Data Science Project 1 2025: Poverty Rate in Malaysia

Luqman Arif Zulkarnain

2025-1-1

1.0 INTRODUCTION

For this project, five datasets will be analyzed to extract valuable insights through descriptive statistical methods and visualizations. The datasets include:

1. **Population**: State-level population data from 1970 to 2024, categorized by sex, age group, and ethnicity.
2. **Basic Amenities**: Proportion of households with access to essential amenities, such as electricity, piped water, and sanitary latrines, by state and district.
3. **Profile**: Number of households and living quarters by state, spanning from 1970 to 2024.
4. **Agriculture**: Production and planted areas of crops from 2017 to 2022.
5. **Household Income and Expenditure Survey (HIES)**: Household-level data on income, expenditure, poverty, and income inequality at the state level, based on the 2022 HIES.

1.1 HYPOTHESIS TEST

These datasets were sourced from the Department of Statistics Malaysia's official website (OpenDOSM). For this analysis, the data was filtered to focus on the specified variables across Malaysian states for the year 2022. The objective of this analysis to identify the relationship between poverty with other variables such as population, income, expenditure, and amenities. Therefore a hypothesis was proposed:

Null Hypothesis (H_0) : There is no relationship between poverty with population, income, expenditure, amenities and agriculture.

Alternative Hypothesis (H_1) : There is a relationship between poverty with population, income, expenditure, amenities and agriculture.

2.0 DATA COLLECTION AND DATA WRANGLING

Five datasets was retrieved from OpenDOSM websites and aggregated as one dataframe providing all the necessary observations from each variables for further analysis. The data was filtered for 2022 for this analysis.

```
# Load required libraries
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(arrow)
```

```
##
## Attaching package: 'arrow'

## The following object is masked from 'package:lubridate':
##
##     duration

## The following object is masked from 'package:utils':
##
##     timestamp
```

```
# URLs for the datasets
urls <- list(
  population = 'https://storage.dosm.gov.my/population/population_state.parquet',
  amenities = 'https://storage.dosm.gov.my/hies/hh_access_amenities.parquet',
  profile = 'https://storage.dosm.gov.my/demography/hh_lq_state.parquet',
  agriculture = 'https://storage.data.gov.my/agriculture/crops_state.parquet',
  hies = 'https://storage.dosm.gov.my/hies/hies_state.parquet'
)

# Load the datasets into a list of data frames
dataframes <- lapply(urls, read_parquet)

# Convert 'date' column to datetime format if it exists
for (name in names(dataframes)) {
  if ("date" %in% colnames(dataframes[[name]])) {
    dataframes[[name]]$date <- ymd(dataframes[[name]]$date)
  }
}
```

```r
# Filter data for the year 2022, group by 'state', and calculate the mean for numeric columns
for (name in names(dataframes)) {
  if ("date" %in% colnames(dataframes[[name]])) {
    filtered_df <- dataframes[[name]] %>%
      filter(year(date) == 2022)
    numeric_cols <- select(filtered_df, where(is.numeric))
    dataframes[[name]] <- filtered_df %>%
      select(state) %>%
      bind_cols(numeric_cols) %>%
      group_by(state) %>%
      summarise(across(where(is.numeric), mean, na.rm = TRUE), .groups = 'drop')
  }
}
```

```
## Warning: There was 1 warning in `summarise()`.
## i In argument: `across(where(is.numeric), mean, na.rm = TRUE)`.
## i In group 1: `state = "Johor"`.
## Caused by warning:
## ! The `...` argument of `across()` is deprecated as of dplyr 1.1.0.
## Supply arguments directly to `.fns` through an anonymous function instead.
##
##   # Previously
##   across(a:b, mean, na.rm = TRUE)
##
##   # Now
##   across(a:b, \(x) mean(x, na.rm = TRUE))
```

```r
# Perform left joins
data <- dataframes$population
for (name in c('amenities', 'profile', 'agriculture', 'hies')) {
  data <- left_join(data, dataframes[[name]], by = "state")
}

# Display the result
data
```

```
## # A tibble: 16 x 14
##    state             population piped_water sanitation electricity households
##    <chr>                  <dbl>       <dbl>      <dbl>       <dbl>      <dbl>
##  1 Johor                   80.8         98.3      100.        100.     1073400
##  2 Kedah                   43.4         99.6      100.        100.      546900
##  3 Kelantan                36.7         72.7       99.9        97.5     376700
##  4 Melaka                  20.2        100        100         100       285000
##  5 Negeri Sembilan         24.2         99.4      100         100.      334000
##  6 Pahang                  32.4         97.1       99.1        99.1     421700
##  7 Perak                   50.4         98.9      100         99.9      695300
##  8 Perlis                   5.81       100        100         100        87700
##  9 Pulau Pinang            34.9        100.       100         100.      519900
## 10 Sabah                   68.5         81.7       98.6        97.7     766600
## 11 Sarawak                 49.6         77.0      100.         94.7     636700
## 12 Selangor               141.         100        100.        100      1952800
## 13 Terengganu              23.8         98.4      100         99.8      298600
```

```
## 14 W.P. Kuala Lumpur      39.3        100        100         100       609800
## 15 W.P. Labuan             1.94       100        100         100        25300
## 16 W.P. Putrajaya          2.34       100        100         100        32200
## # i 8 more variables: living_quarters <dbl>, planted_area <dbl>,
## #   production <dbl>, income_mean <dbl>, income_median <dbl>,
## #   expenditure_mean <dbl>, gini <dbl>, poverty <dbl>
```

All the datasets has been merged into one dataframe using 'state' as the unique primary key to join.

```
data <- as.data.frame(data)
head(data,16)
```

```
##                state population piped_water sanitation electricity households
## 1              Johor  80.765915    98.31818   99.98636    99.96364    1073400
## 2              Kedah  43.370677    99.56154   99.99385    99.96923     546900
## 3           Kelantan  36.703509    72.67500   99.86250    97.51667     376700
## 4             Melaka  20.219048   100.00000  100.00000   100.00000     285000
## 5    Negeri Sembilan  24.214787    99.36250  100.00000    99.97500     334000
## 6             Pahang  32.365163    97.14167   99.06583    99.13333     421700
## 7              Perak  50.415038    98.87692  100.00000    99.93846     695300
## 8             Perlis   5.810526   100.00000  100.00000   100.00000      87700
## 9       Pulau Pinang  34.905013    99.98333  100.00000    99.98333     519900
## 10             Sabah  68.472932    81.67037   98.61481    97.66296     766600
## 11           Sarawak  49.594236    76.96829   99.97073    94.72927     636700
## 12          Selangor 141.357644   100.00000   99.99000   100.00000    1952800
## 13        Terengganu  23.792481    98.41111  100.00000    99.83333     298600
## 14 W.P. Kuala Lumpur  39.324561   100.00000  100.00000   100.00000     609800
## 15       W.P. Labuan   1.940852   100.00000  100.00000   100.00000      25300
## 16    W.P. Putrajaya   2.343108   100.00000  100.00000   100.00000      32200
##    living_quarters planted_area  production income_mean income_median
## 1          1323700 1.155047e+04 18831316.96        8517          6879
## 2           651400 2.549886e+04   193241.16        5550          4402
## 3           451900 1.396037e+04   526430.61        4885          3614
## 4           339600 1.814467e+03    82264.99        8057          6210
## 5           410900 1.454233e+03   325661.67        6788          5226
## 6           481500 8.094078e+03 10935355.78        5777          4753
## 7           840200 1.267304e+04  2843522.64        5779          4494
## 8            75800 7.173533e+03    64936.37        5664          4713
## 9           614100 3.073111e+03    27443.81        8267          6502
## 10          814800 1.156373e+04   798145.86        6171          4577
## 11          821500 2.056748e+04   172822.29        6457          4978
## 12         2227500 6.687778e+03  5310092.94       12233          9983
## 13          331700 3.744789e+03    45059.32        7248          5878
## 14          673500 9.333333e-01    91080.44       13325         10234
## 15           23600 4.590000e+01    10649.01        8250          6904
## 16           41400           NA          NA       13473         10056
##    expenditure_mean    gini poverty
## 1              5342 0.36646     4.6
## 2              3765 0.35938     9.0
## 3              3505 0.38540    13.2
## 4              5707 0.36963     4.2
## 5              4678 0.36853     4.4
## 6              4107 0.30770     6.3
```

```
## 7              3903 0.36769     7.5
## 8              3834 0.33589     4.0
## 9              5322 0.37058     2.0
## 10             3342 0.39491    19.7
## 11             3915 0.38180    10.8
## 12             6770 0.36123     1.5
## 13             4796 0.32631     6.2
## 14             7823 0.37960     1.4
## 15             4176 0.30028     2.5
## 16             8897 0.36780     0.1
```

2.1 DATA STRUCTURES

Based on the obtained results, **the dataframes consist of 16 rows observation** referring to the states in Malaysia and **14 columns variables** referring to states, population, piped_water, sanitation, electricity, households, living_quarters, planted_area, production, income_mean, income_median and expenditure_mean, gini and poverty.

```
dim(data) #identifying number of rows and columns in the dataset
```

```
## [1] 16 14
```

```
colSums(is.na(data)) #identifying total missing values for each column
```

```
##            state       population      piped_water       sanitation
##                0                0                0                0
##      electricity       households  living_quarters     planted_area
##                0                0                0                1
##       production      income_mean    income_median expenditure_mean
##                1                0                0                0
##             gini          poverty
##                0                0
```

2.2 MISSING VALUES IN DATA

It was revealed that the dataframes for ***planted_area*** **and** ***production*** **contained two missing values**, both **associated with the state of W.P. Putrajaya**. This is likely due to W.P. Putrajaya being a well-**developed urban area** with minimal or **no significant agricultural activity**. As a result, data collection for these variables might not have been prioritized or applicable to this region, further justifying the absence of values.

```
#statistical summary in Malaysia 2022
summary(data[ ,-1])
```

```
##    population       piped_water       sanitation      electricity
##  Min.   :  1.941   Min.   : 72.67   Min.   : 98.61   Min.   : 94.73
##  1st Qu.: 22.899   1st Qu.: 98.02   1st Qu.: 99.98   1st Qu.: 99.66
##  Median : 35.804   Median : 99.46   Median :100.00   Median : 99.97
##  Mean   : 40.975   Mean   : 95.19   Mean   : 99.84   Mean   : 99.29
##  3rd Qu.: 49.799   3rd Qu.:100.00   3rd Qu.:100.00   3rd Qu.:100.00
```

```
##  Max.   :141.358   Max.   :100.00   Max.   :100.00   Max.   :100.00
##
##    households      living_quarters    planted_area         production
##  Min.   :  25300   Min.   :  23600   Min.   :    0.933   Min.   :   10649
##  1st Qu.: 295200   1st Qu.: 337625   1st Qu.: 2443.789   1st Qu.:   73601
##  Median : 470800   Median : 547800   Median : 7173.533   Median :  193241
##  Mean   : 541413   Mean   : 632694   Mean   : 8526.851   Mean   : 2683868
##  3rd Qu.: 651350   3rd Qu.: 816475   3rd Qu.:12118.389   3rd Qu.: 1820834
##  Max.   :1952800   Max.   :2227500   Max.   :25498.856   Max.   :18831317
##                                      NA's   :1           NA's   :1
##   income_mean    income_median   expenditure_mean      gini
##  Min.   : 4885   Min.   : 3614   Min.   :3342     Min.   :0.3003
##  1st Qu.: 5778   1st Qu.: 4679   1st Qu.:3886     1st Qu.:0.3535
##  Median : 7018   Median : 5552   Median :4427     Median :0.3677
##  Mean   : 7903   Mean   : 6213   Mean   :4993     Mean   :0.3589
##  3rd Qu.: 8330   3rd Qu.: 6885   3rd Qu.:5433     3rd Qu.:0.3728
##  Max.   :13473   Max.   :10234   Max.   :8897     Max.   :0.3949
##
##     poverty
##  Min.   : 0.100
##  1st Qu.: 2.375
##  Median : 4.500
##  Mean   : 6.088
##  3rd Qu.: 7.875
##  Max.   :19.700
##
```

2.3 STATISTICAL SUMMARY OF DATA

Based on the **statistical summary** obtained for **Malaysia across state in 2022**, the **population ranges** from a **minimum of 1.941** to a **maximum of 141.358** thousand, with a **median of 35.804** thousand. The mean value of 40.975 suggests **a slight right skew**, meaning some areas **might have significantly higher populations**.

In terms of amenities, the access to **piped water** ranges between **72.67% and 100%**, with a high **median of 99.46%**, indicating that **most areas have reliable water supply**. Meanwhile, the **sanitation access** is even higher, **ranging from 98.61% to 100%**, with a **median and mean very close to 100%**. The **electricity availability** shows similar trends, with **values between 94.73% and 100%**, a median of **99.97%**, and a **mean of 99.29%**. These metrics reflect excellent infrastructure coverage overall.

The number of **households and living quarters** demonstrates **significant variation**. The households **range from 25,300 to 1,952,800**, with a **median of 470,800**. Whereas the **living quarters range** from **23,600 to 2,227,500**, with **a median of 547,800**. These variables suggest a mix of densely and sparsely populated areas.

In the context of **agricultural sector**, the **planted area** ranges from **0.933 hectares to 25,498.856 hectares**, with a **median of 7,173.533 hectares**. The large range indicates diversity in agricultural land use. **Agricultural production** spans **from 10,649 tonnes to a massive 18,831,317 tonnes**, with a **median of 193,241 tonnes**. Such variability likely reflects differences in agricultural productivity across regions.

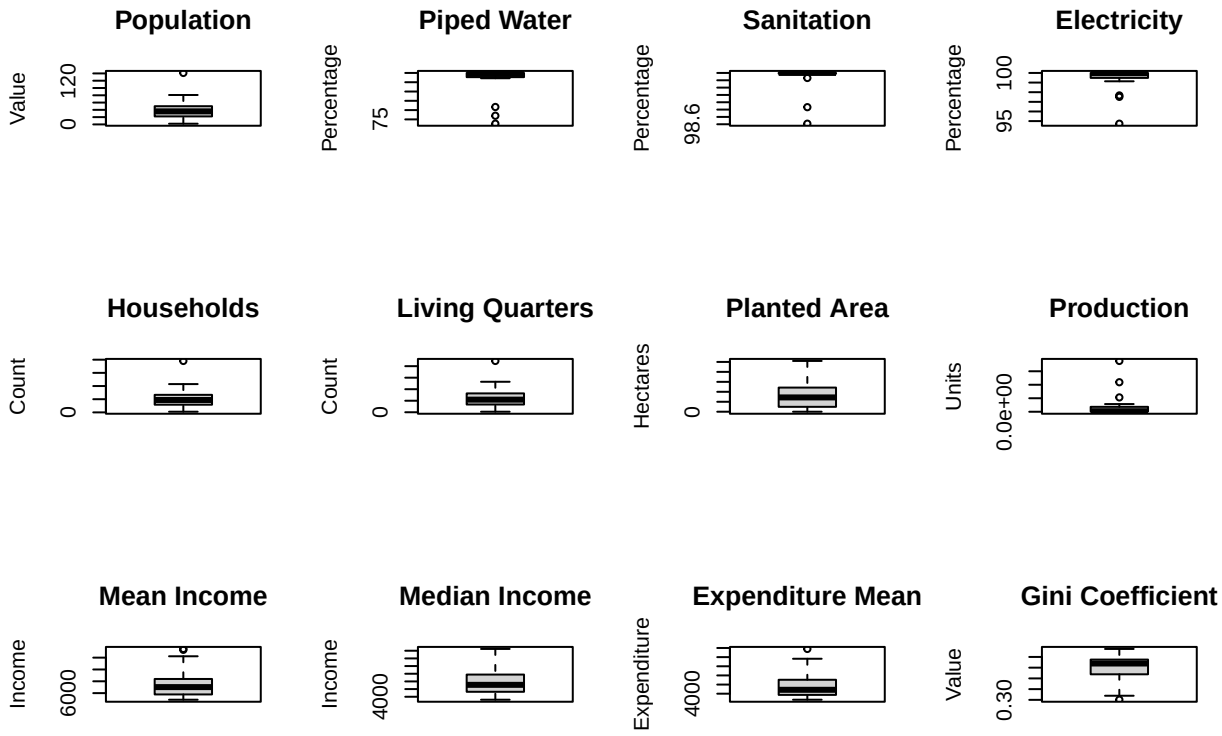
In terms **income and expenditure** in Malaysia the **average income ranges from RM4,885 to RM13,473**, with **a median of RM7,018**. The **income median ranges from RM3,641 to RM10,234**, with **a median of RM5,552**. The differences between mean and median incomes suggest income inequality in some areas. The **average expenditure** in Malaysia **ranges from RM3,342 to RM8,897** with **a median of RM4,427**.

As for **poverty rate and gini coefficient**, the **poverty rate** values **ranges from 19.7 to 0.1**, with a **median of 4.5**. The **gini coefficient** ranges from **0.3003 to 0.3949**, with **a median of 0.3677**. These values **suggest moderate income inequality in the dataset**.

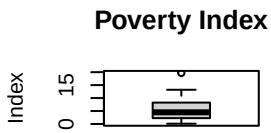
3.0 DATA EXPLORATION (DESCRIPTIVE STATISTICS)

3.1 BOX-PLOTTING

```r
# Adjust layout to fit all plots in one figure
par(mfrow = c(3, 4))

# Boxplots
boxplot(data$population, main="Population", ylab="Value")
boxplot(data$piped_water, main="Piped Water", ylab="Percentage")
boxplot(data$sanitation, main="Sanitation", ylab="Percentage")
boxplot(data$electricity, main="Electricity", ylab="Percentage")
boxplot(data$households, main="Households", ylab="Count")
boxplot(data$living_quarters, main="Living Quarters", ylab="Count")
boxplot(data$planted_area, main="Planted Area", ylab="Hectares")
boxplot(data$production, main="Production", ylab="Units")
boxplot(data$income_mean, main="Mean Income", ylab="Income")
boxplot(data$income_median, main="Median Income", ylab="Income")
boxplot(data$expenditure_mean, main="Expenditure Mean", ylab="Expenditure")
boxplot(data$gini, main="Gini Coefficient", ylab="Value")
```



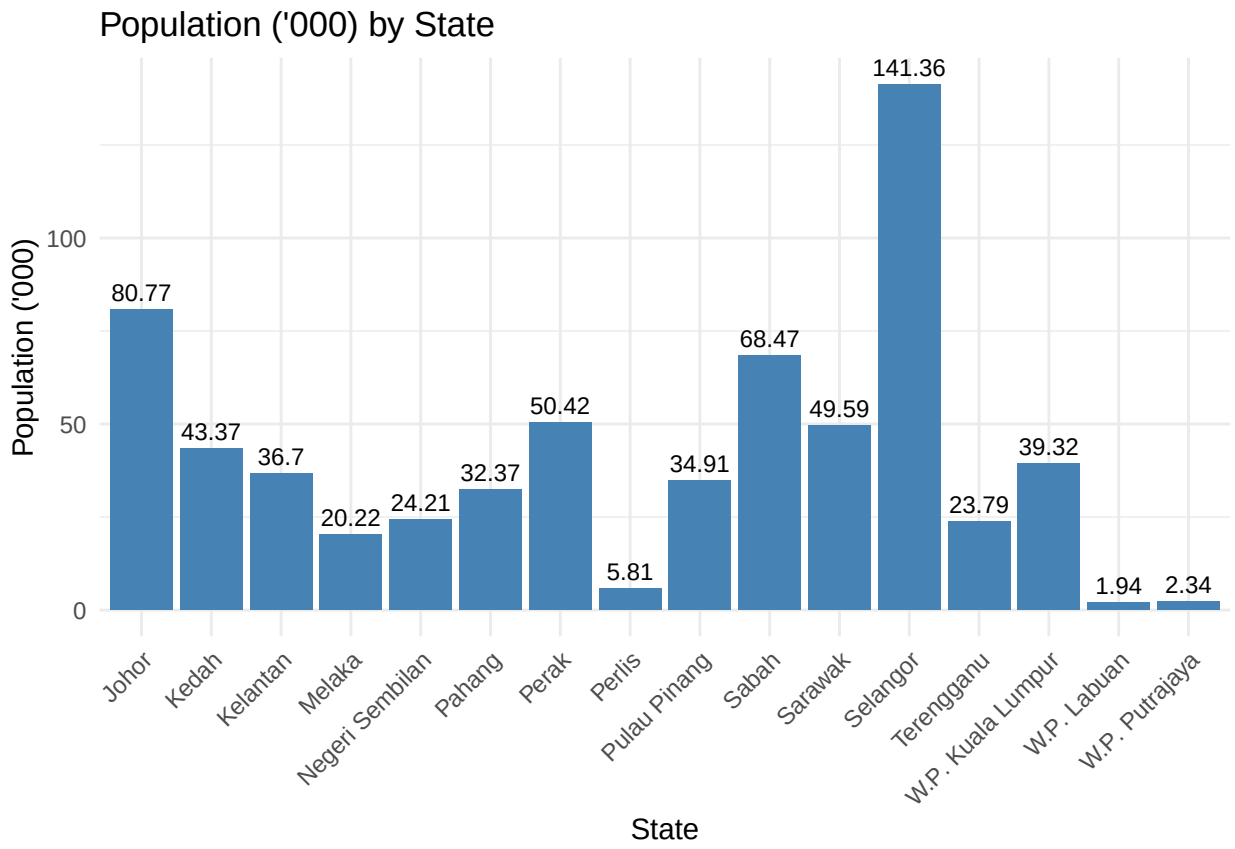

```r
boxplot(data$poverty, main="Poverty Index", ylab="Index")
```



3.2 BAR GRAPH

```r
# Load required libraries
library(dplyr)
library(ggplot2)

# Filter the necessary columns
data_2 <- data %>%
  select(state, population, piped_water, sanitation, electricity,
         income_mean, expenditure_mean, poverty,
         gini, households, planted_area, production)

# Group by state and sum the specified columns
grouped_data <- data_2 %>%
  group_by(state) %>%
  summarise(population = sum(population, na.rm = TRUE))

# Plot the grouped data
p1 <- ggplot(grouped_data, aes(x = state, y = population)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  geom_text(aes(label = round(population, 2)), vjust = -0.5, size = 3) +
  labs(title = "Population ('000) by State", x = "State", y = "Population ('000)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```r
# Display the plot
print(p1)
```

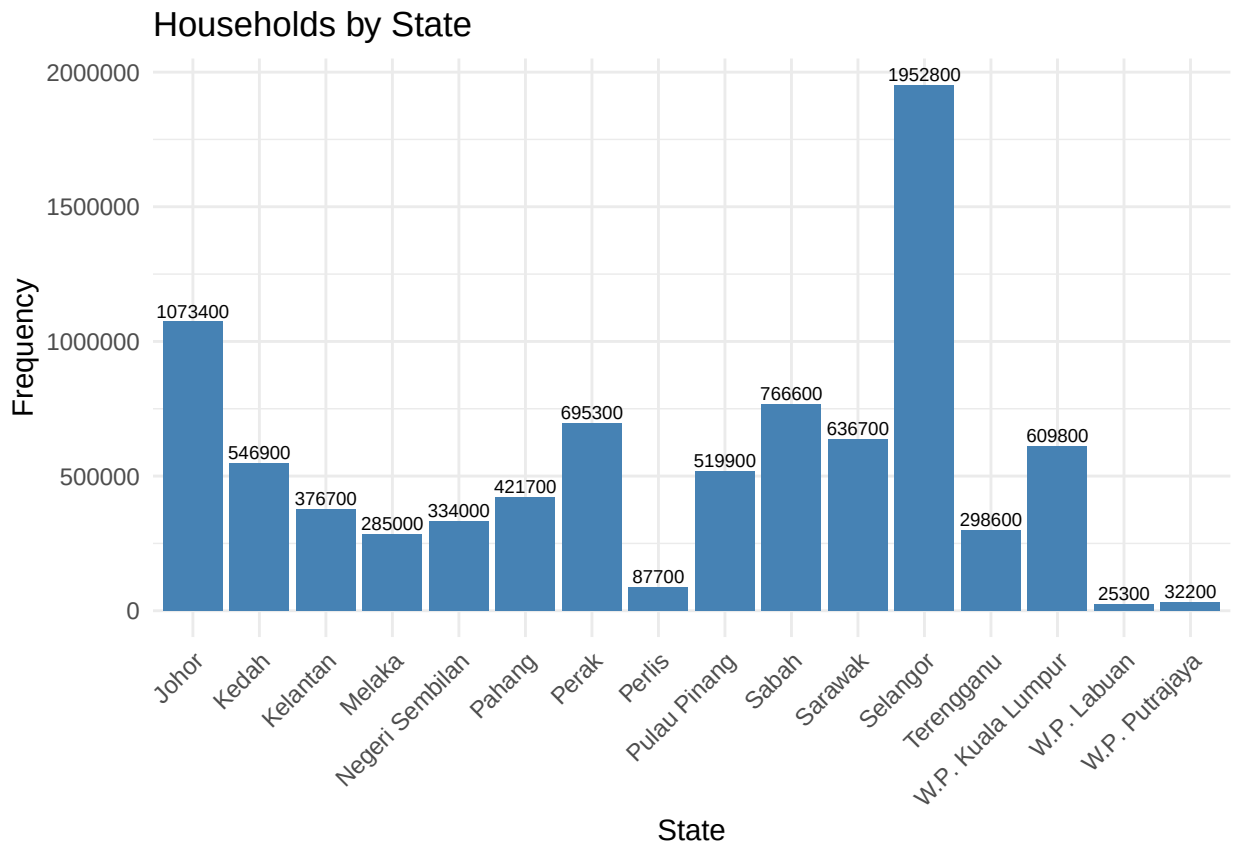


3.2.1 POPULATION

Based on the figure above, it was observed that the state of **Selangor** has the **highest population (141,360)** compared to other states. This can be attributed to Selangor's status as a **major economic and industrial hub**, offering **abundant employment opportunities across various sectors**. Additionally, its **well-developed infrastructure**, **proximity to the capital city of Kuala Lumpur**, and **access to quality education, healthcare, and urban amenities** make it an attractive destination for migration and settlement, further **contributing to its population growth**.

```r
# Group by state and sum the specified columns
grouped_data <- data_2 %>%
  group_by(state) %>%
  summarise(households = sum(households, na.rm = TRUE))

# Plot the grouped data
p2 <- ggplot(grouped_data, aes(x = state, y = households)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  geom_text(aes(label = round(households, 0)), vjust = -0.3, size = 2.5) +
  labs(title = "Households by State", x = "State", y = "Frequency") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```r
# Display the plot
print(p2)
```

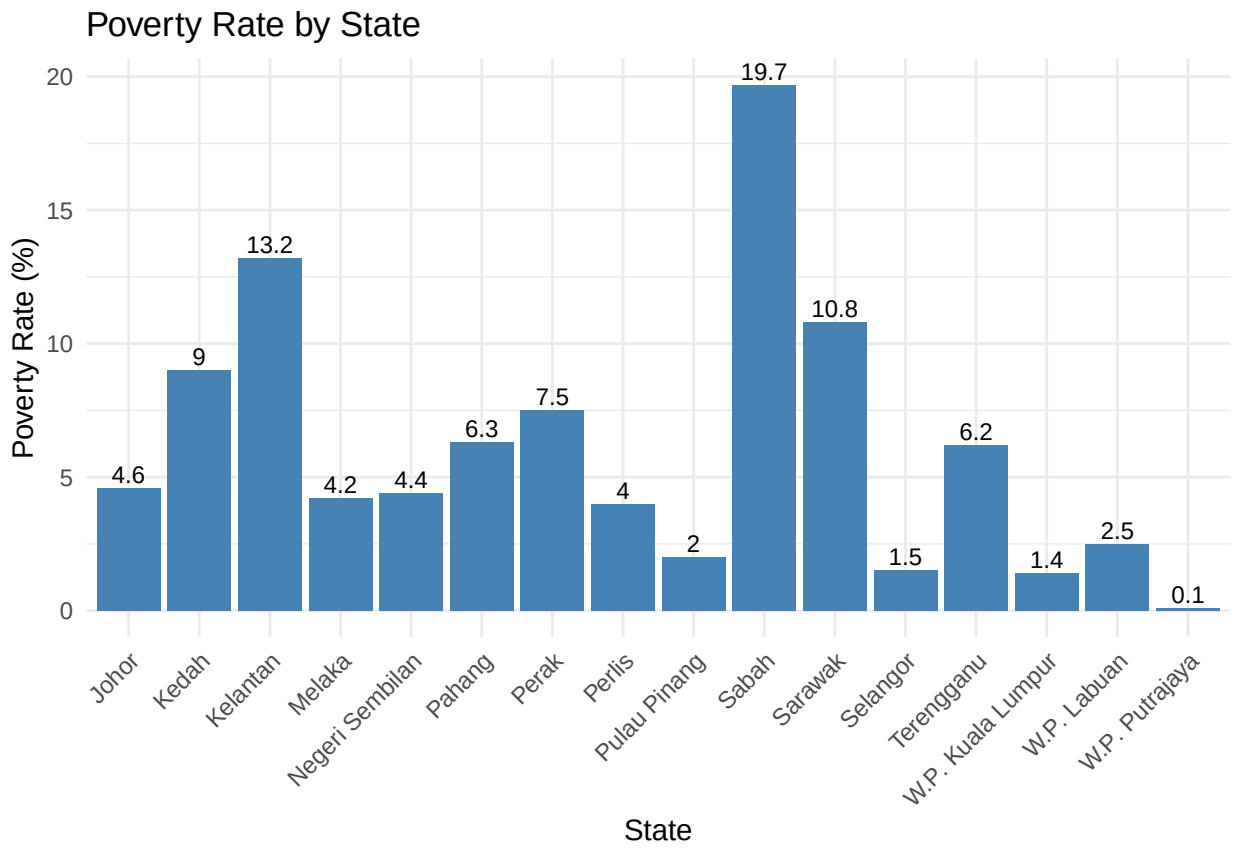


3.2.2 HOUSEHOLDS

In the context of **households**, the results consistently demonstrated that the state of **Selangor** has the **highest number of households compared to other states**. This is primarily due to Selangor's role as an **economic powerhouse, which attracts a large workforce and fosters population growth**. The state's **urbanization**, coupled with its **diverse employment opportunities**, leads to **higher demand for housing**. Furthermore, Selangor's **well-established infrastructure**, access to **essential services**, and **strategic location near the federal capital** make it a **preferred residential area for families** and **individuals**, contributing to the higher household count.

```r
# Group by state and sum the specified columns
grouped_data <- data_2 %>%
  group_by(state) %>%
  summarise(poverty = sum(poverty, na.rm = TRUE))

# Plot the grouped data
p3 <- ggplot(grouped_data, aes(x = state, y = poverty)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  geom_text(aes(label = round(poverty, 2)), vjust = -0.3, size = 3) +
  labs(title = "Poverty Rate by State", x = "State", y = "Poverty Rate (%)") +
  theme_minimal() +
```

```
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Display the plot
print(p3)
```



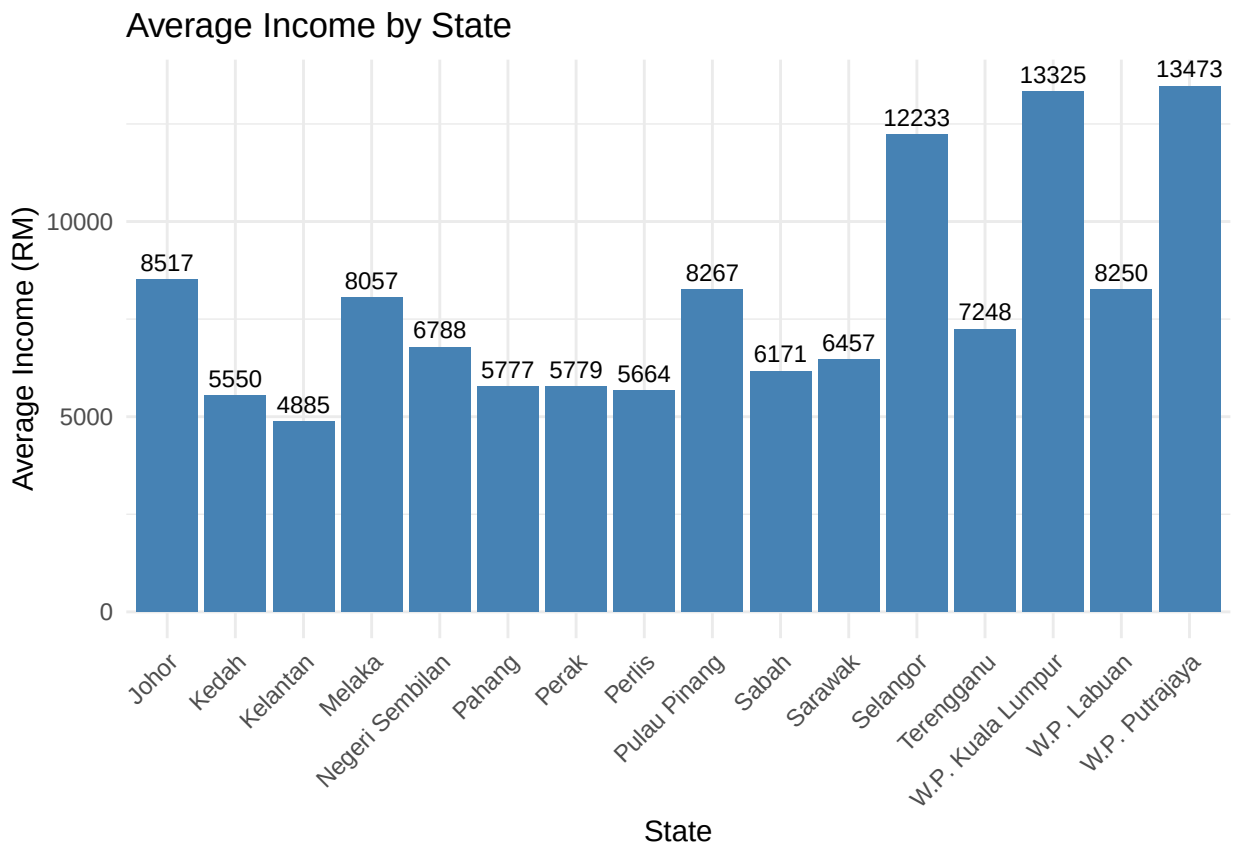
3.2.3 POVERTY

In terms of **poverty rate**, the figure above reveals that the state of **Sabah** has the **highest poverty rate** (**19.7%**), **followed by Kelantan** (**13.2%**) among all states in Malaysia. This is largely attributed to several socio-economic challenges faced by these states. **Sabah**, being **geographically remote** and **predominantly rural, struggles with limited access to infrastructure, education, and healthcare**, which **hampers economic development and livelihood opportunities**.

Similarly, **Kelantan** faces **economic constraints** due to **its focus on agriculture, limited industrialization**, and **lower levels of private sector investment**. Both states also experience challenges in addressing **income inequality** and **ensuring equitable resource distribution**, further **contributing to their higher poverty rates**.

```
# Group by state and sum the specified columns
grouped_data_income <- data_2 %>%
  group_by(state) %>%
  summarise(income_mean = sum(income_mean, na.rm = TRUE))

# Plot the grouped data for income_mean
p4 <- ggplot(grouped_data_income, aes(x = state, y = income_mean)) +
```

```
  geom_bar(stat = "identity", fill = "steelblue") +
  geom_text(aes(label = round(income_mean, 2)), vjust = -0.5, size = 3) +
  labs(title = "Average Income by State", x = "State", y = "Average Income (RM)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Display the plot
print(p4)
```

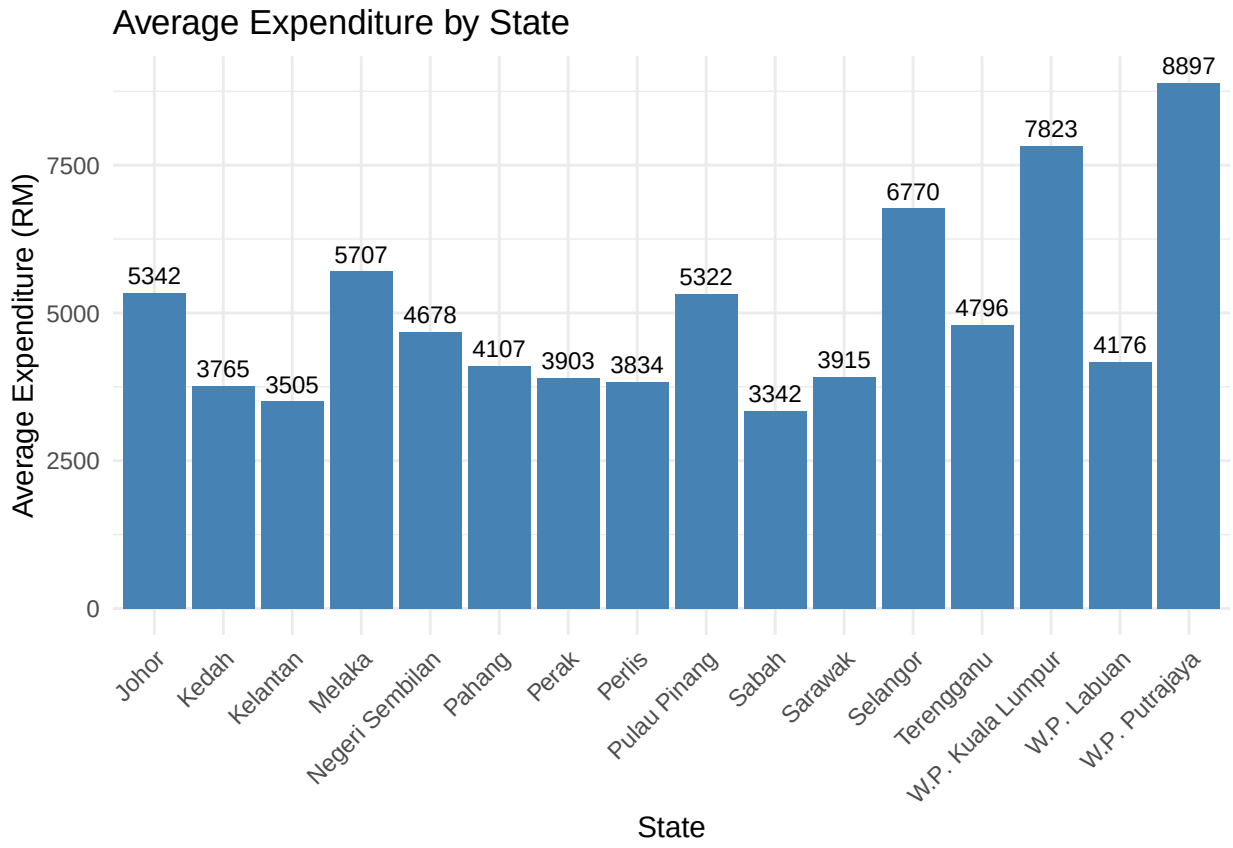


3.2.4 INCOME

In terms of **average income** in Malaysia, it was observed that **W.P. Kuala Lumpur, W.P. Putrajaya, and Selangor** recorded the **highest average incomes**, amounting to **RM13,325, RM13,473, and RM12,233, respectively**. This can be attributed to the **high concentration of economic activities, including finance, technology, and professional services**, in these urban and developed areas. These regions also **benefit from a higher cost of living**, which is often **correlated with better-paying job opportunities.**

Conversely, **Kelantan** recorded the **lowest average income** at **RM4,885**, which reflects its **predominantly rural economy, reliance on agriculture, and limited industrial and commercial development**. The lower level of economic diversification and investment in Kelantan contributes to **fewer high-paying job opportunities**, resulting in a **significantly lower average income** compared to more developed states.

```r
# Group by state and sum the specified columns
grouped_data_expenditure_mean <- data_2 %>%
  group_by(state) %>%
  summarise(expenditure_mean = sum(expenditure_mean, na.rm = TRUE))

# Plot the grouped data for average expenditure
p5 <- ggplot(grouped_data_expenditure_mean, aes(x = state, y = expenditure_mean)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  geom_text(aes(label = round(expenditure_mean, 2)), vjust = -0.5, size = 3) +
  labs(title = "Average Expenditure by State", x = "State", y = "Average Expenditure (RM)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Display the plot
print(p5)
```

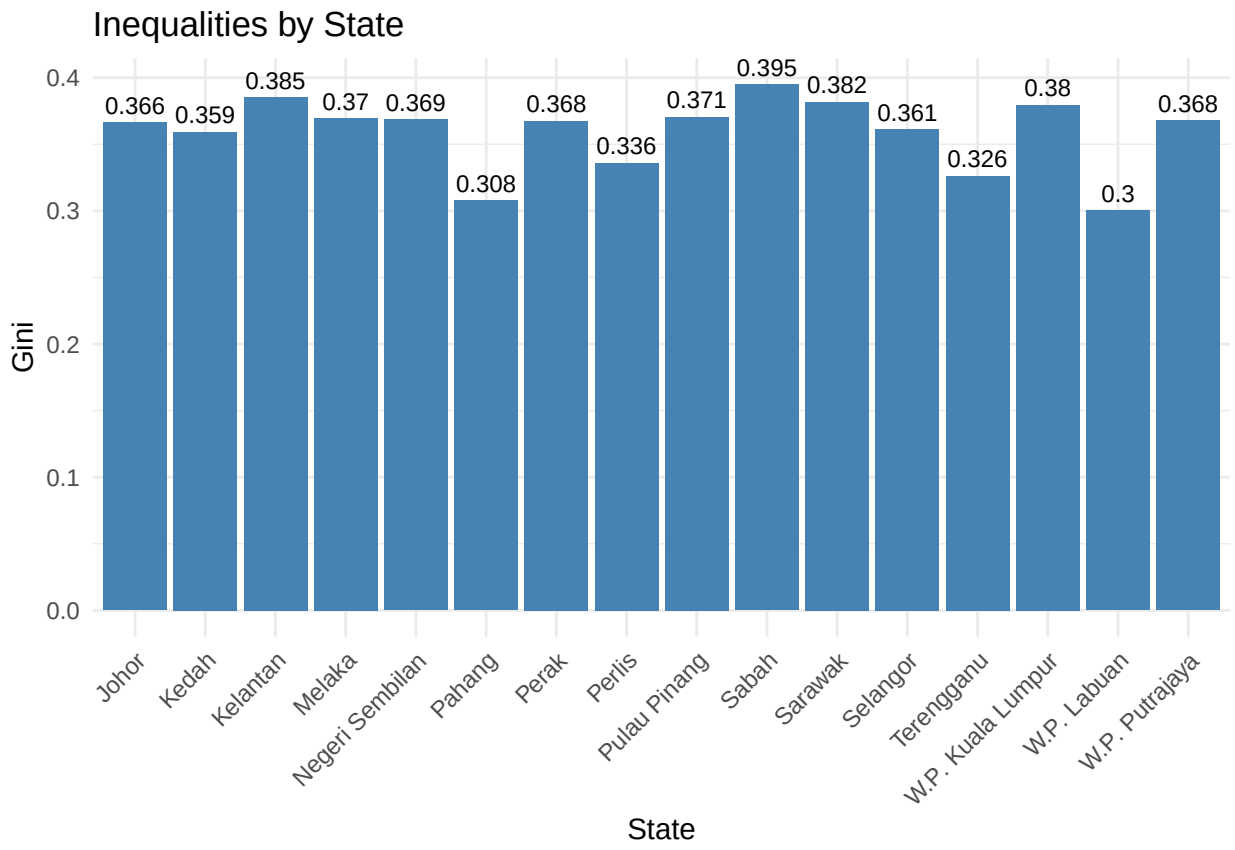


3.2.5 EXPENDITURE

In terms of **average expenditure** in Malaysia, **W.P. Putrajaya** recorded the **highest average expenditure** at **RM8,897**, while **Sabah** had the **lowest at RM3,342**. The **high expenditure in W.P. Putrajaya** can be attributed to **its status as the administrative capital**, where residents often have **higher disposable incomes** due to **employment in government** and **high-skilled sectors**. Additionally, the cost of living in Putrajaya is elevated due to urban amenities, housing costs, and access to premium services.

In contrast, **Sabah's lower average expenditure** reflects its **predominantly rural and agrarian economy**, where residents generally have **lower disposable incomes**. Limited access to urban amenities, **lower consumer purchasing power**, and **a focus on subsistence-based lifestyles** contribute to the **reduced average expenditure** in the state compared to more developed regions in Malaysia.

```r
# Group by state and sum the specified columns
grouped_data_gini <- data_2 %>%
  group_by(state) %>%
  summarise(gini = sum(gini, na.rm = TRUE))

# Plot the grouped data for gini
p6 <- ggplot(grouped_data_gini, aes(x = state, y = gini)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  geom_text(aes(label = round(gini, 3)), vjust = -0.5, size = 3) +
  labs(title = "Inequalities by State", x = "State", y = "Gini") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Display the plot
print(p6)
```



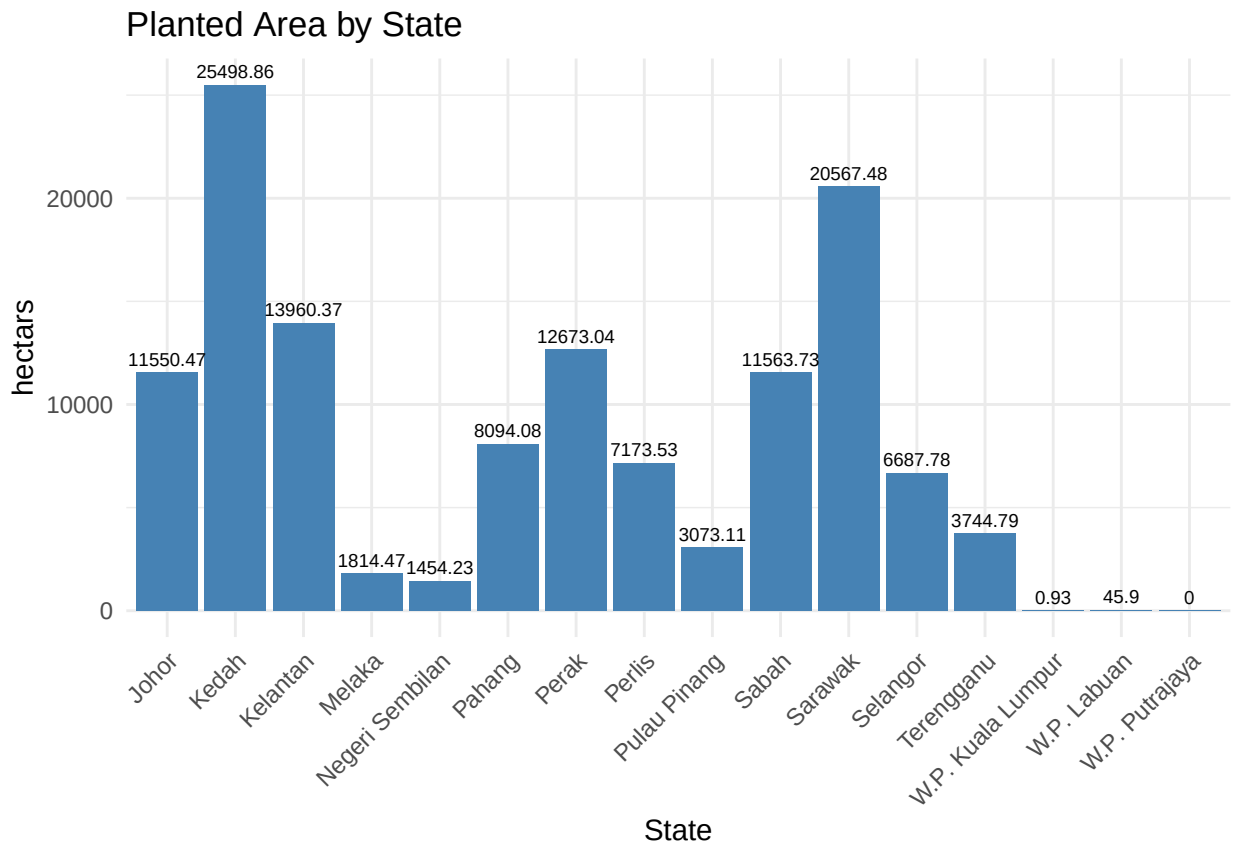
3.2.6 INEQUALITIES

In the context of **inequalities** in Malaysia, the results indicate that **Sabah, Sarawak, and Kelantan** have **the highest Gini coefficients, at 0.395, 0.382, and 0.385 respectively**. These figures highlight

significant income disparities within these states. In Sabah and Sarawak, **geographical challenges, including remote and inaccessible areas, contribute to uneven development and limited access to economic opportunities, education, and healthcare, exacerbating income inequality**.

Similarly, Kelantan's reliance on **agriculture and its slower pace of industrialization lead to fewer high-income opportunities**, further **widening the income gap**. These states also **experience disparities in resource distribution** and **infrastructure development**, which **hinder equitable economic growth** and **perpetuate inequality**. Addressing these structural issues is crucial to **reducing income disparities** and **promoting inclusive development**.

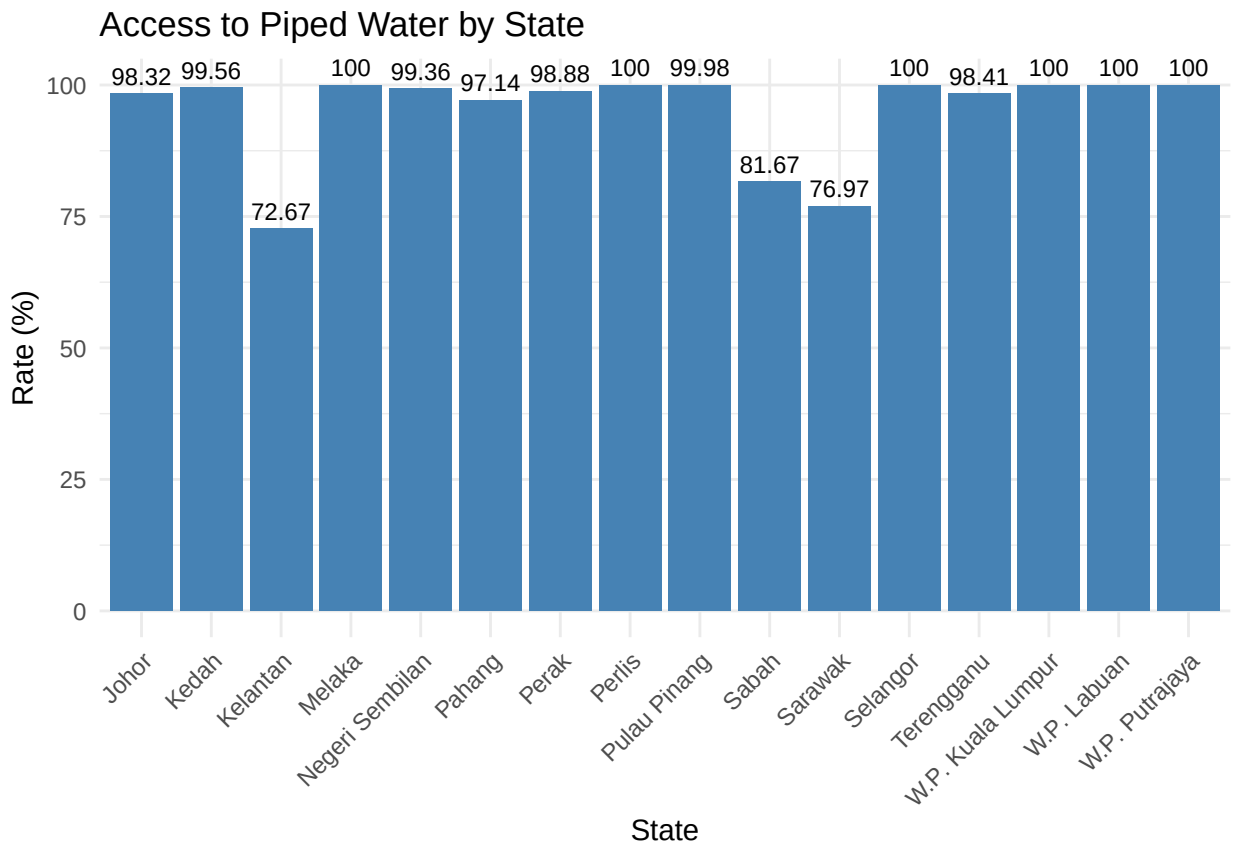
```r
# Group by state and sum the planted area values
grouped_data_planted_area <- data_2 %>%
  group_by(state) %>%
  summarise(planted_area = sum(planted_area, na.rm = TRUE))

# Plot the grouped data for planted area
p7 <- ggplot(grouped_data_planted_area, aes(x = state, y = planted_area)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  geom_text(aes(label = round(planted_area, 2)), vjust = -0.5, size = 2.5, angle = 0) +
  labs(title = "Planted Area by State", x = "State", y = "hectars") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Display the plot
print(p7)
```



3.2.7 AGRICULTURES

In terms of **planted area** in Malaysia, **Kedah** stands out as the **state with the largest planted area**, **covering 25,498.86 hectars**. This significant area highlights **Kedah's pivotal role in the nation's agricultural sector**, particularly in rice cultivation, where it has long been **known as the "Rice Bowl" of Malaysia**. The state's extensive **arable land and favorable climatic conditions** make it an **ideal location for large-scale farming**, contributing to its dominance in planted area compared to other states in the country. This large planted area also underscores Kedah's importance in **ensuring food security** and **sustaining agricultural production** at the national level.

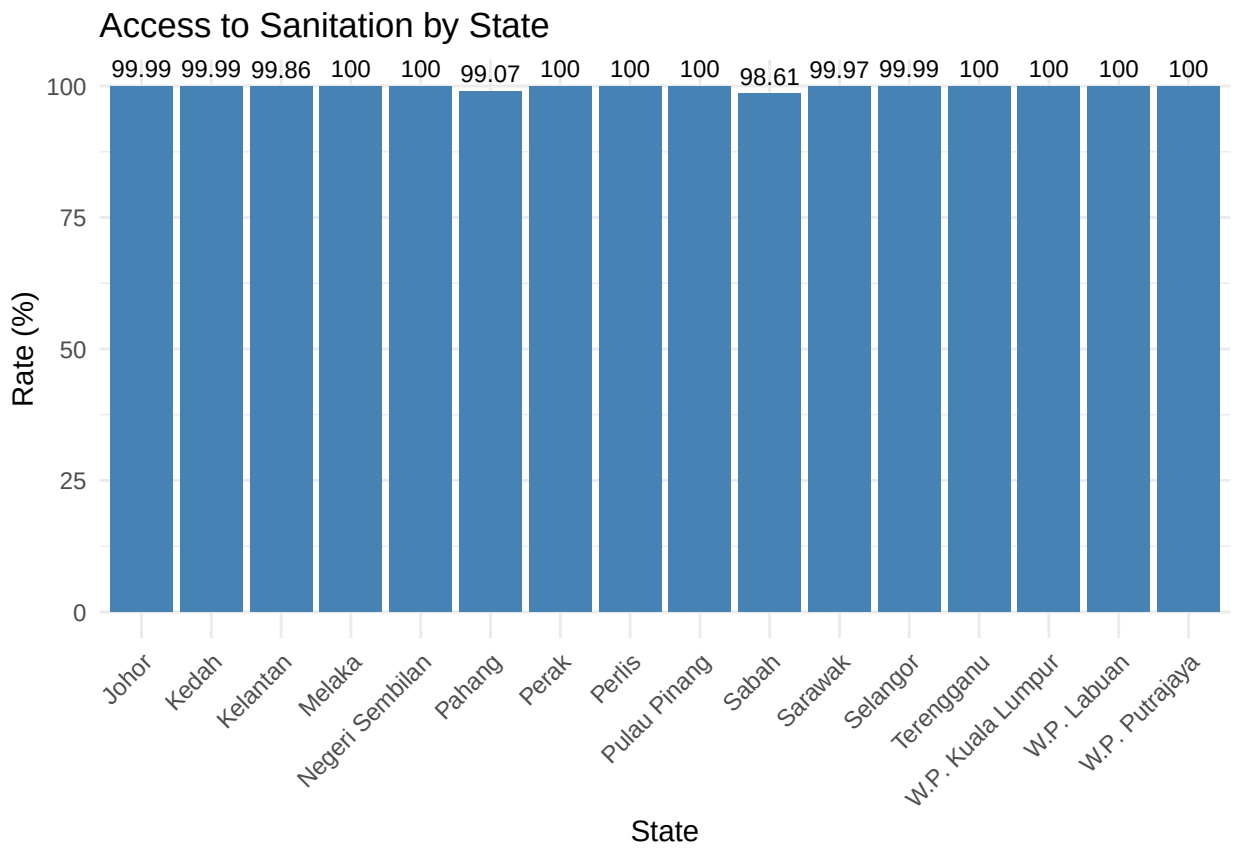
```
# Group by state and sum the access to piped water values
grouped_data_water <- data_2 %>%
  group_by(state) %>%
  summarise(piped_water = sum(piped_water, na.rm = TRUE))

# Plot the grouped data for production
p8 <- ggplot(grouped_data_water, aes(x = state, y = piped_water)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  geom_text(aes(label = round(piped_water, 2)), vjust = -0.5, size = 3) +
  labs(title = "Access to Piped Water by State", x = "State", y = "Rate (%)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Display the plot
print(p8)
```



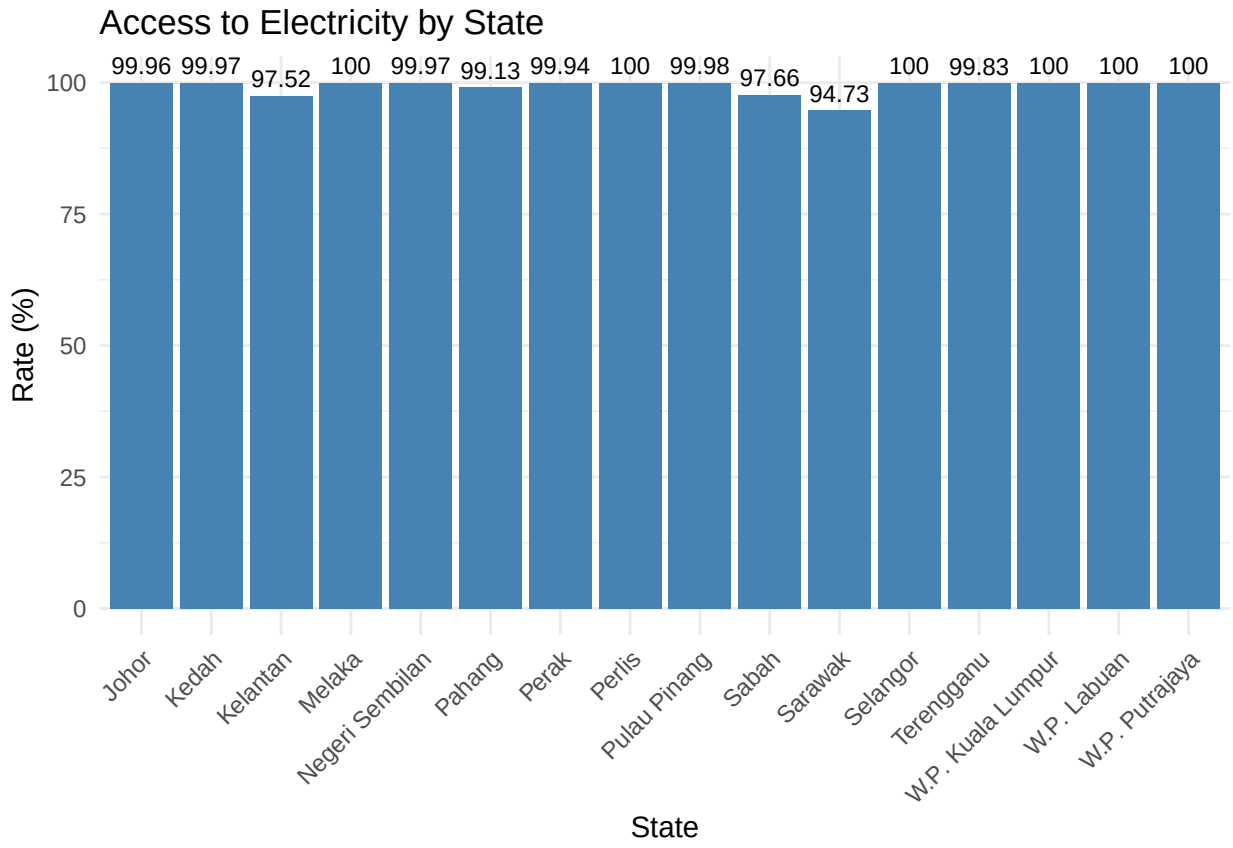
3.2.8 AMENITIES (WATER, SANITATION, ELECTRICITY)

In terms of **access to piped water** in Malaysia, **Kelantan** has the **lowest rate at 72.67%, followed by Sabah** at **81.67%** and **Sarawak** at **76.96%**. This disparity highlights the significant regional challenges in providing reliable infrastructure across the country. **Kelantan's lower rate** can be attributed to its **geographical terrain, rural nature, and historical underinvestment in water distribution systems. Sabah and Sarawak**, while having better access rates, **still face logistical hurdles due to their vast and often remote areas.** These figures reflect the **ongoing need for targeted infrastructure development** in these states to **improve access to essential amenities** and **reduce disparities in living standards.**

```
# Group by state and sum the specified column
grouped_data_sanitation <- data_2 %>%
  group_by(state) %>%
  summarise(sanitation = sum(sanitation, na.rm = TRUE))

# Plot the grouped data for sanitation
p9 <- ggplot(grouped_data_sanitation, aes(x = state, y = sanitation)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  geom_text(aes(label = round(sanitation, 2)), vjust = -0.5, size = 3) +
  labs(title = "Access to Sanitation by State", x = "State", y = "Rate (%)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Display the plot
print(p9)
```



```r
# Group by state and sum the electricity values
grouped_data_electricity <- data_2 %>%
  group_by(state) %>%
  summarise(electricity = sum(electricity, na.rm = TRUE))

# Plot the grouped data for electricity
p10 <- ggplot(grouped_data_electricity, aes(x = state, y = electricity)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  geom_text(aes(label = round(electricity, 2)), vjust = -0.5, size = 3) +
  labs(title = "Access to Electricity by State", x = "State", y = "Rate (%)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Display the plot
print(p10)
```



In terms of **access to sanitation and electricity** in Malaysia, the data indicates that **all states generally have sufficient coverage for both amenities**. However, **further analysis is needed at the district level to better understand the disparities in access within specific regions**. While national averages suggest adequate infrastructure, **there may be localized gaps, particularly in rural or remote areas, where access to sanitation and electricity can vary significantly.** A more **detailed examination at the district level** would provide a clearer picture of the accessibility and quality of these essential services, helping to identify areas that may require targeted improvements or investments to ensure equitable access for all communities.

4.0 CORRELATION ANALYSIS

In this study, the correlation between various independent variables such as average income, average expenditure, sanitation, piped water, electricity, Gini index, living quarters, population, and households and the dependent variable which is poverty rate was examined using both Pearson and Spearman correlation matrices.

```
# Load necessary libraries
library(corrplot)
```

```
## corrplot 0.95 loaded
```

```
library(dplyr)

data_3 <- data %>%
  select(state, population, piped_water, sanitation, electricity,
         income_mean, expenditure_mean, poverty,
         gini, households, living_quarters, planted_area, production)

# Ensure all relevant columns are numeric (while avoiding issues with non-numeric columns)
data_3 <- data_3 %>%mutate(across(everything(), ~ if (is.numeric(.)) . else as.numeric(as.character(.))
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `across(...)`.
## Caused by warning:
## ! NAs introduced by coercion
```

```
# Define target and predictor variables
target_variable <- "poverty"
predictor_variables <- c("population", "piped_water", "sanitation", "electricity",
                         "income_mean", "expenditure_mean", "gini", "households", "living_quarters")

# Function to calculate Pearson and Spearman correlations
calculate_correlations <- function(data, predictors, target) {
  results <- data.frame(Variable = character(),
                        Pearson_Correlation = numeric(),
                        Pearson_p_value = numeric(),
                        Spearman_Correlation = numeric(),
                        Spearman_p_value = numeric(),
                        stringsAsFactors = FALSE)

  for (predictor in predictors) {
    if (predictor %in% colnames(data) && target %in% colnames(data)) {
      try({
        # Pearson Correlation
        pearson_corr <- cor(data[[predictor]], data[[target]], method = "pearson", use = "complete.obs"
        pearson_p_value <- cor.test(data[[predictor]], data[[target]], method = "pearson")$p.value

        # Spearman Rank Correlation
        spearman_corr <- cor(data[[predictor]], data[[target]], method = "spearman", use = "complete.ob
        spearman_p_value <- cor.test(data[[predictor]], data[[target]], method = "spearman")$p.value
```

```
        results <- rbind(results, data.frame(Variable = predictor,
                                             Pearson_Correlation = pearson_corr,
                                             Pearson_p_value = pearson_p_value,
                                             Spearman_Correlation = spearman_corr,
                                             Spearman_p_value = spearman_p_value))
      }, silent = TRUE)
    }
  }
  return(results)
}

# Calculate correlations
correlation_results <- calculate_correlations(data_3, predictor_variables, target_variable)
```

```
## Warning in cor.test.default(data[[predictor]], data[[target]], method =
## "spearman"): Cannot compute exact p-value with ties
```

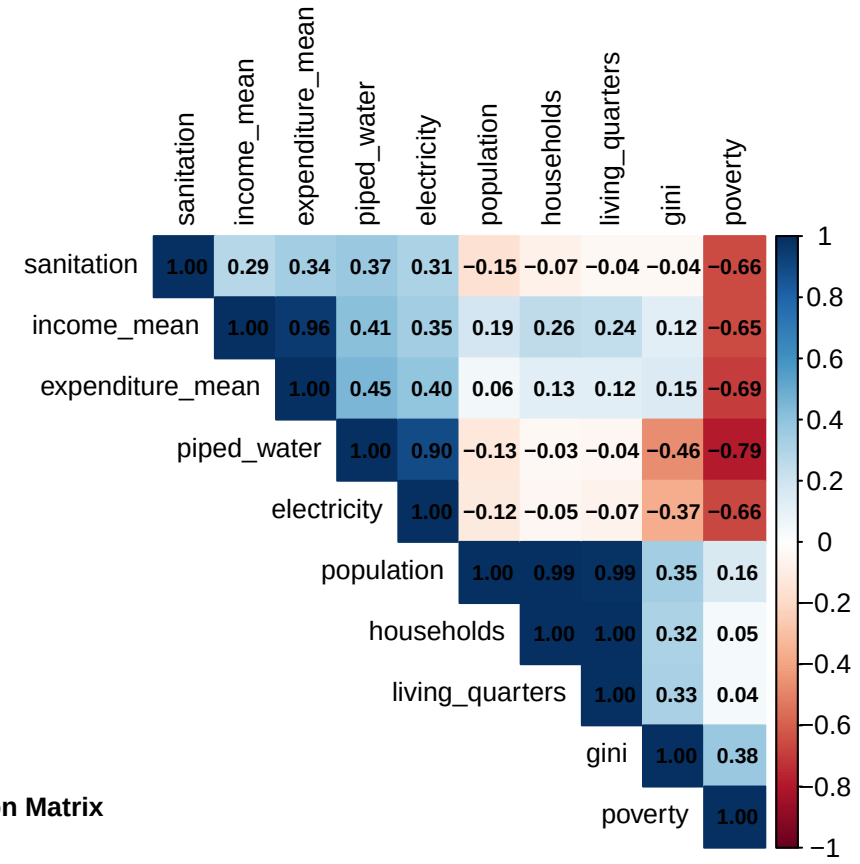
```
## Warning in cor.test.default(data[[predictor]], data[[target]], method =
## "spearman"): Cannot compute exact p-value with ties
## Warning in cor.test.default(data[[predictor]], data[[target]], method =
## "spearman"): Cannot compute exact p-value with ties
```

```
# Display correlation results
print(correlation_results)
```

```
##           Variable Pearson_Correlation Pearson_p_value Spearman_Correlation
## 1       population          0.16093705    0.5515550394            0.3911765
## 2      piped_water         -0.78871035    0.0002823816           -0.8667175
## 3       sanitation         -0.65549486    0.0058397406           -0.6579268
## 4      electricity         -0.66479139    0.0049598077           -0.8908769
## 5      income_mean         -0.65195401    0.0062059277           -0.7852941
## 6 expenditure_mean         -0.69499469    0.0028042517           -0.8323529
## 7             gini          0.37745368    0.1494987227            0.2470588
## 8       households          0.04752911    0.8612429357            0.3147059
## 9  living_quarters          0.04131076    0.8792637885            0.2882353
##   Spearman_p_value
## 1     1.350429e-01
## 2     1.398284e-05
## 3     5.598383e-03
## 4     3.685926e-06
## 5     4.891193e-04
## 6     6.562278e-05
## 7     3.549965e-01
## 8     2.347116e-01
## 9     2.781260e-01
```

```
# Visualize Pearson correlation matrix using a heatmap
if (all(c(target_variable, predictor_variables) %in% colnames(data_3))) {
  correlation_matrix <- cor(data_3[c(predictor_variables, target_variable)], method = "pearson", use =
  corrplot(correlation_matrix, method = "color", type = "upper", order = "hclust",
           addCoef.col = "black", number.cex = 0.7, tl.cex = 0.8, tl.col = "black")
  # Add bold title
```

```
  mtext("Pearson Correlation Matrix", side = 1, line = 3, adj = 0, col = "black", cex = 0.8, font = 2)
}
```



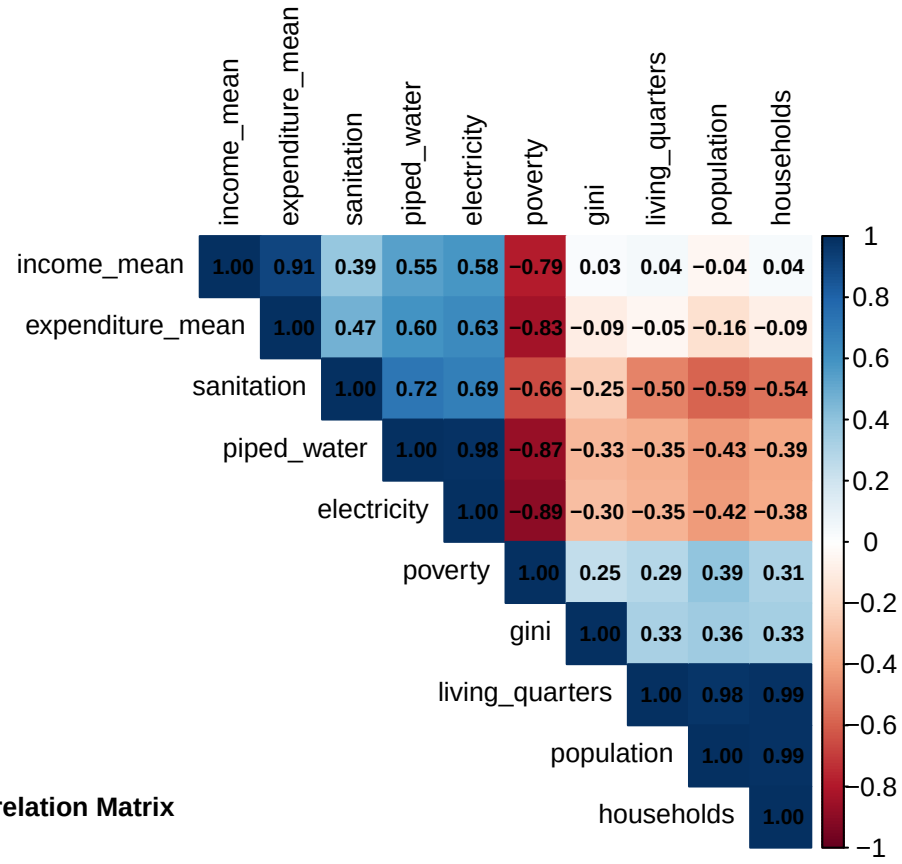
Pearson Correlation Matrix

4.1 PEARSON CORRELATION ANALYSIS

The **Pearson correlation matrix** revealed that **Gini index (0.38), living quarters (0.04), population (0.16), and households (0.05)** exhibited **weak positive correlations with poverty rate**. This suggests that **while there is a slight positive association, the strength of these relationships is not significant enough to imply strong predictive power**. The **weak correlations may be influenced by complex socio-economic factors or regional variations not captured by these variables alone.**

On the other hand, **average income (-0.65), average expenditure (-0.69), sanitation (-0.66), piped water (-0.79), and electricity (-0.66) showed strong negative correlations with poverty rate**. These results **support the notion that improvements in income, expenditure, access to sanitation, piped water, and electricity are strongly associated with a reduction in poverty**, highlighting their role as **critical determinants of socio-economic well-being**.

```
# Visualize Spearman correlation matrix using a heatmap
if (all(c(target_variable, predictor_variables) %in% colnames(data_3))) {
  spearman_matrix <- cor(data_3[c(predictor_variables, target_variable)], method = "spearman", use = "c
  corrplot(spearman_matrix, method = "color", type = "upper", order = "hclust",
           addCoef.col = "black", number.cex = 0.7, tl.cex = 0.8, tl.col = "black")
  # Add bold title
  mtext("Spearman Correlation Matrix", side = 1, line = 3, adj = 0, col = "black", cex = 0.8, font = 2)
}
```



Spearman Correlation Matrix

4.2 SPEARMAN RANK CORRELATION ANALYSIS

In contrast, the **Spearman correlation matrix** indicated somewhat **weaker positive correlations for Gini index (0.25), living quarters (0.29), population (0.39), and households (0.31)**, suggesting a **modest, non-linear relationship with poverty rate**. However, the **negative correlations for average income (-0.79), average expenditure (-0.83), sanitation (-0.66), piped water (-0.87), and electricity (-0.89) were even stronger in this case**.

This emphasizes the **consistent and robust inverse relationship between these variables and poverty, reinforcing the importance of access to resources and higher income levels in alleviating poverty**. The stronger correlations in the Spearman matrix suggest that the relationships between these variables and poverty may be better captured through non-linear associations, potentially revealing more complex dynamics not evident in the linear Pearson correlation analysis.

Overall, **both correlation matrices** highlight the **importance of economic and infrastructural factors in addressing poverty**, with **access to basic services and higher income being key indicators of lower poverty levels.** However, **the differences in correlation strengths between the two methods suggest that further analysis**, **potentially including more advanced statistical techniques**, may be required to fully capture the nuances of these relationships.

```
# Display sorted correlation results (optional)
correlation_results_sorted <- correlation_results %>%
  arrange(desc(Pearson_Correlation))

print("Correlation Results (Sorted by Pearson Correlation):")
```

```
## [1] "Correlation Results (Sorted by Pearson Correlation):"
```

```
print(correlation_results_sorted)
```

```
##           Variable Pearson_Correlation Pearson_p_value Spearman_Correlation
## 1             gini          0.37745368    0.1494987227            0.2470588
## 2       population          0.16093705    0.5515550394            0.3911765
## 3       households          0.04752911    0.8612429357            0.3147059
## 4  living_quarters          0.04131076    0.8792637885            0.2882353
## 5      income_mean         -0.65195401    0.0062059277           -0.7852941
## 6       sanitation         -0.65549486    0.0058397406           -0.6579268
## 7      electricity         -0.66479139    0.0049598077           -0.8908769
## 8 expenditure_mean         -0.69499469    0.0028042517           -0.8323529
## 9      piped_water         -0.78871035    0.0002823816           -0.8667175
##   Spearman_p_value
## 1     3.549965e-01
## 2     1.350429e-01
## 3     2.347116e-01
## 4     2.781260e-01
## 5     4.891193e-04
## 6     5.598383e-03
## 7     3.685926e-06
## 8     6.562278e-05
## 9     1.398284e-05
```

4.3 SIGNIFICANCE OF VARIABLES

The correlation analysis revealed that p-value of **average income (0.0062), average expenditure (0.0028), and access to basic amenities such as sanitation (0.0058), electricity (0.0049), and piped water (0.00028)** have **p-values below the 0.05** significance level. This indicates that these variables are **statistically significant** in **relation to the poverty rate** in Malaysia.

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

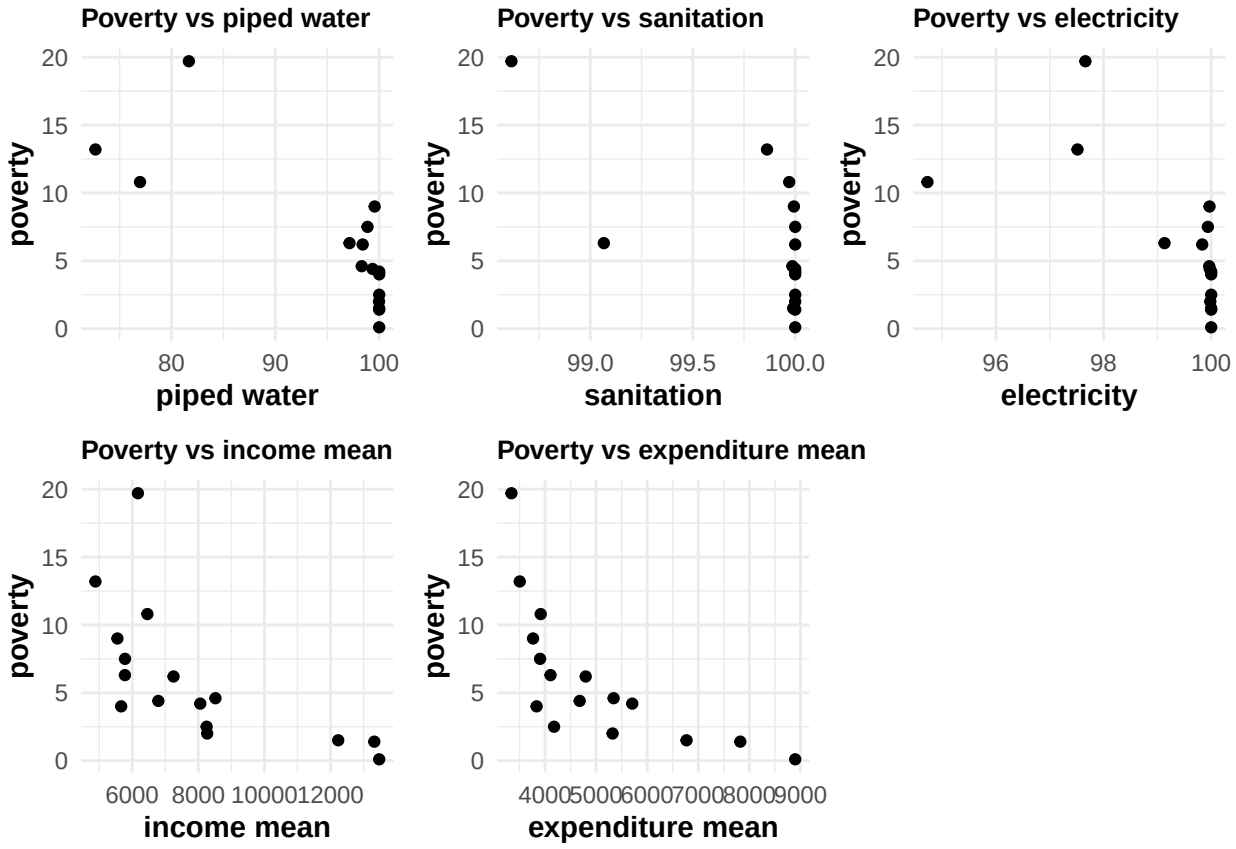
```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
sig_var <- data[, c('piped_water', 'sanitation', 'electricity', 'income_mean', 'expenditure_mean', 'pov

# List of variables to plot
variables <- c('piped_water', 'sanitation', 'electricity', 'income_mean', 'expenditure_mean')

# Create a list to store the individual plots
plots <- list()

# Loop through the variables and create scatter plots
for (var in variables) {
```

```
  plot <- ggplot(sig_var, aes_string(x = var, y = 'poverty')) +
    geom_point() +
   labs(x = gsub("_", " ", tolower(var)), y = "poverty",
         title = paste("Poverty vs", gsub("_", " ", tolower(var)))) +
    theme_minimal() +
    theme(axis.title = element_text(face = "bold"),
          plot.title = element_text(face = "bold", size = 10))

  # Add the plot to the list
  plots[[var]] <- plot
}
```

```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with `aes()`.
## i See also `vignette("ggplot2-in-packages")` for more information.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
# Combine all the plots in a 2x3 grid
grid.arrange(plots[['piped_water']], plots[['sanitation']], plots[['electricity']],
             plots[['income_mean']], plots[['expenditure_mean']], ncol = 3, nrow = 2)
```



5.0 SCATTERPLOT ANALYSIS

Based on the scatter plot shown in the figure above, it is evident that **average income** and **average expenditure exhibit significant trends**, where **higher levels of average income or expenditure are associated with lower poverty rates.** This inverse relationship suggests that **as individuals' income or expenditure increases, they are less likely to fall below the poverty threshold**. Higher income enables **better access to essential goods and services, improving living standards and reducing the likelihood of poverty.** Similarly, **increased expenditure, which often correlates with higher consumption of basic needs and services, further supports the reduction of poverty.** These trends highlight the **critical role of economic well-being in mitigating poverty**, emphasizing that **enhancing income and expenditure levels** can lead to **substantial improvements in socio-economic conditions**.

6.0 SPATIAL ANALYSIS

Mapping the poverty rate, expenditure, average income, sanitation, piped water, and electricity access by state in Malaysia was conducted to provide a clearer demographic visualization of the regional disparities across the country. This spatial analysis allows for a more better understanding of how these key socio-economic factors are distributed geographically, helping to identify areas where improvements are needed most. By visualizing these variables at the state level, we can better assess the effectiveness of existing policies and initiatives aimed at reducing poverty and improving living conditions. The mapping also serves as a tool for policymakers, enabling targeted interventions that address specific regional challenges related to income, infrastructure, and access to basic services.

```
# Load necessary libraries
library(sf)
```

```
## Linking to GEOS 3.12.2, GDAL 3.9.3, PROJ 9.4.1; sf_use_s2() is TRUE
```

```
library(ggplot2)
library(dplyr)
library(viridis)
```

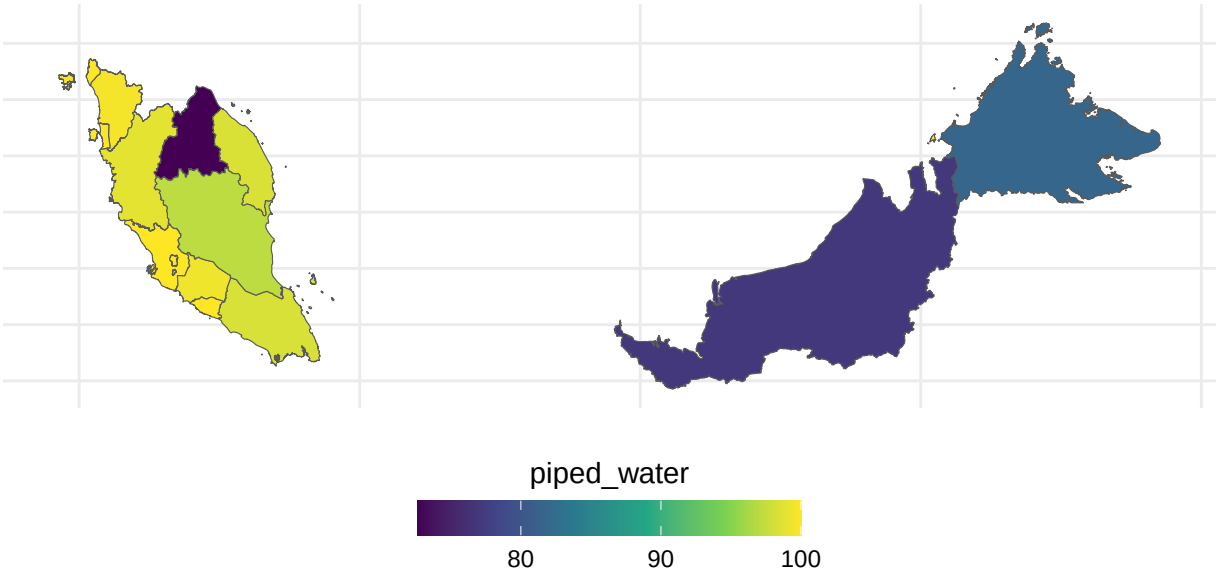
```
## Loading required package: viridisLite
```

```
# Load shapefile
shapefile_path <- "C:/Users/User/OneDrive/Desktop/Shape File Malaysia/malaysia state v2.shp"
gdf <- st_read(shapefile_path)
```

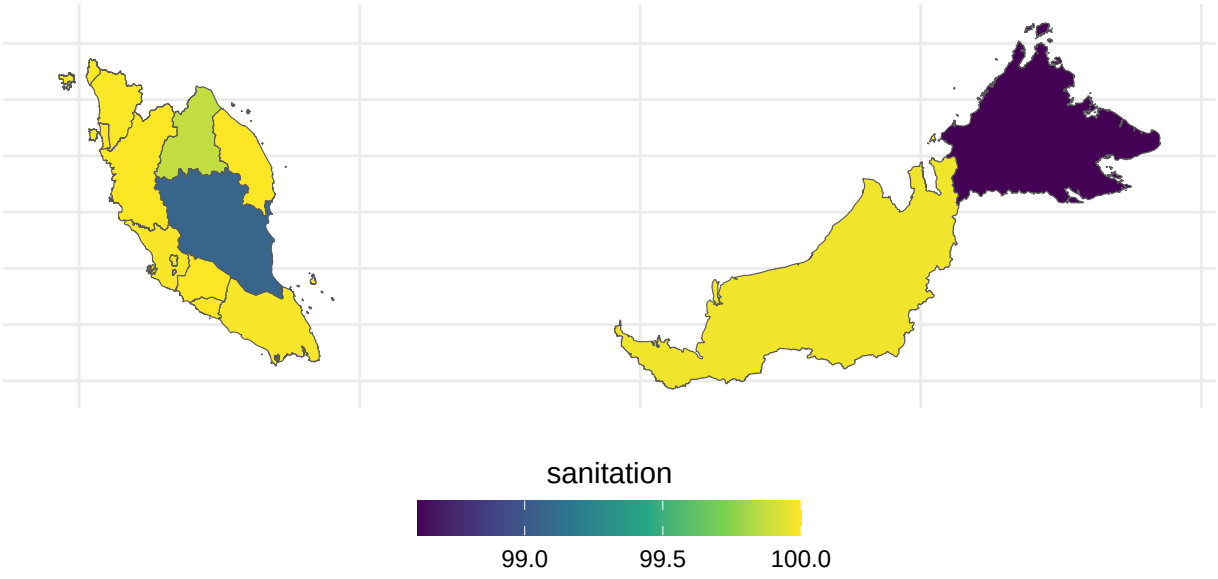
```
## Reading layer 'malaysia state v2' from data source
##   'C:\Users\User\OneDrive\Desktop\Shape File Malaysia\malaysia state v2.shp'
##   using driver 'ESRI Shapefile'
## Simple feature collection with 16 features and 6 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: 99.6405 ymin: 0.853821 xmax: 119.2691 ymax: 7.362818
## Geodetic CRS:  WGS 84
```

```r
# Transform to EPSG:4326
gdf <- st_transform(gdf, crs = 4326)

df <- data %>%
  select(state, piped_water, sanitation, electricity, income_mean, expenditure_mean, poverty)

# Merge geospatial data with poverty data
merged_gdf <- left_join(gdf, df, by = "state")

# List of variables to plot
variables <- c('piped_water', 'sanitation', 'electricity', 'income_mean', 'expenditure_mean', 'poverty'
titles <- c('Piped Water (%)', 'Sanitation (%)', 'Electricity (%)', 'Average Income (RM)', 'Average Exp

# Set up the plotting layout to have 3 rows and 2 columns
par(mfrow = c(3, 2), mar = c(4, 4, 3, 2))  # Adjust margins for better spacing

# Create a plot for each variable
for (i in seq_along(variables)) {
  plot <- ggplot(data = merged_gdf) +
    geom_sf(aes(fill = .data[[variables[i]]])) +
    scale_fill_viridis_c(option = "D") + # Adjust color palette as needed
    guides(fill = guide_colorbar(
      direction = "horizontal",
      title.position = "top",
      title.hjust = 0.5,
      barwidth = 10, # Adjust width of the color bar
      barheight = 0.9 # Adjust height of the color bar
    )) +
    labs(title = paste(titles[i], "by State")) +
    theme_minimal() +
    theme(axis.text = element_blank(),
          axis.ticks = element_blank(),
          plot.title = element_text(face = "bold", color = "darkblue", size = 10),
          legend.position = "bottom")

  # Print each plot explicitly
  print(plot)
}
```

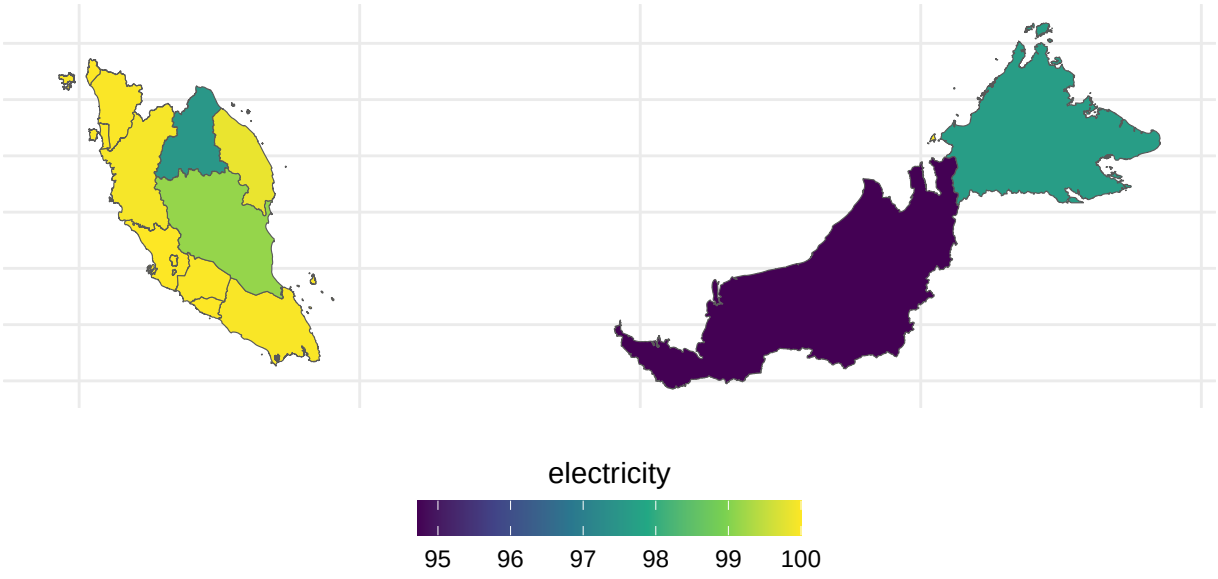
Piped Water (%) by State



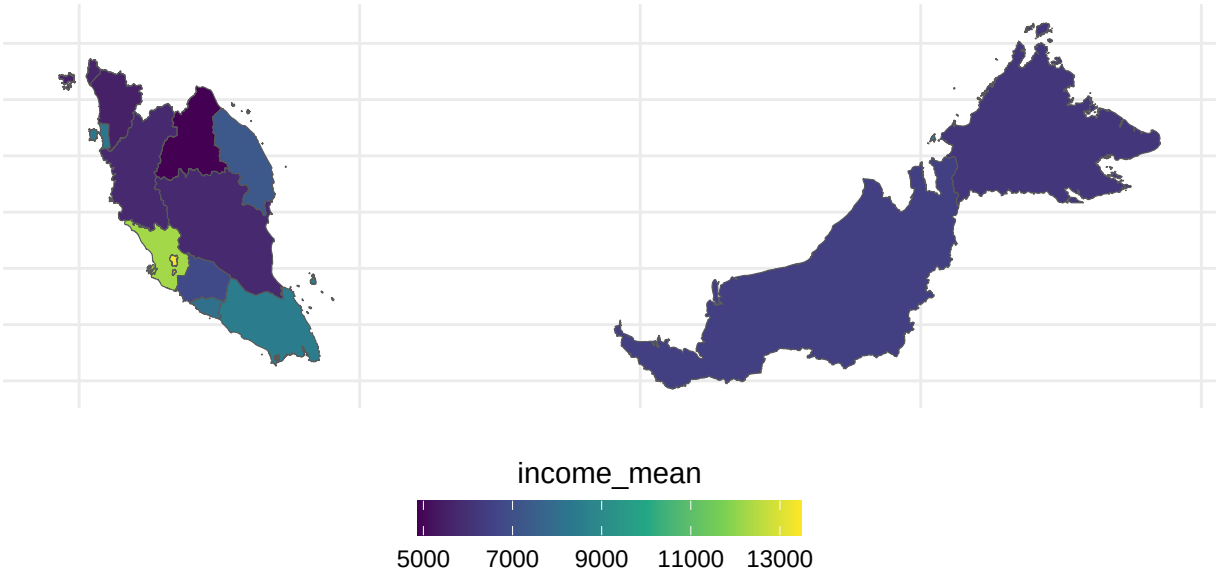
Sanitation (%) by State



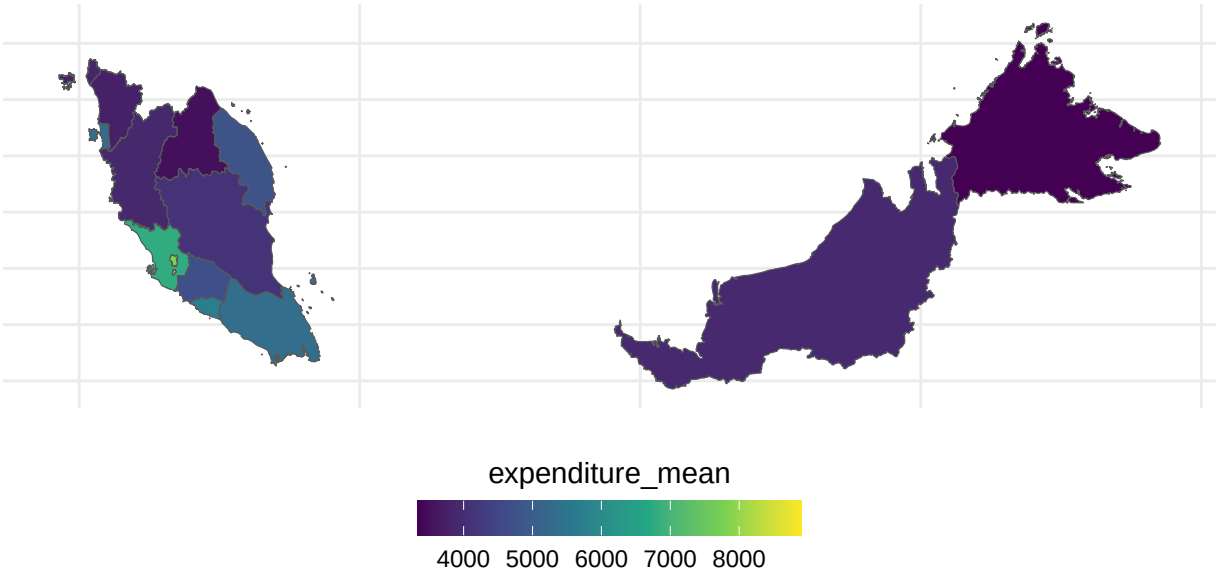
Electricity (%) by State

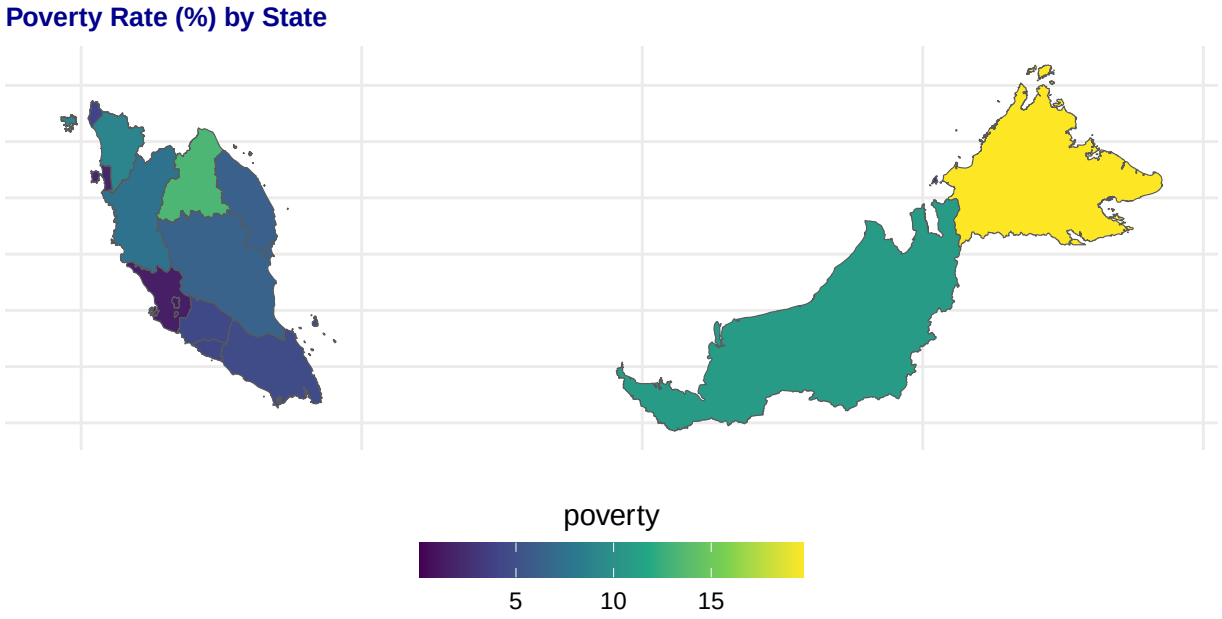


Average Income (RM) by State



Average Expenditure (RM) by State





7.0 CONCLUSION

The analysis of socio-economic and infrastructural factors in Malaysia based on data from 2022 has provided valuable insights into the relationship between poverty and variables such as population, income, expenditure, basic amenities, and agricultural activity. Key findings from the descriptive statistics and correlation analysis highlight the significant role of economic well-being and access to essential services in reducing poverty rates across states.

The Pearson and Spearman correlation matrices consistently showed **strong negative correlations between poverty and variables such as average income, average expenditure, and access to basic amenities (sanitation, piped water, and electricity), with p-values below the 0.05 significance level.** These results indicate that these factors are statistically significant predictors of poverty. On the other hand, variables like Gini index, living quarters, population, and households demonstrated weak or moderate positive correlations with poverty, suggesting limited direct influence on poverty rates.

Given the statistical evidence, **we reject the null hypothesis (H_0)**, which posits that there is no relationship between poverty and the studied variables. Instead, **the results support the alternative hypothesis (H_1), confirming that there is a relationship between poverty and factors such as income, expenditure, and access to basic amenities**.

The findings underscore the importance of improving economic conditions and infrastructure to address poverty effectively. Policymakers should focus on targeted interventions, especially in states with high poverty rates like Sabah and Kelantan, by prioritizing investments in income-generating activities, access to essential services, and equitable resource distribution. Spatial analysis further reinforces the need for region-specific strategies to ensure inclusive and sustainable development across Malaysia.