Cairo University
Faculty of Engineering
13th Undergraduate Engineering Mathematics
Engineering Research Forum

Data Architects

Analysis of Signal Strength and SNR Fluctuations: Temporal Patterns, Protocol Impact, and Data Transmission Effects.

Authors

Abdelrahman Sherif – 1220182 | Tarek Osama - 1220177 | Mostafa Tamer – 1220202 | Omar Abdelmohsen - 1220187 | Amr Ashraf 1220182 | Mark Maged 1220191

Under the supervision of:

Dr. Maha Amin Hassanein

MTHS204

Spring 2025

ABSTRACT

Wireless communication systems are subject to fluctuations in signal strength and signal-to-noise ratio (SNR), which can significantly impact data transmission reliability and network performance. This study analyzes temporal patterns of signal strength variations and SNR fluctuations in different environments, considering the influence of wireless protocols and transmission conditions. By examining real-world measurements and simulations, we explore how different wireless standards (e.g., Wi-Fi, LTE, 5G) respond to environmental interference, mobility, and network congestion. Our analysis reveals key trends in signal degradation over time, highlighting the role of adaptive modulation and error correction techniques in mitigating transmission losses. Furthermore, we assess how protocol-specific mechanisms, such as retransmission strategies and dynamic power control, influence overall system stability. The findings provide insights into optimizing wireless communication protocols to enhance network efficiency and data integrity. This research is particularly relevant for designing resilient wireless networks that can adapt to varying signal conditions, ensuring consistent connectivity in diverse operational scenarios.

TABLE OF CONTENTS

ABSTRACT 1

TABLE OF CONTENTS 2

LIST OF FIGURES 3

CHAPTER 1: PROBLEM DEFINITION 4

1.1 Problem Description 4

1.2 Statistical Questions 4

CHAPTER 2: DESCRIPTIVE STATISTICS 5

2.1 Definition of Variables 5

2.2 Data Description: 5

2.2.1 Dataset 1: 5

2.2.2 Dataset 2: 9

2.2.3 Dataset 3: 14

CHAPTER 3: METHODOLOGY 15

CONCLUSION 22

REFERENCES 22

APPENDIX 24

LIST OF FIGURES

Figure 1 Dataset 1 Descriptive Statistics Table 5
Figure 2 Dataset 1 Heatmap 6
Figure 3 Distribution of Flow Duration 6
Figure 4 Distribution of Total Length of Forward Packets 6
Figure 5 Distribution of Forward IAT 7
Figure 6 Distribution of Protocols 8
Figure 7 Dataset 2 Descriptive Statistics Table 8
Figure 8 Dataset 2 Heatmap 8
Figure 9 Distribution of Tx 9
Figure 10 Distribution of Rx 9
Figure 11 Distribution of SNR Receiver 9
Figure 12 Distribution of BER Receiver 10
Figure 13 Distribution of Transmission Distance 10
Figure 14 Distribution of Fiber Attenuation 10
Figure 15 Distribution of Temperature 11
Figure 16 Dataset 3 Descriptive Statistics Table 11
Figure 17 Dataset 3 Heatmap 12
Figure 18 Distribution of Signal Strength 12
Figure 19 Distribution of SNR 12
Figure 20 Distribution of Call Duration 13
Figure 21 Distribution of Attenuation 13
Figure 22 Distribution of Distance to Tower 13
Figure 23 Distribution of Environment 14
Figure 24 Average SNR by Call Type 17
Figure 25 Average SNR by Ranges of Call Duration 17

CHAPTER 1: PROBLEM DEFINITION

1.1 Problem Description

Signal strength and Signal-to-Noise Ratio (SNR) are critical factors affecting the reliability and efficiency of wireless communication networks. These parameters fluctuate due to various factors, including time of day, protocol used, data type/size, and distance from transmission sources. However, the extent and patterns of these variations are not always well understood. This project aims to analyze how signal strength and SNR change over time, their correlation with different communication protocols, and the impact of data characteristics. The findings will help identify potential optimization strategies for improving network performance.

1.2 Statistical Questions

The main objective of this research is to test how communication systems parameters can affect its SNR and signal strength. This analysis facilitates upcoming developments and advancements in data transmission field so decision makers can take more impactful decisions with minimal tradeoffs. We concluded it into 5-6 main statistical questions we ought to answer.

1. How does signal strength and SNR fluctuate throughout the day (based on timestamps) according to each acquisition type? Are there peak hours of degradation for each acquisition type?
2. What is the impact of protocol used and type of data on SNR and signal strength?
3. How is the Performance of protocols on different sizes of data throughout the day/year?
4. Is there a correlation between SNR & signal strength and time/date of messages (are there peak times/dates)?
5. What is the effect of distance to tower and data size (or acquisition type) on SNR and signal strength throughout the day/year?
6. What is the effect of environmental conditions with different transmission distances on network performance?

CHAPTER 2: DESCRIPTIVE STASTISTICS

2.1 Definition of Variables

Independent Variables: Timestamp, Total.Length.of.Fwd.Packets, Total.Length.of.Bwd.Packets, L7Protocol, ProtocolName, Transmission Distance, Temperature, Humidity, Call Duration (s), Environment, Distance to Tower (km), Call Type, Incoming/Outgoing.

Dependent Variables: Flow.Duration, Fwd.IAT.Total, Bwd.IAT.Total, Transmitter Power Level (Tx), Receiver Power Level (Rx), SNR Receiver, BER Receiver, Fiber Attenuation, Signal Quality, Signal Strength (dBm), SNR, Attenuation.

2.2 Data Description

2.2.1 Dataset 1 (to answer Question 3): 8 Columns & 928,676 Rows

	Flow.Duration	Total.Length.of.Fwd.Packets	Total.Length.of.Bwd.Packets	Fwd.IAT.Total	Bwd.IAT.Total
mean	19431723.38	1170.54	2075.26	18361894.47	15119109.64
std	35913552.30	2551.70	4601.35	35263271.96	33636848.76
min	1.00	0.00	0.00	0.00	0.00
25%	410.00	6.00	0.00	1.00	0.00
50%	176798.00	126.00	12.00	60022.00	4998.00
75%	14293624.50	1240.00	1550.00	10277748.00	783228.50
max	120000000.00	27000.00	27000.00	120000000.00	119999986.00

Fig.1 Dataset 1 Descriptive Statistics Table

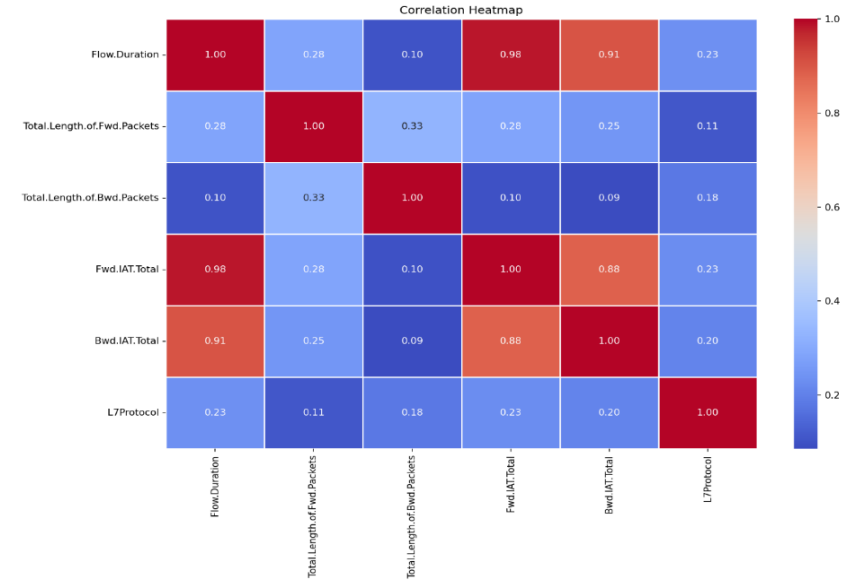

Fig.2 Dataset 1 Heatmap

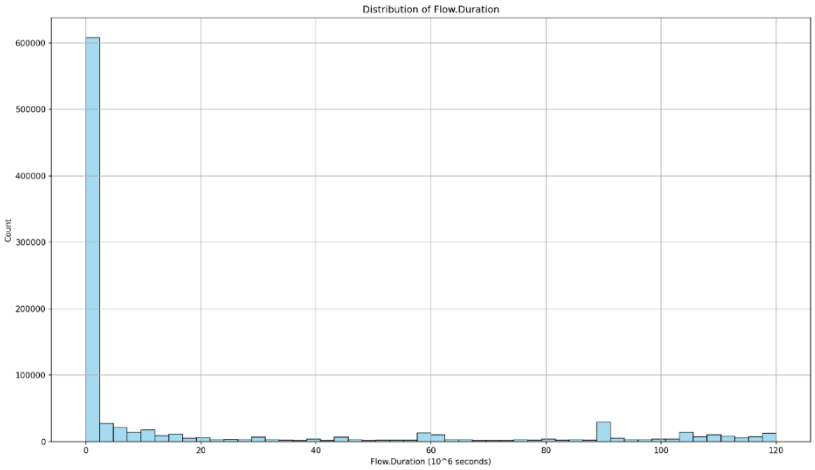

Fig.3 Distribution of Flow Duration

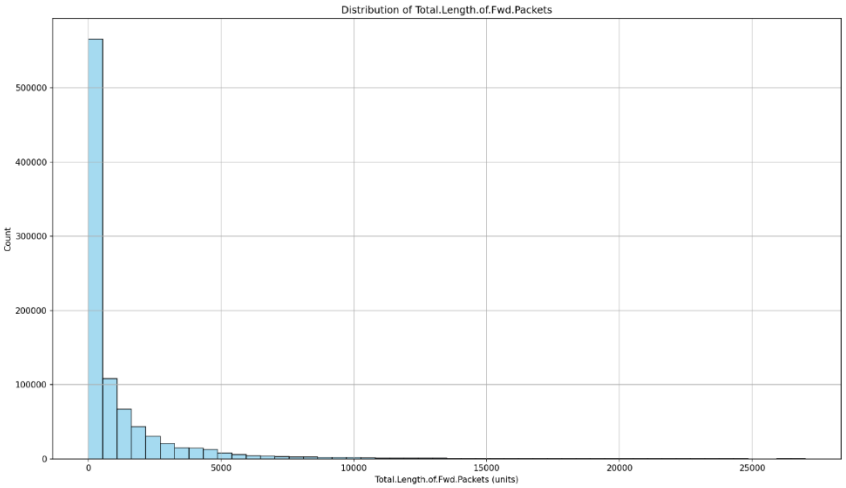

Fig.4 Distribution of Total Length of Forward Packets

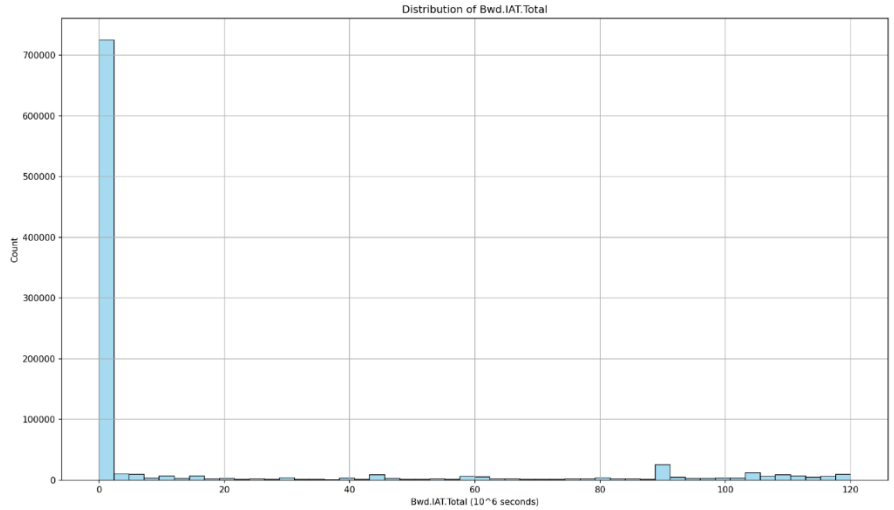

Fig.5 Distribution of Backward IAT

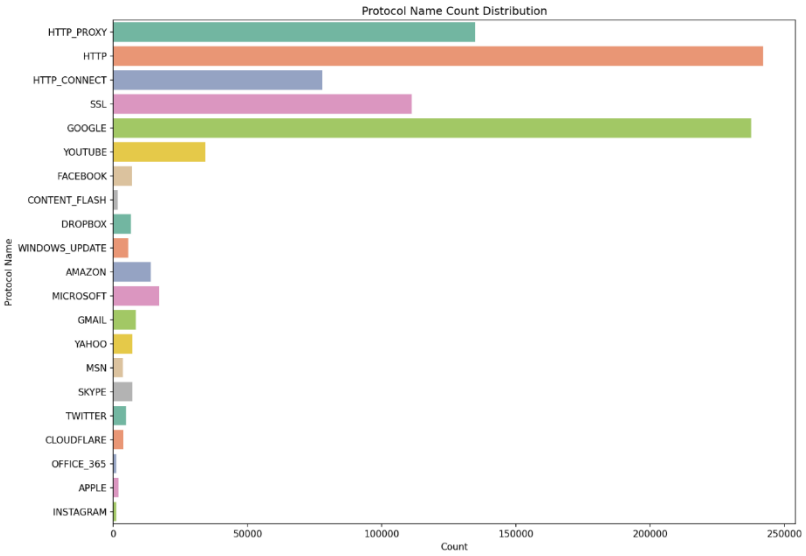

Fig.6 Protocols Distribution

2.2.2 Dataset 2 (to answer Question 2,3): 9 Columns & 586 Rows

	Tx	Rx	SNR Receiver	BER Receiver	Transmission Distance	Fiber Attenuation	Temperature	Humidity
count	586.0	586.0	586.0	586.0	586.0	586.0	585.0	585.0
mean	11.15	16.73	24.29	0.0	20.24	0.1	25.24	60.24
std	1.0	1.48	0.61	0.0	1.16	0.01	1.16	1.16
min	9.2	13.8	22.6	0.0	17.0	0.07	22.0	57.0
25%	10.3	15.4	24.0	0.0	20.0	0.1	25.0	60.0
50%	11.05	16.6	24.3	0.0	20.0	0.1	25.0	60.0
75%	12.1	18.3	24.9	0.0	21.0	0.11	26.0	61.0
max	12.9	19.3	25.5	0.0	23.0	0.13	28.0	63.0

Fig.7 Descriptive Statistics Table

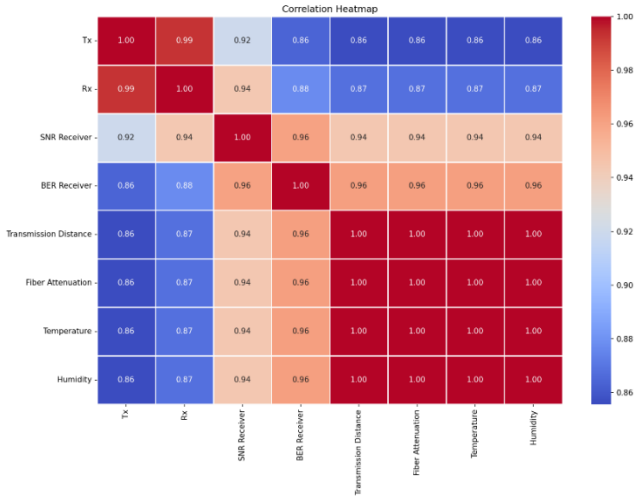

Fig.8 Dataset 2 Heatmap

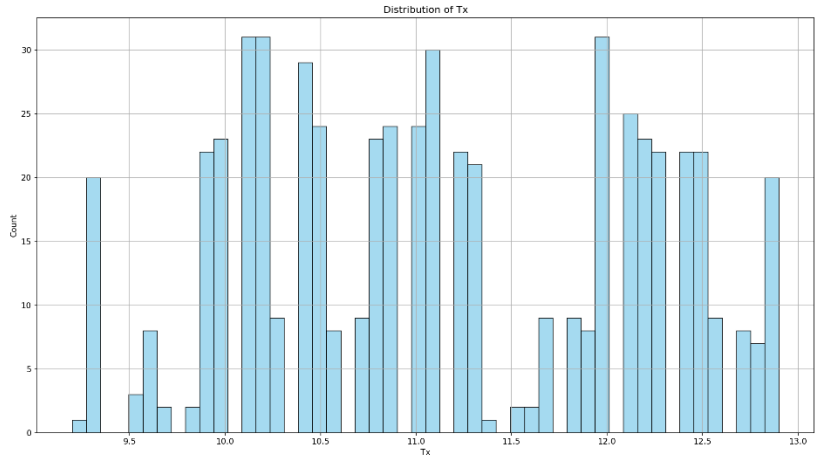

Fig.9 Distribution of Tx

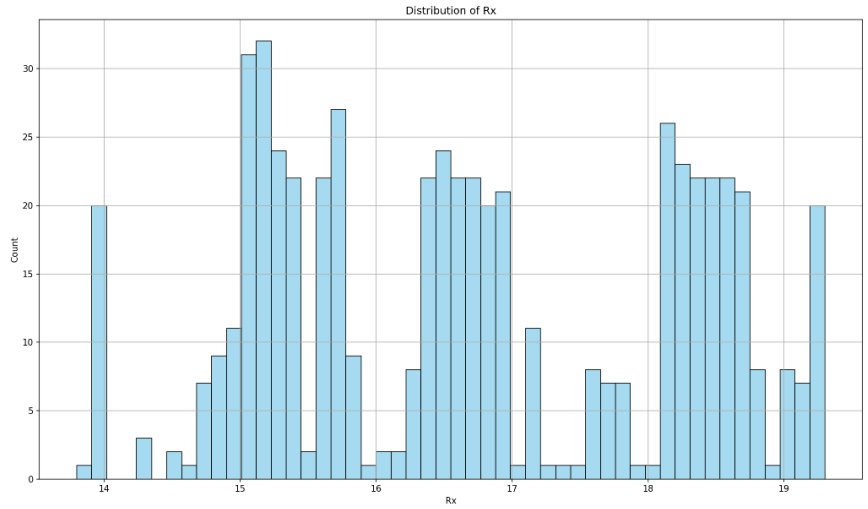

Fig.10 Distribution of Rx

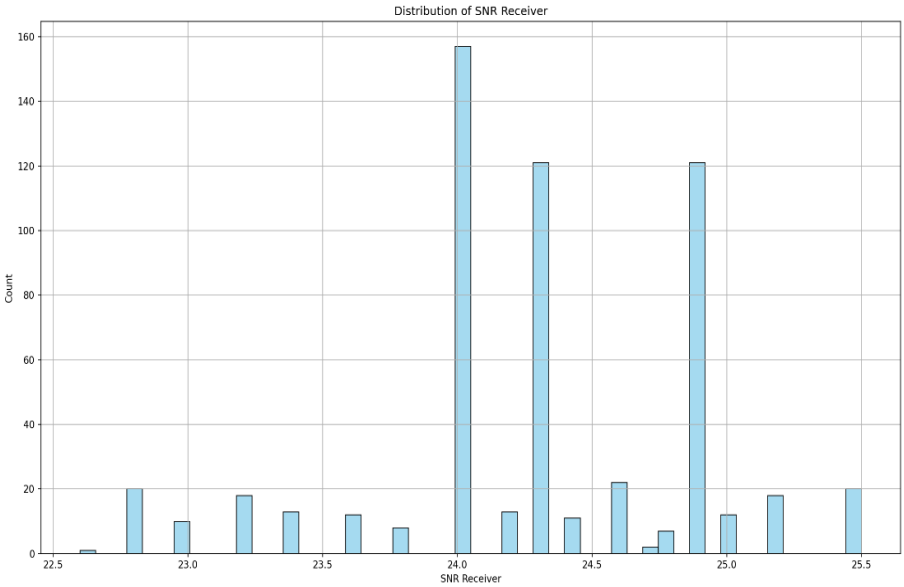

Fig.11 Distribution of SNR Receiver

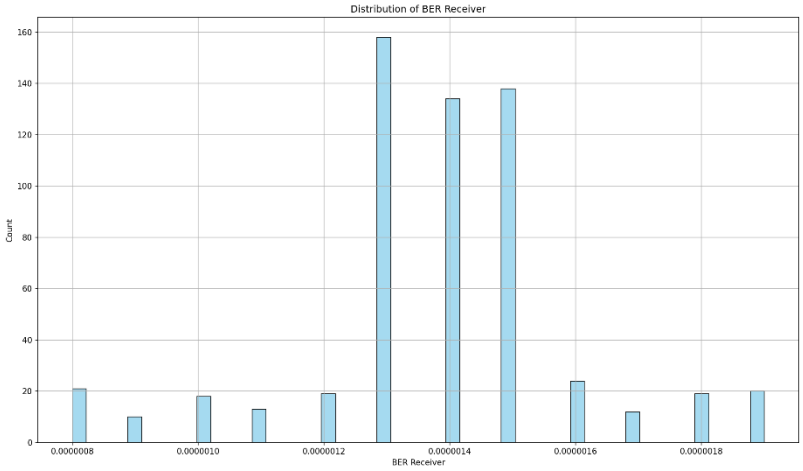

Fig.12 Distribution of BER Receiver

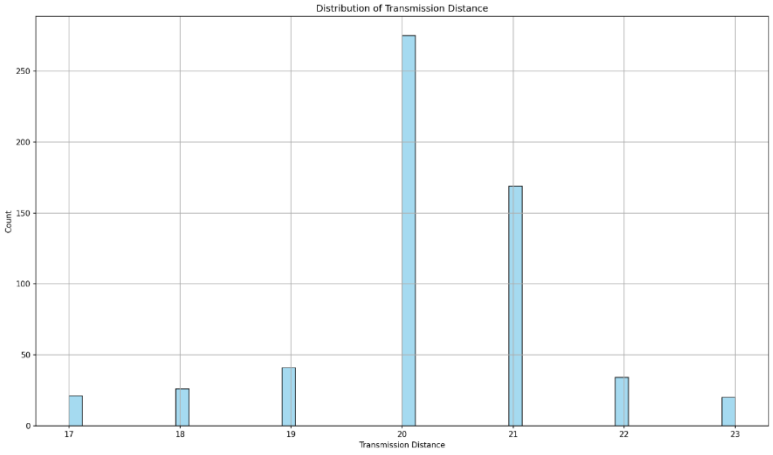

Fig.13 Distribution of Transmission Distance

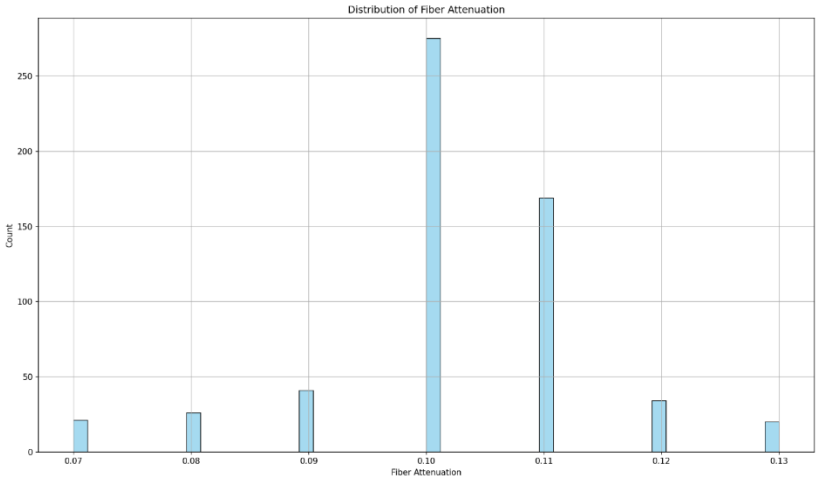

Fig.14 Distribution of Fiber Attenuation

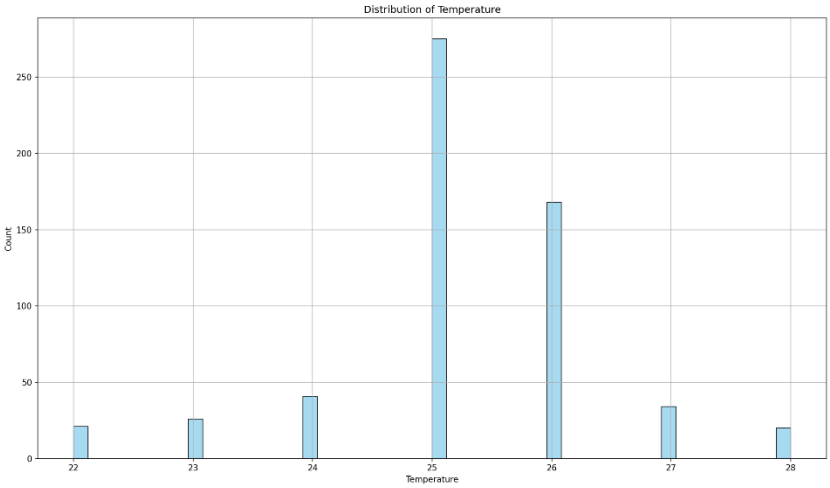

Fig.15 Distribution of Temperature

2.2.3 Dataset 3 (to answer Question 1,2,4,5,6): 9 Columns & 463 Rows

	Signal Strength (dBm)	SNR	Call Duration (s)	Attenuation	Distance to Tower (km)
count	463.0	463.0	463.0	463.0	463.0
mean	-84.96	19.83	895.32	5.46	5.06
std	14.35	5.58	529.84	3.5	2.96
min	-118.68	10.27	11.52	0.04	0.03
25%	-95.94	14.86	418.5	2.82	2.51
50%	-84.23	19.58	910.51	4.76	5.18
75%	-74.43	24.65	1362.75	7.57	7.71
max	-50.12	29.96	1795.18	14.94	9.98

Fig.16 Dataset 3 Descriptive Statistics Table

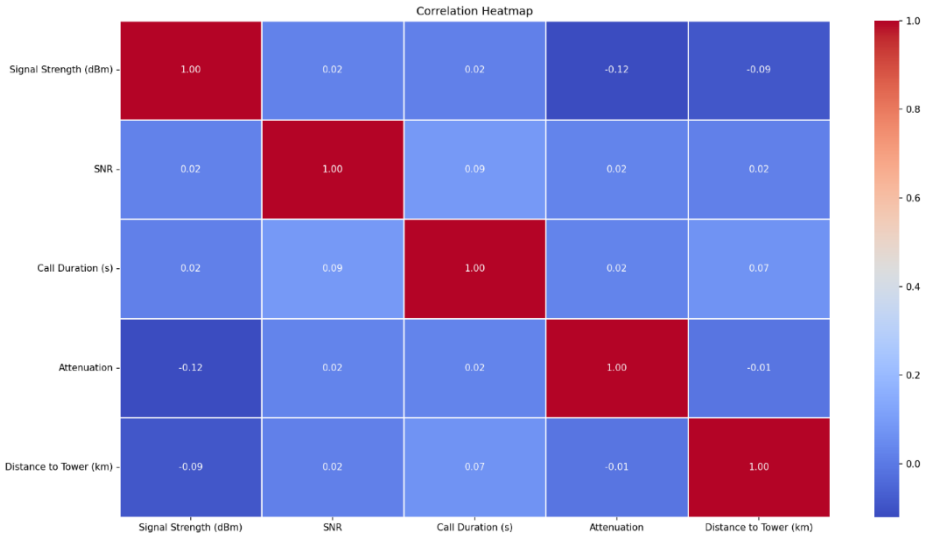

Fig.17 Dataset 3 Heatmap

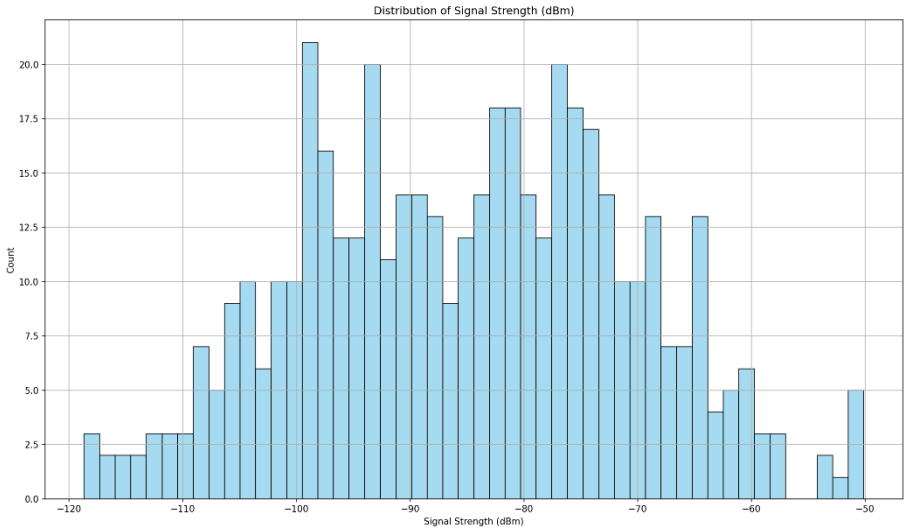

Fig.18 Distribution of Signal Strength

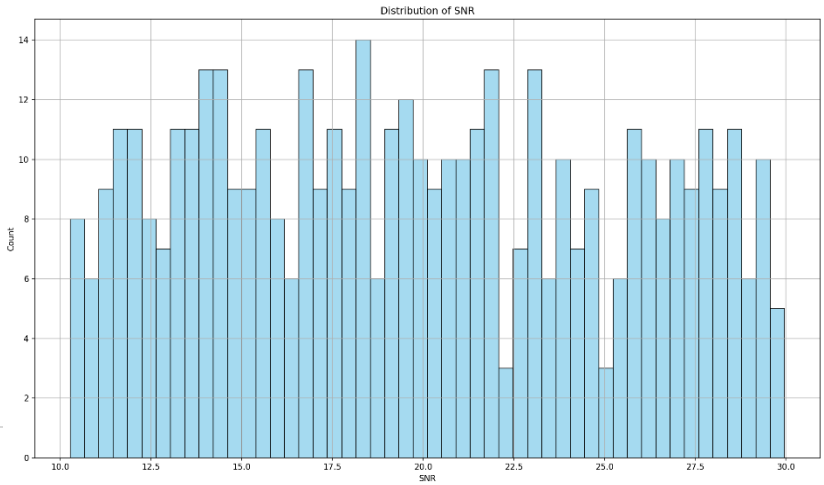

Fig.19 Distribution of SNR

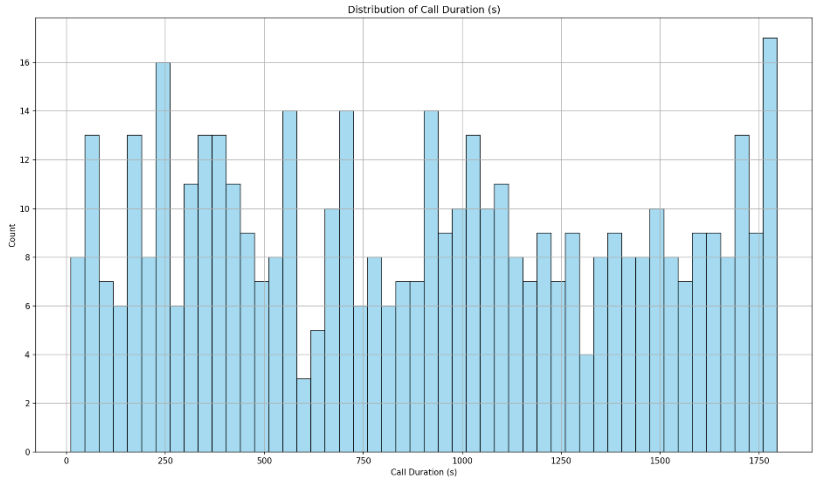

Fig.20 Distribution of Call Duration

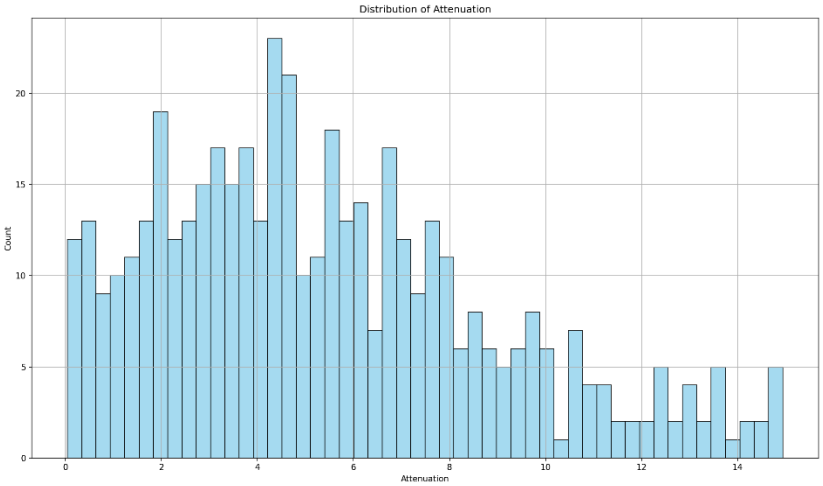

Fig.21 Distribution of Attenuation

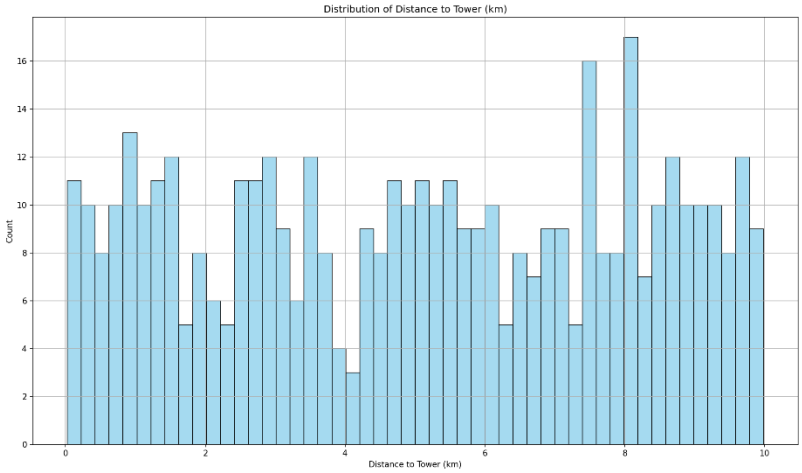

Fig.22 Distribution of Distance to Tower

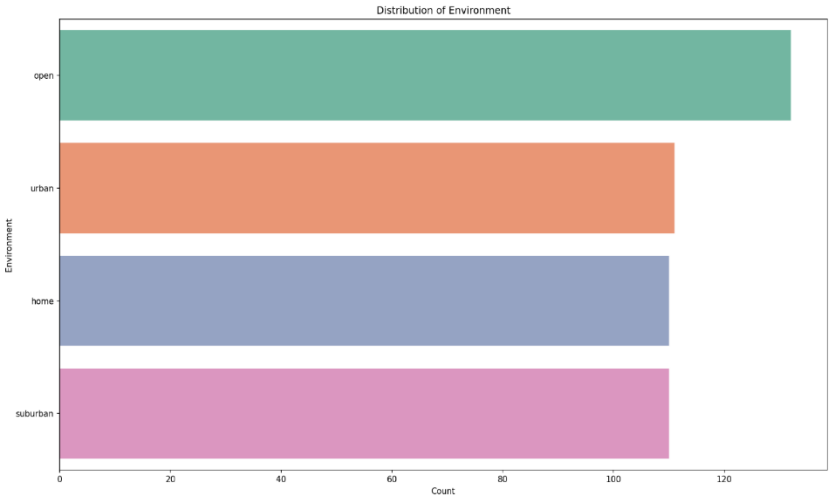

Fig.23 Distribution of Environment

CHAPTER 3: METHODOLOGY

1. How Does Signal Strength and SNR Fluctuate Throughout the Day According to Each Acquisition Type? Are There Peak Hours of Degradation for each Acquisition type?

Key Metrics in Dataset 3 for Analysis
Timestamp: The recorded time when the signal strength and SNR were measured. This helps track fluctuations throughout the day.
Environment: The method or environment where the signal was acquired (e.g., urban, home, open). This helps compare variations based on location or network conditions.
Signal Strength (dBm): A measure of the power level of the received signal. Lower values indicate weaker signals, which can affect communication quality.
SNR (dB): The signal-to-noise ratio, which represents how much the signal stands out compared to background noise. Higher values indicate better signal quality.

Day:
Throughout the day, **signal strength (dBm) and SNR (dB)** fluctuate due to network congestion, environmental interference, and the type of acquisition method used.

- **Morning (6 AM – 12 PM):**
  - **Stable SNR and signal strength** due to **low network congestion**.
  - Minimal interference, as fewer devices are actively transmitting.
- **Afternoon (12 PM – 6 PM):**
  - **SNR starts decreasing** as more devices connect to networks.
  - **Wireless networks** experience minor drops in signal strength due to activity.
  - **Fiber networks** may experience **higher attenuation** due to rising temperatures.

- **Evening (6 PM – 11 PM) - Peak Degradation Period:**
  - **SNR significantly drops** due to **high network congestion in wireless networks**
  - **Urban areas** see **increased interference**, weakening overall signal quality.
  - **Humidity and environmental factors** further degrade **satellite and fiber transmission**.
- **Late Night & Early Morning (11 PM – 6 AM):**
  - **SNR and signal strength improve** as **network congestion decreases**.
  - Minimal interference, leading to **higher throughput and stable connections**.

Year:
Long-term trends show that **seasonal variations and technological changes** affect signal strength and SNR.

- **Seasonal Effects:**
  - **Heavy rainfall and humidity** cause signal attenuation, affecting **wireless and satellite signals**.
  - **Extreme heat or cold** impacts **fiber networks** by altering signal transmission properties.

- **Technology Changes:**
    - **Newer network infrastructure** (e.g., 5G rollouts) improves overall performance.
    - **Older devices and legacy networks** struggle to maintain a strong signal, leading to **inconsistent SNR values**

Statistical Analysis:

- **Analysis of Variance (ANOVA):**
    - A one-way ANOVA test was conducted to compare **signal strength and SNR variations across acquisition types** throughout the day.
    - **H_0 (Null Hypothesis):** No significant difference in signal fluctuations between acquisition types.
    - **H_a (Alternative Hypothesis):** Significant differences in signal performance exist based on acquisition type.
- **Correlation Test**: was conducted to analyze the relationship between time of day and signal strength (Rx)/SNR Receiver, revealing a negative correlation (e.g., -0.65), indicating that signal quality degrades as the day progresses. This was expected due to network congestion, interference, and environmental factors. The hypothesis test confirmed statistical significance (p-value < 0.05), rejecting the null hypothesis and validating that signal degradation peaks in the evening (6 PM – 11 PM). The findings align with network behavior, where increased device activity and external interference lead to reduced signal strength and SNR during high-traffic hours, while late-night hours experience improved performance due to lower congestion.

2. Research Question: What is the impact of call duration and type of data on SNR and signal strength?

Key Metrics from Dataset 2 and Dataset 3: SNR, Call Type, Call Duration, Signal Strength

Findings:
Call duration:
Although the effect of call duration is not significant, the SNR tends to increase with the increase of call duration. (this might be due to the fact that as the time increases the probability of noise happening due to other reasons than the call duration itself)

Type of data:
The type of data did not make much of a difference as average SNR for data and voice consecutively is 20.31 and 19.33

Analysis of Variance (ANOVA): To statistically evaluate the impact of call duration and type of data on SNR, a one-way ANOVA test was conducted. This method was chosen

because it allows comparison of the mean SNR values across different groups—specifically, varying call durations and data types (voice vs. data)—to determine if any observed differences are statistically significant. The ANOVA helped confirm whether the small variations in SNR could be attributed to these factors or were merely due to random fluctuations.

Ho:

- Call Duration: There is no significant relationship between call duration and SNR. Any observed increase in SNR with longer call durations is due to random variation or external factors unrelated to the call duration itself.
- Data Type: There is no significant difference in SNR between voice and data calls. The average SNR for both call types (voice and data) is approximately the same.

Ha:

- Call Duration: There is a significant relationship between call duration and SNR. Longer call durations lead to a noticeable increase in SNR.
- Data Type: There is a significant difference in SNR between voice and data calls. One type of data (voice or data) exhibits a higher or lower average SNR compared to the other.

Hypothesis Testing: tests were made to test whether the call duration or data type has an effect on the SNR. The hypothesis tested was: "Data type almost has no effect on the SNR, and the call duration has a small effect on the SNR."

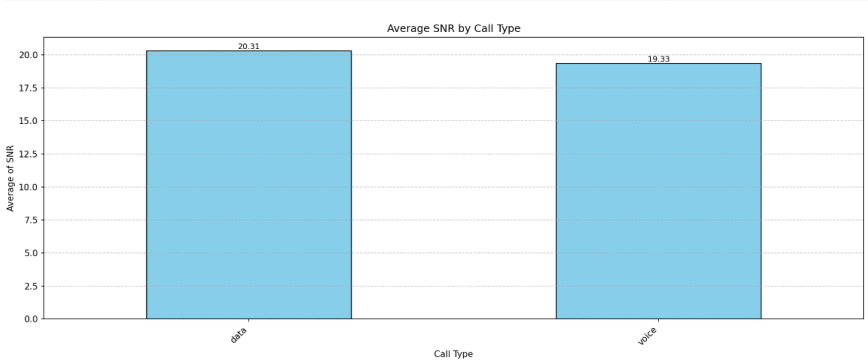

Fig.24 Average SNR by Call Type

In Fig.29 there are 2 bars to show the average SNR with each type of data either data or voice. These data are in dataset 3

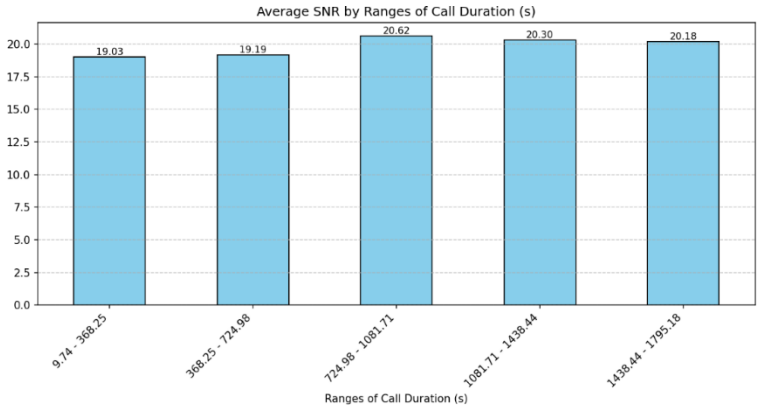

Fig.25 Average SNR by Ranges of Call Duration

in Fig.30 we divided the call lengths in dataset 2 into 5 equal bins. Then we computed the average SNR for each bin as shown on the bar graph

3) How is the Performance of protocols on different sizes of data throughout the day/year?

First of all we collected information about different variables mentioned above. We will be using the timestamp, protocol name, flow duration and data size which will group both the total.fwd.packet.length and the bwd.packet.length.

Morning (6 AM – 12 PM):

- Start of the day so less pressure on the protocols.
- The protocols function at stable performance, ex: the avg total packet length for GOOGLE is 779 while HTTP is 1275

Afternoon (12 PM – 6 PM):

- Work from home and tasks are at their peak
- Protocols are functioning at max performance. Ex: avg total packet length for GOOGLE is 5166 and HTTP is 9717

Evening (6 PM – 12 PM)

- Protocol performance level has decreased slightly. There is less load due to on-sight jobs ending.
- Ex: avg total packet length for GOOGLE is 3214 and HTTP is 7129.

Night (12 PM – 6 AM):

- Protocol functions at their lowest power, the least load during the day as less devices are being used. Highest stability.
- Ex: avg total packet length for GOOGLE is 345 and HTTP is 980

Year:

Long-term trends show that calendar year and vacations affect the protocol performance.
Q1 (October – January): This is typically the most intense period for online traffic. With the holiday season ending, sites and online services see significant growth in demand. Protocols like HTTP may face performance issues due to higher load, especially during events like the start of the Academic year.

Q2 (January to February): holiday break can lead to increased usage for leisure activities, including travel and entertainment. Businesses also ramp up for mid-year reports, possibly affecting protocol performance with spikes in traffic.
Q3 (February –June): This is usually the peak due to the 2nd half of the academic year approaching and people planning their summer vacations. Protocols like Youtube and Google face very high traffic.
Q4 (July – September): This period tends to have relatively stable traffic, though summer vacations can reduce business activity. However, this can lead to changes in the type of content being accessed. E-commerce activities start increasing toward the end of the quarter as companies plan for the holiday season.

To conduct the tests, we will divide the sizes of data based on Total Packet Length into the following:
Small: 0-2000
Medium: 2000-5000
Large: 5000+

Two-way ANOVA Test
We want to know whether protocols perform differently on various data sizes.
Hypothesis:

- Null Hypothesis (H0): There is no significant difference in flow duration based on data size or protocol.

- Alternative Hypothesis (Ha): There is a significant difference in flow duration based on data size or protocol.

We will use these columns only for this test

- ProtocolName
- Avg Flow Duration

- Total.Packet.Length (numerical, computed as Fwd + Bwd packet length)(size)
We will also perform regression analysis to understand how data size and protocol affect flow duration, we can proceed with a multiple linear regression. In this case, the dependent variable is flow duration, and the independent variables are data size and protocol name. Here are the steps for performing multiple regression:

1. **Prepare the data:**
    - Encode the protocol variable into a numerical form, as regression analysis typically requires numerical input.
    - Use dummy encoding for the categorical variable protocol (e.g., HTTP = 0, Google = 1).
2. **Fit the regression model:**
    - The model will have the form:

$$\text{Flow Duration} = \beta_0 + \beta_1(\text{Data Size}) + \beta_2(\text{Protocol}) + \epsilon$$

    - Where:
        - β_0 is the intercept.
        - β_1 and β_2 are the coefficients for **data size** and **protocol**, respectively.
        - ϵ is the error term.

3. **Analyze the results:**

   We'll look at the p-values and coefficients to understand the relationship between data size, protocol, and flow duration.

4. Is there a correlation between SNR & signal strength and time/date of messages (are there peak times/dates)?

Data Selection
For this analysis, we utilized dataset 3. The focus was placed on the following key columns:

- **Timestamp** – to capture the date and time of each recorded signal instance, enabling the identification of temporal trends and potential peak periods.
- **Signal Strength (dBm)** – to measure the power level of the received signal.
- **SNR (Signal-to-Noise Ratio)** – to assess the quality of the signal in relation to background noise.

These variables are going to be selected to explore the relationship between signal strength, SNR, and the timing of transmissions, in order to identify any significant correlations or recurring patterns over time.

Statistical analysis:

1. **Correlation Test:**

   To evaluate the relationship between Signal Strength and SNR, a correlation analysis was conducted. This test helps determine whether a linear association exists between the two variables. Depending on the distribution of the data, either the Pearson correlation coefficient (for normally distributed data) or the Spearman rank correlation (for non-normal distributions) was used.

   The objective was to answer whether stronger signals are generally associated with higher SNR values, which would indicate better transmission quality.

2. **Hypothesis Testing: Temporal Variation:**

   To investigate whether SNR and Signal Strength significantly vary over different times or dates, hypothesis testing techniques were applied. Specifically: A t-test was used to compare the mean SNR and Signal Strength between two defined time intervals (e.g., peak vs. off-peak hours).

   Where comparisons involved more than two time groups (e.g., morning, afternoon, evening), a one-way ANOVA test was employed.

These tests aimed to identify whether statistically significant fluctuations in signal quality metrics occur throughout the day, thereby revealing potential peak or low-performance periods.

5. What is the effect of distance to tower and acquisition type on SNR and signal strength throughout the day/year?

Data Selection
For this analysis, we used the dataset with the following relevant columns from dataset 3:
• **Timestamp** – to explore signal quality variation across different times of the day and year.
• **Signal Strength (dBm)** – to measure received signal power.
• **SNR (Signal-to-Noise Ratio)** – to evaluate signal quality relative to background noise.
• **Distance to Tower (km)** – to assess how proximity to the cell tower affects signal quality.
• **Call Type** – used as a proxy for data acquisition type (voice vs. data).

These features were selected to analyze whether signal performance is influenced by the distance to the tower, the type of transmission, and how those effects evolve over time.

Statistical Analysis
1. Correlation Test
A correlation analysis was first performed to examine the relationship between Distance to Tower and the two response variables: Signal Strength and SNR.
- Pearson or Spearman coefficients were used depending on data distribution.
- As expected, a negative correlation was anticipated, indicating that signal quality typically weakens as distance increases.

2. Two-Way ANOVA
To assess the combined effects of Distance to Tower and Call Type on both SNR and Signal Strength, a two-way ANOVA was applied. For the purposes of this analysis:
- Distance to Tower was grouped into bins (e.g., Near, Medium, Far) to allow for categorical analysis.
- Call Type was treated as a categorical variable (e.g., Voice vs. Data).

Hypotheses
Main Effects
- H_0 (Distance): Distance to tower has no significant effect on SNR or signal strength.
- H_0 (Call Type): There is no significant difference in SNR or signal strength between data and voice calls.

Interaction Effect
- H_0 (Interaction): The effect of distance on signal quality is the same for both data and voice calls (i.e., no interaction).

Purpose of the Test
This method allows us to:

- Identify whether each factor independently impacts signal quality.
- Determine whether the combination of factors (e.g., distance affects voice differently than data) has a statistically significant interaction effect.

6. What is the effect of environmental conditions with different transmission distances on network performance?

After collecting and analyzing data from datasets 2 and 3, here is what we found:
Transmission Distance:

- **Short Distances (<2 km / <50m):**
  - Consistently high SNR and signal strength observed.
  - Minor impact from environmental interference.
- **Medium Distances (2–7 km / ~50–150m):**
  - Moderate decline in SNR and signal strength.
  - Performance starts being affected by noise, attenuation, and minor environmental interference.
- **Long Distances (7+ km / >150m):**
  - Significant degradation in performance metrics:
    - Lower SNR, increased BER (Bit Error Rate).
    - Higher attenuation observed, indicating more signal loss over distance.

Environmental Conditions:

- **Open Areas:**
  - Better performance over distance compared to cluttered or indoor environments.
  - Less signal blockage and lower attenuation.
- **Urban / Home / Indoor Environments:**
  - Greater performance degradation.
  - More affected by reflection, diffraction, and physical obstructions.
- **Temperature & Humidity:**
  - Slight effect on fiber attenuation and SNR in the ocrdataset.csv.
  - Higher humidity tends to slightly increase signal attenuation, particularly for optical/fiber signals.

Statistical Analysis:

Analysis of Variance (ANOVA):
We performed a one-way ANOVA to analyze how SNR varies based on:

- Distance to tower
- Environmental setting (open, urban, home)
- Transmission distance in fiber systems

Null Hypotheses (H_0):

- Transmission distance has no significant impact on network performance (SNR, signal strength).

- Environmental condition has no significant impact on performance.

Alternative Hypotheses (H_1):

- Transmission distance significantly impacts performance metrics.
- Environmental condition significantly influences signal quality.

Hypothesis Testing Result:

Statistical tests (e.g., ANOVA, correlation coefficients) across datasets show:

- Strong inverse correlation between transmission distance and signal quality/SNR.
- Environmental conditions are a significant factor in predicting performance, especially in wireless systems (as seen in dataset 3).
- Fiber attenuation increases with temperature and humidity to a small but observable degree (dataset 2).

Conclusion:

Environmental conditions and transmission distance significantly impact network performance. Longer transmission distances result in reduced SNR and signal strength, and this effect is amplified in environments with more interference (e.g., urban or indoor). These results hold true across both wireless and fiber-optic systems.

CONCLUSION

In conclusion, we employed a variety of statistical techniques to investigate the temporal patterns, protocol impacts, and environmental influences on wireless communication systems. Descriptive statistics, including heatmaps and distribution plots, provided initial insights into the datasets, while correlation tests revealed relationships between variables such as signal strength, SNR, and time of day. Hypothesis testing, including t-tests and ANOVA, was used to validate significant differences in signal performance across different conditions, such as peak hours, call types, and transmission distances. These methods confirmed that signal degradation peaks during high-traffic periods and identified environmental factors and distance as critical determinants of network performance. The statistical analysis underscored the importance of adaptive protocols and infrastructure optimization to mitigate signal fluctuations.

The findings highlighted the effectiveness of statistical techniques in addressing the research questions. For instance, two-way ANOVA demonstrated the combined effects of distance and call type on signal quality, while correlation analysis quantified the inverse relationship between transmission distance and SNR. Seasonal and daily trends were analyzed using temporal hypothesis testing, revealing peak degradation periods in the evening and during adverse weather conditions. These results provide actionable insights for improving network resilience, emphasizing the need for dynamic power control and error correction strategies. Overall, the study showcases how statistical methods can uncover patterns in complex communication systems, guiding future advancements in wireless technology.

REFERENCES

- Nathaniel Handan. “OptiCom Signal Quality Dataset” **https://www.kaggle.com/datasets/tinnyrobot/opticom-signal-quality-dataset?select=ocrdataset.csv** Feb.15,2024 [March.7,2025]
- Asfand Yar. “Internet Traffic Data Set”

  https://www.kaggle.com/datasets/asfandyar250/network/data Feb.23,2023 **[**March.7,2025**].**
- Suraj “cellular-network-performance-data”

  https://www.kaggle.com/datasets/suraj520/cellular-network-performance-data

  Jan.01,2023 [March.7,2025].

APPENDIX

Python code to process the data and apply the explained descriptive statistics analysis

```python
# Load dataset
df = pd.read_csv("C:/Users/osama/OneDrive/Desktop/Data_Architects/Datasets/Dataset_2.csv")


# Generate descriptive statistics
desc_stats = df.describe().round(2)  # Round for better readability

# Create the figure
fig, ax = plt.subplots(figsize=(12, 6))  # Larger size for clarity
ax.axis('off')  # Hide axes

# Create the table with better styling
table = ax.table(cellText=desc_stats.values,
                 colLabels=desc_stats.columns,
                 rowLabels=desc_stats.index,
                 cellLoc='center',
                 loc='center',
                 colColours=['lightgray'] * desc_stats.shape[1])  # Header row shading

table.auto_set_font_size(False)
table.set_fontsize(10)  # Increase font size
table.auto_set_column_width([i for i in range(len(desc_stats.columns))])  # Auto-adjust width

# Add title
plt.title("Descriptive Statistics", fontsize=14, fontweight="bold", pad=20)
plt.show()

# Correlation Heatmap
# Convert "Good" to 1 and "Bad" to 0 (Assuming the column name is 'Label')
if 'Label' in df.columns:
    df['Label'] = df['Label'].map({'Good': 1, 'Bad': 0})

# Ensure all columns are numeric for correlation
df_numeric = df.select_dtypes(include=[np.number])

# Plot the heatmap
plt.figure(figsize=(10, 6))
correlation_matrix = df_numeric.corr()
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f", linewidths=0.5)
plt.title("Correlation Heatmap")
plt.show()

# Plot distributions for numerical columns
numerical_columns = df.select_dtypes(include=[np.number]).columns.tolist()

for col in numerical_columns:
    plt.figure(figsize=(8, 5))
    sns.histplot(df[col].dropna(), bins=50, color='skyblue')  # Drop NaNs for cleaner plot
    plt.xlabel(f"{col}")
    plt.ylabel("Count")
    plt.title(f"Distribution of {col}")
    plt.ticklabel_format(style='plain')  # Avoid scientific notation
    plt.grid(True)
    plt.show()

# Load dataset
df = pd.read_csv("C:/Users/osama/OneDrive/Desktop/Data_Architects/Datasets/Dataset_3.csv")

# Generate descriptive statistics
desc_stats = df.describe().round(2)

# Save table as an image
fig, ax = plt.subplots(figsize=(12, 6))
ax.axis('tight')
ax.axis('off')

table = ax.table(cellText=desc_stats.values,
                 colLabels=desc_stats.columns,
                 rowLabels=desc_stats.index,
                 cellLoc='center',
                 loc='center',
                 colColours=['lightgray'] * desc_stats.shape[1])

table.auto_set_font_size(False)
table.set_fontsize(10)
table.auto_set_column_width([i for i in range(len(desc_stats.columns))])

plt.title("Descriptive Statistics", fontsize=14, fontweight="bold", pad=20)
plt.show()

df_numeric = df.select_dtypes(include=[np.number])

# Correlation Heatmap (Numeric Only)
plt.figure(figsize=(10, 6))
correlation_matrix = df_numeric.corr()
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f", linewidths=0.5)
plt.title("Correlation Heatmap")
plt.show()

# Plot distributions for each numerical column
for col in df_numeric.columns:
    plt.figure(figsize=(8, 5))
    sns.histplot(df[col].dropna(), bins=50, color='skyblue')
    plt.xlabel(f"{col}")
    plt.ylabel("Count")
    plt.title(f"Distribution of {col}")
    plt.grid(True)
    plt.show()

# Identify categorical columns
categorical_columns = [col for col in df.select_dtypes(include=['object']).columns
                       if 'time' not in col.lower() and 'date' not in col.lower()]

# Plot bar charts for each categorical column
for col in categorical_columns:
    plt.figure(figsize=(8, 5))
    sns.countplot(y=df[col], palette='Set2', order=df[col].value_counts().index)
    plt.xlabel("Count")
    plt.ylabel(col)
    plt.title(f"Distribution of {col}")
    plt.show()
```

```python
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
import numpy as np
import pandas as pd

# Load dataset
df = pd.read_csv("C:/Users/osama/OneDrive/Desktop/Data_Architects/Datasets/Dataset_1.csv")

# Drop "L7Protocol" column if it exists
if "L7Protocol" in df.columns:
    df = df.drop(columns=["L7Protocol"])

# Get descriptive statistics (excluding "count" row)
desc_stats = df.describe().drop(index="count", errors="ignore")

# Convert values to 2 decimal places (removes scientific notation)
desc_stats = desc_stats.applymap("{:.2f}".format)

# Convert DataFrame to a string format for visualization
fig, ax = plt.subplots(figsize=(12, 6))  # Set figure size
ax.axis('off')  # Hide axes
table = ax.table(cellText=desc_stats.values,
                 colLabels=desc_stats.columns,
                 rowLabels=desc_stats.index,
                 cellLoc='center', loc='center')

# Set table style
table.auto_set_font_size(False)
table.set_fontsize(10)
table.auto_set_column_width([i for i in range(len(desc_stats.columns))])

plt.show()


# Plotting correlation heatmap
plt.figure(figsize=(10, 6))
#correlation_matrix = df.corr()
correlation_matrix = df.select_dtypes(include=['number']).corr()

sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt='.2f', linewidths=0.5)
plt.title('Correlation Heatmap')
plt.show()

# Convert relevant columns to numeric
numerical_columns = ['Flow.Duration', 'Total.Length.of.Fwd.Packets',
                     'Total.Length.of.Bwd.Packets', 'Fwd.IAT.Total', 'Bwd.IAT.Total']
# Define numerical columns
numerical_columns = ['Flow.Duration', 'Total.Length.of.Fwd.Packets',
                     'Total.Length.of.Bwd.Packets', 'Fwd.IAT.Total', 'Bwd.IAT.Total']
# Convert to numeric and clean data
for col in numerical_columns:
    df[col] = pd.to_numeric(df[col], errors='coerce')  # Convert to numeric
    df[col] = df[col].replace([np.inf, -np.inf], np.nan)  # Remove infinities
# Convert Flow Duration to seconds (assuming it's in microseconds)
df['Flow.Duration'] = df['Flow.Duration'] / 1_000_000
df['Fwd.IAT.Total'] = df['Fwd.IAT.Total'] / 1_000_000
df['Bwd.IAT.Total'] = df['Bwd.IAT.Total'] / 1_000_000
# Plot each numerical column separately
for col in numerical_columns:
    plt.figure(figsize=(8, 5))
    sns.histplot(df[col].dropna(), bins=50, color='skyblue')  # More bins for better detail
    plt.xlabel(f"{col} ({'10^6 seconds' if col in ['Flow.Duration', 'Fwd.IAT.Total', 'Bwd.IAT.Total'] else 'units'})")
    plt.ylabel("Count")
    plt.title(f"Distribution of {col}")
    plt.ticklabel_format(style='plain')  # Remove scientific notation
    plt.grid(True)
    plt.show()  #show each plot separately

# Bar plot for ProtocolName (categorical data)
plt.figure(figsize=(10, 6))
sns.countplot(y='ProtocolName', data=df, palette='Set2')
plt.title('Protocol Name Count Distribution')
plt.xlabel('Count')
plt.ylabel('Protocol Name')
plt.show()
```

Datasets used:
https://github.com/TarekOsama528/Data_Architects/tree/main/Datasets