Tindership
Data Science Intern

Vidhi Agrawal

Preliminary Data Exploration

Objective: To understand user behaviour, actions and engagement on Tinder across different countries, find patterns and state recommendations to increase user engagement and match count

Dataset: 35780 rows(daily user data), 14 columns (demographics, engagement)
Time range: 1 month data (1st Jan 2020-31st Jan 2020)
Number of unique users: 2198

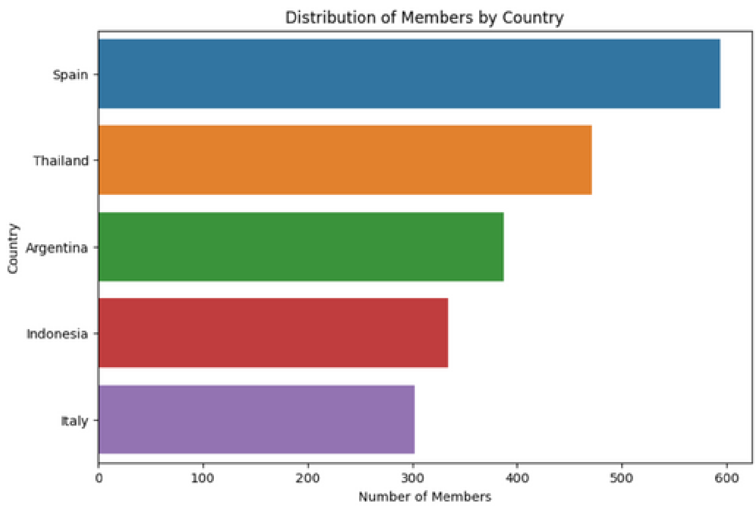
Data Quality Check: No missing values detected, no duplicates detected. No potential outliers detected. The age column has four members with age > 70 (max age 114) which is not very intuitional. Though we have included these users in our analysis, these users can be dropped based on business constraints

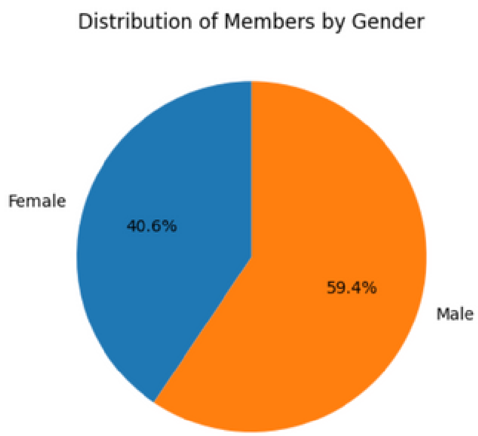
Languages and Tools Used: Python/Jupyter Notebook

Key observations from the data:

- Male gender predominates among active users across all analyzed countries.
- The age range typically centers around mid-to-late 20s.
- Engagement levels (swipes, likes, matches, messages) vary by country, indicating different usage patterns and possibly cultural influences on dating app behavior.
- Though there are more android users, iOS users swipe more often

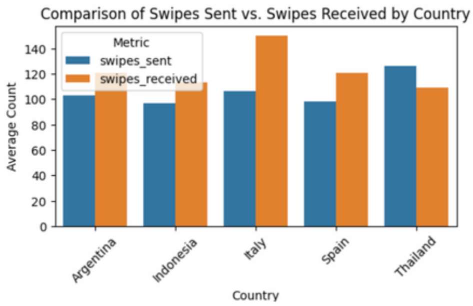
Overview of what the overall data is telling us

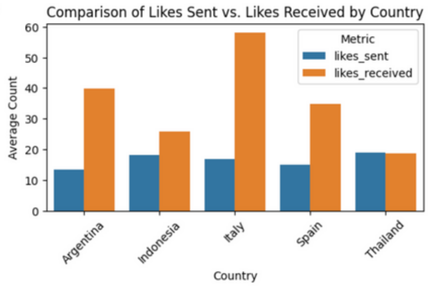


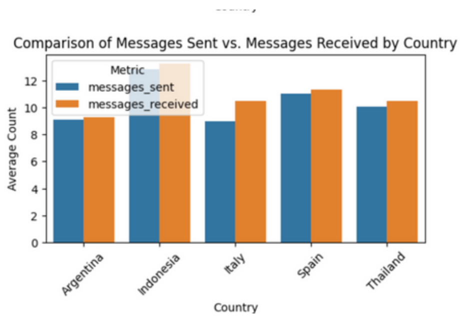


Most active users are using the app from Spain, Thailand and Argentina and the ratio of male to female users is 60% to 40%.

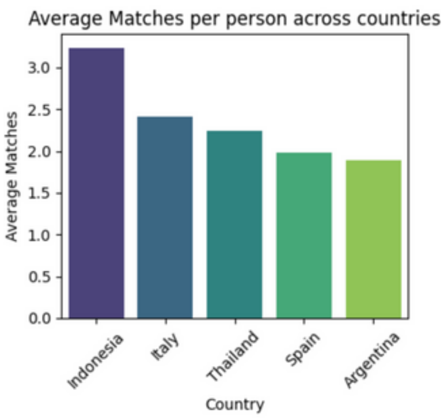
Understanding User Similarities and Differences across countries

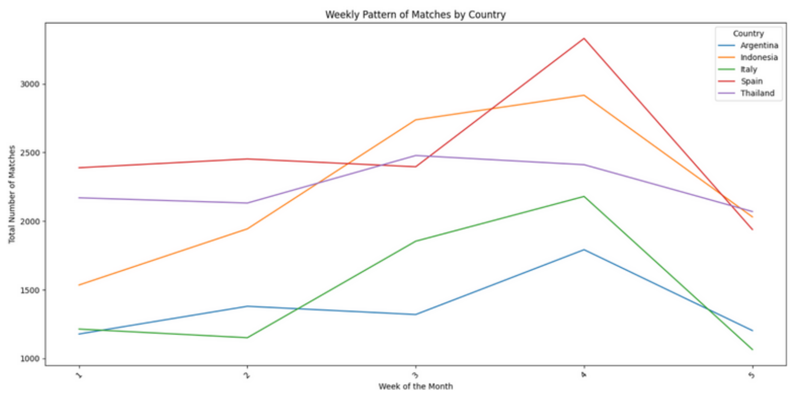






- **Balanced Swiping Activity:** The number of swipes sent and swipes received per person seems relatively balanced across all countries, suggesting a healthy level of user engagement in terms of swiping behavior.
- **Italy:** Shows highest number of swipes and likes received per person
- **Thailand:** Users in Thailand particularly exhibit higher swipes sent than receive which implies users in Thailand are actively swiping
- **Variability in Likes:** There's a more noticeable variation between likes sent and likes received compared to swipes. This could indicate differing levels of engagement or selectivity when it comes to expressing interest with a like.
- **Spain, Italy and Argentina:** Show a significant gap with more likes received than sent, suggesting that users here may be receiving likes from a broader international user base or they may be very selective in sending likes which is not the case for Indonesia and Thailand
- **Message Engagement:** The number of messages sent and received is generally close across all countries with Indonesians being the most communicative

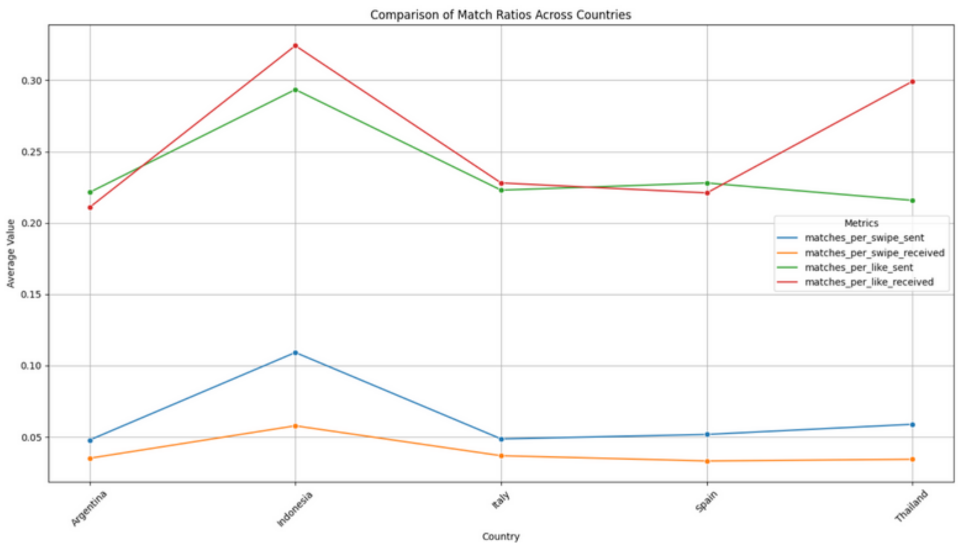




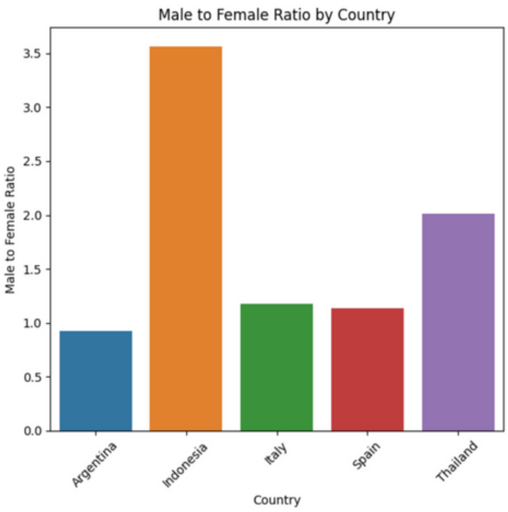
Indonesia has the highest average number of matches per person. It can be observed that though it has lower likes sent/received, the conversion to a match is higher. The graph on the right shows the weekly trend of total matches across countries over the month. We can observe Week 4 spiking on matches almost across all countries with Spain producing highest number of matches almost throughout the month.

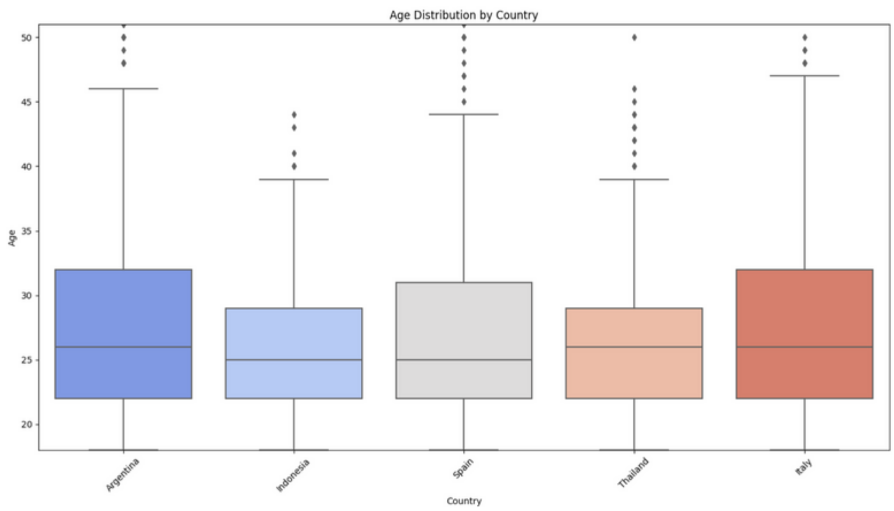
Understanding User Similarities and Differences across countries

It's critical for Tinder to understand how matches interact with swipes and likes. In order to capture how many matches are occurring per swipe/like, we can utilise four metrics:
'matches_per_swipe_sent'(swipe efficiency), ***'matches_per_swipe_received'***, ***'matches_per_like_sent'(match_to_like ratio)***, ***'matches_per_like_received'*** and plotted their average values for all countries to understand which countries have a better engagement. This trend validates our above findings on how Indonesia is doing fairly well on matches (all 4 metrics highest for Indonesia) whereas Spain, Italy and Argentina have lower match rates.



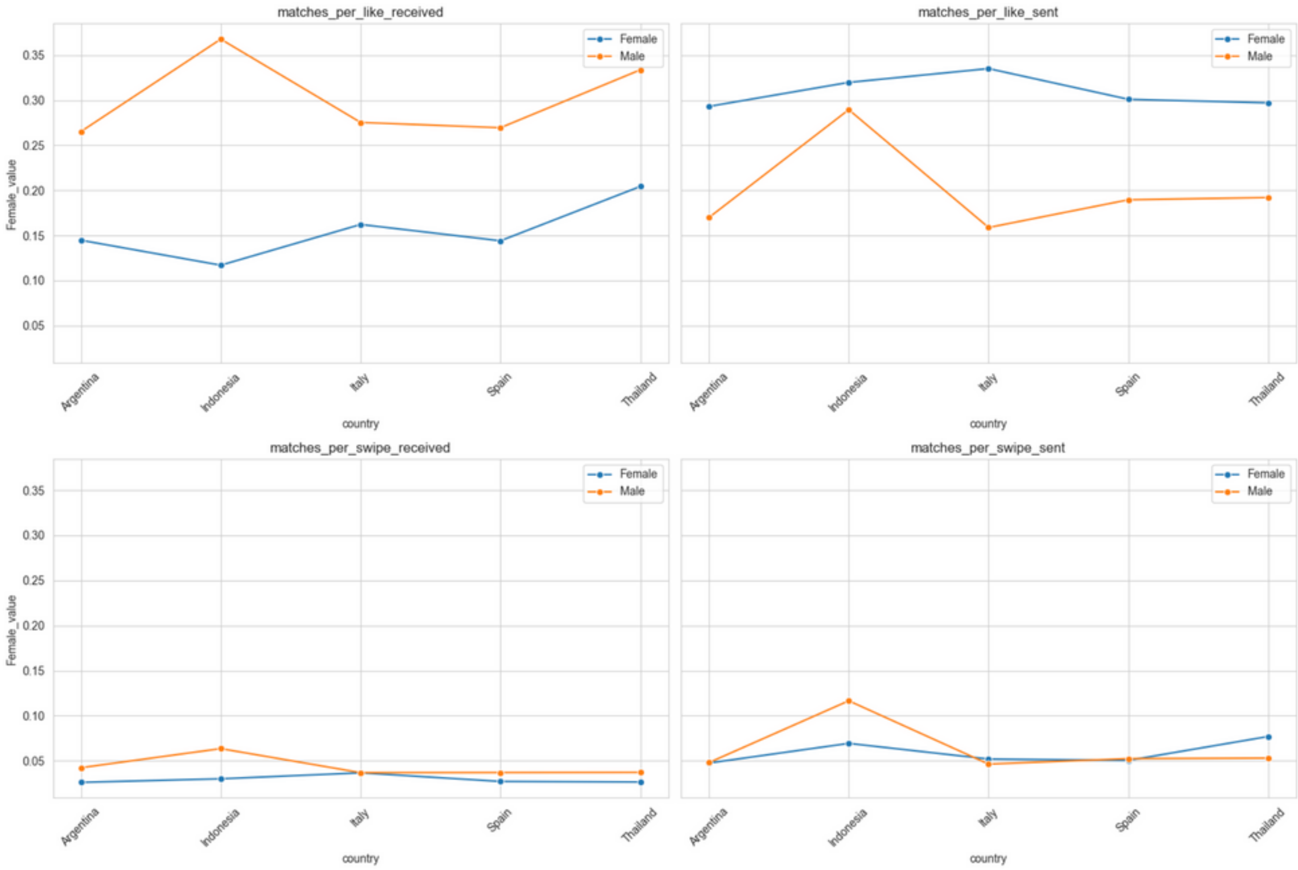
Demographic composition by country





Most countries have a balanced gender ratio but Indonesia(3.5 males : 1 female) and Thailand (2 males : 1 female) seem to have a lot of male users . The age distribution across countries also looks fairly similar. Argentina, Spain and Italy tend to have more number of older people as compared to Indonesia and Thailand (more young users)

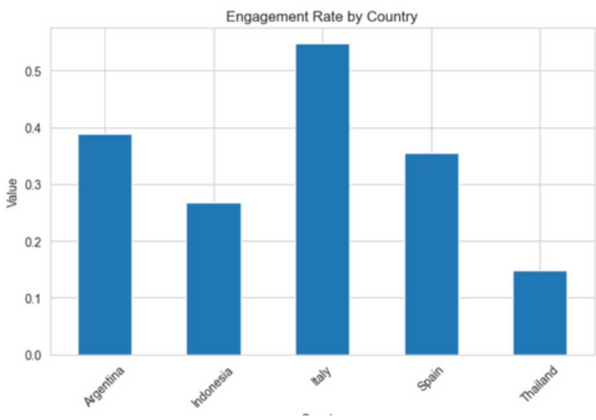
Understanding User Similarities and Differences across countries



The plots above represent how matches metrics are changing for countries based on gender. Swipes aren't really affecting matches. But we can see that likes_sent/received's impact on matches show different patterns for males and females where likes_sent is critcal for females and likes_received is more important for males. We are able to validate the same when we run regression later in the process.

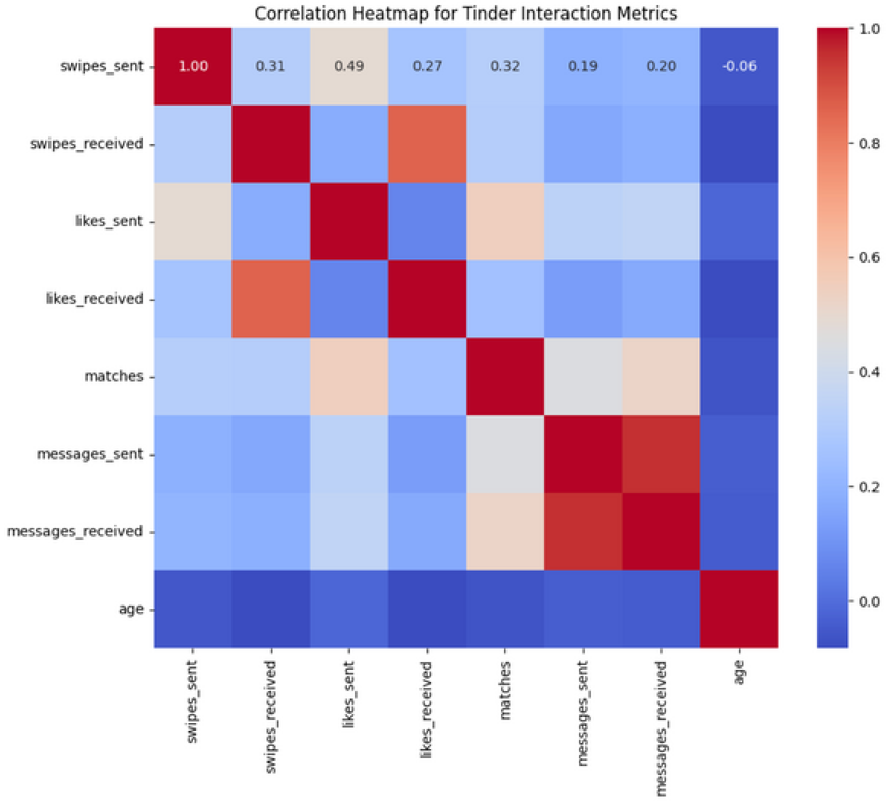
We can create another metric to understand how many swipes it takes to get a like:

engagement_rate=#likes_received/ #swipes sent



Optimizing Metrics for more matches

In order to increase the match count, we first need to understand what is primarily affecting the number of matches. The correlation plot shows the relationship between all numerical variables. We see matches has a decently positive correlation of around 55% with ***likes_sent*** and ***messages_received***. Other factors do not really seem to affect matches a lot. As we know correlation does not imply causation, we perform a regression on matches to see which attributes are critical and significant in explaining the behaviour of the number of matches.



Based on the regression of the overall data of active users on matches, swipes_sent turns out to be a statistically insignificant variable in explaining matches. Other variables are statistically significant but the strength of relationship is not very high. ***messages _received*** and ***likes_sent*** are ***positively*** correlated with a moderate relationship of 12% and 7%. The number of ***messages_sent*** shows a ***negative*** impact on matches. The R-squared and adjusted R-squared is around 47.5% which implies that the given set of variables can explain 47.5% of variance in the value of matches

Dep. Variable:	matches	**R-squared:**	0.475
Model:	OLS	**Adj. R-squared:**	0.475
Method:	Least Squares	**F-statistic:**	3235.
Date:	Sat, 20 Apr 2024	**Prob (F-statistic):**	0.00
Time:	20:28:49	**Log-Likelihood:**	-60970.
No. Observations:	21487	**AIC:**	1.220e+05
Df Residuals:	21480	**BIC:**	1.220e+05
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0743	0.034	2.175	0.030	0.007	0.141
swipes_sent	-0.0001	0.000	-0.727	0.467	-0.000	0.000
swipes_received	0.0027	0.000	12.419	0.000	0.002	0.003
likes_sent	0.0739	0.001	66.388	0.000	0.072	0.076
likes_received	0.0026	0.001	5.210	0.000	0.002	0.004
messages_sent	-0.0750	0.003	-29.390	0.000	-0.080	-0.070
messages_received	0.1268	0.003	49.029	0.000	0.122	0.132

Optimizing Metrics for more matches

The next step was to perform two different regressions based on gender. This thought essentially supported the notion that males and females might have different match behaviour and and there might be different factors that were important to the two segments.

For the Male segment,

1. **Insignificance of Likes Sent:**
   - The metric likes_sent has a p-value greater than 0.05, which indicates that within the active male segment, the number of likes sent does not have a statistically significant effect on the number of matches.
2. **Positive Impact of Messages Received:**
   - Messages_received is positively correlated with the number of matches. This suggests that men who receive more messages tend to have more matches.
3. **Positive Impact of Likes Received:**
   - Despite the insignificance of likes sent, it appears that likes_received has a positive impact on the number of matches, suggesting that the likes that male users receive from others may lead to more matches.
4. **Negative Impact of Messages Sent:**
   - There is a slight negative correlation between messages_sent and the number of matches. This implies that sending more messages does not necessarily lead to more matches for men, and may in fact be counterproductive.
5. **Increase in Model's Explanatory Power:**
   - The R-squared value has increased to 50.2% from a previous 47.5%. This means that the model for the active male segment explains 50.2% of the variability in the number of matches, which is an improvement compared to the combined model that includes all genders.

s

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                matches   R-squared:                       0.502
Model:                            OLS   Adj. R-squared:                  0.502
Method:                 Least Squares   F-statistic:                     2370.
Date:                Sat, 20 Apr 2024   Prob (F-statistic):               0.00
Time:                        18:58:29   Log-Likelihood:                -38610.
No. Observations:               14088   AIC:                         7.723e+04
Df Residuals:                   14081   BIC:                         7.729e+04
Df Model:                           6
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 0.1757      0.039      4.562      0.000       0.100       0.251
swipes_sent          -0.0018      0.000     -8.955      0.000      -0.002      -0.001
swipes_received       0.0002      0.000      0.703      0.482      -0.000       0.001
likes_sent            0.0646      0.001     54.115      0.000       0.062       0.067
likes_received        0.0430      0.001     32.686      0.000       0.040       0.046
messages_sent        -0.0329      0.003    -11.657      0.000      -0.038      -0.027
messages_received     0.0787      0.003     25.860      0.000       0.073       0.085
==============================================================================
Omnibus:                    15148.603   Durbin-Watson:                   1.298
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          5254273.363
Skew:                           4.878   Prob(JB):                         0.00
Kurtosis:                      97.106   Cond. No.                         312.
==============================================================================
```

Optimizing Metrics for more matches

For the Female Segment,

1. **Swipes Sent/Received Are Insignificant:**
    - For active female users, both swipes_sent and swipes_received are statistically insignificant, with p-values greater than 0.05. This suggests that the act of swiping, both sent and received, does not have a meaningful impact on the likelihood of matches for women.
2. **Positive Impact of Messages Received:**
    - Messages_received shows a significant positive correlation with matches. This implies that females who receive more messages are more likely to have more matches, indicating that received messages are an important predictor of match success.
3. **Positive Impact of Likes Sent:**
    - Similarly, likes_sent is significantly correlated with an increase in matches. This indicates that when active female users send likes, it contributes positively to the number of matches they receive.
4. **Negative Impact of Messages Sent:**
    - The regression model indicates a negative relationship between messages_sent and matches for females. This could mean that sending more messages does not lead to more matches for female users and might even decrease the chances of getting matches.
5. **Increase in Model's Explanatory Power:**
    - The R-squared value for the female-specific model is 57.1%, which is higher than the combined model's R-squared of 47.5%. This indicates that the female-specific model has better explanatory power and is more effective at accounting for the variance in the number of matches for females.
6. **Implications for Strategy:**
    - These findings suggest that for active female users, engagement in terms of likes sent and messages received should be encouraged to increase matches, while strategies to increase swiping activity or messages sent may not be as effective.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                matches   R-squared:                       0.571
Model:                            OLS   Adj. R-squared:                  0.571
Method:                 Least Squares   F-statistic:                     1405.
Date:                Sat, 20 Apr 2024   Prob (F-statistic):               0.00
Time:                        19:15:10   Log-Likelihood:                -21064.
No. Observations:                7399   AIC:                         4.214e+04
Df Residuals:                    7391   BIC:                         4.220e+04
Df Model:                           7
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 0.3862      0.190      2.034      0.042       0.014       0.758
swipes_sent          -0.0005      0.000     -2.374      0.018      -0.001   -9.56e-05
swipes_received       0.0007      0.000      1.837      0.066   -4.44e-05       0.001
likes_sent            0.1286      0.003     51.121      0.000       0.124       0.134
likes_received        0.0035      0.001      4.642      0.000       0.002       0.005
messages_sent        -0.1584      0.005    -30.564      0.000      -0.169      -0.148
messages_received     0.1957      0.005     42.320      0.000       0.187       0.205
age                  -0.0218      0.007     -3.229      0.001      -0.035      -0.009
==============================================================================
Omnibus:                    10087.255   Durbin-Watson:                   1.372
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          9511366.389
Skew:                           7.282   Prob(JB):                         0.00
Kurtosis:                     178.042   Cond. No.                     1.85e+03
==============================================================================
```

Optimizing Metrics for more matches

The increased R-squared in the gender-specific model compared to the combined model emphasizes the importance of gender-specific strategies in optimizing user engagement to increase matches. The data suggests that the behavior leading to matches is distinct for female users compared to the aggregate or male users.

For Both Genders:

- **Optimize for Likes Sent and Messages Received:** Since both likes_sent and messages_received show a positive correlation with matches and are statistically significant in the regression models, encouraging users to send more likes and engage in conversations that lead to receiving messages could increase match rates. This could be facilitated through UI/UX design choices that promote these interactions, such as highlighting the like button or notifications for received messages.

For Active Male Segment:

- **Focus on Message Quality:** Since messages_sent shows a slight negative impact, this suggests that indiscriminate messaging may not be effective. Instead, the platform could encourage personalized, meaningful interactions.
- **Likes Sent Insignificance:** The insignificance of likes sent suggests that men may need to be more selective with their likes or that the mechanism of sending likes could be less influential than the response they elicit.

For Active Female Segment:

- **Disregard Swipes:** Since swipes sent and received do not significantly affect matches for females, the focus should be shifted away from just swiping.
- **Encourage Receptive Communication**: With messages_received being critical, strategies that make females more receptive to messages might help. This could be in the form of prompts or notifications when they have unread messages.

General Strategy:

- **User Engagement Features:** Develop features that encourage the behaviors correlated with higher matches, such as highlighting potential matches or incentivizing users to be more active in conversations.
- **Educate Users**: Provide tips on creating engaging profiles and initiating conversations that are more likely to receive responses.
- **Behavioral Insights:** Leverage the data to give personalized insights to users, helping them understand what actions might increase their chances of getting matches based on aggregate user behavior data.

Optimizing Metrics for more matches

Next Steps:

1. **Engagement Score Variable:**
    - Develop an engagement score that combines the most critical metrics identified from the analysis (e.g., swipes, messages, likes).
    - This composite score could be used within the app's algorithms to prioritize users in match suggestions or to identify and reward highly engaged users.
2. **Cultural Customization:**
    - Adapt the app's features and marketing strategies to reflect the cultural nuances and user preferences in different countries.
3. **Gender-Specific Features:**
    - Consider introducing or emphasizing features that cater to the different behaviors observed between genders.
4. **Incentivize Messaging:**
    - Given that messaging behavior has shown to correlate with matches, consider analyzing the impact of incentivizing users to send the first message or respond to messages more quickly or provide message templates as a conversation starter
5. **A/B Testing for Feature Engagement:**
    - Use A/B testing to empirically determine the effect of new features or changes on user engagement and matches.
    - For example, test different methods of incentivizing messaging to see which is most effective at increasing engagement
6. **Longitudinal Data Analysis:**
    - With data spanning multiple years, we can conduct a time-series analysis to discern more stable patterns and trends that transcend seasonal variations.
    - This longer-term data would also allow for cohort analysis to understand the lifecycle of a user's activity and preferences over time.
7. **User Behavior Insights:**
    - Offer users insights into their behavior and how it compares to successful patterns, motivating them to adapt their activity to increase match likelihood.
8. **Personalized Recommendations:**
    - Tailor the app's recommendation algorithms based on the insights from the gender-specific and country-specific analyses, A/B Test results, previous matches to improve match rate

I'm glad we matched!