**Micro Project Report**

On

**Beauty and wellness**

Submitted in Partial Fulfilment of

Award of

**BACHELOR OF TECHNOLOGY**

**In**

**Computer Science and Engineering**

**(Data Analytics)**

By

2023BCSE07AED464 – Nayana S P

2023BCSE07AED510 – Tejaswini S

2023BCSE07AED525 – Meghana M

2023BCSE07AED526 – Shreya Reddy S

Under the

Supervision of

Dr. K Sasi Kala Rani, Professor,

Department of CSE



**ALLIANCE SCHOOL OF ADVANCED COMPUTING**

**ALLIANCE UNIVERSITY**

**BENGALURU**

APRIL 2025



**Department of computer Science and Engineering**

**ALLIANCE SCHOOL OF ADVANCED COMPUTING**

**CERTIFICATE**

This is to certify that the Micro Project work entitled “Beauty and wellness” submitted by Nayana S P[2023BCSE07AED464], Tejaswini S[2023BCSE07AED510], Meghana M[2023BCSE07AED525] and Shreya Reddy S[2023BCSE07AED526] in partial fulfillment for the award of the degree of Bachelor of Technology DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING (DATA ANALYTICS) of Alliance University, is a Bonafide work accomplished under my supervision and guidance during the academic year 2024-2025. This thesis report embodies the results of original work and studies conducted by students and the contents do not form the basis for the award of any other degree to the candidate or anybody else.

**Dr.K.Sasi Kala Rani**

**(Supervisor)**

**Table of content:**

|  |  |  |
| --- | --- | --- |
| Section No. | Title | Page No. |
| 1. | Research Goal | 4 |
| 2. | Data selection | 4 |
| 3. | Data Preparation | 4-5 |
| 4. | Exploratory data analysis | 5-7 |
| 5. | Visualizations | 7-11 |
| 6. | Statistical Analysis | 11-13 |
| 7. | Clustering Analysis | 13-14 |
| 8. | Predictive Modeling | 14-15 |
| 9. | Model Evaluation | 15-16 |
| 10. | Dashboard and Conclusion | 16-19 |

**1.Defining the research goal: project charter.**

This project mainly focuses on exploring and analysing data of the users from the beauty and wellness sector. The *beauty\_wellness\_200\_customer1.csv* dataset contains the data about200 customers.

The major goal of this project is to perform a data analysis to extract the valuable insights of user characteristics, preference and trends. Using some visualizations and statistical techniques we can easily understand the key customer segments, can analyse service usage patterns and understand how demographic influences on purchasing behaviour.

**2. Choose a relevant dataset.**

For this project, we considered a dataset named as *beauty\_wellness\_200\_customer1.csv,* which covers the attributes like user ID, age, skin type, face shape, skin concern, recommender routine and products suggested. This dataset enables us to explore relationship between user characteristics and suggested skin treatments. Here the csv file is getting loaded.

**Code:**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

Import seaborn as sns

from sklearn.preprocessing import LabelEncoder

from sklearn.cluster import KMeans

from sklearn.linear\_model import LinearRegression

from sklearn.metrics import r2\_score

df = pd.read\_csv(“beauty\_wellness\_200\_customers1.csv”)

**3. Data preparation: Data cleaning using Pandas and NumPy like missing values, Detect and remove outliers.**

Data preparation is the process of cleaning, transforming, and organizing raw data into a usable format for analysis.

Here are some steps involved in data preparation:

1.Data Collection: Collecting raw data form different sources.

2.Data Cleaning: Handling missing values, fixing errors, removing duplicates and filtering irrelevant data all these are involved in this step.

3.Data Transformation: Normalization, encoding, parsing dates and aggregation takes place.

4.Data integration: Combining data from multiple sources into a single dataset.

5.Data reduction: Keeping only the most relevant columns and using dimensionality reduction to reduce data complexity by the techniques like PCA.

6.Data validation: Checking the data is accurate and ready for analysis or not.

Here, we are dropping the rows with missing values and converted Age column to int using the below provided code. Now, this Dataset contains only clean rows.

**Code:**

df.dropna(inplace=True)

df[;Age’]=df[‘Age’].astype(int)

**4. Exploratory Data Analysis (EDA)**

This is a process of analysing datasets to summarize their main characteristics, using visual methods. This is about knowing your data before doing actual advanced data modelling or predictions. Purpose of this analysis is to understand the structure and quality of data, identifying patterns and relationships between the attributes, finding outliers and missing values.

Descriptive statistics: includes metrices like count, mean, std, min, max for numerical columns like Age.

Visualization: Age distribution histogram shows how users age is distributed with a KED curve. Skin type countplot shows frequency of each skin type in the dataset.

**Code:**

print(df.describe())

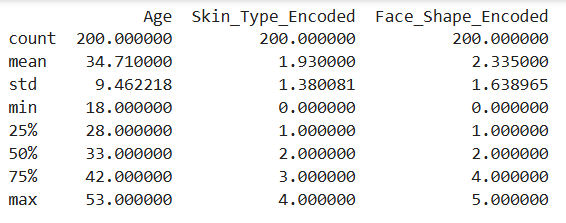
sns.histplot(df[‘Age’],kde=True).set(title=’Age Distribution’)

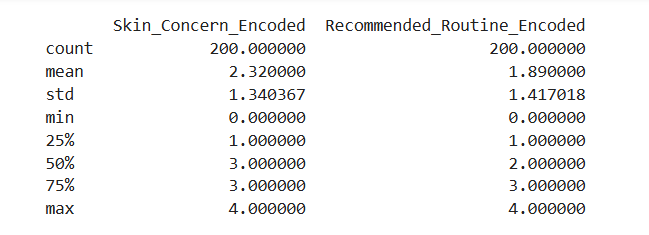
plt.show()

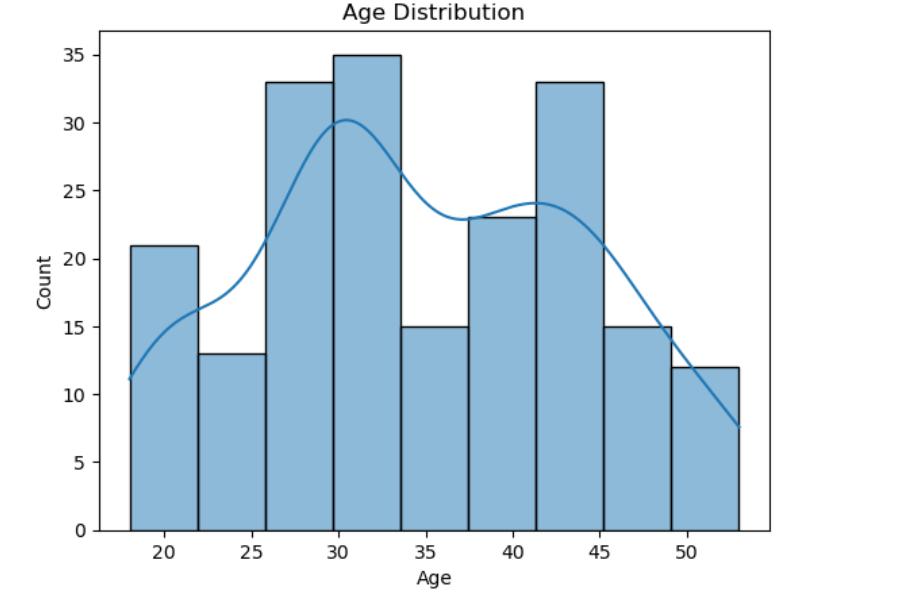
sns.countplot(data=df,x=’Skin\_Type’).set(title=’Skin\_Type Distribution’)

plt.show()

**Output:**







A graph of different sizes of blue bars

AI-generated content may be incorrect.

**5. Visualization:**

Data visualization is graphical representation of the data using graphs, charts and plots to make the data understandable, accessible and visually appealing. There are some common types of visualizations like Bar chart, line chart, histogram, box plot, scatter plot etc.

Here, in this we created a line plot that simulates a sales trend over age. Scatter plot which shows a relationship between simulated price and demand. To display the order frequency by age group histogram is used. Heatmap visualizes the frequency of skin type vs skin concern using a pivot table.

**Code:**

#1. Line Plot: Simulate sales trend over time using Age as a proxy for time (for demo purposes)

df\_sorted = df.sort\_values('Age')

# df\_sorted['Simulated\_Sales'] = np.random.randint(10, 100, size=len(df\_sorted)) plt.figure(figsize=(10, 4))

# plt.plot(df\_sorted['Age'], df\_sorted['Simulated\_Sales'], marker='o')

# plt.title("Simulated Sales Trend Over Age")

# plt.xlabel("Age")

# plt.ylabel("Simulated Sales")

# plt.grid(True)

# plt.tight\_layout()

plt.show()

#2. Scatter Plot: Simulate price vs demand correlation

Df[‘Simulated\_Price’] = np.ramdom.uniform(20,100,size=len(df))

Df[‘Simulated\_Demand’]=200 – df[‘Simulated\_Price’] + np.random.normal(0,10, len(df))

#Inverse relationship

plt.figure(figsize=(8, 5))

sns.scatterplot(x='Simulated\_Price', y='Simulated\_Demand', data=df) plt.title("Price vs Demand (Simulated)")

plt.xlabel("Price")

plt.ylabel("Demand")

plt.grid(True)

plt.tight\_layout()

plt.show()

# 3. Histogram: Order frequency distribution (simulate with Age groups)

df['Age\_Group']=pd.cut(df['Age'],bins=[18, 25, 35, 45, 60],labels=["18-25", "26-35", "36-45","46-60"])

order\_frequency = df['Age\_Group'].value\_counts().sort\_index()

order\_frequency.plot(kind='bar', color='skyblue')

plt.title("Simulated Order Frequency by Age Group")

plt.xlabel("Age Group")

plt.ylabel("Frequency")

plt.tight\_layout()

plt.show()

# 4. Pivot Table & Heatmap: Product category vs Skin Concern (Encoded)

pivot\_table = pd.pivot\_table(df, values='Age', index='Skin\_Type\_Encoded', columns='Skin\_Concern\_Encoded', aggfunc='count')

plt.figure(figsize=(8, 5))

sns.heatmap(pivot\_table, annot=True, cmap='YlGnBu')

plt.title("Heatmap: Skin Type vs Skin Concern Frequency")

plt.tight\_layout()

plt.show()

# 5. Geographic Plot: Simulate geographic data (Not applicable in this dataset)

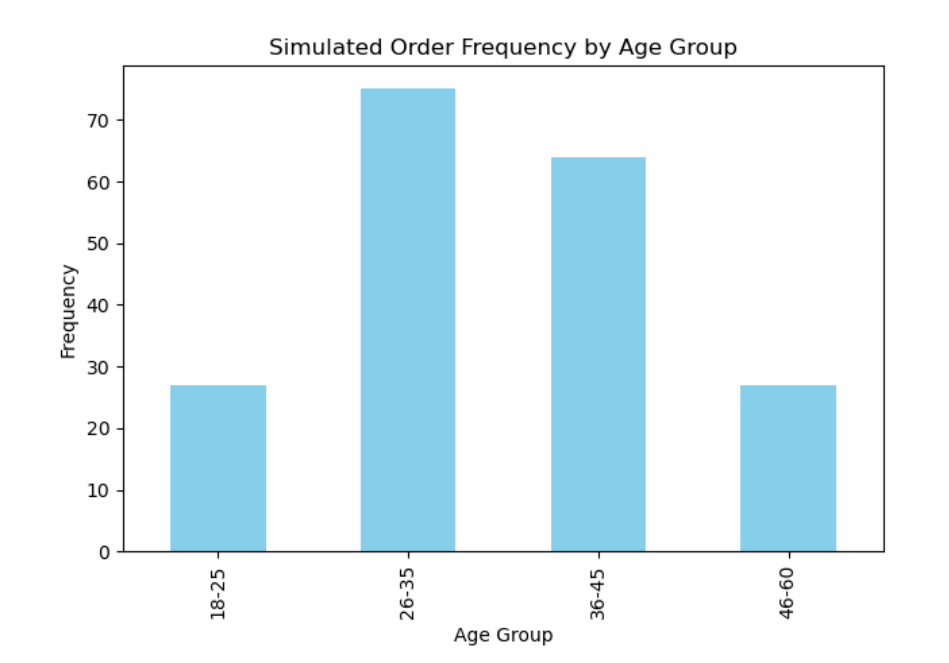
plot\_successful = True

plot\_successful

**Output:**

A graph showing the number of sales

AI-generated content may be incorrect.

A graph with blue dots

AI-generated content may be incorrect.

A black and white text

AI-generated content may be incorrect.A chart of different colors

AI-generated content may be incorrect.

**6. Statistical analysis: Corelation analysis and visualize relationships using scatter plots and regression lines.**

Here, corelation matrix is created that tells how age is corelated with a the skin type and concern.

Scatter plots with regression lines:

Age vs Skin\_concern\_enacodded -shows how does the skin concern changes with age.

Skin\_type\_encoded vs Skin\_concern\_encodded – this shows how specific skin types associated with specific concern.

**Code:**

#Correlation Matrix (Statistical Analysis)

correlation\_matrix = df[['Age', 'Skin\_Type\_Encoded', 'Skin\_Concern\_Encoded']].corr()

#Heatmap Visualization

plt.figure(figsize=(6, 4))

sns.heatmap(correlation\_matrix, annot=True, cmap='coolwarm', fmt=".2f") plt.title("Correlation Matrix: Age, Skin Type, and Skin Concern")

plt.show()

#Scatter Plot + Regression Line:Age vs Skin Concern

sns.lmplot(x='Age', y='Skin\_Concern\_Encoded', data=df, height=5, aspect=1.2, line\_kws={'color': 'red'})

plt.title("Age vs Skin Concern (Encoded) with Regression Line")

plt.show()

#Scatter Plot + Regression Line: Skin Type vs Skin Concern

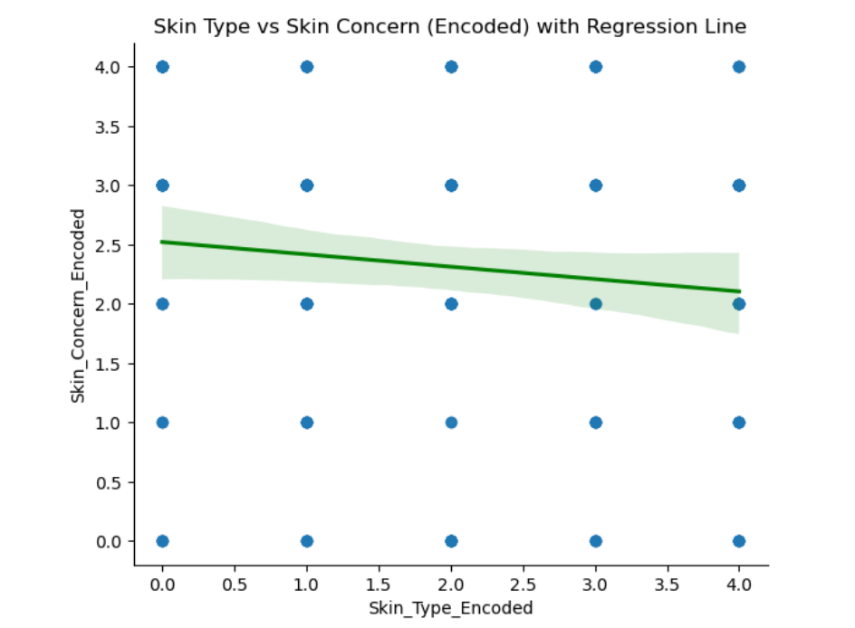
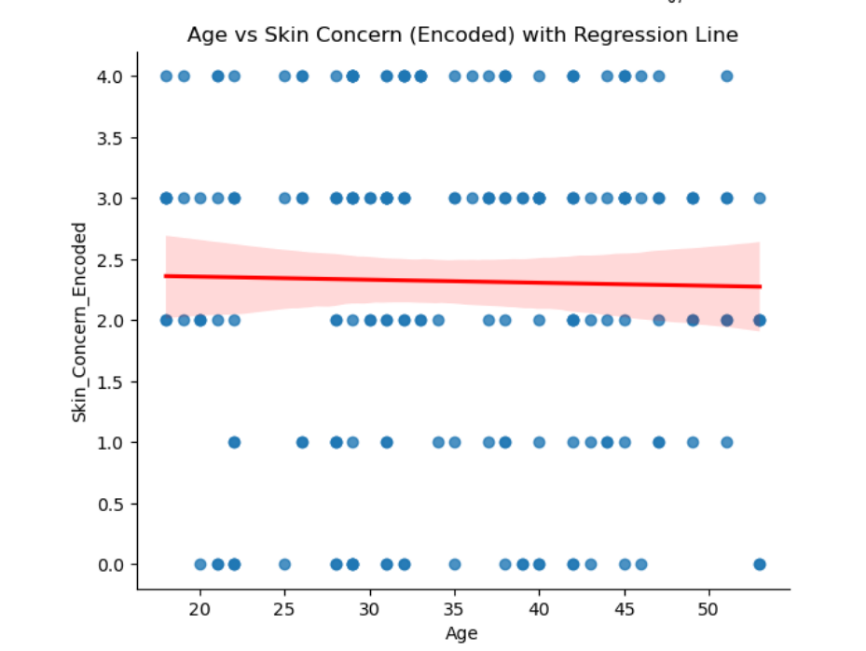
sns.lmplot(x='Skin\_Type\_Encoded', y='Skin\_Concern\_Encoded', data=df, height=5, aspect=1.2, line\_kws={'color': 'green'})

.title("Skin Type vs Skin Concern (Encoded) with Regression Line")

plt.show()

A graph of different colored squares

AI-generated content may be incorrect.**Output:**



**7. Clusters:**

Clusters refers to groups of similar data points that are grouped together based on certain characteristics or patterns.

Here, KMeans clustering to group users into 3 groups. This adds new column cluster to label each user with a cluster number (0,1 or 2). This helps in personalize product recommendations, identifying customer segments and targeted marketing strategies.

**Code:**

X\_cluster = df[[‘Age’, ‘Skin\_Type\_Encodes’, ‘Skin\_Concern\_Encoded’]]

kmeans = KMeans(n\_clusters=3, random\_state=0)

df[‘Cluster’] = kmeans.fit\_predict(X\_cluster)

sns.scatterplot(data = df, x = ‘Age’, y = ‘Skin\_Concern\_Encoded’,

hue= ‘Cluster’, palette= ‘Set2’)

plt.title(“Customer Clusters”)

plt.show()

**Output:**

A chart with dots and numbers

AI-generated content may be incorrect.

**8. Predictive Modeling:**

This is the statistical method used to model the relationship between 2 variables one independent and one dependent variable.

Multiple regression is the extension of the linear regression where we use 2 or more independent variables to predict single dependent variable.

Here, prediction of skin concern based on the attributes age and skin type which will accounts for more complexity and potential interaction between the attributes.

Simple regression R^2 shows the performance of the model using only single feature.

Multiple regression R^2 usually provides a better score using multiple features.

**Code:**

X\_simple = df[[‘Age’]]

model\_simple = LinearRegression()

model\_simple.fit(X\_simple,y)

y\_pred\_simple = model\_simple.predict(X\_simple)

r2\_simple = r2\_score(y, y\_pred\_simple)

print(f“Simple Linear Regression R^2 Score: {r2\_simple:.3f}”)

X\_multi = df[[‘Age’, ‘Skin\_Type\_Encoded’]]

model\_multi = LinearRegression()

model\_multi.fit(X\_multi, y)

y\_pred\_multi = model\_multi.predict(X\_multi)

r2.multi = r2\_score(y, y\_pred\_multi)

print(f“Multiple Linaer Regression R^2 Score: {r2\_multi:.3f}”)

**Output:**

****

**9. Evaluate models using R^2, standard error of estimate. Check for the regression towards mean.**

R^2 score measures how well the model explains the variability of the response variable. Values will range from 0 to 1, if the value is closer to 1 means that better prediction.

Standard error estimation represents that the average distance that the observed values fall from the regression line. Smaller SEE means that more accurate prediction.

Regression towards the mean is where extreme predictions trend to move closer to the average in future predictions. We can access this by comparing mean of actual values vs predicted values. If predictions are closer to the mean, that indicates regression towards the mean.

**Code:**

# 1. R^2 Score (how well the model fits)

r2 = r2\_score(y, y\_pred)

print(f“ R^2 Score: {r2:.3f}”)

# 2. Standard Error of Estimate (SEE)

mse = mean\_squared\_error(y, y\_pred)

see = np.sqrt(mse)

print(f“Standard Error of Estimate (SEE): {see:.3f}”)

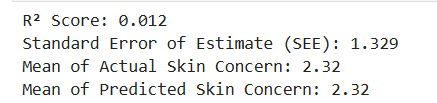
# 3. Regression Toward the Mean (basic check)

# Compare mean of actual vs. Predicted

print(f“Mean of Actual Skin Concern: {np.mean(y):.2f}”)

print(f“Mean of Predicted Skin Concern: {np.mean(y\_pred):.2f}”)

**Output:**



**10. Dashboard (Power BI/Tableau/Stream lit):**

The file is imported into Power BI to create a comprehensive dashboard named as “beauty and wellness Dashboard”. The dashboard provides interactive visualization that includes customer segmentation based on the clustering and Filtering by age groups, skin type, or concern for targeted insights. This visualization toll empowers business stakeholders to explore key insights, sport trends, and make data driven decisions in real time.

link for the Power BI - [Beauty\_and\_wellness\_Dashboard.pbix.pbix](https://alliancebschool-my.sharepoint.com/:u:/g/personal/sshreyabtech23_ced_alliance_edu_in/ETiG_7wg8HBLq3VKefHrNf8B6wXaK_vFsSaAQ8WYrtDqlA?e=h0g4rX)

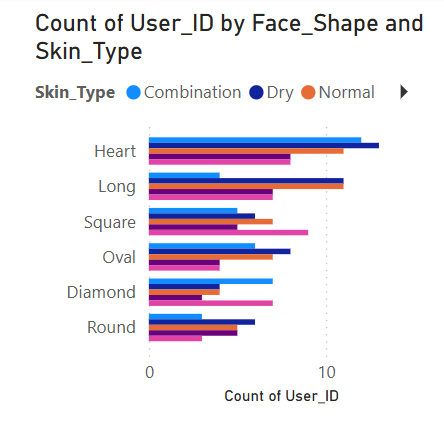
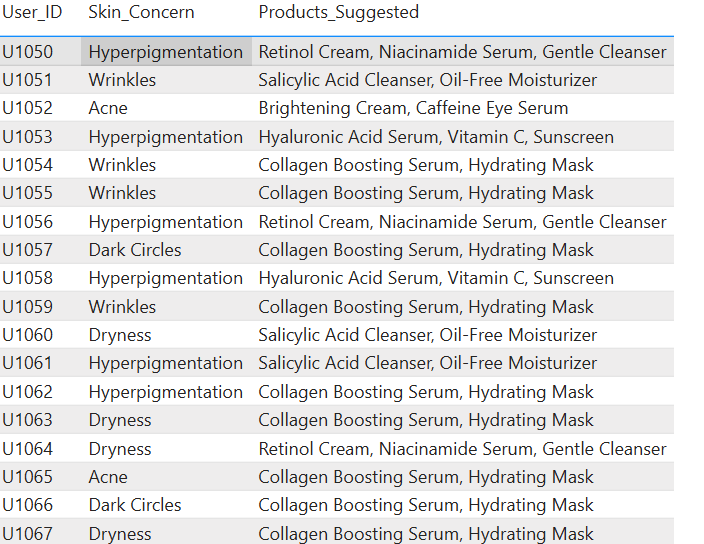
**Code:**

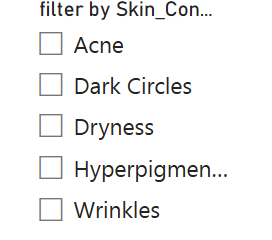
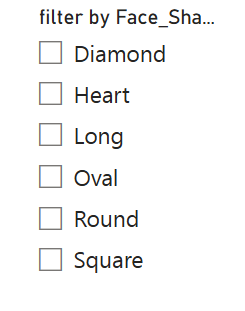
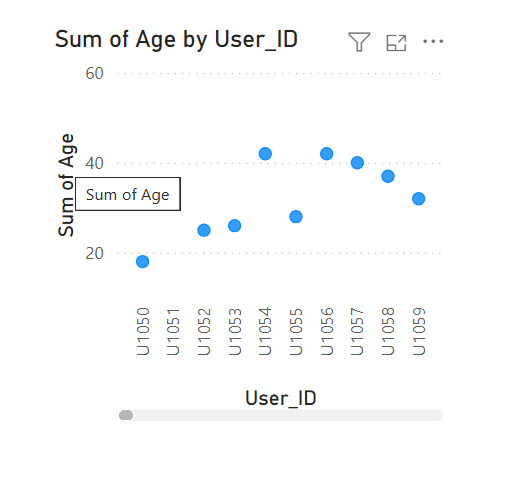
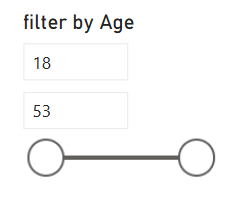
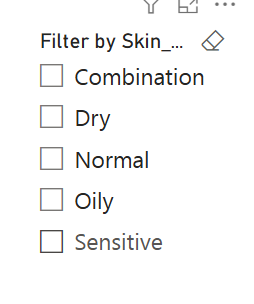
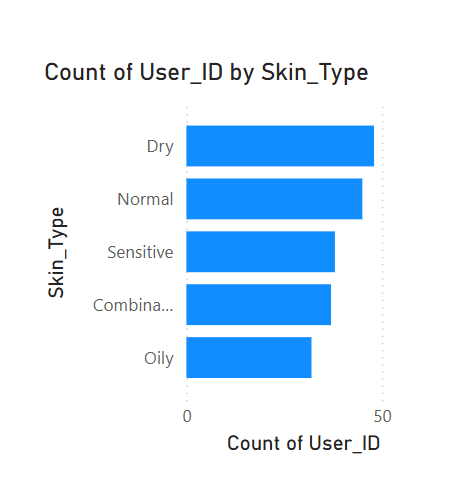
Df.to\_csv(“cleaned\_beauty\_data\_for\_dashboard.csv”,index=False)

print(“Cleaned dataset exported for Power BI/Tableau.”)

A pie chart with numbers and text

AI-generated content may be incorrect.**Snapshots:**





**Conclusion:**

In this micro project, we explored, analysed, and modelled a user dataset from the file beauty and wellness, which focuses on understanding user characters and behaviours.

This project demonstrates how data science techniques from preprocessing and visualization to clustering and regression helps us to understand the data set beauty and wellness clearly by analysing and visualizing the csv file using attributes provided.

With all these insights that we got from this will help us in businesses that can personalize offerings, optimize marketing strategies, and ultimately improve customer satisfaction and retention.

To conclude from the dataset of our csv we can conclude that “Hyperpigmentation is one of the top skin concern (65 customers) and dry is the most common skin type (48 customer)”.