National University of Sciences and Technology (NUST)
School of Electrical Engineering and Computer Science

Department of Computing

CS 366: Data Visualisation

BSCS: 13A

Project Documentation

Due Date: 12th Dec 2025

S.#	Name	Qalam ID
1	Muhammad Hammad	467513
2	Maier Ali	481889

https://github.com/thedevhammad/Data-Visualization-Project.git

Table of Contents

Dataset Source and Description:

All the datasets were collected from the OpenData.pk site. This is a platform offering publicly accessible datasets from different sectors in Pakistan, including education, health, economics, and others. The platform was a part of the government's initiative to promote data transparency and enable citizens, researchers, and organisations to access and use public data for policy development, research and decision making. We gathered 4 different datasets from this platform with one being the major source and the rest of them being supporting datasets.

I) Punjab Annual School Census Report (2017-18)

Source: https://opendata.com.pk/dataset/punjab-annual-school-census-report-2017-18/resource/7baf2669-333c-429a-ba2a-d000ded4ae8f?inner_span=True

Description:

This dataset from the Punjab Annual School Census (2017-18**)** contains 52,471 rows and 108 columns. It provides detailed information on schools (ID, name), district (name, tehsil, markaz, address, etc…), political party information for that school, information about the school (level, head name, status, language medium, reason for non-functionality, gender, location, shift, established year), the school infrastructure information, area, classes, teacher availability**,** school facilities, sports facilities, etc across Punjab. It serves as a foundational resource for understanding the state of education infrastructure in one of Pakistan's largest provinces.

Each row represents a school's above-mentioned data.

It helps to identify regions with resource gaps and provides insights into schooling patterns in Punjab, informing policymakers on areas that need improvement.

II) KPK Non-Functional Primary and Middle Schools:

Sources:

Primary: https://opendata.com.pk/dataset/number-of-govt-primary-schools-non-functional-2021-numbers-kpk-pakistan

Middle: https://opendata.com.pk/dataset/number-of-govt-middle-schools-non-functional-2021-numbers-kpk-pakistan/resource/3c55fceb-3b24-4b6a-8629-eddcba17c6cb

Description:

These datasets provide data on the non-functional primary and middle schools in Khyber Pakhtunkhwa (KPK) for the year 2021, broken down by gender (male and female). It highlights where middle school education is interrupted due to resource deficits or infrastructure failures.

The primary school's dataset contains 24 rows and 4 columns. The middle school contains 10 rows and 4 columns. Both the datasets contain the names of the districts, the number of male non-functional schools, number of female non-functional schools and the total number of male and female non-functional schools for each district.

III) **Key Indicators of Education in Pakistan:**

Source: https://opendata.com.pk/dataset/key-indicators-of-education-in-pakistan/resource/65b65671-55c9-4cb5-8486-420ed451a136

Description:

This dataset provides provincial and national-level educational statistics for Pakistan, including key indicators like literacy rates**,** school enrollment rates, and out-of-school children for the years 2013-14 and 2018-19. It serves as a supporting source for analyzing gender disparities and regional inequalities in education across the country.

The dataset contains 77 rows and 8 columns. These include the name of the province, educational indicator, literacy or enrollment rate for males and females for each province, out of school children in each province, etc.

This dataset compares gender-based education outcomes and regional educational performance across different provinces. It helps identify trends in literacy**,** school enrollment, and out-of-school children, which can be used to evaluate the effectiveness of educational policies and track changes over time.

Dataset Cleaning:

I) Non-Functional Primary schools data:

1. No missing values:

Code snippet:

```
# ✅ 1. DROP completely empty rows (to avoid false missing/inconsistency issues)
df = df.dropna(how='all')
print("Shape after removing empty rows:", df.shape)

# ----------------------------------------------------------------
# ✅ Missing Values Check
# ----------------------------------------------------------------
missing = df.isnull().sum()

print("\n🔳 Missing Values per Column:")
print(missing)
```

```
Initial shape: (24, 4)
Shape after removing empty rows: (23, 4)

🟦 Missing Values per Column:
District    0
Male        0
Female      0
Total       0
dtype: int64
```

Note: Our data contained the last entire empty row and pandas treated them as empty and counted them at first but that is a complete empty row with no entry that is why I removed it.

2. There are no redundant/ derivable attributes in our data:

Code snippet:

```
# Function to check redundant/derivable columns
def check_redundant_values(df, name):
    print(f"\n=== Redundant / Derivable Check for {name} ===")
    redundant_cols = []

    # Check if Total = Male + Female
    if all(col in df.columns for col in ["Male", "Female", "Total"]):
        calc_total = df["Male"] + df["Female"]
        if (calc_total == df["Total"]).all():
            print("✓ 'Total' is derivable as Male + Female.")
            redundant_cols.append("Total")
        else:
            print("⚠ Total does NOT match Male + Female for some rows.")

    # Check for duplicate districts
    if "District" in df.columns:
        dup = df["District"].duplicated().sum()
        print(f"Duplicate District Names: {dup}")

    return redundant_cols
```

```
Duplicate District Names: 0
Redundant / derivable columns: []
```

3. Using the formula for quartiles we detected outliers:

lower = Q1 - 1.5 * IQR

upper = Q3 + 1.5 * IQR

Code snippet:

```
# ----------------------------------------------------------------------
# ✅ Outlier Detection
# Using IQR method on numerical columns
# ----------------------------------------------------------------------
numeric_cols = ["Male", "Female", "Total"]
outliers = pd.DataFrame()

for col in numeric_cols:
    Q1 = df[col].quantile(0.25)
    Q3 = df[col].quantile(0.75)
    IQR = Q3 - Q1

    lower = Q1 - 1.5 * IQR
    upper = Q3 + 1.5 * IQR

    col_outliers = df[(df[col] < lower) | (df[col] > upper)]
    outliers = pd.concat([outliers, col_outliers])

print("\n🔍 Outliers Detected:")
print("Count:", outliers.drop_duplicates().shape[0])
print(outliers.drop_duplicates()[["District", "Male", "Female", "Total"]])
```

```
🔍 Outliers Detected:
Count: 4
            District  Male  Female  Total
12            Kurram  17.0     7.0   24.0
17           Orakzai  15.0     5.0   20.0
19  South Waziristan  38.0    38.0   76.0
11          Kohistan   7.0    53.0   60.0
```

4. There were no inconsistent value since the column name was male and female and it was actually about how many makes and female were in a district:

```
# Example DataFrame
data = {
    "Gender": ["M", "Male", "m", "F", "Female", "f", "MALE", "FEMALE"]
}
df = pd.DataFrame(data)

# Mapping variants to standard values
gender_mapping = {
    "Male": ["m", "male", "M", "MALE"],
    "Female": ["f", "female", "F", "FEMALE"]
}

# Create reverse mapping: variant (lowercase) -> standard
reverse_map = {v.lower(): k for k, variants in gender_mapping.items() for v in
   variants}

# Standardize the column
df["Gender"] = df["Gender"].astype(str).str.strip().str.lower().map(reverse_map)
```

```
print("✅ Standardized Gender column:")
print(df)
```

```
- Type: Total mismatch
  Count: 0
```

II) Non-functional middle schools:

1. No missing values:

```
# ----------------------------------------------------------------------
    # ✅ Missing Values Check
    # ----------------------------------------------------------------------
    missing = df.isnull().sum()
    print("\n🔍 Missing Values per Column:")
    print(missing)
```

```
=============== Processing: Secondary Schools Data ===============

Initial shape: (9, 4)
Shape after removing empty rows: (9, 4)

🟦 Missing Values per Column:
District    0
Male        0
Female      0
Total       0
dtype: int64
```

2. No derivable attribute, based on the formula here the total is derivable from male and female columns by simply adding them:

```
  # Derivable attribute: Total = Male + Female (if columns exist)
    # ----------------------------------------------------------------------
    if all(col in df.columns for col in ["Male", "Female", "Total"]):
        df["Sum"] = df["Male"] + df["Female"]
        inconsistent = df[df["Total"] != df["Sum"]]
        print("\n🔍 Derivable Attribute Issues (Total != Male + Female):")
        print("Count:", inconsistent.shape[0])
        if inconsistent.shape[0] > 0:
            print(inconsistent[["District", "Male", "Female", "Total", "Sum"]])
```

```
🟥 Derivable Attribute Issues (Total != Male + Female):
Count: 0
```

No redundant row:

```
# ✅ Redundant / Derivable Columns Check
    # ----------------------------------------------------------------------
    def check_redundant_values(df, name):
        print(f"\n=== Redundant / Derivable Check for {name} ===")
        redundant_cols = []
        if all(col in df.columns for col in ["Male", "Female", "Total"]):
            calc_total = df["Male"] + df["Female"]
            if (calc_total == df["Total"]).all():
                print("✓ 'Total' is derivable as Male + Female.")
                redundant_cols.append("Total")
            else:
                print("⚠ Total does NOT match Male + Female for some rows.")
        if "District" in df.columns:
            dup = df["District"].duplicated().sum()
```

```
            print(f"Duplicate District Names: {dup}")
        return redundant_cols

    redundant_cols = check_redundant_values(df, name)
    print(f"Redundant / derivable columns: {redundant_cols}")
```

```
Duplicate District Names: 0
Redundant / derivable columns: ['Total']
```

3. There were the outliers based on quartile ranges:

lower = Q1 - 1.5 * IQR

upper = Q3 + 1.5 * IQR

```
    # ✅ Outlier Detection
    # Using IQR method on numeric columns
    # ----------------------------------------------------------------
    numeric_cols = [col for col in ["Male", "Female", "Total"] if col in
   df.columns]
    outliers = pd.DataFrame()
    for col in numeric_cols:
        Q1 = df[col].quantile(0.25)
        Q3 = df[col].quantile(0.75)
        IQR = Q3 - Q1
        lower = Q1 - 1.5 * IQR
        upper = Q3 + 1.5 * IQR
        col_outliers = df[(df[col] < lower) | (df[col] > upper)]
        outliers = pd.concat([outliers, col_outliers])
    print("\n Outliers Detected:")
    print("Count:", outliers.drop_duplicates().shape[0])
    if not outliers.empty:
        print(outliers.drop_duplicates()[["District"] + numeric_cols])
```

```
 Outliers Detected:
Count: 2
          District  Male  Female  Total
8  South Waziristan     7       5     12
6  North Waziristan     0       3      3
```

4. There were no inconsistent value since the column name was male and female and it was actually about how many makes and female were in a district:

```
# Example DataFrame
data = {
    "Gender": ["M", "Male", "m", "F", "Female", "f", "MALE", "FEMALE"]
}
df = pd.DataFrame(data)

# Mapping variants to standard values
gender_mapping = {
    "Male": ["m", "male", "M", "MALE"],
    "Female": ["f", "female", "F", "FEMALE"]
}

# Create reverse mapping: variant (lowercase) -> standard
```

```
reverse_map = {v.lower(): k for k, variants in gender_mapping.items() for v in
   variants}

# Standardize the column
df["Gender"] = df["Gender"].astype(str).str.strip().str.lower().map(reverse_map)

print("✅ Standardized Gender column:")
print(df)
```

```
- Type: Total mismatch
  Count: 0
```

The primary and the secondary KPK non-functional schools' datasets were then integrated together into a single file. The combined dataset now included a column named level, tat indicated whether the district was a primary or a middle school.

III) Punjab Public census 2018:

1. There were missing values:

```
# ----------------------------------------------------------------------
    # ✅ Missing Values Check
    # ----------------------------------------------------------------------
    missing = df.isnull().sum()
    print("\n🔍 Missing Values per Column:")
    print(missing)
```

```
🟦 Missing Values per Column:
school_id                4273
emiscode                    0
school_name                 0
district                    0
tehsil                      0
                          ...
ece_trained_teachers     4291
care_giver               4299
enrollment                 10
Teachers                 4302
NonTeachers             28906
Length: 108, dtype: int64
```

2. No derivable attribute and redundant values:

```
# ✅ Redundant / Derivable Columns Check
    # ----------------------------------------------------------------------
    def check_redundant_values(df, name):
        print(f"\n=== Redundant / Derivable Check for {name} ===")
        redundant_cols = []
        if all(col in df.columns for col in ["Male", "Female", "Total"]):
            calc_total = df["Male"] + df["Female"]
            if (calc_total == df["Total"]).all():
                print("✓ 'Total' is derivable as Male + Female.")
                redundant_cols.append("Total")
            else:
                print("⚠ Total does NOT match Male + Female for some rows.")
        if "District" in df.columns:
            dup = df["District"].duplicated().sum()
            print(f"Duplicate District Names: {dup}")
        return redundant_cols

    redundant_cols = check_redundant_values(df, name)
    print(f"Redundant / derivable columns: {redundant_cols}")
```

```
=== Redundant / Derivable Check for Public Census Oct 2018 ===
Redundant / derivable columns: []
```

3. No outliers detected:

```
    # ✅ Outlier Detection
    # Using IQR method on numeric columns
    # ----------------------------------------------------------------
    numeric_cols = [col for col in ["Male", "Female", "Total"] if col in
   df.columns]
    outliers = pd.DataFrame()
    for col in numeric_cols:
        Q1 = df[col].quantile(0.25)
        Q3 = df[col].quantile(0.75)
        IQR = Q3 - Q1
        lower = Q1 - 1.5 * IQR
        upper = Q3 + 1.5 * IQR
        col_outliers = df[(df[col] < lower) | (df[col] > upper)]
        outliers = pd.concat([outliers, col_outliers])
    print("\n🔍 Outliers Detected:")
    print("Count:", outliers.drop_duplicates().shape[0])
    if not outliers.empty:
        print(outliers.drop_duplicates()[["District"] + numeric_cols])
```

```
🔍 Outliers Detected:
Count: 0
```

4. There were no inconsistent values (row entries):

```
- Type: Total mismatch
  Count: 0
```

But there were inconsistent column names so I made them consistent, for e.g:

Original	Renamed	Change Explanation
uc_name	union_council_name	Expanded abbreviation `uc` to full name for clarity
pp_no	provincial_assembly_number	Expanded abbreviation `pp`
head_name	school_head_name	Added `school_` prefix for context
medium	language_medium	More descriptive, specifies "language" medium
non_func_reason	non_functional_reason	Expanded abbreviation `func`
est_year	school_established_year	More descriptive and consistent with other `_year` columns
bldg_status	building_status	Expanded abbreviation `bldg`
classes	total_classes	Added `total_` to clarify numeric quantity
sections	total_sections	Added `total_` for consistency
other	other_sports	Clarified what "other" refers to
csl_council_meetings	community_council_meetings	Expanded abbreviation `csl`
ece_trained_teachers	trained_preschool_teachers	Clearer wording
Teachers	teachers	Standardized capitalization
NonTeachers	non_teachers	Standardized capitalization

Data Redundancy:

I have removed 3 columns from public-census_oct_2018.csv since they were pointing to the same data. First, we had:

Before removing columns:	**After removing columns (keeping kanal from each):**
• total_area_kanal, total_area_marla	• total_area_kanal

• covered_area_canal, covered_area_marla • uncovered_area_kanal, uncovered_area_marla	• covered_area_kanal • uncovered_area_kanal

Columns like building status and place availability status were redundant since it only contains 1 and does not communicate any significant insight.

Data Reduction Techniques:

Numerosity Reduction:

Numerosity reduction was achieved by summarizing large volumes of school-level data into higher-level aggregates. Instead of analyzing 50,000+ individual schools directly, schools were aggregated at the district levels. For example, district wise average infrastructure, safety, facilities, and total performance scores. This approach significantly reduced the number of records while preserving the overall structure and trends in the data.

Dimensionality Reduction:

Feature Engineering via composite scores:

Multiple raw variables were combines into composite performance scores such as Infrastructure score, Safety score, Facilities score and Total Performance score. Each score compresses several related indicators (e.g., electricity, toilets, classrooms, boundary walls) into a single interpretable metric. This reduced dozens of columns into a small set of meaningful dimensions.

Principal Component Analysis (PCA):

PCA was applied to facility-related variables to further reduce dimensionality. Several correlated facility indicators were transformed into two principal components (PC1 and PC2). These components capture the majority of variance in the data and allow schools to be compared in a 2D space rather than across many separate variables.

Data Aggregation:

Temporal Aggregation:

School establishment and upgrade years were grouped into 5-year bins to reduce noise and highlight long-term trends.

Categorical Aggregation:

Gender-specific values were aggregated into averages or shares (e.g., female literacy share). Binary infrastructure indicators were aggregated into counts and percentages.

Spatial Aggregation:

School-level data was aggregated to district, province, or urban–rural levels for clearer comparison.

Exploratory Data Analysis:

Average Literacy Rate by Gender and Province

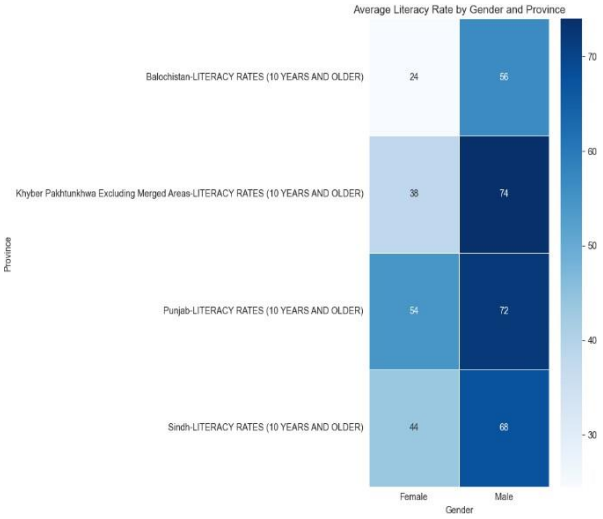


Figure 2: Average Literacy Rate by Gender & Province

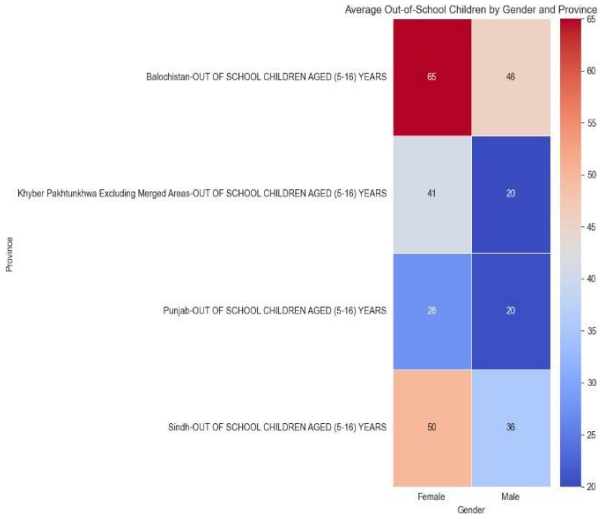


Figure 1: Average Out of School Children by Gender & Provice

Gender and year columns were extracted from the combined "Gender_Year" field, and the data was reshaped from wide to long format for easier analysis. The relevant columns were averaged by gender and province, and the data was pivoted to calculate average literacy rates and out-of-school children's rates.

The literacy heatmap clearly illustrates gender-based literacy disparities across the four provinces: Balochistan, Khyber Pakhtunkhwa, Punjab, and Sindh**.** Punjab shows the highest literacy rates for both genders, with a noticeable gap between female (54%) and male (72%) literacy rates. Balochistan has the lowest literacy rates, especially among females (24%), highlighting the challenges faced in this region regarding female education.

The out-of-school-children heatmap highlights the discrepancy in out-of-school children by gender and province. This supports the literacy rate heatmap. Balochistan has the highest number of out-of-school females (65), while males have a relatively lower rate (46). In contrast, Punjab shows a much lower number of out-of-school children, with only 28 males and 20 females out of school.

This visualization provides an immediate overview and comparison of where gender-based literacy gaps are most severe, supported by the out-of-schools-children heatmap.

Demographic Analysis:

1) Medium of Instruction: Donut Chart

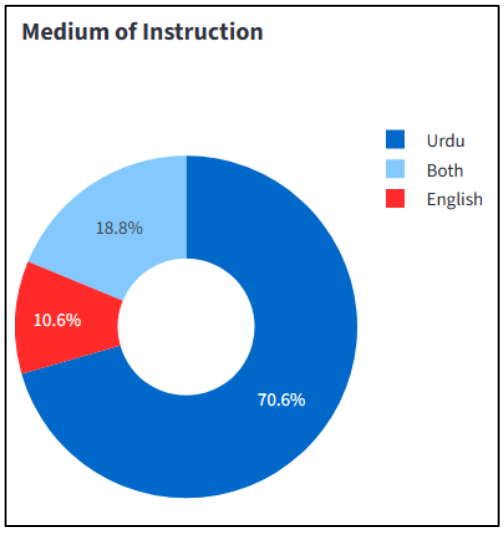


Figure 3: Medium of Instruction

Preprocessing: To prepare the data for the medium (language) of instruction visualisation. The medium column was standardized by converting all the entries to string format and replacing the missing values with unknown. Different representation of the same language (for e.g., 'urdu', 'URDU', etc) were also catered to, to ensure that there are no duplicates. This restructuring allowed accurate frequency counts and prevented duplicate categories in the visualisation.

The donut chart of medium instruction provides an overview of the language used in teaching across schools. This helps us understand whether the schools had one language of instruction or were spread over many. According to this graph, Urdu was a widely used language of instruction amongst the schools across (70.6%). A small percentage of schools (18.8%) used both English and Urdu as their modes of language and a very small percentage (10.6%) used English only. The donut cart in

this case is an appropriate type of chart here, as it allows us to have a clear comparison between the ratios of the different mediums and clearly communicated medium shares.

2) School Location (Urban/Rural): Bar Chart Used

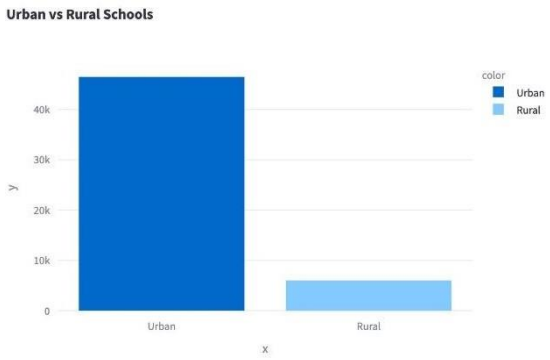


Figure 4: Number of schools in Rural and Urban Punjab.

The data in this case was aggregated into simple frequency counts for both Urban and Rural areas.

The bar chart compares the number of schools in the Rural and the Urban areas telling us the geographic spread of the educational institutions. According to the visualisation, there is a huge difference between the number of schools in the rural and the urban areas. Majority of the schools (approximately 45,000) are present in the urban areas, whereas a small number of schools (approximately 16,000) are found in the rural areas. Bar charts are considered to the be the best type of visualisation in this case, since they clearly display the differences in the category frequencies and allows us to make a comparison between them.

3) School Gender Type: Bar Chart

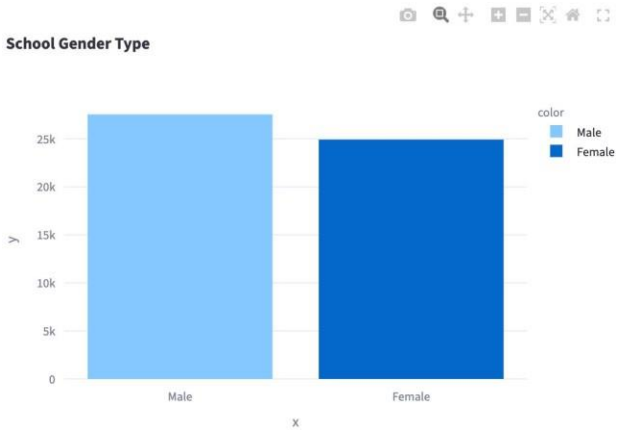


Figure 5: School Gender Distribution.

The gender type bar chart allows us to analyse which gender had the greatest number of schools. By comparing the frequencies of the boys only and girls only we can gain insights on the gender representation and educational access. In this case, there are a greater number of male only schools (approx. 27,500), then female only schools (25,000). The bar chart is an effective way to communicate these findings since it shows categorical data and shows the comparison between the number of school types for different.

4) School Shift Distribution: Bar Chart

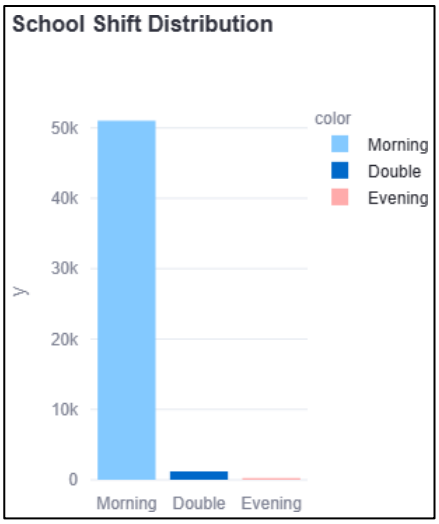


Figure 6: School Shift Distribution

Preprocessing: This graph required some intensive preprocessing, since initially the data had labels like 0,1,2 and 3 for this field. Since these numbers did not correspond to meaningful shift labels, they were replaced with meaningful names. The data was aggregated to produce counts for different shift types, like morning, evening and double.

The school shift bar chart tells us the distribution of the time-based scheduling across institutions. According to the visualisation given above, a majority of schools had the traditional single morning shift, with a small percentage of schools having a double shift and an extremely small number of schools having evening shifts. The bar chart is the most suitable and appropriate type of chart used here, since it allows us to make comparison between the frequencies (the actual number) of the different types of shifts in the schools.

5) School Level Distribution: Horizontal Bar Chart

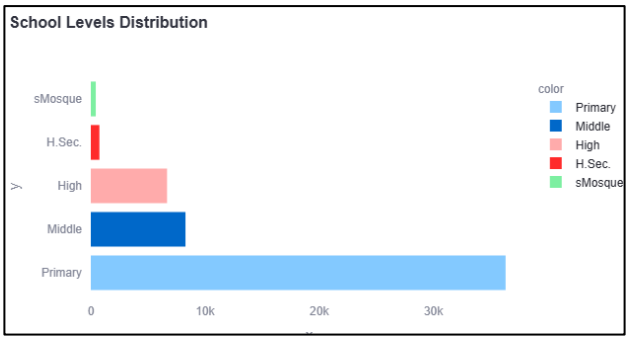


Figure 7: School Level Distribution.

Different labels representing the same category (like 'HS', 'High School', 'H-Sec') were consolidated to maintain uniqueness.

The horizontal bar chart with different colour representing different categories (Primary, Middle, High, High Secondary and Mosque). This visualisation helps us analyze how the education system is layered across different levels. Primary schools have the highest frequency (approximately 38,000) with a drastic reduction in the number of schools as we go up the school level hierarchy. In terms of traditional schooling systems, Higher Secondary Schools have the least frequency. This visualisation also shows another nontraditional schooling system found in Pakistan, the mosques. There are a few mosques in Pakistan as well. A horizontal bar chart is considered as a better option here as it accommodates long category names and shows the frequency comparison of different categories.

6) Non Functional Reasons: Horizontal Bar Chart

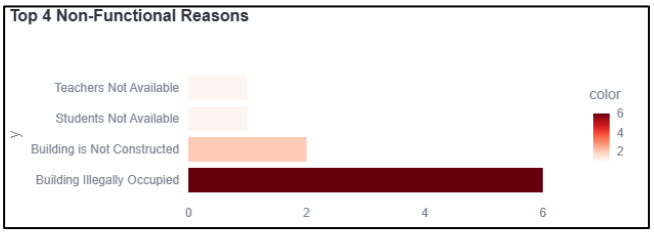


Figure 8: Non Functional Schools Reasons

This horizontal bar chart discovers the primary explanations for why some schools are non-functional. The reasons reveal some challenges such as teachers not being available, students not being available, building not being constructed and building being illegally occupied. The most common reason was the building being illegally occupied. A lot more schools reported to being non-functional because their building was occupied illegally. Horizontal bar chart was preferred in this case as it allowed us to cleanly show the names of the reasons and allowed comparison between different reasons.

Infrastructure Analysis:

1) Building Ownership: Horizontal Bar Chart

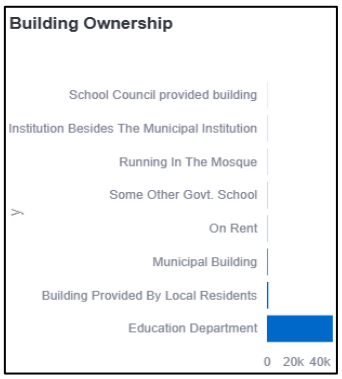


Figure 9: Building Ownership

The horizontal bar chart of building ownership gives us a quick view of who controls the physical school infrastructure. Most schools fall under government education department with a very small number of schools being managed by any other entity. A horizontal bar chart is preferred for this visualisation as it allows to comfortably adjust the labels of the different categories.

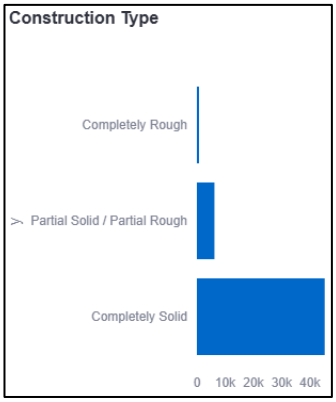
2) Building Condition: Horizontal Bar chart

Figure 10: School Building Condition

This horizontal bar chart reveals the structural health of different schools across the country. Majority of the schools lie in the Satisfying category implying that most schools are in a good condition. There are some schools that fall in the 'Need minor repairing'. Consequently, a small portion of schools also fall in the deteriorating zone, where some schools are in a condition where a part of the building is under dangerous zone. Some schools need complete repair, and a very small portion of the school is completely in a dangerous zone, implying that the school needs urgent attention or complete removal. This visualisation helps us see the structural issues spread across different schools. Use of the horizontal bar chart allows us to

compare the frequencies of different categories and how the infrastructure is spread out across different categories.

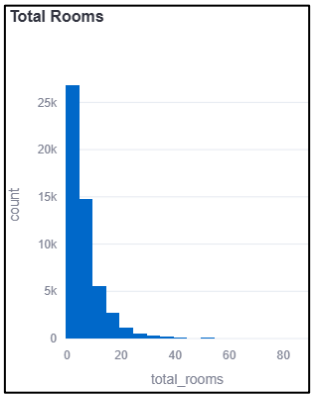
3) Construction Type: Horizontal Bar Chart



This chart helps us understand how the schools are physically built. There are three different categories of construction type including Completely Solid, Partial Solid/Partial Rough and Completely Rough. Most of the schools have a completely solid construction style, whereas some schools have a partial rough and partial solid. A very small number of schools have a completely rough infrastructure. A horizontal bar chart allows fair comparison over different categories.

Figure 11: Building Construction Type

4) Total Rooms: Histogram



The histogram shows total number of rooms shows how school sizes vary across the dataset. The histogram has a negative skewness and many schools (>25,000) appear to have only a small number of classrooms (<10), indicating they are very small-scale schools. This tells us about the limited capacity and infrastructure of the schools. As the number of classrooms increase, the frequency of the classes decreases. The histogram proves to useful ere as it allows us to see clustering and skewness.

Figure 12: Total number of classrooms in a school.

5) Functional and Dangerous Classrooms:

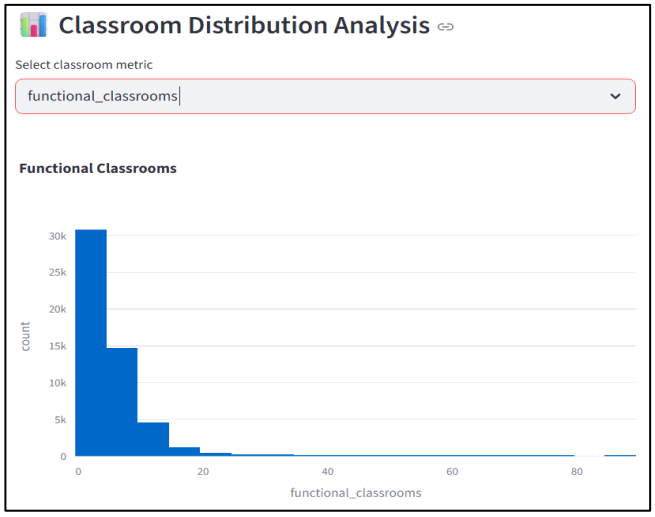


Figure 14: Schools with Functional Classrooms

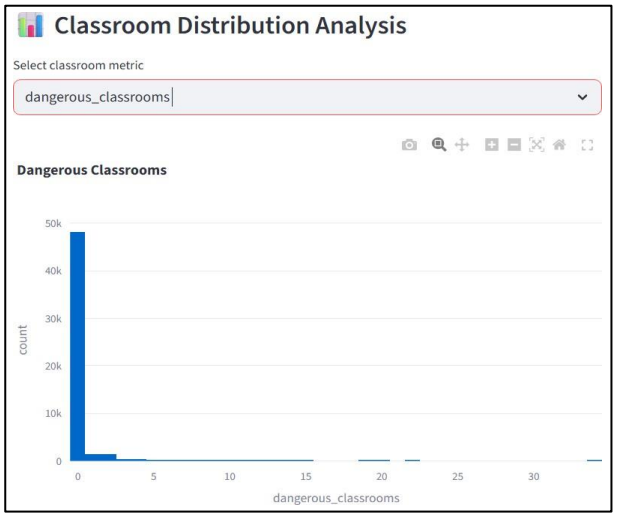


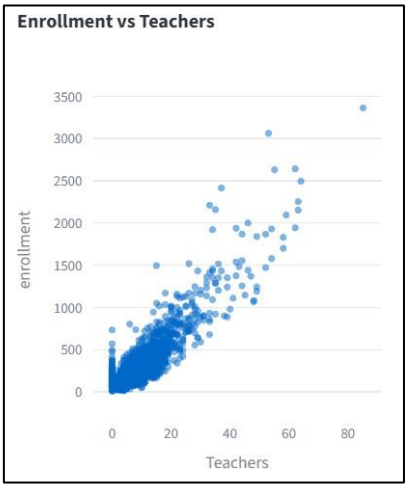
Figure 13: Schools with Dangerous Classrooms

These histogram graphs are interactive and have a filter which allows the user to select between Functional classrooms and dangerous classrooms based on what the user wants to see. The histogram for the functional classrooms reveals more about the usable infrastructure. Many schools have a very small number of functional classrooms, despite having several rooms overall. This gap between total rooms and functional rooms points at possible maintenance issues, incomplete instruction or unsafe conditions. The histogram here helps us analyse this pattern.

The histogram for the dangerous rooms plot tells us about the severity of safety concerns. The histogram shows a long tail, which shows that although many schools report zero dangerous classrooms, some schools have multiple unsafe rooms. This distribution signals that the issue is not isolated and may affect many students.

Staf f & Student Analysis:

1) Enrollment vs Teachers: Scatter Plot

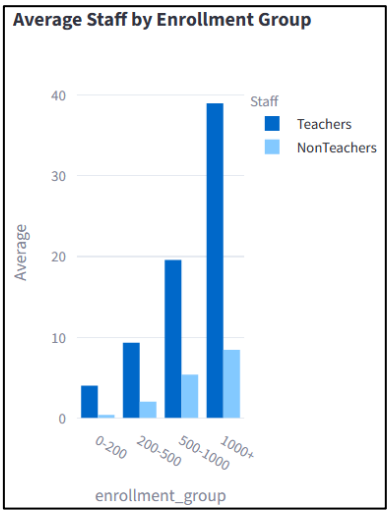


A random sample of 3000 schools was taken from the dataset to avoid overplotting and to ensure the scatter plot remained readable.

This visualisation tells us about the relation between the number of teachers in an institute and the enrollment. This shows us a positive increasing relationship, indicating the more the number of teachers there are, the higher the enrollment for the school is. More number of schools lie in the region where they have less number of teachers (<30) with low enrollment depicting small scale schools. The scatter plot is ideal here as it exposes outliers and systemic imbalance that summary statistics might hides.

Figure 15: Enrollment vs number of teachers

2) Average Staff by Enrollment Group: Grouped Bar chart



To compare the staffing patterns across differently sized schools, enrollment values were grouped into bins of different intervals. This aggregation transforms raw counts into a more interpretable patters across school sizes.

The grouped bar chart clearly illustrates how staffing scales with school sizes. Larger schools have a greater number of teachers and non-teaching staff on average. However, the number of non-teaching staff is very less as compared to the teaching staff.

Figure 16: Average staff by number of enrolled students

3) Teacher Student Ration by Level: Box Plot

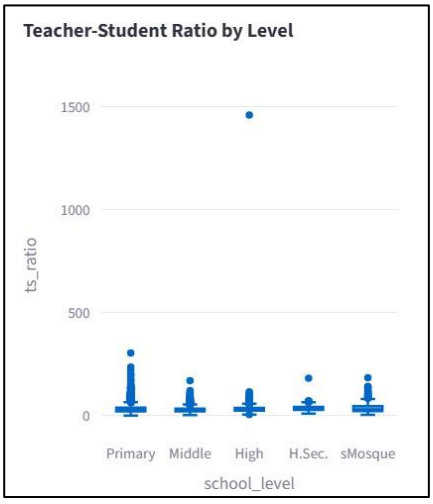


Figure 17: Teacher student ratio by Level

A derived feature ts_ratio (teacher student ratio) was calculated by dividing enrollment by Teachers, with divisions by zero replaced by NaN to prevent distorted ratios. These ratios were then compared across school levels (Primary, Middle, High, Higher Secondary), all standardized into clean categorical labels.

The box plot provides a clear level-wise comparison of teacher student ratios, showing typical ratios within each category and the extreme outliers. In all school levels, the median ratio lies way below but the outliers such as 1500 in high schools, indicate severe staffing deficiencies in certain institutions. These outliers are meaningful as they indicate schools with extremely high enrollment per teacher, pointing to overcrowded classrooms and limited instructional capacity.

Using a box plot helped us analyse how sufficiently institutions are staffed and how they are lacking using the outliers.

4) Total Computers vs Students trained by the Internet: Grouped Barchart

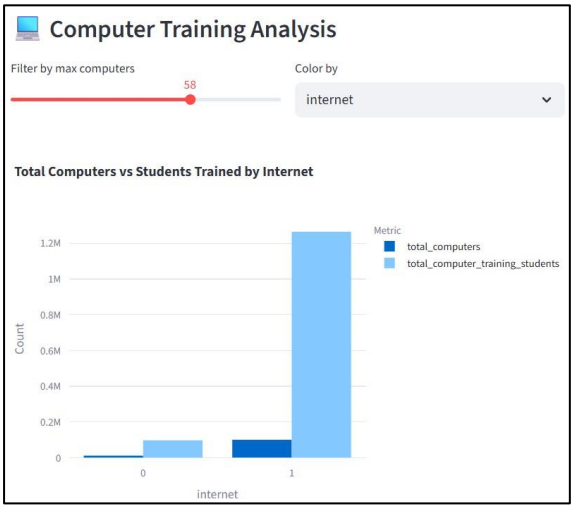


Figure 18: Total Computers vs Students trained by Internet.

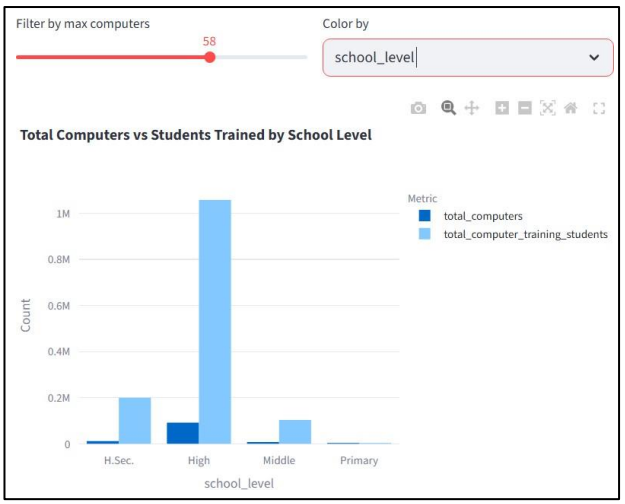


Figure 19: Total Computers vs Students trained by School Level.

For this visualisation, a sample of schools that reported both nonzero computer counts, and nonzero numbers of students trained were included in this chart to ensure meaningful comparisons. On this filtered dataset, total computers and total computer-trained students were aggregated by a user-selected category (e.g., school level, location, or internet availability). 0 and 1 here represent the presence and absence of internet.
This visualisation shows that schools with internet connection had far greater number of computers than the schools in the without internet. This highlights the lack of facilities provided to the schools in the rural areas. This bar chart also reveals a critical issue. Many schools train significantly more students than their computer resources should reasonably support. This uneven distribution points to bottlenecks in equipment availability and in scheduling, staffing, or maintenance.

5) Computer Capacity vs Students Trained: Scatter Plot

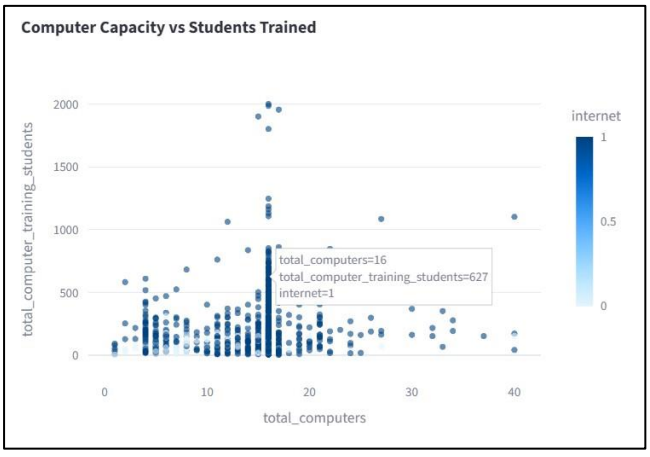


Figure 20: Computer capacity vs Students Trained

This scatter highlights the inconsistency between technological capacity and actual utilization. Instead of a moderate trend displaying schools with more computers training more students, many schools clustered near the bottom-left corner, meaning both computer access and training levels are low. There are also horizontal outliers where there are schools that train many students despite having very few computers, implying shared-device constraints and limited digital infrastructure. The scatter plot helped us identify the outliers and the general trend between number of computers and the computer trainings received.

Timeline Analysis:

1) School Establishment & Upgrades overtime: Line Chart

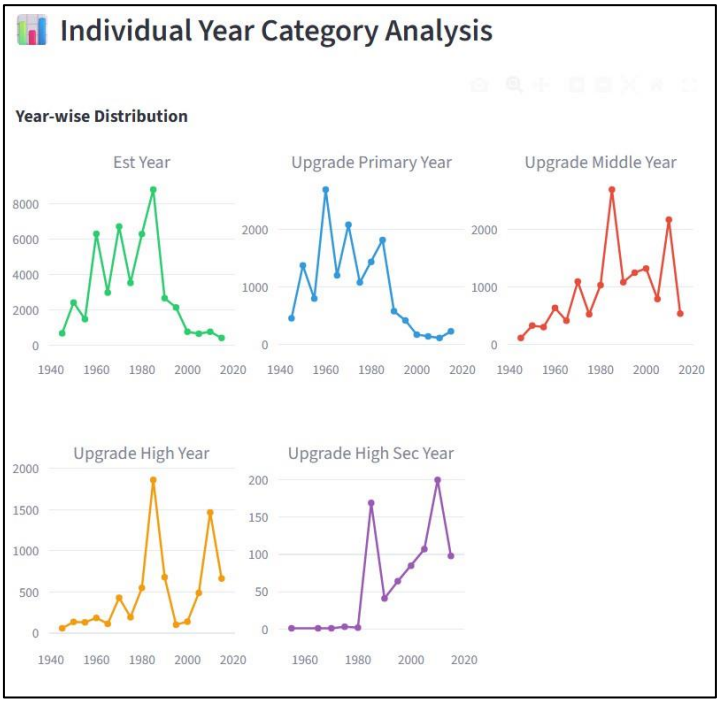


Figure 21: Trend of Educational Institution Upgradation over the years.

The interactive line chart highlights the upgrade of different levels of institutions over several years starting from 1947 till 2020. Each graph represents a different category with the first one representing the number of schools in total over the years.

The establishment plot reveals that there have been frequent waves of educational expansions in which institutes were upgraded. However, in the late 1990's till 2020, the upgradation of institutes has started to decrease.

The upgrade-primary timeline illustrates how many schools transitioned to a primary-level upgrade in each year. The visualisation shows that in the earlier years there was an increased number of upgradations of primary level institutes, but this has declined over the years. The gradual decline in recent years may reflect either improved initial construction standards or changing administrative criteria.

The middle-school upgrade timeline reveals a slightly different pattern compared to primary upgrades, with occasional pronounced spikes. These sudden increases suggest targeted interventions or policy shifts aimed at expanding middle-level access.

The visualisations of the high year and high secondary years, have lesser number of peaks indicating that the upgradation of these institutions was consistent over the years with some occasional spikes. The visualisation of the High Sec Year shows an increasing trend. It implies that the upgradation of the high secondary year institutions increased over time, however, they are still less than other levels of institutions.

Line charts were the most suitable type of graph for this temporal visualisation as they helped us identify a trend between the number of institutes and the year.

2) Building Condition by Establishment Year: Box Plot

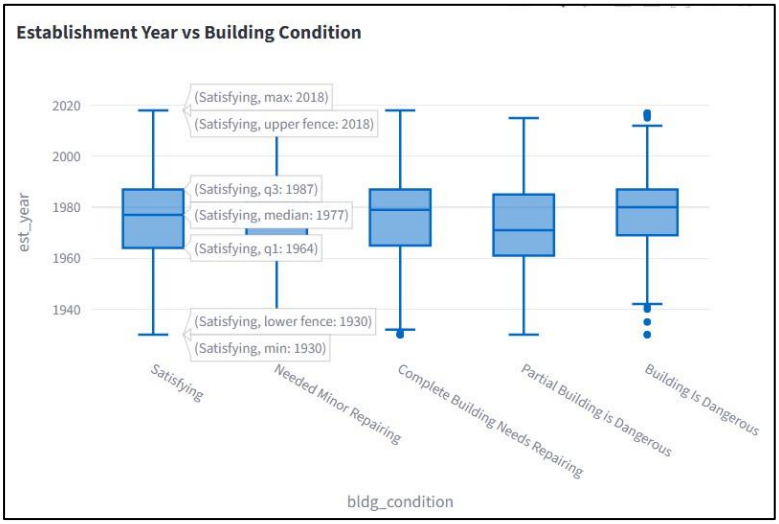


Figure 22: Establishment year vs Building Condition

The interactive box plot demonstrates a clear relationship between building age and condition. Schools classified as "Satisfactory" tend to have much more recent establishment years, while those marked "Major Repair Needed" or "Building Dangerous" skew significantly older. The spread within each category also shows that not all older schools are deteriorated. However, the general trend strongly supports the intuitive link between age and condition.

The box plot is effective because it shows central trends, dispersion, and outliers, giving a complete picture of how structural aging impacts safety and usability.

PCA Analysis:

To prepare the data for PCA, the facility-related variables such as functional classrooms, usable toilets per 100 students, and computers per 100 students were standardized. Each feature was scaled so that they were all on the same range and none of them dominated the analysis simply because they had larger numeric values. Schools with missing values in any of these facility columns were removed from this step so that PCA could be computed properly.

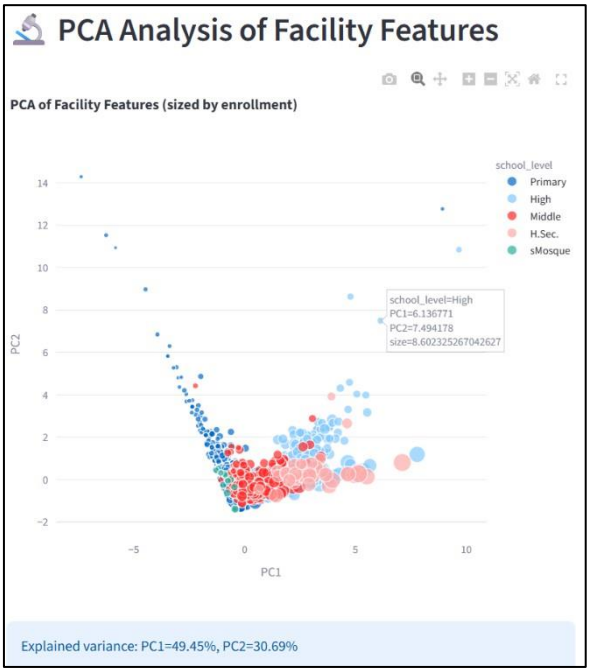


Figure 23: PCA of Facility Features

After standardization, the PCA model generated two new components, PC1 and PC2, which together explain most of the variation in the facility data.

The interactive bubble plot of PC1 versus PC2 gives a compact, visual summary of how schools differ in their facility conditions. Each point represents a school, and its color shows the school level (Primary, Middle, High, etc.), while the size of the point reflects its enrollment. A loose distance between different levels of institutions is apparent. Primary schools tend to cluster toward one region of the PCA space, while middle and high schools stretch into different areas. This means that school levels differ not only in curriculum but also in their overall facility profiles. PC1 appears to behave like an "overall resources" axis. Schools with higher PC1 values generally have better facility indicators, such as more functional classrooms or more computer access.

The PCA visualization is useful because it takes several facility measures and compresses them into two interpretable dimensions, making it easier to spot patterns and groupings that are hard to see from raw data tables alone.

KPK Non-Functional School Analysis (2021):

We had different datasets for KPK non-functional schools which we used to support the analysis. The dataset included a primary and middle KPK government schools' data. The datasets were merged for complete analysis.

1) Primary vs Middle Non-Functional Schools: Bar Chart

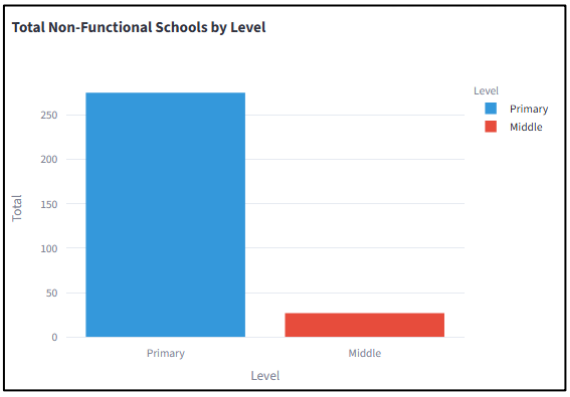


Figure 24: Primary vs Middle Non-Functional Schools in KPK.

The total number of non-functional schools was computed separately for each level, creating an easy to compare structure. This bar chart gives a clear comparison of how many schools are non-functional at the primary level versus the middle level in the KPK province. The data shows a large disproportion. Primary schools make up most non-functional institutions, while middle schools account for a much smaller share. This imbalance suggests that system weaknesses occur very early in the educational pipeline. It also hints at infrastructure neglect at foundational stages, which ultimately affects progression into higher schooling. The bar chart gives an easy to view comparison between the two categories.

2) Gender Breakdown of Non-Functional Schools: Donut Chart

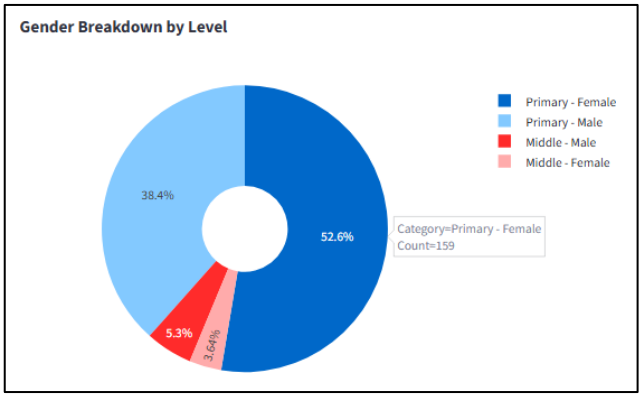


Figure 25: Gender Breakdown of non-functional schools in KPK.

For this visualization, male and female non-functional counts were extracted from both primary and middle datasets. Each value was converted to numeric, and percentages were calculated relative to total non-functional numbers to compute gender shares.

The gender breakdown donut chart reveals a significant imbalance in the composition of non-functional schools. A huge proportion of these institutions, especially at the primary level, serve female students. This means that when a school becomes non-functional, girls are often majorly affected. At the same time, male schools also appear among

the counts, but their share is significantly smaller. This pattern aligns with broader educational trends in Pakistan where female access to education tends to be more vulnerable to institutional failures.
The donut chart helps in easier comparison of all the categories.

3) Non-Functional Schools by District and Level: Grouped Bar chart

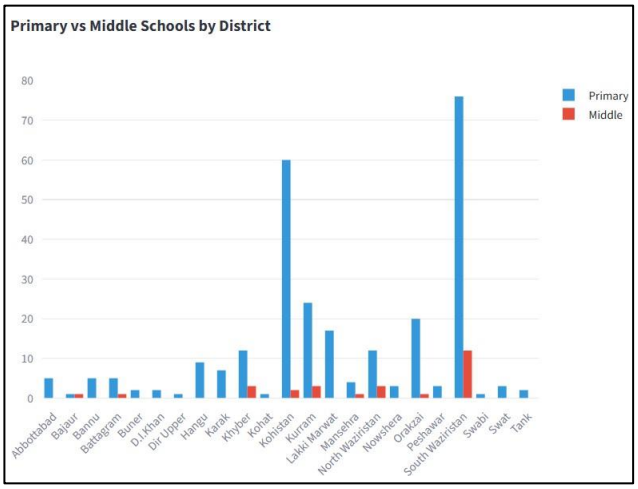


Figure 26: Primary vs Middle Non-Functional Schools by District

This visualisation allows a comparison of non-functional schools over different KPK districts. According to the visualisation, the most non-functional primary and middle schools are in South Waziristan followed by Kohistan. This graph also shows us that there are more number of primary non-functional schools than middle non-functional schools. The grouped bar chart allows a comparison of both primary and middle non-functional school across several KPK districts.

4) Severity Classification of Districts:

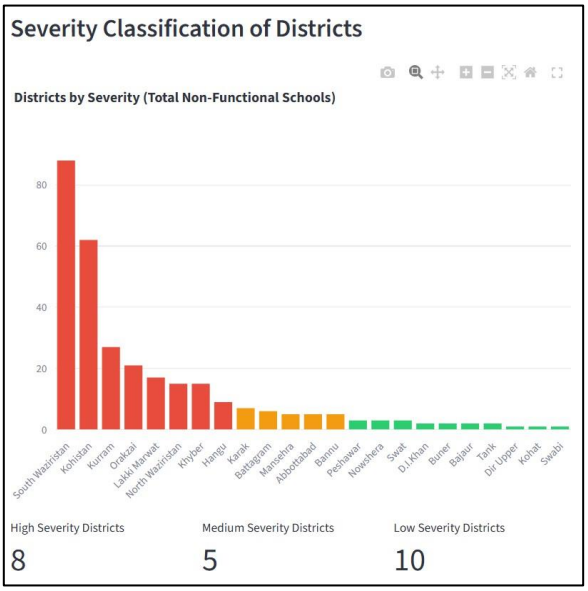


Figure 27: Severity Classification of Districts.

To classify district severity, the total number of non-functional schools in each district was computed by summing primary and middle values. These totals were then divided into three quantiles (Low, Medium, High severity) using statistical cutoffs. This quantile-based labeling ensured that districts were categorized fairly and consistently.
The severity chart summarizes the entire KPK situation by categorizing districts into low-, medium-, and high-risk groups. High-severity districts stand out immediately, showing clusters where the educational system is facing critical failures. Medium-severity districts indicate emerging challenges that may worsen without intervention. Low-severity districts demonstrate relatively stable functioning. By converting raw numbers into qualitative categories, this visualization communicates where policymakers should focus their attention first. Its simple color-coded layout makes it highly effective for decision-making.

5) Education Pipeline Analysis:

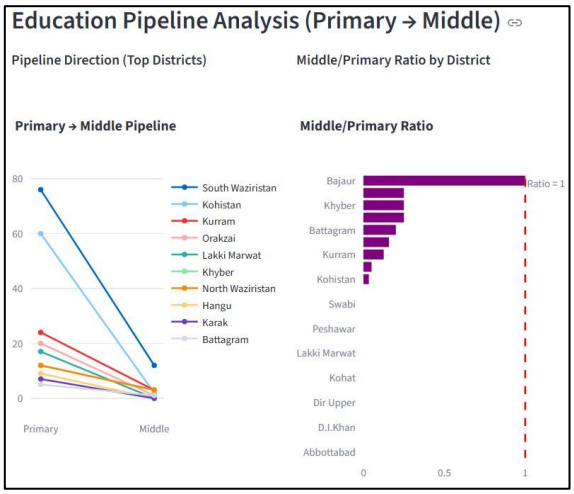


Figure 28: Education Pipeline Analysis

A new derived measure, ratio_middle_to_primary, was calculated for each district by dividing the number of middle-level non-functional schools by the number of primary-level non-functional schools. Cases where primary count was zero were excluded to avoid misleading infinite ratios. These ratios were then sorted to identify districts with values greater than 1, indicating pipeline breakdown.

This ratio-based visualization highlights systemic issues that are not obvious from simple counts. In a healthy education pipeline, primary-level problems should be more common, because they reflect foundational access. However, the districts where the middle-to-primary ratio exceeds 1 show the opposite. More non-

functional middle schools than primary schools are present. This suggests that schools may be opening as primary institutions but failing to support students beyond grade five. Such pipeline breakdowns often signal shortages of subject-specialist teachers, inadequate transition facilities, or weak district-level resource allocation. The visualization is powerful because it captures structural breakdown rather than just raw numbers.

6) Gender Analysis of Non-Functional Schools: Stacked Bar chart

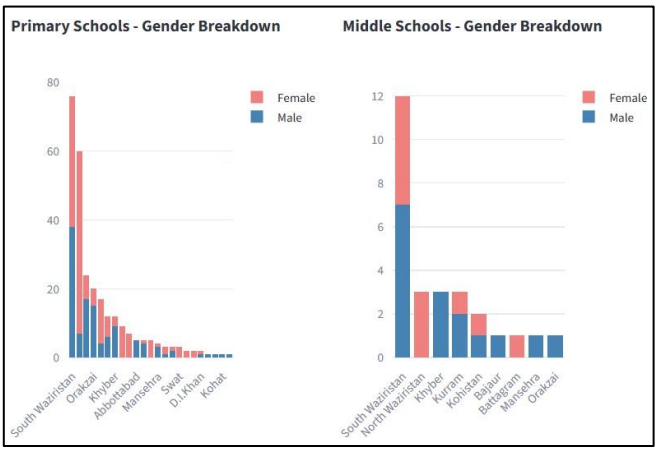
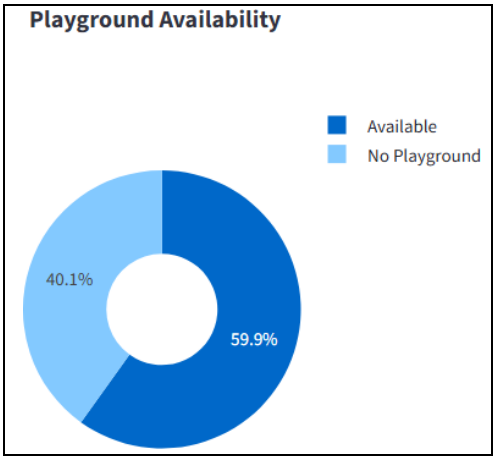


Figure 29: Gender Analysis of Non-Functional Schools

The stacked bar charts reveal how gender disparities differ across districts for both primary and middle non-functional schools. In the primary level chart, several districts, such as South Waziristan, Orakzai, and Khyber, show significantly higher closures among female schools compared to male schools. This suggests that when schools become non-functional, girls are disproportionately affected, reflecting broader access and structural disadvantages. The middle-school chart shows a similar but less pronounced pattern, with female schools again forming a visible share of the closures.
The stacked bar chart is most suited because it allows a direct comparison of male and female counts within each district while also showing the total magnitude of the problem.

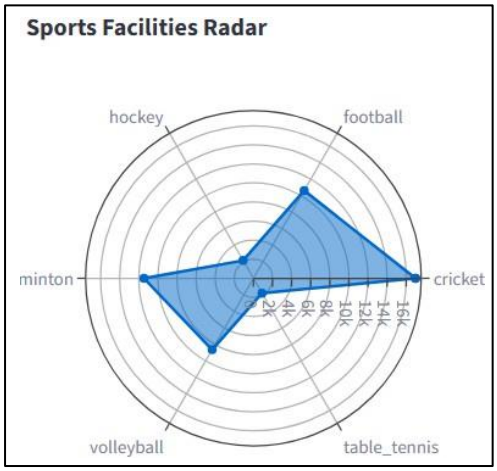
Sports and Labs:

1) Playground Availability (Pie Chart):



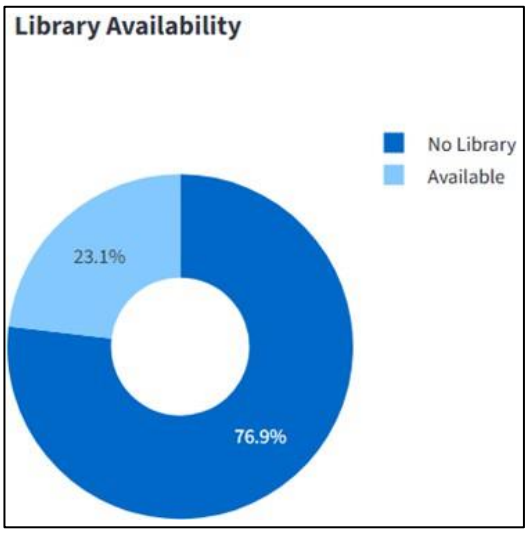
With 21,036 schools lacking playgrounds and 31,431 having them, this pie chart shows approximately 60% availability, highlighting inadequate recreational spaces that may limit physical activity and student well-being in nearly half the schools.

2) Sports Facilities Radar Chart:



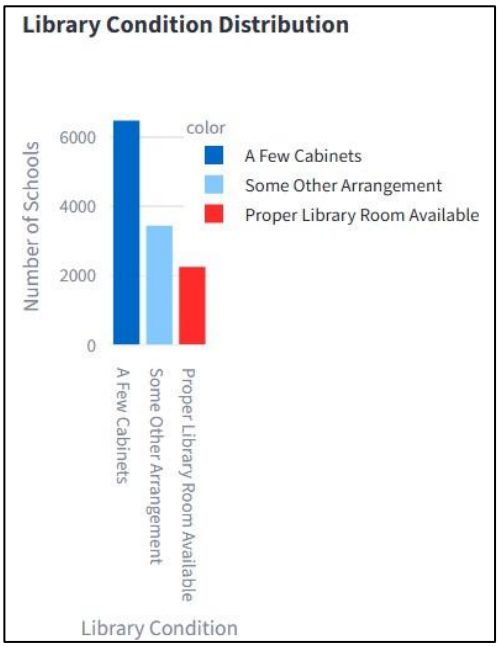
Cricket leads with 16,970 facilities, followed by badminton (11,462) and football (10,623), while hockey (2,159) and table tennis (1,811) are scarce, indicating uneven sports provision favoring popular games over diverse options.

3) Library Availability (Pie Chart):



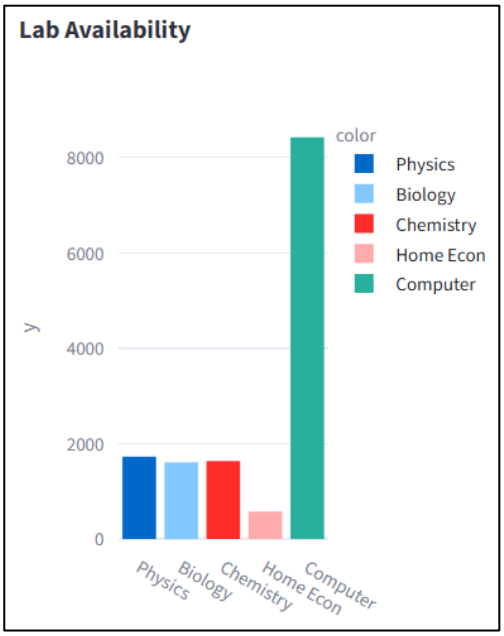
Only 12,125 schools have libraries out of 52,467 total, meaning 76.9% lack access, which restricts reading resources and academic support for most students.

4) Library Condition Distribution (Bar Chart):



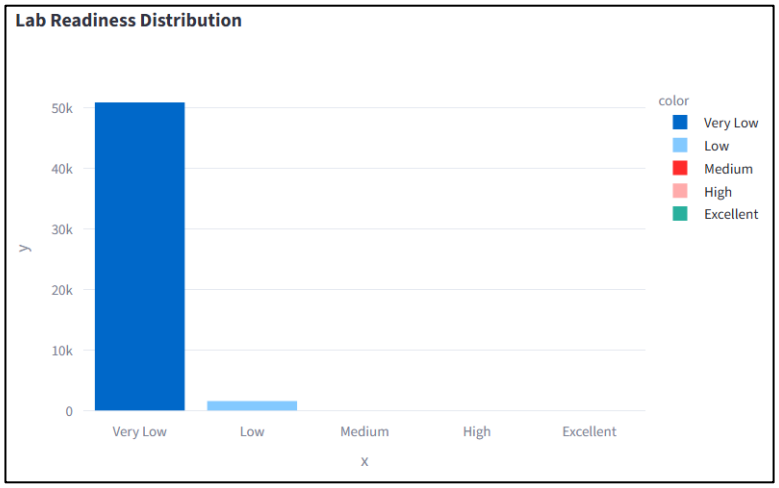
6,478 schools have "A Few Cabinets," 3,439 use "Some Other Arrangement," and 2,244 have "Proper Library Room," showing most libraries are makeshift, with just 18% in dedicated spaces.

5) Lab Availability (Bar Chart):



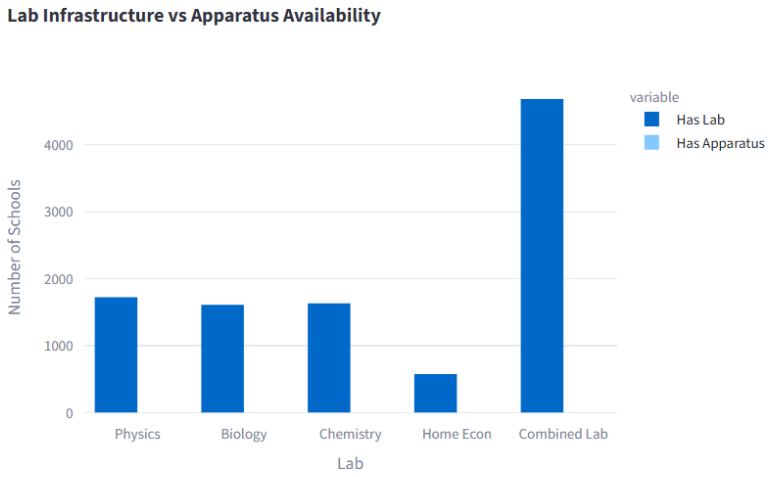
Computer labs dominate with 8,426, followed by chemistry (1,633) and physics (1,725), while home economics (577) is minimal, reflecting a tech-heavy focus over traditional sciences.

6) **Lab Readiness Distribution (Bar Chart)**:



50,883 schools rate "Very Low" and 1,587 "Low," with none in medium or higher, indicating widespread inadequacy in lab infrastructure across 97% of schools.

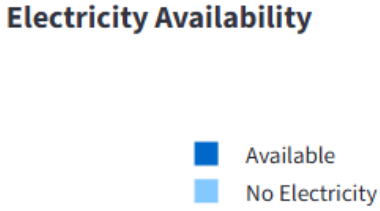
7) **Lab Infrastructure vs Apparatus Availability (Grouped Bar Chart)**:



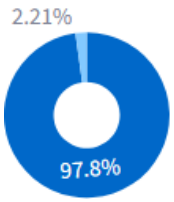
Physics labs (1,725) and biology (1,609) exist but apparatus counts are zero, revealing a critical gap where facilities lack equipment, hindering practical education.

Utilities:

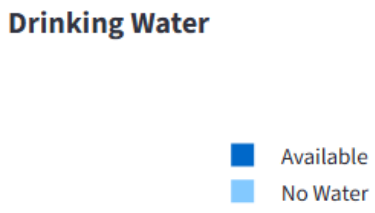
1) Electricity Availability (Pie Chart):



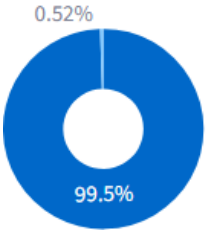
51,312 schools have electricity versus 1,152 without, showing 98% access, but the small gap affects basic operations in over 2,000 institutions.



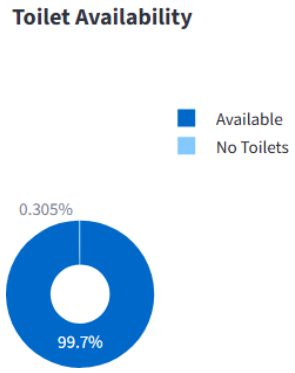
2) Drinking Water (Pie Chart):



52,197 schools provide water, with only 267 lacking, indicating 99.5% availability and strong coverage of this essential utility.

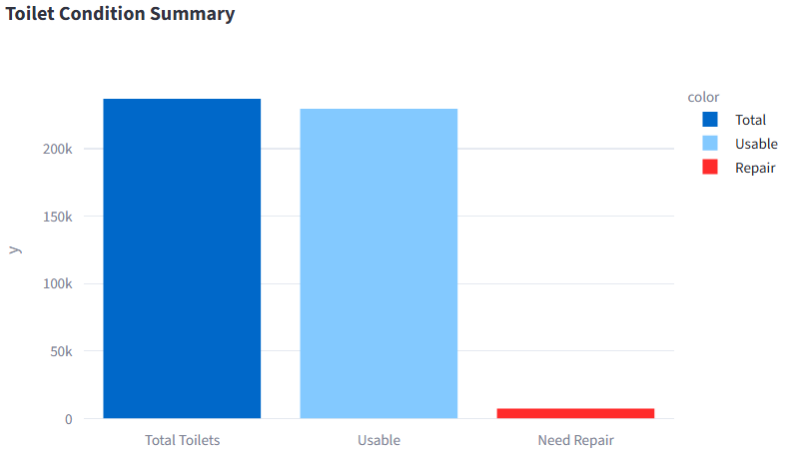


3) Toilet Availability (Pie Chart):



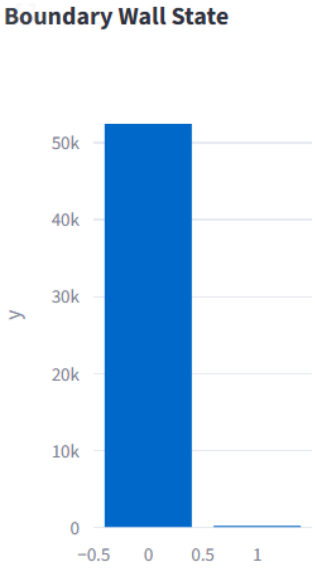
52,299 schools have toilets, compared to 160 without, meaning 99.7% provision, though quality and maintenance remain concerns.

4) Toilet Condition Summary (Bar Chart):



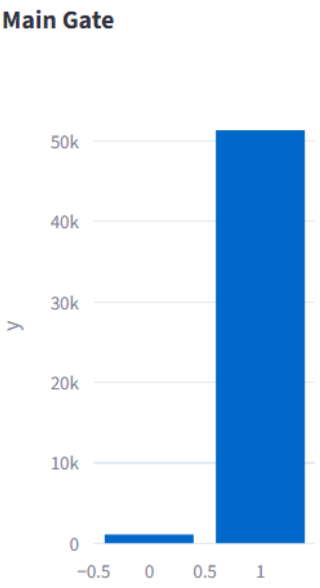
Total toilets sum 237,038, with 229,469 usable and 7,376 needing repair, showing 97% functionality but 3% requiring fixes.

5) Boundary wall (Bar Chart):



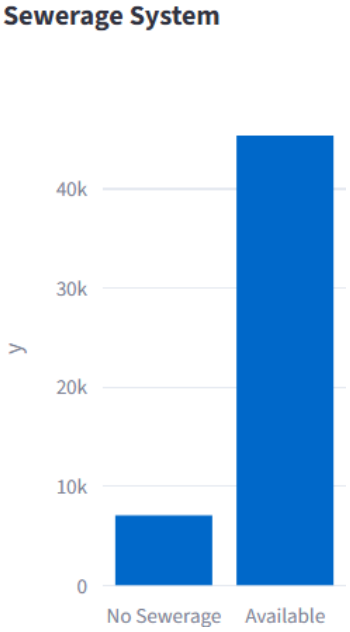
52,468 schools lack proper walls (value 0), with only 2 having them, highlighting severe security deficiencies in nearly all schools.

6) Main Gate (Bar Chart):



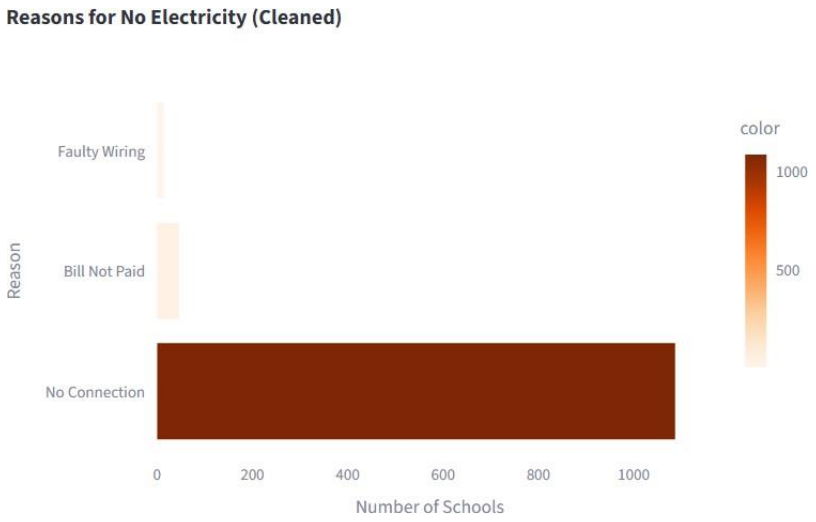
51,367 schools have gates, 1,102 lack them, and 1 is partial, indicating 98% presence but minor gaps in access control.

7) Sewerage System (Bar Chart):



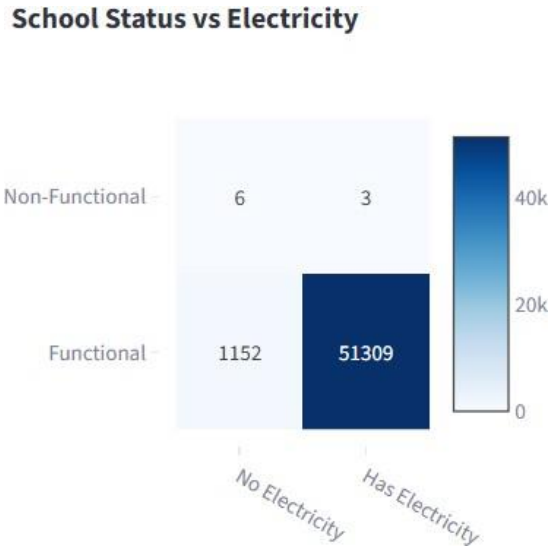
45,381 schools have sewerage versus 7,087 without, showing 86% coverage, with rural or older schools likely underserved.

8) Reasons for No Electricity (Horizontal Bar Chart):



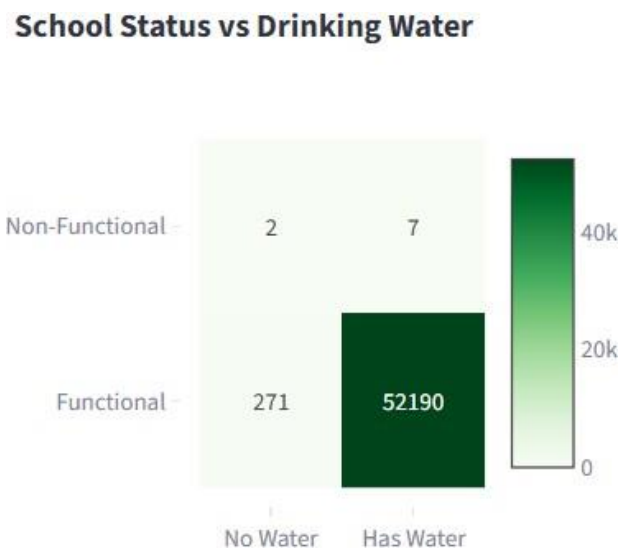
Top reasons include "No Connection" (1,086 schools), "Bill Not Paid" (44), and "Faulty Wiring" (13), pointing to infrastructure and payment issues as primary barriers.

9) School Status vs Electricity (Heatmap):



Functional schools dominate with 51,309 having electricity and 1,152 lacking, while non-functional ones show minimal access, linking utility to operational viability.

10) School Status vs Drinking Water (Heatmap):



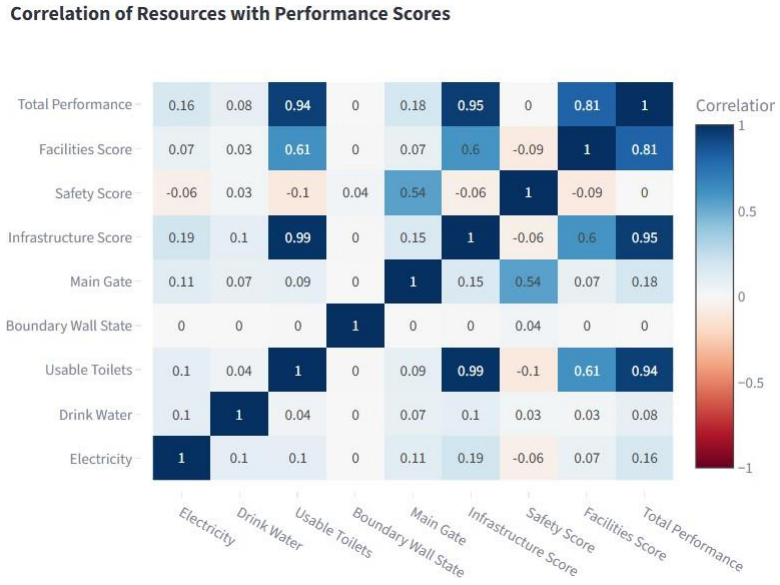
Functional schools have 52,184 with water and 265 without, versus non-functional with 7 and 2, emphasizing water as a key factor in school functionality.

11) Resource Availability Matrix (Area-wise)

This heatmap compares average resource availability in Urban and Rural schools. Urban schools show higher access to key facilities, especially electricity and drinking water, whereas Rural schools have

noticeably lower availability. Both locations score 0 for boundary wall condition, indicating a widespread lack of boundary walls across all areas. Overall, the visualization highlights a clear urban–rural gap in essential utilities, while also revealing a common security infrastructure issue shared by both settings.

12) Resource Correlation with Performance:



This correlation heatmap shows how different school facilities relate to overall performance. The strongest links appear between Infrastructure Score and Total Performance (0.95) and between Electricity and Infrastructure (1.00), indicating that reliable electricity and strong infrastructure are key drivers of school performance. Usable Toilets (0.99) and Drink Water (1.00) also correlate highly with infrastructure, reinforcing their importance for functional learning environments. In contrast, features like the Boundary Wall show weaker or zero correlations with performance-related scores, suggesting they contribute less directly. Overall, the clustering of high values highlights that core utilities and infrastructure upgrades should be prioritized to improve school performance.

District-wise Average Score Comparison: Grouped Bar Chart

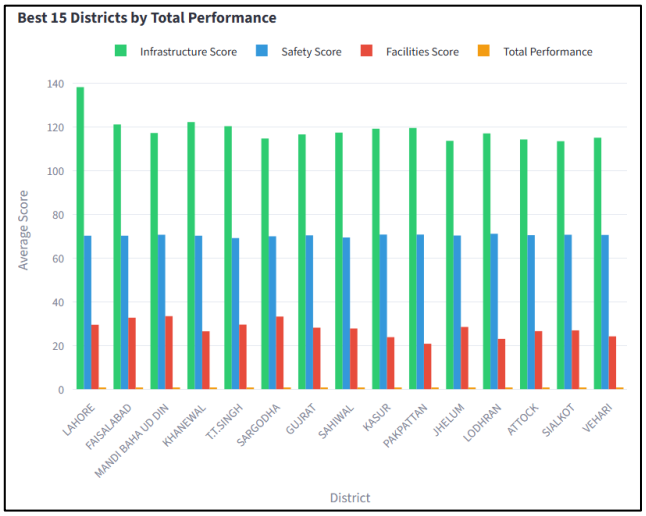


Figure 30: District-wise Average Score Comparison

This grouped bar chart compares the top performing 15 districts across all four composite scores: Infrastructure**,** Safety**,** Facilities, and the combined Total Performance score. Because each of these metrics is constructed from weighted components such as building condition, utilities, classroom functionality, safety measures, and facility availability, this visualization helps us see which factors drive high performance in leading districts.

Overall, this chart shows that top districts excel mainly because of strong infrastructure and safety foundations, while facility-related components remain the weakest dimension even among the best performers.

Resource Availability Heatmap by Location:

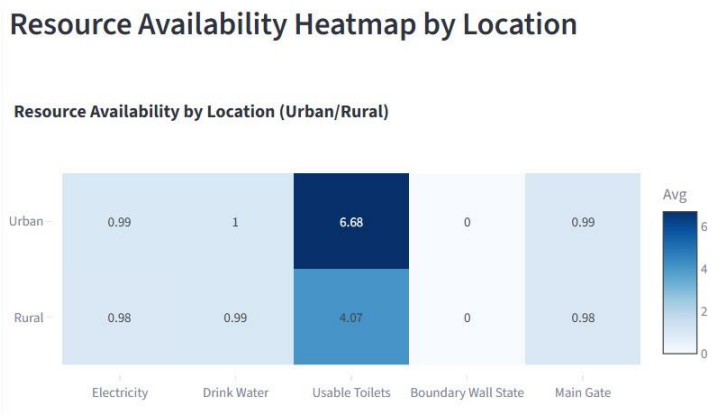


Figure 31: Resource Availability Heatmap by Location

To generate this heatmap, the dataset was grouped by the school_location variable, separating schools into "Urban" and "Rural" categories. Key resource indicators such as electricity, drinking water, usable toilets, boundary wall state, and main gate availability were converted into numeric form so that averages could be computed. Binary variables (e.g., electricity, drink_water, main_gate) were standardized to 0 and 1, while count-based variables like usable_toilets were left in numeric form to preserve true differences in facility levels.

The heatmap compares the availability of essential school resources between urban and rural areas. Averages close to 1 show near-universal availability, and the visualization quickly reveals that electricity and drinking water are consistently accessible across both settings. However, a notable contrast appears in sanitation, urban schools have an average of 6.68 usable toilets, while rural schools average only 4.07. This gap suggests that sanitation facilities lag noticeably in rural regions, likely reflecting maintenance challenges or lower infrastructure investment. Boundary wall values appear as zero in this view due to how the variable was encoded, but main gate availability remains high in both groups. Overall, the heatmap helps highlight that although basic utilities are widely available, sanitation and facility quality differ sharply between urban and rural schools.

Political Party Performance Comparison (PTI vs PMLN vs Independent):

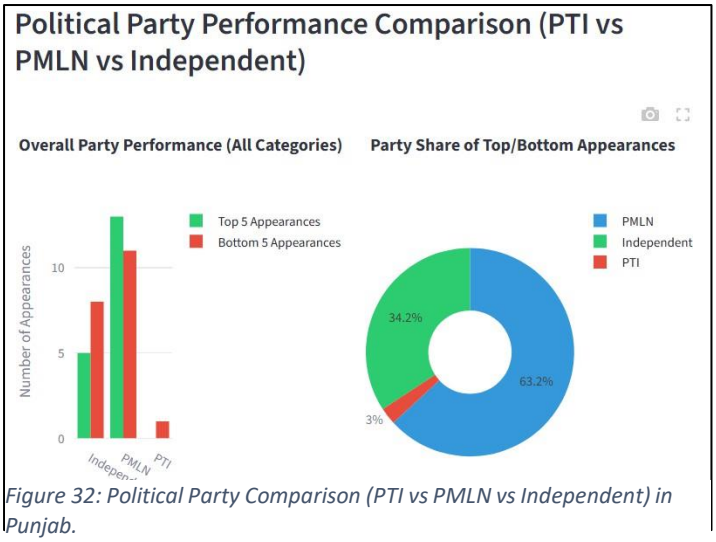


Figure 32: Political Party Comparison (PTI vs PMLN vs Independent) in Punjab.

This visualisation compares how frequently different political parties appear in the top-performing and bottom-performing districts across all score categories. The bar chart shows the raw number of appearances, while the donut chart summarizes the overall share for easier interpretation.

From the bar chart, we see that PMLN candidates appear most often in the top 5, suggesting that many high-performing constituencies are represented by PMLN. However, PMLN also have a high bottom 5 appearances as well. The least number of appearances is of PTI.

This is reinforced by the Donut chart which shows the combined distribution of top and bottom appearances across all parties. PMLN dominates the share due to its high presence across districts, but this dominance includes both strong and weak performers. PTI's share is the smallest, suggesting lower influence or fewer constituencies represented in this dataset.

Top 5 appearances by Party & Category: Grouped Bar chart

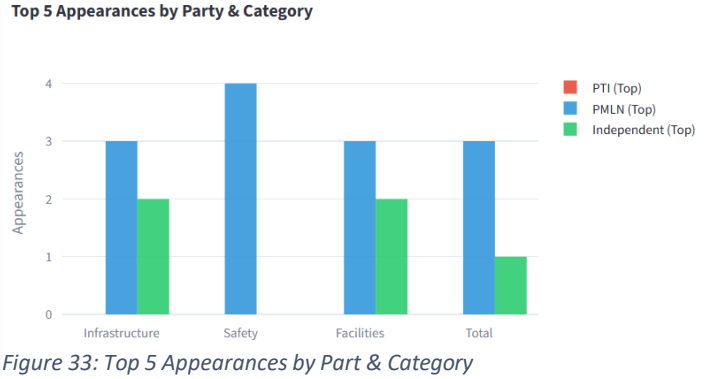


Figure 33: Top 5 Appearances by Part & Category

For this visualization, the data was filtered to focus only on the top 5 appearances for each political party in different categories: Infrastructure, Safety, Facilities, and Total Performance. The data was aggregated by counting the number of appearances for each party in each category.

The bar chart shows the distribution of top appearances across key performance categories for PTI, PMLN, and

Independent candidates. PMLN dominates in all categories, and is the sole winner of the Safety category, where it appears the most in the top 5 rankings. PTI does not appear in any category. Independent candidates, while having some appearances, generally rank lower across all categories. This visualization supports the analysis that PMLN consistently outperforms other parties in terms of infrastructure, safety, and overall performance, making it the top performer in the analyzed dataset.

Explanatory Data Analysis:

District-wise Average Score Comparison: Grouped Bar Chart

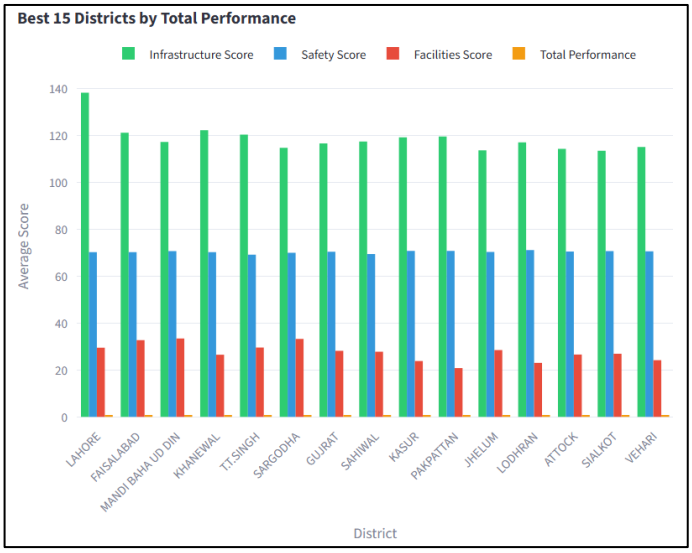


Figure 34: Top 15 Best Performing Districts in Punjab.

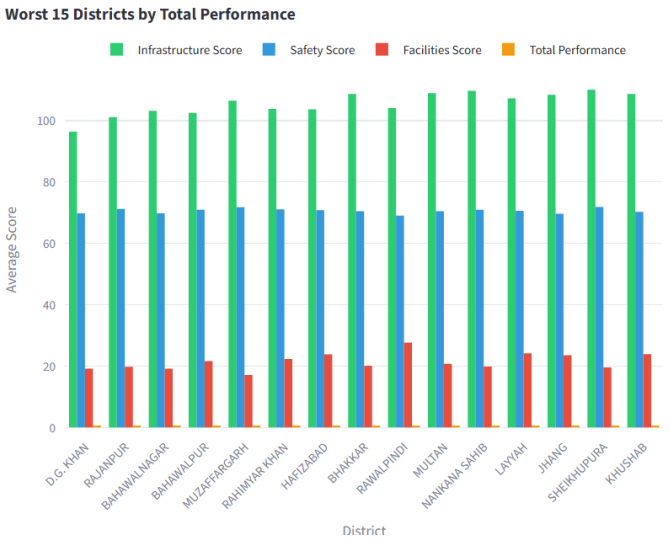


Figure 35: Bottom 15 Worst Performing Districts in Punjab.

This graph provides a detailed comparison of top-performing and bottom-performing schools across three critical performance dimensions: Infrastructure, Safety, and Facilities. By looking at the top and bottom 15 schools, we can identify key areas of strength and weakness.

The first thing that stands out is that Infrastructure is the strongest variable in the performance score, with both the Top 15 and Bottom 15 schools showing relatively higher scores. Lahore and other urban districts appear to perform well in this area, likely due to better school buildings, functional classrooms, and essential utilities like electricity and clean water. However, rural districts such as Kasur and Vehari show significant gaps in infrastructure despite having similar scores compared to Safety and Facilities.

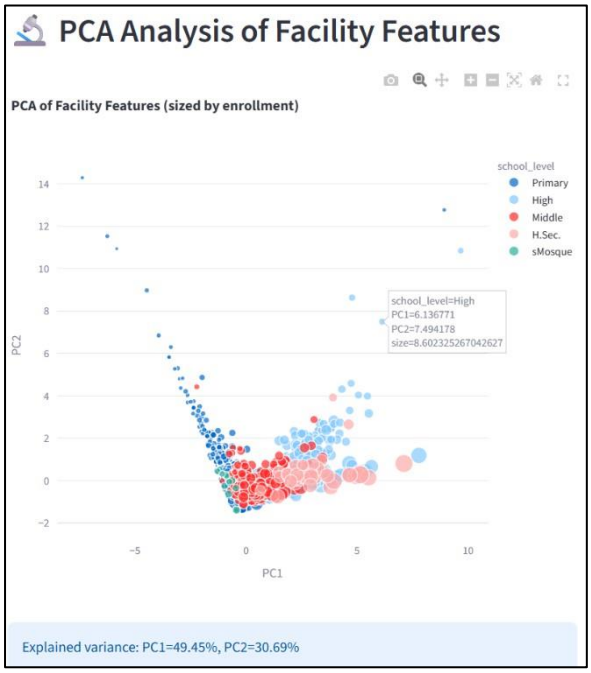
Safety, on the other hand, reveals a more concerning pattern. Even in top-performing schools, safety scores are not as high. This suggests that school security measures, such as boundary walls, gates, and safe classrooms, are insufficient, even when infrastructure is sufficient. Bottom-performing schools, however, display low safety scores, indicating potential structural safety issues that need urgent attention.

Facilities show the second most variation. While top-performing schools (mainly in urban districts) generally score well for library access, computer labs, and sports facilities, the bottom 5 schools have low scores, reflecting the lack of these resources. This shows that, while Lahore and similar districts benefit from better facilities, many rural schools face severe resource shortages, particularly in specialized educational spaces like science labs and sports facilities.

Conclusion:

This graph confirms that while Infrastructure is the strongest performance dimension across all districts, there is a clear gap in Safety and Facilities, especially in rural and low-performing districts. The bottom-performing schools are the most vulnerable to safety issues and lack of adequate facilities, signaling a need for targeted interventions to improve both security and educational resources.

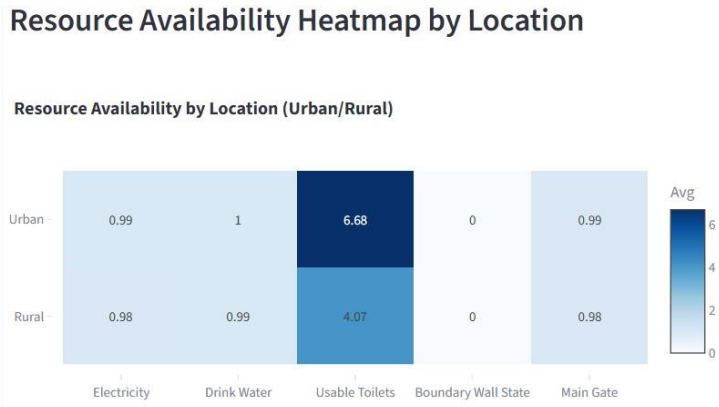
PCA for Facility Features:



The PCA for Facility Features scatter plot reduces multiple facility indicators into two principal components (PC1 and PC2), showing how schools cluster based on resource availability. Primary schools (dark blue) are spread across the plot, generally closer to the lower end, indicating lower facility availability. Secondary schools (red), High schools (light blue), and Higher Secondary (green) schools tend to cluster toward the right side of the plot, where higher PC1 values reflect better access to essential resources like classrooms, toilets, and computers. Mosque schools (orange) appear scattered, indicating varied access to facilities.

The size of the bubbles, reflecting enrollment, shows that larger schools are more likely to be on the right side, suggesting that larger schools tend to have better facilities. This visualization clearly highlights the facility disparity between school types, with higher-level schools (High, Higher Secondary) generally having better resources, while Primary schools and Mosque schools often fall behind. This clustering also emphasizes the resource-poor status of many rural and smaller schools, which require targeted infrastructure investments.

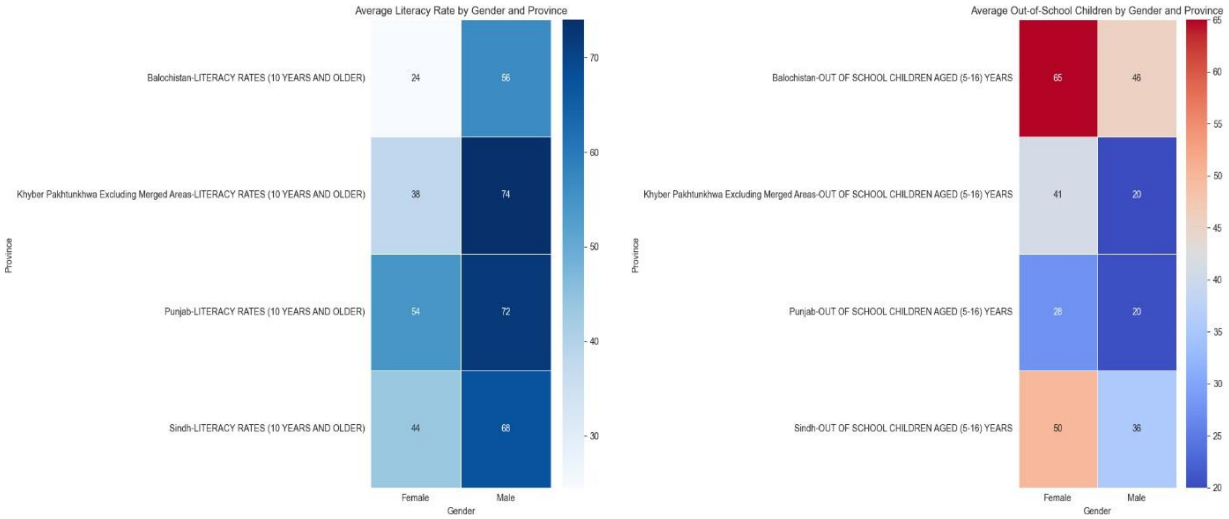
Resource Availability Heatmap (Rural vs Urban):



The Resource Availability Heatmap visually compares the availability of key resources like electricity, drinking water, and usable toilets in urban and rural schools. This heatmap shows that urban schools are consistently better equipped, with almost universal access to electricity and water, as indicated by the high values close to 1. In contrast, rural schools show significantly lower values, particularly for usable toilets, which average around 4.07, compared to the 6.68 in urban schools.

This apparent contrast reinforces the infrastructure gap between urban and rural schools. Rural schools, despite having access to basic utilities like electricity and water, still face critical shortages in sanitation facilities. The heatmap highlights the need for targeted investments in rural school infrastructure, particularly in sanitation and toilets, to ensure that these schools can meet the minimum standards required for a safe and effective learning environment.

Gender Breakdown of Literacy and Out of School Children:



The Gender Breakdown of Literacy Rates and Out-of-School Children visualizations highlight gender-based disparities in education across different provinces. Balochistan stands out with low female literacy rates (24%), vividly contrasting with male literacy rates (56%). This significant gap highlights the challenges faced by females in accessing education in this region which is a reflection of the culture of the entire country. The out-of-school children data further emphasizes this inequality, particularly in rural areas, where female children are more likely to be out of school compared to their male counterparts.

The gender disparities are not as severe in Punjab, where the gap between male and female literacy is narrower. However, the out-of-school children rate remains a concern, particularly for girls in Balochistan and Khyber Pakhtunkhwa. These visualizations clearly demonstrate that female education is disproportionately impacted by regional and gender-specific challenges and point to the need for targeted interventions to improve female access to education in underperforming provinces.

Ethical Considerations:

1) **Avoiding misleading representations:**
   - **Data Integrity:** All data used in visualisations was directly derived from the cleaned datasets, with appropriate transformations and calculations. These calculations were applied carefully to maintain the statistical integrity.
   - **Axis Scaling:** All axes in the charts were properly scaled to avoid data truncation or inconsistent scaling, ensuring that no information was hidden or exaggerated. For example, in bar charts, axis values were adjusted to start at zero to avoid skewing the visual representation of disparities in performance.
   - **Consistent Colour Coding:** The same colors were consistently used across all charts for categories (for e.g., Infrastructure, Safety and Facilities Performance Scores) to avoid confusion. Along with that, in heatmaps, color intensity was mapped clearly to data values with a color legend for accurate interpretation.
   - **Clear Legends and Labels:** Each chart includes easy to understand legends, titles, and axis labels to clarify what is being presented. This ensures that the audience can interpret the data accurately without ambiguity.

2) **Accessibility Features:**
   - **Colour-blind friendly palettes:** Where applicable, colour-blind friendly palettes were chosen, especially for heatmaps and categorical charts. For example, in heatmaps and bar charts, palettes like "Blues", "Coolwarm", and "Viridis" were used, as they are known to be easily distinguishable by individuals with various types of color blindness. We also ensured that contrast was sufficient between adjacent colors to provide clarity for viewers with limited color perception.

- **Readable Labels:** All charts are designed with legible font sizes for titles, axis labels, and legends. For detailed visualizations (like PCA scatter plots and bar charts), font sizes were carefully chosen to ensure clarity, even when visualized on smaller screens or projectors.
- **Text Annotations:** Important data points were made accessible for annotation directly on the charts suing hover tooltip (e.g., outlier points in scatter plots or values on heatmaps) to make the data easy to interpret for people with visual impairments or for those viewing in non-ideal conditions (such as dimly lit rooms).

3) Ethical Concerns:

- **Privacy and Sensitive Data:** All data used in this project is publicly available, such as educational indicators and performance scores from government reports and publicly available datasets. No personal or private information about students, teachers, or schools was included in the analysis. This anonymization of data ensures that no identifiable personal information was exposed.
- **Bias in Data Representation:** While the data comes from publicly available sources, no regional disparities (e.g., Balochistan's lower literacy rates or rural Punjab's resource shortages) are real and are not exaggerated in this project. The visualizations aim to highlight educational inequalities and emphasize areas where targeted interventions are needed, and not to create bias or stigmatize any region or group.
- **Fairness and Accuracy:** Since, this project aims to highlight the educational gaps so that effective policy action can be taken, rather than placing blame on specific groups or regions, gender disparities (e.g., female literacy rates in Balochistan, etc) were presented without reinforcing harmful stereotypes.
- **Regional Focus:** By including the KPK and Punjab datasets, the project is not favoring any province but is rather shedding light on the overall condition of Pakistan by using any publicly available dataset, where education systems face the greatest challenges.

Narrative:

Introduction to the Educational Disparities:

The datasets explored in this project highlight significant regional differences and gender and infrastructural imbalances in the educational sector of Pakistan, with a focus on Punjab and KPK. Through various visualizations, we have uncovered not only the gaps in infrastructure**,** safety, and facilities but also the huge differences between urban and rural schools. While some regions like Lahore perform exceptionally well, others, particularly in Balochistan and KPK, face persistent underinvestment in basic education infrastructure.

Key Insights from the Data:

Infrastructure and Safety Shortcomings

The District-wise Average Scores Comparison and Performance Dimensions visualizations clearly show that while infrastructure is strong in urban districts like Lahore, rural districts still suffer from poor infrastructure and low safety scores. This urban-rural divide stresses on the need for targeted investments in rural educational facilities, with a particular focus on sanitation, boundary walls, and safer classrooms.

Gender Gaps in Education

Another key insight gained, emerges from the gender breakdown of literacy rates and out-of-school children. The female literacy rate in Balochistan (24%) and Khyber Pakhtunkhwa is considerably lower than the male literacy rate, pointing to gender-specific barriers to education. Additionally, backing it up, the statistics for out-of-school children represent that majority are overwhelmingly female, especially in rural regions, where access to education is limited for girls.

Resource Availability

The Resource Availability Heatmap further emphasizes that urban schools in Punjab have better access to basic utilities like electricity and drinking water. In contrast, rural schools face significant shortages in sanitation and toilet facilities, which not only affect health but also hinder the overall learning environment.

Political Influence on School Performance

The Political Party Performance Comparison reveals that certain political parties, particularly PMLN in Punjab, have maintained influence in top-performing districts, suggesting that political stability and governance play a role in securing better educational resources and facilities. However, PTI has fewer top rankings, and Independent candidates show mixed results. This suggests that political orientation should be considered when evaluating the effectiveness of educational policies.

Actionable insights and proposed solutions:

The data shows that rural districts are consistently underperforming in infrastructure and safety. An actionable outcome can be to prioritise infrastructure investments in these districts, especially in schools with unsafe classrooms and insufficient sanitation and other technical facilities. Governments should allocate funds to build or renovate classrooms, install boundary walls, and ensure basic sanitation in schools.

The PCA analysis reveals a clear resource divide between urban and rural schools. Punjab's urban schools show strong performance due to better facilities, but rural districts struggle. One key recommendation is to implement a national program aimed at providing computer labs, libraries, and science labs to rural schools. This could be done through public-private partnerships, leveraging technology to create digital learning hubs in underserved areas.

This increase in these educational resources will attract more students to school and help them gain better knowledge, experience, and awareness of the rapid technological advancement in the world, to improve the literacy rate and secure a better future of the country. They should also start initiatives of allocating more teaching staff to the school, to decrease the burden on the existing staff and improving the quality of education received by the students.

The gender disparities highlighted in the literacy rate and out-of-school children heatmaps must be addressed with policy reforms focused on female education. The government should launch awareness campaigns to combat against cultural barriers to girls' education and invest in female-friendly infrastructure (e.g., separate toilets, female teachers). Programs to encourage female teacher recruitment and scholarships for girls should also be expanded in rural regions.

The Safety Score visualizations suggest that even high-performing schools suffer from security concerns. Therefore, a national school safety program should be introduced to ensure all schools, especially in rural and remote areas, are equipped with adequate security measures like boundary walls and gated entrances. Safety training for teachers and staff should also be provided.

By implementing these recommendations, we can take meaningful steps toward improving the educational outcomes for all students, regardless of gender or geography.