Cairo University Faculty of Engineering

12th 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](#_bookmark0)

[Table of Contents 2](#_bookmark1)

[List of Figures 3](#_bookmark2)

[Chapter 1: Problem Definition 4](#_bookmark4)

* 1. [Problem Description 4](#_bookmark5)
  2. [Statistical Questions 4](#_bookmark7)

[Chapter 2: Descriptive Statistics 5](#_bookmark8)

* 1. [Definition of Variables 5](#_bookmark9)
  2. [Data Description: 5](#_bookmark15)

2.2.1 Dataset 1: 5

2.2.2 Dataset 2: 9

2.2.3 Dataset 3: 14

[Chapter 3: Methodology 10](#_bookmark18)

* 1. [Data Description 10](#_bookmark19)
     1. [Description of Images 11](#_bookmark22)
     2. [Descriptive Statistical Analyses 11](#_bookmark24)
  2. [Model Architecture and Training Procedure 14](#_bookmark27)
     1. [Layers Parameters 14](#_bookmark28)
     2. [Model Training 15](#_bookmark30)
  3. [Results 15](#_bookmark31)

[Conclusion 17](#_bookmark36)

[References 18](#_bookmark37)

[Appendix 19](#_bookmark38)

# List of Figures

[Figure 1 Dataset 1 Descriptive Statistics Table 5](#_bookmark6)

[Figure 2 Dataset 1 Heatmap 6](#_bookmark10)

[Figure 3 Distribution of Flow Duration 6](#_bookmark11)

[Figure 4 Distribution of Total Length of Forward Packets 7](#_bookmark14)

[Figure 5 Distribution of Total Length of Backward Packets 7](#_bookmark16)

[Figure 6 Distribution of Forward IAT 8](#_bookmark20)

[Figure 7 Distribution of Backward IAT 8](#_bookmark21)

[Figure 8 Distribution of Protocols 9](#_bookmark23)

[Figure 9 Dataset 2 Descriptive Statistics Table 9](#_bookmark25)

[Figure 10 Dataset 2 Heatmap 10](#_bookmark26)

[Figure 11 Distribution of Tx 10](#_bookmark29)

[Figure 12 Distribution of Rx 11](#_bookmark33)

[Figure 13 Distribution of SNR Receiver 11](#_bookmark34)

[Figure 14 Distribution of BER Receiver 12](#_bookmark35)

[Figure 15 Distribution of Transmission Distance 12](#_bookmark25)

[Figure 16 Distribution of Fiber Attenuation 13](#_bookmark26)

[Figure 17 Distribution of Temperature 13](#_bookmark29)

[Figure 18 Distribution of Humidity 14](#_bookmark33)

[Figure 19 Dataset 3 Descriptive Statistics Table 14](#_bookmark34)

[Figure 20 Dataset 3 Heatmap 15](#_bookmark35)

[Figure 21 Distribution of Signal Strength 15](#_bookmark34)

[Figure 22 Distribution of SNR 16](#_bookmark35)

[Figure 23 Distribution of Call Duration 16](#_bookmark25)

[Figure 24 Distribution of Attenuation 17](#_bookmark26)

[Figure 25 Distribution of Distance to Tower 17](#_bookmark29)

[Figure 26 Distribution of Environment 18](#_bookmark33)

[Figure 27 Distribution of Incoming/Outgoing 18](#_bookmark34)

[Figure 27 Distribution of Call Type 19](#_bookmark34)

# 

# Chapter 1: Problem Definition

### 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?
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 types/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

### A table with numbers and text AI-generated content may be incorrect.

Fig.1 Dataset 1 Descriptive Statistics Table

### A red and blue squares with white text AI-generated content may be incorrect.

Fig.2 Dataset 1 Heatmap

A graph with blue squares

AI-generated content may be incorrect.

Fig.3 Distribution of Flow Duration

A graph with a number of lines

AI-generated content may be incorrect.

Fig.4 Distribution of Total Length of Forward Packets

A graph with lines in the middle

AI-generated content may be incorrect.

Fig.5 Distribution of Total Length of Backward Packets

A graph with lines in the middle

AI-generated content may be incorrect.

Fig.6 Distribution of Forward IAT

A graph with lines and dots

AI-generated content may be incorrect.

Fig.7 Distribution of Backward IAT

A colorful bars on a white background

AI-generated content may be incorrect.

Fig.8 Protocols Distribution

### 

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

### A table with numbers and text AI-generated content may be incorrect.

Fig.9 Descriptive Statistics Table

A screenshot of a computer screen

AI-generated content may be incorrect.

Fig.10 Dataset 2 Heatmap

A graph of blue lines

AI-generated content may be incorrect.

Fig.11 Distribution of Tx

A graph of blue lines

AI-generated content may be incorrect.

Fig.12 Distribution of Rx

A graph with blue lines

AI-generated content may be incorrect.

Fig.13 Distribution of SNR Receiver

A graph of a bar graph

AI-generated content may be incorrect.

Fig.14 Distribution of BER Receiver

A graph with blue lines

AI-generated content may be incorrect.

Fig.15 Distribution of Transmission Distance

A graph with blue lines

AI-generated content may be incorrect.

Fig.16 Distribution of Fiber Attenuation

A graph with blue lines

AI-generated content may be incorrect.

Fig.17 Distribution of Temperature

A graph with blue lines

AI-generated content may be incorrect.

Fig.18 Distribution of Humidity

### 2.2.3 Dataset 3 (to answer Question ): 9 Columns & 463 Rows

### A table with numbers and letters AI-generated content may be incorrect.

Fig.19 Dataset 3 Descriptive Statistics Table

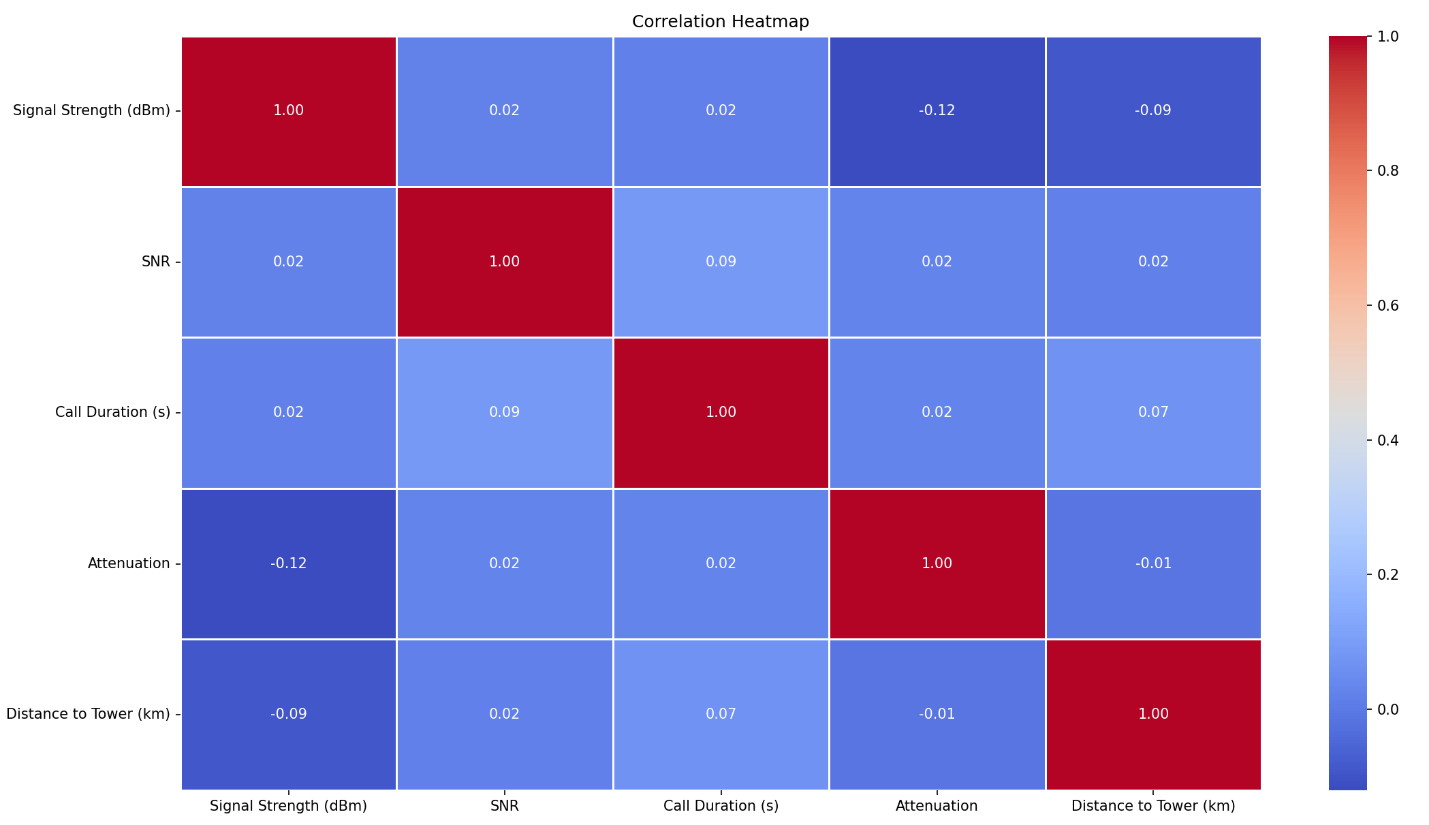


Fig.20 Dataset 3 Heatmap

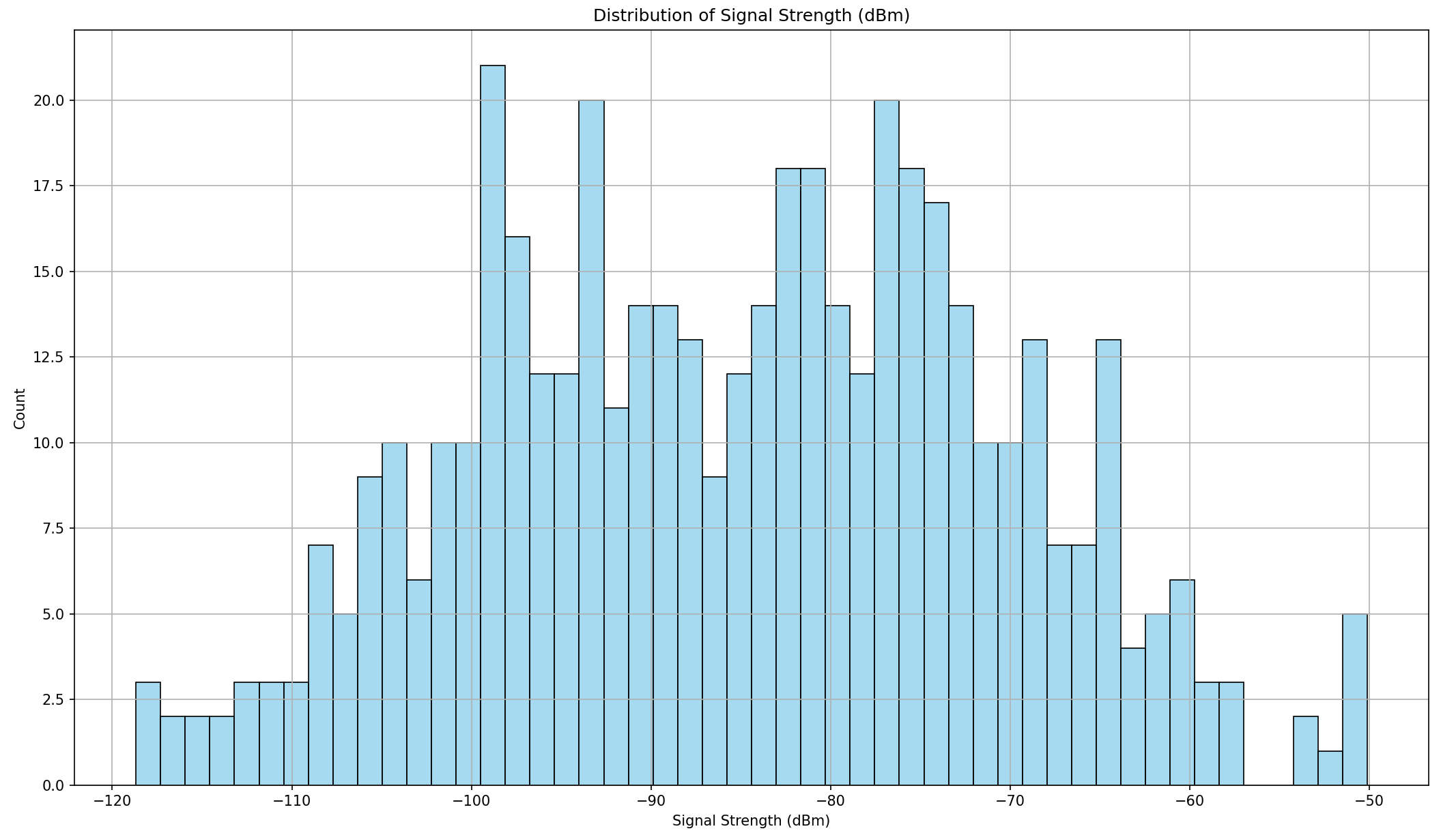


Fig.21 Distribution of Signal Strength

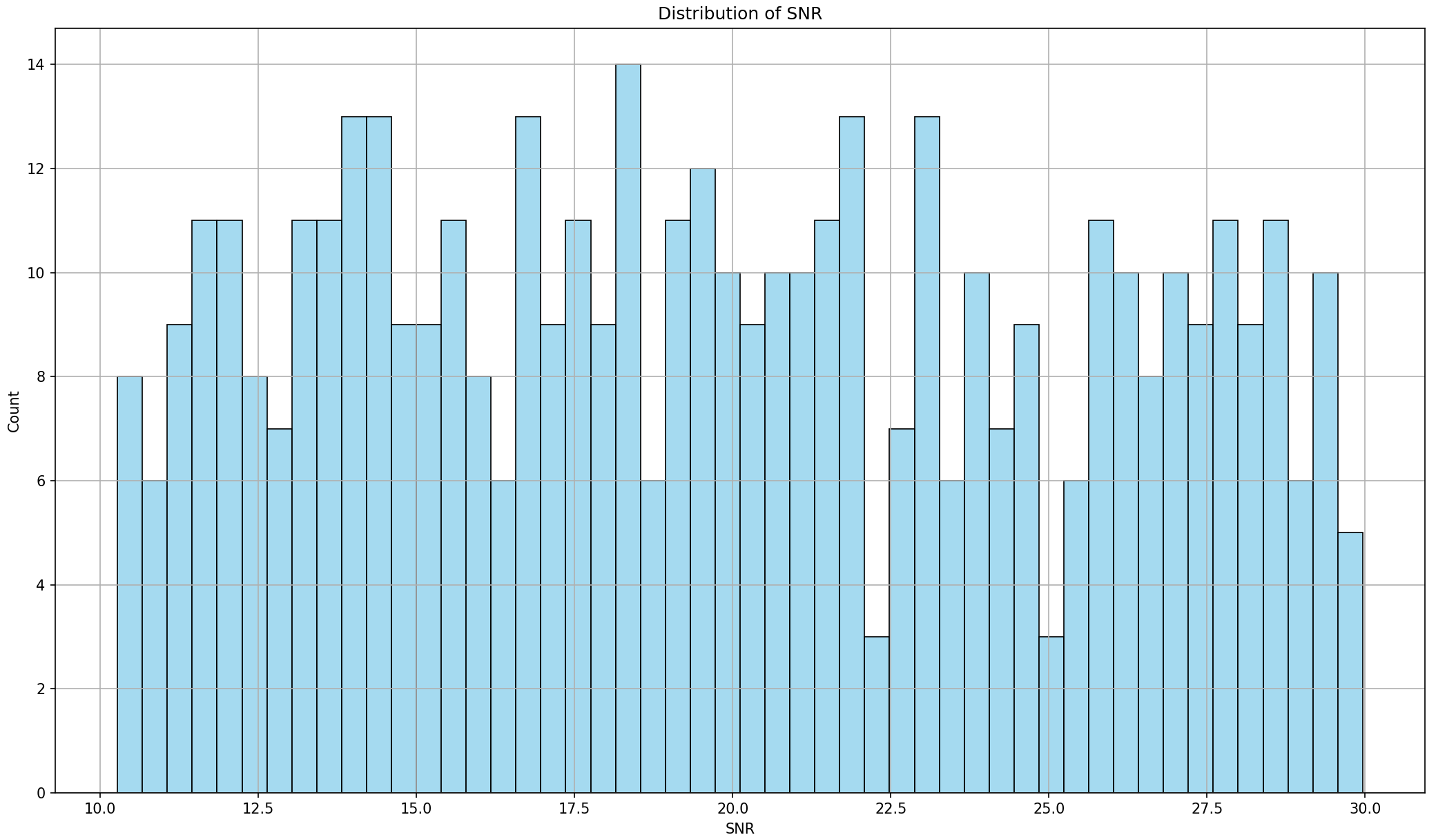


Fig.22 Distribution of SNR

A graph with blue lines

AI-generated content may be incorrect.

Fig.23 Distribution of Call Duration

A graph of a graph

AI-generated content may be incorrect.

Fig.24 Distribution of Attenuation

A graph with blue lines

AI-generated content may be incorrect.

Fig.25 Distribution of Distance to Tower

A colorful rectangular objects with numbers

AI-generated content may be incorrect.

Fig.26 Distribution of Environment

A blue and green rectangles

AI-generated content may be incorrect.

Fig.27 Distribution of Incoming/Outgoing

A screenshot of a computer screen

AI-generated content may be incorrect.

Fig.28 Distribution of Call Type

# Chapter 3: Methodology

The methodology is divided into three main parts: the Artificial Intelligence (AI) model, image preprocessing, and desktop application. The following sections delve deeply into each of these components.

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

**Key Metrics for Analysis**

**Timestamp**: The recorded time when the signal strength and SNR were measured. This helps track fluctuations throughout the day.

**Acquisition Type**: 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 increased 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.
  + **H₀ (Null Hypothesis):** No significant difference in signal fluctuations between acquisition types.
  + **Hₐ (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?**

After collecting data from the schemas here is what we found:

**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):** A one-way ANOVA test was performed to assess how SNR varies across call length and type of data (data, voice).

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.

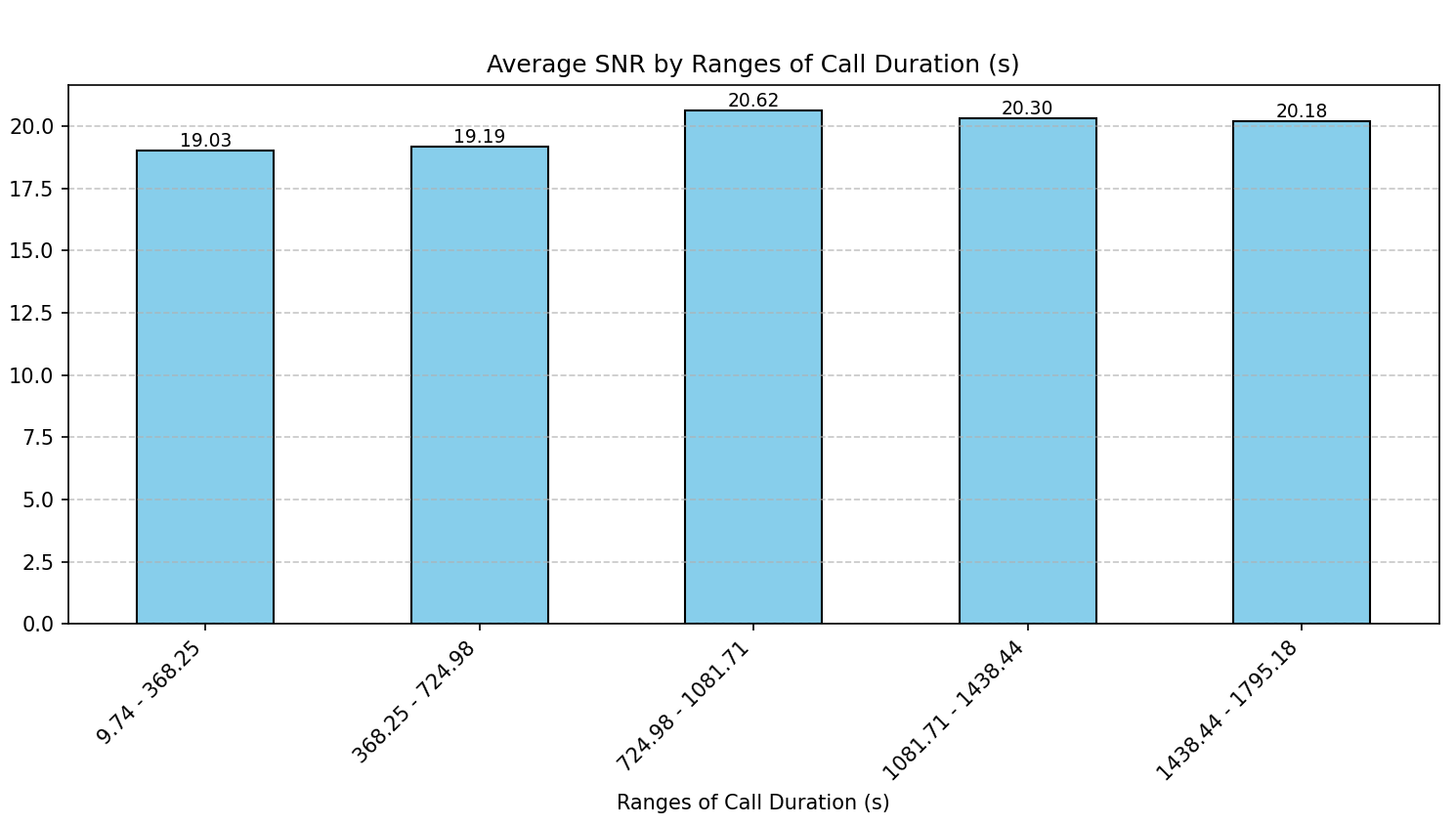
Ha:

* **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:** test 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."

A graph of a graph

AI-generated content may be incorrect.

****

**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 the train.csv dataset. 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.

1. **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 environmental conditions with different transmission distances on network performance?**

**After collecting and analyzing data from the three datasets, 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 (H₀):**

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

**Alternative Hypotheses (H₁):**

* **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 condition is a significant factor in predicting performance, especially in wireless systems (as seen in train.csv).**
* **Fiber attenuation increases with temperature and humidity to a small but observable degree (ocrdataset.csv).**

**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

pllm k

# References

* Nathaniel Handan. “OptiCom Signal Quality Dataset” [**https://www.kaggle.com/datasets/tinnyrobot/opticom-signal-quality-dataset?select=ocrdataset.csv**](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**](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**](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

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-Unicauca-Version2-87Atts.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()

# Load dataset

df = pd.read\_csv("C:/Users/osama/OneDrive/Desktop/Data\_Architects/Datasets/ocrdataset.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/train.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()

Dataset used:

<https://github.com/TarekOsama528/Data_Architects/tree/main/Datasets>