Bumble Profile Data Analysis Report

1. Problem Statement, Goals & Objectives

Problem Statement:

Bumble, a leading dating platform, collects user profile data to enhance matchmaking, optimise engagement, and refine marketing strategies. The challenge is to extract meaningful insights from this data to improve user experience and drive platform growth.

Analyst Role:

As a data analyst at Bumble, Our job is to examine user profile data to find patterns in demographics, preferences, and behaviour. Our insights will help the product and marketing teams make informed decisions to improve the user experience and grow the platform.

Goals & Objectives:

- **Understand user demographics, preferences, and behaviours** to optimize matchmaking algorithms.
- **Identify key lifestyle and financial factors** influencing user engagement.
- **Analyse location and language trends** to improve targeted marketing efforts.
- **Provide data-driven recommendations** to enhance the platform's features and user satisfaction.

2. Dataset Overview

2.1 Dataset Summary

Dataset Source: Provided as part of the Data Analyst course Milestone Project by NextLeap

Dataset Name: Bumble User Profiles

Number of Rows: 59,946

Number of Columns: 17

Categorical Columns: 13

Numerical Columns: 3 [age, height, income]

DateTime Column: 1 [last_online]

Description: This dataset contains user profiles from Bumble, covering demographics, lifestyle preferences, and activity data. The goal is to analyse user behaviour, identify trends, and provide insights to improve matchmaking and user engagement.

2.2 Segregation of Columns:

1. Demographics

age: Age of the user.

status: Relationship status (e.g., single, married, seeing someone).

gender: Gender of the user (e.g., m, f).

2. Physical Attributes

body_type: Descriptions of physical appearance (e.g., athletic, curvy, thin).

height: Height of the user (in inches).

3. Lifestyle Preferences

Diet: Dietary preferences (e.g., vegetarian, vegan, anything).

drinks: Drinking habits (e.g., socially, often).

4. Educational and Financial Details

education: Education level (e.g., college, masters).

income: User-reported annual income.

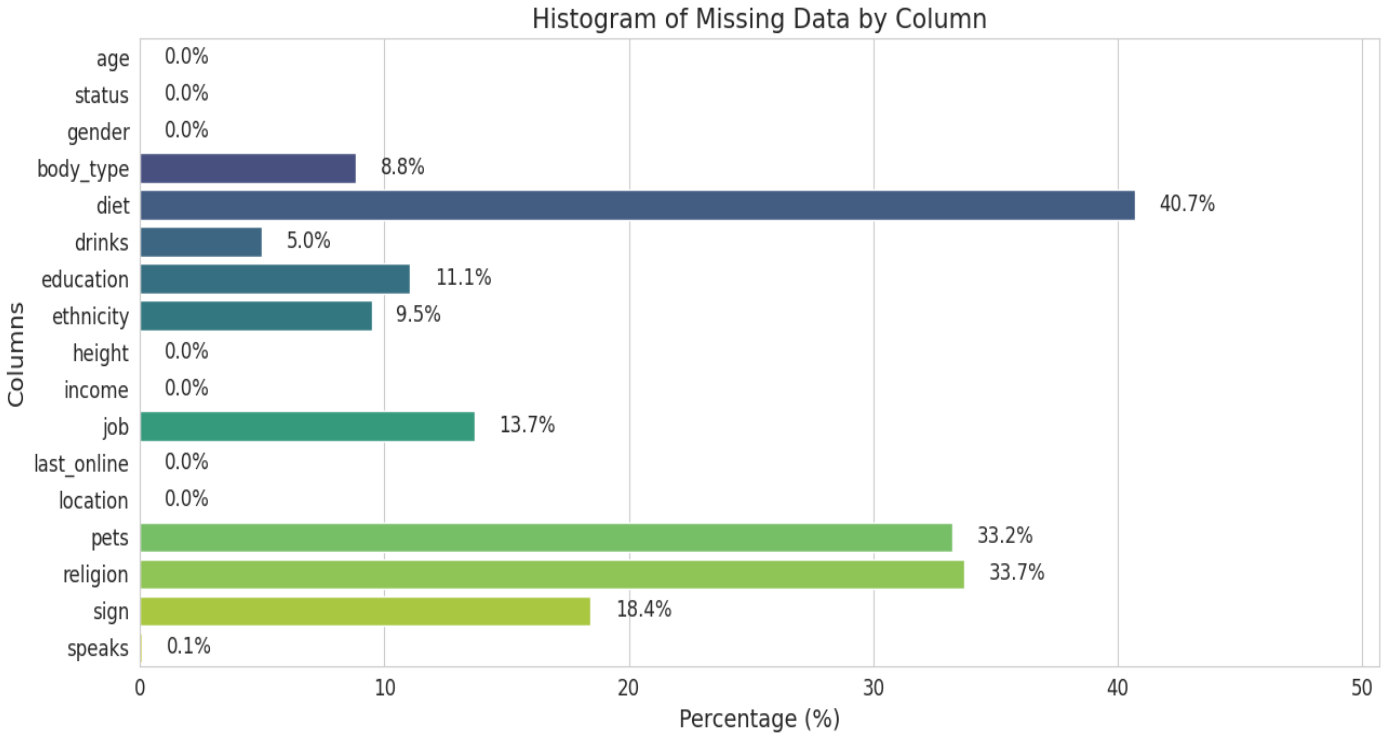
5. Location and Activity

location: City and state where the user resides.

last_online: Date and time when the user was last active.

3. Importing, Loading & Initial Exploration

- **Imported** necessary libraries.
- **Loaded dataset** from the drive.
- Checked dataset **shape,** and **column names,** and **displayed sample rows.**
- Retrieved **dataset information**, **missing values summary**, and **unique value counts**.
- Provided **statistical summary** for age, height, and income.



4. Data Preparation

4.1 Backup Creation

4.2 Handled duplicate rows

4.3 Fixed Incorrect Data Types

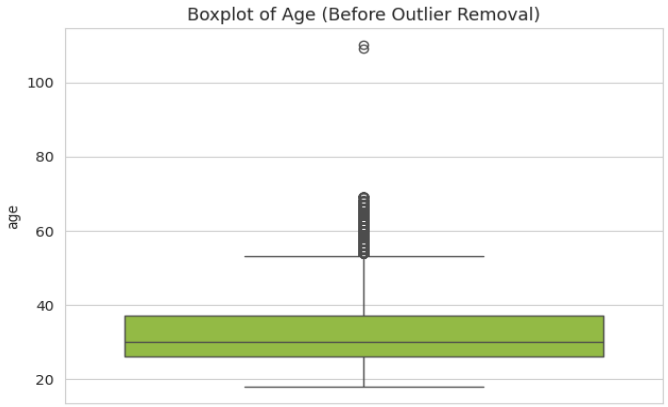
- **Converted** 'last_online' **to datetime format**.
- **Ensured 'income' is numeric** (replaced -1 values with 0).
- **Converted height to centimetres**.

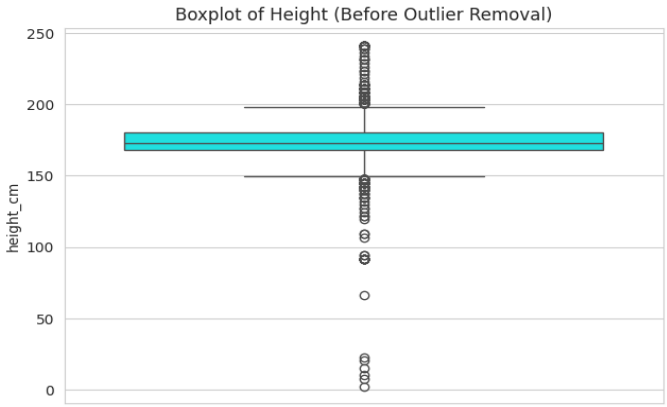
4.4 Missing Value Imputation

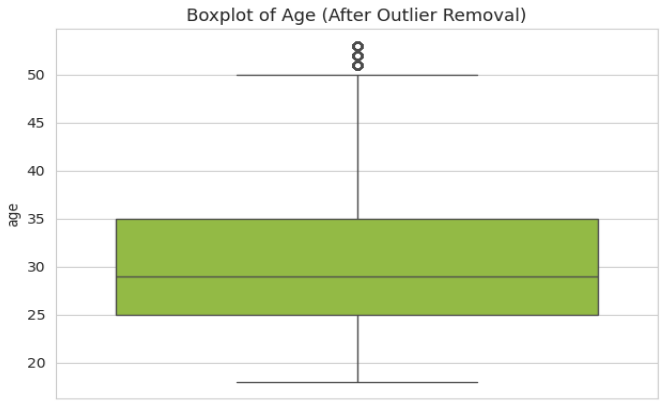
- **Imputed height** using the **gender median**.
- **Replaced missing categorical values** with **"unknown"**.

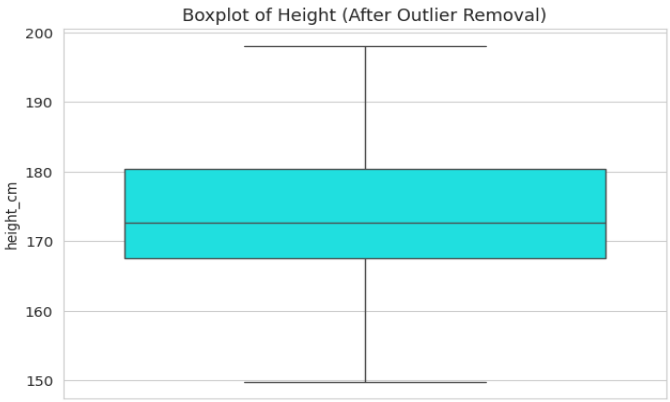
4.5 **Outliers Removal**

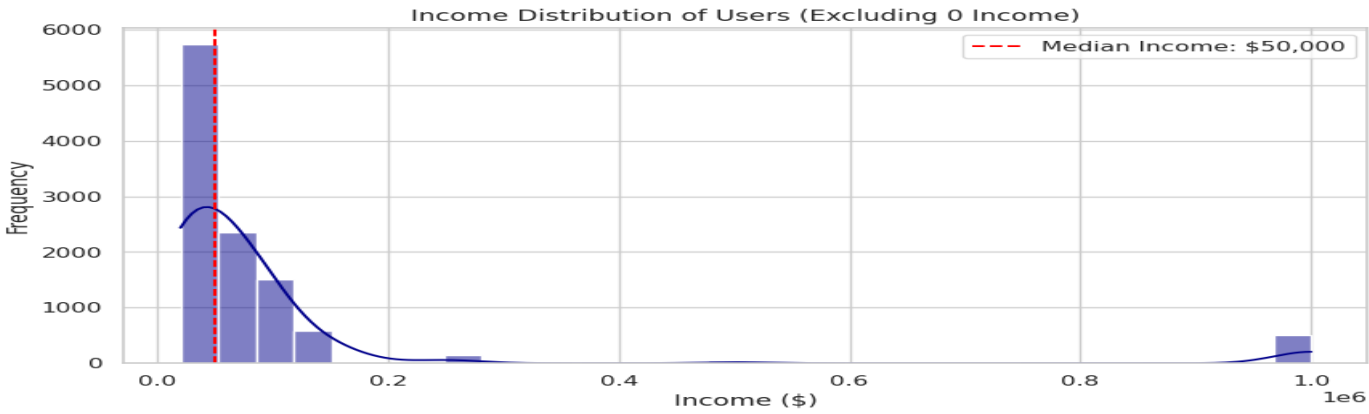
- **Identified outliers** in **age & height** using **the Interquartile Range (IQR)** method.
- **Visualised boxplots before and after outlier removal** to confirm data integrity.
- **Visualised income distribution** chart and **removed outliers manually.**

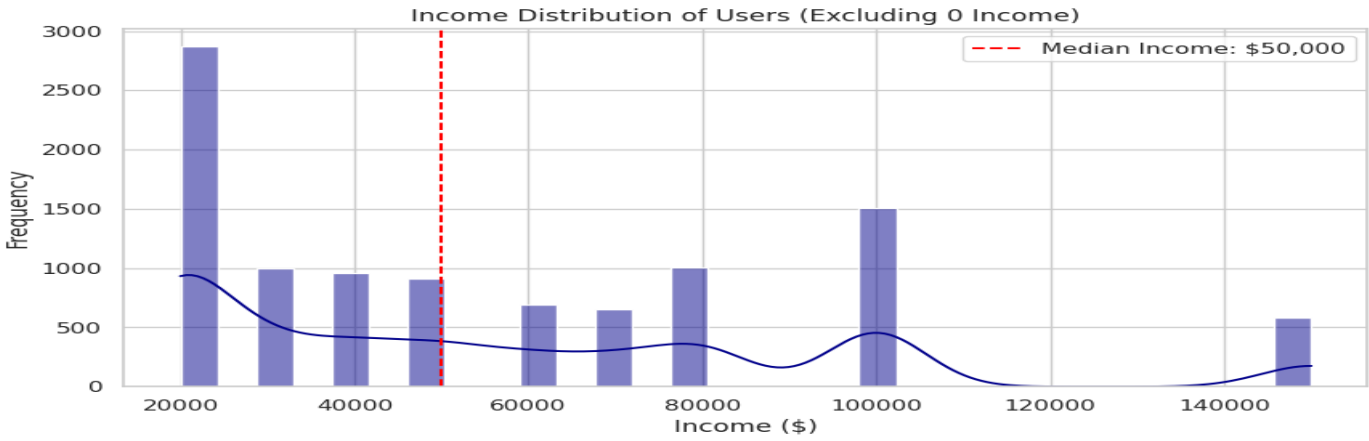












4.6 **Feature Engineering**

- Separate **city & state** from **location**.
- Created **zodiac & zodiac_imp columns** from **sign**.
- Extracted **Primary Language** from the speaks column.
- **Generated primary_language** from **speaks**.
- **Created last_active** from **last_online**.
- **Binned Age into Groups**: 18-25, 26-35, 36-45, 46-60, 60+.
- **Categorized Income Levels**: Low, Medium, High, Very High.

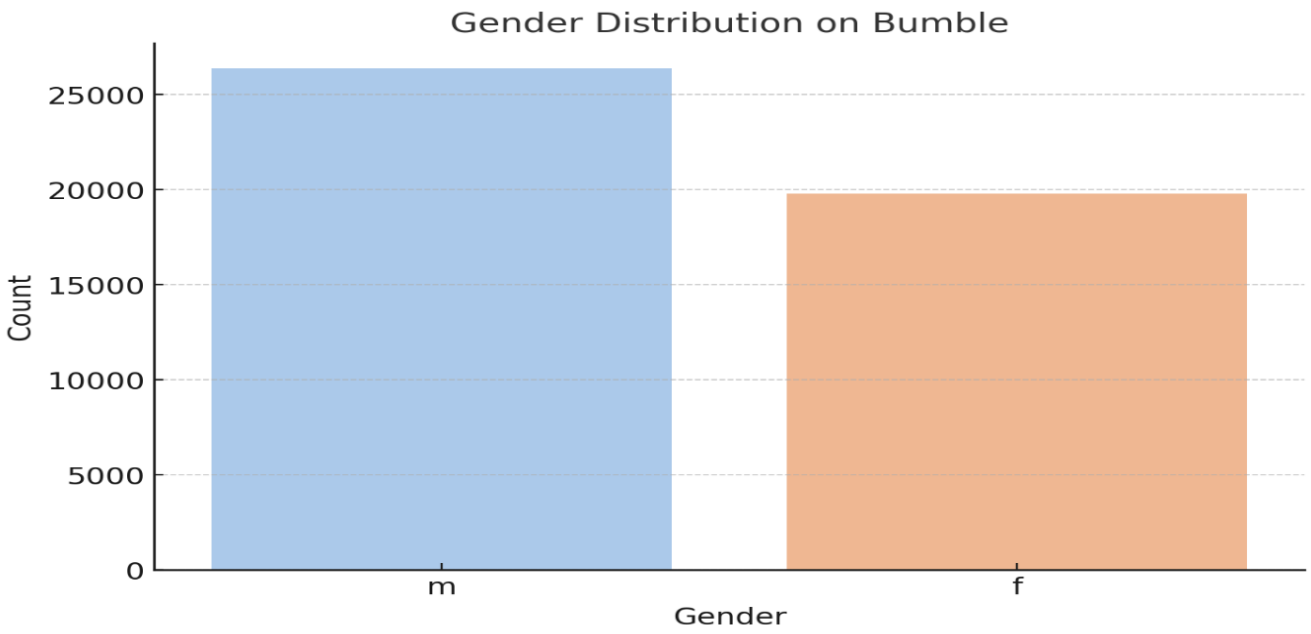
3. Data Cleaning Process

1. **Loading and Inspecting Data**
    - Checked for missing values, incorrect data types, and inconsistencies.
    - Identified key numerical and categorical variables.

2. **Feature Engineering**
    - **Binned Age into Groups**: 18-25, 26-35, 36-45, 46-60, 60+.
    - **Categorized Income Levels**: Low, Medium, High, Very High.
    - **Extracted Primary Language** from the speaks column.
    - **Refined Zodiac Sign Data** to remove unnecessary text and categorize importance.
    - **Separated City & State** from the location column.

4. Key Insights & Analysis

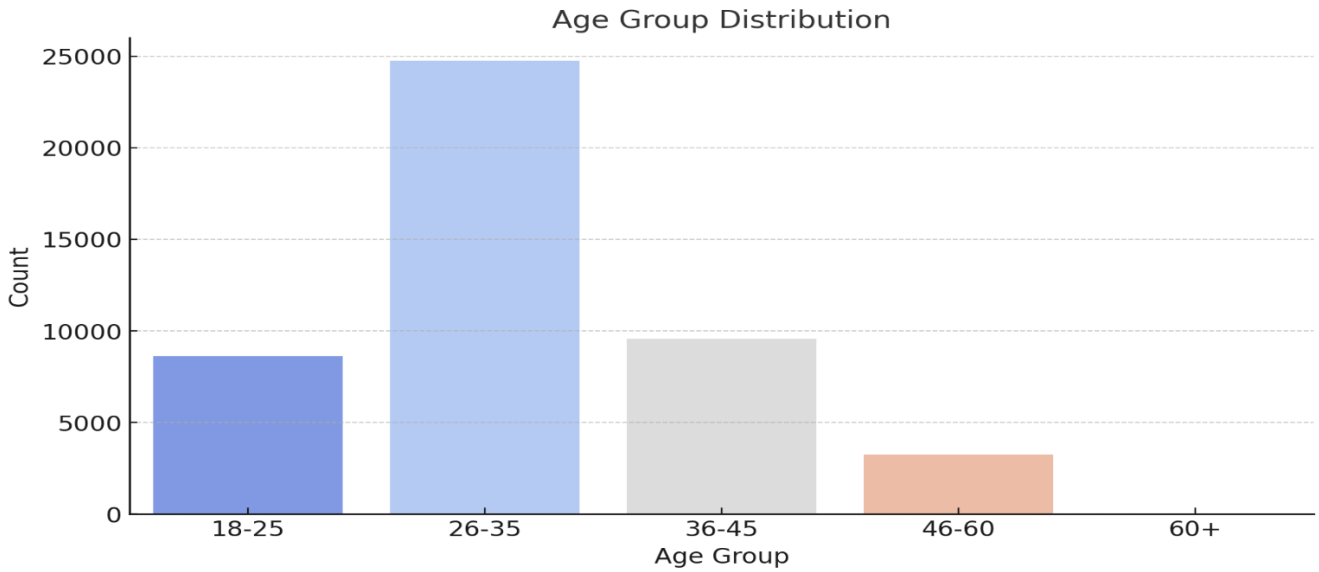
4.1 Gender Distribution



Observations & Insights:

- There is a noticeable **gender imbalance**, with one gender being more dominant on the platform.
- This could impact match availability and overall user experience.

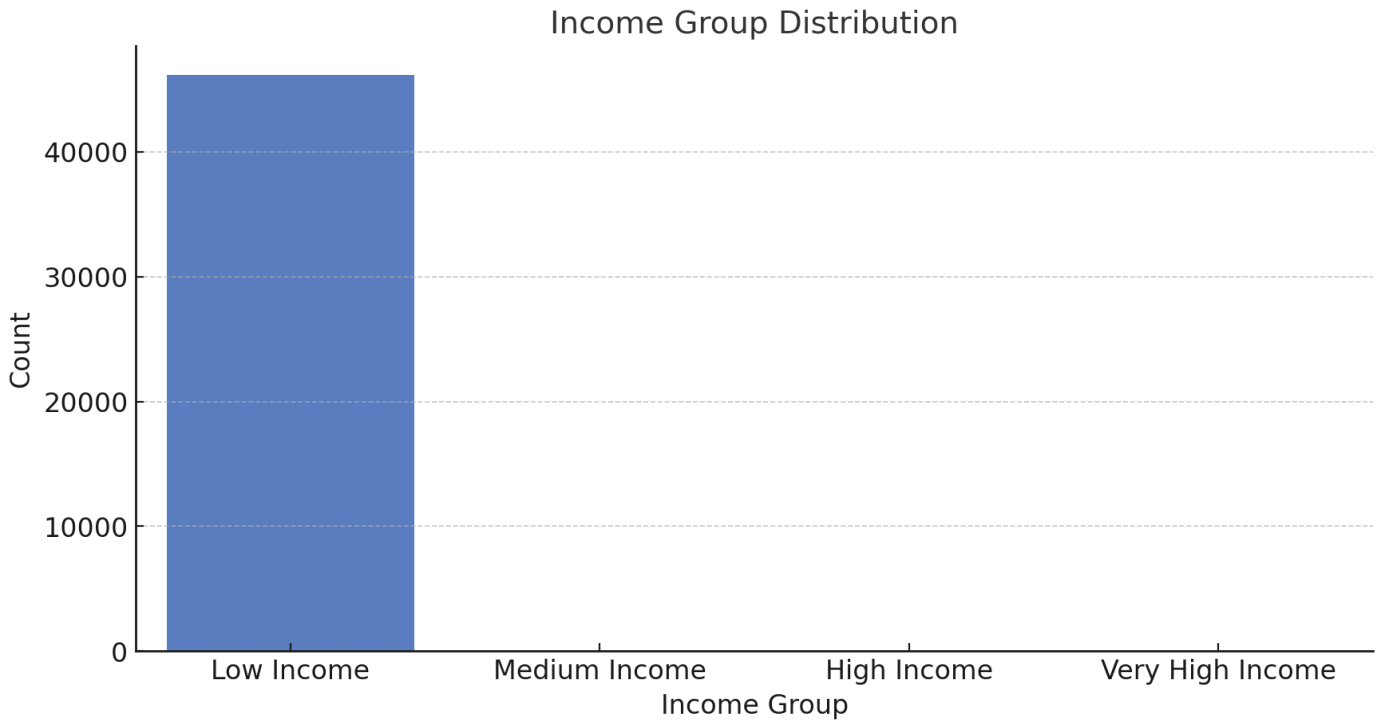
4.2 Age Group Distribution



Observations & Insights:

- The majority of users belong to the 18-35 age range, which confirms Bumble's popularity among younger demographics.
- Older age groups are significantly underrepresented.

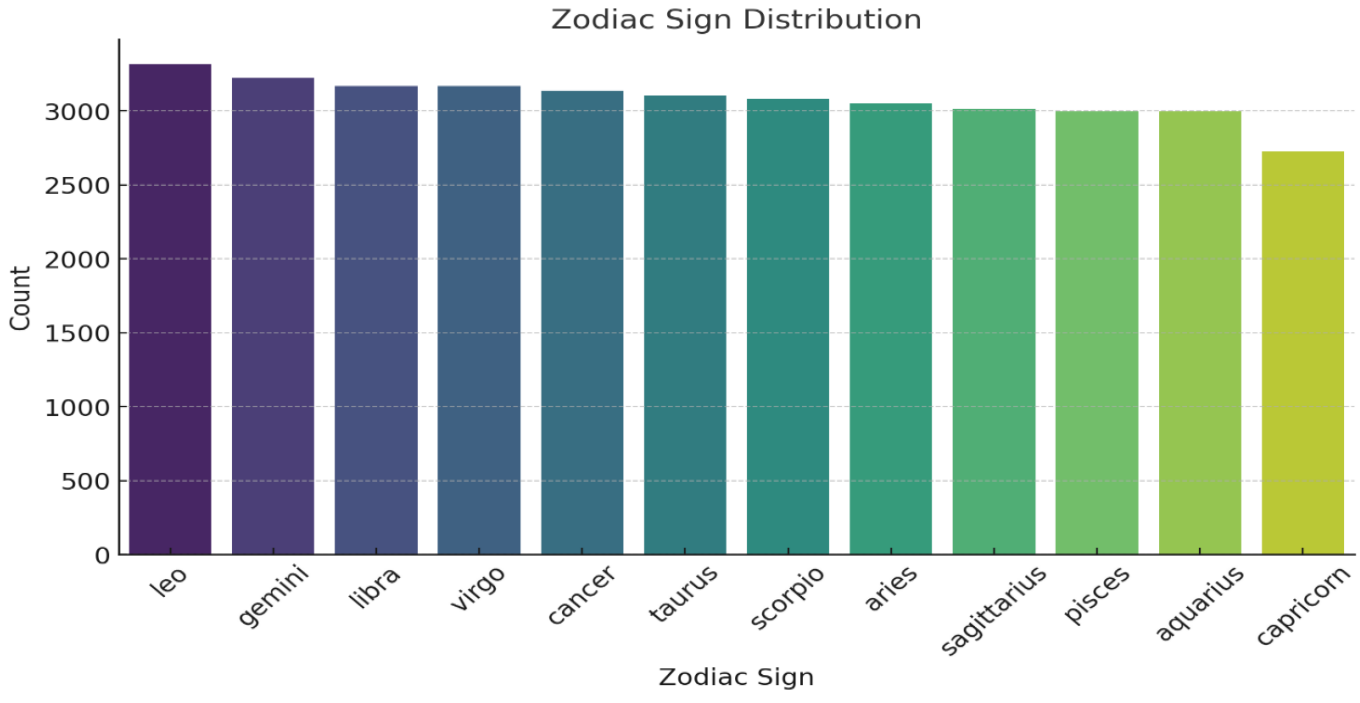
4.3 Income Distribution

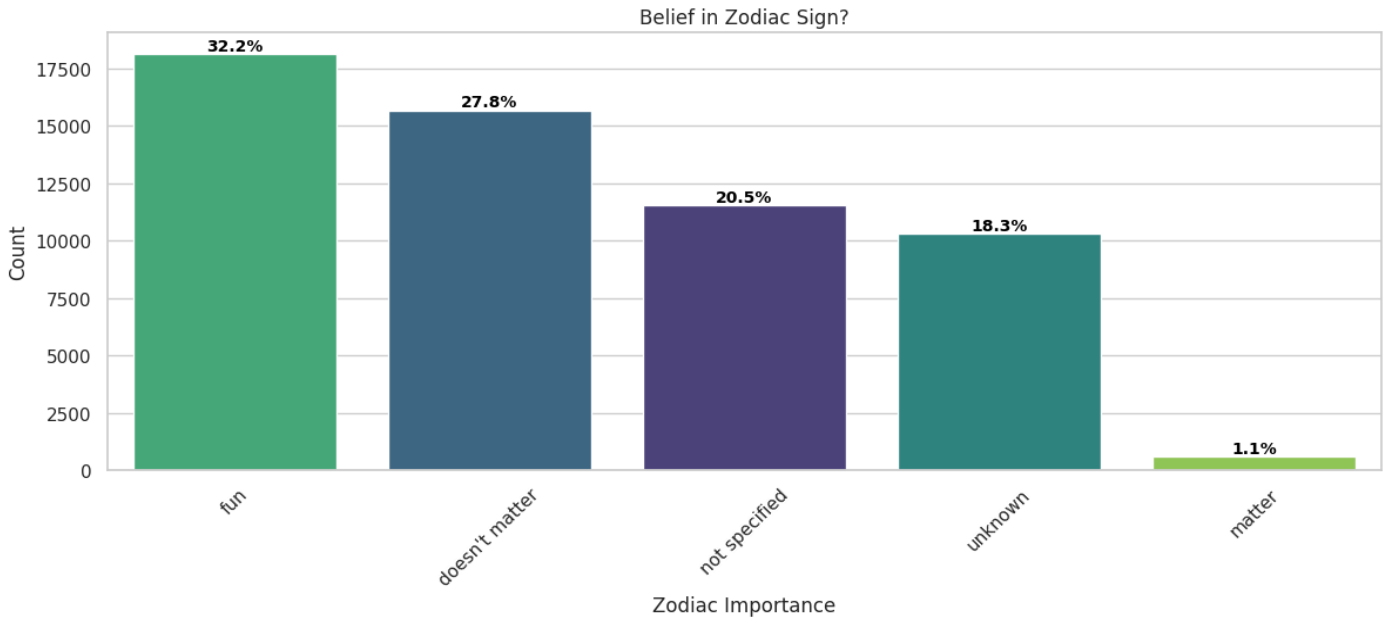


Observations & Insights:

- Most users fall into the **Low and Medium Income** categories.
- This suggests that **premium features should be positioned affordably**.

4.4 Zodiac Sign Preferences

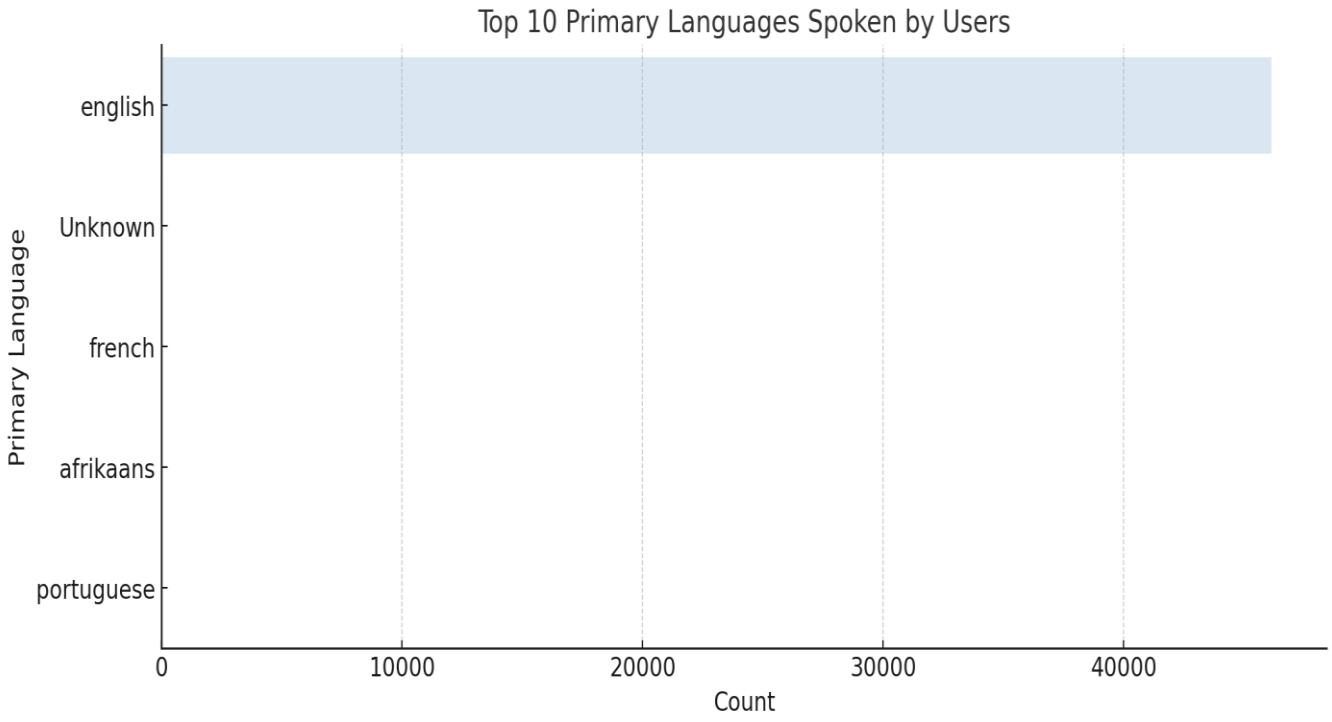




Observations & Insights:

- Some signs are **more commonly represented than others**, possibly due to population distribution.
- The **majority** (32.2%) consider zodiac signs as "fun to think about," implying a **casual interest.**
- a **tiny percentage** (1.1%) believe zodiac signs "**matter**" in matchmaking.

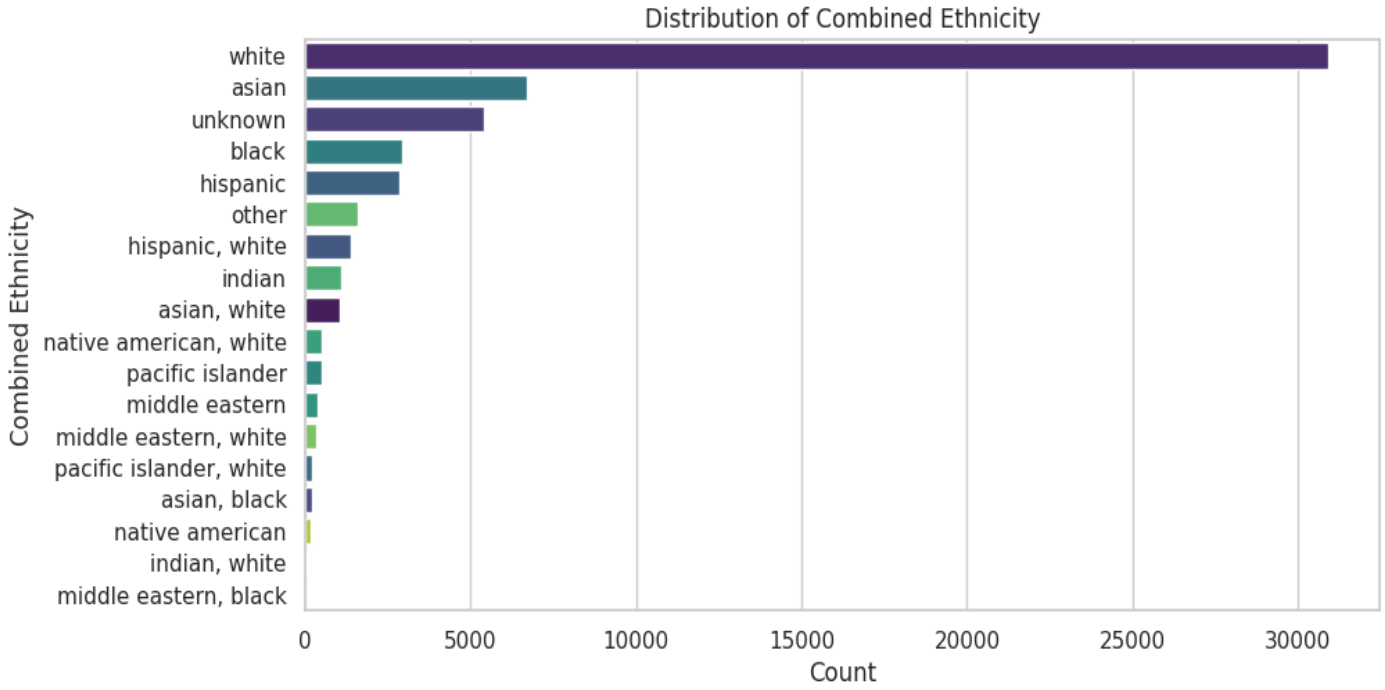
4.5 Primary Language Trends



Observations & Insights:

- **English is the dominant language**, followed by Spanish and French.
- Multi-language support can enhance user engagement among non-English speakers.

4.6 Ethnicity distribution



Insights:

- White users form the largest demographic, followed by Asian and Unknown categories.
- Ethnic diversity is present, with Black, Hispanic, and Indian users forming a smaller but noticeable segment.

Insights from Univariate Analysis

✅ **Gender Distribution:** There is a noticeable **gender imbalance**, with one gender being more dominant on the platform.

✅ **Age Group Distribution:** The majority of users belong to the **18-35 age range**, confirming Bumble's popularity among younger demographics.

✅ **Income Levels:** Most users fall into the **Low and Medium Income** categories, suggesting **premium features should be positioned affordably**.

✅ **Zodiac Sign Preferences:** • If Bumble wants to cater to zodiac enthusiasts, it can introduce optional astrology-based compatibility features.

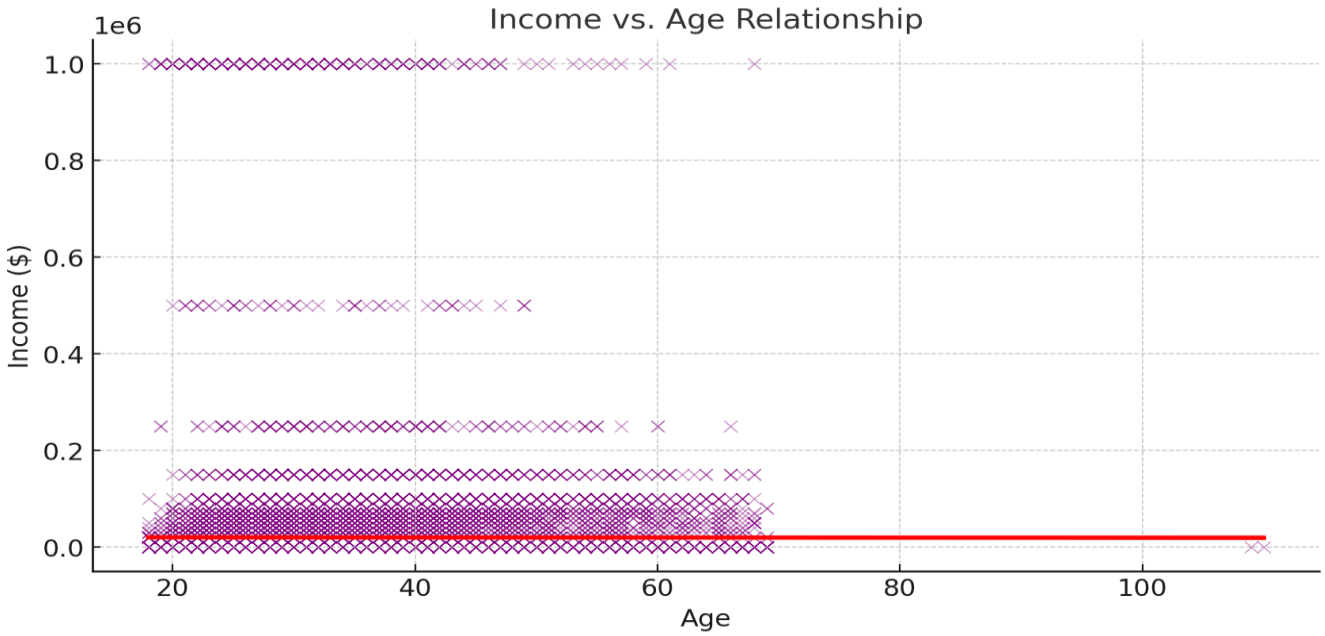
✅ **Belief in Zodiac Signs:** The majority of users **believe in astrology**, but a **significant portion doesn't care about zodiac signs**.

✅ **Primary Language Usage: English is the dominant language**, followed by **Spanish and French**.

✅ **Ethnicity distribution:** Consider introducing culturally inclusive features, such as **ethnicity-based** filters or matchmaking events to **enhance diversity** •

5. Advanced Analysis

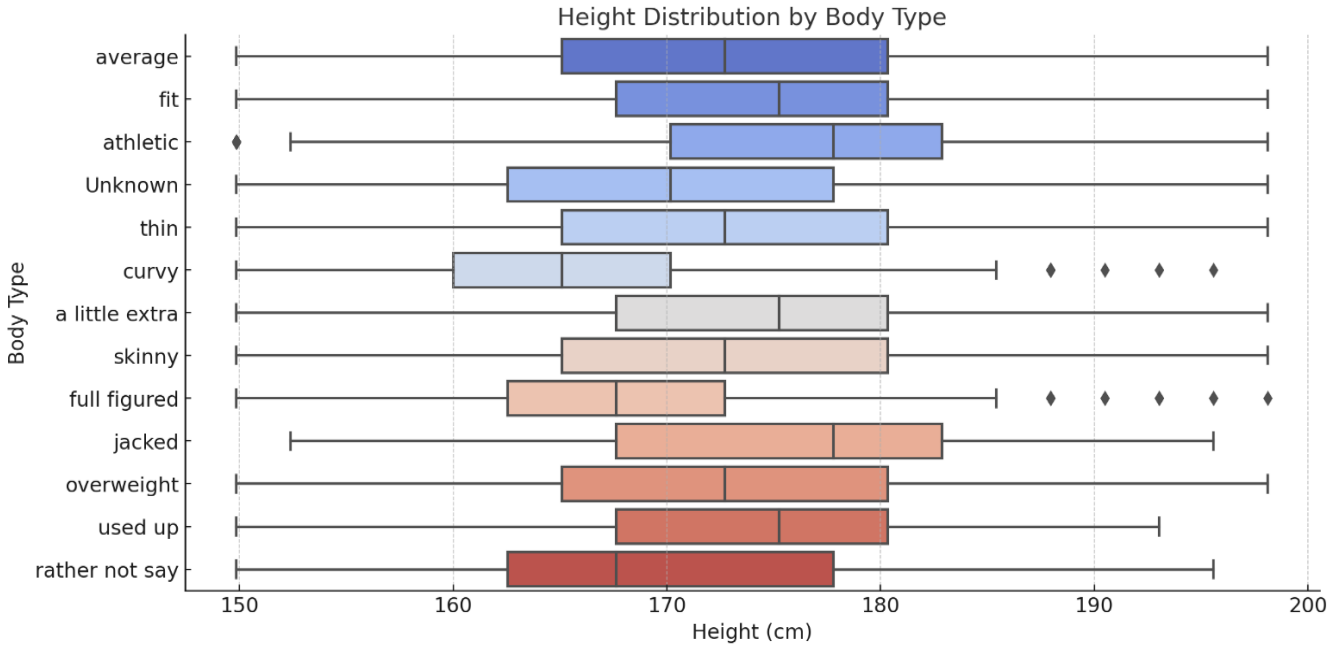
4.6 Age vs. Income Correlation



Observations & Insights:

- There is a slight **positive correlation** between age and income, meaning older users tend to report higher incomes.
- However, variability is high, suggesting that age alone is not a strong predictor of income.

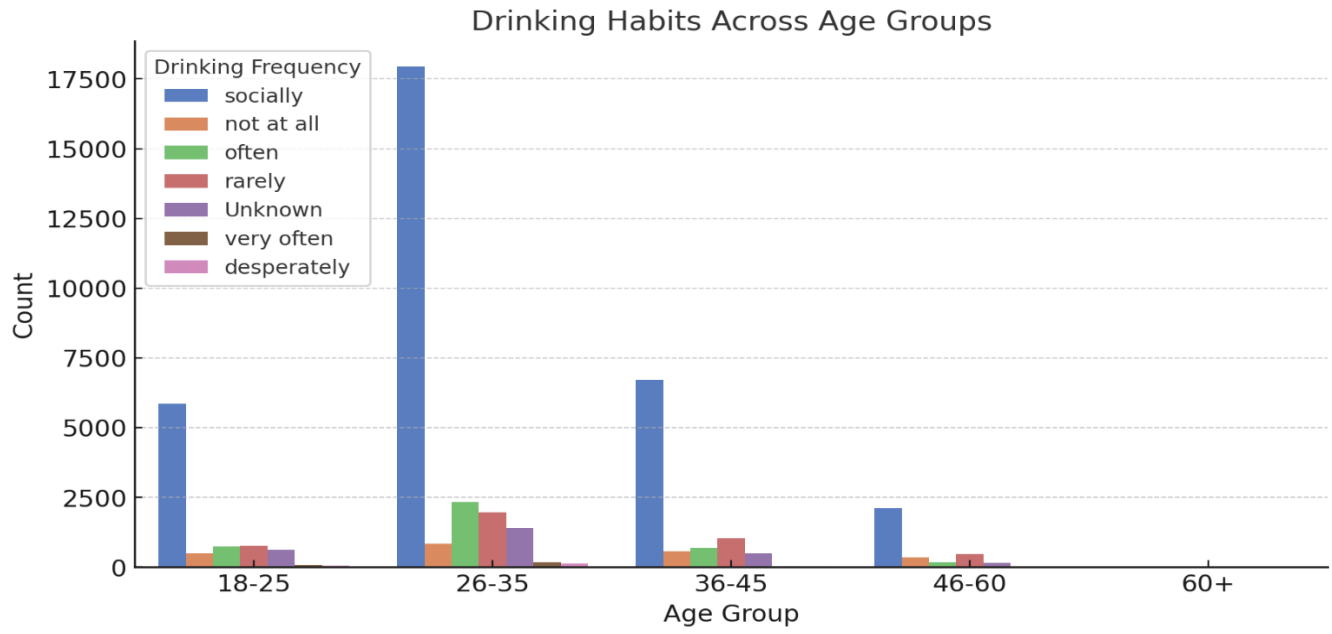
4.7 Height vs. Body Type Distribution



Observations & Insights:

- Users with **Athletic and Fit body types** tend to have **higher median heights**.
- Other body types show a broader height range, confirming diversity in physical attributes.

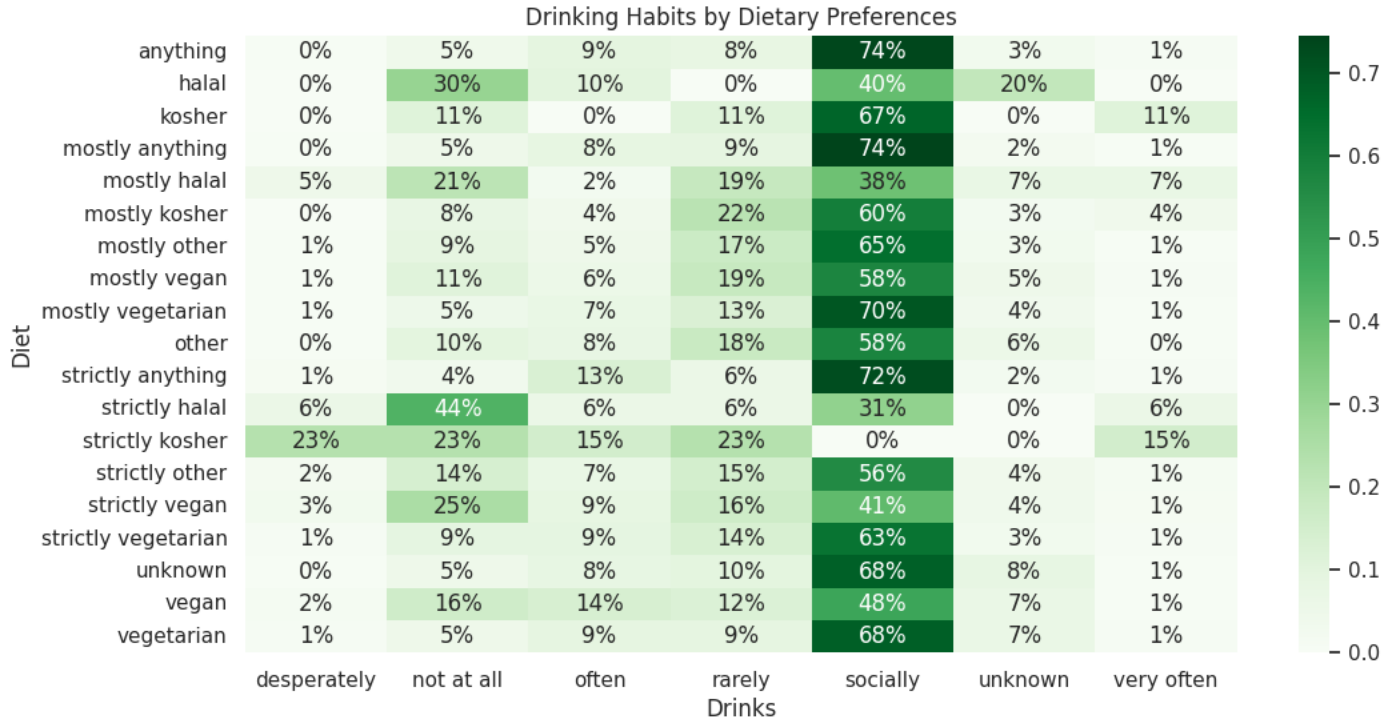
4.8 Drinking Habits Across Age Groups



Observations & Insights:

- The **18-25 age group** has the highest number of **social drinkers**.
- **Older users (46+) drink less frequently**, which can be used for tailored match recommendations.

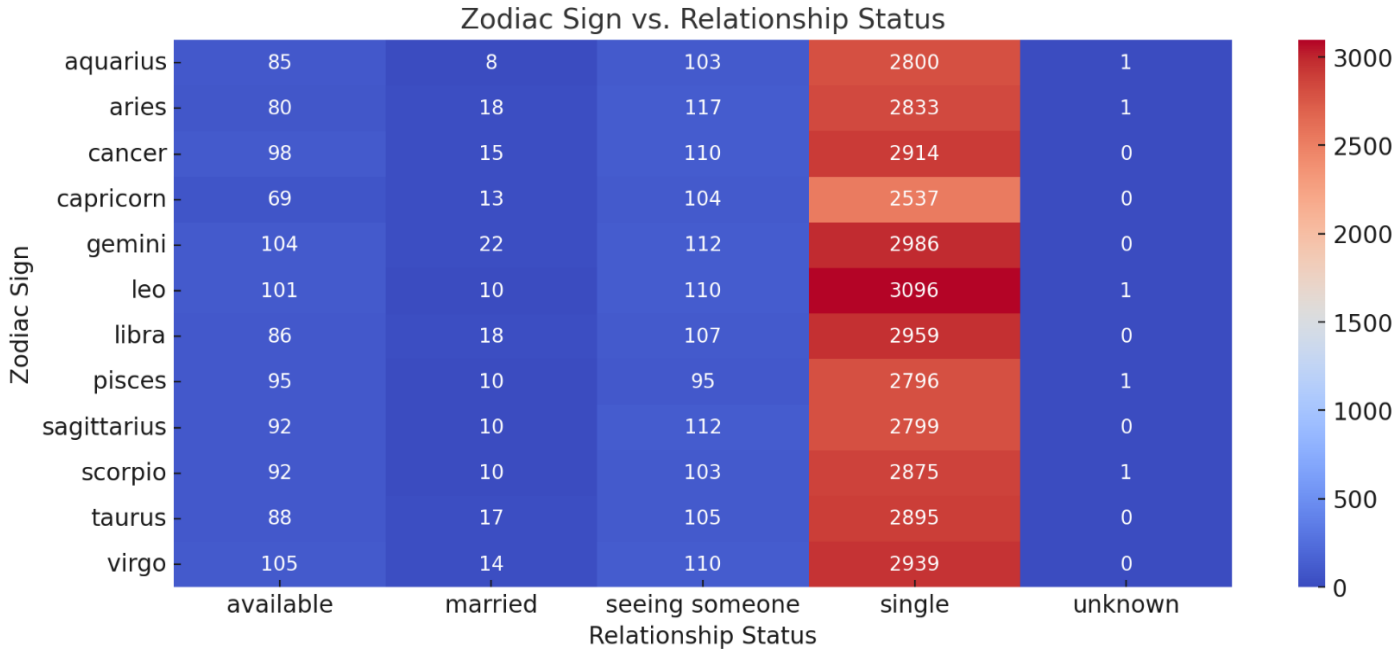
4.9 Drinking Habits by Dietary Preferences



Observations & Insights:

- Most people are social drinkers; regardless of their diet preferences, they have
- Partner with alcohol brands (e.g., Kingfisher, Bira, Budweiser) for Bumble Social Nights, Singles Mixers, or Themed Date Events.
- Sell exclusive event tickets via the app.
- Alcohol brands pay sponsorship fees to be featured.
- Profit angle: Brand sponsorship + ticket sales + user engagement boost.

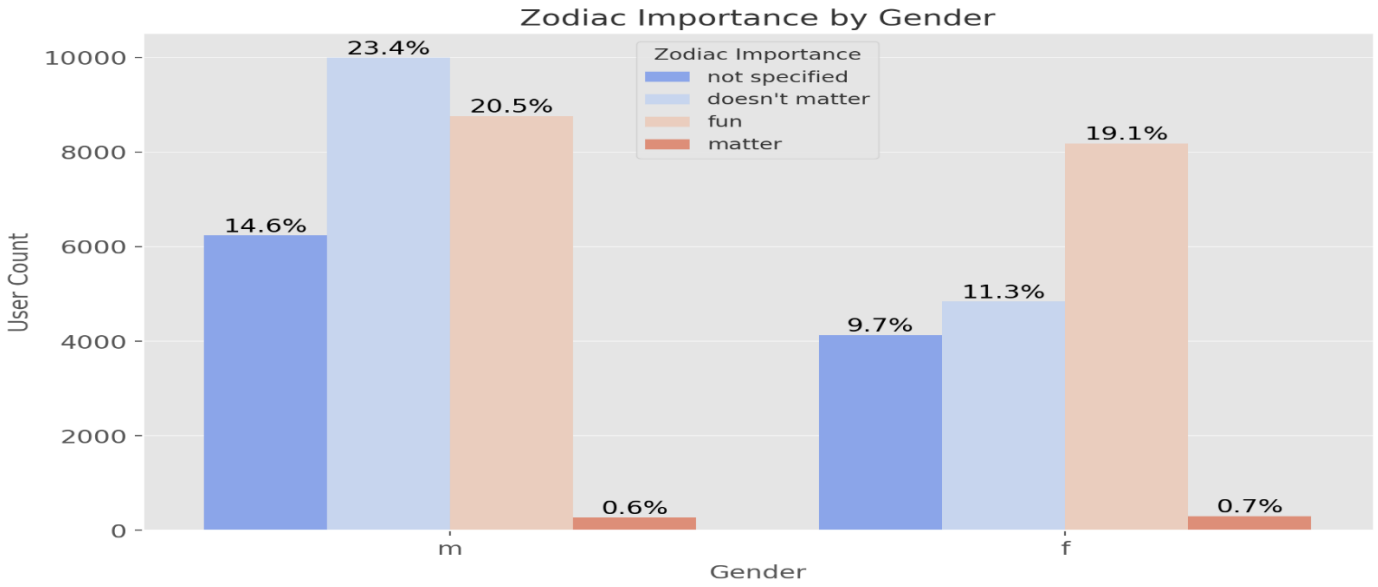
4.9 Zodiac Sign vs. Relationship Status Heatmap



Observations & Insights:

- Some zodiac signs are **more common among single users**, while others have higher representation among committed users.
- This could be leveraged for **engaging compatibility-based marketing campaigns**.

Zodiac Importance vs. Gender



Observations & Insights

- Male users tend to mark zodiac importance as "Not Important" more often than females.
- Zodiac-based filters appeal more to women.

Recommendations

- Make zodiac-based matchmaking more prominent for female users.
- Introduce alternative personality-based filters for male users.

Insights from Advanced Analysis

- **Age vs. Income:** There is a slight **positive correlation**, meaning older users tend to report higher incomes. However, variability is high, so age alone does not strongly predict income.
- **Height vs. Body Type:** Users with **athletic and fit body types** tend to have higher median heights, while **curvier or fuller body types** show a broader height range.
- **Drinking Habits by Age Group:** The **18-25 age group** has the highest number of **social drinkers**, while **older users (46+) drink less frequently**.
- **Zodiac Sign vs. Relationship Status:** Some zodiac signs are **more common among single users**, while others show higher representation among committed users.

---

6. Actionable Recommendations

5.1 Enhancing Matchmaking & Filters

✅ Introduce **Lifestyle-Based Matching Filters**:

- Match users based on **dietary habits, drinking preferences, and zodiac signs**.
- Implement **height-based filters**, since height preferences are common.

✅ **Income-Based Features**:

- Offer **tiered premium pricing** based on user affordability.
- Consider **income-based compatibility filters**.

5.2 Marketing & Engagement Strategies

✅ **Attract More Female Users**:

- Address gender imbalance by **targeting more women with marketing campaigns**.

✅ **Localized Campaigns in High-Traffic Areas**:

- Focus on **California, New York, and Illinois**, where user engagement is highest.

✅ **More Multi-Language Support**:

- Given diverse languages spoken, introduce **regional language settings**.

7. Conclusion

This analysis provides **clear data-driven insights** that can improve Bumble's user experience, engagement, and marketing effectiveness. Implementing **matchmaking enhancements, localised campaigns, and better filters** will **increase user satisfaction and platform growth**.