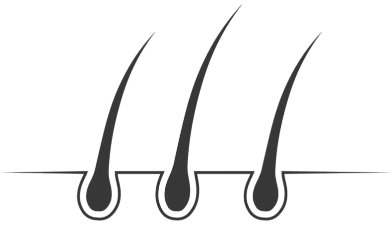


Data Understanding Report

Students Names:

- Alam Bebar - 207415407
- Ryan Thawkho - 305070567

Table of Content

Executive Summary

This report covers the data understanding phase of our study on male pattern baldness progression using a survey dataset from Mendeley and a scalp image dataset from Roboflow.
We checked the data quality, structure, and correlations and found that household water quality, stress levels, and family history had the strongest correlations with baldness in the survey dataset from Mendeley. Some metadata issues, like lighting differences in scalp images, could affect the model's accuracy.
Moving forward, we will focus on data preprocessing, model development, and exploratory data analysis (EDA). Our hybrid model will integrate numerical and image data to generate personalized hair loss prevention recommendations. Additionally, we may incorporate synthetic data and scientific references to enhance the model's accuracy and reliability.

Report Content

1. Data Collection

1.1 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

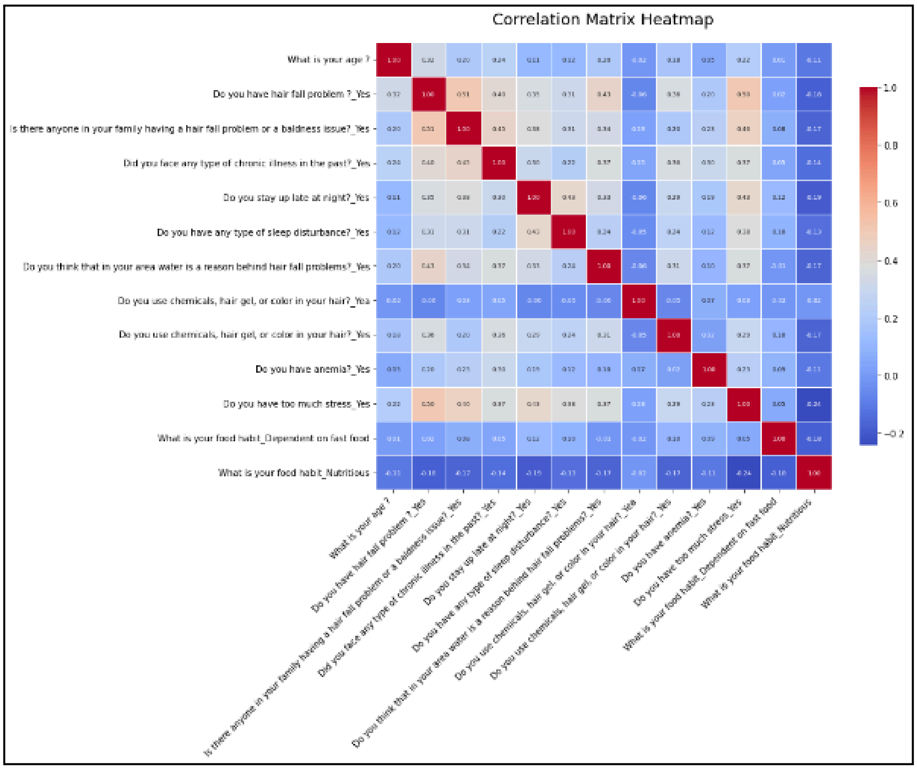
choices that may influence hair loss. The data was collected via Google Forms distributed across social media and online communities. The survey is available for downloading either as a CSV or XML file. The survey dataset is the sole content of the respiratory.

- **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). The dataset will be used for training deep learning models to visually assess hair loss severity.

1.2 Initial data inspection

Upon initial review of the datasets, we considered the following key aspects:

- **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.



*The [Mendeley dataset](#) includes attributes among male individuals such as **stress, family history of hair loss, and household water quality**, which exhibit the highest correlation with hair loss progression and are likely to be valuable for predictive modeling. This correlation matrix was created while using one-hot encoding to transfom the textual data into numerical data.*

- **Irrelevant attributes**: Some survey responses may include unrelated variables that will need to be filtered out. These variables have low correlation levels with our target variable, and one of them is the “Timestamp” which contains the time that the response of the survey was collected.
- **Data sufficiency**: The Mendeley dataset contains 717 responses, 459 of which are males entries. These entries might be limited for reliable machine learning models. Augmentation techniques or additional data sources may be necessary.
- **Merging issues**: Since the two datasets have different formats (numerical vs. image data), we need a structured way to combine them in order to generate recommendation in our hybrid model that we will develop. One approach is to use the numerical dataset only for making recommendations based on the male pattern baldness stage identified from the scalp image analysis.

2. Data Description

2.1 Amount of data

- **Hair Fall Causes Dataset** (Mendeley)
  - **Observations**: 717 (459 males)
- **Scalp Image Dataset** (Roboflow)
  - Multiple images are available for each subject, which capture different angles (front, back, left, right, and top-down). These scalp images are categorized based on male pattern baldness stages. However, it is important to note that **the dataset does not include images for stage 1 of pattern baldness.** (See 4.1 Missing 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

2.2 Value types

The datasets contain a mix of **categorical, numerical, and image data**:

Mendeley dataset

- **Numerical Data**: The following table details all the columns found in the numerical survey dataset. After initially excluding multiple entries and irrelevant columns from the dataset, we plan to apply one-hot encoding to the relevant categorical attributes. Our target variable is “Do you have hair fall problem ?.”

Survey question	Values	Excluded from the dataset?
Timestamp	Date and time	Yes
What is your name ?	Yes/No	Yes
What is your age ?	Numercial	No
What is your gender ?	Male/Female	Only females
Do you have hair fall problem ?	Yes/No	No
Is there anyone in your family having a hair fall problem or a baldness issue?	Yes/No	No
Did you face any type of chronic illness in the past?	Yes/No	No
Do you stay up late at night?	Yes/No	No
Do you have any type of sleep disturbance?	Yes/No	No
Do you think that in your area water is a reason behind hair fall problems?	Yes/No	No
Do you use chemicals, hair gel, or color in your hair?	Yes/No	No
Do you have anemia?	Yes/No	No
Do you have too much stress	Yes/No	No
What is your food habit	Both/Dependent on fast food/Nutritious	No

Roboflow dataset

- **Image Data**: 2075 train and test RGB images of male scalps taken from multiple angles.

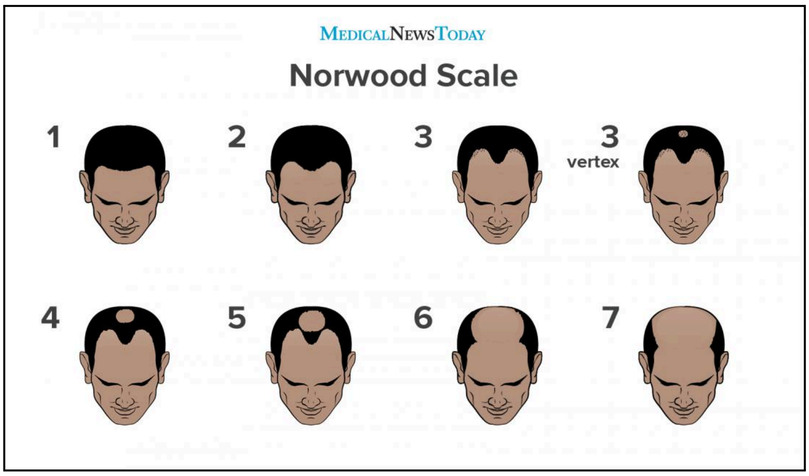
Samples of male scalp images from Roboflow dataset

2.3 Coding schemes

- **Hot-one encoding**: The Mendeley dataset uses categorical values (e.g., "Yes", "No"). These will be encoded as binary values for modeling.
- **Scalp Images**: Labels for hair loss severity may need to be assigned or verified through manual annotation. These labels range from 1-7 (Norwood Scale[1]). It is important to note that in our dataset, Stage 3 encompasses both Norwood Scale Stage 3 and Stage 3 Vertex.
  - The Norwood scale, also known as the Hamilton–Norwood scale, is a classification system used to measure the stages of male pattern baldness. It consists of **seven stages** that describe the severity and pattern of hair loss. Here's an overview of each stage:
    - **Stage 1:** No significant hair loss or recession of the hairline.
    - **Stage 2:** A slight recession of the hairline around the temples, known as an adult or mature hairline.
    - **Stage 3:** The first signs of clinically significant balding appear. The hairline becomes deeply recessed at both temples, resembling an M, U, or V shape. The recessed areas are either bare or sparsely covered with hair.
      - **Stage 3 Vertex:** The hairline remains at Stage 2, but there is significant hair loss on the top of the scalp (the vertex).
    - **Stage 4:** More pronounced recession of the hairline than in Stage 2, with sparse or no hair on the vertex. A band of hair separates the two areas of hair loss, connecting the remaining hair on the sides of the scalp.

[1] https://www.healthline.com/health/norwood-scale

- **Stage 5:** The areas of hair loss are larger than in Stage 4. The band of hair separating them becomes narrower and sparser.
- **Stage 6:** The balding areas at the temples join with the balding area at the vertex. The connecting band of hair across the top of the head is gone or barely noticeable.
- **Stage 7:** The most severe stage of hair loss, leaving only a band of hair around the sides of the head. This remaining hair is usually not dense and may be fine.



Nordwood scale figure depicting the different stages of male pattern baldness accocrding to the Noorwood scale

3. Data Exploration

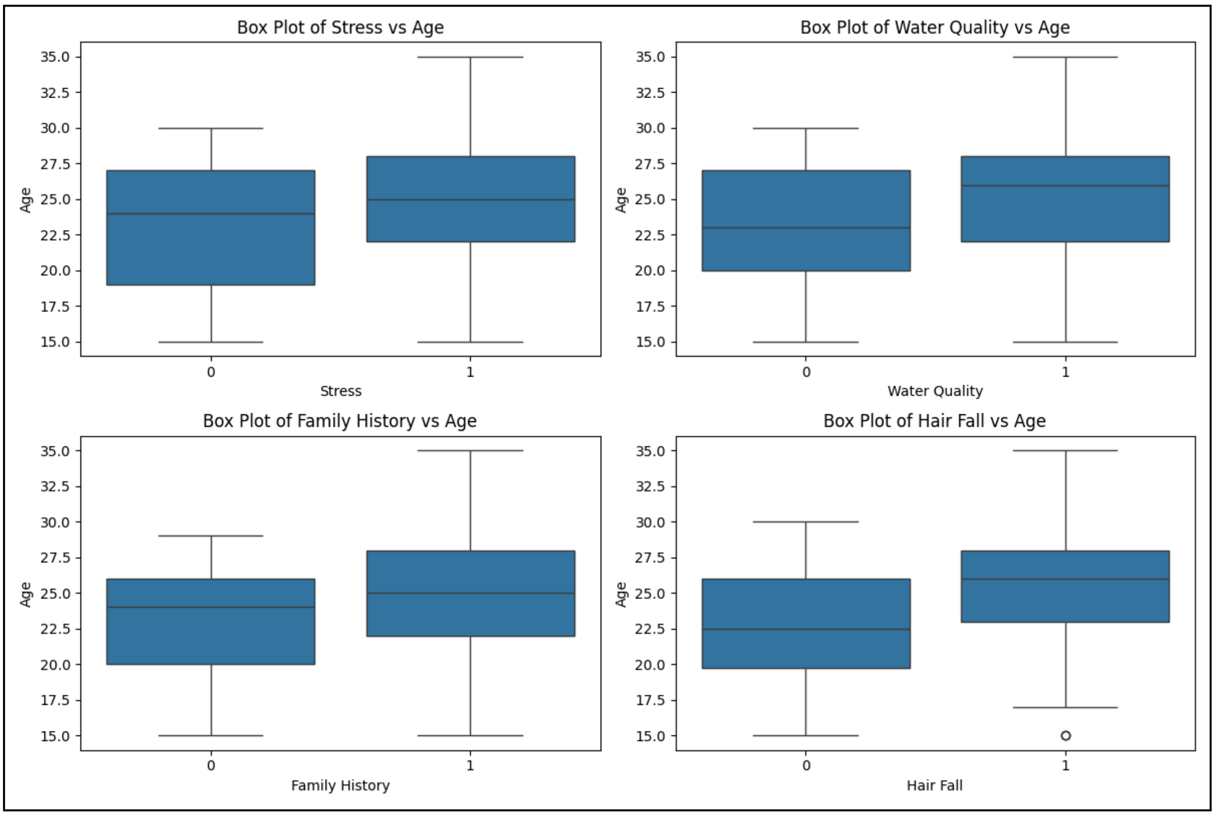
3.1 Hypotheses formation

- Based on the initial data review and existing research, we hypothesize the following relationships between factors contributing to male pattern baldness progression:
  - Family history of hair loss, household water quality, and stress levels may have the strongest influence on hair fall and baldness progression.
  - Age is likely a key factor, with higher baldness stages more common in older individuals.

- ○ Although the survey and image datasets are separate, individuals who report hair fall in the survey dataset may be more likely to be classified in advanced baldness stages in the image dataset.
- ○ External factors such as water quality and stress levels could contribute to hair loss, even in younger individuals with no family history of baldness.
- ○ The image dataset may exhibit a progression pattern, where individuals classified in lower baldness stages share visual similarities with those in adjacent stages.
- ○ While we initially considered dietary habits as a strong predictor of hair loss, our preliminary findings suggest a weaker correlation. However, other lifestyle factors, such as exercise, DHT blocker intake, or supplement use, may have a greater impact and require further analysis.

3.2 Promising attributes for analysis

- We aimed to explore a potential correlation between age and the highly correlated attributes mentioned earlier in the report. As a test case, we created with Python a box plot to visualize the relationships.



Box plots displaying the most correlated features from the report alongside the age column

- Based on our analysis, we have drawn the following key insights.
    - **Family history and hair fall show a clearer trend with age.** People with a family history of hair loss tend to be slightly older, and those experiencing hair fall also have a higher median age. This makes sense since hair loss often increases with age.
    - **Stress and water quality don't seem to have a strong relationship with age.** The distributions look similar across both groups, meaning they might not be the best predictors when using age as a reference.
    - Family history and hair fall could be useful attributes for predicting male pattern baldness, while **stress and water quality might not add much value in this context.**

3.3 Insights from initial data exploration

- The three attributes in the numerical dataset with the highest correlation levels range between 0.43 and 0.51. This suggests that additional methods may be necessary to enhance the correlation for more accurate predictions
- Some survey responses may contain biases due to self-reporting.
- The scalp images will require preprocessing, including cropping and augmentation.
- Since our image dataset lacks images for male pattern baldness Stage 1, we need to source similar images from other databases. These images must match the existing dataset in terms of angles (front, back, left, right, top-down) to ensure consistency in classificatio

3.4 Adjustments to initial hypotheses

After conducting **initial data exploration and correlation analysis**, we refined some of our hypotheses to better align with the dataset's characteristics.

- **Family history, household water quality, and stress levels remain significant factors**, but their correlation with hair loss is **moderate (0.43–0.51)** rather than strongly predictive. Additional data or feature engineering may be required to improve predictive accuracy.
- **Age does influence hair loss progression**, but the correlation is not as strong as initially expected. Some younger individuals also exhibit advanced baldness stages, suggesting that **genetic and environmental factors** play a crucial role.

- **Hair fall reports in the survey dataset** show a clear connection to baldness stages in the image dataset, but **self-reported data may introduce bias**. We may need validation techniques or additional datasets to confirm these findings.
- **Water quality and stress levels may not be direct causes of hair loss** but could act as contributing factors. Further statistical testing is required to determine their exact impact.
- **The image dataset's missing Stage 1 images** create a gap in our analysis. We will need to source additional images or use **data augmentation techniques** to ensure a balanced dataset for training.

4. Data Quality

4.1 Missing data

- It is important to note that **the dataset does not include images for stage 1 of pattern baldness.** To address this gap, we plan to incorporate relevant images from a similar dataset containing these missing classifications.

4.2 Data errors

- Since some stages are missing from our dataset, we might have to use slightly different images from other sources. This could make the model less accurate when predicting those specific stages, as the images won't be an exact match. As a result, the predictions for those stages might not be as reliable.

4.3 Measurement & coding inconsistencies

- Scalp image labeling consistency needs to be verified before training models.
- We need to ensure that the number of images used for training and testing our model is balanced across all male pattern baldness stages. Uneven distribution could lead to biased predictions, where the model performs better on some stages while struggling with others.

4.4 Metadata issues

- The meaning of some survey fields (e.g., hair product usage) needs clarification to ensure correct analysis.
- Some images appear brighter due to variations in the light source, which could affect the model's ability to learn consistent patterns. This may require more advanced preprocessing techniques, such as brightness normalization or adaptive augmentation, to ensure the model trains effectively across all lighting conditions.

5. Next Steps

5.1 Data preprocessing

- Clean and encode survey data.
- Preprocess scalp images (resizing, augmentation).
- Obtain missing scalp image by searching for similar datasets.
- We may plan in later stages to **generate synthetic data** to enhance correlations between features and improve the efficiency of our predictive models, such as exploring SMOTE.[2]

5.2 Exploratory data analysis (EDA)

- Identify feature distributions and correlations.
- Examine whether each stage of male pattern baldness requires unique preventative measures or if a single approach applies across all stages.

5.3 Model development

- **Machine learning models**: Train models (e.g., logistic regression, decision trees) on the survey dataset.
- **Deep learning models**: We'll use convolutional neural networks (CNNs) to analyze scalp images since they're effective at recognizing patterns. We might also use heatmaps to help visualize and understand the results better.
- **Integration**: Develop a **hybrid model** combining numerical and image-based predictions.
  - Our hybrid model will generate recommendations based on the two datasets outputs. Although the two datasets are not directly related, we plan to use the

[2] https://www.machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

image dataset to predict the current male pattern baldness stage, while the survey dataset will provide stage-specific recommendations. These recommendations will be based on our survey data analysis and supplemented with preventative measures. Additionally, we may incorporate further preventative strategies based on scientific literature, even if they are not directly linked to the survey numerical dataset.

- We plan to incorporate other scientifically backed preventative methods to help stabilize male pattern baldness. For each factor, we have provided references to relevant scientific studies, which cover the following treatments and lifestyle factors:
    - Minoxidil[3]
    - Low-Level Laser Therapy (LLLT)[4]
    - DHT Blockers (Finasteride and Dutasteride)[5]
    - Ketoconazole 2% Shampoo[6]
    - Tretinoin[7]
    - Derma Rolling[8]
    - Platelet-Rich Plasma (PRP) Treatments[9]
    - Hair Supplements Intake[10]
    - Protein consumption[11]
    - Smoking[12]
    - Alcohol consumption[13]
    - Sleep quality[14]

[3] https://www.jaad.org/article/S0190-9622(17)30306-7/abstract

[4] https://www.jaad.org/article/S0190-9622(17)30306-7/abstract

[5] https://www.jaad.org/article/S0190-9622(17)30306-7/abstract

[6] https://www.jaad.org/article/S0190-9622(17)30306-7/abstract

[7] https://www.jaad.org/article/S0190-9622(17)30306-7/abstract

[8] https://www.jaad.org/article/S0190-9622(17)30306-7/abstract

[9] https://www.jaad.org/article/S0190-9622(17)30306-7/abstract

[10] https://symbiosisonlinepublishing.com/nutritionalhealth-foodscience/nutritionalhealth-foodscience132.pdf

[11] https://symbiosisonlinepublishing.com/nutritionalhealth-foodscience/nutritionalhealth-foodscience132.pdf

[12] https://symbiosisonlinepublishing.com/nutritionalhealth-foodscience/nutritionalhealth-foodscience132.pdf

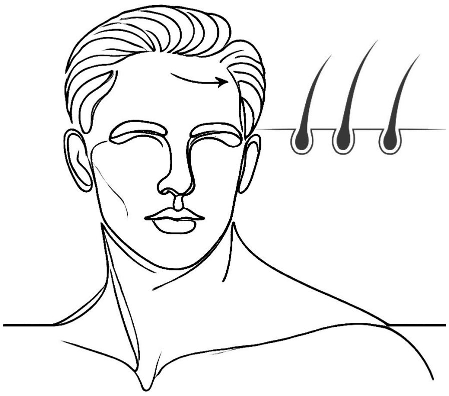
[13] https://symbiosisonlinepublishing.com/nutritionalhealth-foodscience/nutritionalhealth-foodscience132.pdf

[14] https://healthcentre.nz/the-science-of-hair-loss-exploring-the-latest-research/

- Stress managment[15]
- Genetic predisposition to hair loss[16]

5.4 Future improvements

- **Synthetic data generation** to improve correlations and address data imbalances.
- **Additional datasets** to enhance predictive accuracy.
- **Validation** using real-world hair loss progression data.



[15] https://healthcentre.nz/the-science-of-hair-loss-exploring-the-latest-research/
[16] https://healthcentre.nz/the-science-of-hair-loss-exploring-the-latest-research/