Final Project Part 1

Raihan Amin, Rebecca Li, and Ebrahim Mohamud

Friday October 10 2024

Contributions

- Rebecca Li: Data description, Ethics
- Raihan Amin: Introduction, Data Description, Sourcing
- Ebrahim Mohamud: Preliminary results, Ethics, Coding

Introduction

As online dating becomes widespread, understanding the dynamics that influence user engagement is increasingly important. The factors that enhance profile attractiveness can significantly impact users' outcomes in the digital dating landscape. Within this paper, we will explore the following research question: *"How do factors such as age, gender of interest, profile picture count, number of profile visits, and country influence the number of user likes on a dating app profile?"*

Existing literature highlights various elements that contribute to user interactions on dating platforms. For instance, Castro et al. (2020) examine the socio-demographic characteristics of dating app users, showing that younger users and women receive more likes, and that country of origin significantly influences interactions. This highlights the importance of age and cultural factors in understanding user engagement, aligning with the proposed research question. Hitsch et al. (2010) explore mate preferences in online dating, showing that the number of profile pictures significantly affects user clicks and likes, with preferences often based on opposite gender and shared country backgrounds. Their findings highlight the importance of self-presentation in attracting user engagement. Ellison et al. (2017) also investigate self-presentation strategies, finding that users who curate their profiles by displaying more pictures and increasing profile visibility tend to receive more interactions, directly connecting to our research question by highlighting the significance of user visibility through profile visits. Together, these articles reinforce the importance of age, gender, picture counts, profile views, and country preference in online dating, providing a strong foundation for the current research.

To effectively analyze these relationships, linear regression is employed as the statistical tool of choice, as it allows for the examination of the relationship between several predictor variables and a discrete response variable, in this case, the number of user likes. It not only identifies the strength and direction of these relationships but also helps control for confounding factors, providing a clearer understanding of the dynamics in online dating interactions. By leveraging linear regression, we aim to uncover predictions that could inform users and developers about effective strategies for enhancing profile appeal in the competitive landscape of online dating.

Data Description

This dataset, titled *"Dating App User Profiles' stats - Lovoo v3"*, was sourced from the online database Kaggle. The author collected data from the online dating platform Lovoo by creating two male profiles and using software to store data from swiping through recommended matches, in order to understand what makes a good profile, whereas our current research analyzes how factors influence user likes.

Our chosen response variable, *counts_kisses*, is a quantitative variable that represents the number of users that liked a profile, ranging from 0 to 9288. The profile with 9288 kisses was an outlier, thus the mean of *counts_kisses* was 156.6, drastically higher than the median of 44.0. A linear regression model is suitable for the response variable *count_kisses*, as the values it takes rely on multiple predictor variables. Based on the existing literature, we anticipated these variables would be linearly related.

Table 1: Summary Statistics for Predictor Variables

Predictor	**Summary**
Min, 1st Quart, Median, Mean, 3rd Quart, Max	
1. counts_profileVisits (number of clicks on user's profile)	$\{0, 383, 1222, 3705, 4063, 164425\}$
2. counts_pictures (number of pictures on user's profile)	$\{0.00, 2.00, 4.00, 4.79, 6.00, 30.00\}$
3. age (in years)	$\{18.00, 20.00, 22.00, 21.99, 24.00, 28.00\}$
4. country_pop (population of countries in thousands)	$\{39.58, 8850, 68170, 49200, 84480, 1429000\}$
5. genderLooking (gender that user is interested in)	Female: 25, Male: 3580, Both: 21, None: 366

Count_profileVisits ranged from 0 to 164,425 clicks with a mean of 3,704, and *counts_pictures* ranged from 0 to 30 with a mean of 4.79, both predicted to be positively related to *counts_kisses*. Users' ages spanned 18 to 30 years, averaging 21.99, and from the research, we expected younger users to attract more *counts_kisses*. *Country_pop* varied from 3.958×10^4 to 1.429×10^9, with a mean of 4.920×10^7 people; we expected users from populated countries to obtain more *counts_kisses*. Lastly, *genderLooking*, which indicates a user's gender preference, showed a majority male, aligning with the research that female users likely obtain more *counts_kisses*.

We converted the dataset's categorical variable "country" into a numerical one by equating countries to their population, creating a new variable *country_pop*. We redesigned it as such because we found population to be a meaningful factor for our question that could be converted from countries effectively. Additionally, the dataset contained profiles from 32 countries, giving us a reasonable range to work with.

Ethics Discussion

The dataset presents several ethical concerns, particularly with privacy and consent. One issue is the presence of a variable labelled *"name"*, which records the exact names listed on user profiles. This compromises the anonymity of the participants. Additionally, although the data was collected from information that users voluntarily submitted to the app, individuals did not provide explicit consent for their profiles to be used in this research. This lack of informed consent raises questions about whether participants fully understood how their data would be utilized.

Despite these ethical issues, the potential impact on stakeholders appears minimal. The data was collected to analyze online dating optimization, a purpose harmless to users. Additionally, when looking at the trustworthiness of the source, there is no mention of being vetted by an external party; however, use could still be justified if handled with caution. By removing any personally identifiable information, ensuring results are presented in a way that maintains user anonymity, and focusing on the goal of improving user experiences on dating platforms, it is reasonable to consider the dataset ethically acceptable for analysis, provided that privacy concerns are addressed appropriately and safeguards are in place to minimize risks to participants.

Preliminary Results

We began our preliminary model by specifying the predictors of interest; the user's age, gender of interest, country's population, the number of their profile's pictures, and profile visits received. Using the `lm` function in R, we fit a linear regression model using these predictors to the `counts_kisses` variable. When viewing the `qqplot`, the response variables diverged significantly from a normal distribution, suggesting a violation of normality. Additionally, the fitted vs. residual plots showed extreme fanning, suggesting a constant variance violation. However, there were no other patterns to indicate a violation in the linearity assumption, and the data points seemed to maintain the uncorrelated errors assumption.

We proceeded to examine whether our interpretations were valid by checking two conditions: the relationship between predictors using the `pairs` function and the relationship between the response and the fitted values. We noticed a moderately strong linear relationship between the response and fitted values, suggesting a linear model is an apt choice. Additionally, we noticed no complex relationships between the five predictors, which further validates our model.

The preliminary results that we obtained through the dataset seemed consistent with pre-existing literature, suggesting that each of our predictor variables has a relationship to the response variable that is suitable for linear regression. As the research demonstrated, several of the predictor variables, such as country population and number of profile visits, demonstrated a positive relationship with the number of user likes, while `genderLooking` demonstrated that the model did not have a strong bias for predicting values in gender. While our findings generally followed the research, particularly when dealing with inter-relationships between predictors, there were some inconsistencies regarding the assumptions of linear regression. Our residual vs. fitted values plots demonstrated patterns such as fanning and skews, which were not addressed in any research that we investigated.

Bibliography

- 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`

- 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`

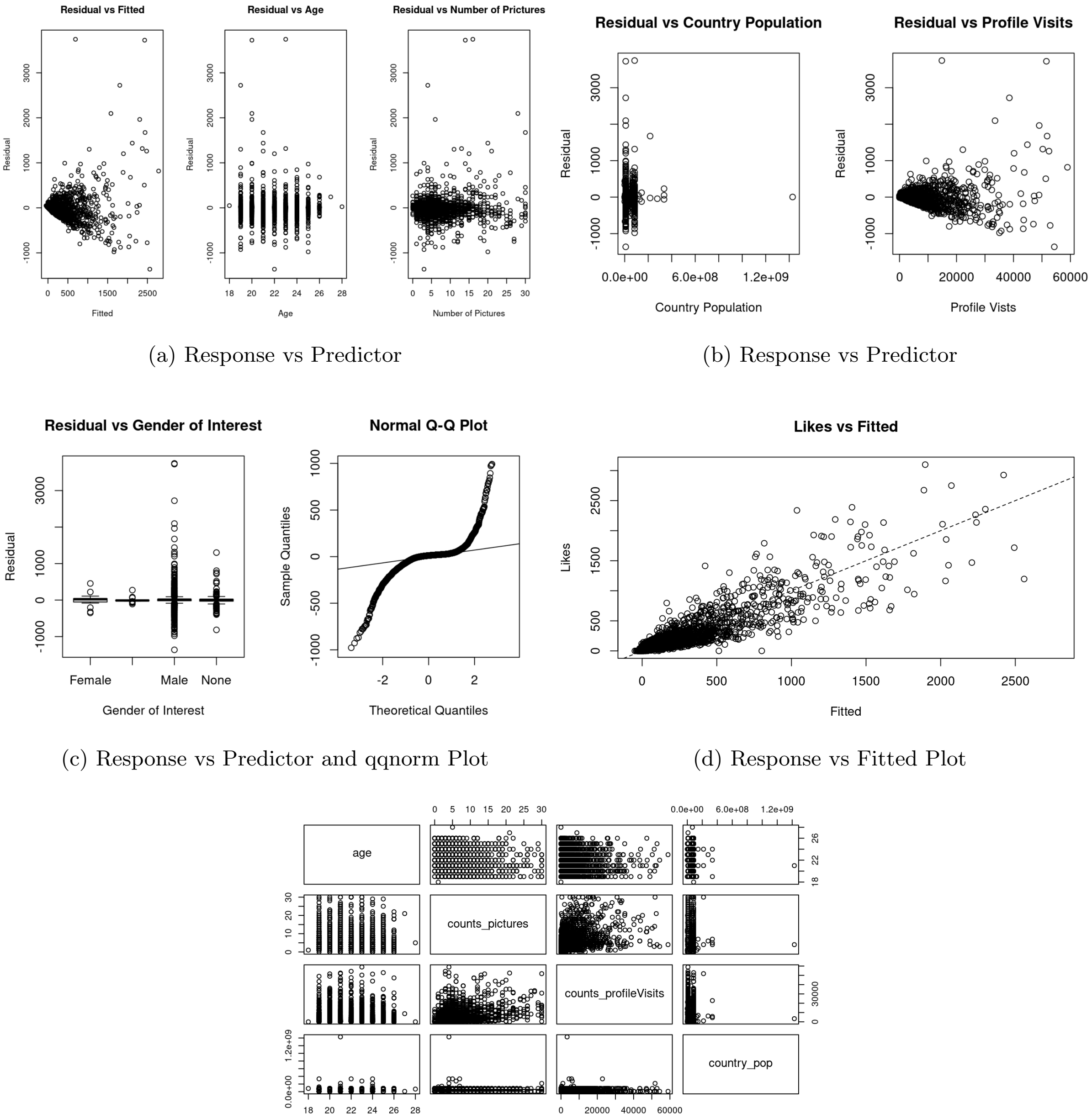
- Ellison, N., Heino, R., & Gibbs, J. (2006). Managing impressions online: Self-presentation processes in the online dating environment. *Journal of Computer-Mediated Communication*, *11*(2), 415-441. `https://doi.org/10.1111/j.1083-6101.2006.00020.x`

- 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`

Residual Plots



(a) Response vs Predictor

(b) Response vs Predictor

(c) Response vs Predictor and qqnorm Plot

(d) Response vs Fitted Plot

(e) Pairwise Scatter-plot

Figure 1: Residual Plot Analysis