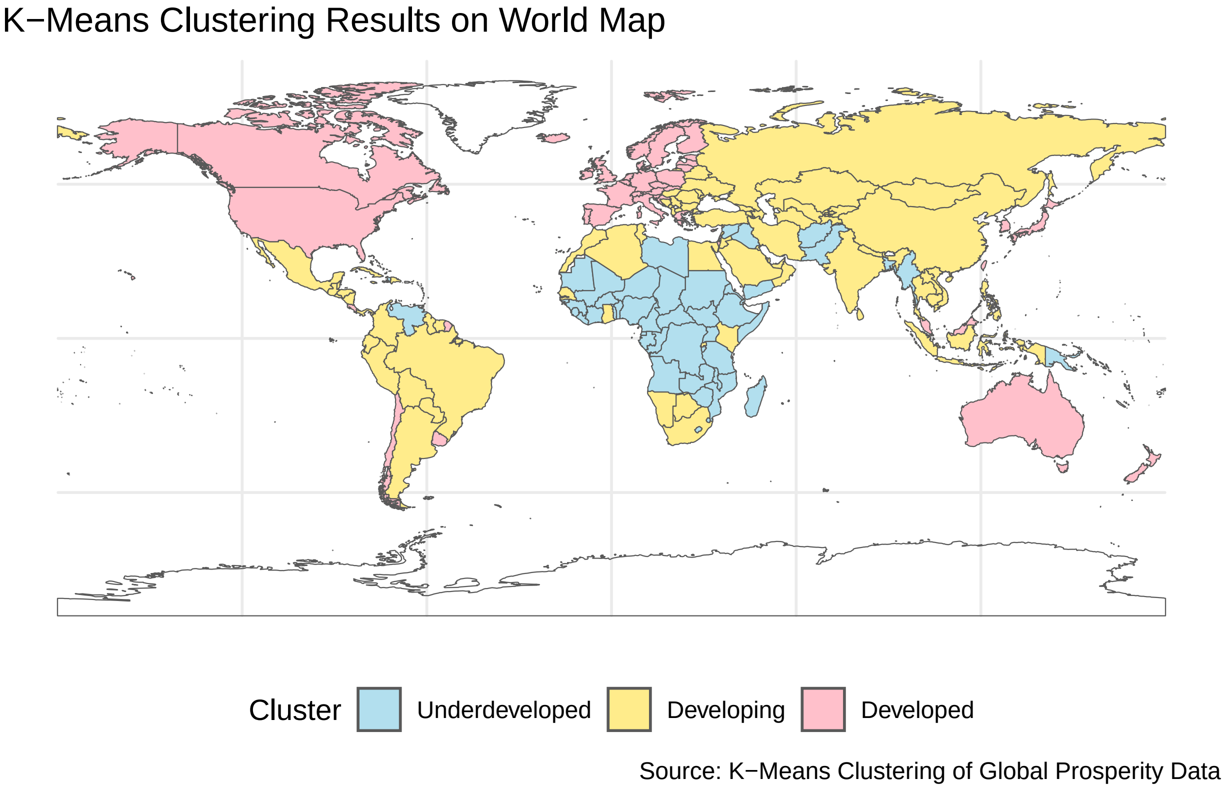
```
##   Group.1 SafetySecurity PersonelFreedom Governance SocialCapital
## 1       1       7.282008        10.79122   9.677700      9.206909
## 2       2      10.551823        16.25754   8.481448      7.121721
## 3       3      16.086650        12.21436   8.559958      7.442817
##   InvestmentEnvironment EnterpriseConditions MarketAccessInfrastructure
## 1              7.399184             7.785493                   4.906537
## 2              7.929676             7.301863                   7.894316
## 3              7.004395             7.247080                   5.643467
##   EconomicQuality LivingConditions   Health Education NaturalEnvironment
## 1        7.280882         3.423754 3.020291  4.913869           6.301603
## 2        8.238869         9.870290 4.885521 10.321538           6.466131
## 3        5.718043        12.442174 7.858516 10.303038           6.733957
```

- Cluster 1 - Developed Nations: This cluster is characterized by high average scores in most dimensions, indicating well-developed and prosperous countries. Safety and security, personal freedom, and governance scores are significantly higher, reflecting stable and free societies with effective governance structures. Economic quality, living conditions, and education are also high, suggesting a high standard of living and strong educational systems. The lower standard deviation values indicate that countries in this cluster are more homogenous, with similar high scores across various dimensions. This cluster can be named "Developed Nations," as it includes countries that are leaders in multiple areas of prosperity.
- Cluster 2 - Developing Nations: This cluster includes countries with moderate scores across most dimensions. The average scores for safety and security, personal freedom, and governance are balanced, indicating a transitional state of development. Economic quality and living conditions are relatively strong, suggesting a decent standard of living and economic environment. The standard deviation values suggest moderate variability within this cluster, with countries having consistent scores across different dimensions. This cluster can be named "Developing Nations," representing countries that are progressing towards higher levels of prosperity but still face significant challenges.
- Cluster 3 - Underdeveloped Nations: Countries in this cluster exhibit lower scores across most dimensions compared to Cluster 1. The average scores for safety and security, governance, and economic quality are notably lower, reflecting substantial challenges in these areas. Personal freedom and social capital are also lower, indicating issues related to individual liberties and societal trust. The higher standard deviation values in this cluster indicate greater variability among countries, highlighting significant differences within the group. This cluster can be referred to as "Underdeveloped Nations," as it comprises countries that are still working towards improving their overall prosperity.

Overall, the clustering analysis suggests that 3 clusters effectively capture the differences in global prosperity among countries. These clusters align with the classic literature categorization of countries into underdeveloped, developing, and developed nations. This natural separation within the data confirms that the traditional distinctions are grounded in the actual metrics of prosperity and development. The analysis reinforces the validity of these categories by showing that the differences in dimensions such as safety and security, governance, and economic quality are significant enough to group countries naturally into these three well-defined clusters.

K-means Clustering Results on World Map

```r
country_raw <- read_xlsx("Global Prosperity 2023.xlsx")
world <- ne_countries(scale = "medium", returnclass = "sf")

country_raw$Country <- str_trim(country_raw$Country)

country_raw$Cluster <- country_k3$cluster

country_raw <- country_raw %>%
  mutate(Country = case_when(
    Country == "United States" ~ "United States of America",
    Country == "Korea" ~ "South Korea",
    Country == "North Korea" ~ "Dem. Rep. Korea",
    Country == "Ivory Coast" ~ "Côte d'Ivoire",
    Country == "Czech Republic" ~ "Czechia",
    Country == "Dominican Republic" ~ "Dominican Rep.",
    Country == "Bosnia and Herzegovina" ~ "Bosnia and Herz.",
    Country == "Cape Verde" ~ "Cabo Verde",
    Country == "São Tomé and Príncipe" ~ "São Tomé and Principe",
    Country == "Equatorial Guinea" ~ "Eq. Guinea",
    Country == "Swaziland" ~ "eSwatini",
    Country == "Democratic Republic of Congo" ~ "Dem. Rep. Congo",
    Country == "Central African Republic" ~ "Central African Rep.",
    Country == "South Sudan" ~ "S. Sudan",
    Country == "Lao PDR" ~ "Laos",
    Country == "Macedonia" ~ "North Macedonia",
    Country == "Syrian Arab Republic" ~ "Syria",
    TRUE ~ Country
  ))

country_raw$Cluster <- factor(country_raw$Cluster, levels = c("3", "2", "1"), labels = c("Underdevelope

world_data <- left_join(world, country_raw, by = c("name" = "Country"))

cluster_colors <- c("Underdeveloped" = "lightblue2",
                    "Developing" = "lightgoldenrod1",
                    "Developed" = "pink")

ggplot(data = world_data) +
  geom_sf(aes(fill = Cluster)) +
  scale_fill_manual(values = cluster_colors, name = "Cluster", na.translate = FALSE) +
  theme_minimal() +
  labs(title = "K-Means Clustering Results on World Map",
       caption = "Source: K-Means Clustering of Global Prosperity Data") +
  theme(legend.position = "bottom")
```



K−Means Clustering Results on World Map

Source: K−Means Clustering of Global Prosperity Data

The map highlights the clustering results of the K-Means algorithm applied to the Global Prosperity Data:

- **Developed (pink)**: These regions include North America, Western Europe, and parts of Oceania, reflecting high prosperity and well-being.
- **Developing (Green)**: Central and South America, parts of Eastern Europe, and much of Asia fall into this category, showing moderate levels of prosperity.
- **Underdeveloped (blue)**: Most of Africa and parts of South Asia are classified here, indicating lower levels of prosperity and development.