HDI in Indonesia: Revealing Gaps and Equality Between Regions in Indonesia

Stephanie Nadya - 2702264576

Matthew Sebastian Junus - 2602104125

Bryan Ferdinand Teddy Fiersdy - 2602075296

Darren Nathaneil - 2602081216

22 June 2023

1.Introduction

This report discusses the HDI in Indonesia. HDI is an important indicator to assess the development of a country by examining the level of human development in a region. HDI encompasses health, income, education, and other factors within a region. The purpose of selecting this data is to identify and analyze the HDI across various regions in Indonesia and determine whether it is evenly distributed.

We chose this topic because HDI provides insights into the quality of life of residents in different regions of Indonesia. By understanding the HDI of a region, we and others can identify areas needing attention to improve the well-being of the local population.

This report is intended for researchers and other interested parties in findings related to the HDI in Indonesia. The target audience includes policymakers, academics, and organizations involved in human development and social welfare.

The objectives of this analysis are to answer questions such as:

- What is the state of HDI in each region?
- What are the key factors influencing HDI in each region?

Metode yang digunakan untuk analisis:

- EDA (Exploratory Data Analysis)
- Statistical methods such as analyzing correlations between variables
- Data visualization to provide an overview of the HDI data

This analysis is significant because it allows others, including students like us, to understand the HDI conditions in different regions. While we may be well-off in our area, how are conditions in other regions? Is development evenly distributed or not? The main goal is to provide a clear picture of whether the HDI in a region meets the expected standards.

2.Data Description

The dataset used by our group is the “IPM INDONESIA 2021” from Kaggle. The dataset consists of 519 rows and includes 12 variables. Dataset link : https://www.kaggle.com/datasets/fhadhai/dataipmindonesia/data

The variables in the dataset are as follows:

1. Provinsi
   - Description : Name of the province
   - Type : Character
   - Example : Lampung, West Java, Central Java, etc. Tengah, dll.
2. Kab/Kota
   - Description : Name of the district/city within the province
   - Type : Character
   - Example : Bogor City, Tasikmalaya, Subang, etc.
3. Persentase Penduduk Miskin (P0) Menurut Kabupaten/Kota (Persen)
   - Description : Percentage of the population living below the poverty line by district/city, in percent
   - Type : Numeric
   - Example : 11.67, 13.66, 16.24, etc.
4. Rata-rata Lama Sekolah Penduduk 15+ (Tahun)
   - Description : Average years of schooling for the population aged over 15 years, in years
   - Type : Numeric
   - Example : 8.64, 8.71, 9.34, etc.
5. Pengeluaran per Kapita Disesuaikan (Ribu Rupiah/Orang/Tahun)
   - Description : Adjusted expenditure per capita in thousand Rupiah per person per year
   - Type : Integer
   - Example : 7148, 8776, 8030, etc.
6. Indeks Pembangunan Manusia
   - Description : Combined index measuring human development
   - Type : Numeric
   - Example : 69.22, 68.74, 71.46, etc.
7. Umur Harapan Hidup (Tahun)
   - Description : Life expectancy at birth, in years
   - Type : Numeric
   - Example : 69.26, 70.56, 65.53, etc.
8. Persentase rumah tangga yang memiliki akses terhadap sanitasi layak
   - Description : Percentage of households with access to adequate sanitation facilities
   - Type : Numeric
   - Example: 66.75, 90.58, 74.30, etc.
9. Persentase rumah tangga yang memiliki akses terhadap air minum layak
   - Description : Percentage of households with access to safe drinking water
   - Type : NumeriC
   - Example : 83.16, 89.24, 91.09, etc.
10. Tingkat Pengangguran Terbuka
    - Description: Percentage of the labor force that is unemployed and actively seeking work
    - Type : Numeric
    - Example : 5.71, 8.36, 6.46, etc.
11. Tingkat Partisipasi Angkatan Kerja
    - Description : Percentage of the working-age population that is working or actively seeking work
    - Type : Numeric
    - Example : 71.15, 62.85, 60.05, etc..
12. PDRB atas Dasar Harga Konstan menurut Pengeluaran (Rupiah)
    - Description : Gross Domestic Product at constant prices by expenditure, in Rupiah
    - Type : Numeric
    - Example : 1780419, 16924103, 1152875, etc.

3. Data Preprocessing

In this section, we performed preprocessing on the dataset. The preprocessing steps included:

- Changing the data types of variables
- Renaming variables
- Removing missing values
- Removing duplicate data

```r
# import library
library(dplyr)
library(skimr)
library(ggplot2)
library(pander)
library(plotly)
library(corrplot)
```

Above is the library that will be used for the data preprocessing, EDA and visualization stages.

1. dplyr
   - A library used for data manipulation.
2. skimr
   - A library that provides statistical summaries.
3. ggplot2
   - A library for data visualization.
4. pander
   - A library for creating easy-to-read summaries.
5. plotly
   - A library for creating interactive graphs.
6. corrplot
   - A library for creating a better correlation plot.

```r
df <- read.csv("ipm-indonesia2021-cluster.csv")
str(df)
```

```
## 'data.frame':    519 obs. of  12 variables:
##  $ Provinsi                                                          : chr  "ACEH" "ACEH" "ACEH" "ACEH" ...
##  $ Kab.Kota                                                          : chr  "Simeulue" "Aceh Singkil" "Aceh Selatan" "Aceh Tenggara" ...
##  $ Persentase.Penduduk.Miskin..P0..Menurut.Kabupaten.Kota..Persen.     : num  19 20.4 13.2 13.4 14.4 ...
##  $ Rata.rata.Lama.Sekolah.Penduduk.15...Tahun.                       : num  9.48 8.68 8.88 9.67 8.21 ...
##  $ Pengeluaran.per.Kapita.Disesuaikan..Ribu.Rupiah.Orang.Tahun.        : int  7148 8776 8180 8030 8577 10780 9593 9644 9860 8867 ...
##  $ Indeks.Pembangunan.Manusia                                        : num  66.4 69.2 67.4 69.4 67.8 ...
##  $ Umur.Harapan.Hidup..Tahun.                                        : num  65.3 67.4 64.4 68.2 68.7 ...
##  $ Persentase.rumah.tangga.yang.memiliki.akses.terhadap.sanitasi.layak : num  71.6 69.6 62.5 62.7 66.8 ...
##  $ Persentase.rumah.tangga.yang.memiliki.akses.terhadap.air.minum.layak: num  87.5 78.6 79.7 86.7 83.2 ...
##  $ Tingkat.Pengangguran.Terbuka                                       : num  5.71 8.36 6.46 6.43 7.13 2.61 7.09 7.7 7.28 4.32 ...
##  $ Tingkat.Partisipasi.Angkatan.Kerja                                : num  71.2 62.9 60.9 69.6 59.5 ...
##  $ PDRB.atas.Dasar.Harga.Konstan.menurut.Pengeluaran..Rupiah.           : int  1648096 1780419 4345784 3487157 8433526 5953118 7485861 10261585 7975099 10374480 ...
```

Variable names will be renamed to make them easier to read.

```r
names(df) <- c("provinsi", "kab_kota", "persentase_penduduk_miskin", "lama_sekolah", "pengeluaran", "ipm", "umur_harapan_hidup", "persentase_akses_sanitasi", "persentase_akses_air", "tingkat_pengangguran", "tingkat_partisipasi_kerja", "PDRB")
str(df)
```

```
## 'data.frame':    519 obs. of  12 variables:
##  $ provinsi                  : chr  "ACEH" "ACEH" "ACEH" "ACEH" ...
##  $ kab_kota                  : chr  "Simeulue" "Aceh Singkil" "Aceh Selatan" "Aceh Tenggara" ...
##  $ persentase_penduduk_miskin: num  19 20.4 13.2 13.4 14.4 ...
##  $ lama_sekolah              : num  9.48 8.68 8.88 9.67 8.21 ...
##  $ pengeluaran               : int  7148 8776 8180 8030 8577 10780 9593 9644 9860 8867 ...
##  $ ipm                       : num  66.4 69.2 67.4 69.4 67.8 ...
##  $ umur_harapan_hidup        : num  65.3 67.4 64.4 68.2 68.7 ...
##  $ persentase_akses_sanitasi : num  71.6 69.6 62.5 62.7 66.8 ...
##  $ persentase_akses_air      : num  87.5 78.6 79.7 86.7 83.2 ...
##  $ tingkat_pengangguran      : num  5.71 8.36 6.46 6.43 7.13 2.61 7.09 7.7 7.28 4.32 ...
##  $ tingkat_partisipasi_kerja : num  71.2 62.9 60.9 69.6 59.5 ...
##  $ PDRB                       : int  1648096 1780419 4345784 3487157 8433526 5953118 7485861 10261585 7975099 10374480 ...
```

There are 2 categorical variables that have inappropriate data types (province and district). We change these two data types into factors.

```r
chr_to_factor <- c("provinsi", "kab_kota")

for (col in chr_to_factor) {
  df[[col]] <- as.factor(df[[col]])
}
```

Next, we checked for anomalies in the data such as missing values. duplicate data and delete the data.

```r
# Checking Missing Value
colSums(is.na(df))
```

```
##                   provinsi                   kab_kota
##                          0                          0
## persentase_penduduk_miskin               lama_sekolah
##                          5                          5
##                pengeluaran                        ipm
##                          5                          5
##         umur_harapan_hidup  persentase_akses_sanitasi
##                          5                          5
##       persentase_akses_air       tingkat_pengangguran
##                          5                          5
##  tingkat_partisipasi_kerja                       PDRB
##                          5                          5
```

```r
# Delete missing value
df <- na.omit(df)

# Delete duplicate data
df <- df[!duplicated(df), ]

# Check whether the data still contains missing values or not
colSums(is.na(df))
```

```
##                   provinsi                   kab_kota
##                          0                          0
## persentase_penduduk_miskin               lama_sekolah
##                          0                          0
##                pengeluaran                        ipm
##                          0                          0
##         umur_harapan_hidup  persentase_akses_sanitasi
##                          0                          0
##       persentase_akses_air       tingkat_pengangguran
##                          0                          0
##  tingkat_partisipasi_kerja                       PDRB
##                          0                          0
```

4. Data Exploration

In this section, we carry out exploratory data analysis and display data visualization.

```r
# Summary of numeric variables
select_df <- df %>% select(-provinsi, -kab_kota)
summary_df <- summary(select_df)
pander(summary_df)
```

Table continues below

persentase_penduduk_miskin	lama_sekolah	pengeluaran	ipm
Min. : 2.38	Min. : 1.420	Min. : 3976	Min. :32.84
1st Qu.: 7.15	1st Qu.: 7.510	1st Qu.: 8574	1st Qu.:66.64
Median :10.46	Median : 8.305	Median :10196	Median :69.61
Mean :12.27	Mean : 8.437	Mean :10325	Mean :69.93
3rd Qu.:14.89	3rd Qu.: 9.338	3rd Qu.:11719	3rd Qu.:73.11
Max. :41.66	Max. :12.830	Max. :23888	Max. :87.18

Table continues below

umur_harapan_hidup	persentase_akses_sanitasi	persentase_akses_air
Min. :55.43	Min. : 0.00	Min. : 0.00
1st Qu.:67.39	1st Qu.:70.22	1st Qu.: 79.04
Median :69.97	Median :81.80	Median : 89.80
Mean :69.66	Mean :77.20	Mean : 85.14
3rd Qu.:72.04	3rd Qu.:89.88	3rd Qu.: 96.40
Max. :77.73	Max. :99.97	Max. :100.00

tingkat_pengangguran	tingkat_partisipasi_kerja	PDRB
Min. : 0.000	Min. :56.39	Min. : 147485
1st Qu.: 3.180	1st Qu.:65.07	1st Qu.: 3654292
Median : 4.565	Median :68.95	Median : 8814926
Mean : 5.059	Mean :69.46	Mean : 21964077
3rd Qu.: 6.530	3rd Qu.:72.34	3rd Qu.: 19735101
Max. :13.370	Max. :97.93	Max. :460081046

The table above displays a summary of the statistical measures for each numeric column. The table provides information on the minimum value, first quartile, median, mean, third quartile, and maximum value.

Here, we create a visualization of the 10 provinces with the highest average HDI

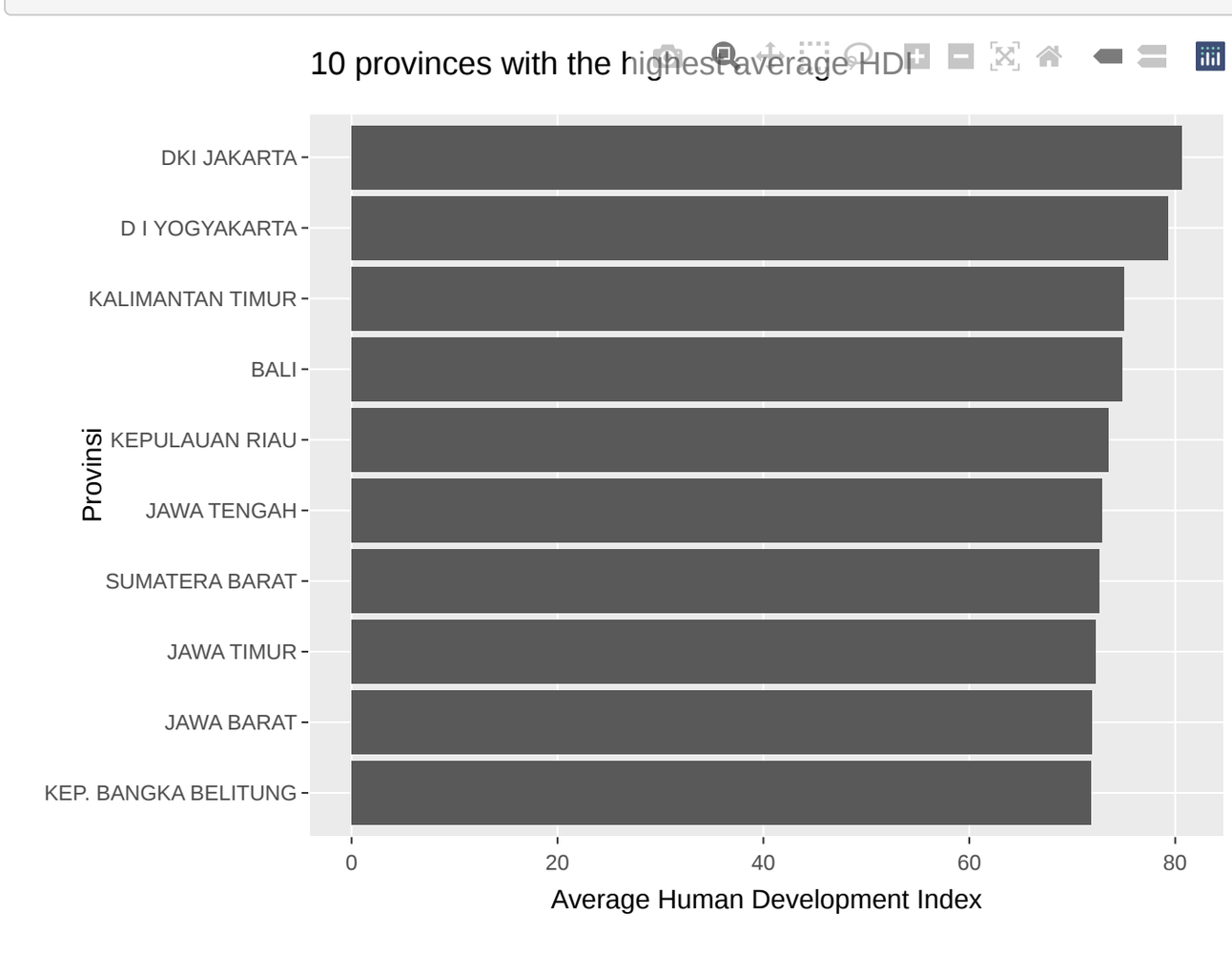
```r
# Calculate the average HDI per province
provinsi_ipm_tertinggi <- aggregate(df$ipm, by = list(df$provinsi), FUN=mean)

# Rename columns
colnames(provinsi_ipm_tertinggi)[1] ="provinsi"
colnames(provinsi_ipm_tertinggi)[2] ="ratarata"

# Make a table with the 10 provinces with the highest average HDI
top10_ipm_tertinggi <- provinsi_ipm_tertinggi[order(-provinsi_ipm_tertinggi$ratarata), ][1:10, ]

# Plotting graph bar
p <- ggplot(top10_ipm_tertinggi, aes(x = reorder(provinsi, ratarata), y = ratarata)) +
  geom_bar(stat = "identity") +
   labs(title = "10 provinces with the highest average HDI", x = "Provinsi", y = "Average Human Development Index") +
  coord_flip()

ggplotly(p)
```



From the graph above, we can see the 10 provinces with the highest average HDI. Most of the provinces shown are provinces on the island of Java, showing the development gap between regions in Indonesia.

Here, we create a visualization of the 10 provinces with the lowest HDI average to compare with the highest.

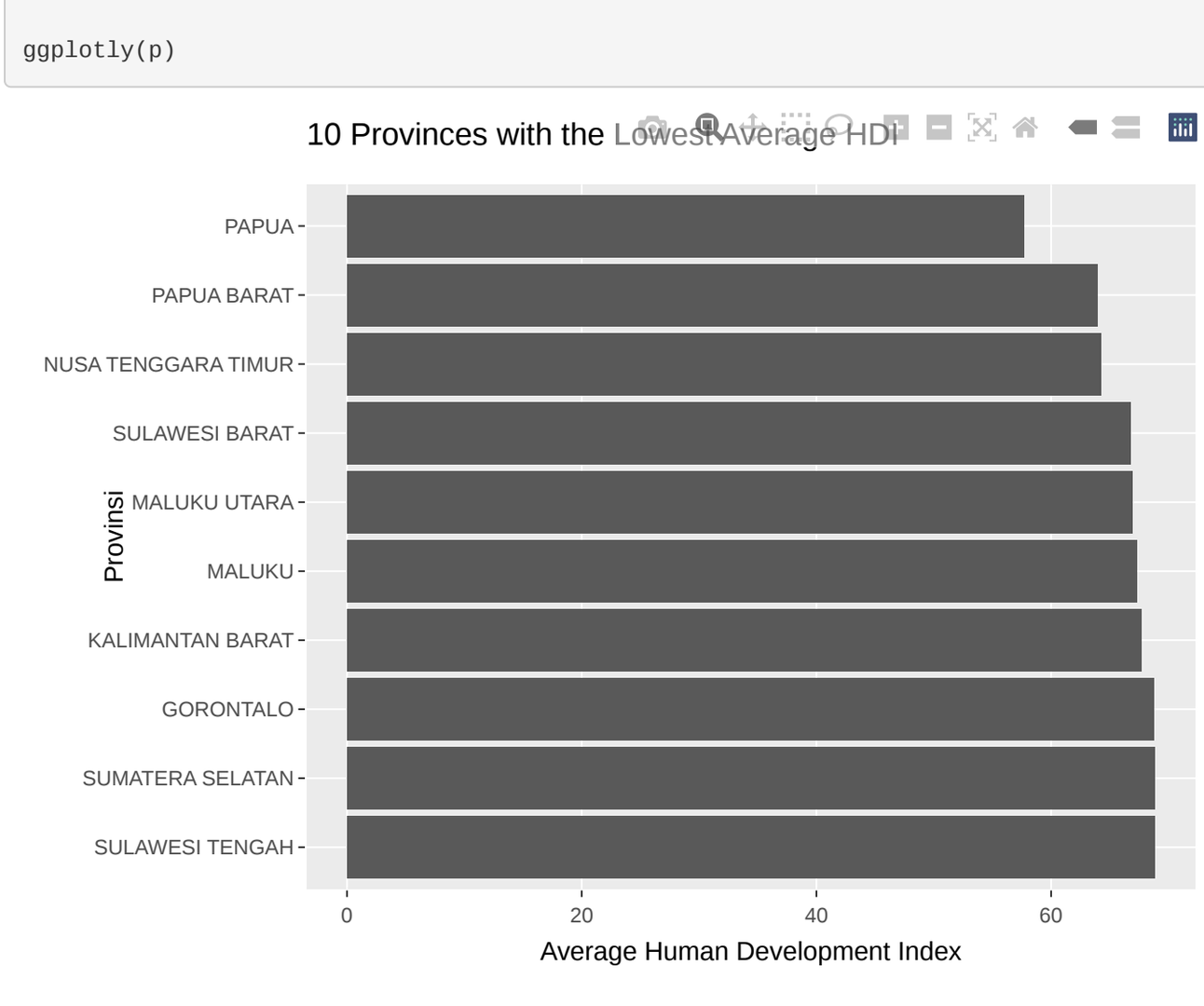
```r
# Calculate the average HDI per province
provinsi_ipm_terendah <- aggregate(df$ipm, by = list(df$provinsi), FUN=mean)

# Rename Columns
colnames(provinsi_ipm_terendah)[1] ="provinsi"
colnames(provinsi_ipm_terendah)[2] ="ratarata"

# Make a table with 10 provinces with the lowest average HDI
top10_ipm_tertinggi <- provinsi_ipm_tertinggi[order(provinsi_ipm_tertinggi$ratarata), ][1:10, ]

# Plotting graph bar
p <- ggplot(top10_ipm_tertinggi, aes(x = reorder(provinsi, -ratarata), y = ratarata)) +
  geom_bar(stat = "identity") +
  labs(title = "10 Provinces with the Lowest Average HDI", x = "Provinsi", y = "Average Human Development Index")
+
  coord_flip()

ggplotly(p)
```

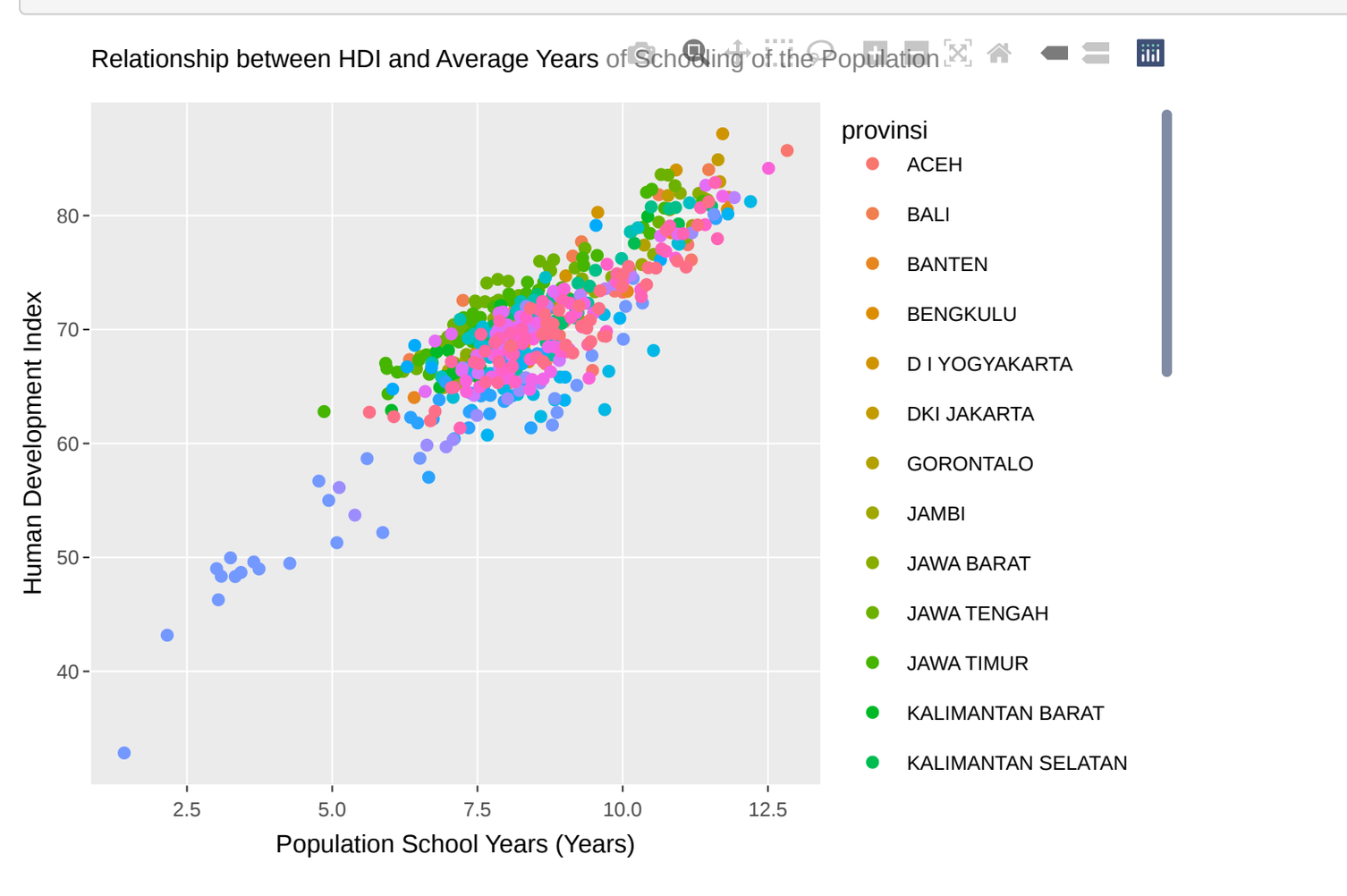


The graph above shows the 10 provinces in Indonesia that have the lowest average Human Development Index. The graph also shows a significant gap between provinces with the highest and lowest HDI.

Scatterplot visualization of the relationship between HDI and average years of schooling

```r
# plotting scatterplot
p <- ggplot(df, aes(x = lama_sekolah, y = ipm, color=provinsi)) +
  geom_point() +
  labs(title = "Relationship between HDI and Average Years of Schooling of the Population",
       x = "Population School Years (Years)",
       y = "Human Development Index") +
  theme(plot.title = element_text(size=11))

ggplotly(p)
```



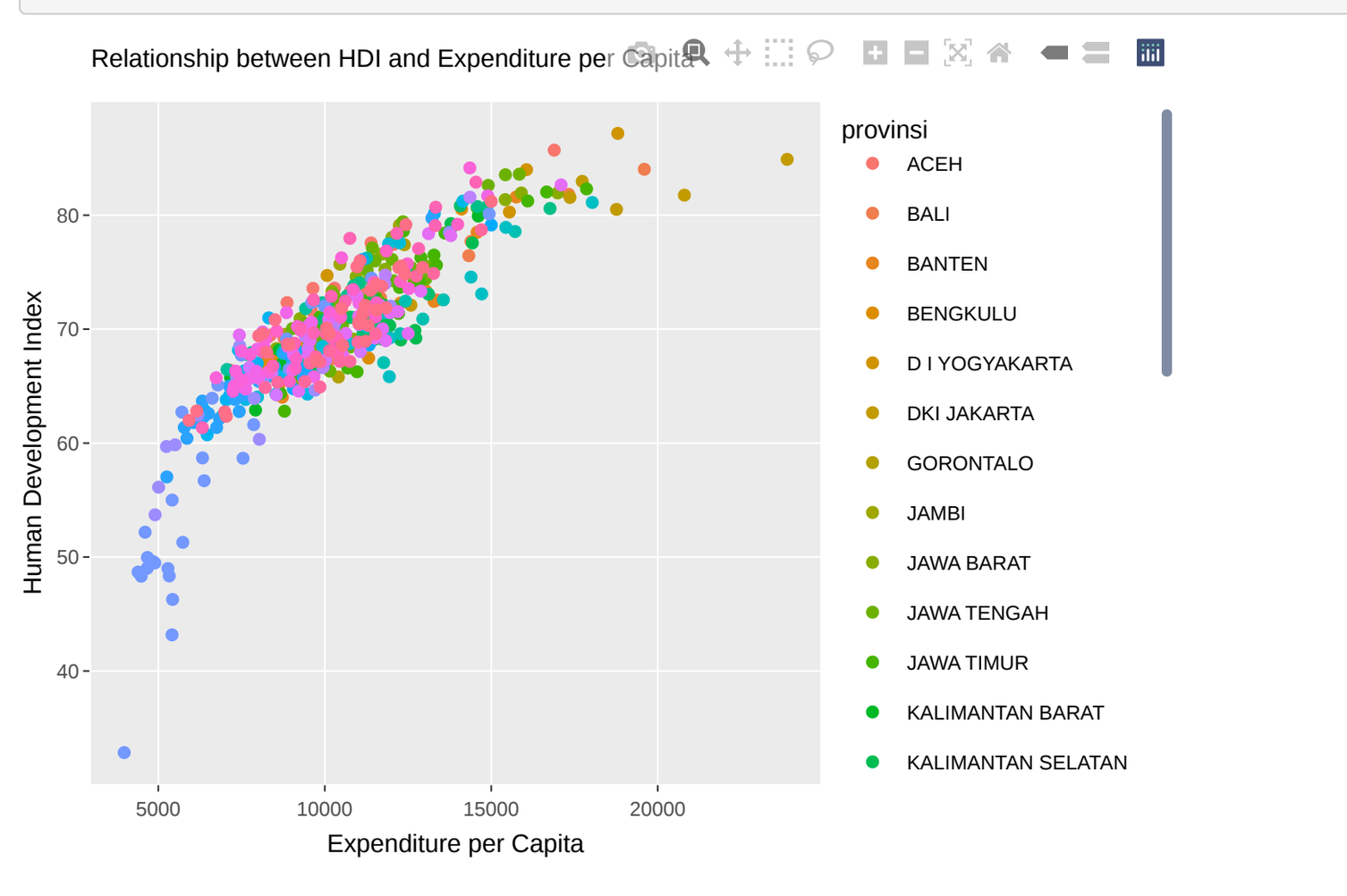
From this graph we can see that there is a clear positive correlation between the population's years of schooling and the human development index. When the average length of schooling increases, the HDI also increases. This suggests that higher levels of education are associated with better human development outcomes.

The graph shows that education plays an important role in improving human development.

Scatterplot visualization of the relationship between per capita expenditure and HDI

```r
# plotting scatterplot
p <- ggplot(df, aes(x = pengeluaran, y = ipm, color=provinsi)) +
  geom_point() +
  labs(title = "Relationship between HDI and Expenditure per Capita",
       x = "Expenditure per Capita",
       y = "Human Development Index") +
  theme(plot.title = element_text(size=11))

ggplotly(p)
```

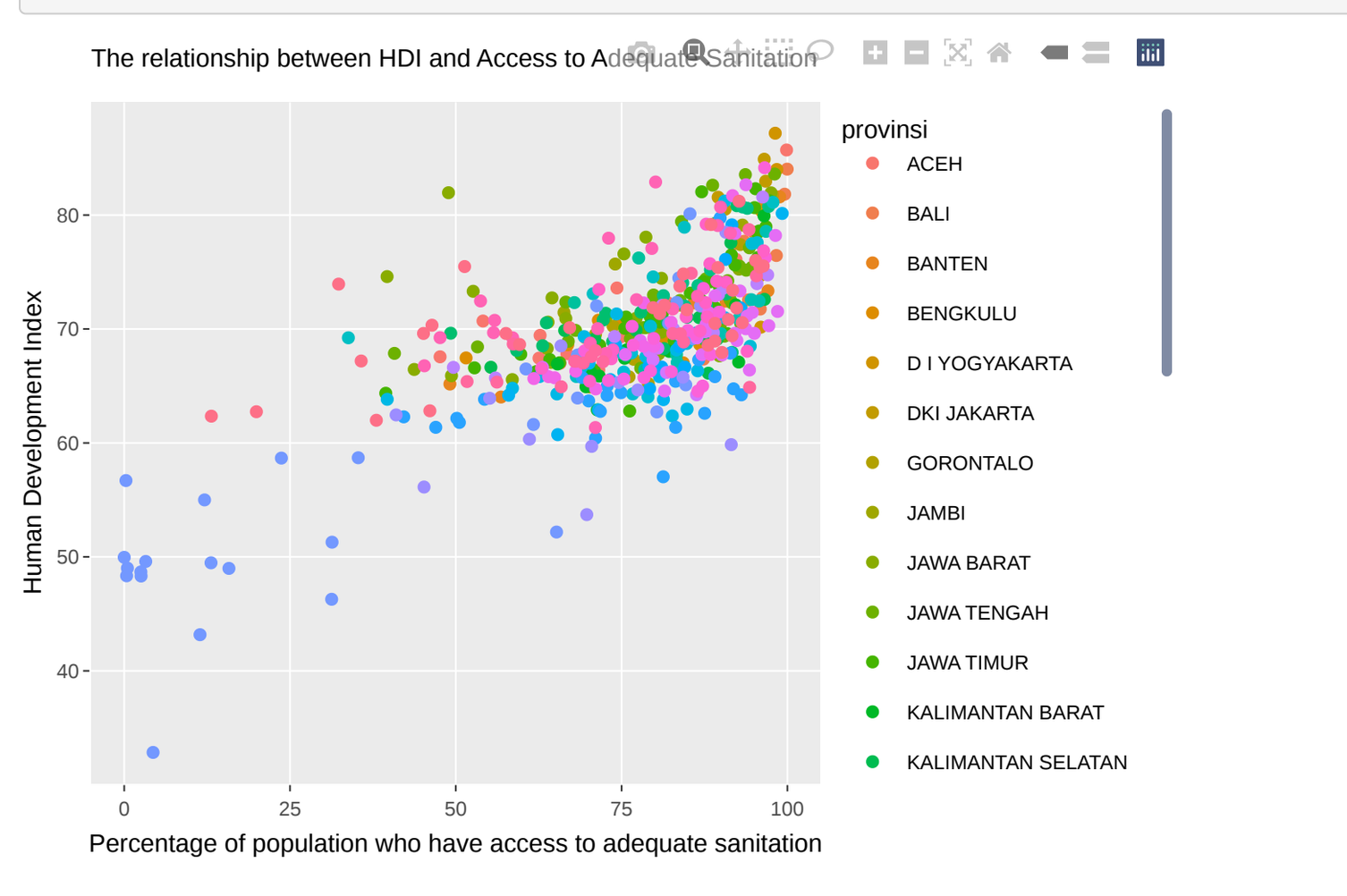


The graph shows the relationship between HDI and per capita expenditure. It can be seen that there is a positive relationship between HDI and expenditure, the higher the expenditure, the higher the HDI. This shows that an increase in per capita expenditure is correlated with an increase in quality of life as measured by the HDI.

Scatterplot visualization of the relationship between HDI and access to adequate sanitation

```r
# plotting scatterplot
p <- ggplot(df, aes(x = persentase_akses_sanitasi, y = ipm, color=provinsi)) +
  geom_point() +
  labs(title = "The relationship between HDI and Access to Adequate Sanitation",
       x = "Percentage of population who have access to adequate sanitation",
       y = "Human Development Index") +
  theme(plot.title = element_text(size=11))

ggplotly(p)
```



The graph shows the relationship between HDI and the percentage of the population who have access to adequate sanitation. There is a positive relationship between adequate sanitation and HDI. Through this graph, we can see that access to adequate sanitation contributes positively to increasing HDI.

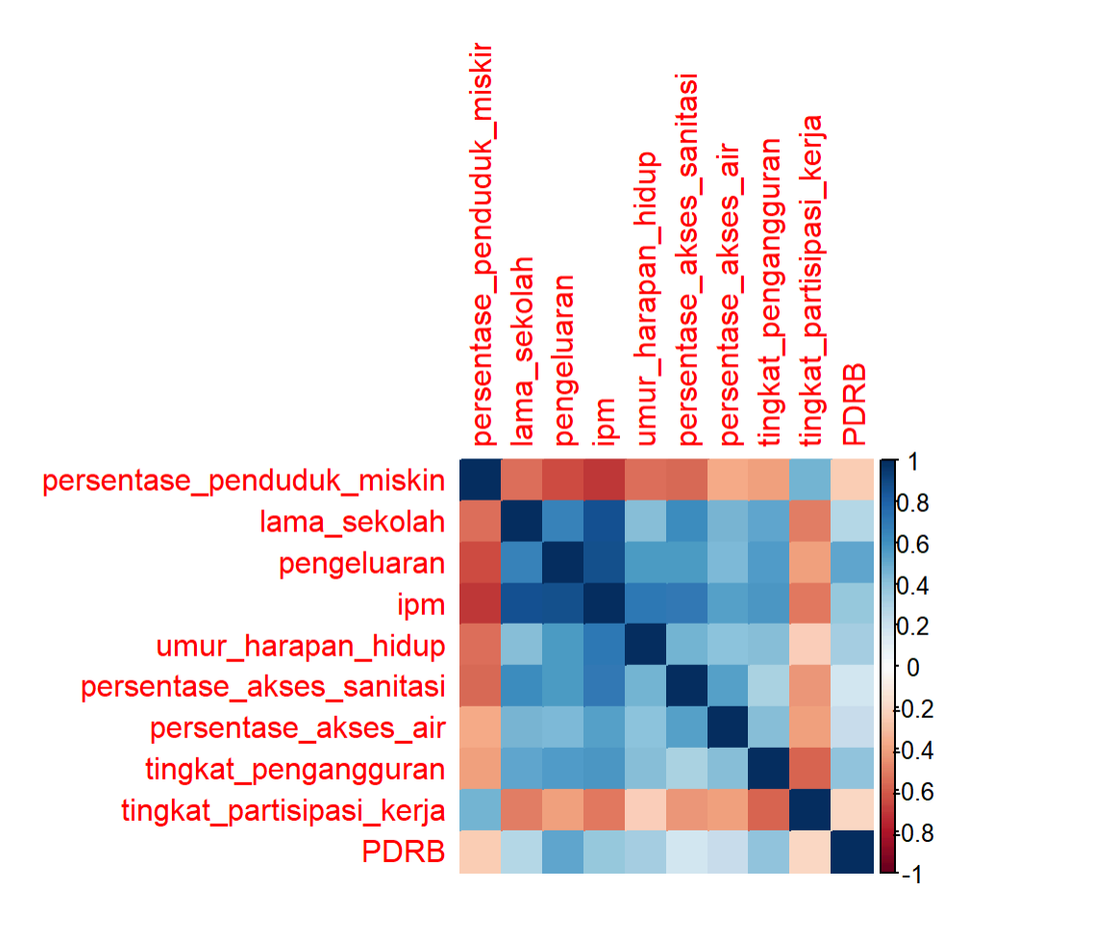
5. Statistical Analysis

In this section, a statistical analysis will be conducted in the form of correlation analysis of all numerical variables.

```r
numerical_variables <- df %>% select_if(is.numeric)

matrix = cor(numerical_variables)

corrplot(matrix, method = "color", mar=c(1,1,1,1))
```



The percentage of poverty exhibits a strong negative correlation with HDI, life expectancy, and access to clean water. This suggests that as poverty decreases, these indicators tend to improve, highlighting an inverse relationship. Similarly, HDI is strongly positively correlated with life expectancy, years of schooling, and expenditure, indicating that regions with higher human development index also tend to have higher life expectancy, more years of schooling, and higher expenditure level.

Additionally, unemployment rate shows a negative correlation with labor participation rate, suggesting that as unemployment decreases, labor participation increases. The PDRB (Gross Regional Domestic Product) has moderate to strong positive correlations with expenditure and years of schooling, indicating that regions with higher economic output also have higher expenditure and education levels.

6. Discussion

Based on the analysis, it is evident that provinces with the highest HDI such as DKI Jakarta, D.I. Yogyakarta, and East Kalimantan have good access to quality education, adequate healthcare services, and developing infrastructure. Conversely, provinces with the lowest HDI like Papua, West Papua, and East Nusa Tenggara face challenges such as geographic constraints, limited resources, and high poverty rates that hinder their progress.

The factors significantly influencing HDI in each region include access to education, healthcare services, infrastructure, economic conditions, and government policies. Education and healthcare emerge as pivotal factors, where regions with robust educational facilities and healthcare services tend to exhibit higher HDI. Additionally, adequate infrastructure and strong economic conditions also contribute significantly to HDI improvement. On the other hand, remote and hard-to-reach areas generally have lower HDI due to limited access to basic services and economic opportunities.

7. Conclusion

Based on the analysis conducted :

- Provinces with the highest HDI such as DKI Jakarta, D.I. Yogyakarta, East Kalimantan, and others have demonstrated good access to quality education, adequate healthcare services, and developed infrastructure. However, some provinces like Papua, West Papua, East Nusa Tenggara, and others still lack sufficient access.
- Key factors influencing HDI in each region include access to education, healthcare services, infrastructure, economic conditions, and government policies. Education and healthcare are pivotal, where regions with strong educational and healthcare facilities tend to have higher HDI. Additionally, adequate infrastructure and a robust economy significantly contribute to HDI improvement. Conversely, remote and hard-to-reach areas generally have lower HDI due to limited access to basic services and economic opportunities.

To enhance HDI across Indonesia, it is recommended that the government and stakeholders prioritize investment in education and healthcare, especially in provinces with low HDI. Improving infrastructure in remote areas is also crucial to ensure better access to basic services. Furthermore, policies focusing on inclusive and equitable human development are needed to address regional disparities. With a deeper understanding of HDI and its influencing factors, development efforts can be more effective in achieving more equitable and sustainable progress across Indonesia.