Final Report - Group 01

HDI and Broadband - Data Analysis

Amie Leung, Ivan Nguyen, Janna Cameron, Kaye Ann Caroningan, and Syed Hassan

Nov 29, 2025

Abstract

This report analyzes the relationship between Human Development Index (HDI) and broadband access across various countries, highlighting trends and insights from the data. Lorem ipsum dolor sit amet, consectetur adipiscing elit. In egestas lobortis porta. Phasellus tincidunt metus sed mollis imperdiet. Aliquam blandit nibh in fermentum consequat. Fusce blandit magna quis nulla ornare, non mattis enim tempor. Nulla et odio dui. Nunc ut elit venenatis, porttitor dui et, placerat tortor. Sed aliquet sodales magna, sit amet porta odio tempor in. CHANGE THIS AT THE END

Contents

1. Motivation

Lorem ipsum dolor sit amet, consectetur adipiscing elit. In egestas lobortis porta. Phasellus tincidunt metus sed mollis imperdiet. Aliquam blandit nibh in fermentum consequat. Fusce blandit magna quis nulla ornare, non mattis enim tempor. Nulla et odio dui. Nunc ut elit venenatis, porttitor dui et, placerat tortor. Sed aliquet sodales magna, sit amet porta odio tempor in.

Sed eget luctus tellus. Vestibulum dui orci, commodo a tincidunt sed, aliquet sed elit. Aenean mauris dolor, luctus sed sapien ut, pellentesque ornare nibh. Phasellus velit mauris, interdum eu massa id, pretium

condimentum nisi. Sed at pellentesque mi. Vivamus id justo non elit lacinia efficitur. Sed commodo vehicula nisi, a tristique lectus cursus at. Nullam finibus elementum justo, a molestie augue iaculis sit amet. Praesent quis ante sit amet arcu condimentum commodo quis eu lectus. Curabitur pellentesque mattis enim, eu elementum lectus sollicitudin pellentesque. CHANGE THIS OF COURSE

2. Review of Similar Research

Lorem ipsum dolor sit amet, consectetur adipiscing elit. In egestas lobortis porta. Phasellus tincidunt metus sed mollis imperdiet. Aliquam blandit nibh in fermentum consequat. Fusce blandit magna quis nulla ornare, non mattis enim tempor. Nulla et odio dui. Nunc ut elit venenatis, porttitor dui et, placerat tortor. Sed aliquet sodales magna, sit amet porta odio tempor in.

Sed eget luctus tellus. Vestibulum dui orci, commodo a tincidunt sed, aliquet sed elit. Aenean mauris dolor, luctus sed sapien ut, pellentesque ornare nibh. Phasellus velit mauris, interdum eu massa id, pretium condimentum nisi. Sed at pellentesque mi. Vivamus id justo non elit lacinia efficitur. Sed commodo vehicula nisi, a tristique lectus cursus at. Nullam finibus elementum justo, a molestie augue iaculis sit amet. Praesent quis ante sit amet arcu condimentum commodo quis eu lectus. Curabitur pellentesque mattis enim, eu elementum lectus sollicitudin pellentesque. CHANGE THIS OF COURSE

3. Research Questions

3.1. Question 1

Lorem ipsum dolor sit amet, consectetur adipiscing elit. In egestas lobortis porta. Phasellus tincidunt metus sed mollis imperdiet. Aliquam blandit nibh in fermentum consequat. Fusce blandit magna quis nulla ornare, non mattis enim tempor. Nulla et odio dui. Nunc ut elit venenatis, porttitor dui et, placerat tortor. Sed aliquet sodales magna, sit amet porta odio tempor in. CHANGE THIS OF COURSE

3.2. Question 2

Lorem ipsum dolor sit amet, consectetur adipiscing elit. In egestas lobortis porta. Phasellus tincidunt metus sed mollis imperdiet. Aliquam blandit nibh in fermentum consequat. Fusce blandit magna quis nulla ornare, non mattis enim tempor. Nulla et odio dui. Nunc ut elit venenatis, porttitor dui et, placerat tortor. Sed aliquet sodales magna, sit amet porta odio tempor in. CHANGE THIS OF COURSE

3.3. Question 3

Lorem ipsum dolor sit amet, consectetur adipiscing elit. In egestas lobortis porta. Phasellus tincidunt metus sed mollis imperdiet. Aliquam blandit nibh in fermentum consequat. Fusce blandit magna quis nulla ornare, non mattis enim tempor. Nulla et odio dui. Nunc ut elit venenatis, porttitor dui et, placerat tortor. Sed aliquet sodales magna, sit amet porta odio tempor in. CHANGE THIS OF COURSE # 4. Data {#sec-data}

We are using three files from the Gapminder dataset:
HDI, Broadband, and GDP. (*Gapminder Datasets* 2023)

4.1 Loading the files

```r
hdi <- read.csv("data/hdi_human_development_index.csv")
broadband <- read.csv("data/broadband_subscribers_per_100_people.csv")
gdp <- read.csv("data/gdp_pcap.csv")
```

4.2. Data Structure

In this section we will clean the header names of each dataset and explore and describe the datasets one by one.

1. Human Development Index (HDI): Based on data from the United Nations Development Programme website, this dataset contains the HDI scores for each country between the years 1990 and 2023. There are 193 rows and 35 columns. The data contains some missing values for certain periods.

Let's explore the shape of the dataset.

Rows x Columns

```
## [1] 193  35
```

Since there are 35 columns, let's only see the first and last five columns to see more detail of column names (after cleaning).

```
## [1] "country" "x1990"   "x1991"   "x1992"   "x1993"
```

```
## [1] "x2019" "x2020" "x2021" "x2022" "x2023"
```

The year range of this dataset is 1990 - 2023. We can also see that the column names have "x" characters that need to be cleaned further.

2. Broadband subscribers (per 100 people): This dataset contains the number of fixed broadband subscriptions per 100 people between the years 1998 and 2023. Included in this measurement are all wired connections that provide high-speed internet access (>= 256 kbits/s), covering both residential and organizational subscriptions but excluding mobile-cellular networks. There are 193 rows and 28 columns. The data contains some missing values, mainly concentrated in the earlier years.

Rows x Columns

```
## [1] 206  27
```

And the first and last five column names:

```
## [1] "country" "x1998"   "x1999"   "x2000"   "x2001"
```

```
## [1] "x2019" "x2020" "x2021" "x2022" "x2023"
```

The year range of this dataset is 1998 - 2023. Again, the column names have "x"s that aren't helpful. Several columns have chr data type, which we will need to clean up.

3. Gross Domestic Product (GDP): Adjusted for inflation and cost-of-living differences, this dataset measures the value of goods and services produced per person for each country, compiling data from the early 19th century to the present year. There are 194 rows and 303 columns.

Rows x Columns

```
## [1] 193 303
```

And the first and last five column names:

```
## [1] "geo"   "name"  "x1800" "x1801" "x1802"
```

```
## [1] "x2096" "x2097" "x2098" "x2099" "x2100"
```

This dataset has the most number of years ranging from 1800 - 2100. The future year values might be predicted or empty which we can analyze further when standardizing and cleaning rows. The *geo* column is also interesting as after standardizing the country names, and merging the datasets, we can also map the data if needed.

Please note: all datasets were sourced from the Gapminder repository. Data from the year 2022 coincides with the tail end of the COVID-19 pandemic; however, for the purposes of this report, we will not account for its potential impacts on GDP and HDI, particularly with regard to health.

4.3. Selecting the required columns

Since the dataset has values from 100+ years, we will narrow them down to the most recent decade starting from 2010 to 2023. Will also clean the years by removing the 'x' character before we select the columns of interest ("X2010" -> 2010).

HDI

```
##       country  2011  2012  2013  2014  2015  2016  2017  2018  2019  2020  2021
## 1 Afghanistan 0.474 0.484 0.492 0.497 0.496 0.495 0.496 0.498 0.507 0.501 0.486
## 2      Angola 0.545 0.557 0.567 0.577 0.603 0.609 0.610 0.611 0.611 0.610 0.609
## 3     Albania 0.781 0.790 0.793 0.797 0.797 0.797 0.798 0.801 0.805 0.794 0.794
##    2022  2023
## 1 0.495 0.496
## 2 0.615 0.616
## 3 0.806 0.810
```

Need to talk about the values. WHAT DO THE VALUES MEAN

Broadband

```
##       country   2011    2012     2013     2014    2015   2016   2017   2018
## 1       Aruba     NA      NA 18.70000 18.60000 18.2000     NA     NA     NA
## 2 Afghanistan     NA 0.00491  0.00474  0.00457  0.0209 0.0254 0.0257 0.0435
## 3      Angola 0.0653 0.08150  0.08520  0.32300  0.5450 0.2900 0.3210 0.3500
##     2019   2020    2021    2022   2023
## 1 17.700 17.700 17.7000 17.5000     NA
## 2  0.052  0.068  0.0664  0.0796 0.0801
## 3  0.368  0.363  0.3910  0.3870 0.3740
```

Need to talk about the values. WHAT DO THE VALUES MEAN

GDP

```
##       country      2011      2012      2013      2014      2015      2016
## 1 Afghanistan  1962.057  2123.871  2166.402  2145.500  2109.747  2102.452
## 2      Angola  7663.286  8011.050  8099.679  8183.165  7966.886  7487.925
## 3     Albania 11052.778 11227.950 11361.252 11586.817 11878.438 12291.842
##        2017      2018      2019      2020      2021      2022      2023
## 1  2097.120  2061.709  2080.941  1969.306  1517.016  1386.755  1359.020
## 2  7216.061  6878.590  6602.269  6029.692  5911.836  5906.116  5778.834
## 3 12770.992 13317.119 13653.182 13278.370 14595.944 15491.992 16209.877
```

Need to talk about the values. WHAT DO THE VALUES MEAN

4.4. Standardizing country names - rows

Since the datasets might have countries that are not listed in the others, we need to make sure that we standardize all the country names to match one of the datasets as our source. We will take the HDI dataset as our source, and compare the names, and check for the difference in rows/country names. If we find any difference, we will make sure to either match those rows to the source dataset country names, and also remove territories and records that are not common amongst all.

```
## [1] "Number of rows in HDI, Broadband, and GDP datasets:"
```

```
## HDI:  193 , Broadband:  206 , GDP:  193
```

```
## [1] "Countries in Broadband but NOT in HDI:"
```

```
##  [1] "Aruba"                  "Bermuda"                "Cayman Islands"
##  [4] "Faeroe Islands"         "Gibraltar"              "Greenland"
##  [7] "Guam"                   "Macao, China"           "Monaco"
## [10] "New Caledonia"          "Curaçao"                "French Polynesia"
## [13] "British Virgin Islands"
```

```
## [1] "Countries in GDP but NOT in HDI:"
```

```
## [1] "Monaco"      "North Korea"
```

We can see some diffences in countries within each dataset. We will now standardized and filter the data to only focus on countries that are common amongst all three datasets.

After standardizing:

```
## Number of rows in HDI dataset: 191
```

```
## Number of rows in Broadband dataset: 191
```

```
## Number of rows in GDP dataset: 191
```

Now let's check for duplicates.

4.5. Handling duplicates (if any):

```
## 0 rows removed.
```

```
## 0 rows removed.
```

```
## 0 rows removed.
```

Now that we have standardized names, and checked for duplicates, let's explore the missing values.

4.6. Handling missing values - rows

```
## [1] "Empty values in HDI"
```

```
## [[1]]
## country    2011    2012    2013    2014    2015    2016    2017    2018    2019
##       0       1       1       1       1       1       1       1       1       1
##    2020    2021    2022    2023
##       1       1       1       0
```

```
## [1] "Empty values in Broadband"
```

```
## [[1]]
## country    2011    2012    2013    2014    2015    2016    2017    2018    2019
##       0      14       7       6       7       5       7       7      20       4
##    2020    2021    2022    2023
##       6       3       2      40
```

```
## [1] "Empty values in GDP"
```

```
## [[1]]
## country    2011    2012    2013    2014    2015    2016    2017    2018    2019
##       0       0       0       0       0       0       0       0       0       0
##    2020    2021    2022    2023
##       0       0       0       0
```

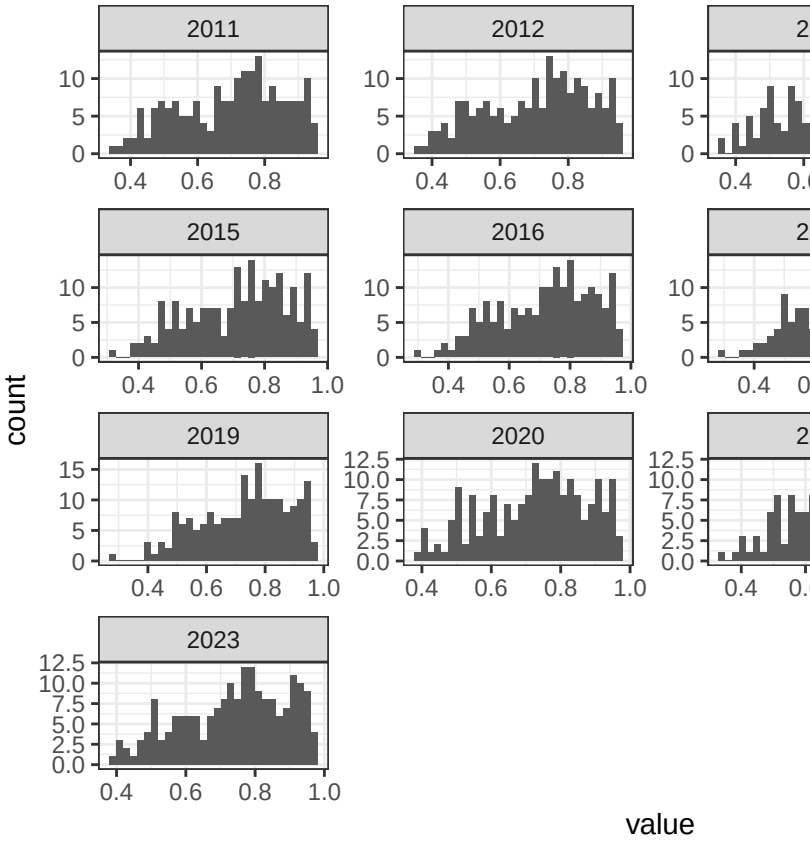
Not dropping any at this point

5. Methodology

Lorem ipsum dolor sit amet, consectetur adipiscing elit. In egestas lobortis porta. Phasellus tincidunt metus sed mollis imperdiet. Aliquam blandit nibh in fermentum consequat. Fusce blandit magna quis nulla ornare, non mattis enim tempor. Nulla et odio dui. Nunc ut elit venenatis, porttitor dui et, placerat tortor. Sed aliquet sodales magna, sit amet porta odio tempor in. CHANGE THIS OF COURSE

5.2. Data Distribution

5.2.1. Distribution of HDI



#SYED: Can we show the 2011 and 2022 values only?
Table: Data summary

Name	hdi_clean
Number of rows	191
Number of columns	14
Column type frequency:	
character	1
numeric	13
Group variables	None

Variable type: character

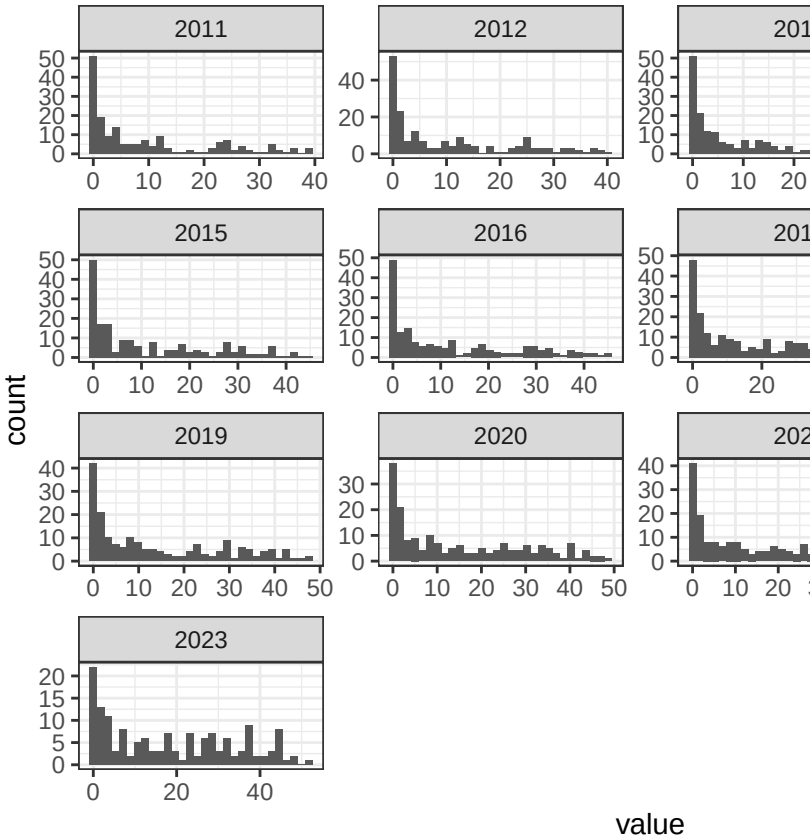
skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
country	0	1	2	30	0	191	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
2011	1	0.99	0.70	0.15	0.35	0.57	0.72	0.82	0.95	
2012	1	0.99	0.71	0.15	0.36	0.58	0.73	0.82	0.95	
2013	1	0.99	0.71	0.15	0.36	0.58	0.74	0.83	0.95	
2014	1	0.99	0.71	0.15	0.32	0.59	0.74	0.83	0.96	
2015	1	0.99	0.72	0.15	0.32	0.60	0.74	0.84	0.96	
2016	1	0.99	0.72	0.15	0.30	0.61	0.75	0.84	0.96	
2017	1	0.99	0.72	0.15	0.29	0.61	0.75	0.84	0.96	
2018	1	0.99	0.73	0.15	0.37	0.61	0.75	0.85	0.97	
2019	1	0.99	0.73	0.15	0.28	0.61	0.75	0.85	0.97	
2020	1	0.99	0.73	0.15	0.39	0.61	0.74	0.84	0.97	
2021	1	0.99	0.73	0.15	0.34	0.61	0.74	0.84	0.97	
2022	1	0.99	0.74	0.15	0.38	0.62	0.76	0.85	0.97	
2023	0	1.00	0.74	0.15	0.39	0.62	0.76	0.86	0.97	

Comment on the distribution

5.2.2. Distribution of Broadband



#SYED: Can we show the 2011 and 2022 values only?
Table: Data summary

Name	broadband_clean
Number of rows	191
Number of columns	14

Column type frequency:	
character	1
numeric	13
Group variables	None

Variable type: character

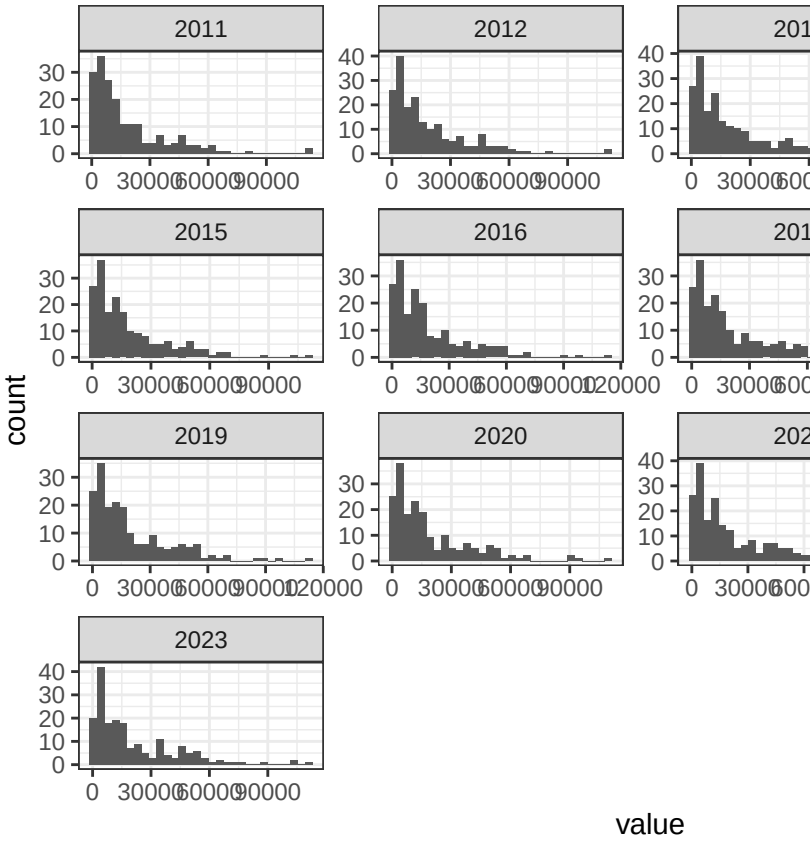
skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
country	0	1	2	30	0	191	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
2011	14	0.93	9.61	11.31	0	0.40	4.26	16.20	38.9	
2012	7	0.96	10.06	11.66	0	0.38	4.73	17.45	40.2	
2013	6	0.97	10.63	12.08	0	0.47	4.94	18.40	42.5	
2014	7	0.96	11.24	12.47	0	0.54	5.62	19.80	43.2	
2015	5	0.97	11.78	12.90	0	0.61	6.31	20.70	44.7	
2016	7	0.96	12.40	13.25	0	0.69	7.11	21.97	45.1	
2017	7	0.96	13.23	13.99	0	0.83	8.14	22.98	62.2	
2018	20	0.90	14.43	14.03	0	1.33	10.00	27.20	47.4	
2019	4	0.98	14.00	14.28	0	1.06	8.85	26.55	47.5	
2020	6	0.97	15.07	14.70	0	1.30	9.40	27.10	48.7	
2021	3	0.98	15.48	14.98	0	1.37	10.60	27.20	50.3	
2022	2	0.99	16.08	15.25	0	1.48	11.00	29.00	51.2	
2023	40	0.79	18.41	15.52	0	2.78	16.10	31.55	51.7	

Comment on the distribution

5.2.3. Distribution of GDP



#SYED: Can we show the 2011 and 2022 values only?
Table: Data summary

Name	gdp_clean
Number of rows	191
Number of columns	14
Column type frequency:	
character	1
numeric	13
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
country	0	1	2	30	0	191	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
2011	0	1	18175.43	19634.09	807.66	3705.53	10933.21	24937.33	111840.3	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
2012	0	1	18359.22	19518.19	506.72	3770.56	11098.89	25689.84	110892.4	
2013	0	1	18472.83	19488.91	624.18	3968.33	11361.25	25872.48	110354.4	
2014	0	1	18700.32	19561.46	614.46	4033.83	11767.16	26482.08	110611.0	
2015	0	1	18979.66	19815.30	594.92	4172.21	12030.38	27189.90	110483.3	
2016	0	1	19226.13	19995.38	501.11	4321.96	12336.90	27857.71	113510.3	
2017	0	1	19559.76	20168.35	458.77	4518.15	12497.82	28710.15	112243.4	
2018	0	1	19916.38	20450.98	435.41	4530.59	13218.92	29622.61	111441.6	
2019	0	1	20184.39	20670.50	425.94	4803.56	13215.57	30351.18	112461.8	
2020	0	1	19053.94	19977.86	386.68	4691.03	12407.79	26943.40	109597.0	
2021	0	1	20164.82	21367.89	395.80	4829.15	13045.93	30234.91	115683.5	
2022	0	1	20922.87	22044.01	364.70	4736.43	13148.07	33000.12	114938.7	
2023	0	1	21088.15	21712.32	354.18	4859.07	13416.93	33417.31	111043.4	

Comment on the distribution

5.2.4. Reshape the data

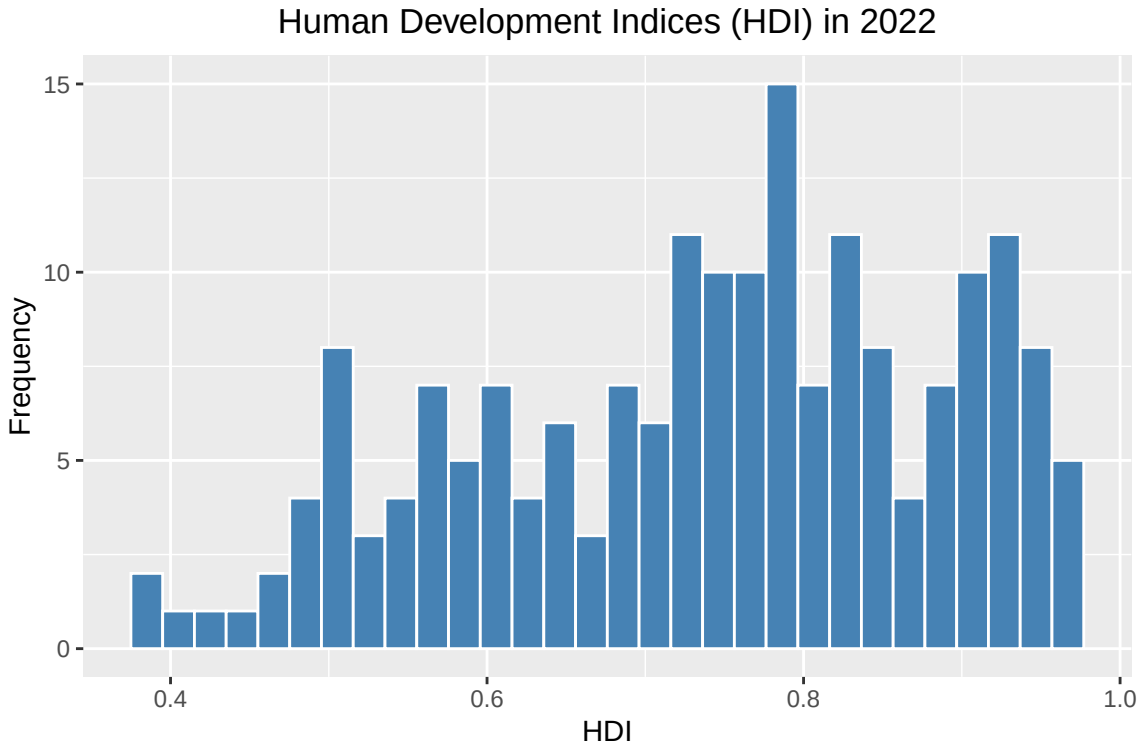
Merge the data

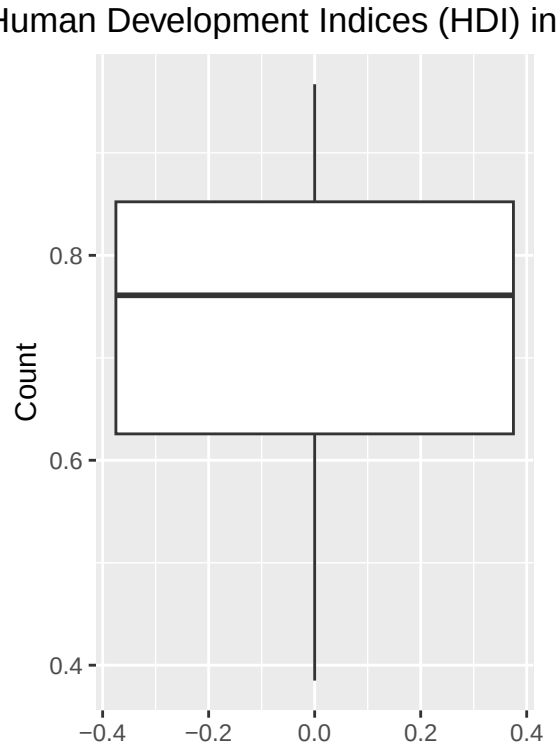
Skewness

Although the original data spanned multiple years, this analysis focuses only on 2011 and 2022. We chose the former column as this coincides with the year that the UN declared internet access a basic human right (United Nations Human Rights Council, 2011). Although 2023 is the most recent year in our dataset, we decided to use 2022 because this year had the fewest missing values. More than 10% of the values for broadband subscriptions in 2023 (49/193) were missing.

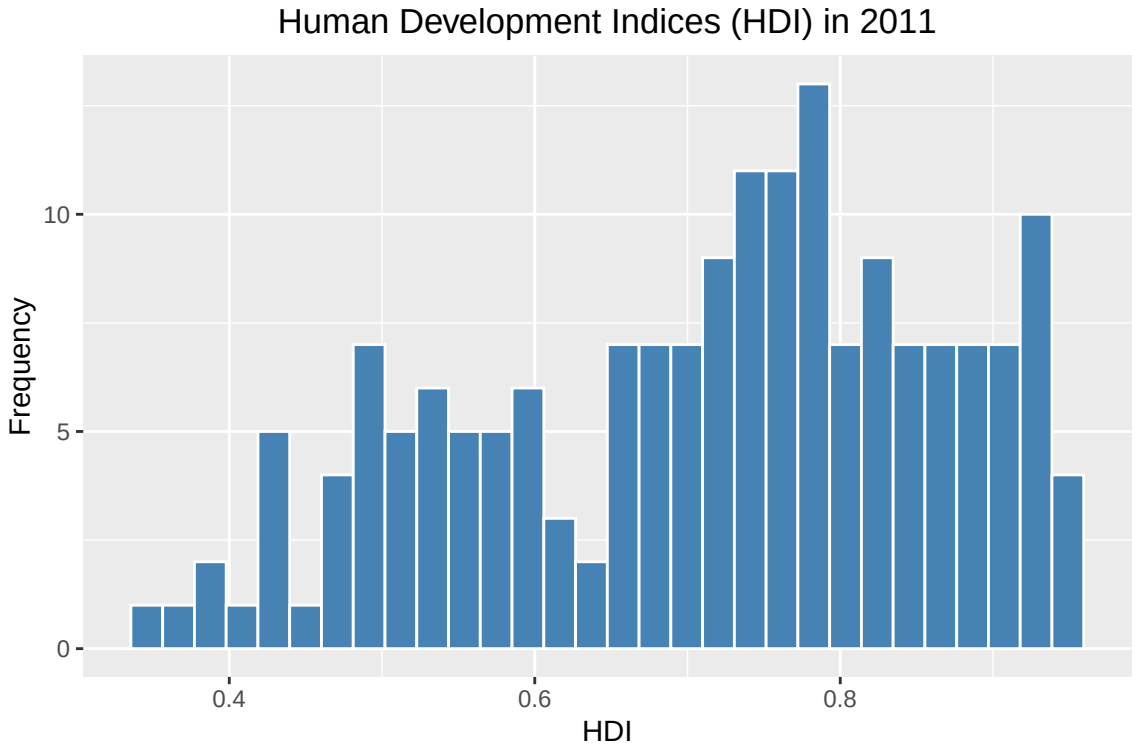
Before modelling, we conducted an exploratory data analysis on the 2011 and 2022 distributions of HDI, broadband subscriptions, and GDP using histograms and boxplots to assess skewness and normality.

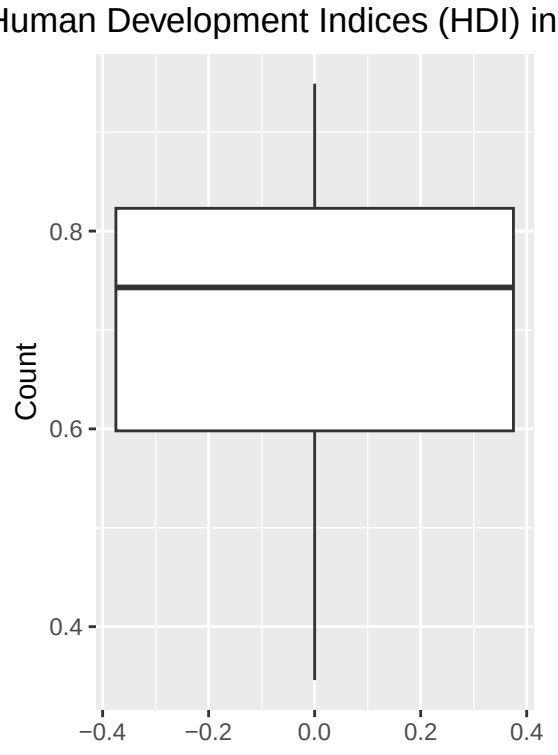
Distribution of Human Development Index (2022) The 2022 HDI histogram and boxplot below indicate a mild left skew. However, because the skewness is not substantial, we chose not to transform the data.



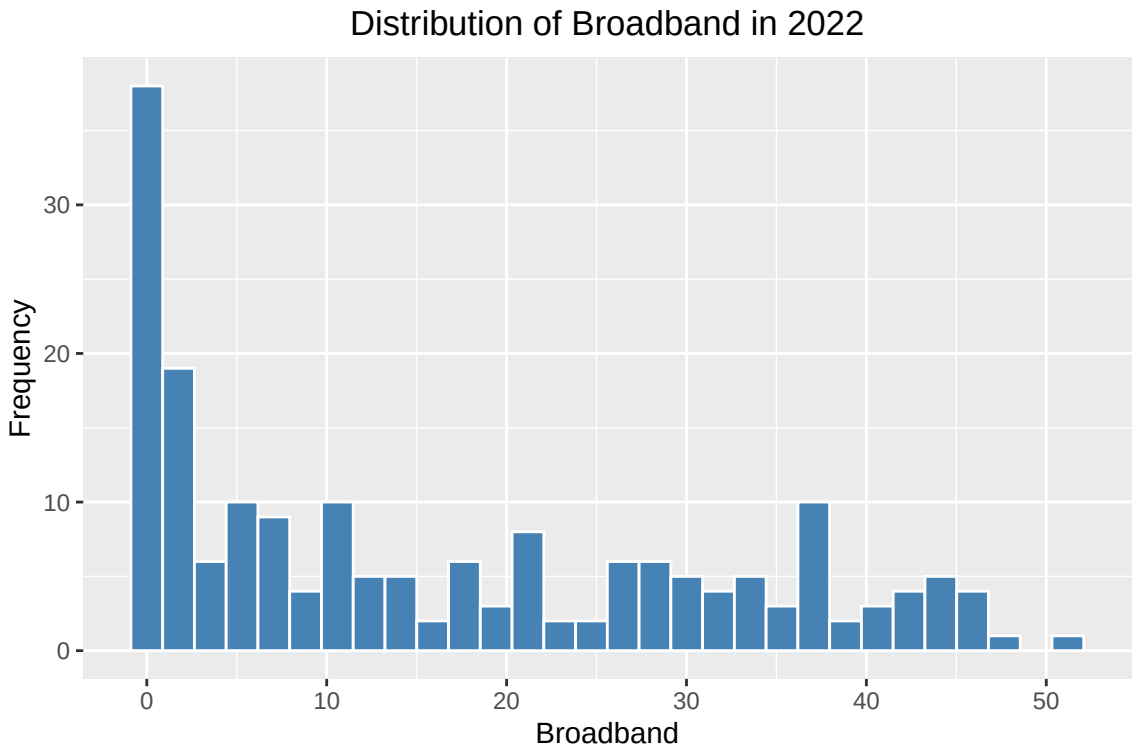


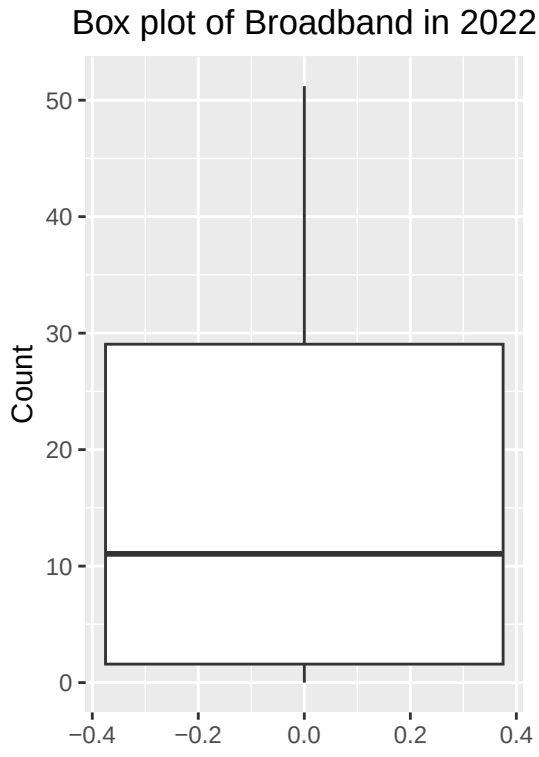
Distribution of Human Development Index (2011) The 2011 HDI histogram and boxplot also show a mild left skew. Similar to the 2022 HDI data, the skewness is not substantial, therefore we chose not to transform the data.



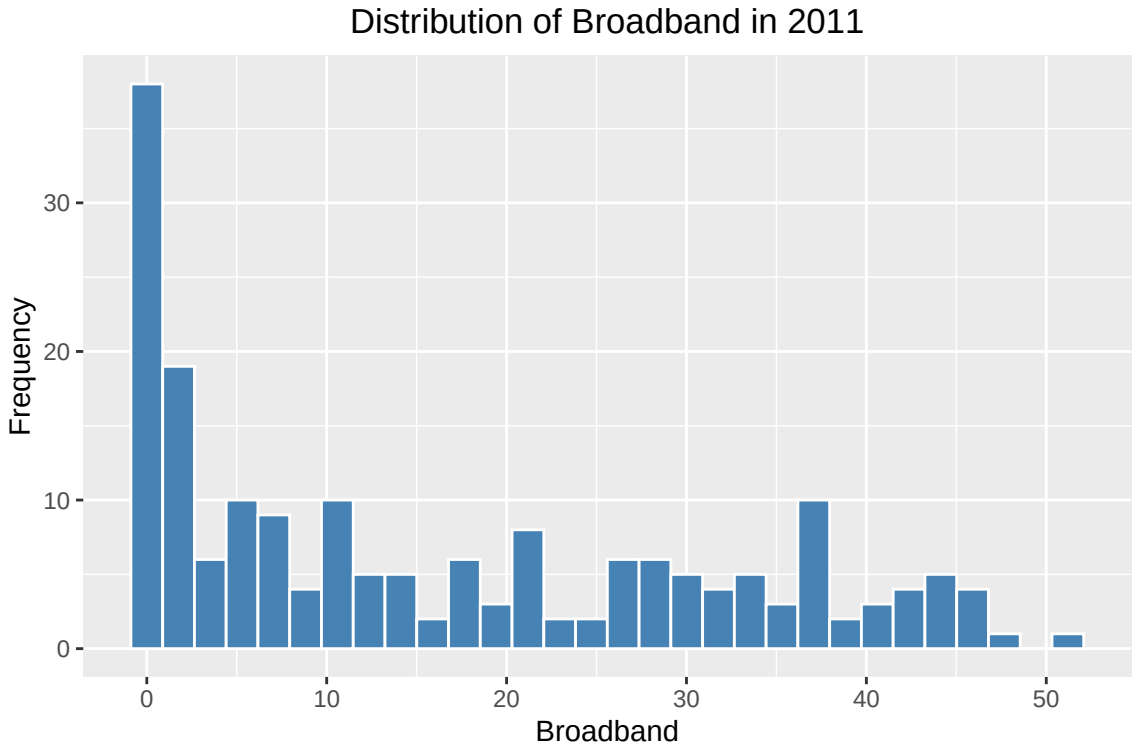


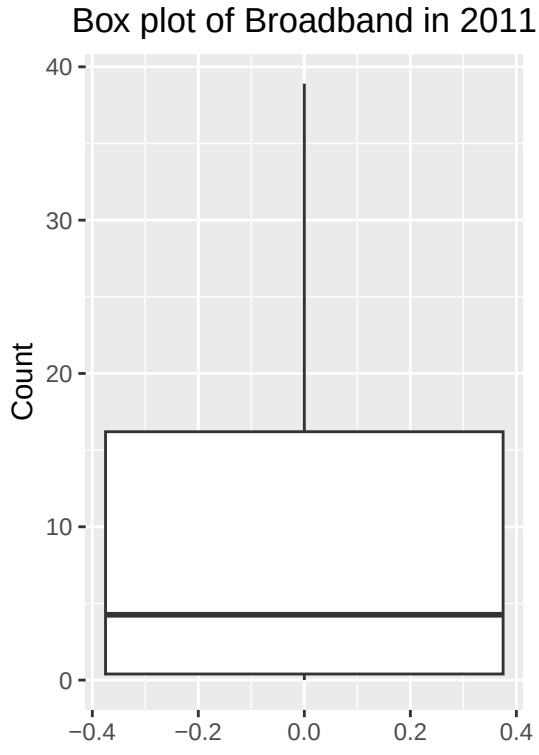
Distribution of Broadband (2022) The 2022 broadband subscription histogram and boxplot show a strong right skew, indicating the need to transform this variable before running the linear regression model.





Distribution of Broadband (2011) The 2011 broadband subscription data is also strongly right skewed based on the histogram and boxplot below. Therefore, we also need to transform this variable before including them in our model.

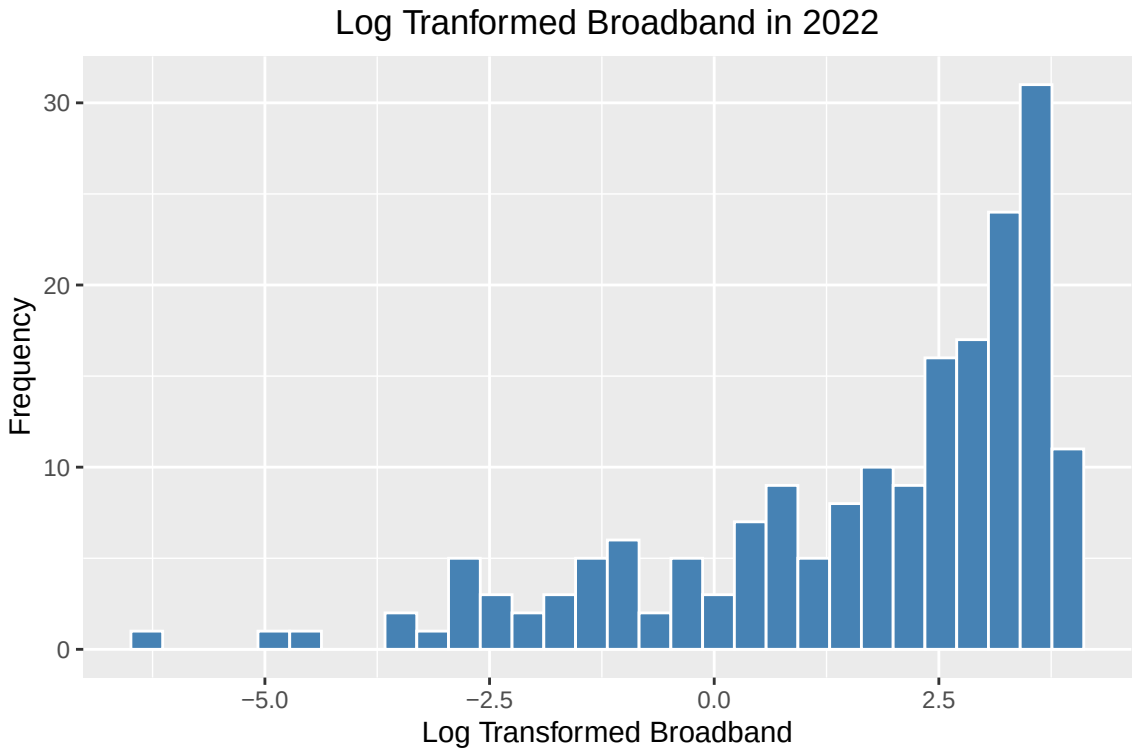




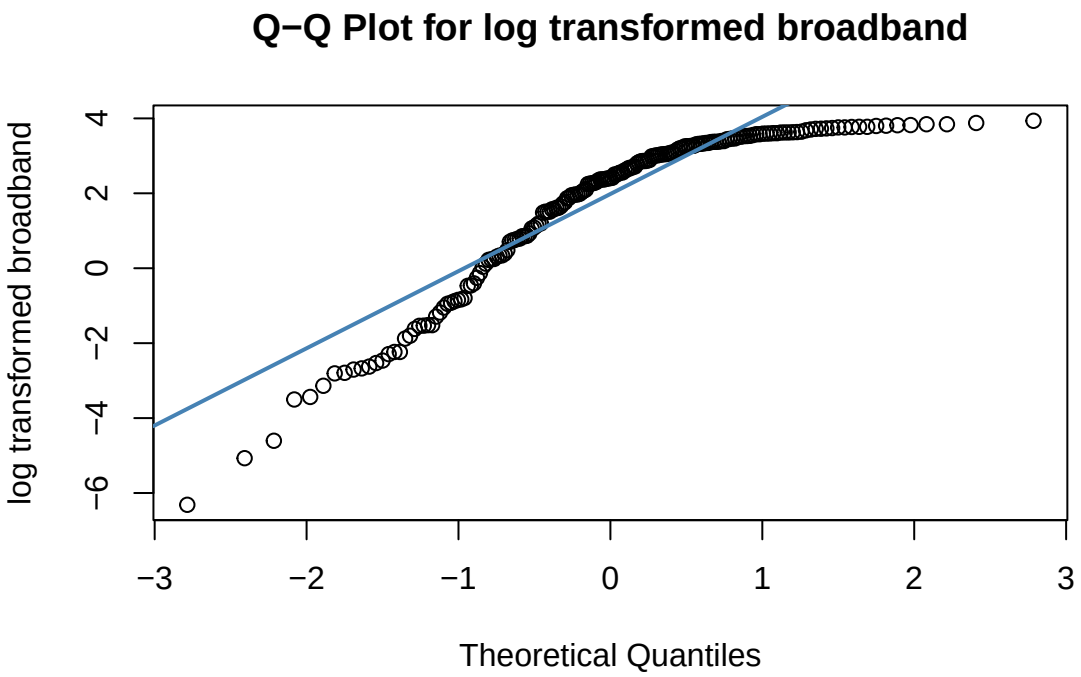
Option 1: Log Transform To address the severe right skew in the 2022 broadband subscription values, we applied a log transformation. First, all countries with a broadband value of zero were removed, as the

logarithm of zero is undefined. We then created a new column, log_broadband, by taking the natural log of the broadband subscription values in the data set.

After the transformation, we visualized the distribution of log_broadband using a histogram and Quantile - Quantile (Q-Q) plot to assess whether the skewness had improved. However, the histogram and Q-Q plot indicate that the log transformation did not adequately normalize the data.

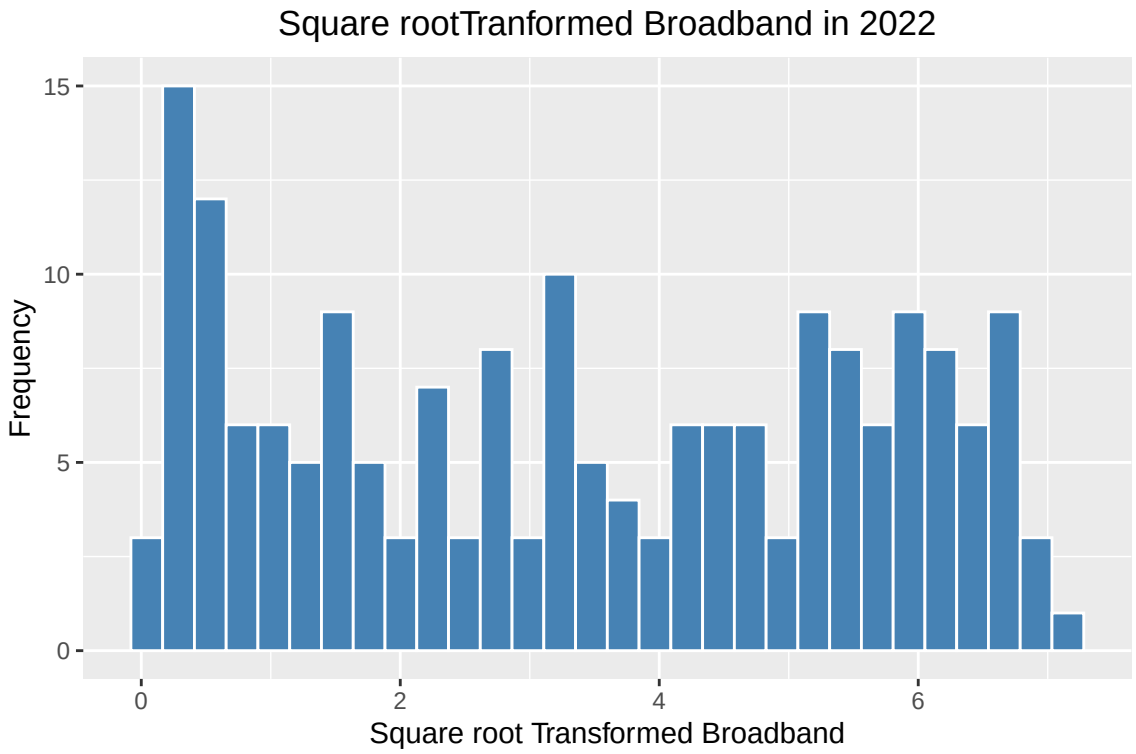



```
##
##  Shapiro-Wilk normality test
##
## data:  data_2022$log_broadband
## W = 0.85633, p-value = 2.711e-12
```

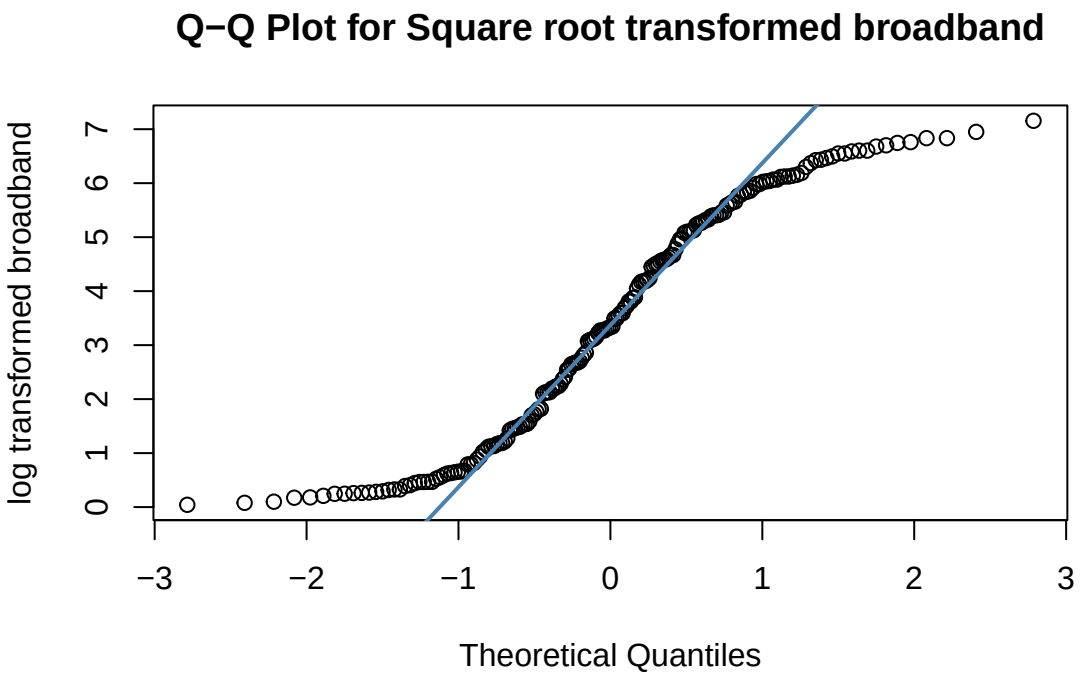


The histogram shows that the log-transformed broadband values in 2022 is left-skewed. Furthermore, the S-shaped curve on the Q-Q plot provides further evidence that the log transformed data do not appear to have a normal distribution.

Option 2: Square root Transform Because the log transformation did not sufficiently improve normality, we then consider a square root transformation instead. We created another column, sqrt_broadband, by taking the square root of the broadband subscription values in the 2022 dataset. After that, we created the histogram and Q-Q plot to visualize the distribution, which show a noticeable improvement in symmetry compared to both the original and log-transformed values




```
##
##  Shapiro-Wilk normality test
##
## data:  data_2022$sqrt_broadband
## W = 0.9293, p-value = 6.888e-08
```

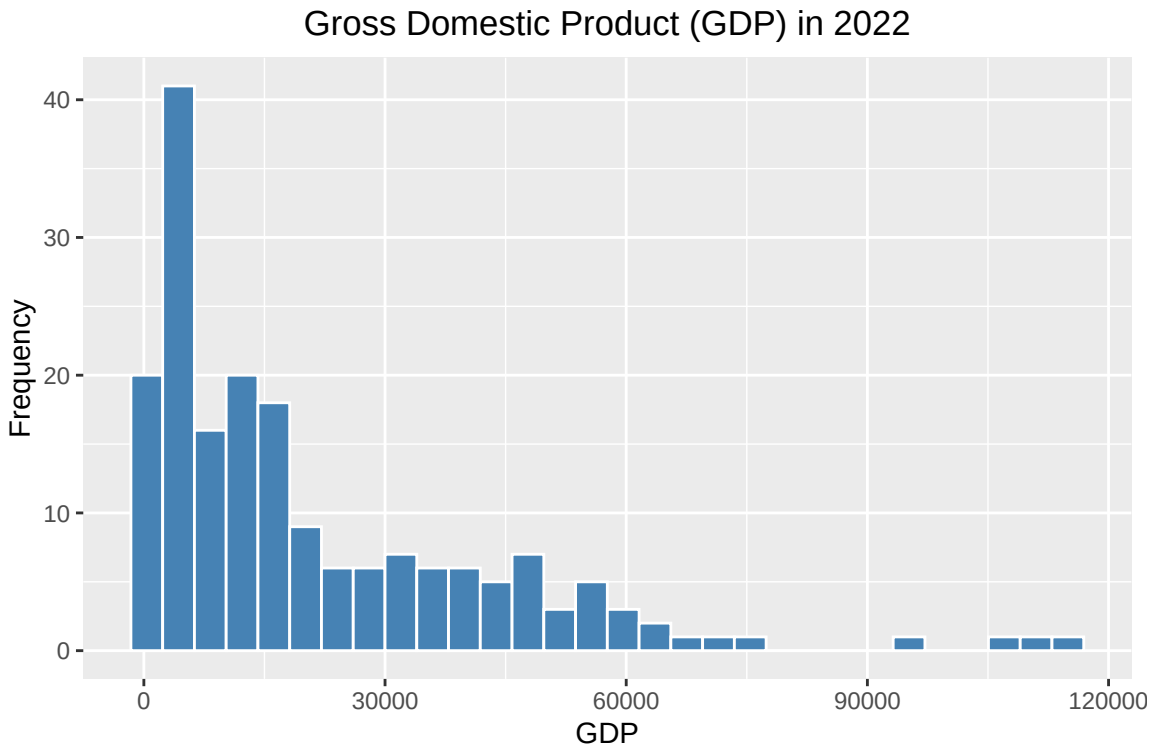


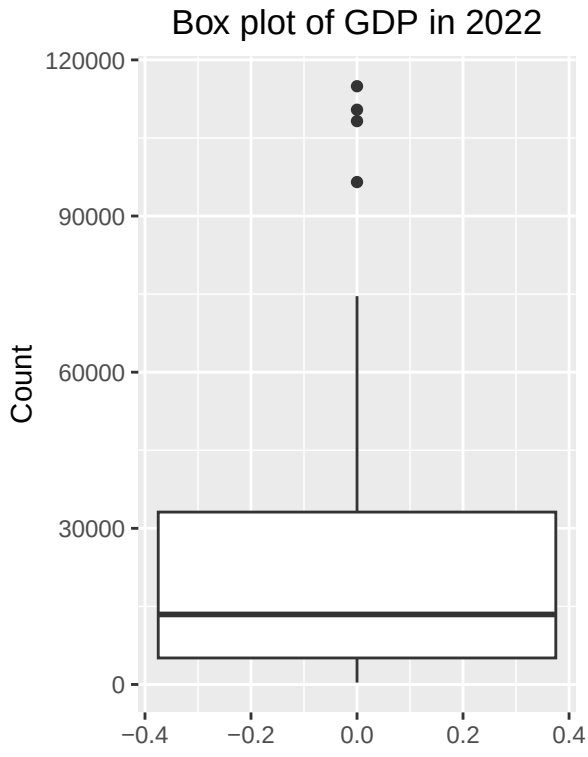
The histogram of the square root-transformed broadband values appears more evenly distributed across the range. Besides that, the Q-Q plot also confirms that the square root transformation brings the data closer to normality, where the points closely follow the theoretical normal line, indicating good alignment with a normal distribution.

Overall, these results make square root transformation a more appropriate approach for the 2022 broadband subscription variable.

Square root Transform for 2011 broadband Because we applied a square-root transformation to the 2022 broadband subscription data, we can apply the same transformation to the 2011 broadband values for consistency. The histogram and Q-Q plot show that the square-root transformation effectively reduces skewness and moves the 2011 broadband subscription data closer to a normal distribution.

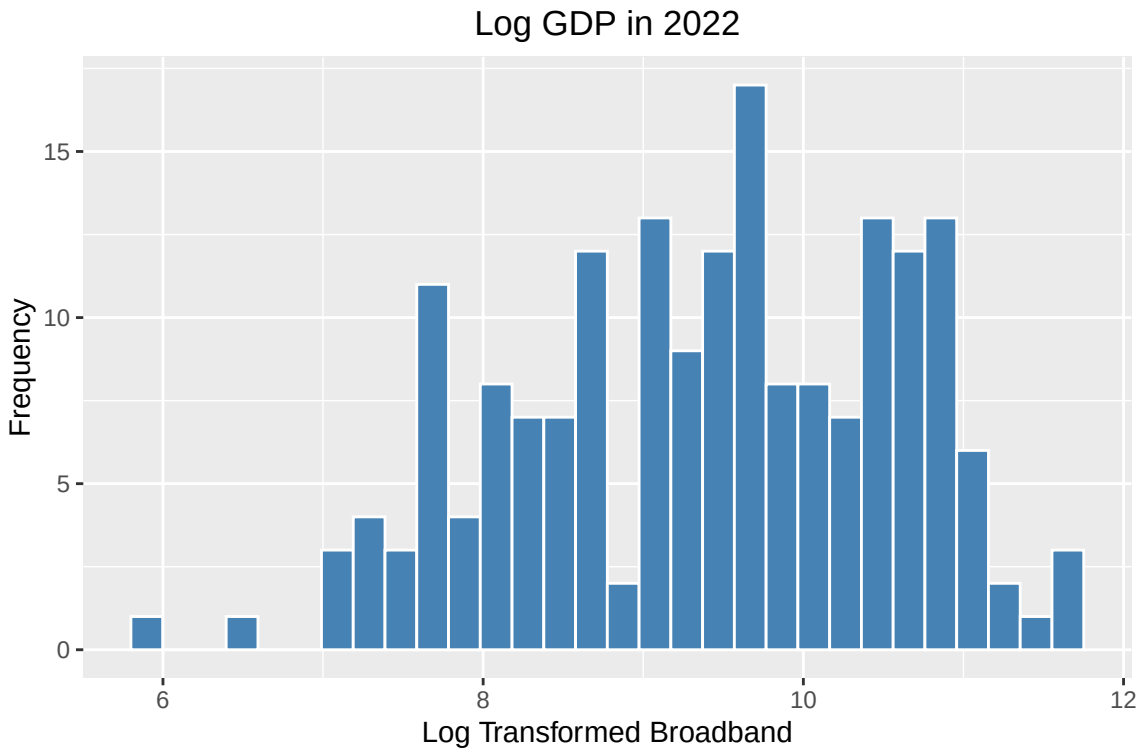
Distribution of Gross Domestic Product (2022) GDP values in 2022 shows a highly skewed distribution as shown in the histogram and boxplot below. The box plot also shows numerous statistical outliers. This suggests that GDP values in 2022 will need to be transformed before including in the regression model.



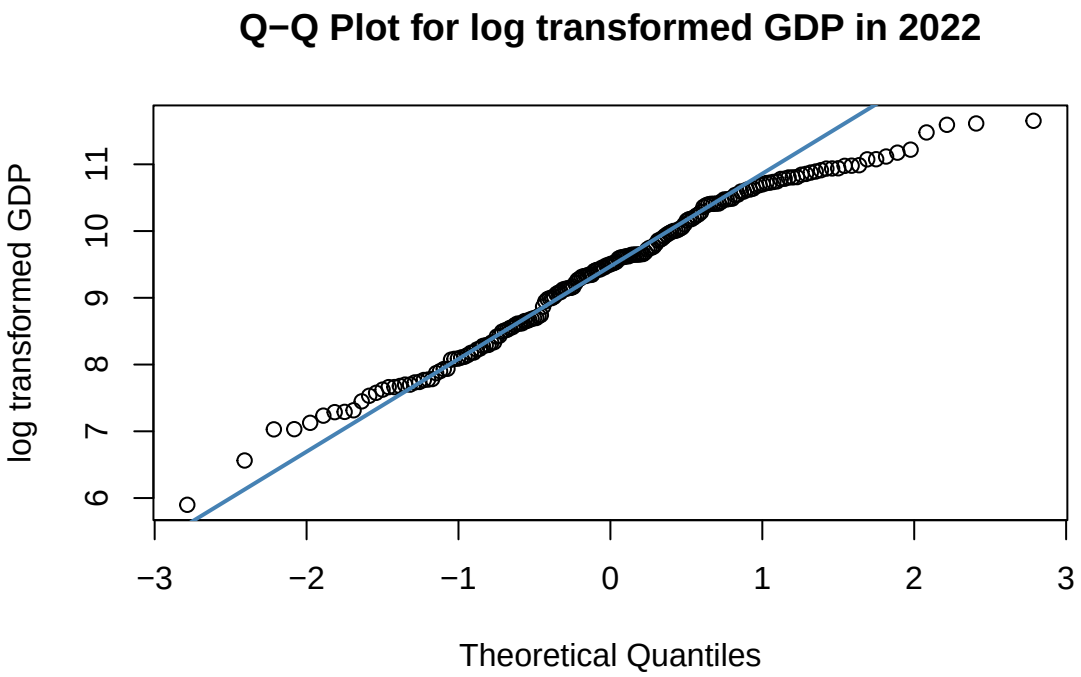


Log Transform for 2022 GDP After applying a log transformation to 2022 GDP values, the distribution becomes more symmetric and approximately normal. The histogram indicates reduced skewness compared to the original data. The Q-Q plot confirms this with most points falling closely along the theoretical normal

line. Overall, the log transformation is effective in improving the suitability of GDP for the regression modelling.

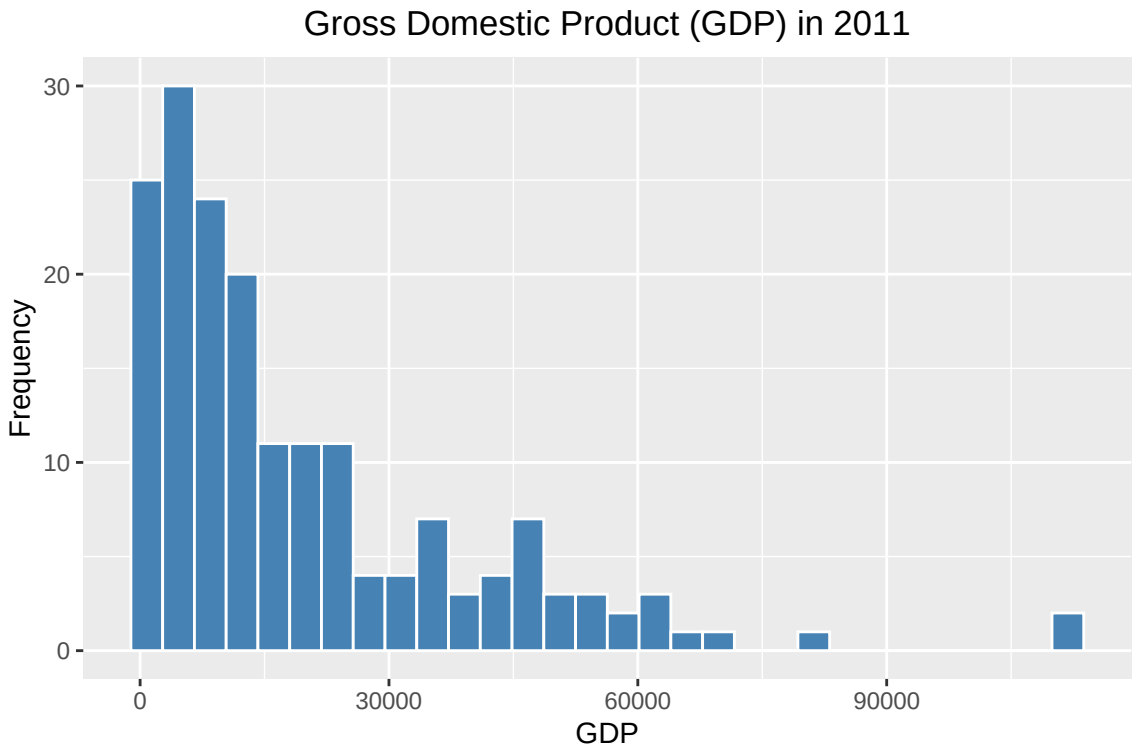


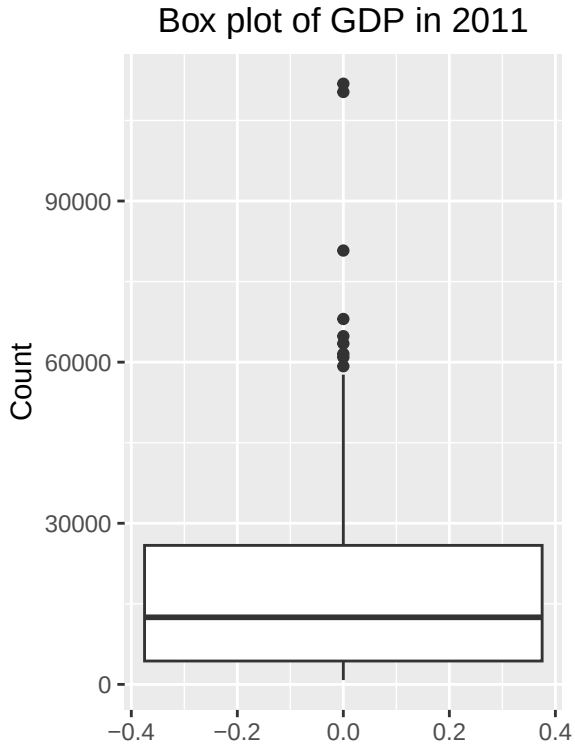

```
##
##  Shapiro-Wilk normality test
##
## data:  data_2022$log_gdp
## W = 0.97764, p-value = 0.004284
```



Distribution of GDP (2011)

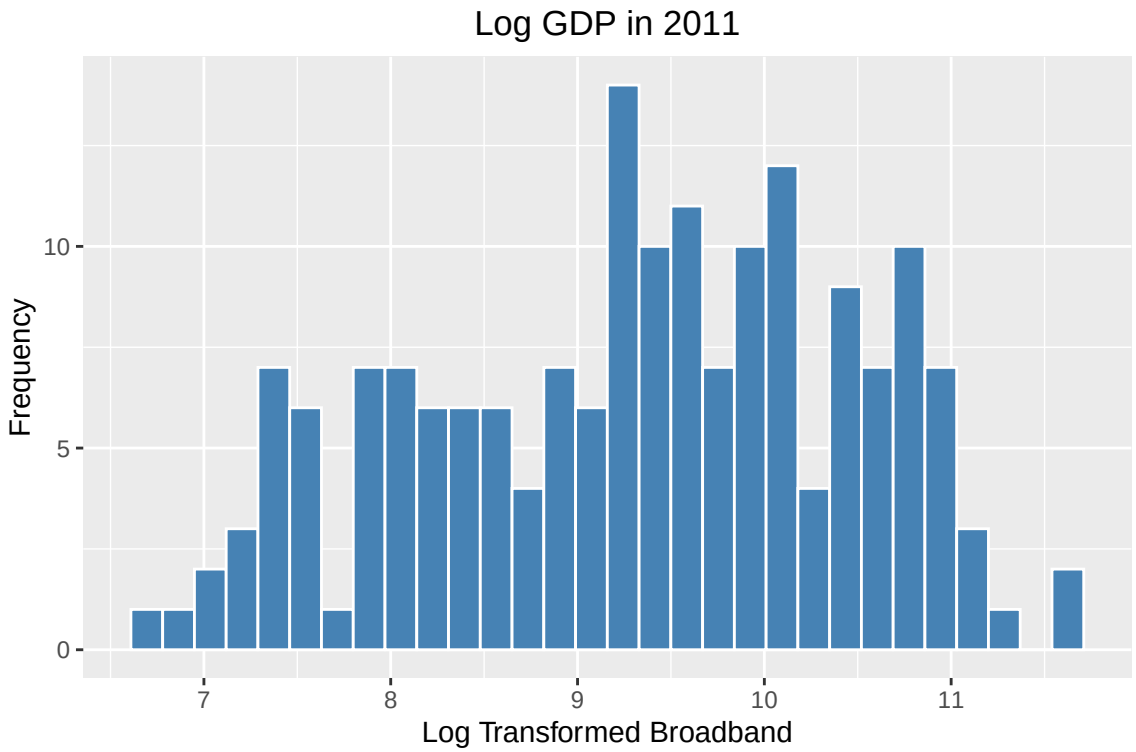
The distribution of GDP in 2011 is highly right-skewed, as shown by both the histogram and the box plot. There are several extreme values appear as outliers well above the upper whisker. We will need to transform the data before including it in the model.



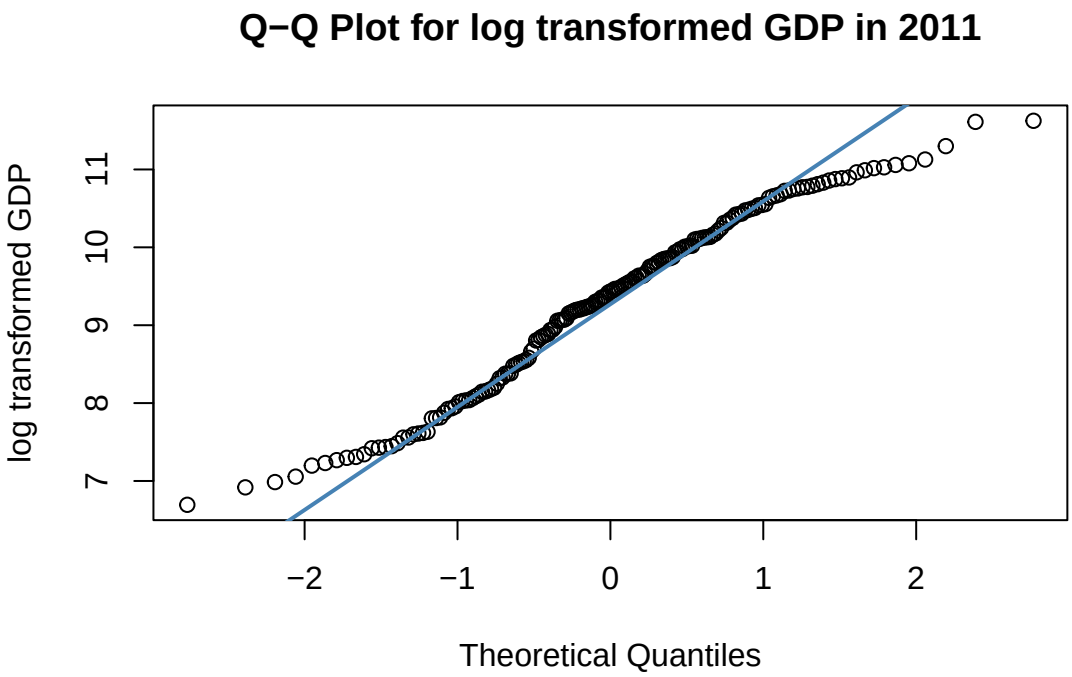


Since a log transformation was successfully applied to the 2022 GDP values, the same approach will be used for the 2011 data. After applying a log transformation, the distribution of GDP in 2011 becomes much more symmetric and closer to a normal shape. The Q–Q plot also shows most points fall closely

along the theoretical normal line, indicating that the transformed data approximates normality. Overall log transformation effectively reduces skewness and makes the data more suitable for modelling.




```
##
##  Shapiro-Wilk normality test
##
## data:  data_2011$log_gdp
## W = 0.97339, p-value = 0.001804
```



Standardization

We standardized our data using z-scores to ensure that all variables were equally weighted in our analysis. Standardization is the process of transforming raw values into z-scores, which express each observation relative to the mean and standard deviation of its variable.

For each year (2011 and 2022), we calculated the mean and standard deviation of each variable. Z-scores were computed using the formula z= (x-)/ , where x is the observation value, is the mean, and is the standard deviation of the variable.

This transformation produced standardized variables with a mean of 0 and standard deviation of 1, allowing for comparisons across our three variables HDI, Broadband access, and GDP.

Question 1

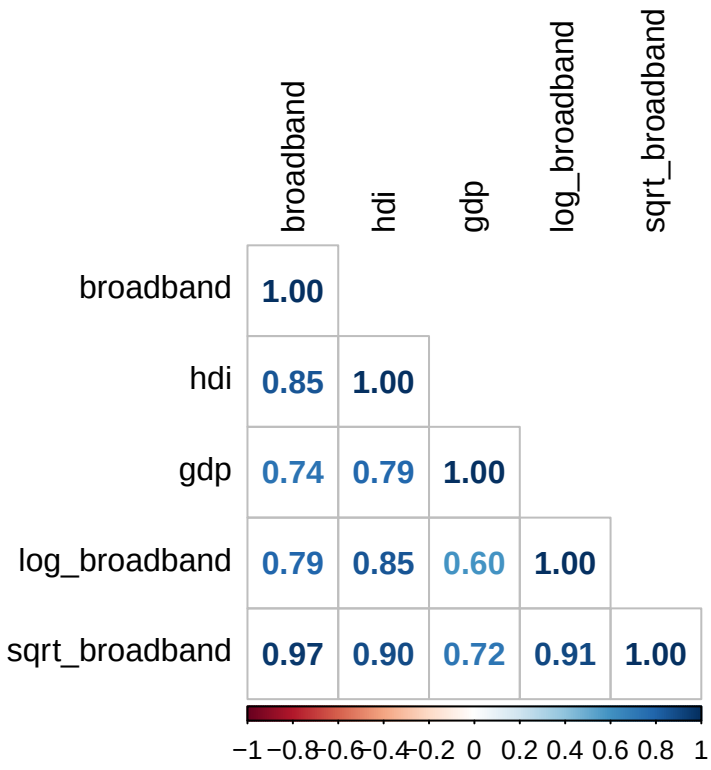
What is the relationship between HDI and broadband subscriptions, based on the latest available data, which is 2022.

Descriptives and EDA

Let's peek at means and standard deviations.

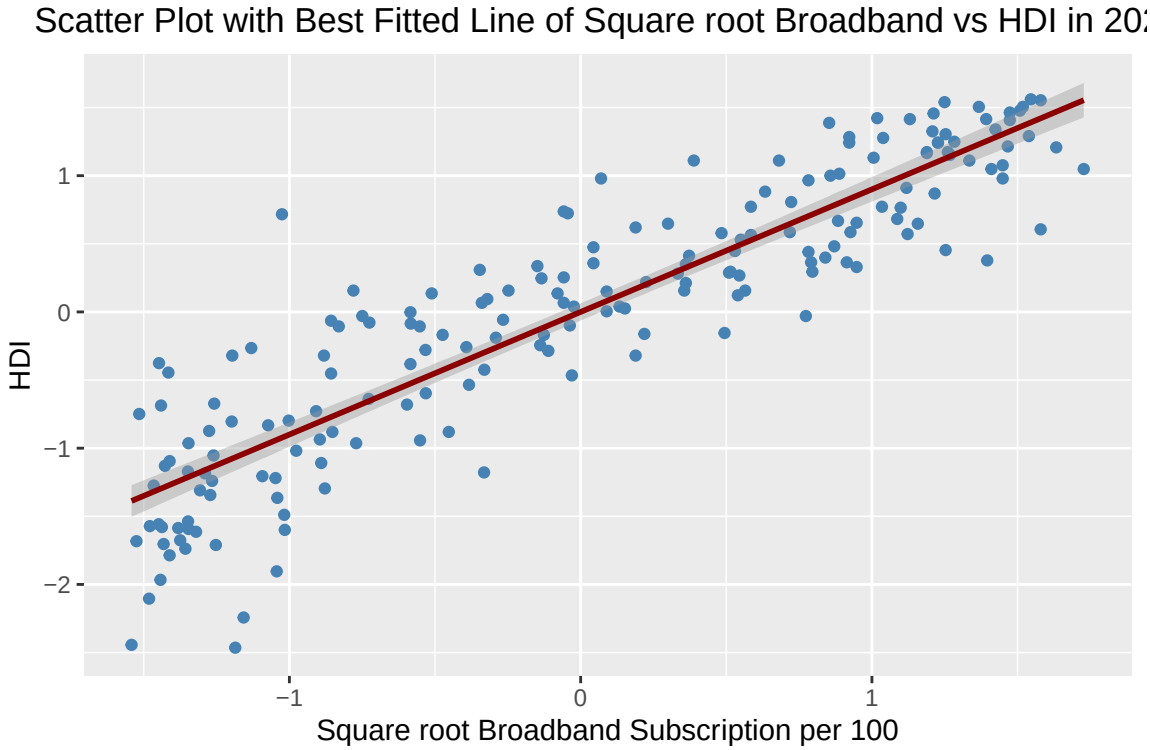
Correlation matrix

Let's look at the correlations between the independent variables. We want there to be low correlations for no multicollinerity.



Plots

Scatterplot



```
##
## Call:
## lm(formula = hdi ~ sqrt_broadband + log_gdp, data = data_2022_standardized)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.87118 -0.14524  0.02077  0.13574  0.79086
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -6.138e-16  1.926e-02    0.00        1
## sqrt_broadband  2.940e-01  3.869e-02    7.60 1.46e-12 ***
## log_gdp         6.991e-01  3.869e-02   18.07  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2633 on 184 degrees of freedom
## Multiple R-squared:  0.9314, Adjusted R-squared:  0.9307
## F-statistic:  1249 on 2 and 184 DF,  p-value: < 2.2e-16
```

Question 2

Are there countries that significantly deviate from this relationship between HDI and broadband subscription rate?

```
## # A tibble: 2 x 10
##   country  broadband    hdi    gdp log_broadband sqrt_broadband log_gdp
##   <chr>        <dbl>  <dbl>  <dbl>         <dbl>          <dbl>   <dbl>
## 1 Kiribati     -1.06 -0.687 -0.874        -2.03           -1.44   -1.51
## 2 Mali         -1.02 -2.24  -0.870        -0.903          -1.16   -1.48
## # i 3 more variables: fitted_hdi <dbl>, residuals <dbl>, std_resid <dbl>
```

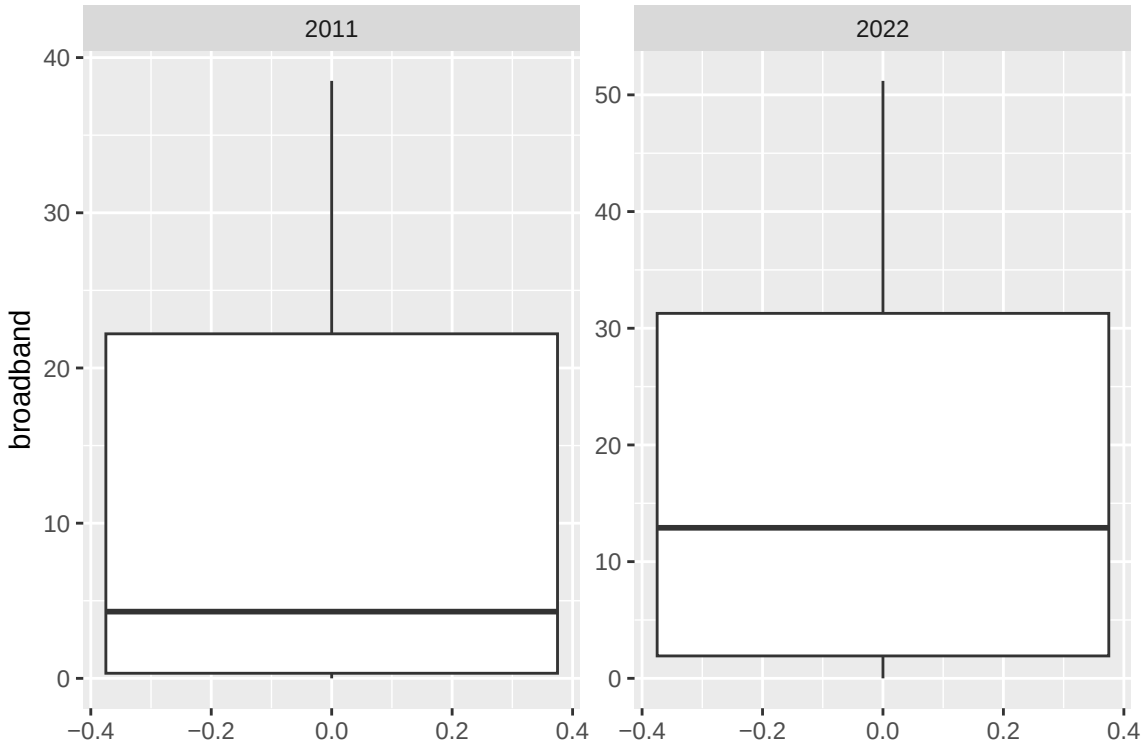
Our outliers in 2022 are Kiribati and Mali

Question 3

How has broadband access by country changed between 2011, when the UN declaration was announced, and 2022?

Descriptives and EDA

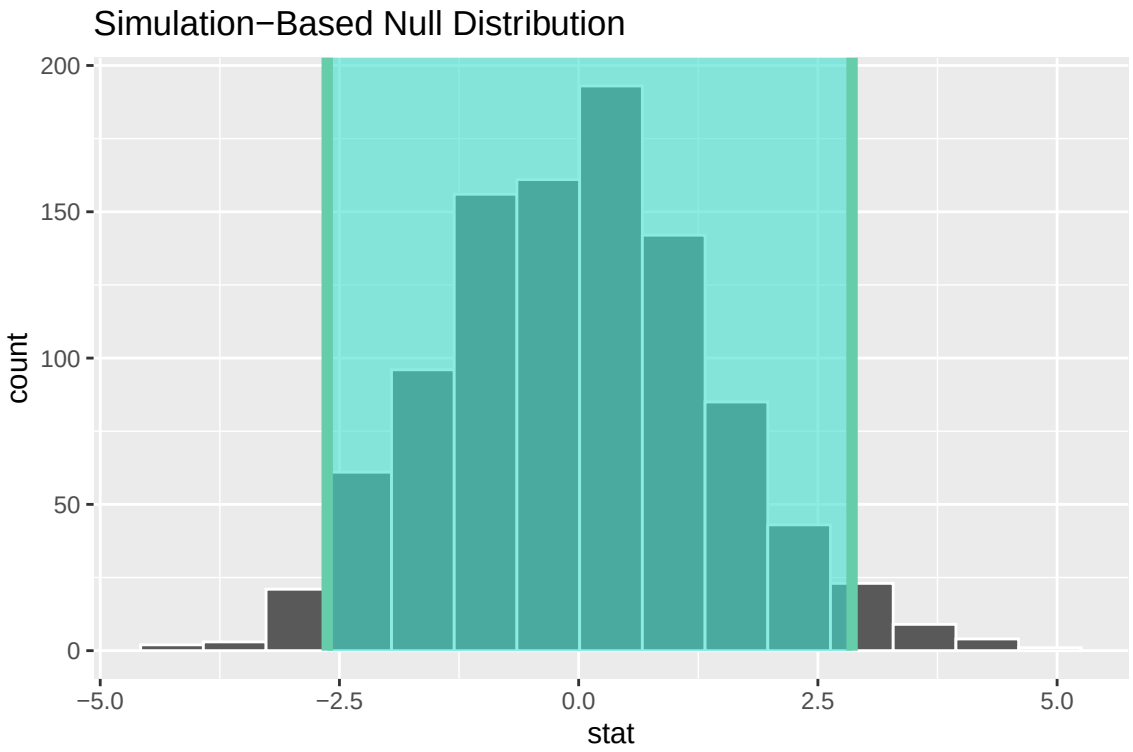
Plots



T-test

```
## [1] -9.393962 -3.880547
## attr(,"conf.level")
## [1] 0.95
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    -2.63     2.86
```

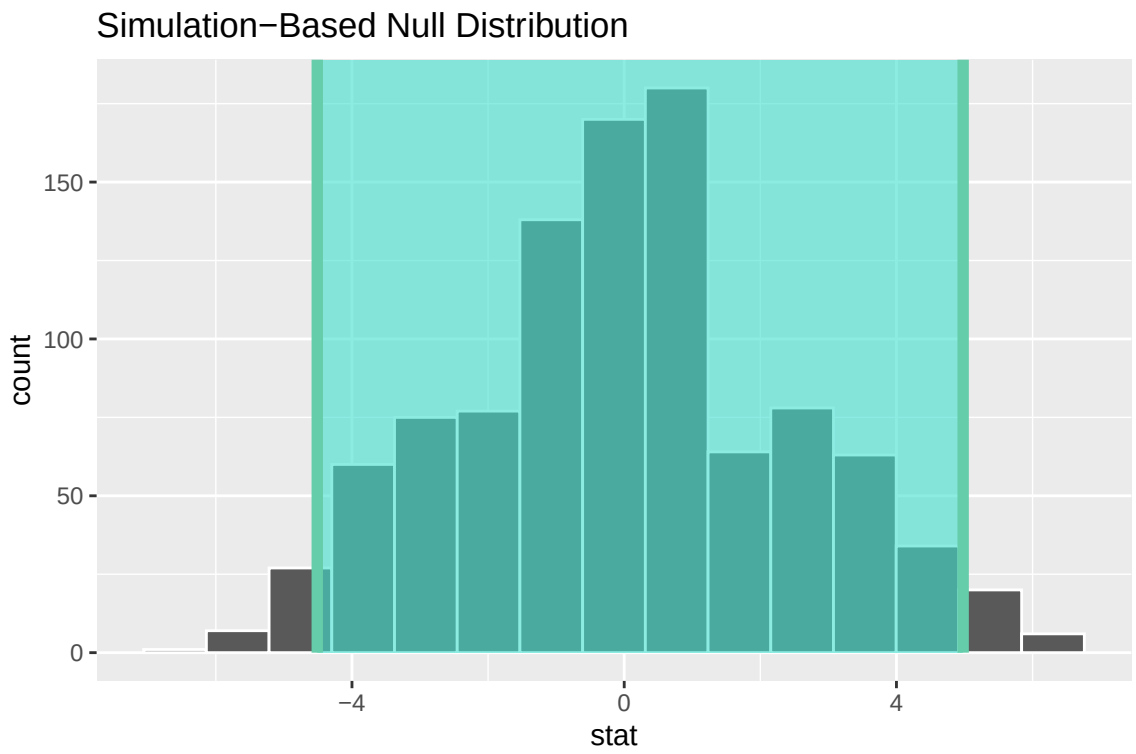


Wilcoxin-test

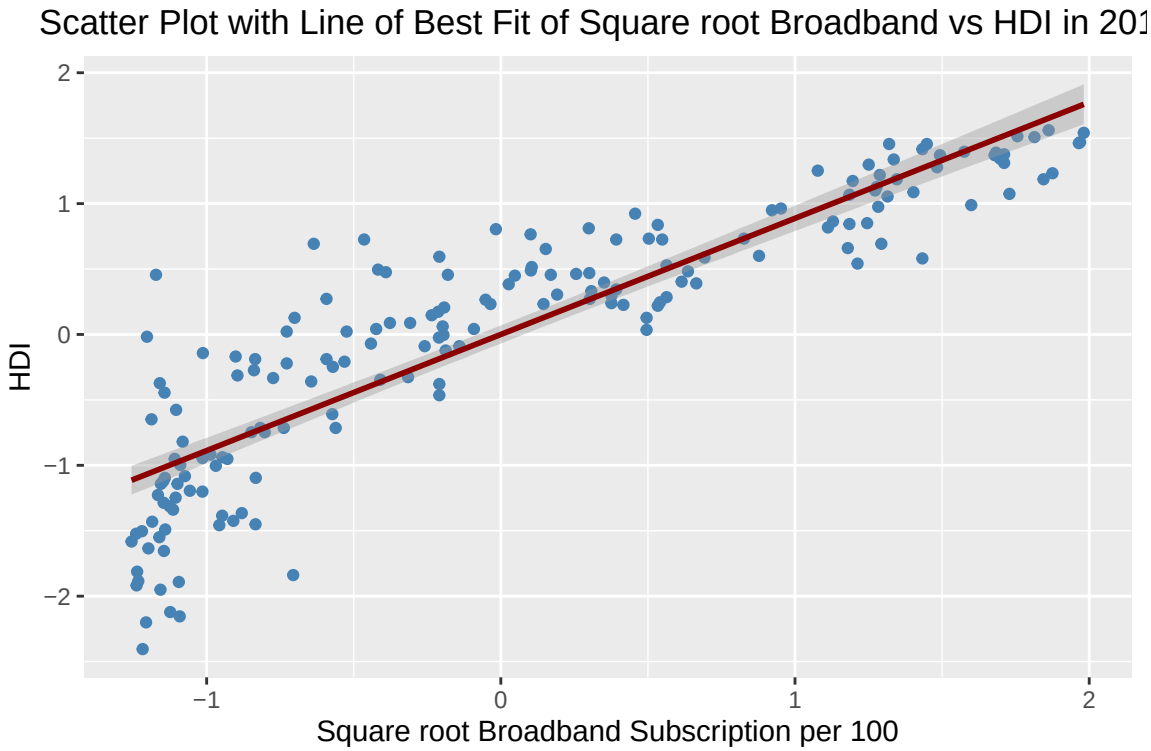
```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  data_2011$broadband and data_2022$broadband
## W = 12189, p-value = 1.391e-05
## alternative hypothesis: true location shift is not equal to 0
```

```
## NULL
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    -4.51     4.98
```



Scatterplot



##

```
## Call:
## lm(formula = hdi ~ sqrt_broadband + log_gdp, data = data_2011_standardized)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.94560 -0.13112 -0.00058  0.15436  1.04002
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.379e-16  2.292e-02   0.000        1
## sqrt_broadband 3.406e-01  4.293e-02   7.934 2.49e-13 ***
## log_gdp        6.477e-01  4.293e-02  15.089  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.305 on 174 degrees of freedom
## Multiple R-squared:  0.908,  Adjusted R-squared:  0.907
## F-statistic: 859.1 on 2 and 174 DF,  p-value: < 2.2e-16
```

##Outlier country in 2011

```
## # A tibble: 2 x 9
##   country  broadband    hdi    gdp sqrt_broadband log_gdp fitted_hdi residuals
##   <chr>        <dbl>  <dbl>  <dbl>          <dbl>   <dbl>      <dbl>     <dbl>
## 1 Cuba        -0.846  0.456 -0.570         -1.17   -0.285     -0.584     1.04
## 2 Djibouti    -0.745 -1.84  -0.795         -0.706  -1.01      -0.893    -0.946
## # i 1 more variable: std_resid <dbl>
```

Our outlier is 2011 is Cuba

Question ??

Create a multi regression with gdp as the control variable

```
##
## Call:
## lm(formula = hdi ~ sqrt_broadband + log_gdp, data = data_2022_standardized)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.87118 -0.14524  0.02077  0.13574  0.79086
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -6.138e-16  1.926e-02    0.00        1
## sqrt_broadband  2.940e-01  3.869e-02    7.60 1.46e-12 ***
## log_gdp         6.991e-01  3.869e-02   18.07  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2633 on 184 degrees of freedom
## Multiple R-squared:  0.9314, Adjusted R-squared:  0.9307
## F-statistic:  1249 on 2 and 184 DF,  p-value: < 2.2e-16
```

Checking assumptions

Lack of Multicollinearity

```
## sqrt_broadband        log_gdp
##       4.015438       4.015438
```

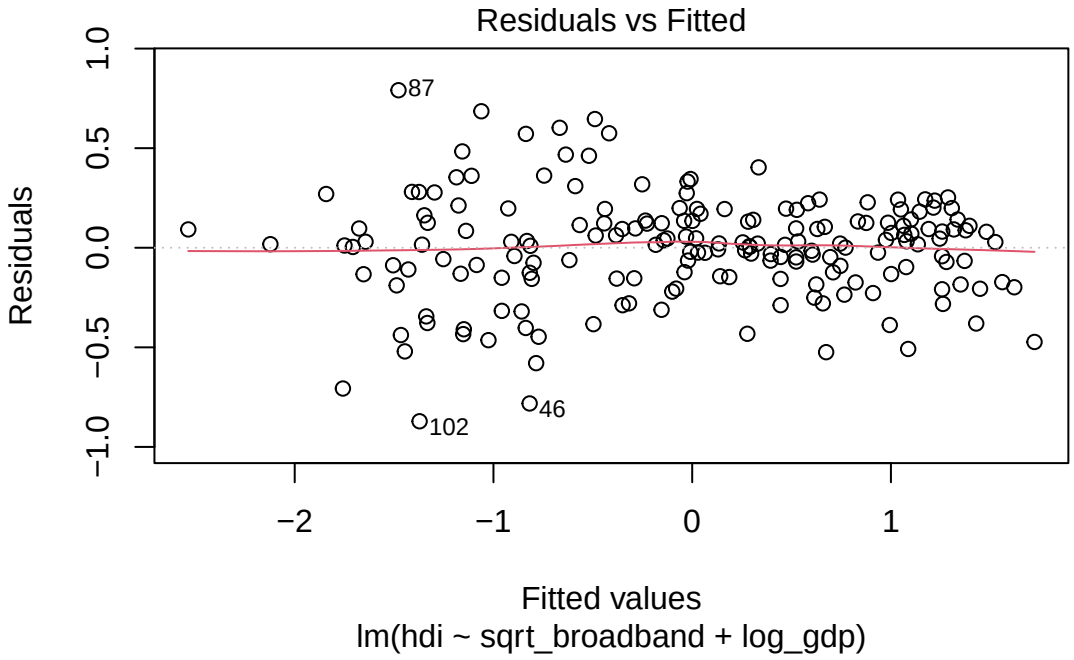
The variance inflation factors are 4.01 for sqrt_broadband and 4.01 for log_gdp. This indicates that there is not a strong linear relationship between these factors.

Independence

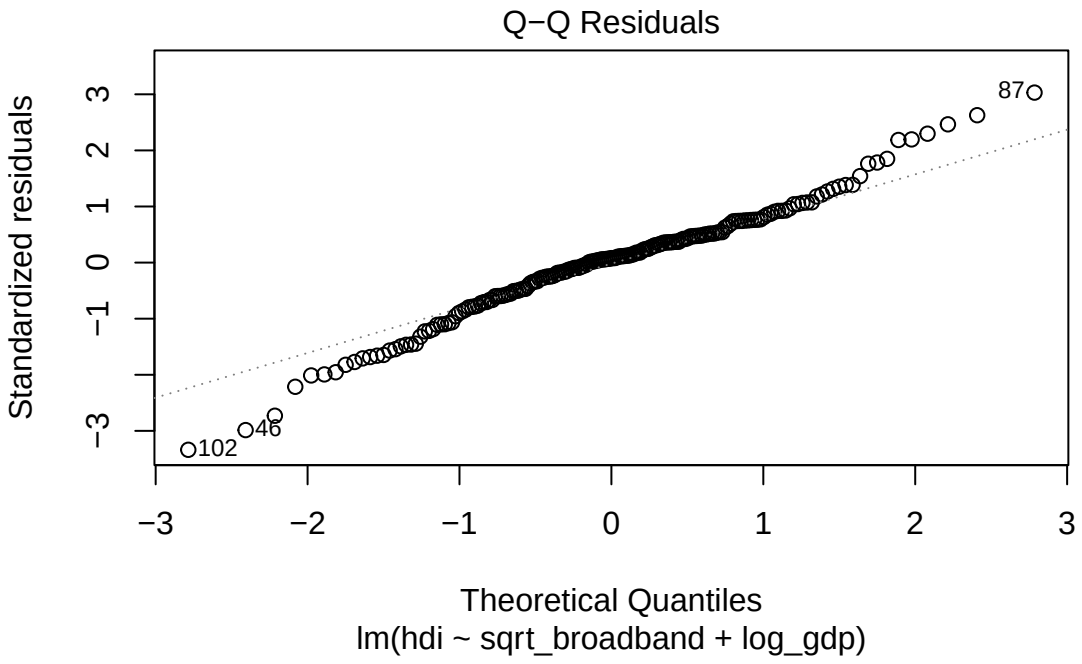
```
##
##  Durbin-Watson test
##
## data:  multi_reg
## DW = 1.9102, p-value = 0.2686
## alternative hypothesis: true autocorrelation is greater than 0
```

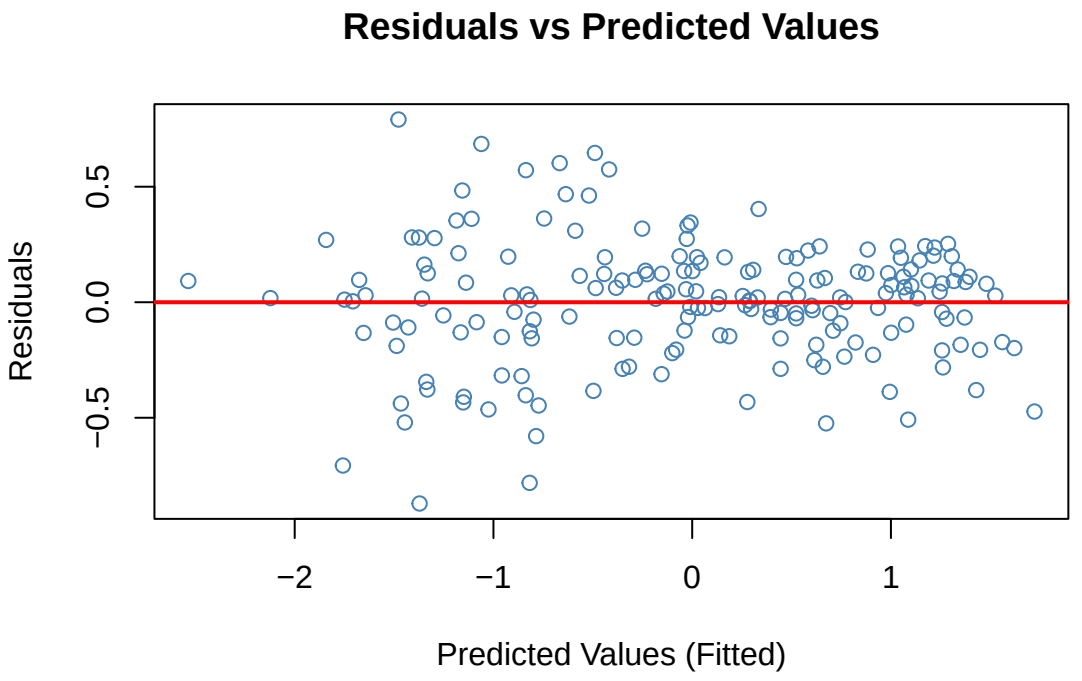
??

Normaility of residuals




```
##
##  Shapiro-Wilk normality test
##
## data:  res_model
## W = 0.98015, p-value = 0.009251
```

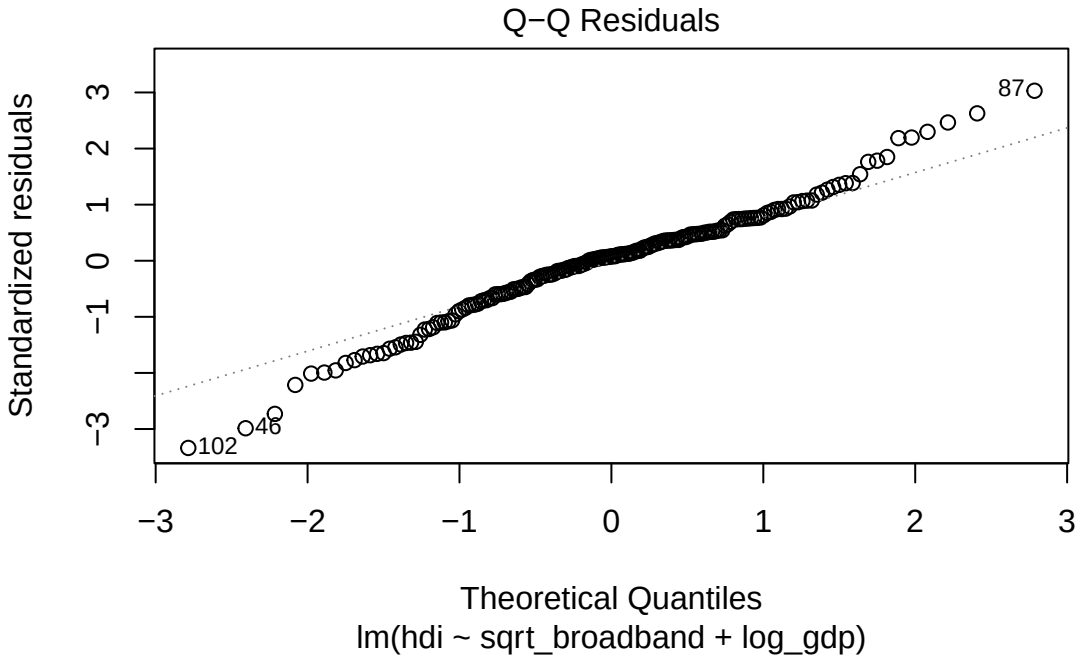





```
## integer(0)
```

Homoscedascity Next, let's look for the independence of the residuals.

```
##
##  Shapiro-Wilk normality test
##
## data:  res_model
## W = 0.98015, p-value = 0.009251
```



6. Results and Interpretation

Lorem ipsum dolor sit amet, consectetur adipiscing elit. In egestas lobortis porta. Phasellus tincidunt metus sed mollis imperdiet. Aliquam blandit nibh in fermentum consequat. Fusce blandit magna quis nulla ornare, non mattis enim tempor. Nulla et odio dui. Nunc ut elit venenatis, porttitor dui et, placerat tortor. Sed aliquet sodales magna, sit amet porta odio tempor in. CHANGE THIS OF COURSE

7. Discussions

Lorem ipsum dolor sit amet, consectetur adipiscing elit. In egestas lobortis porta. Phasellus tincidunt metus sed mollis imperdiet. Aliquam blandit nibh in fermentum consequat. Fusce blandit magna quis nulla ornare, non mattis enim tempor. Nulla et odio dui. Nunc ut elit venenatis, porttitor dui et, placerat tortor. Sed aliquet sodales magna, sit amet porta odio tempor in. CHANGE THIS OF COURSE

Appendix

Code

The report was created using *R* (R Core Team 2020) and *R Studio* (RStudio Team 2020) with *R Markdown* (RStudio, n.d.). The main library utilized for this purpose is *Tidyverse* (Wickham et al. 2019). Its used sub-packages include *dplyr* (Wickham et al. 2022) to enable query-like syntax, and *ggplot* (Wickham 2016) to create graphs and charts. Other packages and tools include *janitor* (Firke 2021), *knitr* (Xie 2022), *kableExtra* (Zhu 2021), and *scales* (Wickham and Seidel 2022).

References

Firke, Sam. 2021. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Gapminder Datasets. 2023. Gapminder. https://www.gapminder.org/data/.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

RStudio. n.d. *Your Data Tells a Story. Tell It with r Markdown*. https://rmarkdown.rstudio.com.

RStudio Team. 2020. *RStudio: Integrated Development Environment for r*. Boston, MA: RStudio, PBC. http://www.rstudio.com/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2022. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, and Dana Seidel. 2022. *Scales: Scale Functions for Visualization*. https://CRAN.R-project.org/package=scales.

Xie, Yihui. 2022. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.