Indonesia’s HDI Landscape: Exploring Regional Disparities and it’s Influencing Factors

Group 5

2025-06-13

Background

A country’s development is not only measured from economic growth, but also from the quality of life of it’s residents. Human Development Index (HDI) is one of the key indicator that measure overall public welfare. Developed by United Nations in 1990, HDI designed to highlight that humans and their capabilities should be the main criteria in evaluate a country’s development besides just from economic growth. HDI consists of 3 basic dimension of human development including standard of living, health and longevity, and education.

Since it was introduced, HDI has become an important tool for government and international organization to evaluating development progress and identify which area that need more attention. In Indonesia, although HDI shows consistent increasing trend from year to year in aggregate, this progress has not occurred evenly across all regions. Data shows that there is in fact a significant regional disparities in achieving HDI between regions. As example in 2023, “Papua Pegunungan” Province recorded as having lowest HDI in Indonesia at approximately 52.45. This highlights a striking gap compared to the provinces in the western part of the country which usually have much more higher HDI.

These disparities not only reflecting the differences in access to basic services like education and healthcare, but also indicating the presences of complex factors like social, economy and geographic which contributes to those disparities. Understanding the root cause of these inequality problems becoming crucial and important for formulating more targeted and inclusive development policies. Therefore, this research will study deeper into Indonesia’s HDI landscape with focus in identifying the main contributing factors and regional disparities pattern.

Objectives

Based on the background that has been described, this research has some main objectives:

1. Identifying key variables that influences HDI in Indonesia: Analyze HDI data to determine which factors that have the most significant influence on HDI level across regions in Indonesia.
2. Analyzing regional HDI disparities in Indonesia: Explore patterns and disparities in HDI level across various regions, and identify areas with lowest HDI level that need more attention in Indonesia
3. Providing policy recommendation for reducing HDI disparities: Formulating more targeted and inclusive development policies for government and any other stakeholders to reduce human development gaps between regions in Indonesia.

Data Collection

In this research, the data used for analyzing HDI and its factors in Indonesia collected from 2 main source:

1. Poverty Levels Classification Dataset in Indonesia

   This dataset was obtained and downloaded from Kaggle platform with the title of "Klasifikasi Tingkat Kemiskinan di Indonesia.csv". Although the dataset title refers to poverty classification, this dataset is also highly relevant for HDI analysis as it contains per city-level data including variables that have strong correlation with HDI such as per capita expenditure, life expectancy, average years of schooling, etc. This dataset will serve as main base for statistical analysis to identifying main factors that influence HDI and seeing trends or pattern in HDI disparities between regions in Indonesia. (Source: https://www.kaggle.com/datasets/ermila/klasifikasi-tingkat-kemiskinan-di-indonesia)

2. Geographic Spatial Data of Indonesia

   This spatial data was obtained and downloaded from the official website of Global Administrative Areas. It is provided in the shapefile format that represent the administrative boundaries of Indonesia's provinces. This data is important for visualization especially in creating thematic maps that illustrate the HDI disparities between provinces in Indonesia. By using this spatial data, we can map the results of the analysis geographically so that regional disparities can be clearly visualized and better understood. (Source: https://gadm.org/download_country.html)

By combining these two types of data, tabular data containing HDI predictor variables and spatial data of Indonesia, this research aims to provide a comprehensive and insightful analysis of the HDI in Indonesia both in terms of its key factors and the regional disparities across the country.

Data Exploration and Preparation

Before building any analytical model, it is essential to understand the dataset thoroughly. This section focuses on exploring the structure and contents of the data, identifying any inconsistencies or missing values, and understanding the distribution of key variables. Exploratory Data Analysis (EDA) is performed to uncover patterns, detect anomalies, and gain initial insights. The section also includes basic preprocessing steps needed to ensure the data is ready for further analysis.

Data Content and Structure Inspection

Dimensions

```
The dataset has 999 rows and 13 columns.
```

Each row represents one data observation for each city/regency covered in the dataset, meanwhile each column represents a different variable in the poverty classification. Understanding these dimensions is the initial foundation for deeper data exploration.

Column Names

The next step is to identify the names of the columns.

```
Provinsi, Kab.Kota, Persentase.Penduduk.Miskin..P0..Menurut.Kabupaten.Kota..P
ersen., Rata.rata.Lama.Sekolah.Penduduk.15...Tahun., Pengeluaran.per.Kapita.D
isesuaikan..Ribu.Rupiah.Orang.Tahun., Indeks.Pembangunan.Manusia, Umur.Harapa
n.Hidup..Tahun., Persentase.rumah.tangga.yang.memiliki.akses.terhadap.sanitas
i.layak, Persentase.rumah.tangga.yang.memiliki.akses.terhadap.air.minum.layak
, Tingkat.Pengangguran.Terbuka, Tingkat.Partisipasi.Angkatan.Kerja, PDRB.atas
.Dasar.Harga.Konstan.menurut.Pengeluaran..Rupiah., Klasifikasi.Kemiskinan
```

The output above displays 13 column names in the dataset. At first glance, it appears that these column names tend to be long and less intuitive, potentially complicating the analysis process. Therefore, to improve readability, code writing efficiency, and ease of interpretation, it is a wise choice to rename or replace the column names to be shorter, more consistent, and more descriptive such as:

```
Provinsi, Kab/Kota, Persentase Penduduk Miskin, Lama Sekolah Penduduk, Pengel
uaran per Kapita, IPM, Harapan Hidup, Akses Sanitasi, Akses Air Minum, Pengan
gguran Terbuka, Partisipasi Angkatan Kerja, PDRB, Kategori Kemiskinan
```

This way, column names will be much more efficient and easier to understand, making further exploration and analysis easier.

Dataset Preview

The next important step is to preview the dataset. This preview will show the first few (head) and last few (tail) rows of the dataset, allowing us to examine the data values, and identify potential problems such as missing values or invalid entries.

Dataset Preview

	Provinsi	Kab/Kota	Persentase Penduduk Miskin	Lama Sekolah Penduduk	Pengeluaran per Kapita	IPM	Harapan Hidup	Akses Sanitasi	Akses Air Minum	Pengangguran Terbuka	Partisipasi Angkatan Kerja	PDRB	Kategori Kemiskinan
1	ACEH	Simeulue	18,98	9,48	7148	66.41	65.28	71.56	87.45	5.71	71.15	1648096	0
2	ACEH	Aceh Singkil	20,36	8,68	8776	69.22	67.43	69.56	78.58	8.36	62.85	1780419	1
3	ACEH	Aceh Selatan	13,18	8,88	8180	67.44	64.4	62.55	79.65	6.46	60.85	4345784	0
4	ACEH	Aceh Tenggara	13,41	9,67	8030	69.44	68.22	62.71	86.71	6.43	69.62	3487157	0
5	ACEH	Aceh Timur	14,45	8,21	8577	67.83	68.74	66.75	83.16	7.13	59.48	8433526	0
6	…	…	…	…	…	…	…	…	…	…	…	…	…
995					NA	NA	NA	NA	NA	NA	NA	NA	NA
996					NA	NA	NA	NA	NA	NA	NA	NA	NA
997					NA	NA	NA	NA	NA	NA	NA	NA	NA
998					NA	NA	NA	NA	NA	NA	NA	NA	NA
999					NA	NA	NA	NA	NA	NA	NA	NA	NA

Notes: The preview dataset visualization is obtained from knitting the code in HTML.

Based on the preview conducted by looking at the first and last 5 rows of the dataset, it was found that the head of the dataset looks safe and clean with completely filled data, while the tail of the dataset shows a significant problem where all the last rows (tail) contain missing values. This phenomenon most likely indicates that there are a number of rows at the end of the dataset that do not have information. Given that the total rows of this dataset is 999 and the number of cities/regencies in Indonesia does not reach that number, it is very likely that there is a certain range of rows that contain all missing values. To confirm and address this issue, a missing value check will be performed to identify the extent and location of incomplete entries in the dataset.

Number of Missing Values per Column

Column	Missing.Values
Provinsi	0
Kab/Kota	0
Persentase Penduduk Miskin	0
Lama Sekolah Penduduk	0
Pengeluaran per Kapita	485
IPM	485
Harapan Hidup	485

Column	Missing.Values
Akses Sanitasi	485
Akses Air Minum	485
Pengangguran Terbuka	485
Partisipasi Angkatan Kerja	485
PDRB	485
Kategori Kemiskinan	485

This confirms that the majority of the columns in the dataset have the same number of missing values, which is 485. This finding supports the previous observation from the dataset preview where the tail is completely empty (some are empty strings and NA). This indicates that the last 485 rows of the dataset are likely to be completely empty or have no valid data. Therefore, we should check the dataset preview again from row 515 (999 - 485 + 1) to the end.

Dataset Preview from Row 515

	Provinsi	Kab/Kota	Persentase Penduduk Miskin	Lama Sekolah Penduduk	Pengeluaran per Kapita	IPM	Harapan Hidup	Akses Sanitasi	Akses Air Minum	Pengangguran Terbuka	Partisipasi Angkatan Kerja	PDRB	Kategori Kemiskinan
515					NA	NA	NA	NA	NA	NA	NA	NA	NA
516					NA	NA	NA	NA	NA	NA	NA	NA	NA
517					NA	NA	NA	NA	NA	NA	NA	NA	NA
518					NA	NA	NA	NA	NA	NA	NA	NA	NA
519					NA	NA	NA	NA	NA	NA	NA	NA	NA
520	…	…	…	…	…	…	…	…	…	…	…	…	…
995					NA	NA	NA	NA	NA	NA	NA	NA	NA
996					NA	NA	NA	NA	NA	NA	NA	NA	NA
997					NA	NA	NA	NA	NA	NA	NA	NA	NA
998					NA	NA	NA	NA	NA	NA	NA	NA	NA
999					NA	NA	NA	NA	NA	NA	NA	NA	NA

Notes: The preview dataset visualization is obtained from knitting the code in HTML.

Upon further inspection of the dataset from row 515 to the end (row 999), it is confirmed that all rows in this range do indeed contain missing values in most of the columns. This strengthens the suspicion that the last 485 rows (999 - 514 = 485) of the dataset do not contain valid and reliable information for analysis. To confirm this in more detail and see the unique values in the columns.

Number of Unique Values in Dataset Starting from Row 515

Column	Unique.Values
Provinsi	1
Kab/Kota	1
Persentase Penduduk Miskin	1

Column	Unique.Values
Lama Sekolah Penduduk	1
Pengeluaran per Kapita	1
IPM	1
Harapan Hidup	1
Akses Sanitasi	1
Akses Air Minum	1
Pengangguran Terbuka	1
Partisipasi Angkatan Kerja	1
PDRB	1
Kategori Kemiskinan	1

Based on this, it is definitively confirmed that each column in this row range has only one unique value. This finding proves that the last 485 rows of the dataset all contain the same value (empty string and NA), consistent with the previous preview results. This clearly indicates that these rows are not part of the valid or relevant data for analysis. Thus, these empty rows will be removed as they do not contain any information.

Dataset Preview after Handling Missing Values

```
The dataset has 514 rows after handling missing values.
```

	Provinsi	Kab/Kota	Persentase Penduduk Miskin	Lama Sekolah Penduduk	Pengeluaran per Kapita	IPM	Harapan Hidup	Akses Sanitasi	Akses Air Minum	Pengangguran Terbuka	Partisipasi Angkatan Kerja	PDRB	Kategori Kemiskinan
1	ACEH	Simeulue	18,98	9,48	7148	66.41	65.28	71.56	87.45	5.71	71.15	1648096	0
2	ACEH	Aceh Singkil	20,36	8,68	8776	69.22	67.43	69.56	78.58	8.36	62.85	1780419	1
3	ACEH	Aceh Selatan	13,18	8,88	8180	67.44	64.4	62.55	79.65	6.46	60.85	4345784	0
4	ACEH	Aceh Tenggara	13,41	9,67	8030	69.44	68.22	62.71	86.71	6.43	69.62	3487157	0
5	ACEH	Aceh Timur	14,45	8,21	8577	67.83	68.74	66.75	83.16	7.13	59.48	8433526	0
6	…	…	…	…	…	…	…	…	…	…	…	…	…
510	PAPUA	Puncak	36,26	2,16	5412	43.17	65.86	11.43	85.03	0.94	89.43	831070	1
511	PAPUA	Dogiyai	28,81	4,94	5415	55	65.85	12.11	71.24	5.68	78.2	906904	1
512	PAPUA	Intan Jaya	41,66	3,09	5328	48.34	65.69	0.36	35.01	1.43	75.75	767101	1
513	PAPUA	Deiyai	40,59	3,25	4673	49.96	65.36	0	85.23	0.79	85.01	841296	1
514	PAPUA	Kota Jayapura	11,39	11,57	14937	80.11	70.52	85.31	97.1	11.67	63.75	22852202	0

Notes: The preview dataset visualization is obtained from knitting the code in HTML.

Number of Missing Values per Column After Handling Missing Values

Column	Missing.Values
Provinsi	0
Kab/Kota	0
Persentase Penduduk Miskin	0
Lama Sekolah Penduduk	0
Pengeluaran per Kapita	0
IPM	0
Harapan Hidup	0
Akses Sanitasi	0
Akses Air Minum	0
Pengangguran Terbuka	0
Partisipasi Angkatan Kerja	0
PDRB	0
Kategori Kemiskinan	0

After a careful data cleaning step including the removal of identified blank rows, the number of observations in the dataset is now reduced to 514 rows. The dataset preview at both the head and tail shows that there are no more missing value issues. All rows are now filled with complete and consistent data. Further confirmation by counting the number of missing values (NA) across the entire dataset also confirms that there are no more missing values detected in any columns. This indicates that the missing values issue has been fully addressed and the dataset is now clean and ready for the next stage of analysis.

The next problem is the “Poverty Category” column is currently still in numeric type. Most likely this is intended as a boolean value that represents a condition of 0 for not poor and 1 for poor. If so, converting to a factor type with clear labels would facilitate further interpretation and analysis. Last but not least, in the numeric columns that have decimal values (float), two versions of the decimal separator are identified, namely using a comma (in columns “Poverty Rate” and “Average Years of Schooling”) and a dot (from column “HDI” to “Labor Force Participation”). This indicates that there is a possibility of inconsistency in the data type in the dataset which can cause numeric columns to be read as character or object data types by R, which can hinder the statistical analysis process. These issues will be checked further in the next section.

Data Type of Each Column

The next step is to check the data type of each column. This is crucial to ensure that each variable is interpreted correctly by R, especially for columns discussed in the previous section.

Data Types per Column

Column	Type
Provinsi	character
Kab/Kota	character
Persentase Penduduk Miskin	character
Lama Sekolah Penduduk	character
Pengeluaran per Kapita	integer
IPM	numeric
Harapan Hidup	numeric
Akses Sanitasi	numeric
Akses Air Minum	numeric
Pengangguran Terbuka	numeric
Partisipasi Angkatan Kerja	numeric
PDRB	integer
Kategori Kemiskinan	integer

From the table given, it turns out that “Poverty Category” column type is integer.

```
These are the unique values found in the 'Poverty Category' column:

- 0
- 1
```

It was confirmed that the “Poverty Category” column is essentially a boolean variable as its unique values are limited to 0 and 1, indicating that it should represent logical values such as true and false. Given this, it is advisable to convert this column into a factor data type. This is because treating it as a categorical variable makes it more appropriate in analysis and modeling and also improves interpretability by making the analysis results clearer and more meaningful.

```
The 'Poverty Category' column has been successfully converted to a factor wit
h the following levels:

- Tidak
- Ya
```

Next, it turns out that there is indeed a data type inconsistency problem in the numeric columns. Specifically, the columns “Poverty Rate” and “Average Years of Schooling” which should be numeric, are identified as character types (chr). This happens because the decimal values in these columns use commas as decimal separators instead of dots which is the default standard for decimal numbers in R. This issue requires careful data type conversion.

Data Types per Column after Data Conversion

Column	Type
Provinsi	character
Kab/Kota	character
Persentase Penduduk Miskin	numeric
Lama Sekolah Penduduk	numeric
Pengeluaran per Kapita	integer
IPM	numeric
Harapan Hidup	numeric
Akses Sanitasi	numeric
Akses Air Minum	numeric
Pengangguran Terbuka	numeric
Partisipasi Angkatan Kerja	numeric
PDRB	integer
Kategori Kemiskinan	factor

The results of this verification show that all column data types are now correct and appropriate, including the “Poverty Category”, “Poverty Rate” and “Average Years of Schooling” columns which previously did not have appropriate data types.

Descriptive Statistics

This section presents descriptive statistical analysis to provide a comprehensive overview of the characteristics of the dataset. This analysis is essential to understand the distribution, variation, and initial patterns of each variable before moving on to more deep analysis. The coverage of descriptive statistics includes a general summary of the dataset, frequencies of data types per column, and detailed descriptive statistics for each data type, providing crucial initial insights into the quality and structure of the data.

Dataset Summary

Attribute	Value
Name	data
Number of rows	514
Number of columns	13

Column Data Type Frequency

Column Type	Frequency
character	2
factor	1
numeric	10

Numerical Descriptive Statistics

Variable	count	mean	median	min	max	sd	skewness	kurtosis	na
Persentase Penduduk Miskin	514	12.273	10.455	2.38	4.16600e+01	7.459	1.496	2.283	0
Lama Sekolah Penduduk	514	8.437	8.305	1.42	1.28300e+01	1.631	-0.351	1.533	0
Pengeluaran per Kapita	514	10324.788	10196.500	3976.00	2.38880e+04	2717.144	0.728	1.720	0
IPM	514	69.927	69.610	32.84	8.71800e+01	6.497	-0.776	3.555	0
Harapan Hidup	514	69.657	69.975	55.43	7.77300e+01	3.447	-0.466	0.587	0
Akses Sanitasi	514	77.202	81.800	0.00	9.99700e+01	18.584	-1.812	4.058	0
Akses Air Minum	514	85.137	89.795	0.00	1.00000e+02	15.702	-2.028	5.743	0
Pengangguran Terbuka	514	5.059	4.565	0.00	1.33700e+01	2.637	0.785	0.153	0
Partisipasi Angkatan Kerja	514	69.464	68.955	56.39	9.79300e+01	6.396	1.180	2.556	0
PDRB	514	21964077.482	8814925.500	147485.00	4.60081e+08	47904920.444	5.819	40.494	0

Notes: The summary table visualization is obtained from knitting the code in HTML.

For the purposes of a more stable analysis, the “GRDP” and “Per Capita Expenditure” variables are treated specifically because both reflect economic indicators in currency units and tend to have a wide distribution. In the context of cities/regencies in Indonesia, the variation in economic values is very sharp due to the disparities between regions, especially between economic centers and underdeveloped areas.

“GRDP” undergoes a logarithmic transformation to overcome extreme skewness (> 5) that can interfere with statistical interpretation and modeling. Meanwhile, in “Per Capita Expenditure”, no transformation is carried out, but capping is applied to outliers using the three-sigma rule. This means that values that are outside three standard deviations from the average are limited to maximum and minimum values that are still within these limits.

This step maintains the original scale of the data so that it can still be interpreted economically, while reducing the influence of outliers that can distort the model. Capping is preferred over deletion because it retains all observations and reduces the potential for bias due to data omissions.

Numerical Descriptive Statistics after Preprocessing

Variable	count	mean	median	min	max	sd	skewness	kurtosis	na
Persentase Penduduk Miskin	514	12.273	10.455	2.380	41.660	7.459	1.496	2.283	0
Lama Sekolah Penduduk	514	8.437	8.305	1.420	12.830	1.631	-0.351	1.533	0
Pengeluaran per Kapita	514	10306.366	10196.500	3976.000	18476.220	2647.862	0.499	0.608	0
IPM	514	69.927	69.610	32.840	87.180	6.497	-0.776	3.555	0
Harapan Hidup	514	69.657	69.975	55.430	77.730	3.447	-0.466	0.587	0
Akses Sanitasi	514	77.202	81.800	0.000	99.970	18.584	-1.812	4.058	0
Akses Air Minum	514	85.137	89.795	0.000	100.000	15.702	-2.028	5.743	0
Pengangguran Terbuka	514	5.059	4.565	0.000	13.370	2.637	0.785	0.153	0
Partisipasi Angkatan Kerja	514	69.464	68.955	56.390	97.930	6.396	1.180	2.556	0
PDRB	514	16.012	15.992	11.901	19.947	1.268	0.230	0.303	0

Notes: The summary table visualization is obtained from knitting the code in HTML.

Overall, these descriptive statistics confirm that the preprocessing steps on the “GDRP” and “Per Capita Expenditure” column have succeeded in creating a more stable and representative dataset. The transformation of “GRDP” and handling of outliers in “Per Capita Expenditure” have reduced data distortion and prepared these variables for more deep analysis.

Categorical Descriptive Statistics

Character Descriptive Statistics

skim_type	skim_variable	n_missing	complete_rate	character.min	character.max	character.empty	character.n_unique	character.whitespace
character	Provinsi	0	1	4	20	0	34	0
character	Kab/Kota	0	1	4	26	0	514	0

Notes: The summary table visualization is obtained from knitting the code in HTML.

Factor Descriptive Statistics

skim_type	skim_variable	n_missing	complete_rate	factor.ordered	factor.n_unique	factor.top_counts
factor	Kategori Kemiskinan	0	1	FALSE	2	Tid: 452, Ya: 62

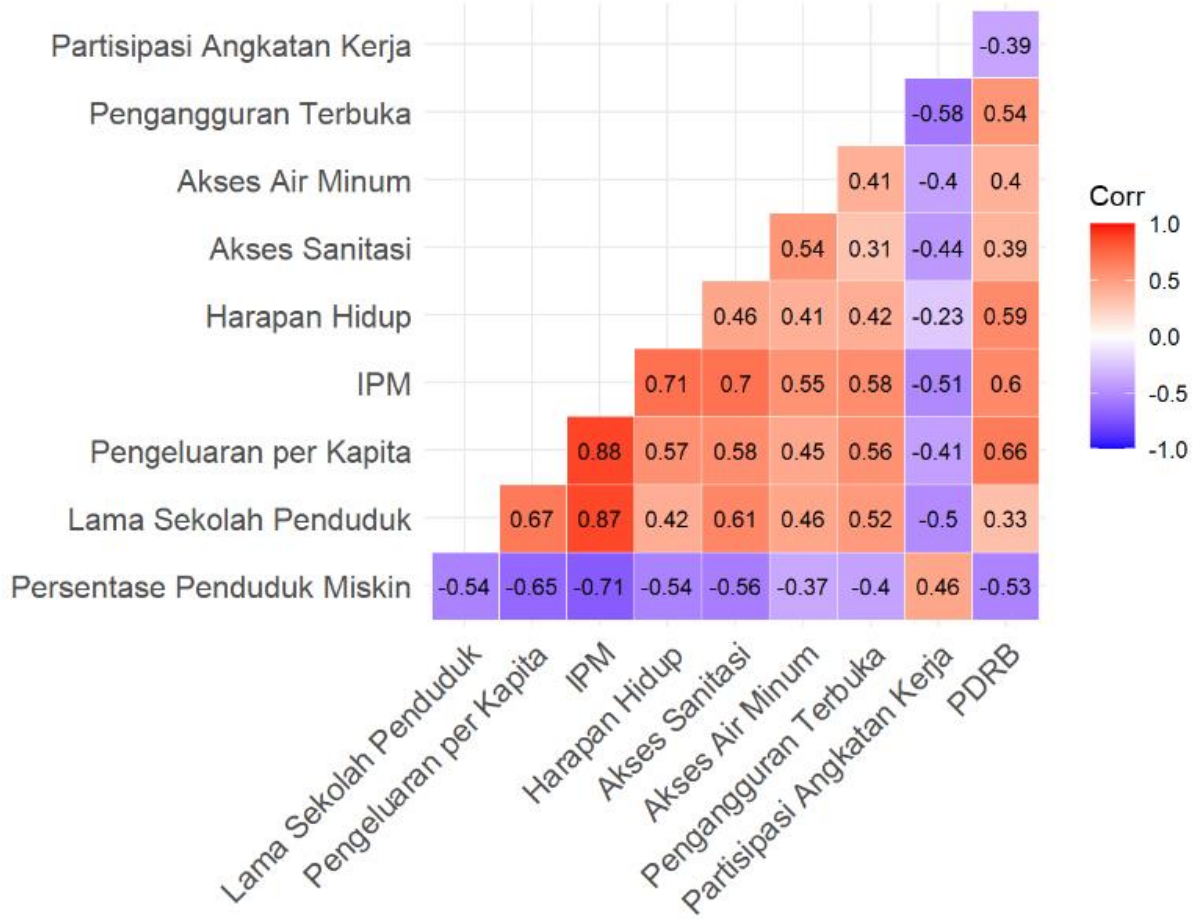
Notes: The summary table visualization is obtained from knitting the code in HTML.

Visualize Correlation

The results of the descriptive statistical analysis have provided a deep understanding of the characteristics and distribution of each variable in the dataset. This information is very

valuable as a foundation, but to truly understand how these variables are related to each other and which ones have the most significant influence on HDI, the next essential step is correlation visualization. Through this visualization, we can identify the strength and direction of the relationship between variables intuitively, which will be key in answering the research objectives regarding the factors that influence HDI in Indonesia.

Numerical Variables



Notes: The heatmap visualization is obtained from knitting the code in HTML.

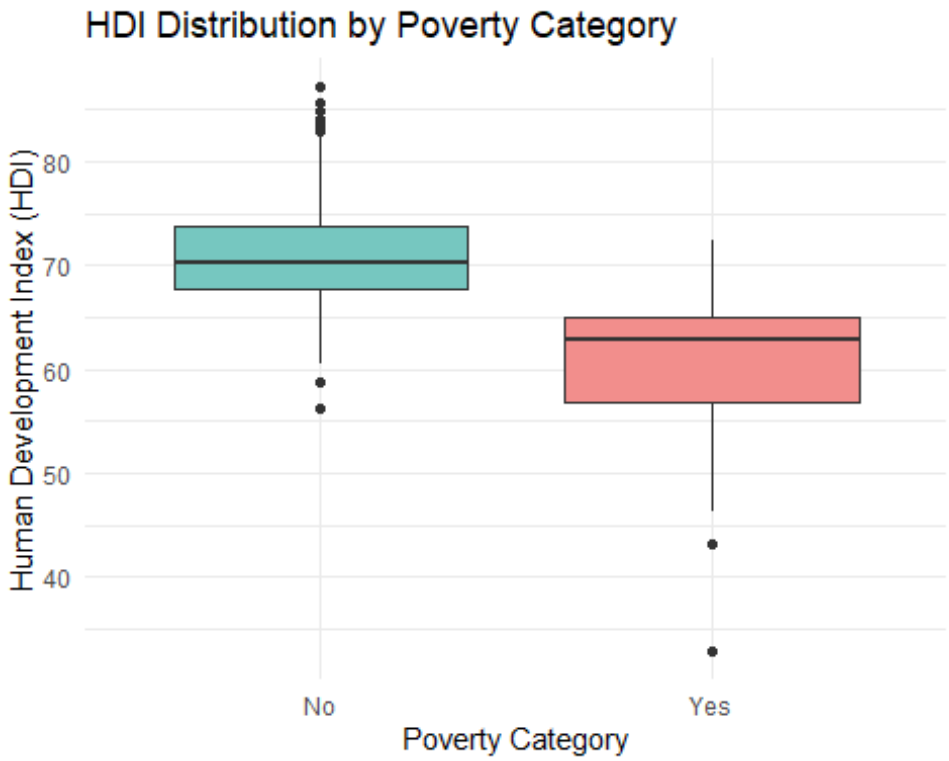
Based on the heatmap visualization of the correlation of numerical variables, we can identify 3 main factors that show a strong positive correlation with the HDI, namely: Per Capita Expenditure (0.88), Average Years of Schooling (0.87), and Life Expectancy (0.71). This finding is very consistent and strongly supported by the definition and framework of the HDI indicator itself. As previously explained, the HDI is built on three basic dimensions of human development.

This very strong correlation not only proves the relevance of the three indicators as the main pillars in calculating the HDI, but also empirically shows that efforts to improve the economic, education, and health sector will have a significant and direct impact on

increasing the HDI in various regions. This finding further strengthens the urgency of policies that focus on these three aspects to encourage holistic and equitable human development.

Although Poverty Rate also shows a strong correlation with HDI at 0.71, we chose not to include it as one of the three main factors. The reason is because conceptually, poverty rate is often considered as a consequence or reflection of low achievement in the HDI dimensions. In other words, high poverty rate is most likely caused by limited access to education, poor health services, and low income, all of which are already captured in the three core dimensions of HDI.

Categorical Variable



From the graph it can be seen that the variables “Poverty Category” and “HDI” variable are actually very closely related. In fact, poverty is one of the key components that intrinsically influences and is strongly correlated with the indicators that form the HDI.

However, in the context of this study which focuses on identifying the true independent variables that influence the HDI, variables such as Poverty Category will be set aside as these variables tend to be non-independent and are highly intertwined with the HDI components themselves, potentially leading to multicollinearity problems and not providing new predictive insights beyond what is already covered by the other HDI components.

Final Preprocessing

After identifying key variables that are strongly correlated with HDI and ensuring data quality, the final preprocessing step can now begin.

Feature Engineering

In this section, there are 2 new variables created.

First, the “HDI Category” variable is created from the numeric value of the HDI. This variable groups regions into three categories of human development based on predetermined thresholds: “Low” for HDI < 60, “Medium” for HDI >= 60 and < 70, and “High” for HDI >= 70. This categorization aims to simplify the interpretation of human development levels in various regions, allowing for more intuitive identification of groups of areas that require special attention and facilitating policy reporting.

Second, the “Education Level” variable is created from the Average Years of Schooling column that grouped into more familiar and contextual education levels: “ES” stands for Elementary School (up to 6 years), “JHS” stands for Junior High School (above 6 to 9 years), “SHS” stands for Senior High School (above 9 to 12 years), and “Uni” stands for University (above 12 years). The creation of this “Education Level” variable allows for the analysis of HDI patterns based on education levels becoming more easily understood in general and providing more applicable insights for stakeholders.

Matching Provinces Value with Spatial Dataset

The next crucial step is to combine our tabular data with a spatial dataset (geographic map). The main goal is to visualize the HDI and related indicators geographically, so that regional patterns and disparities can be clearly seen in map form. If there is a mismatch in the naming of the provinces, the data merging process will not be successful, and the map cannot be created correctly.

Duplicate Data

```
There are 0 duplicate rows in the dataset.
```

Categorical Descriptive Statistics after Final Preprocessing

Since feature engineering introduced 2 new variables of factor type, it’s a good idea to revisit the descriptive statistics for categorical (factor) variables.

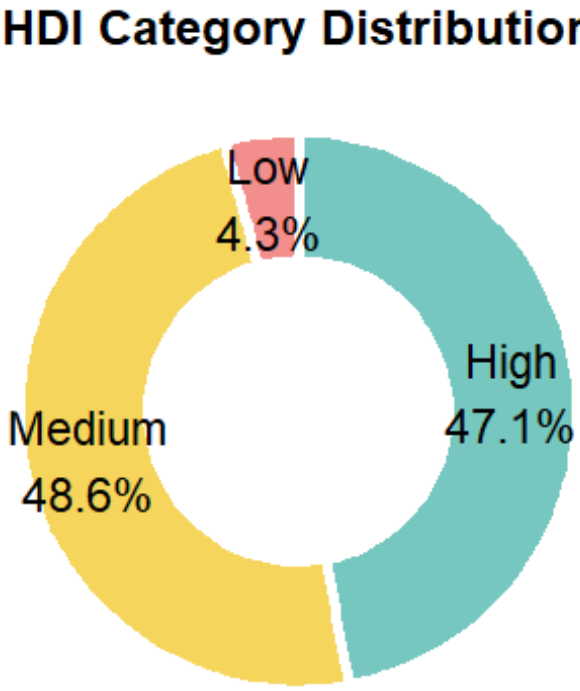
skim_type	skim_variable	n_missing	complete_rate	factor.ordered	factor.n_unique	factor.top_counts
factor	Kategori Kemiskinan	0	1	FALSE	2	Tid: 452, Ya: 62
factor	IPM_Kategori	0	1	TRUE	3	Med: 250, Hig: 242, Low: 22
factor	jenjang	0	1	FALSE	4	SMP: 334, SMA: 154, SD: 23, PT: 3

Notes: The summary table visualization is obtained from knitting the code in HTML.

Result Analysis and Visualization

HDI Overview in Indonesia

After categorizing the HDI into 3 categories in the previous section, the following visualization shows the proportion of each category in the dataset. This provides an overview of the distribution of human development levels across the observed regions.



Almost half of the observed areas (47.1%) are included in the “High” HDI category. This indicates that a large number of cities/regencies in Indonesia have achieved a good level of human development, exceeding 70.

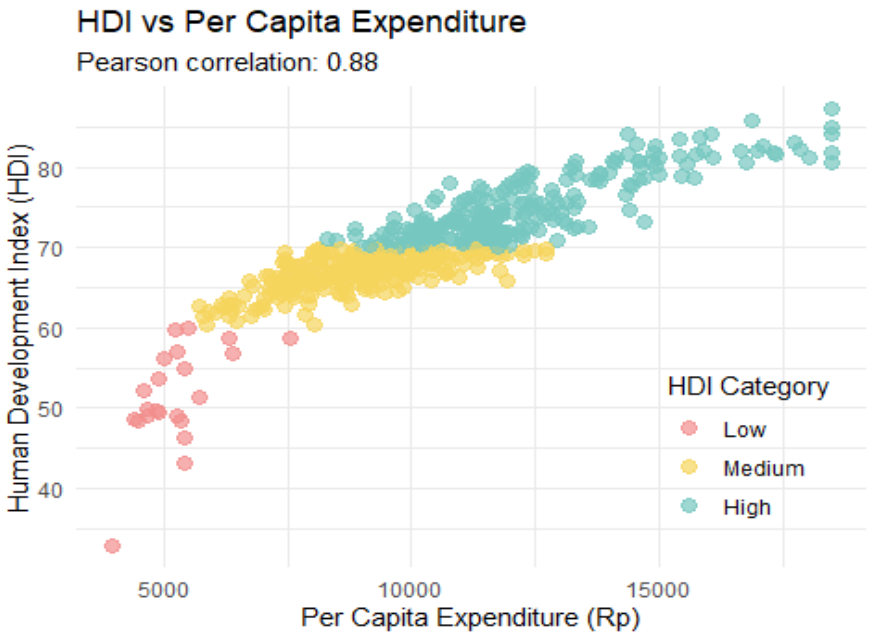
The largest proportion, namely 48.6% of the areas, are in the “Medium” HDI category (between 60 and 70). This shows that the majority of areas in Indonesia are still in the medium human development stage, which means there is still significant room for improvement towards the high category.

Only a small portion, namely 4.3% of the region, is in the “Low” HDI category (below 60). Although the percentage is small, the existence of this category highlights the existence of several regions that are still very far behind in human development and require urgent attention and policy intervention.

Overall, this visualization confirms the disparity in HDI achievements in Indonesia. The majority of regions are in the “Medium” and “High” categories, which are positive indications of development progress. However, the existence of a small percentage in the “Low” category is an important alarm to focus on the regions that need the most development encouragement so that they do not fall further behind. This data will be an important basis for further analysis of the factors that encourage or hinder regions from being in a particular category.

Relationship Between HDI and Per Capita Expenditure

One of the main indicators in the decent living standard dimension of the HDI is Per Capita Expenditure. The scatter plot visualization below illustrates the relationship between these two variables, with data points colored based on the HDI categories.



From the scatter plot above, it is clear that there is a very strong positive correlation between Per Capita Expenditure and the HDI, with a Pearson correlation coefficient of 0.88. Positive linear pattern is showed by most of the data points forming an upward trend from bottom left to top right. This means that along with the increase in Per Capita Expenditure, the HDI value also tends to increase significantly. The HDI grouping also suggests the existence of an indirect threshold such that regions with higher Per Capita Expenditure have a greater chance of achieving a “Medium” or “High” HDI.

This very strong correlation underlines the importance of Per Capita Expenditure as a vital indicator of human development. Increasing people's purchasing power and economic well-being directly contribute to improving the quality of life reflected in the HDI. Therefore, policies aimed at increasing people's income and expenditure, such as job creation, increasing the minimum wage, or social assistance programs, have great potential to drive HDI improvement.

This visualization visually reinforces the understanding that improving the standard of living, as measured by Per Capita Expenditure, is one of the key drivers in achieving better human development in Indonesia.

Education's Influence on HDI

Before visualizing the relationship between education level and HDI, data aggregation is needed to obtain summary statistics per education level category. This step aims to calculate the average HDI and the number of observations for each level, as well as prepare the data for a unique visualization.

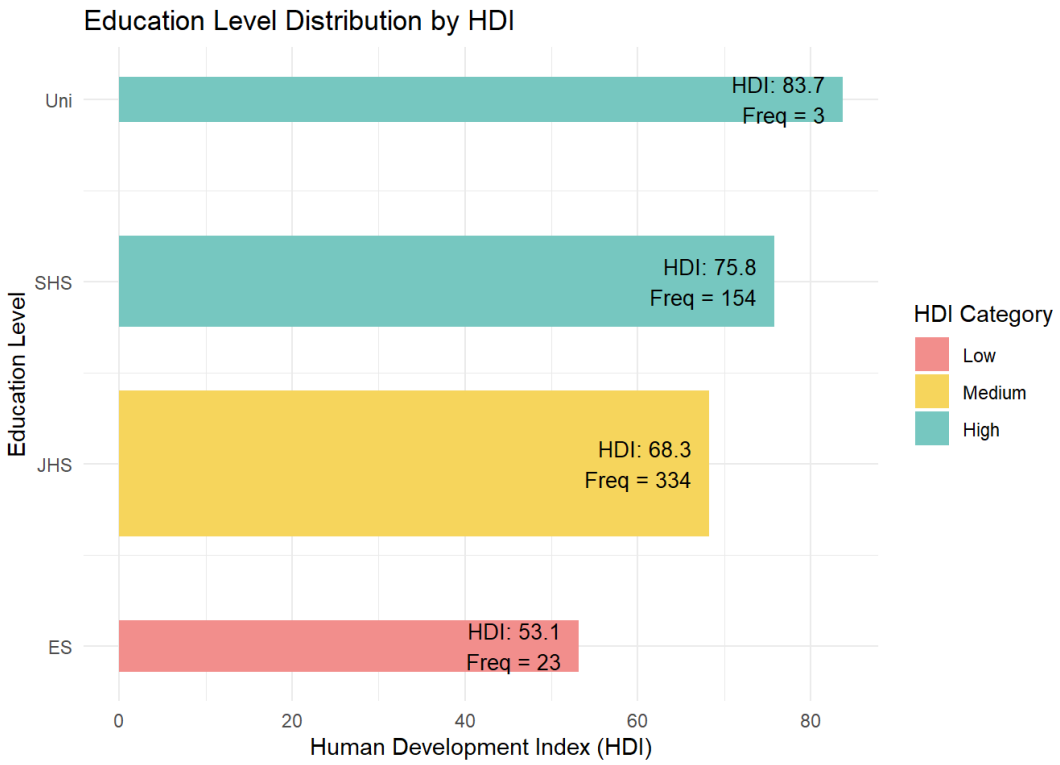
jenjang	count	IPM	IPM_Kategori
ES	23	53.12522	Low
JHS	334	68.25296	Medium
SHS	154	75.79805	High
Uni	3	83.69667	High

The result of this section is a summary table showing how many cities/regencies are at a particular education level, what their average HDI is, and their average HDI category. Next, further calculations are necessary to adjust for the differences in bar thickness, ensuring that the visualization of frequency distribution across categories is accurate and easy to interpret.

jenjang	count	IPM	IPM_Kategori	norm_count	bar_thickness	y_mid	ymin	ymax
SD	23	53.12522	Low	0.0604230	0.2832326	1	0.8583837	1.141616
SMP	334	68.25296	Medium	1.0000000	0.8000000	2	1.6000000	2.400000
SMA	154	75.79805	High	0.4561934	0.5009063	3	2.7495468	3.250453
PT	3	83.69667	High	0.0000000	0.2500000	4	3.8750000	4.125000

Notes: The table visualization is obtained from knitting the code in HTML.

The visualization below presents the relationship between the average education level of the population and the HDI value in various regions. This plot uniquely uses the length of the bar to represent the average HDI and the thickness of the bar to indicate the number of observations of cities/regencies in each education level, and the color of the bar corresponds to the average HDI category of that level.



There is a very strong positive relationship between education level and HDI. The higher the average education level of the population in a region, the higher the average HDI.

The University level shows the highest average HDI at 83.7, placing it in the “High” category. Although it has the fewest observations (only 3), its HDI value is notably high. Regions where the average years of schooling correspond to Senior High School have an average HDI of 75.8, also in the “High” category, with a total of 154 observations. For areas with schooling levels equivalent to Junior High School, the average HDI drops to 68.3, included in “Medium” category. This group has the highest frequency, with 334 observations, suggesting that most areas in the dataset are still at the Junior High School level. At the other end of the spectrum, areas with an average schooling level equivalent to Elementary School have the lowest HDI at 53.1, classified as “Low”, with 23 observations.

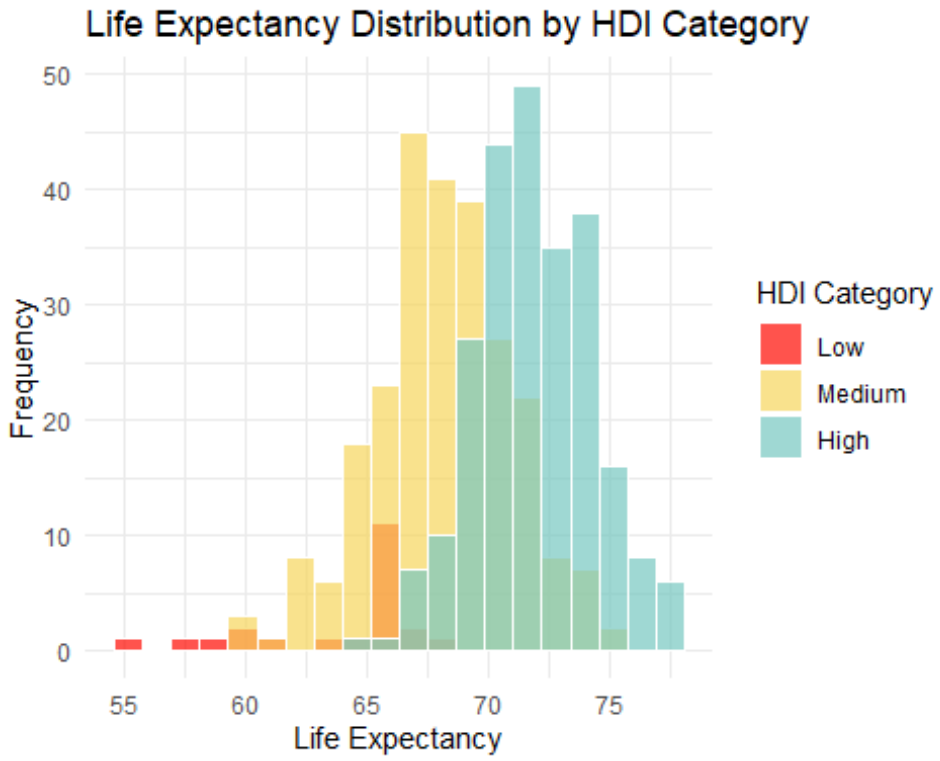
The Junior High School bar appears the thickest, corresponding to its large number of observations (334), highlighting that this level dominates the dataset. The bar for Senior High School is also relatively thick with 154 entries, while both Elementary School (23) and University (3) have thinner bars due to their smaller number of observations, representing the two extremes in education levels.

This visualization strongly supports the importance of education as a key pillar of human development. Increasing access and quality of education to higher levels has been shown to be directly correlated with increasing HDI. This finding underscores the need for continued investment in the education sector to drive improvements in the quality of human resources and, in turn, regional development.

Overall, this plot provides an intuitive and comprehensive illustration of how the level of education achieved by the population directly contributes to the progress of the HDI while also showing where the majority of Indonesia's regions are concentrated on the education spectrum.

Longevity Differences Across HDI Groups

The following overlay histogram visualizes the distribution of Life Expectancy, with colors representing HDI categories. This plot will show how life expectancy varies across regions with different HDI categories.



Regions classified as having a "Low" HDI tend to exhibit the lowest life expectancy distribution, mostly concentrated between 55 and 65 years. This indicates that areas with lower levels of human development are generally associated with shorter lifespans. In contrast, regions with a "Medium" HDI display a higher range, typically between 60 and 70 years, with the peak distribution around 67–68 years. Areas in the "High" HDI category have the highest life expectancy distribution, with most values above 70 years and a peak near 72–73 years. A few even reach close to 80 years, reflecting significantly better health outcomes.

Despite these clear distinctions, the distributions overlap to some extent. For instance, some areas in the "Medium" HDI category have life expectancies comparable to those in the "High" group, and vice versa. This suggests that while life expectancy is a key driver

of HDI, other contributing factors also play a role in determining an area's overall human development level.

This data highlights the need to invest in health services, sanitation, nutrition, and other factors that contribute to increased life expectancy, especially in areas that are still in the "Low" and "Medium" HDI categories. Improving the quality of life and health of the population will directly contribute to improving the overall HDI.

Overall, these plots visually confirm the strong relationship between population health and the level of human development reflected by HDI, suggesting that progress in one dimension tends to go hand in hand with progress in the other.

Mapping HDI per Province in Indonesia

This section requires verification of the province name.

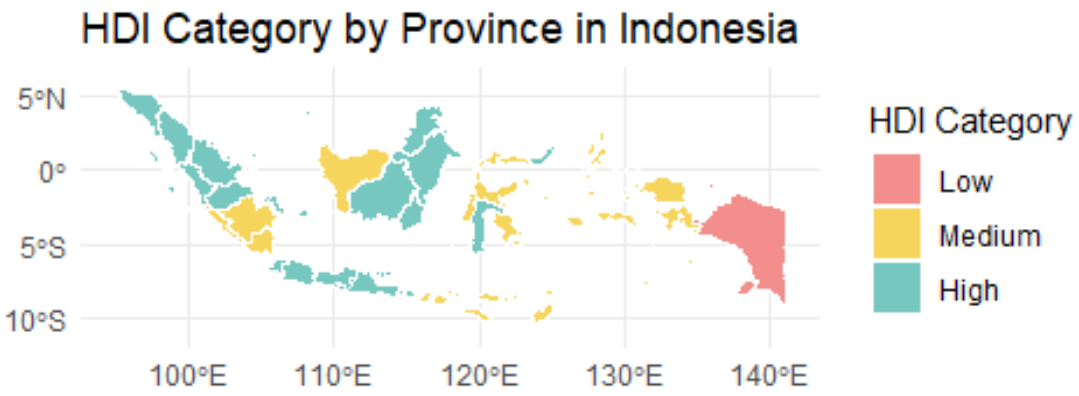
```
Unique province names in the spatial data (geographic map):

##  [1] "Aceh"                "Bali"                "Bangka Belitung"
##  [4] "Banten"              "Bengkulu"            "Gorontalo"
##  [7] "Jakarta Raya"        "Jambi"               "Jawa Barat"
## [10] "Jawa Tengah"         "Jawa Timur"          "Kalimantan Barat"
## [13] "Kalimantan Selatan"  "Kalimantan Tengah"   "Kalimantan Timur"
## [16] "Kalimantan Utara"    "Kepulauan Riau"      "Lampung"
## [19] "Maluku"              "Maluku Utara"        "Nusa Tenggara Barat"
## [22] "Nusa Tenggara Timur" "Papua"               "Papua Barat"
## [25] "Riau"                "Sulawesi Barat"      "Sulawesi Selatan"
## [28] "Sulawesi Tengah"     "Sulawesi Tenggara"   "Sulawesi Utara"
## [31] "Sumatera Barat"      "Sumatera Selatan"    "Sumatera Utara"
## [34] "Yogyakarta"


Unique province names in the tabular data:

##  [1] "Aceh"                "Sumatera Utara"      "Sumatera Barat"
##  [4] "Riau"                "Jambi"               "Sumatera Selatan"
##  [7] "Bengkulu"            "Lampung"             "Bangka Belitung"
## [10] "Kepulauan Riau"      "Jakarta Raya"        "Jawa Barat"
## [13] "Jawa Tengah"         "Yogyakarta"          "Jawa Timur"
## [16] "Banten"              "Bali"                "Nusa Tenggara Barat"
## [19] "Nusa Tenggara Timur" "Kalimantan Barat"    "Kalimantan Tengah"
## [22] "Kalimantan Selatan"  "Kalimantan Timur"    "Kalimantan Utara"
## [25] "Sulawesi Utara"      "Sulawesi Tengah"     "Sulawesi Selatan"
## [28] "Sulawesi Tenggara"   "Gorontalo"           "Sulawesi Barat"
## [31] "Maluku"              "Maluku Utara"        "Papua Barat"
## [34] "Papua"
```

This spatial visualization maps HDI categories across Indonesia, provides a clear picture of development disparities between provinces. The map uses the same coloring system as the previously defined HDI categories.



Most provinces in the western part of Indonesia, such as Java, Sumatra, and Kalimantan, are dominated by the “High” HDI category. This indicates that these regions, which are also the centers of economy and population, have achieved a high level of human development.

Several provinces in central and eastern Indonesia, such as Nusa Tenggara, Sulawesi, and Maluku, are mostly in the “Medium” HDI category. This indicates that despite progress, these regions are still in the intermediate stage of human development and have the potential and need to continue to improve.

Strikingly, Papua is the only region dominated by the “Low” HDI category. This confirms that Papua still faces the most significant development challenges compared to other provinces in Indonesia, requiring special attention and intervention to improve the quality of life of its population.

This map visually strongly shows that western Indonesia is far more advanced in human development compared to the eastern part. The geographic disparity is a major challenge in efforts to realize inclusive and equitable development. This pattern also highlights the urgency of more focused and intensive policies to encourage development in areas that are still lagging behind especially in eastern Indonesia.

Summary and Recommendation

Although Indonesia’s Human Development Index (HDI) shows an overall upward trend, this progress has not been evenly distributed across regions. Significant disparities are still exist especially in some eastern regions such as Papua, which still lag behind in access to education, health services, and decent living standards. This HDI gap highlights that development is not just about economic growth, but also ensuring that every citizen

has a fair chance to achieve a better quality of life. Ignoring this disparity risks leaving millions of people behind on the journey towards national progress.

To close this gap and ensure equitable development for all Indonesians, here are some key recommendations and solutions:

1. Equitable Access: Ensure equal access to health services, education, and decent work in areas that are still lagging and underdeveloped.
2. Local Economic Support: Promote local economic growth through skills training, entrepreneurship development, and supportive infrastructure development.
3. Community Empowerment: Directly involve and empower communities in the planning and decision-making process related to development in their areas.
4. Data-Driven Strategy and Inter-Regional Collaboration: Leverage data-driven strategies to effectively target and address HDI gaps, and foster inter-regional collaboration to share best practices and resources.