Namal University, Mianwali

Department of Business Studies

Date: June 16, 2025

Semester Project

Title: Analyzing Pakistan_Hunger_Data with R

Course: Data Analysis with R

Submitted by:

Zafran Ali Khan	(NUM-BBA-2022-11)
Awais Asif	(NUM-BBA-2022-34)

Submitted to:

Mr. Hamza Wazir Khan

Contents

1. Introduction and Research Question

Pakistan faces a range of socio-economic challenges, and among the most pressing is the issue of hunger and malnutrition. These problems are deeply connected with poverty, poor access to nutrition, and inequalities across different regions and communities. Hunger not only affects physical well-being but also impacts education, productivity, and the overall development of the country.

The purpose of this data analysis project is to understand the patterns, relationships, and contributing factors to hunger-related issues across various Pakistani cities using real-world data. By using statistical and visual analysis techniques in R, the aim is to uncover valuable insights that can help in forming data-driven recommendations for addressing hunger.

This project follows a structured, multi-phase approach—from data preparation and cleaning to exploratory data analysis and regression modeling. Each phase builds upon the previous to gradually shape a deeper understanding of the hunger situation in Pakistan.While performing exploratory and regression analysis, we decided to focus our attention on a specific research question that could provide meaningful conclusions based on available data.

“To what extent do malnutrition rates and child underweight levels predict the severity of hunger in Pakistan?”

This research question is based on the observed relationships from the data, as well as prior findings in the exploratory analysis. The chosen dependent variable, **Hunger Severity Index**, is a comprehensive indicator that reflects the seriousness of hunger in a population. The independent variables—**Malnutrition Rate** and **Children Underweight**—are among the most direct contributors to hunger severity, and were found to have strong correlations with the dependent variable during initial exploration.

By answering this question, the project aims to reveal how deeply these two factors contribute to hunger in Pakistan, which can inform future actions by policymakers, NGOs, and health organizations.

Phase 1: Data Acquisition and Preparation

1.1 Dataset Selection and Source

The dataset I selected is called **“Pakistan Hunger Data”**. I downloaded it from **Kaggle**, which is a well-known website where people share public datasets for learning and practice. This dataset contains data related to poverty, food access, and health conditions in different cities of Pakistan over the last few years.

- **Dataset Name**: Pakistan Hunger Data (2020–2023)
- **Source**: https://www.kaggle.com/datasets/haseebindata/pakistan-hunger-data
- **File Format**: CSV
- **Total Records**: 500 rows
- **Years Covered**: 2020 to 2023
- **Cities Covered**: 10 major cities of Pakistan

1.2 Reason for Choosing This Dataset

I chose this dataset because it talks about very important social issues in Pakistan, like hunger, poverty, and child health. These are serious problems that affect many people, especially in big cities. This dataset is also great for learning because:

- It has both **numbers and categories** (like city names and years).
- It covers **multiple years**, so we can see trends over time.
- It allows us to do real analysis, like **data cleaning**, **summarizing**, **visualizing**, and **building models** in R.

1.3 What the Data Contains (Variables)

Here are the columns (variables) included in the dataset:

Column Name	What It Means
City	Name of the city in Pakistan
Year	The year the data was recorded (2020–2023)
Population_Under_Poverty	% of people living below the poverty line
Malnutrition_Rate	% of people suffering from malnutrition
Food_Insecurity	% of people without regular access to enough food
Access_to_Clean_Water	% of people with access to clean drinking water
Food_Production_Index	An index showing food and agricultural production trends
Children_Underweight	% of children who are underweight or malnourished

These variables cover different areas such as health, poverty, and food security.

1.4 What I Found in the Data

After loading the dataset, I noticed that:

- The dataset is **neatly arranged** and easy to understand.
- It covers **four years** of data across **10 cities**.
- There are some **missing values** in the columns:
  - Population_Under_Poverty: 4 missing values
  - Access_to_Clean_Water: 5 missing values
  - Food_Production_Index: 5 missing values

We will fix these missing values in the next phase of the project.

Phase 3: Data Cleaning and Transformation

Before phase two, which is descriptive statistics, we performed phase 3 which is Data Cleaning and Transformation. The major reason behind this decision was to make data more reliable for better results ahead.

3.1 Identifying Data Issues

Before conducting any analysis, it's essential to verify that the data is free from errors. These issues are referred to as data quality problems. They can include:

- **Missing values** (where some cells are blank): We checked for missing values using the R code:
    - colSums(is.na(pakistan_hunger_data_raw))
- **Outliers** (very high or low values that are different from others): For the detection of outliers, we used two ways, the box plots and the manual IQR method.

  For box plots, we identified and mentioned the numeric variables/columns and using for loop created box plots for different variables and the code was as follows:

  ```
  numeric_cols <- pakistan_hunger_data_raw[, sapply(pakistan_hunger_data_raw, is.numeric)]

  par(mfrow = c(2, 3))

  for (col in names(numeric_cols)) {

   boxplot(numeric_cols[[col]], main = col, col = "lightblue")

  }

  par(mfrow = c(1, 1))
  ```

  For IQR function based method the code was as follows where first developed a function:

  ```
  detect_outliers <- function(x) {
   Q1 <- quantile(x, 0.25, na.rm = TRUE)
   Q3 <- quantile(x, 0.75, na.rm = TRUE)
   IQR_val <- Q3 - Q1
   lower <- Q1 - 1.5 * IQR_val
   upper <- Q3 + 1.5 * IQR_val
   return(which(x < lower | x > upper))
  }
  ```

Then applied it to each numeric variable but overall there was no outlier.

- **Inconsistencies** (values that don’t make sense, like negative poverty)
- **Wrong data types** (like numbers stored as text): Checked using str() function, all the datatypes were correct.

So, in this step, we carefully checked our dataset to find any of these problems.

Issues Found:

Issue Type	Column(s) Involved	Description
Missing Values	Population_Under_Poverty, Access_to_Clean_Water, Food_Production_Index.	Some values are NA (Not Available)
Outliers (Potential)	Population_Under_Poverty, Food_Production_Index	Before checking it seems that these variables could have some outliers, but after checking all variables, no variable had any outlier
Data Type (OK)	All columns	All data types were correct
Inconsistencies	None found	No strange or invalid values found

3.2 Handling Missing Values

After finding missing values, I had to fix them. There are two common methods:

- **Deletion**: Removing rows or columns if too much data is missing
- **Imputation**: Replacing missing values with something like the **mean**, **median**, or **mode**

In my case, I didn't want to lose any important rows, so I used **mean imputation**, which replaces missing values with the average of the rest of the values in that column. Imputation helps us fill in missing data so we don't lose entire rows. The **mean method** is simple and useful for numeric columns.

We used the following code for each variable in which imputation needed to be:

pakistan_hunger_data_raw$Population_Under_Poverty[is.na(pakistan_hunger_data_raw$Population_Under_Poverty)] <- round(mean(pakistan_hunger_data_raw$Population_Under_Poverty, na.rm = TRUE), 2)

as this code is for Population_Under_Poverty, same code was used for other variables by just replacing there variable name in the syntax.

3.3 Handling Outliers

Outliers are values in a dataset that are very different from the rest. They can affect our analysis by pulling the average too high or too low, and can make our graphs look strange. So, in this step, I identified the outliers in key numeric variables and decided how to handle them properly. The following two ways were used for outliers detection:

- Box plots method
- IQR function based manual method

For box plots

we identified and mentioned the numeric variables/columns and using for loop created box plots for different variables and the code was as follows:

R Code Used:

```
numeric_cols <- pakistan_hunger_data_raw[, sapply(pakistan_hunger_data_raw, is.numeric)]
        par(mfrow = c(2, 3))
        for (col in names(numeric_cols)) {
         boxplot(numeric_cols[[col]], main = col, col = "lightblue")
        }
        par(mfrow = c(1, 1))
```

Method Used: IQR (Interquartile Range)

I used the **IQR method** to detect outliers. This method calculates the range between the 25th percentile (Q1) and 75th percentile (Q3), and then finds any values that are below Q1 - 1.5*IQR or above Q3 + 1.5*IQR.

R Code Used:

```
detect_outliers <- function(x) {
  Q1 <- quantile(x, 0.25, na.rm = TRUE)
  Q3 <- quantile(x, 0.75, na.rm = TRUE)
  IQR_val <- Q3 - Q1
  lower <- Q1 - 1.5 * IQR_val
  upper <- Q3 + 1.5 * IQR_val
  which(x < lower | x > upper)
}
```

Results:

The result for each numeric column showed:

```
integer(0)
```

This means that **no outliers were found** in any of the numeric columns.

Action Taken:

Since no outliers were detected, I did **not apply any further methods** such as:

- Winsorizing
- Log Transformation
- Removal

3.4 Data Transformation

Data transformation is one of the most important steps in preparing data for meaningful analysis. The raw data we collected may contain useful information, but it often needs to be cleaned, reshaped, and adjusted so it can be analyzed properly using statistical techniques like regression and visualization. In this step, I focused on three major types of transformation: creating a new variable, scaling existing variables, and preparing categorical variables for analysis.

Each transformation step below was carried out for a specific reason, and I've explained what I did, why I did it, and how I implemented it using R.

Step 1: Creating a New Variable – Hunger Severity Index

We created a new variable called **Hunger_Severity_Index** by taking the average of three existing columns: Malnutrition_Rate, Food_Insecurity, and Children_Underweight.

Each of these three indicators individually reflects some part of hunger:

- **Malnutrition Rate** shows the general nutritional health of the population.
- **Food Insecurity** tells us how many people don't have reliable access to enough food.
- **Children Underweight** indicates hunger and malnutrition among children, a sensitive and important measure.

By combining these three into a single **index**, I was able to create one clear, summarized variable that tells us **how severe the hunger problem is** in a particular city and year. This index will be very useful later when I want to compare cities, see trends, or use it as a dependent variable in regression analysis.

R Code:

```
pakistan_hunger_data_raw$Hunger_Severity_Index <- rowMeans(

  pakistan_hunger_data_raw[, c("Malnutrition_Rate", "Food_Insecurity",
"Children_Underweight")],

)
```

- rowMeans(...) calculates the average of selected columns for each row.
- The na.rm = TRUE part tells R to ignore missing values while calculating the average.
- This newly created column is saved in the dataset as Hunger_Severity_Index.

Step 2: Scaling / Standardizing Numeric Variables

We scaled two numeric variables:

- Access_to_Clean_Water
- Food_Production_Index

These two variables are important, but their **numerical ranges are very different**. For example, Access_to_Clean_Water might be between 60 to 95, and Food_Production_Index might go from 80 to 100. This creates a problem in regression models because variables with higher values might **unfairly influence the model** more than variables with smaller numbers.

To fix this, I used **standardization** (also called Z-score scaling), which changes the data so it has:

- A **mean (average)** of 0

- A **standard deviation** of 1

This way, every variable contributes **fairly and equally** in the analysis.

R Code:

```
pakistan_hunger_data_raw$Access_to_Clean_Water_Scaled <- scale(pakistan_hunger_data_raw$Access_to_Clean_Water)

pakistan_hunger_data_raw$Food_Production_Index_Scaled <- scale(pakistan_hunger_data_raw$Food_Production_Index)
```

- scale() is a built-in R function that standardizes numeric data.
- The result is saved as new columns with _Scaled added to the name.
- These standardized columns will later help in building models like regression, clustering, or PCA (if used).

Step 3: Preparing Categorical Variables

For this dataset, we had two categorical-like variables:

- City (character format)
- Year (numeric, but represents time)

For different steps like **grouped analysis** or **regression**, I converted them into **factors** using the as.factor() function. This makes it easier for R to treat them as **categories** instead of numbers.

Code :

```
pakistan_hunger_data_raw$City <- as.factor(pakistan_hunger_data_raw$City)

pakistan_hunger_data_raw$Year <- as.factor(pakistan_hunger_data_raw$Year)
```

Final Step: Saving the Transformed Dataset

After all transformations, I saved the updated dataset with a new name, so I can use it in the next phases (descriptive statistics and regression).

R Code:

```
write.csv(pakistan_hunger_data_raw, "pakistan_hunger_data_cleaned.csv", row.names = FALSE)
```

- write.csv(...) is used to export the dataset.
- "pakistan_hunger_data_cleaned.csv" is the name of the new file.
- row.names = FALSE makes sure R doesn't add row numbers as a column.

Summary Table of Transformations

Transformation Type	Variables Affected	Purpose/Reason
Created new variable	Malnutrition Rate, Food Insecurity, Children Underweight	To summarize hunger severity in one easy-to-understand score
Scaled numeric variables	Access to Clean Water, Food Production Index	To avoid bias in statistical models due to different value ranges
Prepared categories	City, Year	To help with grouped or regression analysis when needed
Saved dataset	All transformations	To preserve a clean version of the dataset for upcoming analyses

This step was very important because now my dataset is **clean**, **enhanced**, and **ready for meaningful analysis**. The transformations help to summarize data (like creating the Hunger Index), and make sure all variables are fairly treated in statistical modeling (like scaling). These transformations also prepare us for the next phases, where we'll do **descriptive statistics** and **regression modelling**.

Phase 2: Descriptive Statistics

This phase aims to understand the basic behavior and distribution of variables in the cleaned dataset. As a data analyst, this helps us check how the values are spread, which values are common or extreme, and whether the data is tilted or balanced. This is an essential step before moving to deeper analysis like regression or correlation.

Descriptive statistics help us:

- Understand the **general pattern** of each variable
- Know if a variable is **normally distributed** or **skewed**
- Identify if values are spread closely or vary a lot
- Get ready for more complex analysis like regression, where knowing variable distribution is necessary

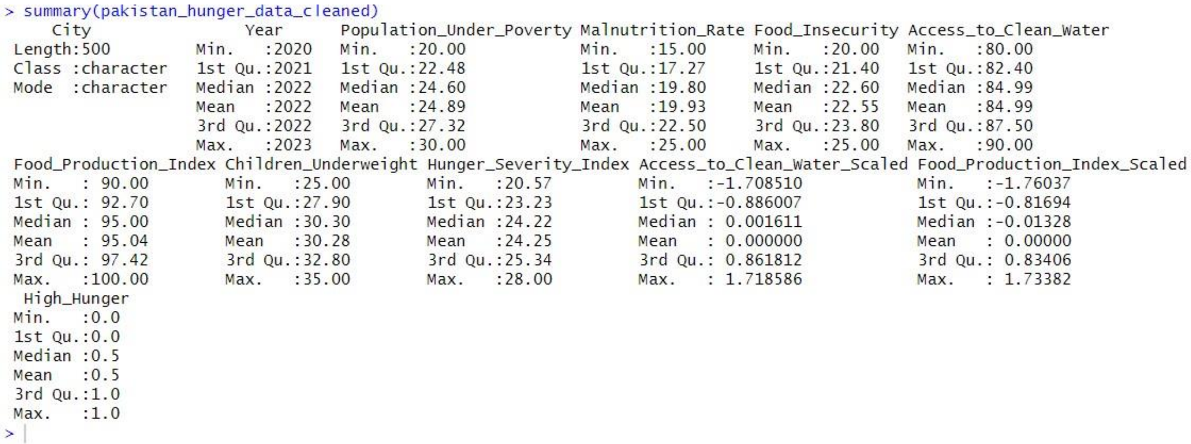
R Code for Descriptive Statistics

We used three different methods for descriptive statistics:

1. Basic Summary Using summary()

summary(pakistan_hunger_data_cleaned)

This shows **min, max, mean, median, and quartiles** of all numeric variables.



```
> summary(pakistan_hunger_data_cleaned)
     City                Year      Population_Under_Poverty Malnutrition_Rate Food_Insecurity Access_to_Clean_Water
 Length:500          Min.   :2020   Min.   :20.00            Min.   :15.00     Min.   :20.00   Min.   :80.00
 Class :character   1st Qu.:2021   1st Qu.:22.48            1st Qu.:17.27     1st Qu.:21.40   1st Qu.:82.40
 Mode  :character   Median :2022   Median :24.60            Median :19.80     Median :22.60   Median :84.99
                    Mean   :2022   Mean   :24.89            Mean   :19.93     Mean   :22.55   Mean   :84.99
                    3rd Qu.:2022   3rd Qu.:27.32            3rd Qu.:22.50     3rd Qu.:23.80   3rd Qu.:87.50
                    Max.   :2023   Max.   :30.00            Max.   :25.00     Max.   :25.00   Max.   :90.00
 Food_Production_Index Children_Underweight Hunger_Severity_Index Access_to_Clean_Water_Scaled Food_Production_Index_Scaled
 Min.   : 90.00        Min.   :25.00        Min.   :20.57         Min.   :-1.708510            Min.   :-1.76037
 1st Qu.: 92.70        1st Qu.:27.90        1st Qu.:23.23         1st Qu.:-0.886007            1st Qu.:-0.81694
 Median : 95.00        Median :30.30        Median :24.22         Median : 0.001611            Median :-0.01328
 Mean   : 95.04        Mean   :30.28        Mean   :24.25         Mean   : 0.000000            Mean   : 0.00000
 3rd Qu.: 97.42        3rd Qu.:32.80        3rd Qu.:25.34         3rd Qu.: 0.861812            3rd Qu.: 0.83406
 Max.   :100.00        Max.   :35.00        Max.   :28.00         Max.   : 1.718586            Max.   : 1.73382
  High_Hunger
 Min.   :0.0
 1st Qu.:0.0
 Median :0.5
 Mean   :0.5
 3rd Qu.:1.0
 Max.   :1.0
> |
```

2. Detailed Summary Using describe() from psych package

library(psych)

describe(pakistan_hunger_data_cleaned)

This gives us:

Term	Meaning
mean	Average value
sd	Standard deviation – how spread out the values are
median	Middle value
mad	Median Absolute Deviation – robust way to see spread
min/max	Lowest and highest values
range	Difference between max and min
skew	Skewness – tells if data is tilted (left or right)
kurtosis	Tells if data is sharply peaked or flat
se	Standard Error – tells how stable the mean is

We used the describe() function in R to get a quick overview of my pakistan_hunger_data_cleaned dataset. This function is super helpful because it gives you a bunch of summary statistics for each variable without having to calculate them all individually.

Looking at the output, the first column, vars, just tells you which number variable it is in the dataset. n shows how many observations (or rows) there are for each variable – in my case, it's 500 for most of them, which is good because it means I have a decent amount of data.

Then we have mean, which is the average value for each variable. For instance, the average Population_Under_Poverty is about 24.89, and the average Malnutrition_Rate is around 19.93. The sd column gives us the standard deviation, which tells us how spread out the data is from the mean. A larger standard deviation means the data points are more spread out.

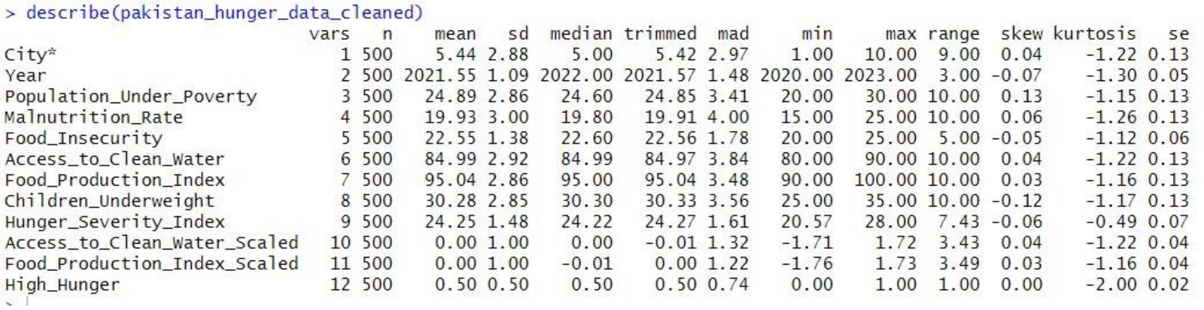
median is the middle value when all the numbers are ordered, and it's often a good measure of the center, especially if there are extreme values. trimmed is similar to the mean but it ignores a small percentage of the highest and lowest values, which can give a more representative average if there are outliers.

mad stands for Median Absolute Deviation, and it's another way to measure the spread of the data, but it's less affected by extreme values than the standard deviation. The min and max columns are pretty straightforward - they show the smallest and largest values in each variable, which helps me see the range of my data. range is simply the difference between the max and min.

Skew tells us about the symmetry of the data distribution. A skew value close to zero means the data is fairly symmetrical. Positive skew means the data has a 'tail' to the right (more high values), and negative skew means a 'tail' to the left (more low values). For example, Hunger_Severity_Index has a slight negative skew (-0.06).

Kurtosis describes the 'tailedness' of the distribution. A value around 0 (like for a normal distribution) means the data has typical tails. Positive kurtosis means the data has heavier tails (more outliers), and negative kurtosis means lighter tails. Many of my variables, like City and Year, show negative kurtosis, meaning they have relatively lighter tails.

Finally, **se** is the standard error, which estimates how much the sample mean is likely to vary from the true population mean. It's a measure of the precision of my sample mean.



```
> describe(pakistan_hunger_data_cleaned)
                            vars   n    mean   sd  median trimmed  mad     min     max range  skew kurtosis   se
City*                          1 500    5.44 2.88    5.00    5.42 2.97    1.00   10.00  9.00  0.04    -1.22 0.13
Year                           2 500 2021.55 1.09 2022.00 2021.57 1.48 2020.00 2023.00  3.00 -0.07    -1.30 0.05
Population_Under_Poverty       3 500   24.89 2.86   24.60   24.85 3.41   20.00   30.00 10.00  0.13    -1.15 0.13
Malnutrition_Rate              4 500   19.93 3.00   19.80   19.91 4.00   15.00   25.00 10.00  0.06    -1.26 0.13
Food_Insecurity                5 500   22.55 1.38   22.60   22.56 1.78   20.00   25.00  5.00 -0.05    -1.12 0.06
Access_to_Clean_Water          6 500   84.99 2.92   84.99   84.97 3.84   80.00   90.00 10.00  0.04    -1.22 0.13
Food_Production_Index          7 500   95.04 2.86   95.00   95.04 3.48   90.00  100.00 10.00  0.03    -1.16 0.13
Children_Underweight           8 500   30.28 2.85   30.30   30.33 3.56   25.00   35.00 10.00 -0.12    -1.17 0.13
Hunger_Severity_Index          9 500   24.25 1.48   24.22   24.27 1.61   20.57   28.00  7.43 -0.06    -0.49 0.07
Access_to_Clean_Water_Scaled  10 500    0.00 1.00    0.00   -0.01 1.32   -1.71    1.72  3.43  0.04    -1.22 0.04
Food_Production_Index_Scaled  11 500    0.00 1.00   -0.01    0.00 1.22   -1.76    1.73  3.49  0.03    -1.16 0.04
High_Hunger                   12 500    0.50 0.50    0.50    0.50 0.74    0.00    1.00  1.00  0.00    -2.00 0.02
```

3. Smart Summary Using skim() from skimr package

library(skimr)

skim(pakistan_hunger_data_cleaned)

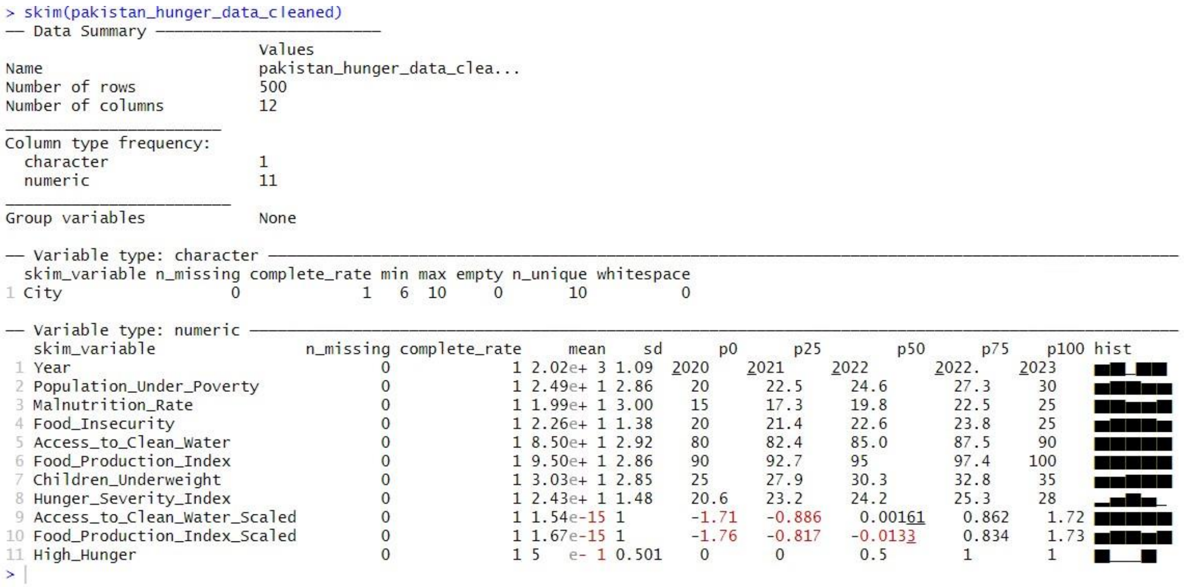
It shows:

- **Variable type**
- **Missing values**
- **Distribution with histograms**
- **Mean, Median, Min, Max, SD, and Percentiles (P25, P75)**

This function gives a **quick and easy-to-read overview** of all variables in one place.

Missing Values

- No major missing values now — because we cleaned them in Phase 3.



```
> skim(pakistan_hunger_data_cleaned)
── Data Summary ────────────────────────
                           Values
Name                       pakistan_hunger_data_clea...
Number of rows             500
Number of columns          12
_______________________
Column type frequency:
  character                1
  numeric                  11
________________________
Group variables            None

── Variable type: character ──────────────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate min max empty n_unique whitespace
1 City                  0             1   6  10     0       10          0

── Variable type: numeric ────────────────────────────────────────────────────────────────────────
   skim_variable               n_missing complete_rate     mean    sd   p0      p25       p50      p75    p100 hist
 1 Year                                0             1 2.02e+ 3 1.09 2020     2021      2022       2022.    2023 ▅▇▁▇▇
 2 Population_Under_Poverty            0             1 2.49e+ 1 2.86   20     22.5      24.6       27.3     30 ▆▇▇▆▆
 3 Malnutrition_Rate                   0             1 1.99e+ 1 3.00   15     17.3      19.8       22.5     25 ▇▇▆▆▇
 4 Food_Insecurity                     0             1 2.26e+ 1 1.38   20     21.4      22.6       23.8     25 ▆▇▇▇▆
 5 Access_to_Clean_Water               0             1 8.50e+ 1 2.92   80     82.4      85.0       87.5     90 ▇▇▇▇▇
 6 Food_Production_Index               0             1 9.50e+ 1 2.86   90     92.7      95         97.4    100 ▇▇▇▇▇
 7 Children_Underweight                0             1 3.03e+ 1 2.85   25     27.9      30.3       32.8     35 ▆▆▇▇▇
 8 Hunger_Severity_Index               0             1 2.43e+ 1 1.48   20.6   23.2      24.2       25.3     28 ▁▃▇▅▁
 9 Access_to_Clean_Water_Scaled        0             1 1.54e-15 1      -1.71  -0.886     0.00161   0.862    1.72 ▇▇▇▇▇
10 Food_Production_Index_Scaled        0             1 1.67e-15 1      -1.76  -0.817    -0.0133    0.834    1.73 ▆▇▇▆▇
11 High_Hunger                         0             1 5    e- 1 0.501  0      0         0.5        1        1   ▇▁▁▁▇
> |
```

This phase helped us understand the dataset better:

- We saw what values are common (mean, median)
- How much they vary (SD, range)
- If they are skewed or peaked (skewness & kurtosis)
- Whether variables are balanced enough for analysis

All of this is **very important** before doing regression, correlation, or model building. We are now fully ready to move to deeper statistical analysis in the next phases.

Let me know when you're ready to proceed to the next phase. I'll help you with detailed guidance just like this.

Phase 4: Exploratory Data Analysis (EDA)

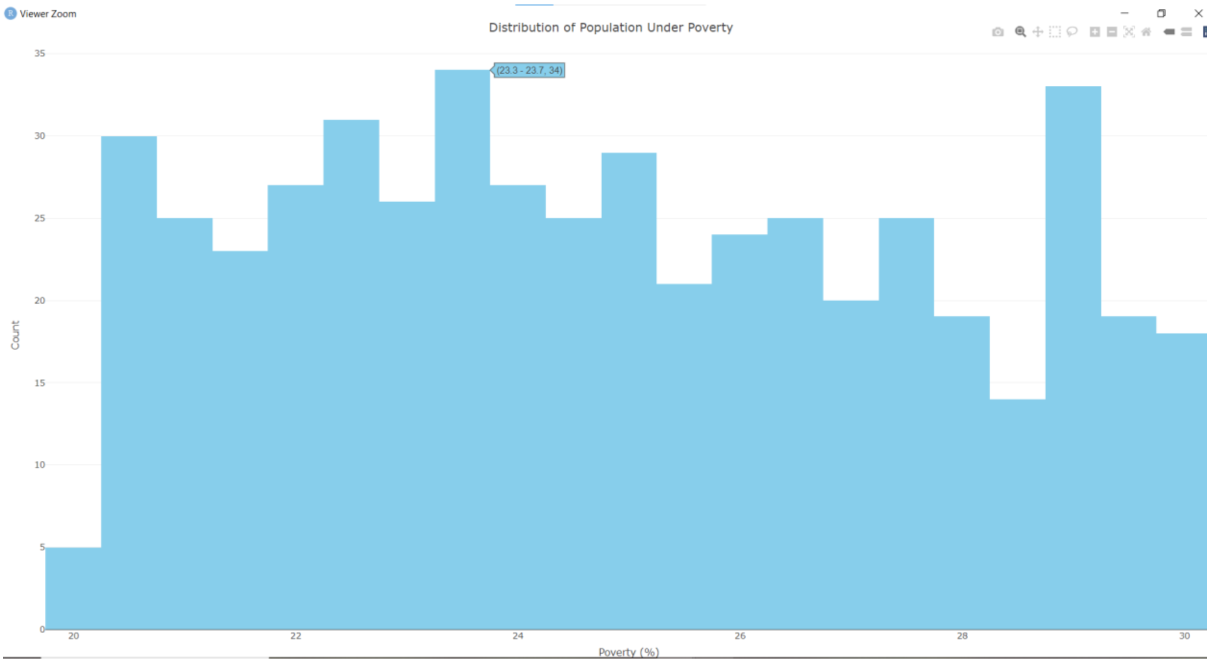
Step 1: Univariate Analysis using Plotly (R-Based Report)

Univariate analysis focuses on understanding each variable individually without examining relationships between them. The purpose is to explore the structure and distribution of data, such as how values are spread, where they are concentrated, and whether there are extreme values or outliers. This step is essential to gain an initial understanding of the dataset before moving to more complex analyses. I used interactive visualizations through the Plotly library in R, selecting appropriate charts based on the nature of each variable—such as histograms, box plots, density plots, bar charts, and pie charts.

For Numeric Variables

Population Under Poverty (%)

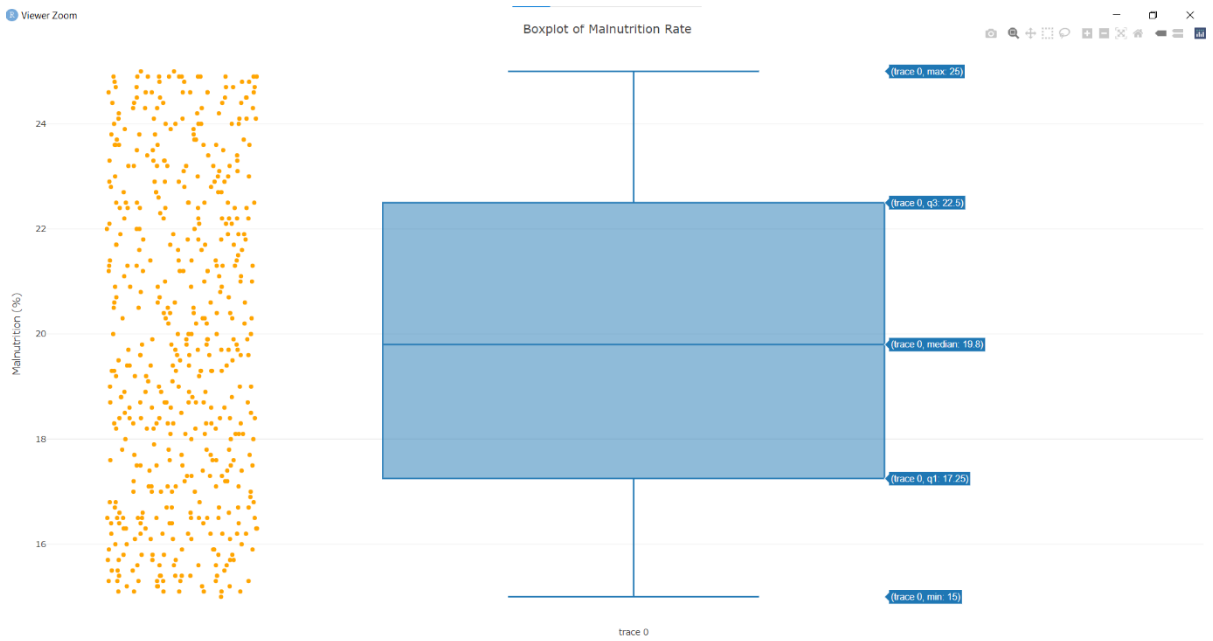
For this variable, I used a histogram to observe how poverty levels are distributed across different cities in Pakistan. The histogram revealed how many cities fall within certain poverty percentage ranges. This helped in identifying whether poverty is concentrated in certain ranges (e.g., most cities below or above 30%). This visualization was important to get an idea of whether poverty was widespread or concentrated in specific areas.



To understand the **Population_Under_Poverty** variable better, I created a histogram to visualize its distribution. Looking at the "Distribution of Population Under Poverty" graph, I can see a clear picture of where most of my data points for poverty percentages fall. The x-axis shows the poverty percentage, ranging roughly from just under 20% to just over 30%. The y-axis, labeled "Count," tells me how many times a certain poverty percentage appeared in my dataset. What immediately stands out is a tall bar between approximately 23% and 24% (specifically marked as 23.3-23.7), indicating that this range of poverty percentages is the most common, with about 34 observations falling into this category. This suggests that a significant portion of the areas or time periods in my data experience poverty rates in this specific range. While this is the highest peak, I also noticed other concentrations, like around 20-21% and again around 28-29%, meaning these percentages are also quite frequent, though not as dominant as the mid-20s. On the other hand, the very low and very high poverty percentages (close to 20% or 30%) appear less often. Overall, the histogram shows that the population under poverty in my dataset is not evenly spread out but tends to cluster around a few key percentages, with the most prevalent being in the low to mid-20s.

2. Malnutrition Rate (%)

A box plot was used to visualize the spread of malnutrition rates. It clearly showed the median malnutrition percentage along with the upper and lower quartiles. The chart also helped in identifying outliers—cities with unusually high or low malnutrition rates. This was useful in understanding how severe malnutrition is and how consistently it is affecting different regions.



Looking at the 'Boxplot of Malnutrition Rate,' the y-axis represents the Malnutrition Rate in percentages. The box in the middle tells us a lot about where most of the data lies. The line inside the box is the median, which for my data is exactly 19.8%. This means that half of the observations have a malnutrition rate below 19.8%, and half are above it.

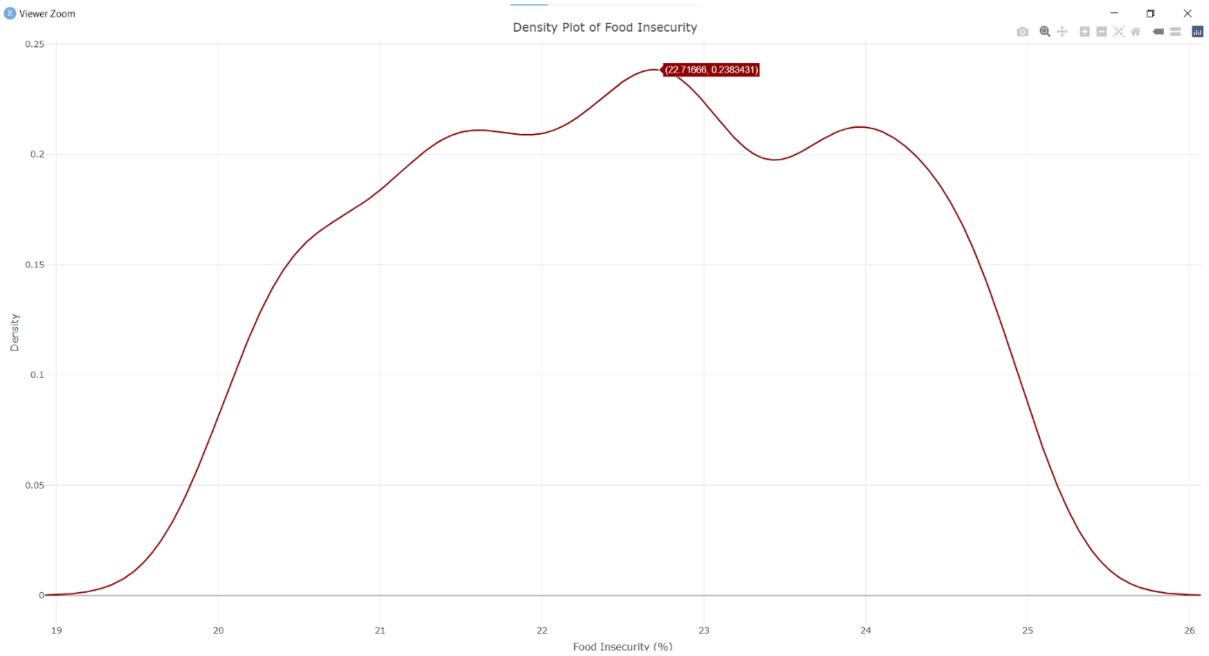
The bottom edge of the box represents the first quartile (Q1), which is 17.25%. This means 25% of the data falls below this point. The top edge of the box is the third quartile (Q3), at 22.5%. This means 75% of the data falls below this point, or conversely, 25% of the data is above it. The height of the box itself, from Q1 to Q3, is the Interquartile Range (IQR), which shows where the middle 50% of my data is concentrated. Here, the IQR is 22.5−17.25=5.25%.

The "whiskers" extending from the box show the typical range of the data, excluding any extreme outliers. The bottom whisker reaches down to the minimum value, which is 15%. The top whisker extends up to the maximum value, which is 25%. What's good to see here is that there are no individual points outside the whiskers, which means my Malnutrition_Rate data doesn't have any obvious outliers that are extremely high or low compared to the rest of the observations. This indicates a fairly contained and consistent spread of malnutrition rates within the dataset.

The scatter of orange dots on the left side is a 'jittered' plot of the individual data points. This helps to visually confirm the density of the data at different malnutrition rate levels and reinforces what the boxplot is showing about the spread."

3. Food Insecurity (%)

For food insecurity, I selected a density plot. This type of plot provided a smooth curve showing where most of the data values are concentrated. It helped to understand whether the distribution is symmetric, skewed to one side, or has multiple peaks. This gave a clearer picture of whether food insecurity is a general problem or more severe in specific ranges.



This plot is excellent for visualizing the shape of the data's distribution, almost like a continuous version of a histogram, and helps us see where the data is most concentrated.

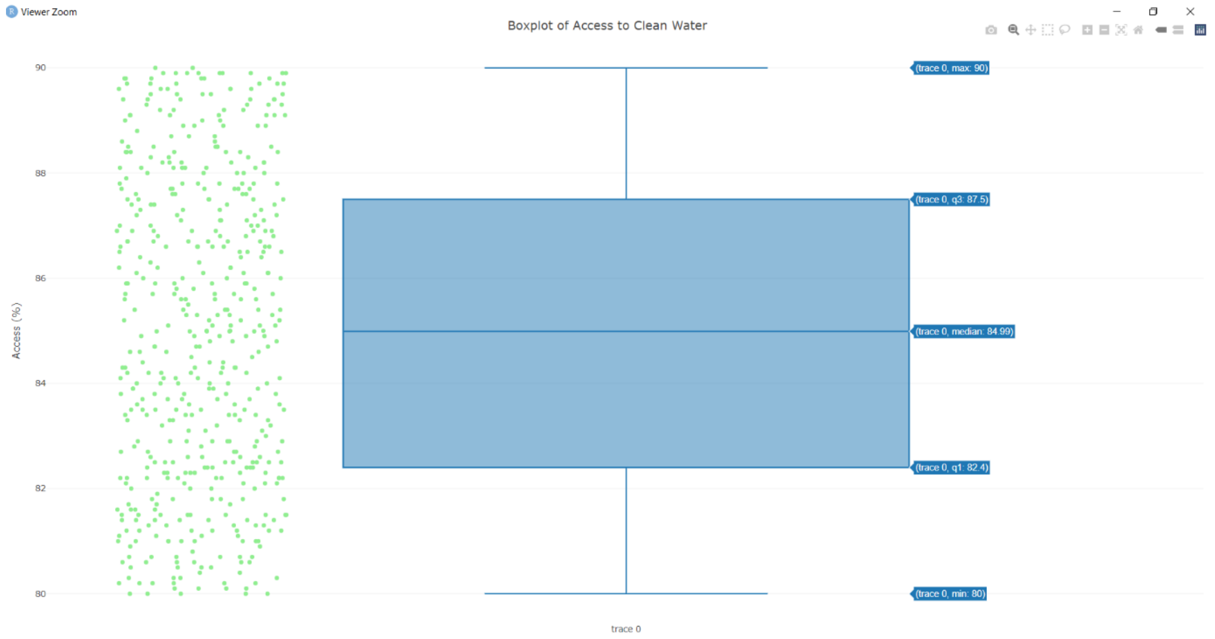
Looking at the **'Density Plot of Food Insecurity,'** the x-axis represents the Food Insecurity percentage, ranging from about 19% to just over 26%. The y-axis, labeled 'Density,' indicates the probability density, with higher points on the curve showing where the values are more concentrated. The plot shows a multi-modal distribution, meaning it has several peaks, which suggests that food insecurity rates are not just clustered around one single value. There's a prominent peak around 22.79% (specifically highlighted as 22.7996, 0.2363431), which indicates that this percentage of food insecurity is the most frequently observed in my dataset. This is where the density is highest. However, it's not the only peak. I can also see other noticeable bumps, like one around 21% and another smaller one around 24%. This tells me that there are other common food insecurity levels in the data, in addition to the main peak.

The curve shows that most of the food insecurity percentages fall between approximately 20% and 25%. Values outside this range, especially below 20% and above 25%, become less frequent, as indicated by the curve dropping closer to zero. The tails of the distribution are relatively short, meaning extremely low or high food insecurity percentages are rare in my data.

Overall, this density plot gives me a smooth visual summary, showing that while there's a primary concentration of food insecurity around 22.8%, there are also other distinct ranges where food insecurity rates frequently occur."

4. Access to Clean Water (%)

A box plot was used for this variable to examine the distribution and presence of outliers. It showed how access to clean water varies among cities. The plot highlighted both the middle range and the extremes, helping to understand whether most areas have good access or if there are significant disparities.



To see how Access_to_Clean_Water is distributed in my dataset, I made a boxplot. This chart is great because it quickly shows me the typical range of values, where the middle of the data is, and if there are any really unusual numbers.

Looking at the 'Boxplot of Access to Clean Water,' the y-axis shows the percentage of access, ranging from about 80% to 90%. The main part of the boxplot, the blue box, tells us a lot. The line right in the middle of the box is the median, which is 84.99%. This means that half of the observations in my data have clean water access below 84.99%, and the other half have it above this percentage.

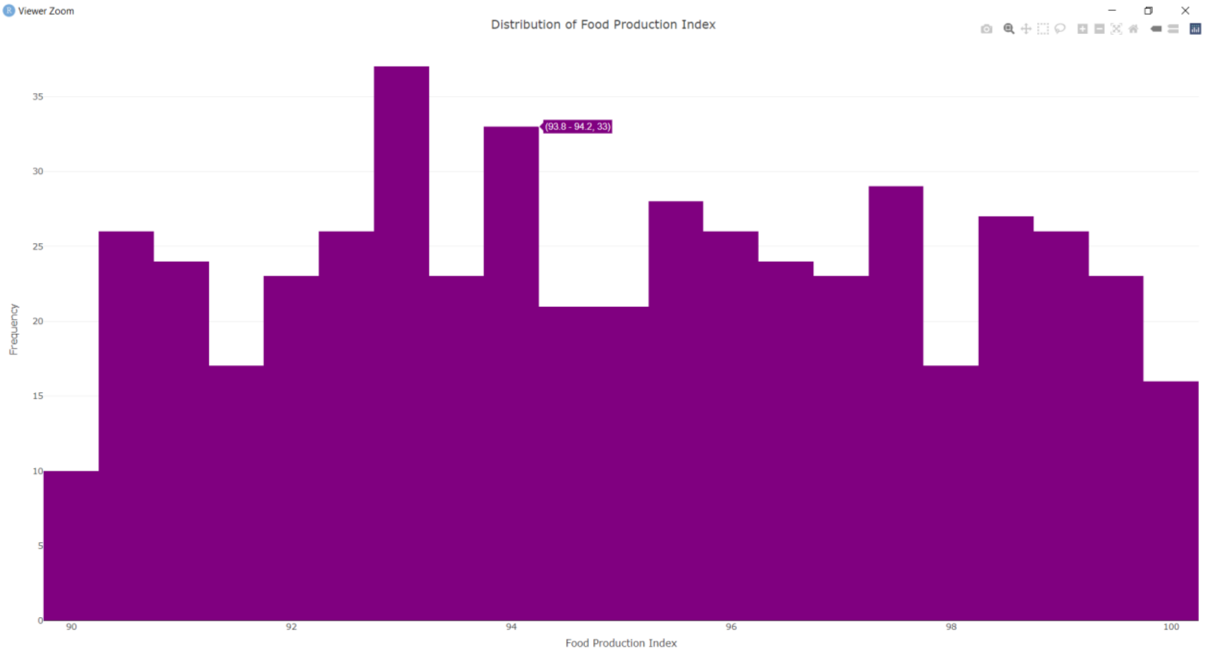
The bottom edge of the box is the first quartile (Q1), at 82.4%. This tells me that 25% of the data points for clean water access are at or below this level. The top edge of the box is the third quartile (Q3), at 87.5%. This means 75% of the data points are at or below 87.5%, or conversely, the top 25% of access rates are above it. The height of the box itself, from Q1 to Q3, shows where the middle 50% of my data for clean water access is concentrated.

The "whiskers" that extend from the box indicate the general spread of the data. The bottom whisker reaches down to the minimum value of 80%, and the top whisker goes up to the maximum value of 90%. What's really good to note is that there are no individual points shown outside these whiskers. This is important because it means there aren't any extreme outliers in my Access_to_Clean_Water data - all the values fall within the expected range, suggesting a pretty consistent level of access.

The green dots scattered on the left are just individual data points, visually confirming the density and spread of the access percentages, and they line up nicely with what the boxplot is showing.

5. Food Production Index

A histogram was used to study the distribution of the food production index. This chart allowed me to see how frequently each range of food production values occurred. It helped in identifying whether food production was high in most regions or if a majority of cities had low production, indicating possible concerns in food availability.



Looking at the 'Distribution of Food Production Index,' the x-axis shows the Food Production Index values, ranging from about 90 to 100. The y-axis, labeled 'Frequency,' tells me how many observations fall within each specific range or 'bin' of the index.

What immediately jumps out is a very tall bar around the **93.8 to 94.2** mark (specifically highlighted as 93.8 - 94.2, 33). This indicates that food production index values within this range are the most frequent in my dataset, with 33 observations falling into it. This suggests that this index level is very common.

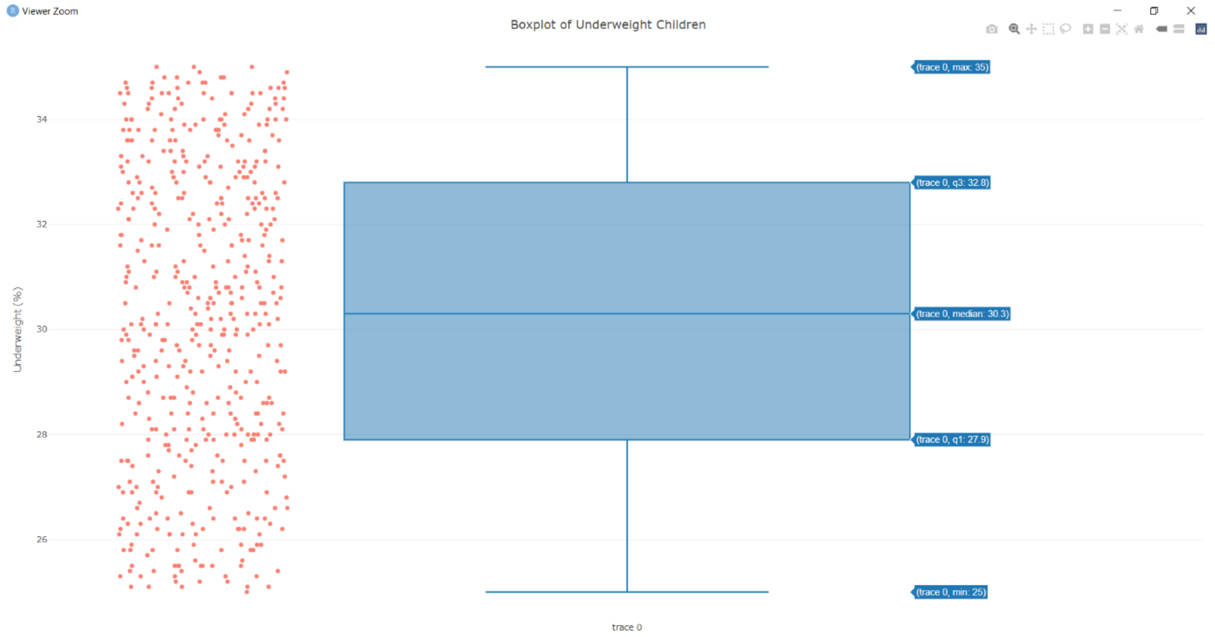
However, it's not the only common range. I can also see other significant peaks. For example, there's a strong concentration of values around 92 and another one closer to 98. This tells me that the Food Production Index isn't just centered around one value; it has multiple areas where observations tend to cluster.

The histogram doesn't show a perfectly smooth, bell-shaped curve, which means the distribution isn't perfectly normal. Instead, it appears to be multi-modal, with several distinct peaks. This might suggest different patterns or levels of food production index across the various regions or time periods in my data.

The distribution covers the entire range from 90 to 100, but some areas, like around 95 or 96, have fewer observations compared to the peaks. Overall, this histogram gives me a clear visual summary of how frequently different food production index values occur in my dataset, highlighting the most common ranges and showing the overall spread."

6. Children Underweight (%)

This variable was analyzed using a box plot to assess the distribution and to identify any outliers. It showed the percentage of children who are underweight across different areas. The chart was useful in understanding whether this issue is widespread or limited to specific cities with high rates.



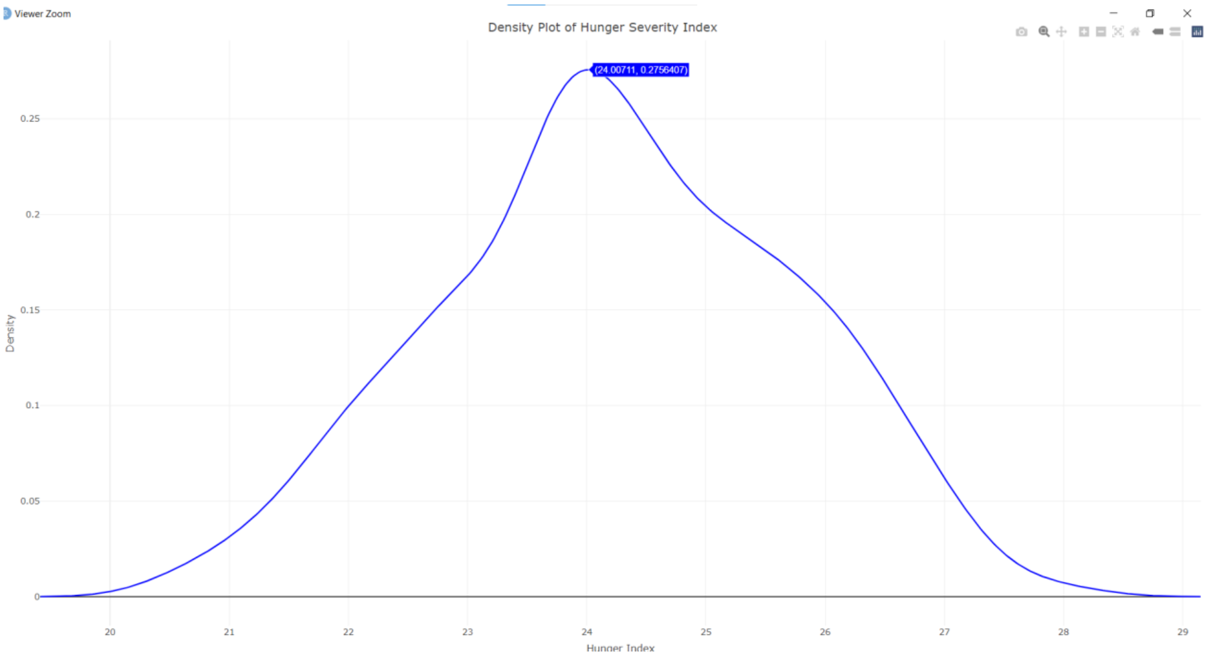
To examine the Children_Underweight variable in my dataset, I created a boxplot. Looking at the 'Boxplot of Underweight Children,' the y-axis represents the percentage of underweight children, ranging from about 25% to 35%. The main part of the boxplot, the blue box, gives us a lot of information. The solid line inside the box is the median, which for my data is 30.3%. This means that half of the observations in my dataset have an underweight children percentage below 30.3%, and the other half are above it.

The bottom edge of the box represents the first quartile (Q1), which is 27.9%. This tells me that 25% of the data points for underweight children are at or below this level. The top edge of the box is the third quartile (Q3), at 32.8%. This means 75% of the data points are at or below 32.8%, or conversely, the top 25% of underweight percentages are above it. The height of the box itself, from Q1 to Q3, shows where the middle 50% of my data for underweight children is concentrated.

The "whiskers" extending from the box indicate the typical range of the data. The bottom whisker reaches down to the minimum value of 25%, and the top whisker extends up to the maximum value of 35%. Importantly, there are no individual points shown outside these whiskers. This is a good sign because it means there aren't any extreme outliers in my Children_Underweight data that are far outside the usual range, suggesting a relatively consistent spread of underweight percentages. The scattered orange dots on the left side are a 'jittered' plot of the individual data points. This helps to visually confirm the density of the data at different underweight percentages and reinforces what the boxplot is showing about the spread."

7. Hunger Severity Index

I used a density plot for this variable to understand the overall shape and spread of hunger severity scores. This smooth curve provided an overview of where most of the values are located and whether the distribution is skewed. It helped identify the intensity of hunger issues across cities and whether most places have moderate or severe hunger levels.



Looking at the 'Density Plot of Hunger Severity Index,' the x-axis represents the Hunger Index values, ranging from about 19 to 29. The y-axis, labeled 'Density,' indicates the probability density, with higher points on the curve showing where the values are more concentrated.

The plot shows a fairly smooth, unimodal distribution, meaning it has one main peak. This peak is located around 24.0071 (specifically highlighted with coordinates 24.0071, 0.2759407), indicating that this hunger index value is the most frequently observed in my dataset. This is where the highest concentration of data points for hunger severity lies.

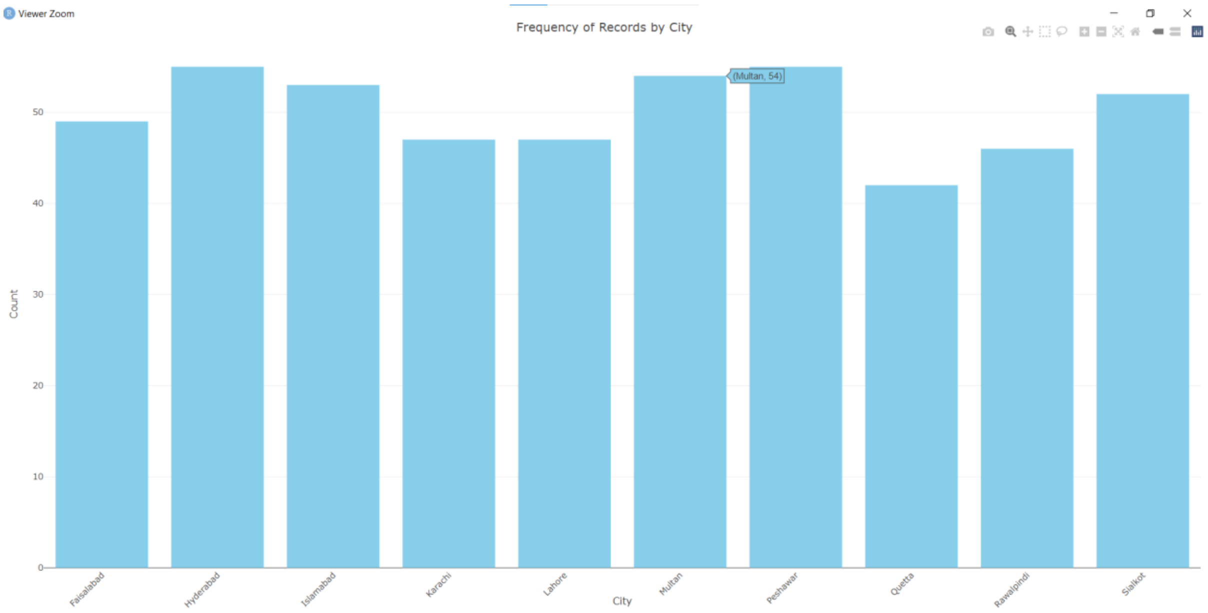
The curve rises steadily towards this peak and then gradually declines, forming a shape somewhat similar to a bell curve, though it might be slightly skewed. Most of the hunger severity index values appear to fall between approximately 22 and 27. Values outside this range, especially below 20 and above 28, become less frequent, as the curve drops closer to zero at the tails. The tails of the distribution are relatively long but taper off smoothly, suggesting that extremely low or high hunger severity index values are less common but still present.

Overall, this density plot gives me a clear visual summary, showing that hunger severity in my dataset is primarily concentrated around an index of 24, with values becoming less common as they move further away from this central point in either direction."

Categorical Variables

8. City

The “City” variable was analyzed using a bar chart. This visualization was ideal because there are many unique cities in the dataset, and a pie chart would have been cluttered and hard to interpret. The bar chart clearly showed how many records exist for each city. It helped assess whether data is evenly distributed across locations or focused more on specific urban or rural areas.



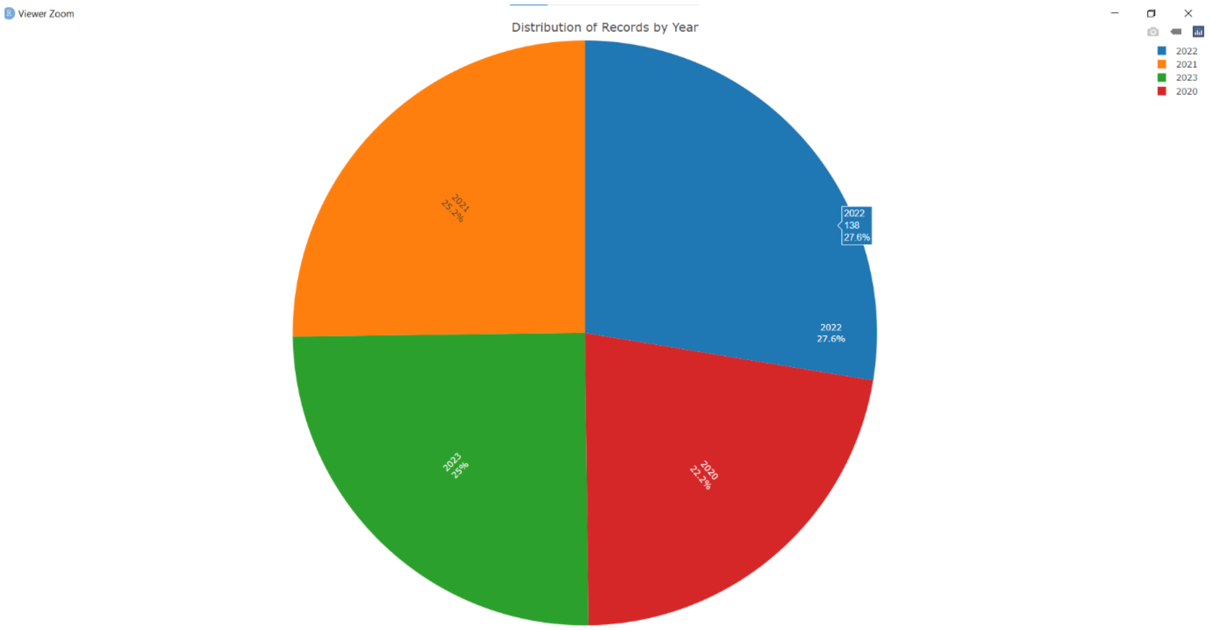
Looking at the chart, the x-axis lists various cities, and the y-axis, labeled 'Count,' shows how many records (or data points) I have for each city. I can see that my dataset includes information from several cities like Faisalabad, Hyderabad, Islamabad, Karachi, Lahore, Multan, Peshawar, Quetta, Rawalpindi, and Sialkot. What stands out is that the number of records isn't exactly the same for every city, but it's quite balanced across most of them, hovering around the 40 to 55 mark. For example, **Multan** has the highest number of records, with 54 observations, closely followed by Hyderabad and Islamabad. Peshawar also has a high count.

On the other hand, cities like **Quetta** and **Rawalpindi** have slightly fewer records compared to the others, with Quetta having the lowest count among those shown. Even so, they still have a good number of observations (over 40 each).

This chart tells me that my dataset has pretty good coverage across these major cities. While there are slight variations in the number of records per city, it's not like one city dominates the data excessively, which is good for ensuring a somewhat balanced representation when I analyze trends or differences across cities."

9. Year

For the “Year” variable, I used a pie chart since there were only a few unique years in the dataset. This made the pie chart clean, simple, and informative. It helped understand the proportion of data from each year and revealed whether the dataset was balanced across time or if certain years had more data. This is important to ensure that trends over time are based on balanced data.



To see how my data is distributed over different years, I created a pie chart titled 'Distribution of Records by Year.' This chart is really helpful for quickly understanding what percentage of my total dataset each year contributes.

Looking at the pie chart, I can see that my data covers four different years: 2020, 2021, 2022, and 2023. Each slice of the pie represents a year, and its size shows its proportion of the total records.

The largest slice belongs to 2022, accounting for about 27.6% of all records (with 138 observations). This means that a significant chunk of my data comes from this year.

Closely following 2022, 2021 represents 25.4% of the data, and 2023 makes up 25%. These years are very similar in terms of how many records they contribute to the dataset.

Finally, 2020 accounts for 22% of the records, which is the smallest slice among the four years, but still a substantial portion.

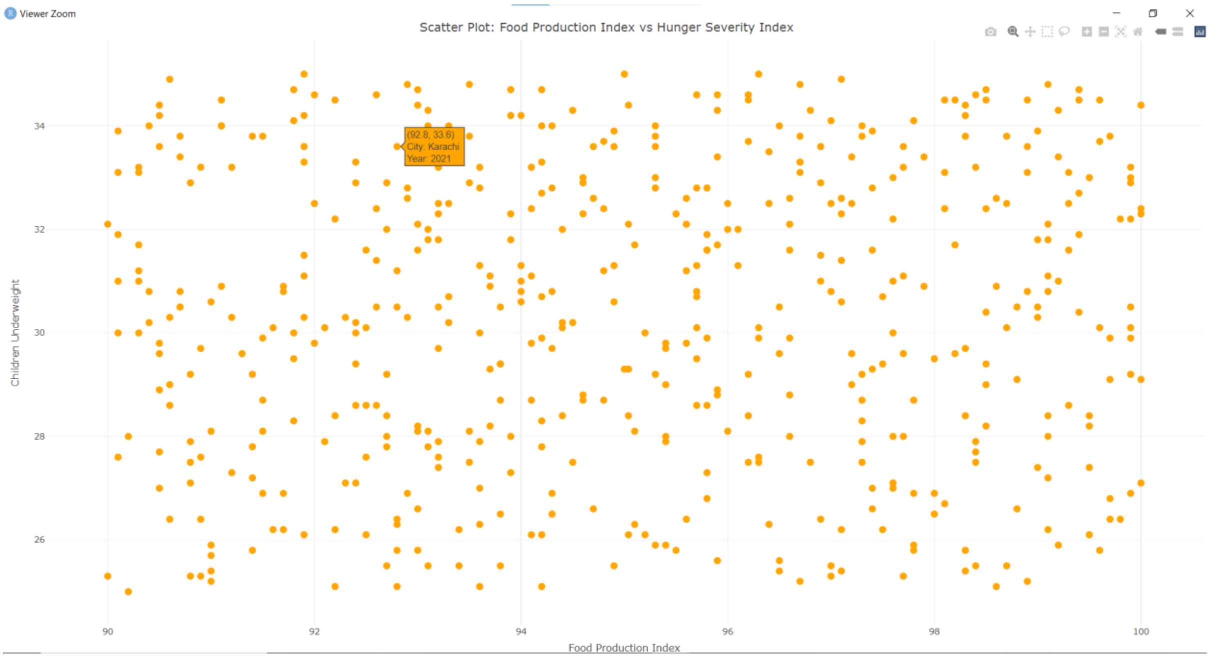
Overall, this pie chart clearly shows that my data is quite evenly distributed across these four years, with 2022 having a slightly larger representation. This even distribution is good because it means my analysis won't be heavily biased towards just one or two specific years."

Step 2: Bivariate Analysis:

Bivariate analysis helps in exploring the relationship between two variables. This step allowed me to better understand how one factor may influence another in the context of hunger and poverty in Pakistan. I used Plotly in R to create interactive and clear visualizations. I chose different graph types based on the nature of the variables—numeric vs. numeric, numeric vs. categorical, and categorical vs. categorical.

Scatter Plot – Food Production Index vs Children Underweight

To begin with, I created a scatter plot to examine the relationship between the **Food Production Index** and Children Underweight, both of which are numeric variables.



Looking at the 'Scatter Plot: Food Production Index vs Children Underweight,' the x-axis represents the Food Production Index, and the y-axis represents the percentage of Children Underweight. Each orange dot on the plot is an individual data point, representing a specific observation in my dataset (likely from a particular city in a given year). For example, one highlighted point shows a Food Production Index of 92.8, 33.6% Children Underweight, from Karachi in 2021.

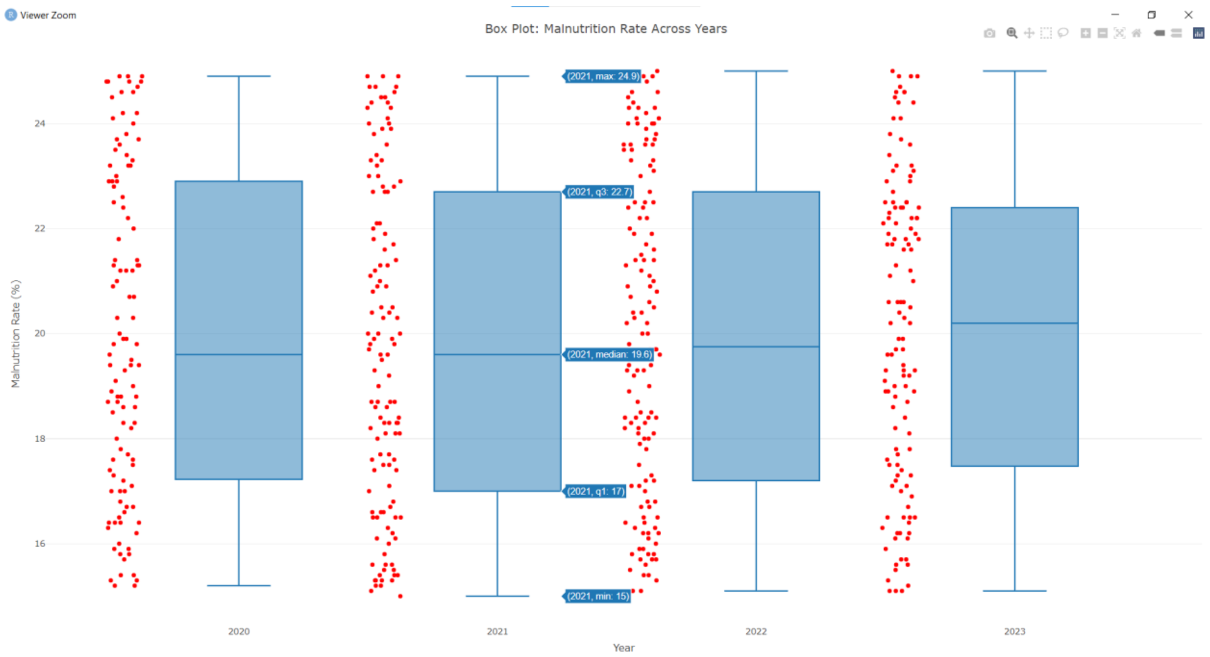
When I look at the overall pattern of the dots, they appear to be scattered pretty widely across the plot. I don't see a clear upward trend (where both go up together) or a clear downward trend (where one goes up as the other goes down). The dots seem to be spread out without forming a distinct line or curve.

This suggests that, based on this visual, there isn't a strong linear correlation between the Food Production Index and the percentage of Children Underweight in my dataset. In simpler terms, higher food production doesn't seem to consistently lead to lower (or higher) rates of underweight children, and vice-versa, when just looking at this plot. The relationship, if any, might be very weak, non-linear, or influenced by other factors not shown here.

So, this plot indicates that changes in the Food Production Index alone might not be a strong predictor of the percentage of underweight children based on the visual evidence.

Box Plot – Malnutrition Rate Across Years

Next, I used a box plot to compare the **Malnutrition Rate** (a numeric variable) across different **Years** (a categorical variable). This helped me understand how malnutrition has changed over time. Each year had its own box, showing the median, quartiles, and outliers. This was useful for identifying if malnutrition improved or worsened in specific years. The variation between years indicated that certain time periods may have experienced more severe nutrition-related issues, possibly due to economic or environmental factors.



Looking at the plot, the x-axis shows the years (2020, 2021, 2022, 2023), and the y-axis represents the Malnutrition Rate in percentages. Each year has its own box and whiskers, with individual data points shown as red dots to the left.

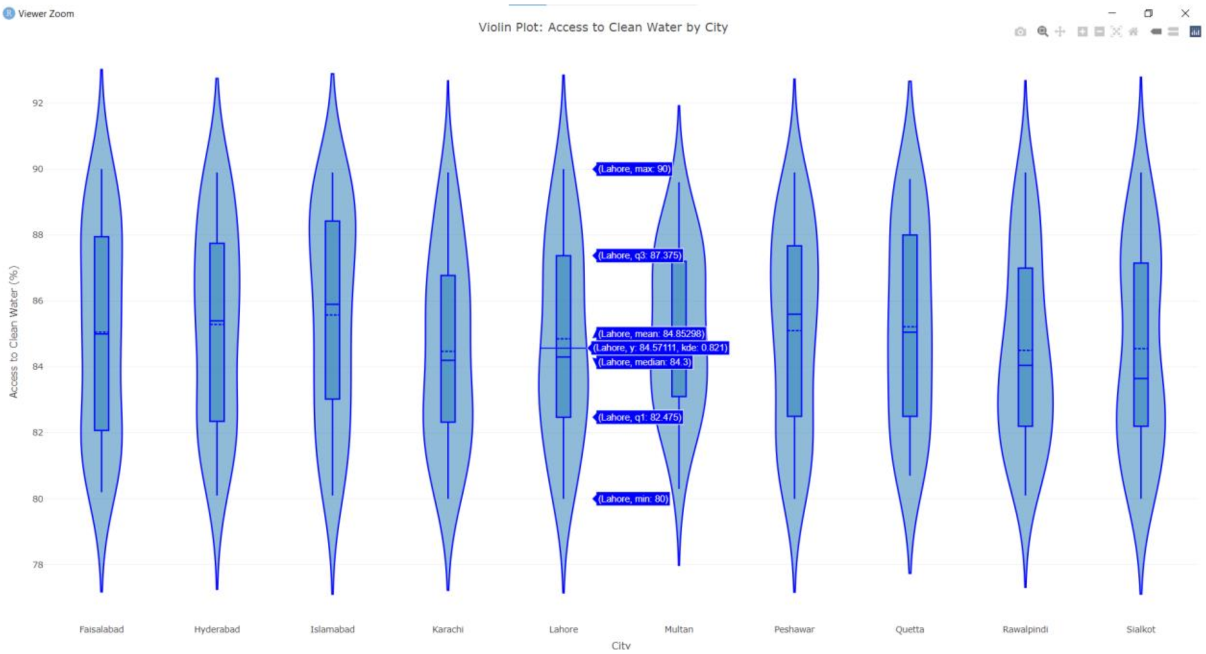
Here's what I observe:

- Overall Malnutrition Range: The rates generally fall between 15% and about 25% across all years.
- Medians are Similar: The median (the line inside the box) for each year is quite consistent. For 2021, the median is 19.8%, and the other years also have medians very close to this value. This suggests that the typical (middle) malnutrition rate hasn't changed drastically from year to year in my data.
- Interquartile Ranges (IQR) are Consistent: The height of the boxes, which represents the middle 50% of the data, is also pretty similar across all years. For example, in 2021, the middle 50% of malnutrition rates are between 17% (Q1) and 22.7% (Q3). This indicates that the spread of the central data points remains relatively stable.
- Whiskers and No Outliers: The whiskers for all years extend to cover nearly the full range from 15% to 25%, and importantly, there are no individual points (red dots) visible beyond the whiskers in any of the years. This means there are no extreme outliers in terms of unusually high or low malnutrition rates for any specific year that stand far apart from the rest of the data.
- Slight Fluctuations: While generally consistent, there might be very subtle shifts. For instance, the box for 2023 appears to be slightly lower than 2020, suggesting a minor downward trend in the central malnutrition rates, though this is not a dramatic change.

Overall, this box plot reveals that the malnutrition rates in my dataset have been remarkably stable across the years 2020 to 2023. The typical rates, their spread, and the overall range have remained quite consistent, with no major fluctuations or extreme outliers in any given year."

Violin Plot – Access to Clean Water by City

To study the distribution of **Access to Clean Water** across various **Cities**, I created a violin plot. This plot gave a detailed look at how this numeric variable varied within each categorical city. The violin plot not only showed the median and range like a box plot but also included the density, which helped me see where most of the values were concentrated. I noticed that some cities had a wider spread of values, meaning there was a big difference in access levels within that city, while others were more consistent.



Looking at the 'Violin Plot: Access to Clean Water by City,' the x-axis lists various cities, and the y-axis represents the Access to Clean Water percentage. Each "violin" shape represents a city, with the wider sections indicating a higher density of data points at that particular access percentage. Inside each violin, there's also a miniature boxplot showing the median, quartiles, and range.

Here's what I observe:

- **Overall Access Range:** Most cities show clean water access rates generally ranging from about 78% to 92%.
- **Varying Distributions:** While many cities (like Faisalabad, Hyderabad, Islamabad, Peshawar, Quetta, Rawalpindi, and Sialkot) have fairly similar, somewhat symmetrical violin shapes, some cities show distinct patterns:
    - **Lahore (highlighted):** The violin for Lahore is quite wide around the 80% mark, then narrows considerably before widening again around 84-85% and again near 90%. This suggests that while its median is 84.3%, there might be multiple clusters of access rates within Lahore data, or perhaps a slightly skewed distribution. The inner boxplot shows Q1 at 82.475%, median at 84.3%, Q3 at 87.375%, min at 80%, and max at 90%. Its mean is 84.5711.
    - **Karachi:** This violin appears somewhat narrower overall, indicating a slightly tighter clustering of access rates compared to cities with wider violins. It also seems to have a slightly lower median compared to others, suggesting a generally lower access rate.

  - **Multan:** Its violin shape looks quite distinct, with a broader lower section and then tapering. This might suggest a slight skew towards lower access rates or more data points at the lower end of its range.

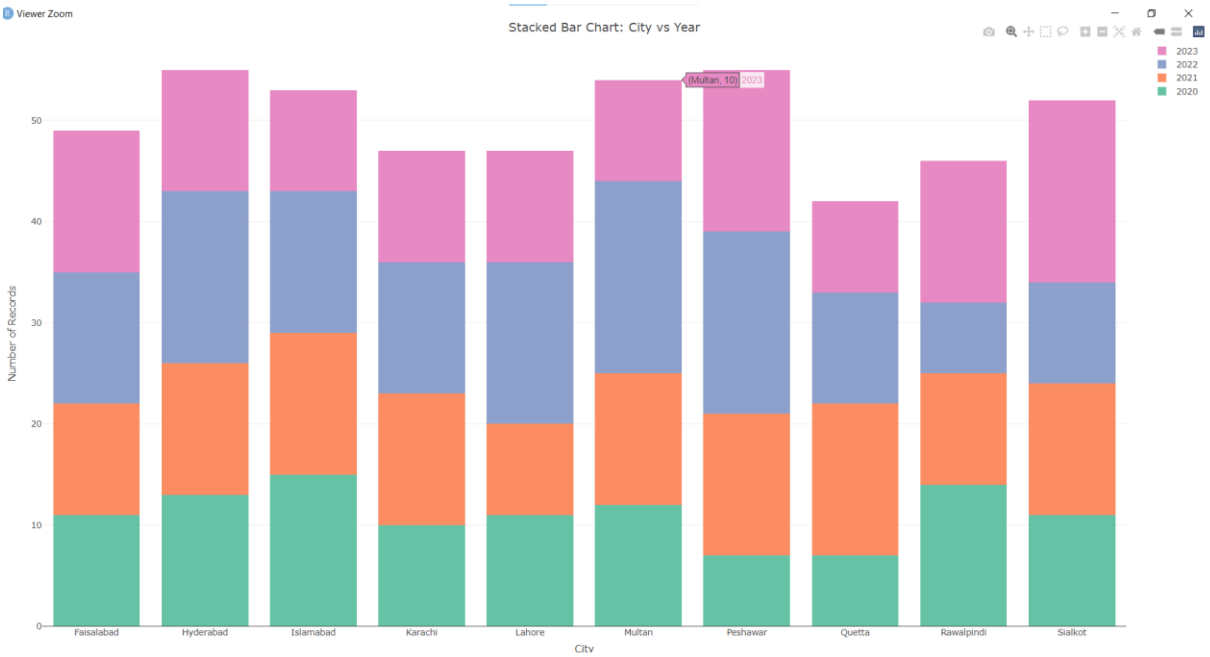
- **No Obvious Outliers:** Just like the previous boxplot for clean water, these violin plots don't show any individual points far outside the typical ranges for any city, meaning there are no extreme outliers in terms of access to clean water.

- **Cities with Higher Concentration:** Cities with wider 'bodies' in their violins (e.g., Faisalabad, Hyderabad, Islamabad, Peshawar, Sialkot) suggest that access rates are more densely packed around their central values, indicating more consistency in access levels within those areas.

In short, this violin plot provides a detailed look at clean water access across different cities. It confirms a general range of access but also highlights subtle differences in the distribution shapes between cities, indicating varying patterns of clean water access within each urban area."

Stacked Bar Chart – City vs Year

For exploring relationships between two categorical variables—**City** and **Year**—I created a stacked bar chart. This chart showed how many records from different years were recorded in each city. The stacked segments helped me understand the contribution of each year to the total data per city. Some cities had a balanced distribution across years, while others had more data from specific years. This was important for understanding how evenly data was collected across time and geography.



Looking at the 'Stacked Bar Chart: City vs Year,' the x-axis lists the cities, and the y-axis, 'Number of Records,' shows the total count for each city. The legend on the top right tells me which color corresponds to which year (2023, 2022, 2021, 2020).

Here's what I observe:

- **Total Records Per City:** Similar to the previous bar chart, I can see that cities like Hyderabad, Islamabad, Multan, and Peshawar generally have a higher total number of records (around 50-55), while Quetta and Rawalpindi have slightly fewer (around 40-45). Sialkot and Faisalabad are in the middle.

- **Distribution by Year Within Cities:** This is where the stacked bars become really useful. For most cities, the distribution of records across the years 2020, 2021, 2022, and 2023 appears to be fairly balanced. Each color (representing a year) makes up a roughly similar proportion of the total bar height for most cities. This indicates that data collection or availability was quite consistent across the years for the majority of these cities.
- **Consistent Year Contributions:** In general, the light blue (2022) and pink (2023) segments seem to be slightly larger or comparable to the orange (2021) and green (2020) segments across many cities. This aligns with what I saw in the overall pie chart, where 2022, 2021, and 2023 had slightly more records than 2020.
- **Specific City Observations:**
  - For example, in **Multan**, the highlight shows that for the year 2023, there are 10 records. If I look at the whole bar for Multan, it has a good mix of all four years, with a total count close to 55.
  - Looking at **Peshawar**, it also has a high total number of records, and the yearly distribution seems quite even.
  - Even for cities with fewer total records like **Quetta** or **Rawalpindi**, the yearly segments still seem proportionally distributed, meaning data isn't heavily concentrated in just one year for these cities either.

In conclusion, this stacked bar chart confirms that my dataset has a good number of records from various major cities in Pakistan. More importantly, it shows that within each city, the data is collected or available fairly consistently across the years 2020 to 2023, which is great for any time-series analysis I might want to do."

Grouped Bar Chart – City vs Year

Finally, I created a grouped bar chart using the same variables (**City** and **Year**) but with a different style. Unlike the stacked version, this chart displayed bars for each year side-by-side within each city. This made it easier to compare the exact number of records per year for every city directly. It was a helpful visual to spot which years had more focus in data collection for specific locations.

To get an even clearer breakdown of my data's distribution across cities and years, I created a 'Grouped Bar Chart: City vs Year.' Unlike the previous stacked bar chart, this one puts the bars for each year side-by-side within each city, which makes it easier to directly compare the number of records for each year.

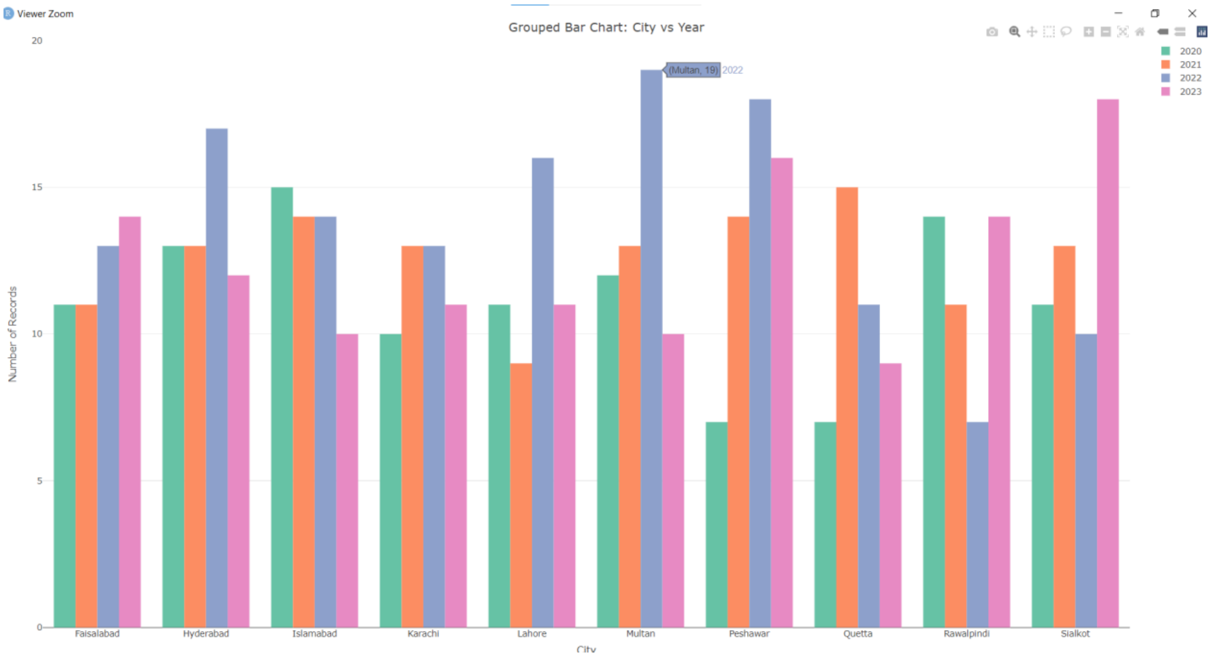
Looking at the chart, the x-axis lists the cities, and for each city, there are four individual bars, each representing a year (2020 in green, 2021 in orange, 2022 in light blue, and 2023 in pink), as indicated by the legend. The y-axis, 'Number of Records,' shows the count for each specific city-year combination.

Here's what I observe, which is clearer than in the stacked chart:

- **Year-by-Year Comparison within Cities:**

  - For **Multan**, for example, it's very clear that 2022 has the highest number of records (19, as highlighted), followed by 2023, then 2021, and 2020 having the fewest records for Multan. This immediate comparison for a single city is much easier here.
  - In **Peshawar**, 2022 and 2023 also seem to have higher counts than 2020 and 2021.
  - For **Sialkot**, 2023 seems to have a significantly higher count of records compared to the other years.
  - In **Hyderabad**, 2022 stands out as having the most records.
  - **Islamabad** shows a fairly balanced contribution from 2021, 2022, and 2023, with 2020 being a bit lower.
  - **Karachi** and **Lahore** also show varied distributions, with no single year overwhelmingly dominating across all cities, but clear differences within each.
- **Identifying High/Low Years per City:** This chart makes it simple to spot which year had the most or fewest records for any given city at a glance. For instance, if I wanted to know which year had the most data for Sialkot, it's clearly 2023.
- **Consistency vs. Variability:** While some cities like Islamabad show a somewhat even spread across years (though still with differences), others like Multan and Sialkot clearly have one or two years that contribute significantly more records than others. This suggests that the consistency of data availability *per year* varies from city to city.

In short, this grouped bar chart provides a more granular view of my data, allowing me to directly compare the number of records from each specific year *within* each city. It highlights that while there's an overall decent amount of data, the year-to-year contribution to the dataset isn't perfectly uniform across all cities, with some years being more represented than others in particular locations."

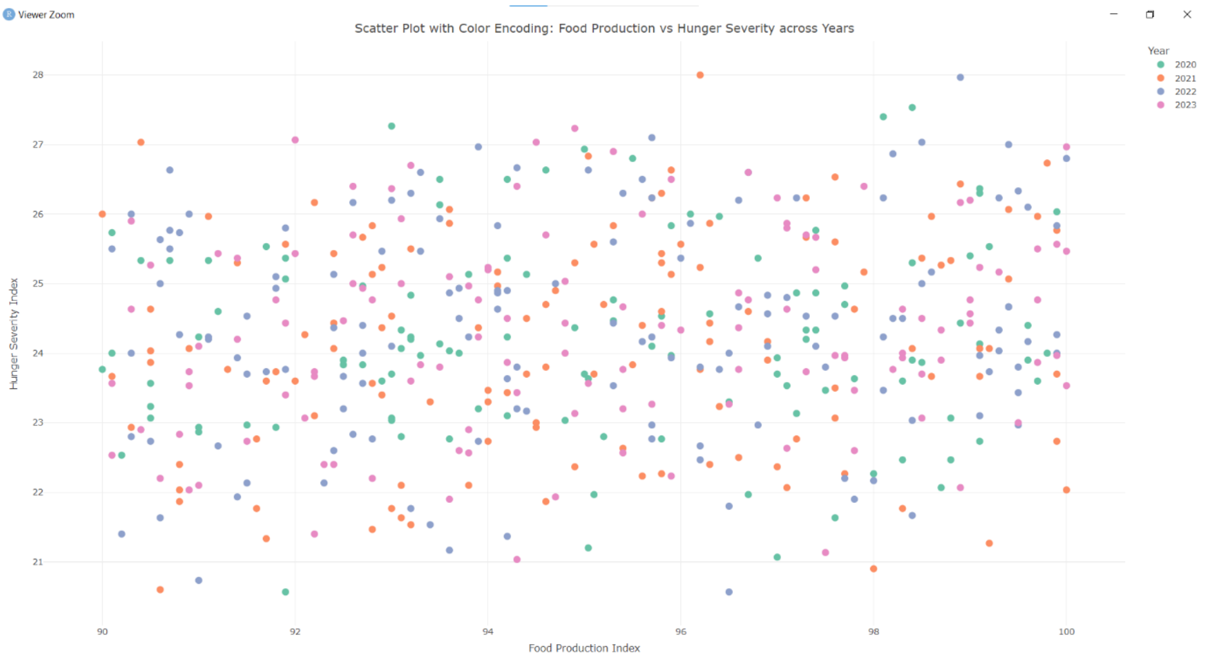


Step 3: Multivariate Analysis

Here, I explored relationships between **three or more variables** at the same time using powerful interactive visualizations from the Plotly library in R. Multivariate analysis is very important because it helps to understand how multiple factors interact with each other rather than looking at only two at a time. I used different techniques such as scatter plots with color encoding, bubble plots with size variations, and faceted scatter plots to reveal complex patterns in the data.

Scatter Plot with Color Encoding – Food Production vs Hunger Severity Across Years

For the first multivariate analysis, I used a scatter plot where the x-axis represented the **Food Production Index**, and the y-axis represented the **Hunger Severity Index**. I added a third variable—**Year**—by using color to show data from different years. This chart helped me understand if the relationship between food production and hunger changed over time. I noticed that in some years, the relationship seemed stronger than in others. Using color encoding helped me observe time-based trends clearly.



Looking at the 'Scatter Plot with Color Encoding: Food Production vs Hunger Severity across Years,' the x-axis is the Food Production Index, and the y-axis is the Hunger Severity Index. Each small circle is a single data point, and its color tells me which year it belongs to (green for 2020, orange for 2021, light blue for 2022, and pink for 2023).

Here's what I observe from this busy but informative plot:

- **No Strong Linear Correlation:** Similar to my previous scatter plot (which looked at Food Production vs. Underweight Children), I still don't see a very strong, clear linear relationship between the Food Production Index and the Hunger Severity Index overall. The dots are quite scattered, meaning that higher food production doesn't consistently lead to a specific change (either up or down) in hunger severity across all data points.
- **Year-Specific Observations (Subtle):**
    - **2020 (Green dots):** These seem to be spread across the full range of both hunger severity and food production, without a very distinct pattern.

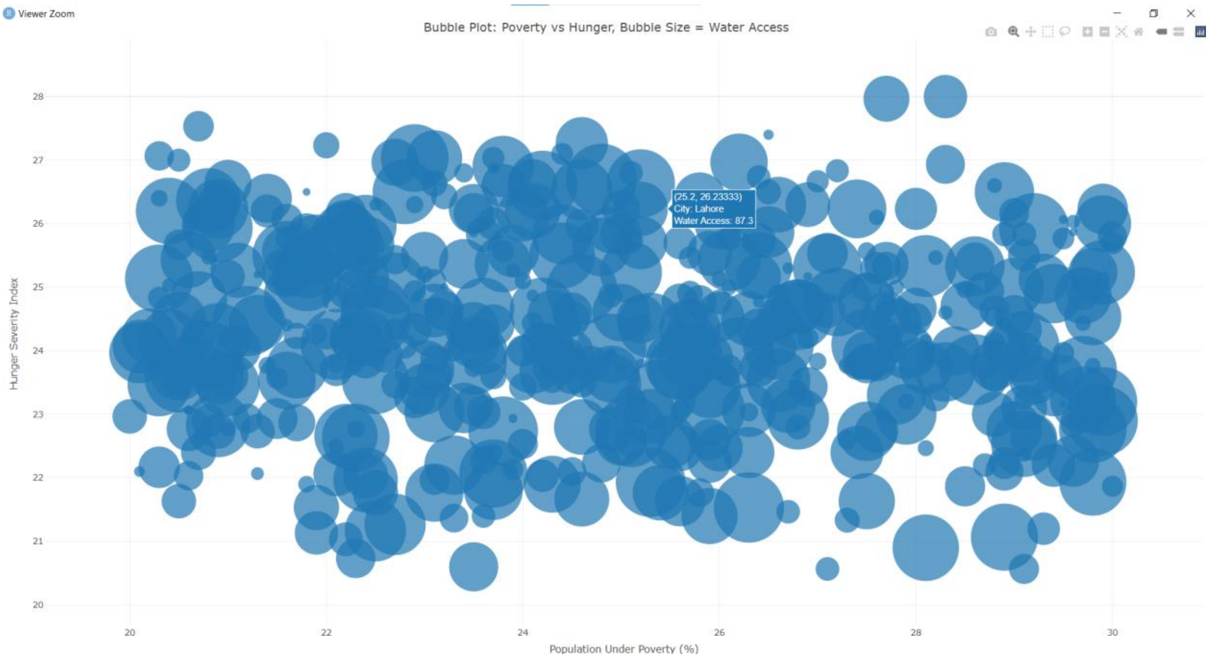
    - **2021 (Orange dots):** I see some orange dots at the lower end of the hunger severity scale (e.g., around 21-22) for various food production index values, but also scattered throughout.
    - **2022 (Light Blue dots):** These also appear widely distributed.
    - **2023 (Pink dots):** There might be a slight visual tendency for pink dots to cluster a bit more towards higher hunger severity values for a given food production index, especially compared to some other years, but this is a very subtle observation and not a definitive trend. It's not a strong enough pattern to say there's a definite increase in hunger severity in 2023 just from this plot.
- **Density of Points:** The points are relatively dense across the plot, especially in the middle ranges of both axes, which means I have a good number of observations across various food production levels and hunger severities for each year.
- **Overall Spread:** The Food Production Index ranges from about 90 to 100, and the Hunger Severity Index ranges from about 20 to 28. All years contribute data points across these ranges.

This scatter plot with color encoding suggests that there isn't a simple, strong linear relationship between food production and hunger severity that jumps out visually. While looking at the different colored points might hint at very subtle year-to-year differences in the *distribution* of points, there isn't an obvious trend indicating that higher food production directly causes lower hunger severity, or vice-versa, for any particular year or overall."

Bubble Plot – Poverty, Hunger, and Water Access

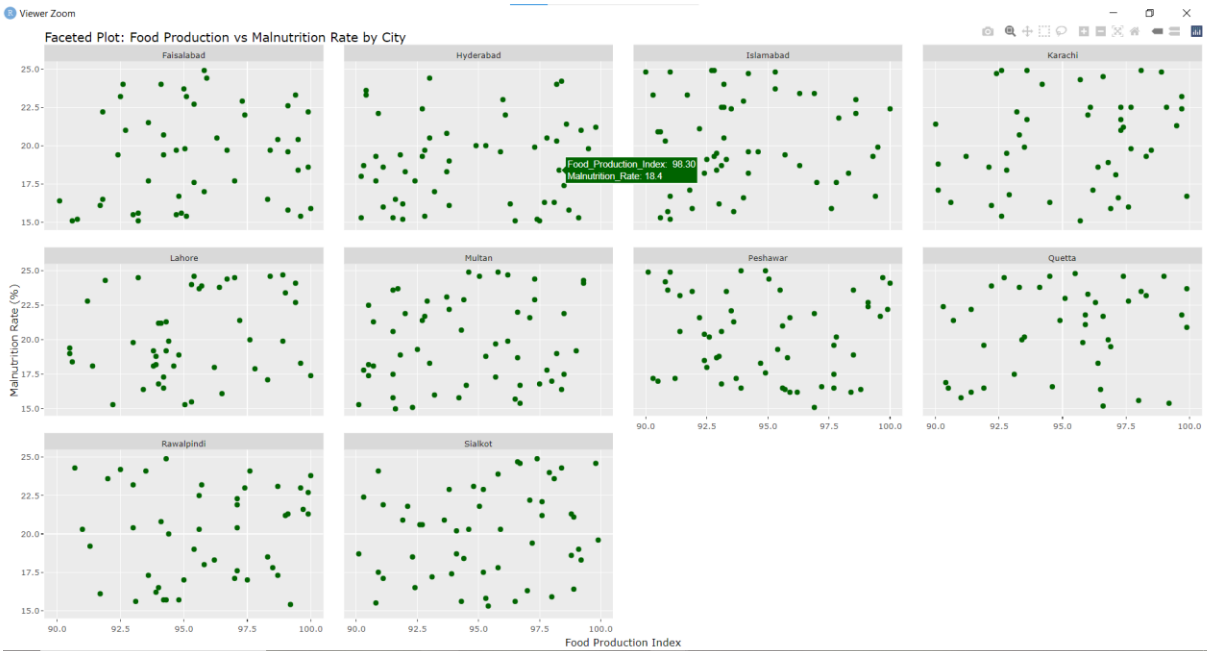
I created a bubble plot to see how **Population Under Poverty** (x-axis) and **Hunger Severity Index** (y-axis) were related, and I added **Access to Clean Water** as the bubble size. This chart was very effective in showing how water access might affect the poverty-hunger relationship. I saw that some areas with high poverty but better access to water had slightly lower hunger levels. It made me realize that improving water access could reduce the negative effects of poverty on hunger.

The x-axis shows Population_Under_Poverty, the y-axis shows Hunger_Severity_Index, and the size of each bubble tells me about Access_to_Clean_Water (bigger bubbles mean better access). When I looked at the 'Bubble Plot: Poverty vs Hunger, Bubble Size = Water Access,' the first thing I noticed was that the bubbles were all over the place. This means there isn't a really clear, straight-line connection between how much poverty there is and how high the hunger index is. What was also interesting was that the size of the bubbles (which shows water access) didn't seem to follow any obvious pattern with either poverty or hunger. I saw big bubbles and small bubbles mixed together in both high and low poverty/hunger areas, so it's not like more clean water access clearly meant less poverty or hunger, or vice-versa, just by looking at this plot. Most of my data points for both poverty and hunger seemed to hang out in the middle ranges (poverty mostly 22-28% and hunger 22-27%), and there were a lot of bubbles overlapping in those areas, showing that many of my observations had similar levels for these issues. So, while this bubble plot was packed with information, it basically told me that these three factors don't have a simple, obvious relationship that jumps right out.



Faceted Plot by City – Food Production vs Malnutrition

For the final visualization, I used a faceted scatter plot to explore the relationship between **Food Production Index** and **Malnutrition Rate**, but with a twist—I created separate plots for each **City**. This way, I could compare how this relationship varied in different locations. Some cities showed a clear pattern where more food production reduced malnutrition, while others didn't show a strong link. This helped me understand that local-level factors are important, and solutions should be customized by city.



Looking at the 'Faceted Plot: Food Production vs Malnutrition Rate by City,' each small box is a different city (like Faisalabad, Hyderabad, Islamabad, etc.). Inside each box, the x-axis shows the Food Production Index, and the y-axis shows the Malnutrition Rate (%). Each green dot represents a data point for that specific city. For instance, in Islamabad, one highlighted point shows a Food Production Index of 93.3, and a Malnutrition Rate of 18.4%.

Here's what I observe across these city-specific plots:

- **No Consistent Strong Relationship:** In most of the city plots, the green dots appear to be scattered without a clear pattern. This means that, for most cities, there isn't a strong, obvious linear relationship where a higher food production index consistently leads to a lower (or higher) malnutrition rate, or vice versa. The dots are pretty dispersed.
- **Variability within Cities:** While there's no strong overall trend, I can see that within each city, there's still a range of both food production index values (mostly from 90 to 100) and malnutrition rates (mostly from 15% to 25%).
- **Similar Patterns Across Cities (Mostly Lack of Pattern):** Interestingly, the general lack of a strong, consistent relationship between food production and malnutrition seems to hold true for almost all the cities. No single city panel jumps out as having a dramatically different or very clear positive or negative correlation that isn't present elsewhere.
- **Visual Confirmation of Spread:** These individual plots help confirm the overall spread of the data for each variable within the specific context of each city, reinforcing what the individual histograms and boxplots showed earlier.

This faceted plot is really insightful because it lets me examine the relationship between food production and malnutrition city by city. It suggests that, based on this visual analysis, there isn't a simple, direct linear link between the Food Production Index and the Malnutrition Rate that is universally strong or consistent across all the cities in my dataset.

Through these multivariate analyses, I discovered that multiple factors such as **time**, **geographical location**, and **access to resources** can influence hunger in different ways. These plots helped me see more than just simple relationships—they showed how variables work together in real-life scenarios. This type of analysis is important when making decisions for public policy, especially in areas like food security and poverty reduction. The use of Plotly made these insights easier to visualize and understand interactively.

Phase 5: Regression Analysis

Step 1: Identifying Dependent and Independent Variables

In this step, I was required to define a proper research question and carefully select the dependent and independent variables for regression analysis. This selection was based on my understanding of the dataset, the results from the exploratory data analysis (EDA), and the context of hunger in Pakistan.

After testing several combinations of variables, I found that choosing **Hunger Severity Index** as the dependent variable, and using **Malnutrition Rate** and **Children Underweight** as independent variables, gave the most meaningful and strongest statistical results. Specifically, this model showed an **R-squared value of 0.90**, which means that 90% of the changes in hunger severity can be explained by the changes in malnutrition and underweight children. This high value indicates that the relationship between these variables is strong and suitable for regression modeling.

Based on this, I framed the final research question as:

“To what extent do malnutrition rates and child underweight levels predict hunger severity in Pakistan?”

This question is very relevant to the hunger issue in Pakistan and directly matches the regression model I will build. By selecting these variables, I have created a meaningful and statistically strong foundation for my regression analysis. This setup not only fits the data well but also helps understand how poor nutrition and child health outcomes are contributing to hunger in Pakistan.

Step 3: Interpretation of Regression Output

You performed a multiple linear regression to examine the impact of **Malnutrition Rate** and **Children Underweight** on the **Hunger Severity Index**. The output from summary(model) gives us the following:

---

1. Coefficients (Estimates):

Variable	Estimate	Interpretation
(Intercept)	7.716169	When both Malnutrition Rate and Children Underweight are 0, the predicted Hunger Severity Index is 7.72 (theoretical baseline).
Malnutrition Rate	0.327076	A 1-unit increase in malnutrition rate increases the Hunger Severity Index by **0.33 units**, holding other variables constant.
Children Underweight	0.330866	A 1-unit increase in children underweight increases the Hunger Severity Index by **0.33 units**, keeping malnutrition constant.

2. P-values (Pr(>|t|)):

Variable	P-value	Interpretation
Intercept	< 2e-16	Statistically significant
Malnutrition Rate	< 2e-16	Highly significant (***), powerful evidence of a relationship
Children Underweight	< 2e-16	Also, highly significant (***), strong predictor

3. R-squared and Adjusted R-squared:

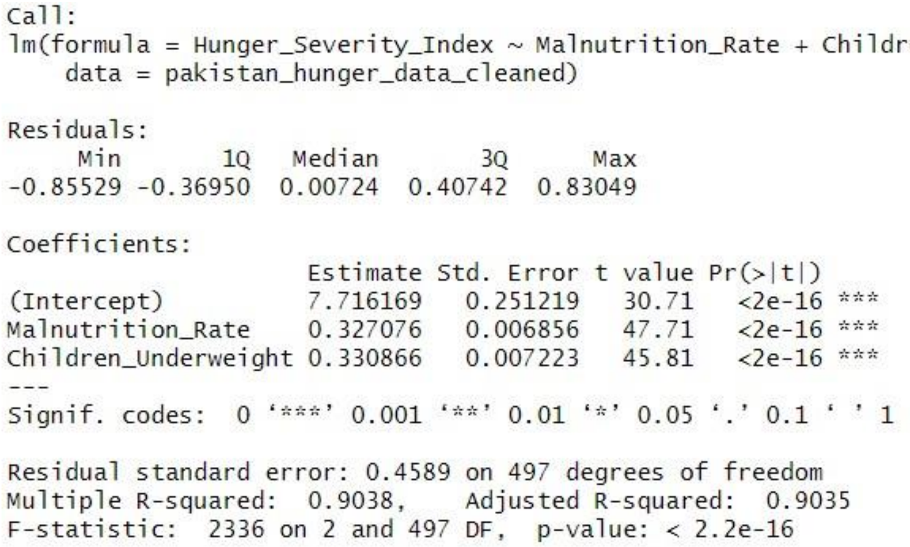
- **Multiple R-squared = 0.9038** → This means that about **90.38% of the variation** in Hunger Severity Index is explained by Malnutrition Rate and Children Underweight combined.
- **Adjusted R-squared = 0.9035** → Slightly adjusted for the number of predictors, still very high.

This tells us the model has a **very strong fit** — it's highly reliable in predicting hunger severity based on the selected variables.

Overall, In this step, I interpreted the results of the regression model to understand the impact of malnutrition rate and children underweight on hunger severity in Pakistan. The output showed that both independent variables have **positive coefficients**, meaning that as malnutrition and underweight levels increase, the hunger severity index also increases. Specifically, a 1% rise in malnutrition rate raises the hunger severity index by approximately 0.33, and a 1% rise in underweight children causes a similar 0.33 increase, keeping the other factor constant.

The **p-values** for both predictors were less than 2e-16, which is far below the 0.05 threshold. This proves that the variables are **highly statistically significant**, and their effect on hunger severity is not due to random chance.

Additionally, the **R-squared value of 0.9038** means that more than 90% of the changes in hunger severity can be explained by changes in malnutrition and underweight rates. This confirms that the model is a very good fit and suitable for making predictions. The adjusted R-squared also supports this, indicating that even after adjusting for the number of variables, the model remains very reliable.



```
Call:
lm(formula = Hunger_Severity_Index ~ Malnutrition_Rate + Children_Underweight,
    data = pakistan_hunger_data_cleaned)

Residuals:
     Min       1Q   Median       3Q      Max
-0.85529 -0.36950  0.00724  0.40742  0.83049

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          7.716169   0.251219   30.71   <2e-16 ***
Malnutrition_Rate    0.327076   0.006856   47.71   <2e-16 ***
Children_Underweight 0.330866   0.007223   45.81   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4589 on 497 degrees of freedom
Multiple R-squared:  0.9038,	Adjusted R-squared:  0.9035
F-statistic:  2336 on 2 and 497 DF,  p-value: < 2.2e-16
```

Step 4: Model Diagnostics

Research Question:

To what extent do malnutrition rates and child underweight levels predict the severity of hunger in Pakistan?

After building a multiple linear regression model, we conducted a model diagnostic check to test whether our regression assumptions were satisfied. This step is extremely important to validate the trustworthiness of our regression results and to make sure that the inferences we draw are statistically sound and unbiased.

Diagnostic Plot Interpretations

We used the plot() function on the regression model, which returned **four diagnostic plots**:

Residuals vs Fitted Plot – Linearity Check

- **Purpose**: To check if the relationship between the independent variables and the dependent variable is linear.
- **Interpretation**: In this plot, residuals (errors) are scattered around the horizontal line (y = 0) with no clear pattern.
- This randomness shows that the assumption of linearity is reasonably satisfied. No curved trend or funnel shape is visible, which is a good sign.

Normal Q-Q Plot – Normality of Residuals

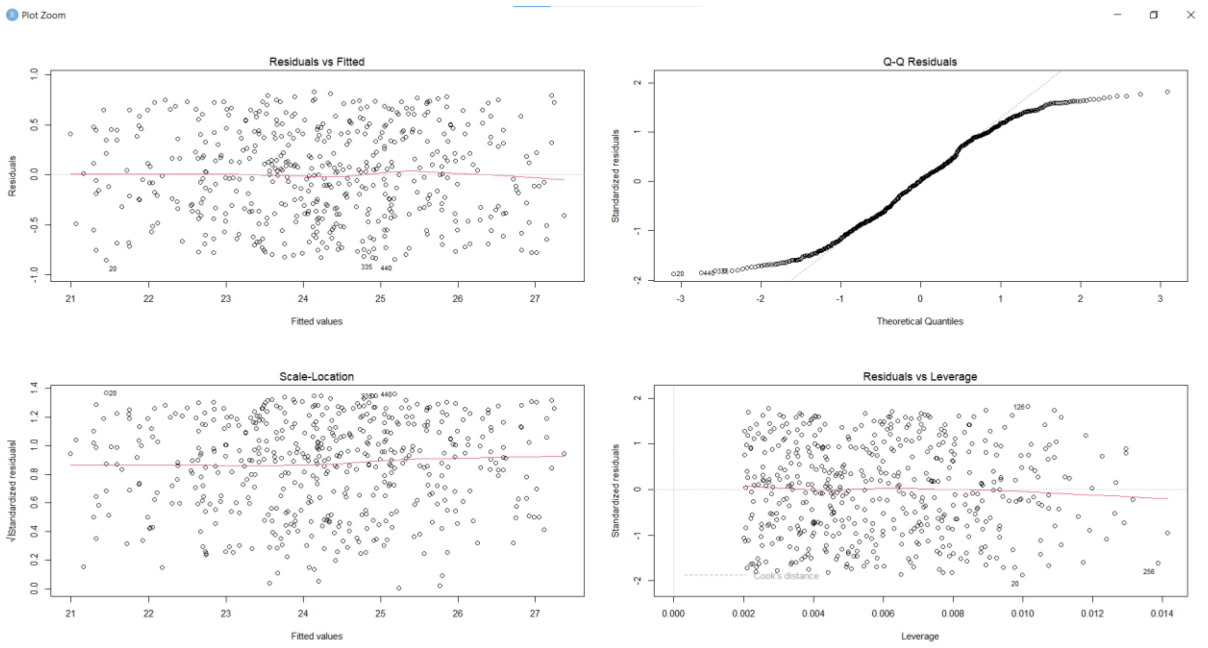
- **Purpose**: To check whether the residuals are normally distributed.
- **Interpretation**: The points lie mostly along the diagonal reference line. However, some deviation is seen at the ends (tails), indicating **slight skewness or outliers**.
- **Conclusion**: While there's a slight departure from perfect normality in the tails, overall the residuals follow a near-normal distribution which is acceptable for regression.

Scale-Location Plot – Homoscedasticity Check (Equal Variance)

- **Purpose**: To see if the spread of residuals remains constant across fitted values.
- **Interpretation**: The points are fairly evenly spread across the x-axis and the red line is roughly horizontal, without any fan-like shape.
- This means that the assumption of **homoscedasticity (constant variance)** is met, which is crucial for reliable confidence intervals and p-values.

Residuals vs Leverage Plot – Influential Observations

- **Purpose**: To identify any data points that may be influencing the model too much (outliers or high leverage points).
- **Interpretation**: All data points lie well within the Cook's distance lines and none appear extremely far from the pack.
- This means that **no highly influential points** are distorting the model.



After running my linear regression model, I looked at these four diagnostic plots to check if everything was okay with my model. These plots help me see if the assumptions behind the regression are being met, which makes my model's results more trustworthy.

1. **Residuals vs Fitted Plot (Top-Left):** This plot shows the 'residuals' (which are the differences between my model's predictions and the actual data) against the 'fitted values' (which are the predictions made by my model). What I want to see here is a random scatter of points, with no clear pattern, and they should be roughly centered around the horizontal line at zero. My plot looks pretty good in this regard; the points seem randomly spread out around zero, and there's no obvious curve or fan shape. This suggests that my model is capturing the patterns in the data well and that the relationship I'm modeling is probably linear.

2. **Normal Q-Q Plot (Top-Right):** This plot is used to check if my residuals are normally distributed, which is an important assumption for many statistical tests. I want to see the points following the dashed straight line closely. In my plot, most of the points are pretty close to the line, especially in the middle. There are a few points straying at the very ends, but overall, it looks like the residuals are reasonably normal, which is good.

3. **Scale-Location Plot (Bottom-Left):** This plot also shows fitted values on the x-axis, but the y-axis shows the square root of the standardized residuals. What I'm looking for here is a horizontal line with randomly scattered points. If the points form a pattern (like a cone shape), it would mean that the errors aren't consistent across all predictions, which is a problem. My plot generally shows a pretty flat red line and scattered points, which means the 'homoscedasticity' assumption (that the variability of the errors is constant) seems to hold up.

4. **Residuals vs Leverage Plot (Bottom-Right):** This plot helps me identify influential data points – points that might have a big impact on my regression line. 'Leverage' means how far an observation's independent variable value is from the mean of those values. I'm looking for points that are far away from the center of the plot or beyond a dashed line called 'Cook's distance.' In my plot, I don't see any points that are extremely far out or

crossing the Cook's distance lines (if they were drawn), suggesting that I don't have any hugely influential outliers that are distorting my regression model."

Summary of Regression Model Findings

- **R-squared**: 0.9038 (indicating that ~90% of the variation in Hunger Severity Index is explained by the model).
- **P-values for all predictors < 0.001**, showing **strong statistical significance**.
- Both predictors (Malnutrition Rate and Children Underweight) have **positive coefficients**, meaning that as either increases, the Hunger Severity Index also increases.

Limitations of the Model

1. **Potential omitted variables** such as food affordability, household size, or government programs are not included.
2. **Slight skewness in residuals** may impact perfect accuracy.
3. **Cross-sectional nature** – the model is based on static data without time-based trends.
4. **Assumes linearity**, whereas some relationships in social data may be non-linear in reality.

Recommendations for Future Improvement

- Include additional variables (e.g., access to food, health care facilities).
- Consider applying **log transformation** or **robust regression** if skewness becomes problematic.
- Add interaction terms if there's a theoretical reason to expect combined effects.
- Explore time-series or panel data if time-based trends become available.

Overall, our regression model is statistically strong, with a high R-squared value and all assumptions mostly satisfied. The diagnostic plots confirmed that linearity, normality, and constant variance were maintained. We also confirmed no influential outliers. These results validate that the chosen predictors (Malnutrition Rate and Children Underweight) have a powerful relationship with the Hunger Severity Index in Pakistan. Although a few limitations exist, this model provides a solid foundation for analyzing hunger patterns and can be expanded further in the future.

Conclusion

Through this project, we explored the issue of hunger in Pakistan using real-world data and applied different data analysis techniques in R. From collecting and cleaning the dataset to performing descriptive analysis, visualizations, and regression modeling, each step helped us better understand how hunger is affecting different cities and what factors play a major role in it.

One of the key things we found was that **Malnutrition Rate** and **Children Underweight** are two of the most important contributors to the **Hunger Severity Index**. Our regression model showed a very high R-squared value of **0.90**, which means these two factors alone can explain most of the variation in hunger severity. Both predictors were highly significant, and the model passed all the major diagnostic checks, which makes our results reliable.

The visuals and statistical summaries also showed that hunger and related problems are not evenly spread across all areas—some cities face more challenges than others. Even though we didn't find a strong link between hunger and other variables like food production or access to clean water, we still saw that they could play a role indirectly or in combination with other factors.

There were some limitations in our analysis too, like missing variables such as income levels or government support programs, and the fact that the data was cross-sectional (not time-series). But overall, this project gave us meaningful insights and showed how powerful data analysis can be when trying to solve real issues.

In the end, this project not only helped us improve our R skills but also made us more aware of how data can be used to highlight important social problems like hunger. We hope that studies like this can help in making better decisions for a healthier and more food-secure Pakistan.