Data Preparation Report

Students Names:

- Alam Bebar - 207415407
- Ryan Thawkho - 305070567

Table of Content

Executive Summary

This report outlines the data preparation process for our project on male pattern baldness by using two main sources: a survey-based dataset from Mendeley and a scalp image dataset from Roboflow. In the first section, we explain how and why we selected these datasets. We then detail the cleaning process, including handling missing data, fixing errors, and addressing coding inconsistencies in both datasets. To enhance our analysis, we created new data by designing a Google Form that mirrors the original features, which allows us to collect additional responses. We also describe how the datasets will be integrated using Python, and how we formatted the data to ensure compatibility with future modeling, including encoding features and resizing images. Finally, we performed exploratory data analysis to identify key patterns, with a focus on visualizations and correlation with the target variable.

Report Content

1. Selecting Data

Data Sources:

For this project, we utilize two datasets from different sources:

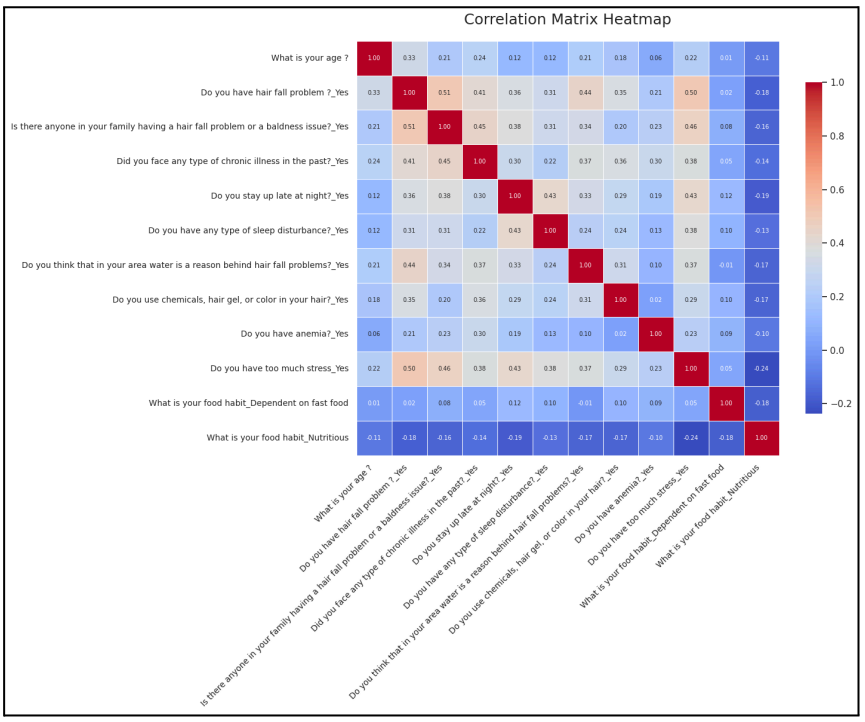
- **Hair fall causes dataset (Mendeley)**
  - **Source:** Mendeley data repository
  - **Description:** This dataset consists of **717 survey responses from male and female participants** regarding hair care habits, health conditions, and lifestyle choices that may influence hair loss. The data was collected via Google Forms distributed across social media and online communities. The survey is available for

downloading either as a CSV or XML file. The survey dataset is the sole content of the respiratory.

- **Scalp image dataset (Roboflow)**
  - **Source:** [Roboflow dataset repository]
  - **Description:** This dataset contains **images of male scalps** taken from different angles (front, back, left, right, top-down). These images illustrate the Norwood scale male pattern baldness stages from 2-7. The dataset will be used for training deep learning models to visually assess hair loss severity.

Data Selection:

- **Rows (Observations)**:
  - We **excluded all female entries** since the focus is on male pattern baldness.
  - We **kept 459 male responses** from the survey dataset.
  - **Survey dataset separation led to higher correlation:** During our initial correlation analysis, we observed weak relationships between features in the dataset. However, given that our model aims to predict male pattern baldness, and based on scientific research indicating that males are significantly more prone to this condition, we decided to exclude all female entries from the dataset. After filtering the data to include only male participants, we recalculated the correlation matrix and observed higher correlation rates between key features. This refinement strengthens our belief that focusing solely on male data will enhance our model's predictive efficiency for hair loss progression.
- **Columns (Attributes)**:
  - We **removed irrelevant attributes**, including timestamps and names.
  - Retained key variables such as **family history, water quality, stress levels, and sleep patterns**, which showed moderate correlation with hair loss.
  - Categorical variables (e.g., "Do you stay up late?") were prepared for encoding.
  - After reviewing the datasets collected in the data understanding phase, we selected the most relevant data to support our analysis of male pattern baldness progression. We have calculated the correlation values of the columns only after we preprocessed and cleaned the dataset. The following **columns had the highest correlation rate.**



Survey question	Values	Correlation value
Do you have hair fall problem ? [target variable]	Yes/No	1
Is there anyone in your family having a hair fall problem or a baldness issue?	Yes/No	0.51
Do you have too much stress	Yes/No	0.50
Do you think that in your area water is a reason behind hair fall problems?	Yes/No	0.43
Did you face any type of chronic illness in the past?	Yes/No	0.40
Do you use chemicals, hair gel, or color in your hair?	Yes/No	0.36
Do you stay up late at night?	Yes/No	0.35
What is your age ?	Numercial	0.32

2. Cleaning Data

We identified and addressed various data issues in both datasets:

Hair fall causes dataset (Mendeley)

Handling Missing Data:

- In the survey dataset we conducted an analysis to identify missing values in the dataset to ensure completeness. We used a function to count the number of missing values in each column.

Fixing Data Errors:

- During the data cleaning phase, we identified an anomaly in the Age column, one entry was 218 which seems to be a data entry error. We assume that the user meant to enter the number 18 so we correct this entry manually.
- We also identified an inconsistency in the “Do you use chemicals in your hair” column. While most responses were recorded as Yes or No there was an entry labeled as Yea which deviated from the standard format. To ensure consistency in categorical responses we changed this entry to Yes.
- In the “Do you have too much stress” column we found an entry recorded as “\No” which could cause errors in data processing. We corrected the entry by removing the leading slash to standardize it as No

Addressing Coding Inconsistencies:

- Binary categorical responses (Yes/No) were converted into **0s and 1s**.
- Multi-category attributes (e.g., food habits) were **one-hot encoded**.
- The scalp image labels were **standardized** according to the Norwood scale.

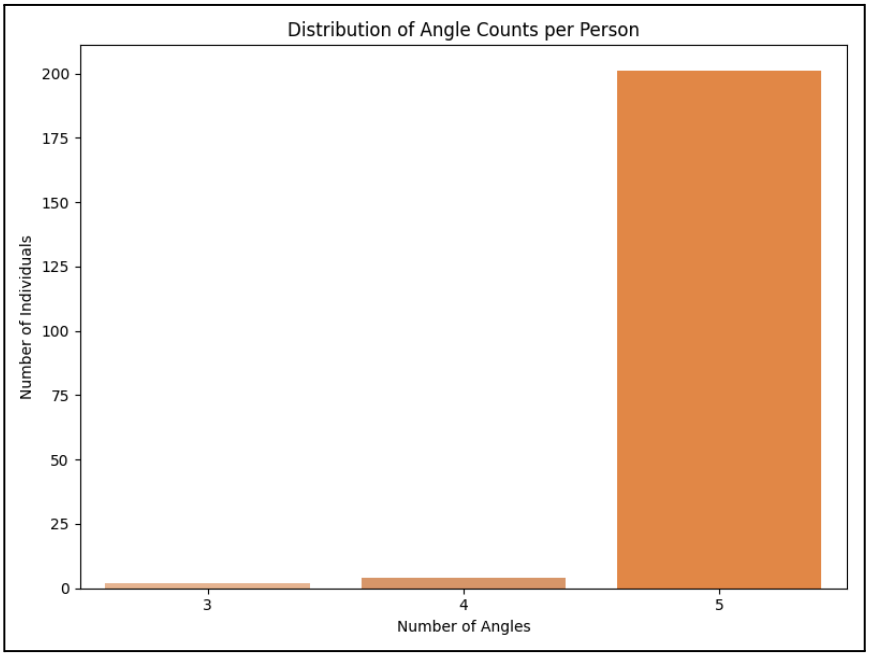
Scalp image dataset (Roboflow)

Handling Missing Data:

- Since our dataset of male scalp images lacks examples of Norwood Scale Stage 1 male pattern baldness, we decided to group Stage 1 (no significant hair loss or hairline

recession) together with Stage 2 (slight recession of the hairline around the temples). As a result, both will be classified under a combined category: **Stage 1–2**.

- The **scalp image dataset lacked Stage 1 baldness images**, requiring augmentation or sourcing from other datasets.
- In the scalp image dataset not every individual has images in all 5 angles. We decided to keep it this way knowing there might be accuracy differences between angles.
- While examining the scalp image dataset, we found that most individuals have five images representing different angles. However, in the training folder, each image is duplicated multiple times with a 90-degree rotation. This means that instead of having truly distinct angles, the dataset includes rotated versions of the same image, which may mislead the model into learning from artificially augmented data rather than genuine visual diversity. We decided to keep the rotated versions and treat them as natural variation, and allow the model to learn from them, since CNNs benefit from such data.



This graph shows that most of the participants in the scalp images data set five images representing five different angles

This image shows the contents of the 'Training' folder in the scalp image dataset, where the same image appears multiple times, each rotated at different angles.

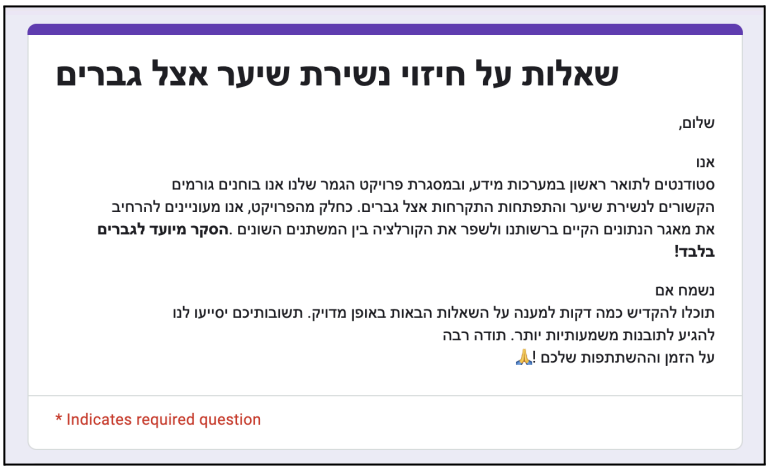
Fixing Data Errors:

- While exploring the scalp image dataset, we noticed that some individuals had multiple images from the same angle (e.g., right side), but the images were rotated differently — for example, one version with a straight head and another tilted 90 degrees. Since the actual content of these images is the same, just rotated, we considered whether to remove them or treat them as valuable training data. In the end, we decided to **keep these rotated images** in the dataset because they introduce natural variations in head orientation. This can actually help our CNN model learn to generalize better and be more robust when identifying hair loss patterns, even when the head is tilted in real-world scenarios.
- Some photos have high exposure or lighting, which may affect the model's ability to accurately identify the stage of male pattern baldness and may require additional enhancements.

3. Constructing New Data

As part of our effort to improve the quality and performance of our analysis on hair fall causes, we constructed new data by generating additional records. Specifically, we designed and distributed a

Google Forms questionnaire aimed at replicating the most relevant features from the existing real-world dataset sourced from Mendeley ("Dataset for Evaluating Hair Fall Causes Using Machine Learning Techniques", See Data Sources).



שאלות על חיזוי נשירת שיער אצל גברים

שלום,

אנו
סטודנטים לתואר ראשון במערכות מידע, ובמסגרת פרויקט הגמר שלנו אנו בוחנים גורמים
הקשורים לנשירת שיער והתפתחות התקרחות אצל גברים. כחלק מהפרויקט, אנו מעוניינים להרחיב
את מאגר הנתונים הקיים ברשותנו ולשפר את הקורלציה בין המשתנים השונים. **הסקר מיועד לגברים בלבד!**

נשמח אם
תוכלו להקדיש כמה דקות למענה על השאלות הבאות באופן מדויק. תשובותיכם יסייעו לנו
להגיע לתובנות משמעותיות יותר. תודה רבה
על הזמן וההשתתפות שלכם !🙏

* Indicates required question

Google Forms questionnaire conducted to collect records for the existing Mendeley dataset

After analyzing the original dataset, we identified the columns with the highest correlation to hair fall severity and selected them to be included in our custom survey. This allowed us to preserve the structure and relevance of the dataset while collecting more representative and up-to-date responses from our target audience, mainly students. We succeeded in collecting responses from 24 participants.

The responses collected from Google Forms will be exported and **appended to the existing dataset** using the **Pandas** library in Python, while effectively increasing the number of records (rows) and enhancing the dataset's reliability and statistical power.

We ensured that:

- The data types in the form match the original dataset to maintain consistency.
- The responses collected are categorical, and any necessary transformations (such as encoding) will be done in later preprocessing steps to support machine learning models.
- No normalization was required at this stage since all responses reflect qualitative or ordinal data.
- No new attributes were derived at this point, but future iterations may include constructed features (e.g., a stress score) based on combined inputs.

4. Integrating Data

Since our project involves both survey data (numerical) and scalp images (visual), we needed to establish a structured way to integrate them:

- The **image dataset** is used for classification of baldness stages.
- The **survey dataset** is used to identify contributing factors and provide prevention recommendations.
- The two datasets are **linked indirectly**, where the model first classifies baldness stage from images, then suggests recommendations based on the survey insights.
- We've documented the Python code used to preprocess the Mendeley dataset with Pandas and other libraries in a Google Colan notebook. As part of the process, we appended the responses from our Google Form to the original dataset, ensuring that the column structure and formatting were consistent between the two sources. (See [Alam_Ryan Datasets Preparation Code.ipynb])

Data preprocessing and correlation matrix of male records

```
[19] import pandas as pd
     import seaborn as sns
     import matplotlib.pyplot as plt
     from sklearn.preprocessing import MinMaxScaler

     data = pd.read_csv("hairfall_problem3592.xlsx - hairfall_problem3592 (1).csv")  #read and preprocess the data
     #data = pd.read_csv('/content/survey data.csv')
     male_records = data[data['What is your gender ?'] == 'Male']
     df = male_records.drop(columns=["Timestamp", "What is your name ?"], axis=1)  #drop irrelevant columns for analysis

     # check the age column for incorrect values
     invalid_ages = df[(df['What is your age ?'] < 0) | (df['What is your age ?'] > 120)]
     #print the error value and its index
     print("Out of range values in age column:")
     print(invalid_ages[['What is your age ?']])

     # replace the incorrect age to 18
     df.loc[503, 'What is your age ?'] = 18  # Correcting 218 to 18

     # Columns to exclude from checking
     excluded_columns = [
         "What is your age ?",
         "What is your gender ?",
         "What is your food habit"
     ]
```

Example of a code block for the initial step for preprocessing and cleaning our survey Mendeley dataset

5. Formatting Data

- **Survey Data**: Converted all categorical values to numerical representations.
- **Image Data**: May be preprocessed to have **consistent dimensions** (resized, normalized brightness, and augmented where necessary).

- We have handled missing data, corrected data errors, and addressed coding inconsistencies in both the image and survey datasets to ensure clean and reliable data for analysis and modeling.

Models

Hair Fall Causes Dataset (Mendeley)

This is a **tabular, structured** dataset derived from survey responses. It includes binary (Yes/No) and categorical values that have been one-hot encoded.

Preprocessing Steps:

- **One-hot encoding** applied to all categorical variables (Yes: 1, No: 0)
- **Dropped irrelevant fields** (e.g., name, timestamp)
- **Handled missing values**
- **Optional:** Normalize features for algorithms sensitive to scale (e.g., KNN, Logistic Regression)

Planned models and requirements:

1. Logistic Regression

- Requires numeric features
- Perform better with normalized data
- Output interpretable coefficients

2. K-Nearest Neighbors (KNN)

- Sensitive to **feature scaling** → use StandardScaler or MinMaxScaler
- Distance-based , which means that high dimensionality can reduce performance

3. Random Forest

- Handles categorical and numerical features
- No need for normalization
- Can handle missing values (in some implementations)

4. Gradient Boosting (XGBoost/LightGBM)

- Works well with default settings
- Boosting can handle weak feature interactions
- Normalization optional but beneficial for convergence speed

Roboflow Scalp Image Dataset

This dataset consists of **labeled scalp images** depicting various stages (2–7) of male pattern baldness based on the **Norwood scale**. The images are taken from five angles: front, back, left, right, and top-down.

Preprocessing Requirements:

- **Image size:** Resize all images to a uniform input shape, commonly **224×224 pixels**, which is compatible with most pre-trained CNN models
- **Color channels:** Ensure images are **RGB (3 channels)**
- **Pixel value normalization:** Scale pixel values to range **[0, 1]** or **[-1, 1]**
- **File format:** PNG or JPEG (consistent format)
  Data Augmentation: Apply transformations such as:
  - Horizontal/vertical flip
  - Rotation
  - Zoom
  - Brightness/contrast adjustment (to mitigate lighting variance)

Planned Models and Requirements:

1. ResNet (e.g., ResNet50)

- Input shape: **(224, 224, 3)**
- Requires normalized images
- Best used with **transfer learning** (pre-trained weights on ImageNet)

2. MobileNet

- Designed for performance on mobile/embedded devices
- Input shape: typically **(224, 224, 3)**
- Lightweight and efficient → ideal for deployment

3. VGG16

- Input shape: **(224, 224, 3)**
- Deep network with uniform structure (3×3 conv layers)
- Large model size and slower inference

4. Custom CNN

- Flexible input size, but **(224, 224, 3)** is standard
- Can design architecture tailored to the dataset
- Ideal for experimenting with smaller, faster models

Evaluation Metrics for Both Datasets:

- **Accuracy**: General performance
- **Precision & Recall**: Class-specific performance
- **F1-Score**: Harmonic mean of precision and recall
- **Confusion Matrix**: Error breakdown
- (Optional) **ROC-AUC** for binary classifiers

6. Exploratory Data Analysis (EDA)

Summary of Processed Data

- **Final survey dataset:** 484 male respondents, with cleaned and encoded features **+ additional records from the Google Form questionnaire.**
- **Final image dataset:** 2,075 scalp images categorized into 6 baldness stages (Stage 1 missing, but we will treat stage 1, and 2 as one stage since there is little difference between both, and the model is intended to assist people who have already lost their hair).

Key Visualizations

The following charts and graphs present the values included in the final dataset after merging the records from the Google Forms responses.

Scalp image dataset (Roboflow)

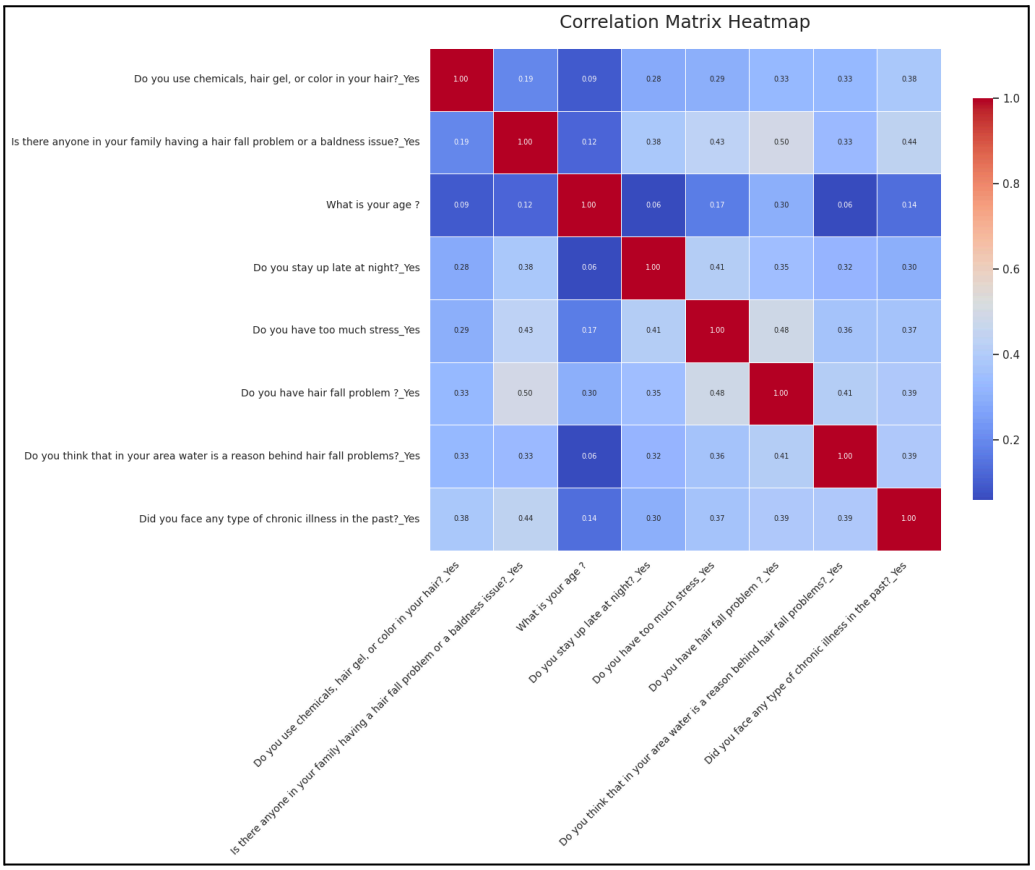
(See angles count graph in 2. Cleaning Data)

- The table below shows the number of images for each male pattern baldness stage in the train and test dataset files.

Pattern baldness stage	Test	Train
2	22	351
3	30	529
4	22	352
5	14	248
6	8	333
7	7	159
	Sum: 103	**Sum:** 1972

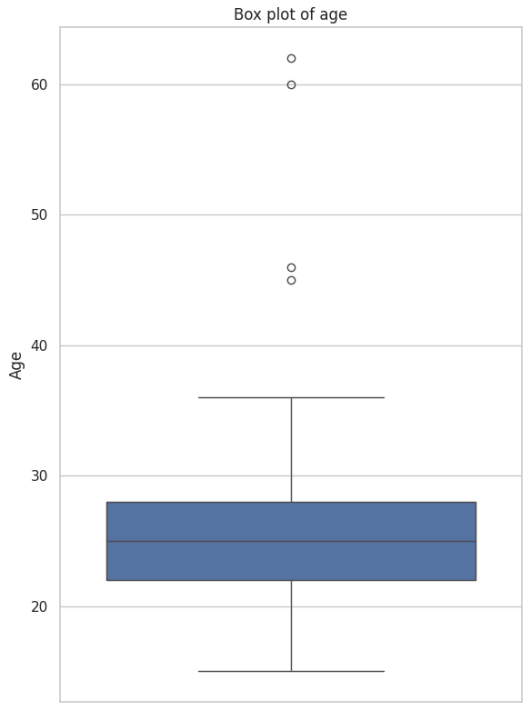
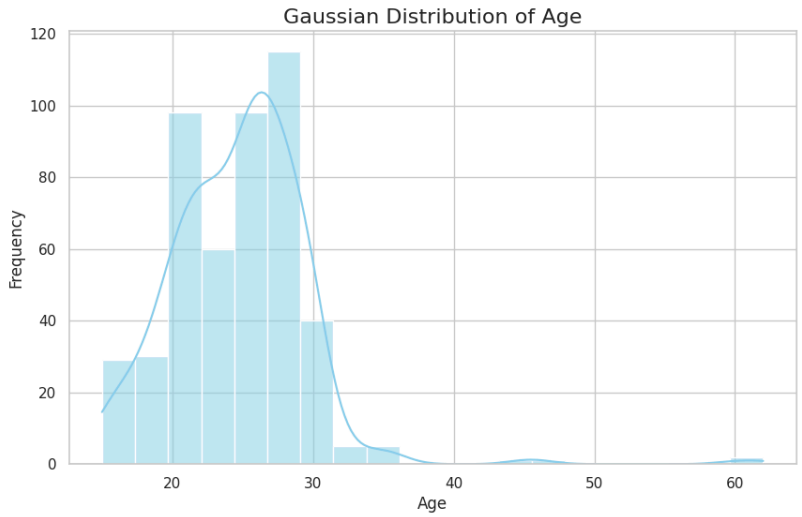
Hair fall causes dataset (Mendeley)

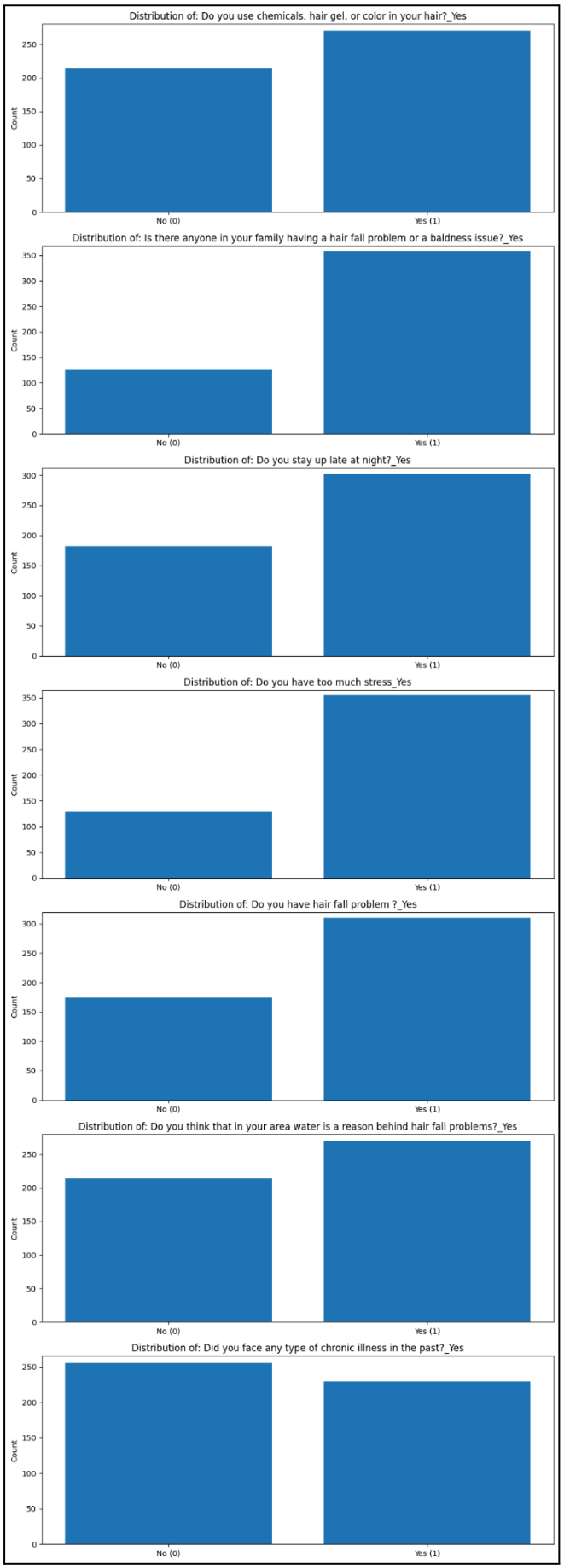
- The following charts and graphs present the values included in the final dataset after merging the records from the Google Forms responses. **Interestingly, after adding the new records, we observed a surprising decrease in the correlation values between the features.**



Feature	Correlation with Target
Do you have hair fall problem ?	1
Is there anyone in your family having a hair fall problem or a baldness issue?	0.497499
Do you have too much stress?	0.483251
Do you think that in your area water is a reason behind hair fall problems?	0.408046
Did you face any type of chronic illness in the past?	0.390889
Do you use chemicals, hair gel, or color in your hair?	0.351738
Do you stay up late at night?	0.330020
What is your age?	0.296860

- Additional visualizations for the final combined Mendeley dataset:

Visualization	Description
	This graph illustrates the distribution of ages **in the final dataset after merging it with the Google Forms responses,** using a box plot. It highlights the median age, interquartile range, and any potential outliers, which provides a clear overview of how the ages are spread within the dataset.
	This graph illustrates the distribution of participants' ages, following a bell-shaped Gaussian curve, with most responses concentrated around the 20s to early 30s.



This graph shows how many participants answered “Yes” or “No” for each question in the dataset, which helps seeing the distribution, as well as which factors were more common across the responses

Next Steps

- **Finalize preprocessing:** Ensure balanced datasets and augment missing scalp images.
- Keep collecting new records for the Google Forms dataset.
- **Begin model development:** Train machine learning models on survey data and deep learning models on scalp images.
- **Test hybrid model:** Integrate both data sources for comprehensive hair loss predictions.