Mini-Project Report: Data Mining & Analysis of NITJ WiFi Speeds

Student: Rahul Deb Nath
Roll No: 23103114
Student: Rittik Ganchaudhuri
Roll No: 23103123

1. Introduction

A reliable and high-performance WiFi network is essential infrastructure for a modern academic institution like NITJ. It supports everything from online learning and research to daily communication. However, network performance is often variable, influenced by factors like user load, physical location, and time of day.

This project aims to analyze and mine patterns from a dataset of WiFi speed tests conducted at various locations across the NITJ campus. The primary objective is to use the R programming language to clean, process, and visualize the data, uncovering insights into the network's performance. The analysis seeks to answer practical questions:

- What is the typical download/upload speed?
- How does performance and reliability vary by location?
- Does time of day impact speed?
- Are there other correlations, such as between speed and packet loss?

2. Methodology

The entire analysis was conducted using the R programming language (Version 4.x) and RStudio. The workflow was divided into four main stages.

2.1. Tools and Packages
The analysis relied on a set of modern R packages, primarily from the tidyverse:

- **tidyverse**: A collection of packages for data science, including:
  - **readr**: For loading the CSV file.
  - **dplyr**: For all data manipulation and aggregation.
  - **ggplot2**: For creating all static visualizations.
- **lubridate**: Used to parse and manipulate date and time columns.
- **janitor**: Used to clean and standardize all column names for easier analysis.

2.2. Data Loading and Cleaning
The dataset (NITJ_WiFi_Speed.csv) was loaded into R. The first step was to clean the column names using janitor::clean_names(), which converts names like "d_speed(mbps)" to a usable d_speed_mbps. The spot_location column also required cleaning to remove extra quotation

marks.

2.3. Feature Engineering

To enable time-based analysis, new features were engineered from the existing date and time columns:

1. A single datetime object was created by parsing and combining the two columns.
2. Using this datetime object, two new columns were created:
   - hour_of_day: An integer from 0 to 23.
   - day_of_week: A factor (e.g., "Monday", "Tuesday") to analyze weekly trends.

2.4. Data Analysis and Visualization

The core of the project involved aggregating the cleaned data to find patterns. The dplyr package was used to group data by location and hour, calculating the mean download speed for each group. ggplot2 was then used to create six key visualizations (histograms, bar charts, boxplots, line charts, and scatter plots) to present these findings.

3. Results and Discussion

The analysis generated the following visual patterns:

Result 1: Overall Download Speed Distribution

(Insert plot_1_histogram.png here)

The histogram shows the frequency of all recorded download speeds. The distribution is centered around 20-22 mbps, with the majority of tests falling between 15 and 25 mbps. This suggests that while speeds vary, a user can typically expect performance in this range.

Result 2: Average Performance by Location

(Insert plot_3_bar_location.png here)

This bar chart reveals that location is a significant factor in average speed. The "New Library Building" and "Boys hostel 7E" show the highest average download speeds. Conversely, the "IT Block" and "Science Block" show notably lower average speeds.

Result 3: Performance Distribution & Reliability by Location

(Insert plot_2_box_location.png here)

While the bar chart (Result 2) shows the average speed, this boxplot provides a much deeper insight into the distribution and reliability of speeds at each location. The 'box' represents the middle 50% of all tests, and the line inside is the median.

This plot confirms the "New Library Building" has a high median speed. More importantly, it shows that the "IT Block" is highly unreliable: it has a very wide spread of speeds and many low-speed outliers (the dots), making it a poor location for consistent internet.

Result 4: Performance by Hour of Day

(Insert plot_4_line_hour.png here)

The line chart of hourly performance shows a clear temporal pattern. Speeds remain high and stable during the afternoon and evening (13:00 - 23:00) but experience a sharp decline in the early morning, bottoming out around 2:00-4:00 AM. Speeds begin to recover around 5:00 AM and peak in the afternoon. This could be due to scheduled network maintenance.

Result 5 & 6: Speed Correlations

(Insert plot_5_scatter_d_vs_u.png and plot_6_scatter_d_vs_loss.png here)

Two scatter plots were generated to find other relationships:

- **Download vs. Upload:** (Left) There is a clear positive correlation. When download speed is high, upload speed tends to be high as well.
- **Packet Loss vs. Download:** (Right) Packet loss is consistently near 0% for all tests, regardless of download speed. This is an excellent result, indicating a highly stable and reliable connection with minimal data loss.

4. Conclusion

This analysis successfully identified several key patterns in the NITJ WiFi network. The main findings are:

1. The typical download speed is a solid **15-25 mbps**.
2. **Location is the most critical factor.** The "New Library Building" is a prime spot for both high speed and high reliability. Conversely, the "IT Block" is not only slower on average but also **highly unreliable**, with a wide variance in performance and many low-speed outliers.
3. **Time of day matters**, with speeds dropping significantly in the early morning hours (2-4 AM).
4. The network is **highly stable**, with near-zero packet loss across almost all tests.

Future work could involve collecting a larger dataset with more locations to build an even more comprehensive performance and reliability map of the campus network.

5. GitHub Link

The dataset and R script used for this analysis are available at the following repository:

nitj_wifi_speed_test