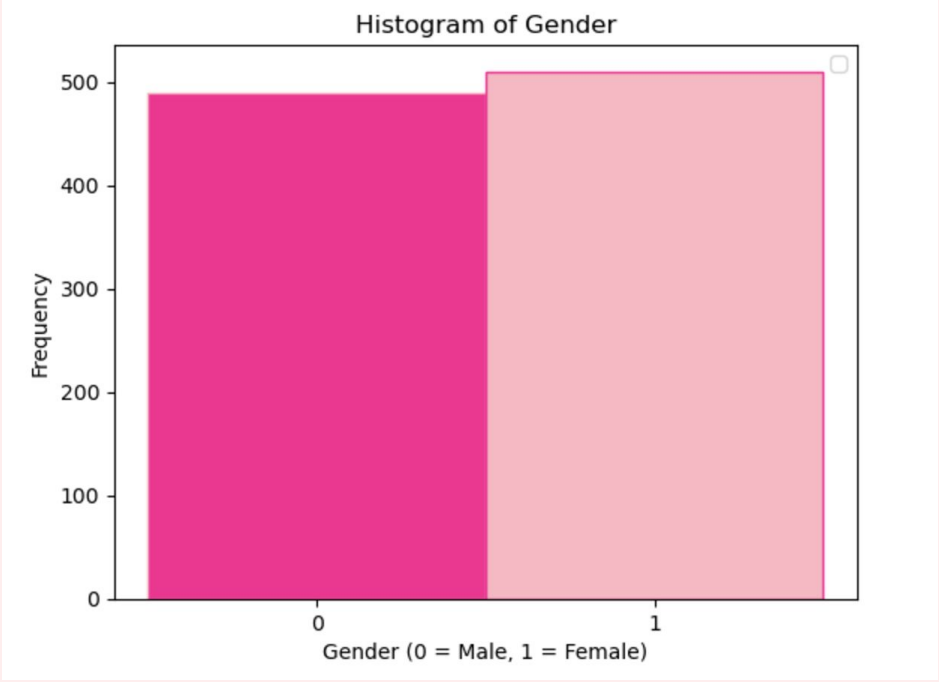
Exploratory Data Analysis of Online Dating Matches DataSet

Amanda Wyse
DSC 530

01

The Variables

Gender



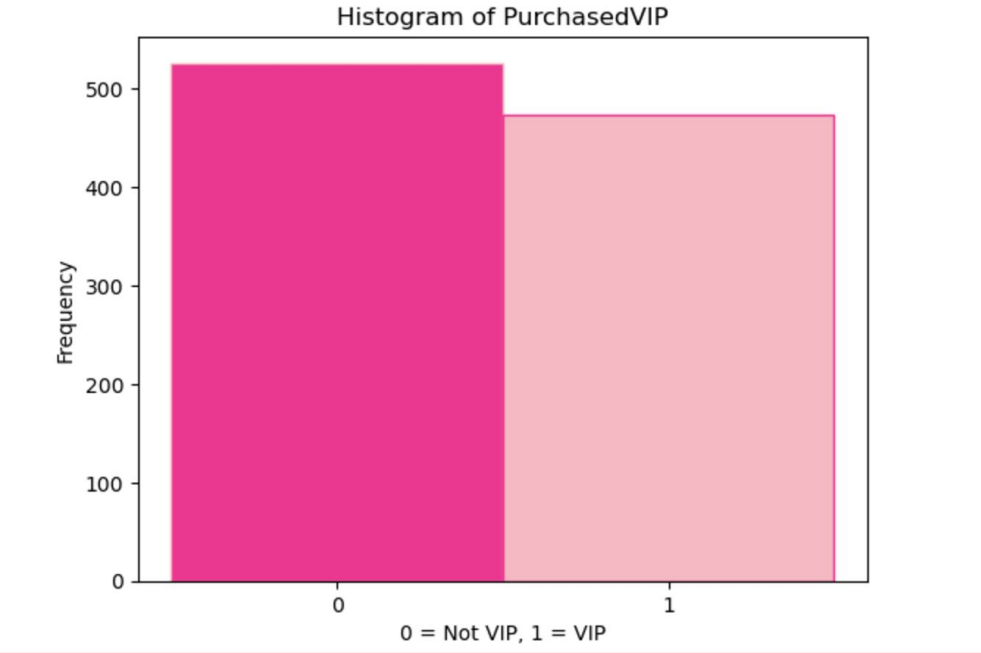
Male

For this variable, 0 represents a male in the population.

Female

For this variable, 1 represents a male in the population.

VIP Member



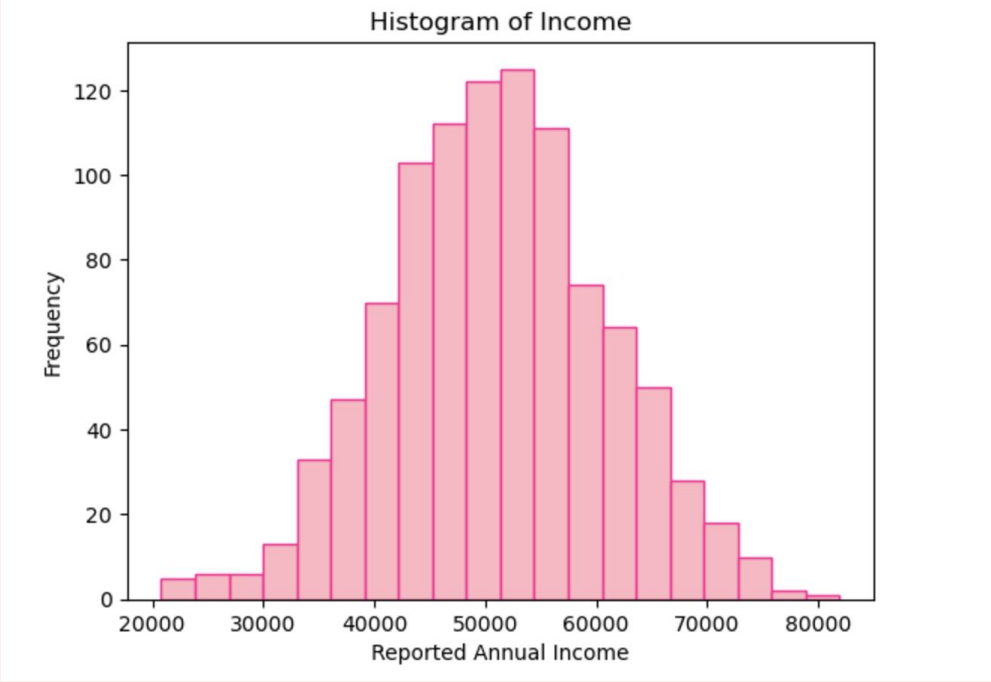
Not VIP

For the variable, 0 represents a member that did not purchase VIP.

VIP

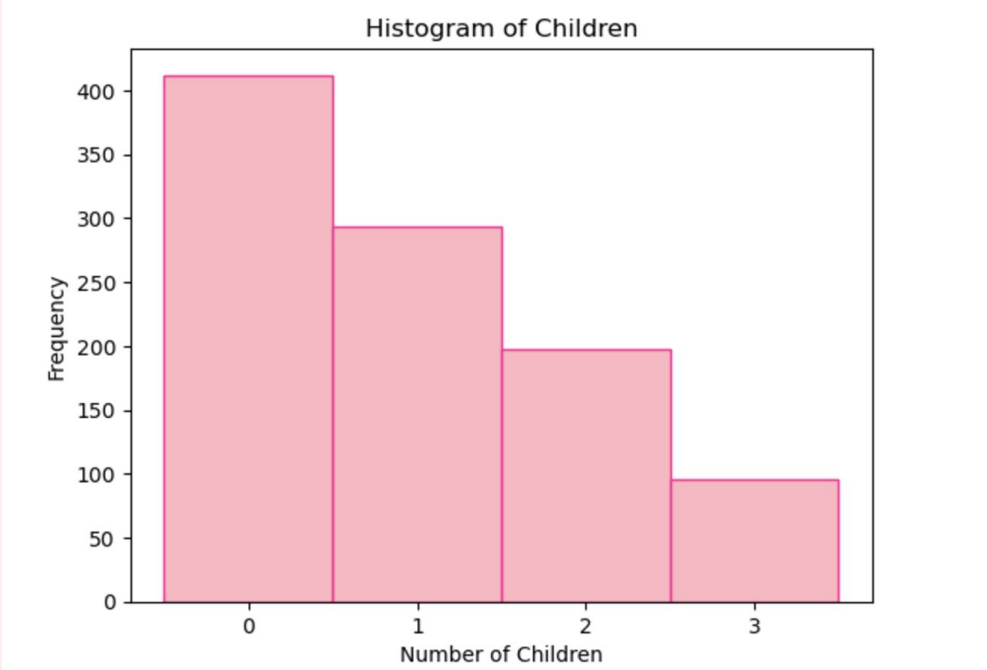
For this variable, 1 represents a member that did purchase VIP.

Self-Reported Annual Income



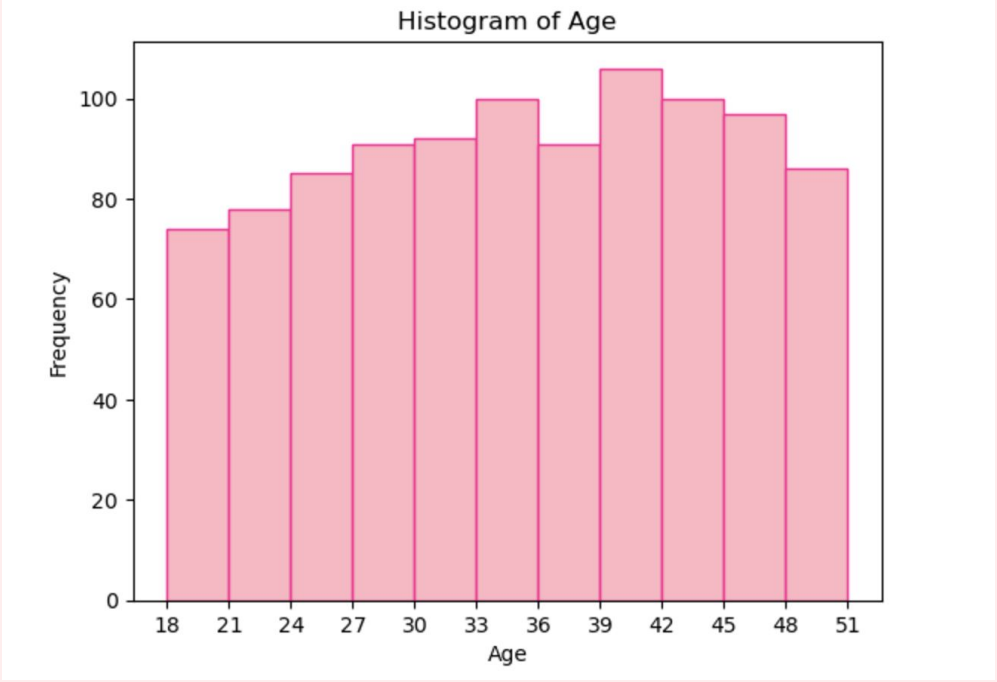
This variable captures self-reported annual income for each member. There are no null values for the variable, and no outliers.

Number of Children



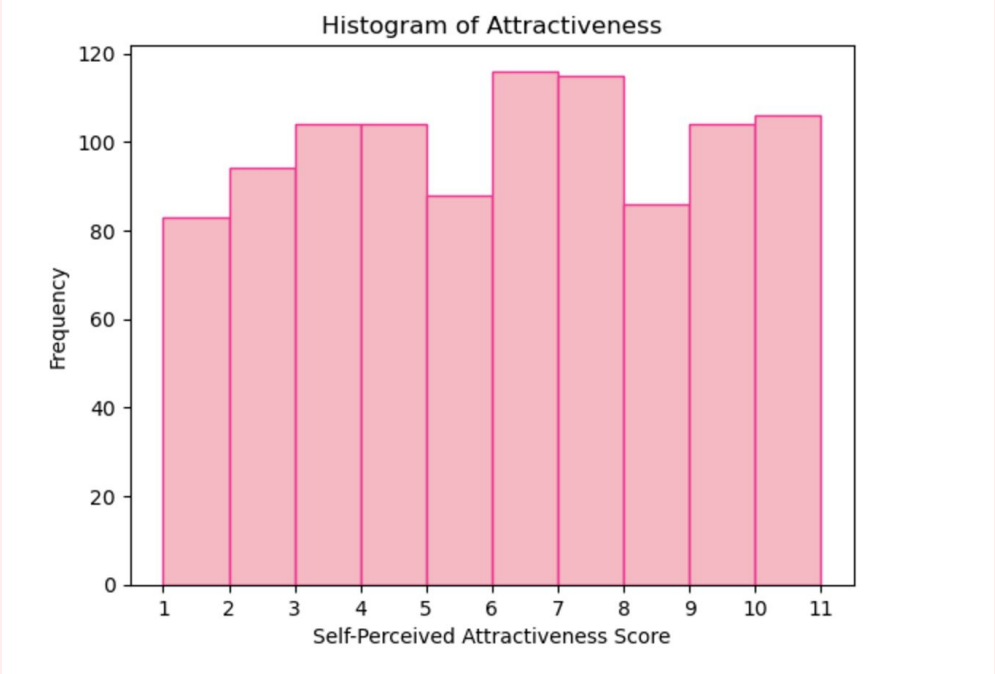
This variable captures the number of children the member has. There are no null values for this variable and no outliers.

Age



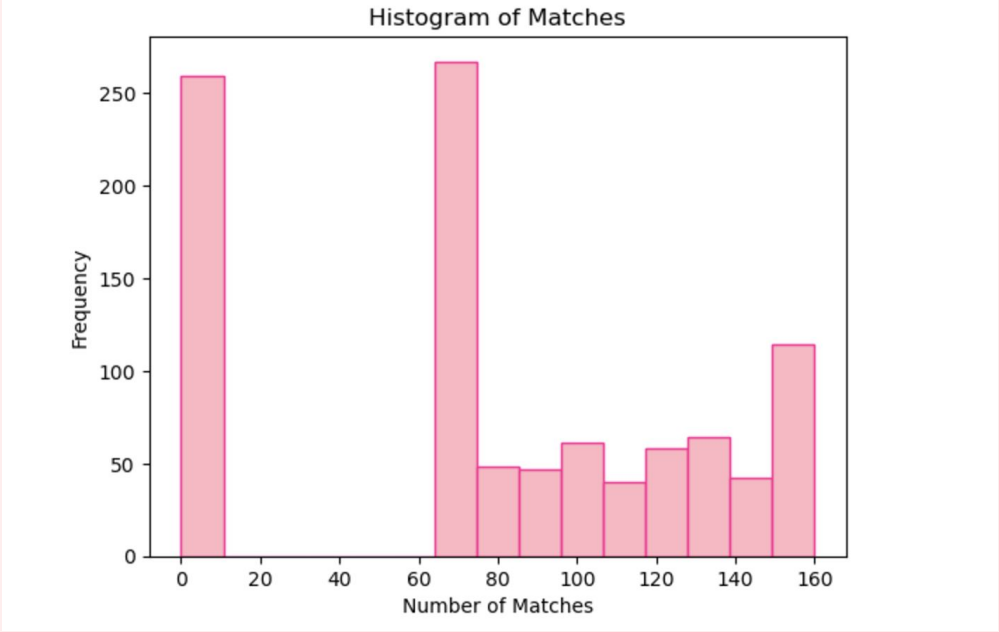
This variable captures the age of the member. There are no null values for this variable and no outliers.

Self-Perceived Attractiveness Score



This variable captures the self-perceived attractiveness score of each member on a 1-10 scale. There are no null values and no outliers.

Number of Matches



This variable captures the number of matches the member has made on the app. There are no null values and no outliers. A 0 represents in inactive member.

02

Descriptive Statistics

Gender

Mean	0.51
Mode	1
Spread	0.50
Skewness	-0.04
Kurtosis	-1.998

The Gender variable has a mean of 0.51, indicating a roughly equal distribution of males (0) and females (1). The mode is 1, meaning females are slightly more prevalent. The spread, with a standard deviation of 0.50, suggests a balanced variation. The skewness is -0.04, showing the distribution is nearly symmetrical, while the kurtosis of -2.00 reflects a flatter, less concentrated distribution compared to a normal distribution.

VIP Member

Mean	0.474
Mode	0
Spread	0.4996
Skewness	0.10
Kurtosis	-1.989

The PurchasedVIP variable has a mean of 0.474, indicating slightly fewer VIP purchasers (1) compared to non-VIP (0). The mode is 0, showing the majority of users did not purchase VIP. The standard deviation of 0.50 reflects a balanced variation. A skewness of 0.10 suggests a slight right skew, while the kurtosis of -1.99 indicates a flatter distribution compared to a normal curve.

Self-Reported Annual Income

Mean	50,988
Mode	37,903
Spread	9,899
Skewness	-0.037
Kurtosis	-0.027

The Income variable has a mean of $50,988, with the most common income (mode) being $37,903. The standard deviation of $9,889 indicates moderate variability in income levels. The skewness of -0.037 suggests a nearly symmetrical distribution, while the kurtosis of -0.027 reflects a distribution close to normal in terms of tail behavior.

Number of Children

Mean	0.978
Mode	0
Spread	0.997
Skewness	0.626
Kurtosis	-0.774

The Children variable has a mean of 0.978, indicating an average of about one child per individual. The mode is 0, meaning most individuals have no children. The standard deviation of 0.997 reflects moderate variability. A skewness of 0.626 indicates a right-skewed distribution, with more individuals having fewer children. The kurtosis of -0.774 suggests a flatter distribution with fewer extreme values compared to a normal curve.

Age

Mean	34.6
Mode	25
Spread	9.15
Skewness	-0.10
Kurtosis	-1.17

The Age variable has a mean of 34.6, indicating the average age of individuals is around 35. The mode is 25, meaning the most common age is 25. The standard deviation of 9.15 reflects a moderate spread in ages. The skewness of -0.10 shows a slightly left-skewed distribution, suggesting a minor concentration of younger individuals. The kurtosis of -1.17 indicates a flatter distribution with fewer extreme age values compared to a normal distribution.

Self-Perceived Attractiveness Score

Mean	5.62
Mode	6
Spread	2.82
Skewness	-0.03
Kurtosis	-1.28

The Attractiveness variable has a mean of 5.62, suggesting that the average self-perceived attractiveness rating is slightly above the middle of the scale. The mode is 6, indicating that the most common rating is 6. The standard deviation of 2.82 shows moderate variability in attractiveness ratings. The skewness of -0.03 indicates a nearly symmetrical distribution, while the kurtosis of -1.18 reflects a flatter distribution with fewer extreme values compared to a normal curve.

Number of Matches

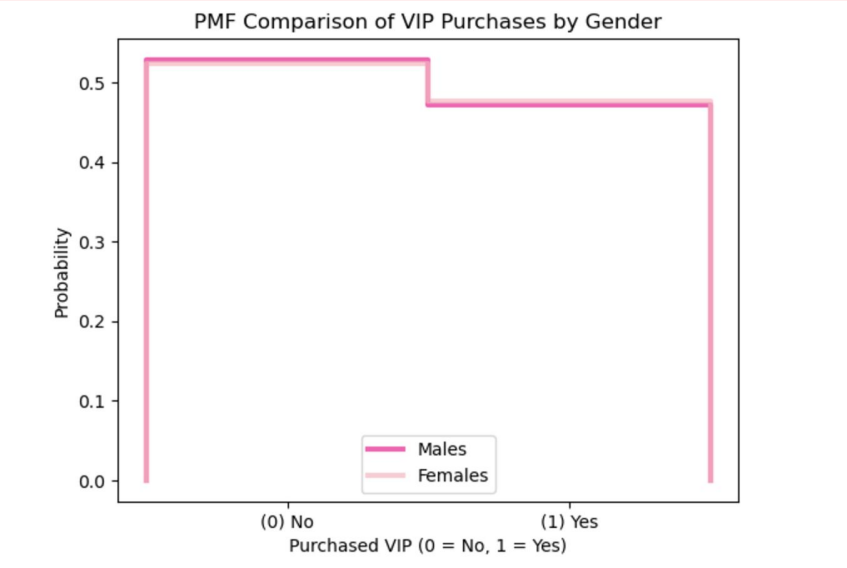
Mean	76.05
Mode	70
Spread	52.71
Skewness	-0.20
Kurtosis	-1.10

The Matches variable has a mean of 76.05, indicating an average of about 76 matches. The mode is 70, meaning the most common number of matches is 70. The standard deviation of 52.71 reflects significant variability in the number of matches. The skewness of -0.20 suggests a slightly left-skewed distribution, with more individuals having fewer matches. The kurtosis of -1.10 indicates a flatter distribution with fewer extreme values than a normal distribution.

03

Probability Mass Function

VIP Among Males and Females



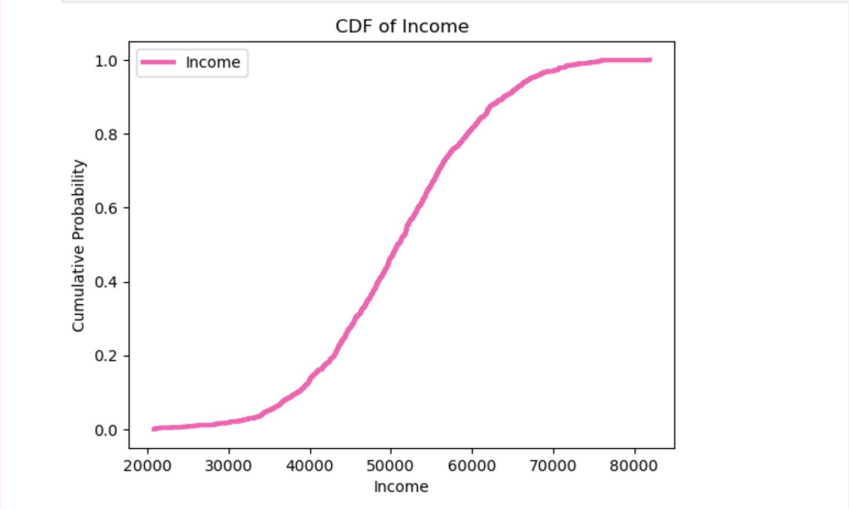
The PMF chart shows no significant difference in VIP purchase behavior between males and females, as their probabilities overlap entirely.

Both genders have an equal likelihood (50%) of purchasing or not purchasing VIP access.

This indicates that gender does not appear to influence the decision to buy VIP access in this dataset.

04
Cumulative Distribution Function

Income



The CDF shows that the majority of incomes are concentrated between 40,000 and 60,000, as indicated by the steep slope in this range.

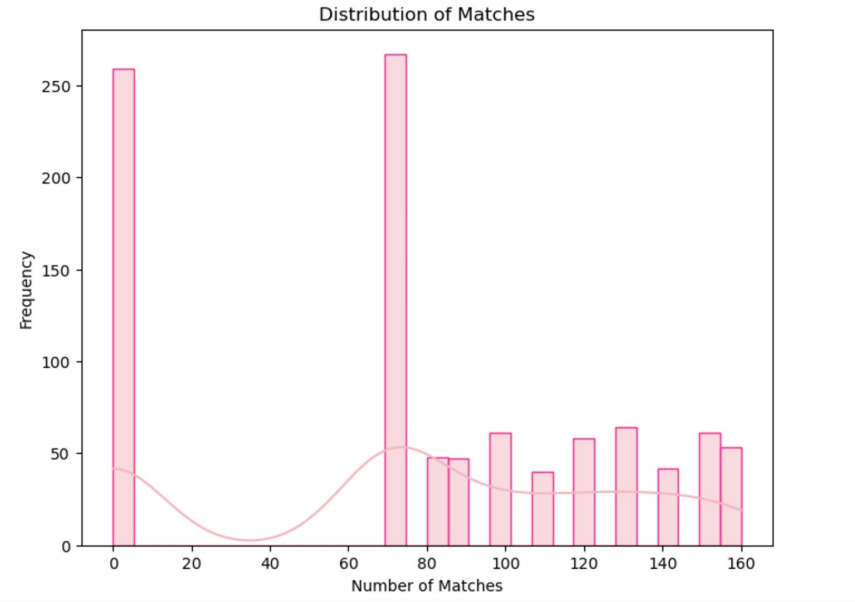
The median income is approximately 50,000, where the cumulative probability reaches 0.5.

There are fewer individuals with incomes below 30,000 or above 70,000, as evidenced by the flatter regions at the extremes.

05

Analytical Distribution

Matches



The distribution shows a clear bimodal pattern, with one large group having zero matches and another concentrated around 80 matches, indicating two distinct participant behaviors or outcomes.

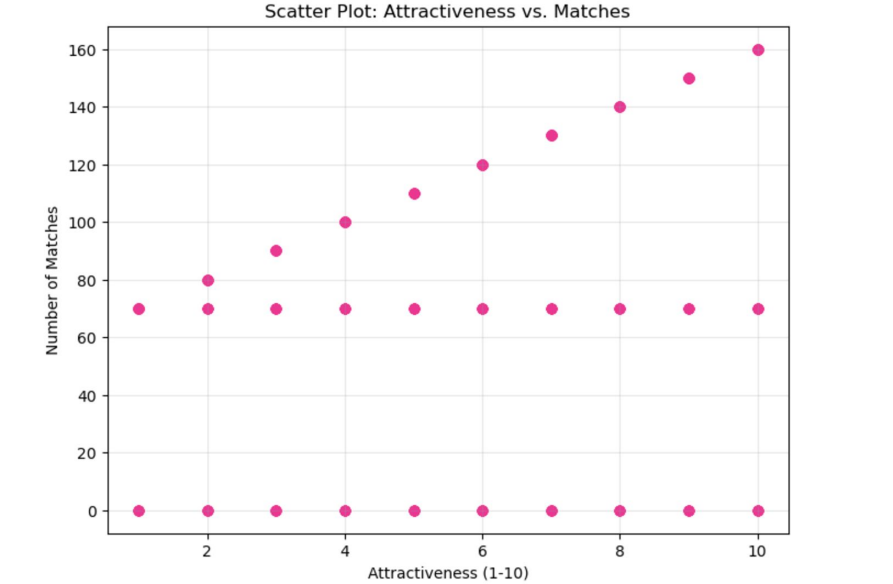
The long tail in the distribution suggests that some participants achieved significantly higher numbers of matches, perhaps representing outliers.

This pattern highlights disparities in success or participation, suggesting the need for further analysis to understand underlying factors driving these differences.

06

Scatterplots

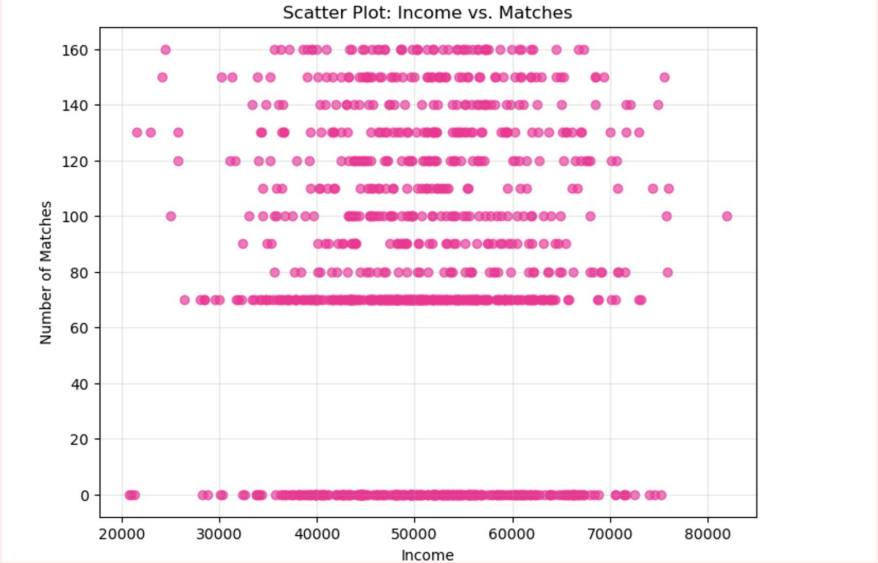
Attractiveness VS Number of Matches



Covariance (Attractiveness vs. Matches): 46.68
Pearson's Correlation (Attractiveness vs. Matches): 0.31

The scatterplot shows that inactive members (0 matches) exist across all attractiveness levels, indicating no correlation between inactivity and attractiveness. Among active members, there is a positive trend where higher attractiveness ratings generally lead to more matches. Outliers with exceptionally high matches are concentrated at the upper end of the attractiveness scale, highlighting the strong engagement linked to higher attractiveness.

Income VS Number of Matches



Covariance (Income vs. Matches): 3721.89
Pearson's Correlation (Income vs. Matches): 0.01

The scatterplot shows no clear linear relationship between income and the number of matches, indicating that income does not strongly influence match frequency. Clusters of matches are observed at specific levels (e.g., 60, 80, and 100 matches), suggesting common thresholds across all income groups. A large proportion of individuals with 0 matches are present across all income levels, highlighting that inactivity is unrelated to income.

07

Hypothesis Testing

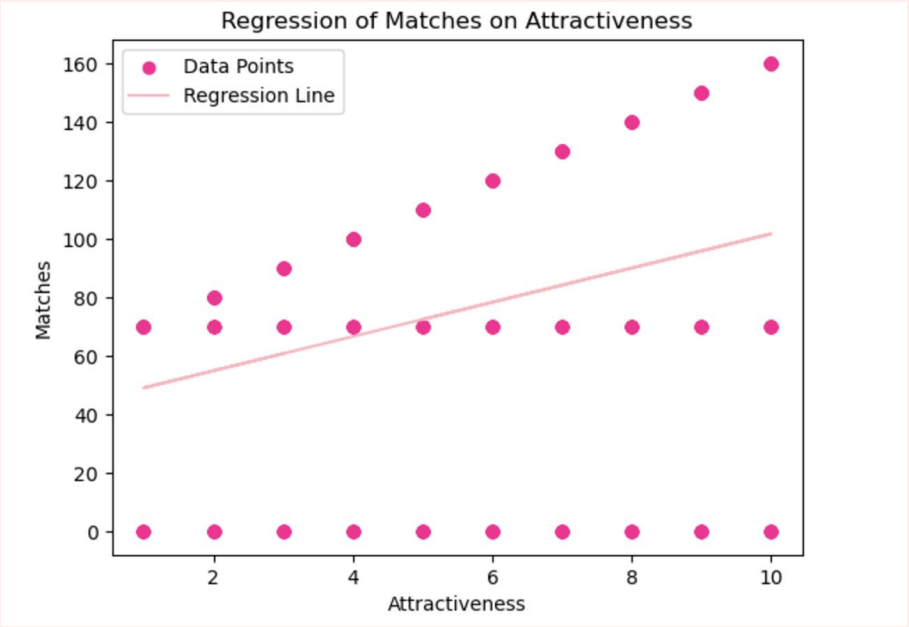
P-VALUE = 0.00

The hypothesis test aimed to examine the relationship between Attractiveness and the Number of Matches. The null hypothesis (H^0) suggests that there is no correlation between Attractiveness and Matches, meaning the two variables are independent. The alternative hypothesis (H^1) suggested that there is a non-zero correlation, indicating a relationship between the two variables. The test used the correlation coefficient as the test statistic, with data permuted over 1,000 iterations. The resulting p-value was 0, which strongly suggests rejecting the null hypothesis and supports a statistically significant correlation between Attractiveness and the Number of Matches.

08

Regression Analysis

Number of Matches with Attractiveness



The regression plot shows a positive relationship between attractiveness and the number of matches, with higher attractiveness generally correlating to more matches. However, there is considerable variability in the data, suggesting other factors influence the number of matches. Outliers with high matches at higher attractiveness levels and inactive individuals (0 matches) across all attractiveness levels highlight that attractiveness is not the only factor at play. Further analysis may be needed to explore other contributing factors.