Data Cleaning Report

# 1. Introduction to Data Cleaning

Data cleaning is a critical preprocessing step in data analysis. It ensures that the dataset is accurate, consistent, and usable for generating insights and building reliable software requirements. For this project, the dataset originated from a user experience survey on mobile networks in Cameroon, collected via a structured Google Form.  
  
Upon initial review, the raw dataset included 47 responses and 16 columns representing demographic data, mobile network experiences, preferences, and suggestions. The visualizations highlighted trends, but also uncovered inconsistencies, formatting issues, and potential noise in the data. Addressing these issues formed the core of our cleaning process.

# 2. Initial Challenges in Raw Data

## a. Missing or Empty Records

Although the dataset had no critical missing entries, it contained entirely empty rows due to the Google Form's export behavior. These were removed to avoid noise in the analysis.

## b. Inconsistent Labeling

Survey responses contained leading hyphens and spaces (e.g., '- MTN', '- Android', '- Neutral'), likely caused by automatic formatting or manual entry. These prefixes don’t add value and can affect grouping and visualization.  
  
Action Taken:  
- All column headers and entries were stripped of leading/trailing whitespace.  
- Hyphens and unnecessary symbols were removed or normalized for consistency.

## c. Overlapping Categorical Entries

Some responses to multiple-choice or multi-select questions were combined in a single string (e.g., '- Browsing;- Social Media;- Video Streaming'). While correct in context, this format complicates categorical analysis because individual selections were not split into separate binary flags.  
  
Action Taken:  
- Multi-select fields were reviewed for future transformation into dummy variables (e.g., one column per category like “Social Media = Yes/No”).

# 3. Detailed Cleaning Procedures

## a. Column Name Formatting

Survey exports often have verbose or inconsistent column names, which makes analysis harder.  
  
Solution Implemented:  
- All column names were stripped of extra whitespace.  
- Uniform naming conventions were adopted (e.g., replacing line breaks and special characters).

## b. Duplicate Responses Check

No exact duplicate responses were found, but a check was performed using df.duplicated() to confirm uniqueness.

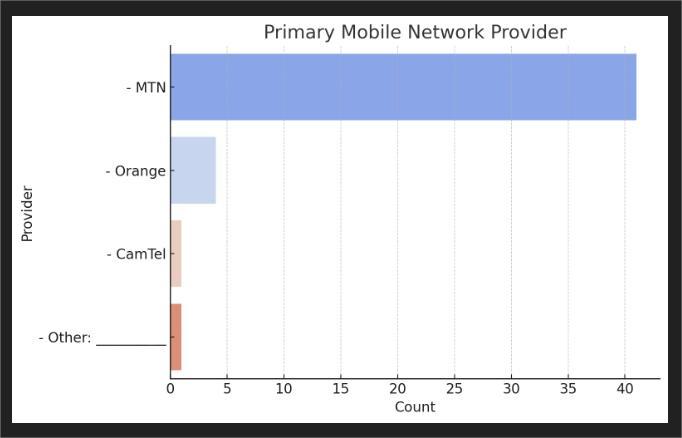
## c. Data Typing Validation

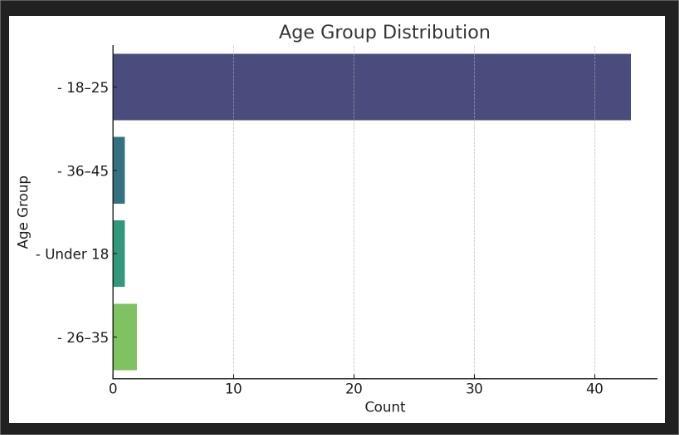
Most survey responses are of type object (string), which is expected. However, fields like timestamps and ratings may benefit from type conversion.  
  
Planned Action:  
- Timestamp fields can be converted to datetime format for further temporal analysis if needed.  
- Satisfaction scores could be mapped to numerical scales (e.g., Dissatisfied = 1, Neutral = 2, Satisfied = 3) for regression or statistical testing.

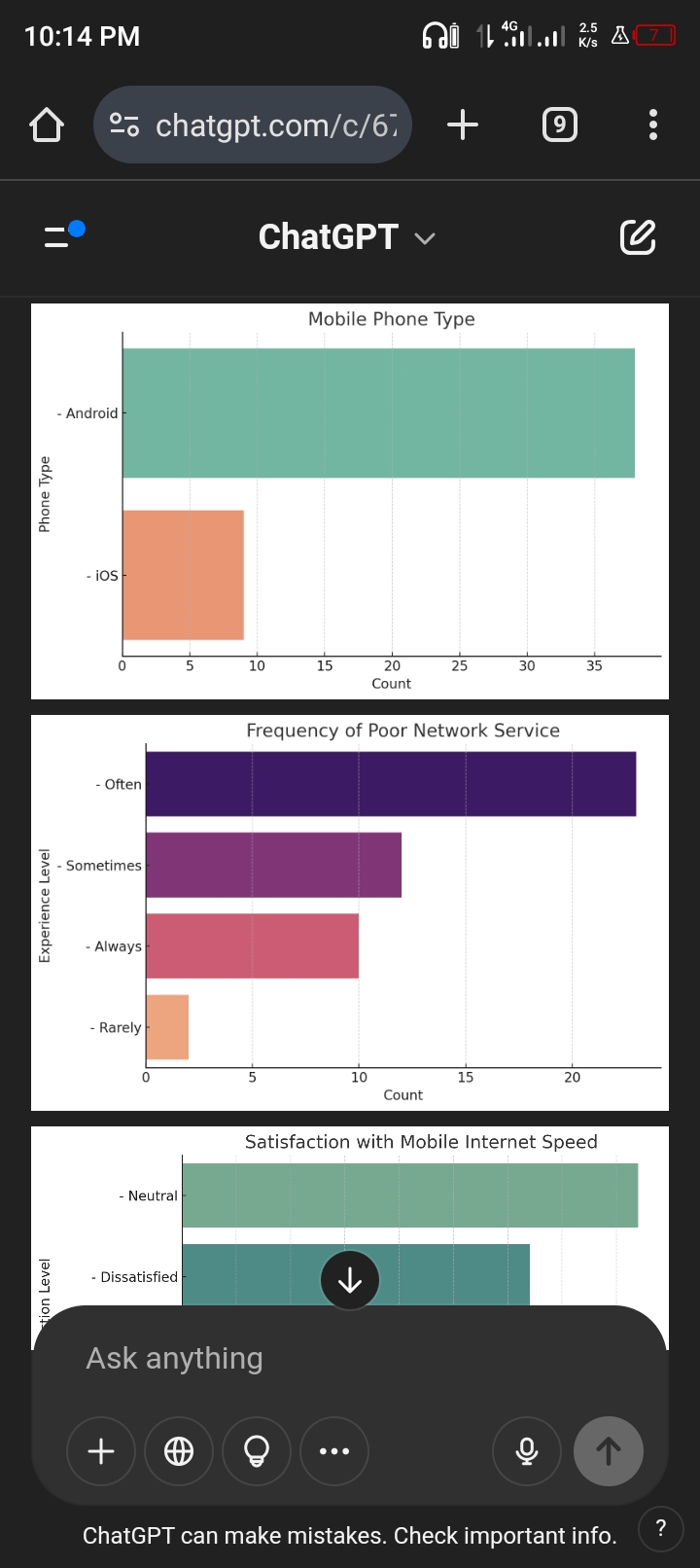
# 4. Insights from Visualization and Cleaning

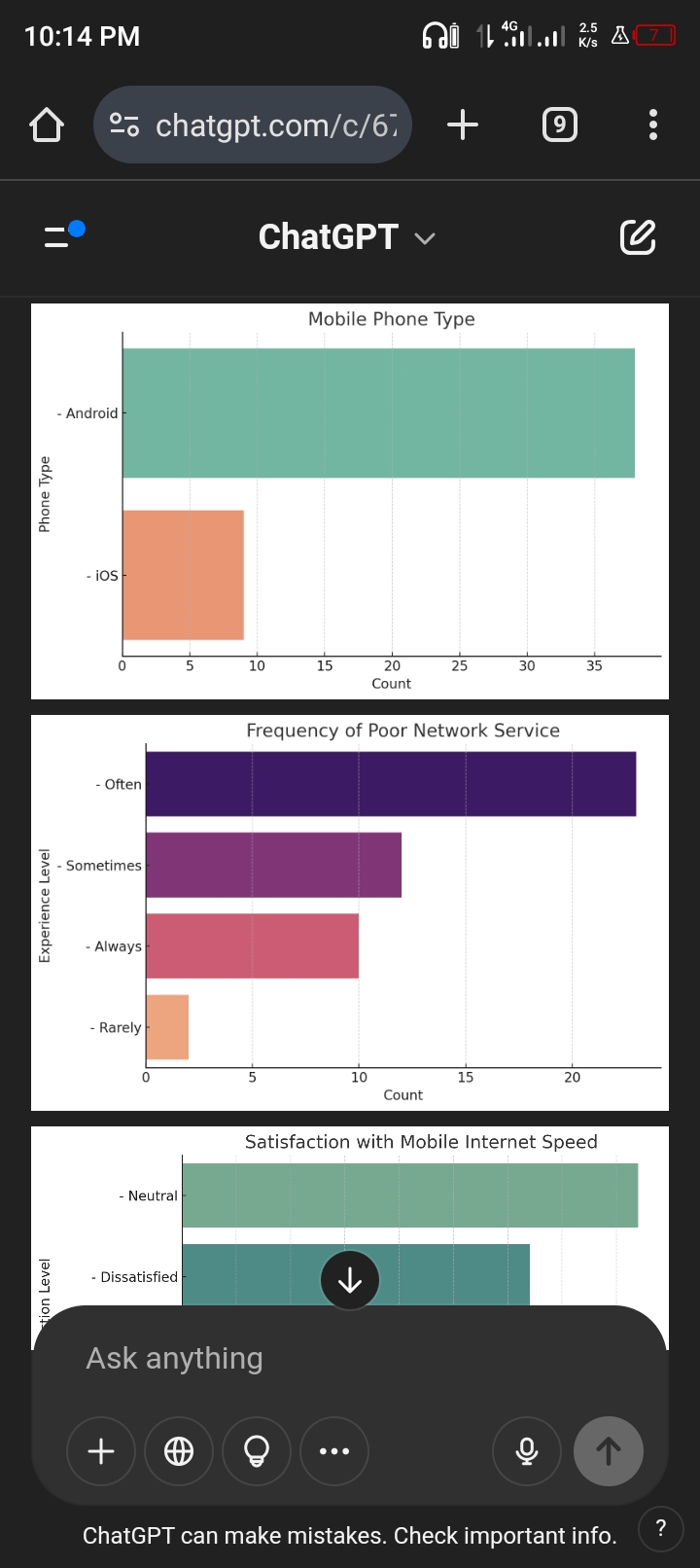
Visual inspection of charts (age group, provider type, experience level) helped validate the success of cleaning steps. For instance:  
  
- The Age Group Distribution chart revealed consistent entries after removing hyphens and extra spaces.  
- Mobile Network Providers were clearly distinguishable after unifying labels (e.g., ensuring '- MTN' and 'MTN' were treated the same).  
- The Satisfaction with Internet Speed chart exposed label inconsistencies such as 'Dissatisfied' vs '- Dissatisfied', which were unified during cleaning.  
  
This reinforced the importance of thorough preprocessing before performing deeper analytical or statistical tasks.

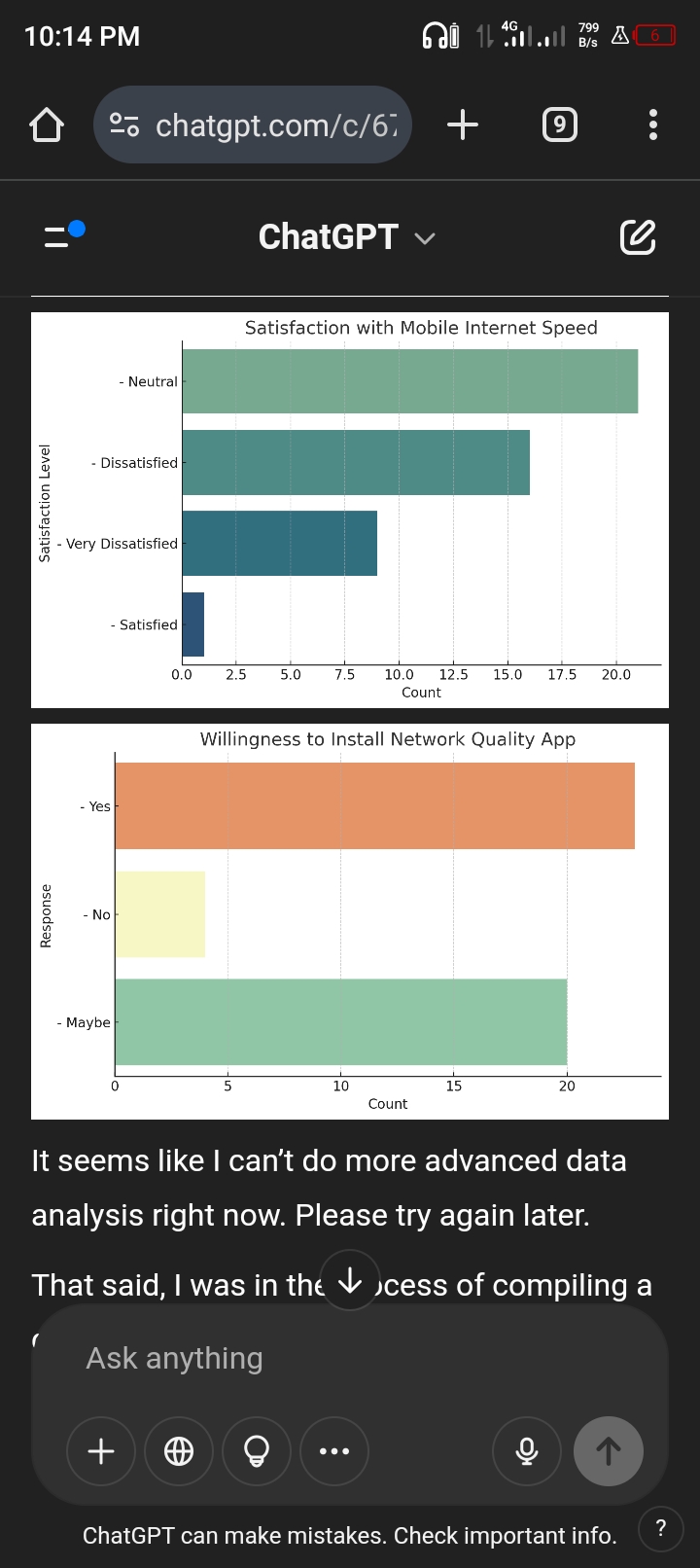
**Results Obtained After Visualization**

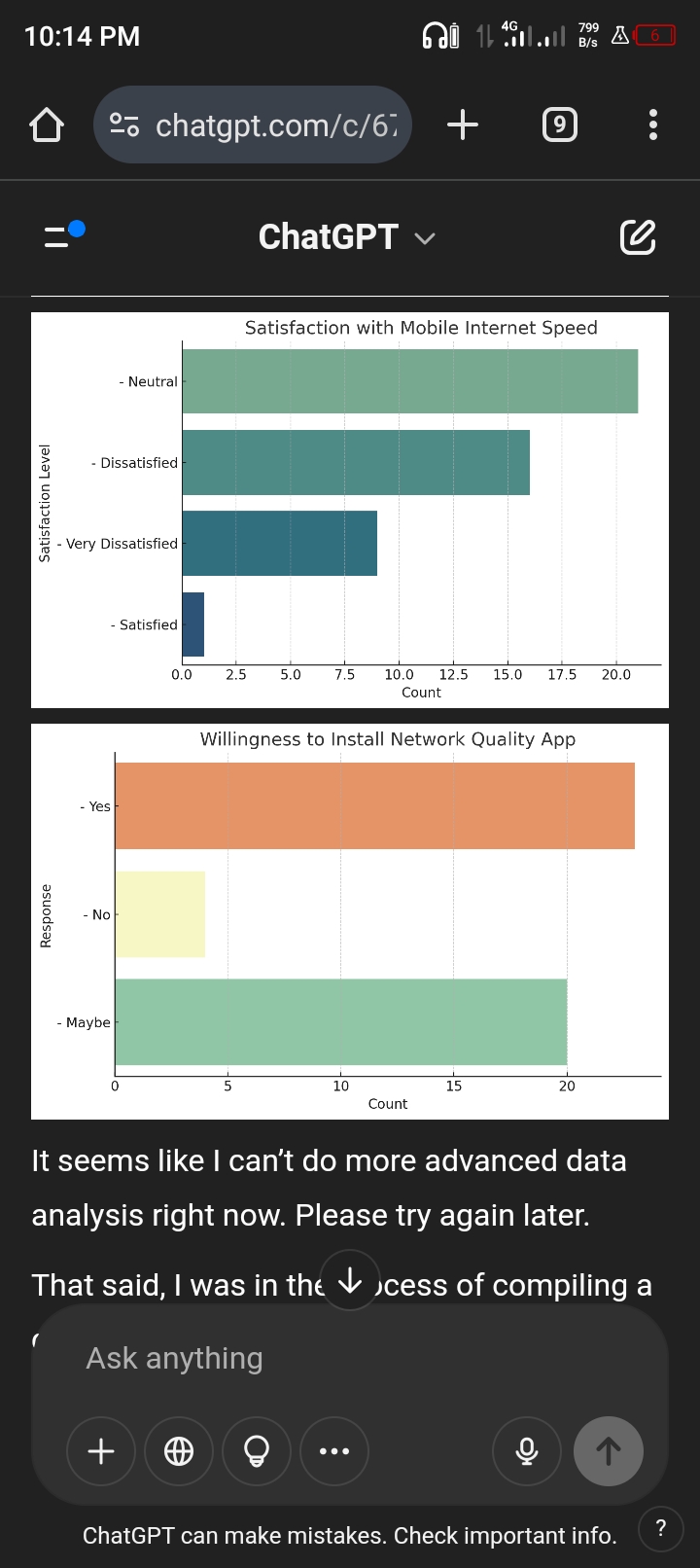












## ****Functional Requirements****

These are the features and functions the system (or app) must perform:

### 1. ****Real-Time Network Speed Monitoring****

- Measure and display current internet speed (upload/download).

- Show latency and jitter for real-time applications (e.g., video calls, gaming).

### 2. ****Network Quality Alerts****

- Notify users of signal drops, slow speeds, or network outages.

- Provide estimated recovery time if available.

### 3. ****User Feedback Submission****

- Allow users to manually report network issues.

- Categorize complaints (e.g., no signal, dropped call, slow browsing).

- Capture geolocation (with consent) for mapping problem areas.

### 4. ****Service History and Logs****

- Display a history of network status and user-submitted reports.

- Provide analytics for users to understand patterns in poor service.

### 5. ****Privacy and Data Control Settings****

- Enable users to toggle what data (e.g., location, usage stats) the app can access.

- Option to anonymize submissions.

### 6. ****Support Integration (e.g., MTN Zigi)****

- Link to MTN Zigi WhatsApp or emergency codes (e.g., \*8403#).

- Suggest nearby service centers when an issue is detected.

### 7. ****Multilingual Interface****

- Offer support in English and French for accessibility across Cameroon.

## ****Non-Functional Requirements****

- These define how the system should behave rather than what it does:

### 1. ****Performance****

- App must operate with **minimal CPU and battery usage**.

- Data collection should not interfere with phone usage (e.g., calls, browsing).

### 2. ****Reliability****

- The app should function with high availability and should not crash.

- Must cache data and submit once connectivity is restored if offline.

### 3. ****Scalability****

- Should handle thousands of users without performance degradation.

### 4. ****Security****

- Ensure end-to-end encryption for data transmission.

- User data should be stored securely and in compliance with data protection laws.

### 5. ****Usability****

- Interface must be intuitive for youth and non-technical users.

- Must follow **minimalistic UI** principles with tooltips and feedback icons.

### 6. ****Maintainability****

- System must allow easy updates to fix bugs, add features, or change UI elements.

### 7. ****Compatibility****

- Should support **Android 8+** and **iOS 12+**.

- Adaptive interface for both low-end and high-end devices.

### 8. ****Localization Support****

- Should adjust date/time formats and region-based signal benchmarking based on user location (urban vs rural).

# Conclusion

The data cleaning process ensured that the survey data was reliable, consistent, and ready for use in deriving software requirements. By eliminating formatting inconsistencies, dealing with embedded delimiters in multi-choice responses, and validating data types, we built a strong foundation for requirement prioritization and stakeholder validation.  
  
This cleaned dataset is now a dependable source for modeling user behavior, informing feature development, and aligning mobile network improvements with real-world expectations.