Exploratory Data Analysis (EDA) for HAM10000 Skin Cancer Dataset

Step 1&2: Install and import Required Libraries

Before starting the analysis, ensure that necessary Python libraries such as **pandas, numpy, matplotlib, seaborn, scipy** and **sci-kit learn** are installed and imported.

Step 3: Load the Dataset

- Download the dataset from **Kaggle** (HAM10000 Dataset).
- Load the metadata CSV file(**HAM10000_metadata.csv**) into a **pandas DataFrame** for analysis.

NOTE: Set the correct file path for the dataset:

```
data_path = "/content/HAM10000_metadata.csv"
```

Ensure that the path is updated to match the actual location of the file on your system.

Step 4: Basic Exploration

- Display basic information about the dataset (e.g., column names, data types, and missing values).

Step 5: Data Preprocessing

- Handle missing values if present.
- Encode categorical variables if needed.

Step 6: Exploratory Data Analysis

- **Age Distribution**: Plot a histogram to analyze the age range of patients.
- **Lesion Type Distribution**: Visualize the frequency of different lesion types.
- **Gender Distribution**: Compare the number of male and female patients.

Step 7: Statistical Tests

- Compare the **age distribution** between benign and malignant lesions using statistical tests like **t-tests**.

Step 8: Hypothesis Evaluation

- Perform a **Chi-Square test** to analyze relationships between **lesion type and gender/localization**.

Step 9: Observations & Summary

- Highlight potential **challenges** (e.g., class imbalance, data bias).
- Suggest next steps for model building or further feature engineering.