Decoding Online Attraction: Key Factors Influencing User Likes on Dating App Profiles

"Using a linear regression model, can we determine if age, gender of interest, profile picture count, number of profile visits, verification status, the number of languages spoken, and country increase the number of likes on a dating app profile?"

INTRODUCTION

- Online dating platforms are increasingly popular, with millions of users seeking meaningful connections. However, many users struggle to optimize their profiles to maximize likes, a key measure of engagement on these apps.
- Profile characteristics like age, gender of interest, profile pictures, and verification status are believed to influence user engagement, yet there's limited empirical research quantifying these effects.

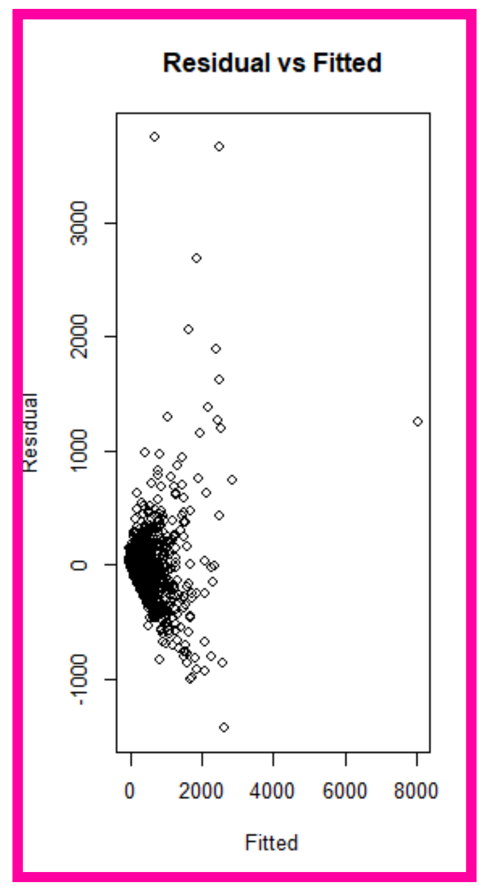
- Users lack clear, data-driven insights into which profile attributes significantly impact likes.
- Existing studies focus on general trends (e.g., attractiveness, messaging behaviour) but fail to offer predictive models for specific platforms.
- Can we build a linear regression model to identify the significant predictors of profile likes, helping users optimize their dating app profiles?

- Understanding these factors could improve user experience and app design.
- The study builds on existing literature (e.g., Castro et al., 2020; Hitsch et al., 2010) by providing actionable, predictive insights for users and app developers.

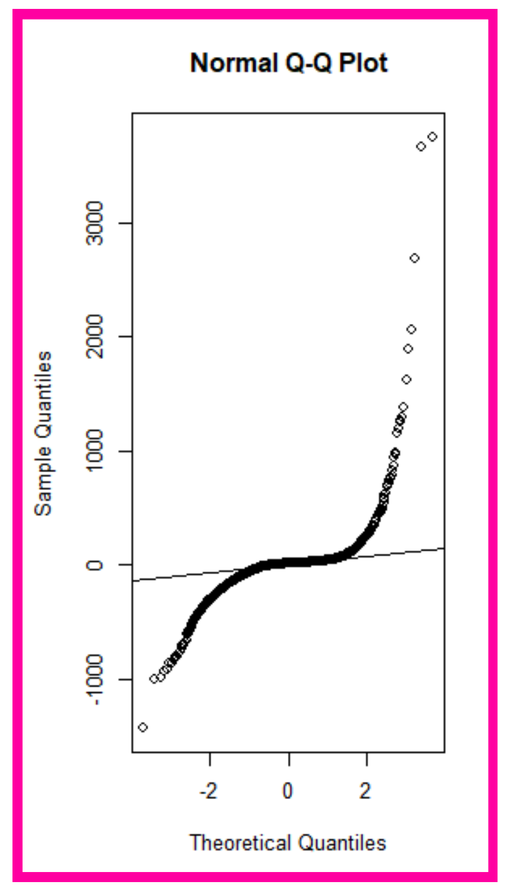
METHODS

Model Construction:

- Built preliminary multiple linear regression model using predictors identified from EDA and literature.
- Predictors: *profile visits, number of profile pictures, verification status, number of languages spoken, & country population.*



Residual vs Fitted plot to evaluate LR assumptions: Linearity, Constant Variance, & Uncorrelated Errors



Normal Q-Q Plot to evaluate LR Normality assumptions

Model Refinement:

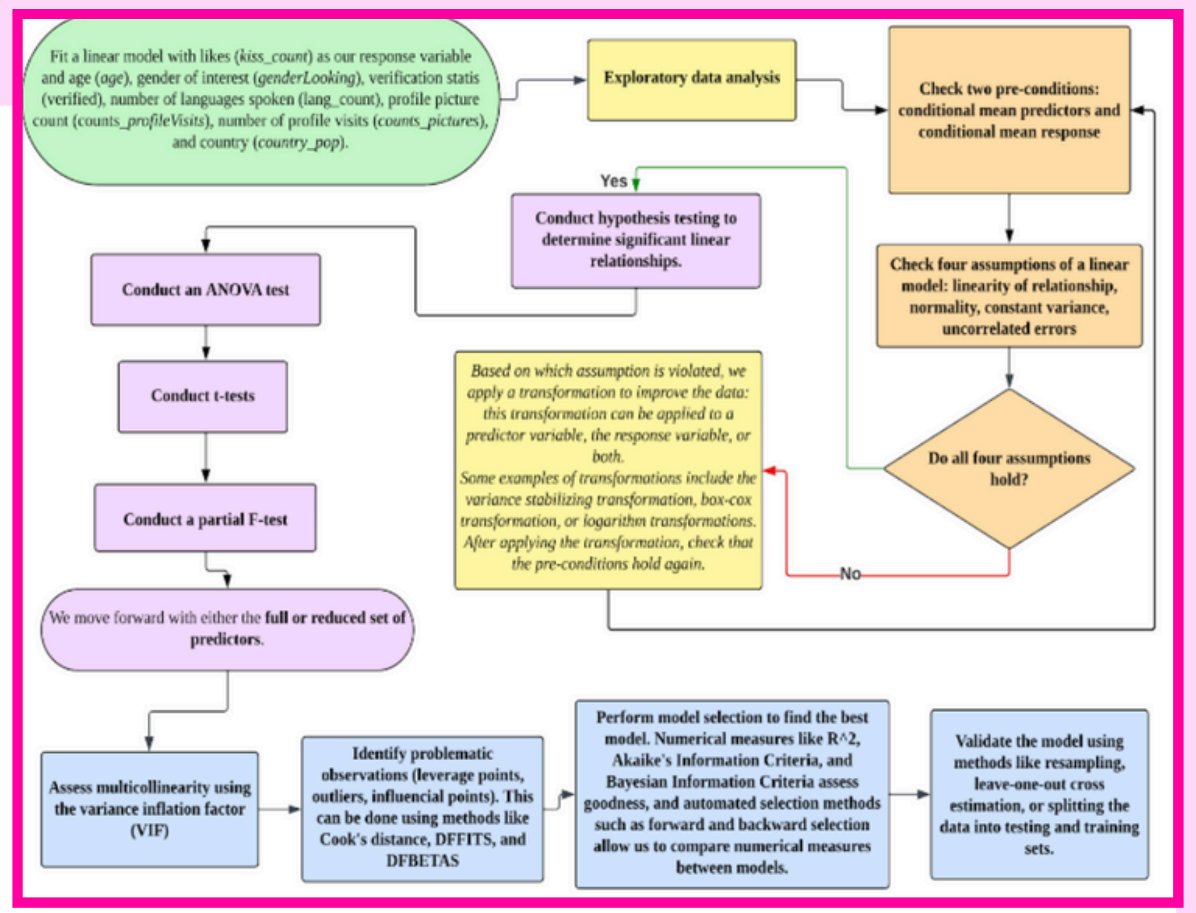
- Conducted ANOVA test to determine if selected predictors explained significant variation in likes.
- Used Partial F-test to compare full model with reduced models, retaining the simpler model if it explained as much variability.

Multicollinearity Check:

- Calculated Variance Inflation Factors (VIF) to detect multicollinearity. Predictors with VIF > 5 were iteratively removed to ensure reliable coefficient estimates.

Influence Diagnostics:

- Identified leverage points, outliers, and influential observations using metrics like Cook's Distance, DFFITS, and DFBETAS.
- Removed problematic data points only if there was contextual justification.



Fit a linear model with likes (*kiss_count*) as our response variable and age (*age*), gender of interest (*genderLooking*), verification statis (verified), number of languages spoken (lang_count), profile picture count (counts_*profileVisits*), number of profile visits (*counts_pictures*), and country (*country_pop*).

Exploratory data analysis

Check two pre-conditions: conditional mean predictors and conditional mean response

Check four assumptions of a linear model: linearity of relationship, normality, constant variance, uncorrelated errors

Do all four assumptions hold?

Yes

No

Based on which assumption is violated, we apply a transformation to improve the data: this transformation can be applied to a predictor variable, the response variable, or both. Some examples of transformations include the variance stabilizing transformation, box-cox transformation, or logarithm transformations. After applying the transformation, check that the pre-conditions hold again.

Conduct hypothesis testing to determine significant linear relationships.

Conduct an ANOVA test

Conduct t-tests

Conduct a partial F-test

We move forward with either the **full or reduced set of predictors.**

Assess multicollinearity using the variance inflation factor (VIF)

Identify problematic observations (leverage points, outliers, influencial points). This can be done using methods like Cook's distance, DFFITS, and DFBETAS

Perform model selection to find the best model. Numerical measures like R^2, Akaike's Information Criteria, and Bayesian Information Criteria assess goodness, and automated selection methods such as forward and backward selection allow us to compare numerical measures between models.

Validate the model using methods like resampling, leave-one-out cross estimation, or splitting the data into testing and training sets.

Analysis Flowchart

Model Selection:

- Employed manual selection techniques to refine the model based on statistical metrics:
  - R² (coefficient of determination)
  - Akaike's Information Criterion (AIC)
  - Adjusted Akaike's Information Criterion (AICc)
  - Bayesian Information Criterion (BIC)

Validation and Generalizability:

- Split the data into training and testing sets to assess generalizability.
- Calculated adjusted R² and Mean Squared Error (MSE) for both sets to validate predictive performance.

Tools and Techniques:

- Analysis conducted in R, leveraging packages for regression modeling, diagnostics, and visualization.

RESULTS

Initial Data Cleaning and Exploration:

- Renamed & created variables for clarity and consistency.
- Conducted EDA using histograms to identify variable distributions and potential issues.

Preliminary Model and Assumption Checks:

- Selected seven predictors: *age, gender of interest, profile picture count, profile visits, verification status, language count, and country population.*
- Assessed LR assumptions revealing normality and constant variance violations in the initial model
- Box-Cox transformations on key variables to solve:
  - Counts of kisses (power = 0.0882).
  - Country population (power = 0.1184).
  - Profile visits (power = 0.2539).

Significance Testing and Model Refinement:

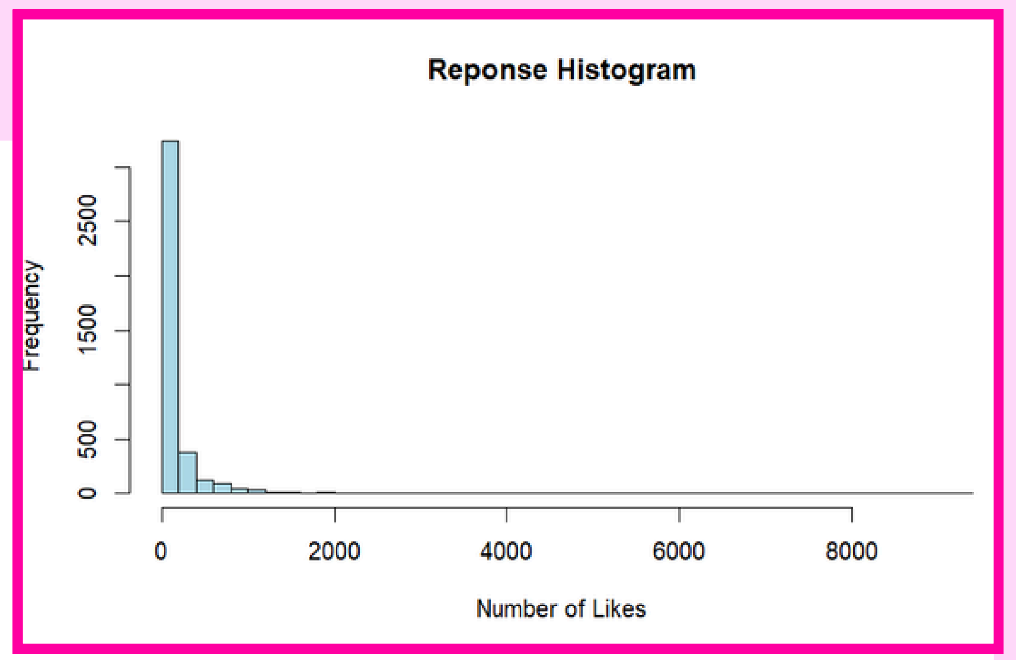
- ANOVA test, yielded a highly significant p-value (2.2e^-16), confirming at least one predictor was significantly related to the response.
- T-tests and Partial F-tests, found all predictors were significant and no variables could be removed without loss of explanatory power.

Multicollinearity and Diagnostic Checks:

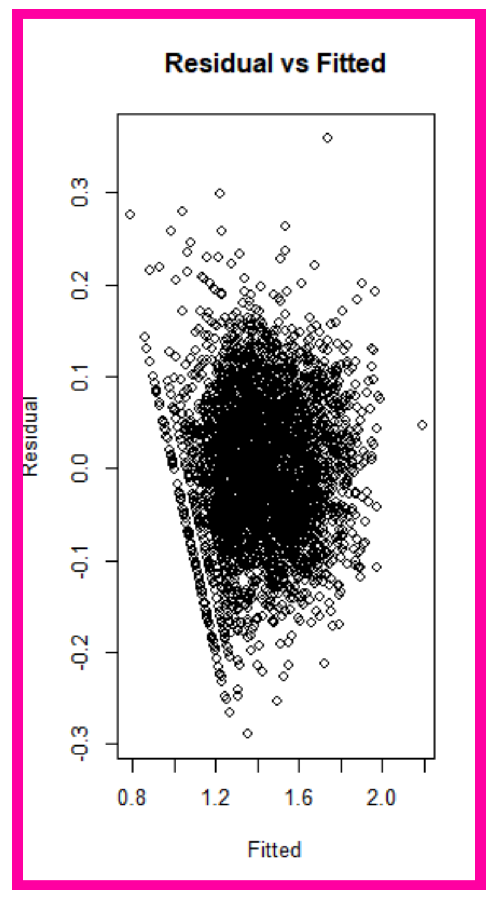
- All predictors had VIF values below the threshold of 5, with a maximum of 1.378.
- Identified 189 leverage points and 1 outlier; none were influential or required removal based Cook's Distance and DFBETAS.

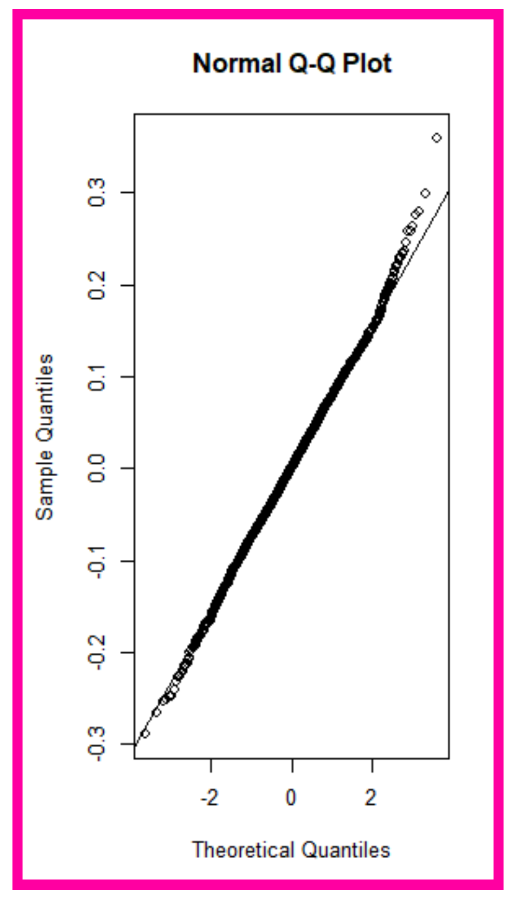
Final Model Selection:

- Compared models using adjusted R², AIC, adjusted AIC, and BIC.

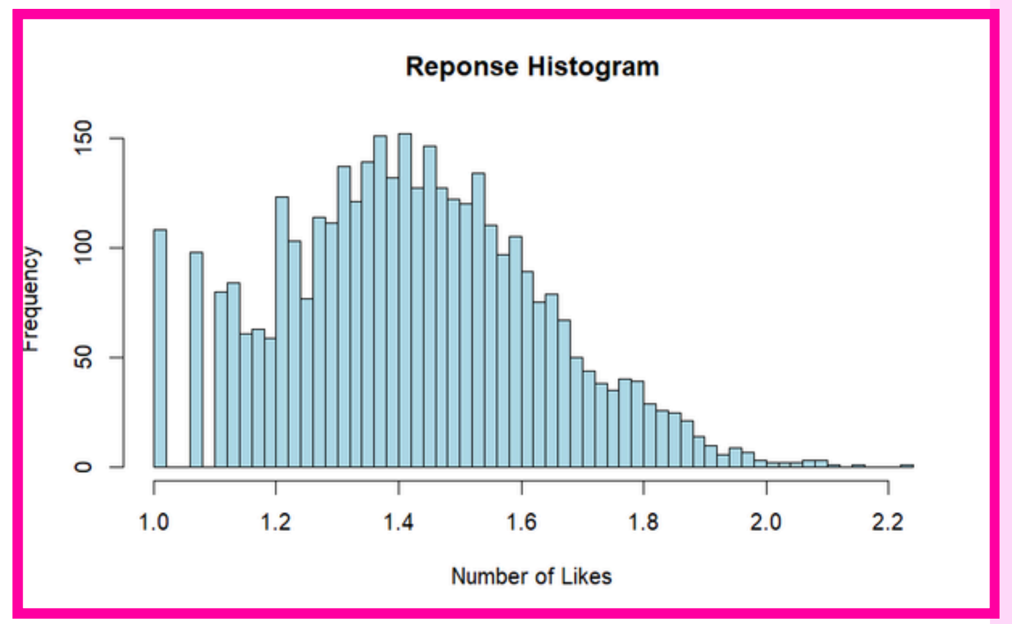


Issue: Right-skewed distributions across response

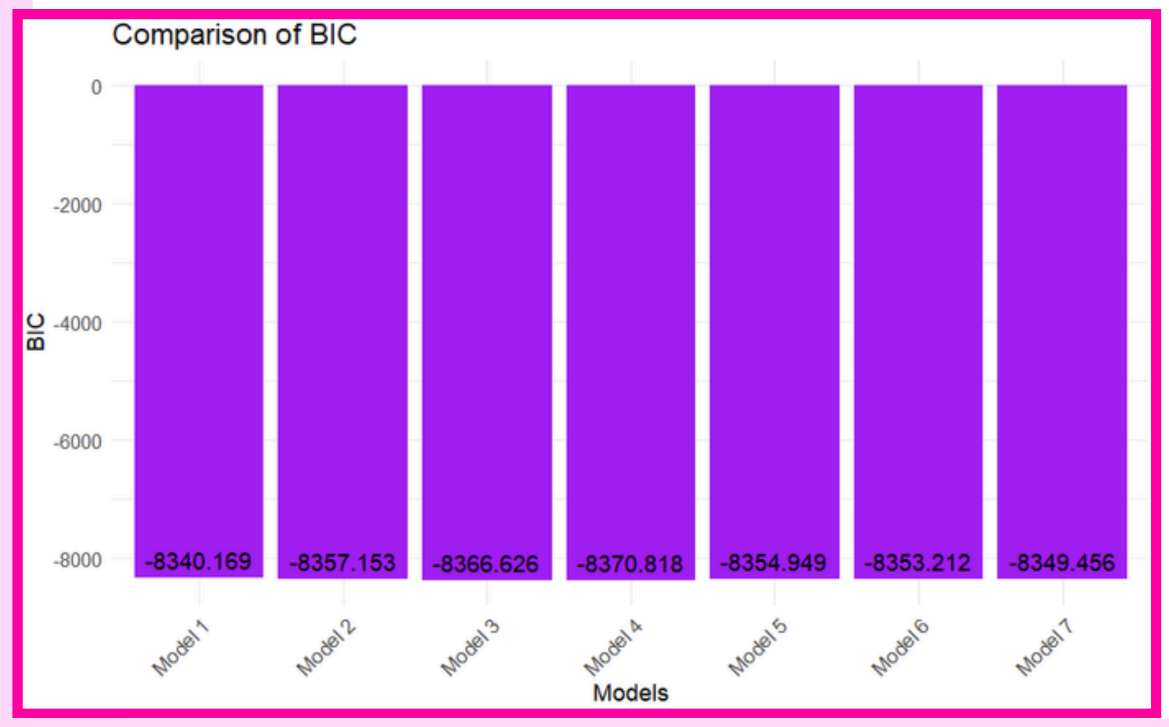




Post-Transformation Residual & Q-Q Plots showed significant improvements in normality and constant variance.



Response After Box-Cox Transformation



The full model and a four-predictor model (*age, country population, profile visits, verification status*) performed equally well, with the four-predictor model slightly outperforming on BIC.

Model Validation:

- Split data into training and testing sets to validate the final model with results:
  - Adjusted R²: 0.8625 (training), 0.8627 (testing).
  - MSE: 0.00596 (training), 0.0839 (testing).
- Demonstrated strong model fit and generalizability.

CONCLUSIONS

Summary of Findings:

- The final four-predictor model (*age, country population, profile visits, and verification status*) demonstrated the strongest ability to predict the number of likes on a dating app profile.
- A significant linear relationship was identified between the response variable (likes) and the predictors in this model.
- Predictors *gender of interest, profile picture count,* and *number of languages spoken* showed no statistically significant relationship with user likes.

Significance of Results:

- Hypothesis Supported: The analysis confirms that a linear regression model can effectively identify factors influencing user likes on dating apps.
- Practical Implications:
  - Highlights key features (e.g., profile visits, verification status) users and app developers might prioritize for optimizing user engagement and experience.
  - Provides actionable insights into what contributes to a successful profile in the online dating ecosystem.

Model Limitations and Considerations:

- Problematic Data Points:
  - A total of 189 leverage points were identified, with 180 flagged as influential on their own coefficients. These points may affect coefficient reliability.
- Data Collection Bias:
  - Data was sourced from a controversial method using two male profiles, resulting in a dataset disproportionately composed of women with a male preference in the "gender of interest" variable.
  - This sampling method limits the generalizability of the findings across broader dating app user demographics.

Suggestions for Future Work:

- Address data quality issues by collecting a more balanced dataset using less biased sampling methods.
- Explore other potential predictors and interactions, or consider non-linear models to capture complex relationships.
- Incorporate robust diagnostic techniques to better handle problematic points in the data.

REFERENCES

Castro, Á., Barrada, J. R., Ramos-Villagrasa, P. J., & Fernández-del-Río, E. (2020). Profiling dating apps users: Sociodemographic and personality characteristics. International Journal of Environmental Research and Public Health, 19(3), 1575. https://doi.org/10.3390/ijerph17103653

Ellison, N., Heino, R., & Gibbs, J. (2006). Managing impressions online: Self-presentation processes in the online dating environment. Journal of Computer-Mediated Communication, 11(2), 415-441. https://doi.org/10.1111/j.1083-6101.2006.00020.x

Hitsch, G. J., Hortaçsu, A., & Ariely, D. (2010). What makes you click? Mate preferences in online dating. Quantitative Marketing and Economics, 8(4), 393-427. https://doi.org/10.1007/s11129-010-9088-6

Kaggle dataset: Mabilama Jeffrey Mvutu . (2015). Dating App Lovoo User Profiles. Kaggle. https://www.kaggle.com/datasets/jmmvutu/dating-app-lovoo-user-profiles

Original source of the Kaggle dataset: Jfreex. (2015). Dating App User Profiles' Stats - Lovoo v3. Data World. https://data.world/jfreex/dating-app-user-profiles-stats-lovoo-v3

Population data: World Bank. (2024). Population, total (SP.POP.TOTL). https://data.worldbank.org/indicator/SP.POP.TOTL?locations=1W