Natalie LePera 22 Apr 2025
STA512 – Exam 2

**INTRODUCTION**

As the world becomes an ever-globalized landscape, the frequency and profitability of online dating apps has skyrocketed. A new heavy hitter has entered the saturated market of online dating apps, Pizzaz.com. Pizzaz sets out to utilize real world data as obtained from an in-person speed dating event to create a prediction model for use in their online dating platform. By analyzing each set of responses, the Pizzaz team is hoping to create two gender specific prediction models based on the dater rankings on key attributes to predict how much the dater will like the potential match on a scale of 1 – 10.

Employing this level of prediction not only ensures that Pizzaz.com users are presented with the highest quality potential matches, but also allows for user customization of the app experience. Through application settings, the user can select their desired "Like" cutoff value to include or exclude potential matches with lower predicted like scores. For example, one user may select a like minimum of 6 ensuring they will be presented with all potential matches that score a predicted like rating of 6 or higher, while another user may raise that like minimum to 9 ensuring only the matches with the highest predicted like rating are presented.

**DESCRIPTIVE STATISTICS**

The data collected for model creation was compiled from a group of 276 heterosexual couples paired at random during a short speed dating event. Each participant in each couple was asked to provide their personal information (age & race) as well as their opinions of their matched partner on the following characteristics on a scale of 1 - 10: attractiveness, sincerity, intelligence, fun, ambitiousness, & shared interests. Additionally,

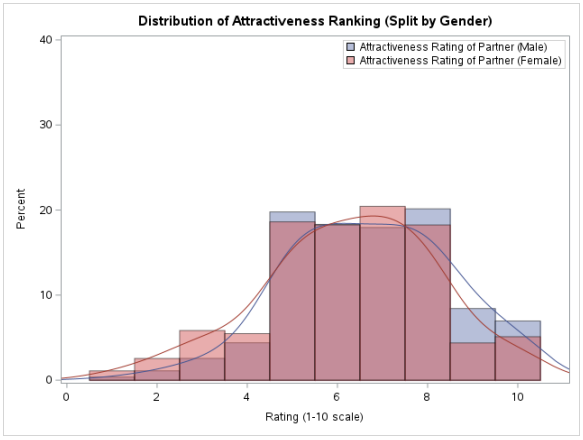
participants were asked to complete 2nd date rankings to indicate their decision of if they would like to see their matched partner again for another date and their best guess on a scale of 1 – 10 of if their matched partner would also like a second date. Unfortunately, due to data export errors these second date rankings were omitted from the data set and subsequent analysis.

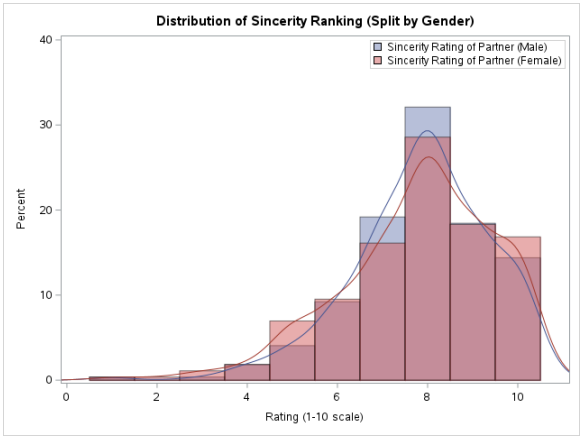
The data was split into two groups for analysis: male and female, as it has been previously determined that each gender may prioritize different characteristics when selecting a potential partner. The distribution of the personality characteristic variables are included below in Figure 1 with the scores recorded by male participants being represented in blue while the scores recorded by female participants are represented in red.

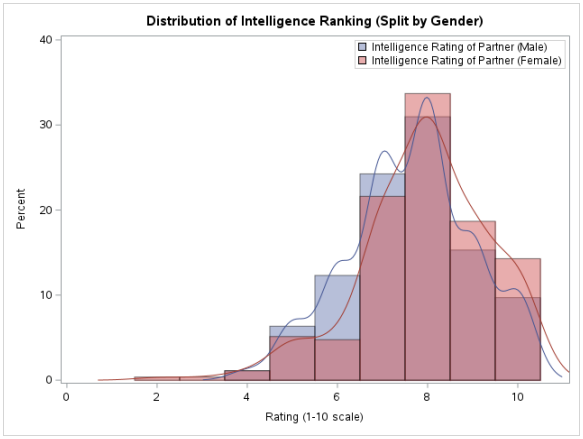
Natalie LePera

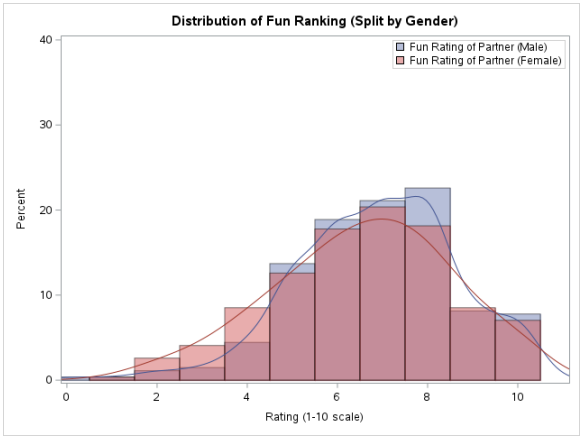
STA512 – Exam 2

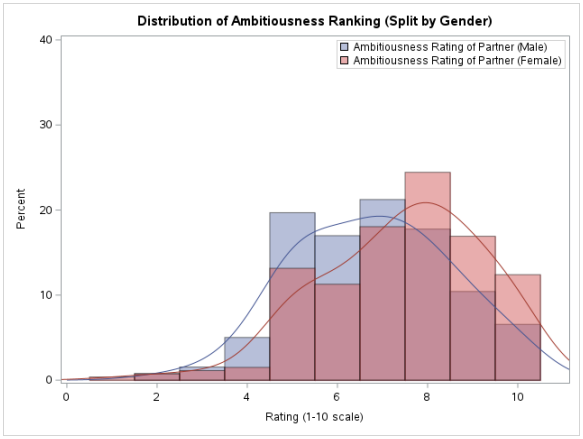
**Figure 1: Distribution Data for Personality Charachteristic Variables by Gender**

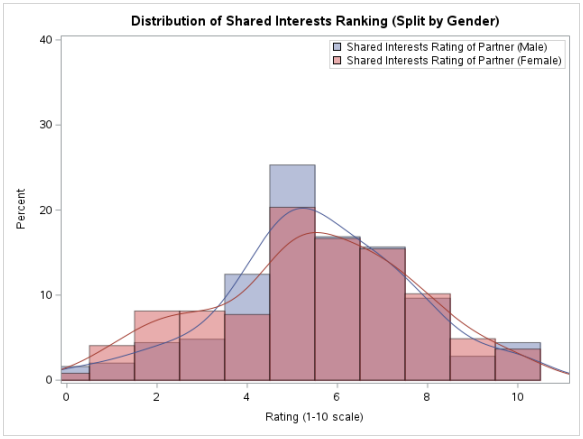


Natalie LePera 22 Apr 2025
STA512 – Exam 2

**Table 1: Missing Values**

**The MEANS Procedure**

| Variable | N Miss | N |
|---|---|---|
| LikeM | 2 | 274 |
| LikeF | 4 | 272 |
| AgeM | 3 | 273 |
| AgeF | 5 | 271 |
| AttractiveM | 3 | 273 |
| AttractiveF | 2 | 274 |
| SincereM | 5 | 271 |
| SincereF | 3 | 273 |
| IntelligentM | 8 | 268 |
| IntelligentF | 3 | 273 |
| FunM | 6 | 270 |
| FunF | 6 | 270 |
| AmbitiousM | 17 | 259 |
| AmbitiousF | 10 | 266 |
| SharedInterestsM | 27 | 249 |
| SharedInterestsF | 30 | 246 |
| smrace | 0 | 276 |
| closea | 0 | 276 |

Upon cursory analysis of the provided data there appeared to be no entry errors resulting in impossible values, but a number of missing values were identified. It is unknown weather these values were left blank by the speed dating participants or if the blank values are resultant of entry / export errors. A summary of missing observations per variable are included to the left in Table 1. N Miss indicates the number of blanks (missing observations) per variable while N indicates the number of complete observations per variable. As no imputation process was requested by the client, the missing observations were simply omitted from the model creation process rather than "filling in the gaps" through imputation. Future analysis may be conducted to determine the best possible imputation approaches for improved model creation and tuning.

As part of the client request, summary statistics were calculated to determine the impact of the couples' age and the impact of couples being of same or different races on the overall like ratings. A "close" age gap was defined as both parties' ages being within 2 years of each other. There were 155 interracial couples (56.16%) and 121 same race couples (43.84%) randomly paired for this speed dating event. In reviewing the correlation of the reported "like" scores for each group, there was found to be <0.1 correlation between race match and like scores, indicating negligible to no impact on reported like scores based on interracial vs same race parings. Additionally, 157 couples not close in age (56.88%) and 119 couples with close ages (43.12%) were randomly paired. The correlations for the age

gap closeness and reported “like” scores for each group were found to be <0.02 indicating near 0 impact on like ratings based on close or far age gaps.

**SELECTION OF THE MODELS AND TYPE OF ANALYSIS EMPLOYED**

Prior to any model creation the data was split into relevant male and female subgroups as outlined in Section 2, followed by subsequent splits into 80:20 (training : test) groups for each gender. This split into 221 training observations and 55 test observations for both male and female groups allowed for cross validation of the created models to ensure the models produce the best prediction values without over fitting to the data.

The initial male prediction model was created and found to violate the assumptions of regression, leading to squaring of the y variable and subsequent new model creation through stepwise forward selection to ensure no assumption violations.

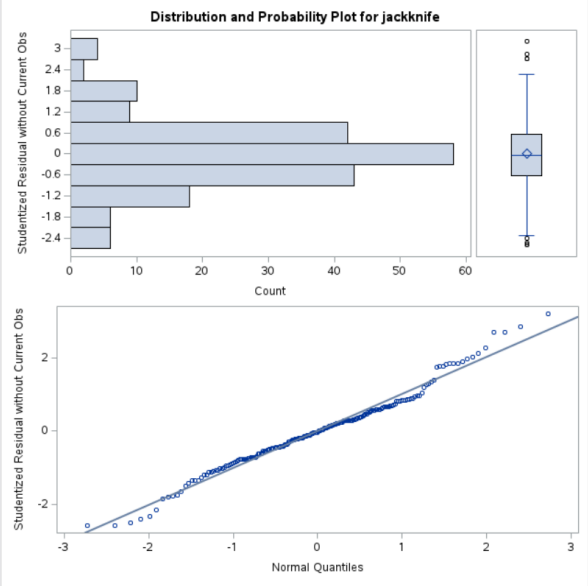
$$\text{Male Like Rating (M)} = \text{Fun} + (\text{Attractiveness} * \text{Sincerity}) + (\text{Attractiveness} * \text{Shared Interests}) + \text{Intercept}$$

Through analysis of this model, it was found that all assumptions of regression were met (residuals normally distributed with a mean of 0 and constant variance with independence between variables), and various outliers were identified based on the approach utilized (Figure 2 below). All the tests carried out in the below Figure 2 underscore the efficacy of the model created to predict male participants’ like ratings. Additionally, a collinearity test was conducted to ensure any variables in the model are not dependent on one another, which would result in inaccurate predictions. The variance indication factor (VIF) and condition index were calculated and found to be in passing ranges (VIF < 10 and Condition Index < 30), indicating no collinearity in the model. Finally shrinkage value was calculated

to assess the reliability of the model resulting in a shrinkage score of 0.0623. As the shrinkage was found to be < 0.1 this model can be considered reliable for predicting the male participants' like scores.

**Figure 2: Male Model Residual Analysis**

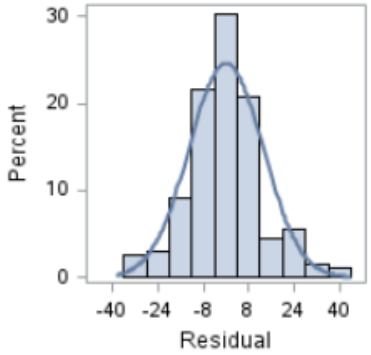
Normal Distribution Proof (Histogram, Box & Whisker Plot, & Q-Q Residual Plot)




Outlier Summary:

1. Jackknife = 13 (net -2 outliers from original)
2. Leverage = 10 (net -17 outliers from original)
3. Cooks distance = 0 outliers
4. Boxplot = 6 outliers

Proof of Normal Distribution




Residual distribution plot with normal curve overlaid indicates corrected skew and reduced kurtosis. Lower kurtosis value indicates fewer outliers and improved stability

Skew value = 0.24006157
Kurtosis = 0.86456372

Normal Distribution Tests

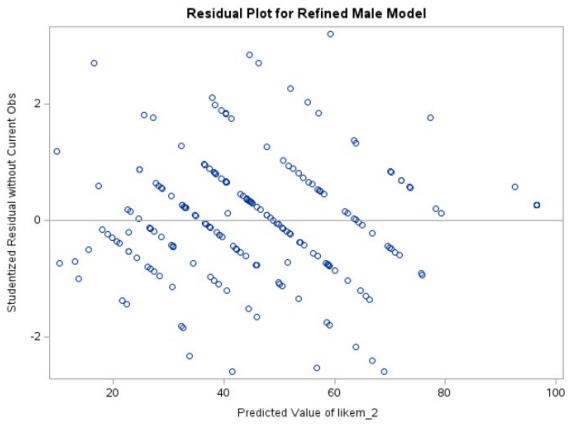
| Tests for Normality | | | | |
|---|---|---|---|---|
| Test | Statistic | | p Value | |
| Shapiro-Wilk | W | 0.980805 | Pr < W | 0.0082 |
| Kolmogorov-Smirnov | D | 0.062748 | Pr > D | 0.0559 |
| Cramer-von Mises | W-Sq | 0.164675 | Pr > W-Sq | 0.0165 |
| Anderson-Darling | A-Sq | 1.149228 | Pr > A-Sq | 0.0052 |

Kolmogorov-Smirnov test p >0.05 indicating normal distribution

*(H0: Normal Distribution, p not <0.05 -> cannot reject H0)*

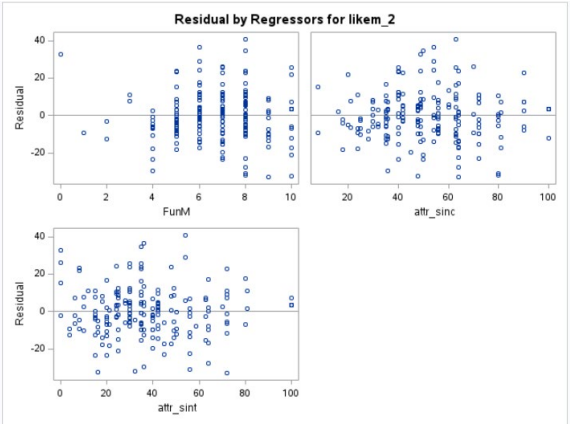
**Figure 2: Male Model Residual Analysis**

Residuals vs Predictions Test




Shows improved centering around y = 0, supports assumption mean of residuals = 0

Residuals vs Regressors Tests




Shows reduced funneling and improved homoscedasticity (assumption variance is constant, met)

The initial female prediction model was also found to violate the assumptions of regression, leading to squaring of the y variable and subsequent new model creation through stepwise forward selection to ensure no assumption violations.

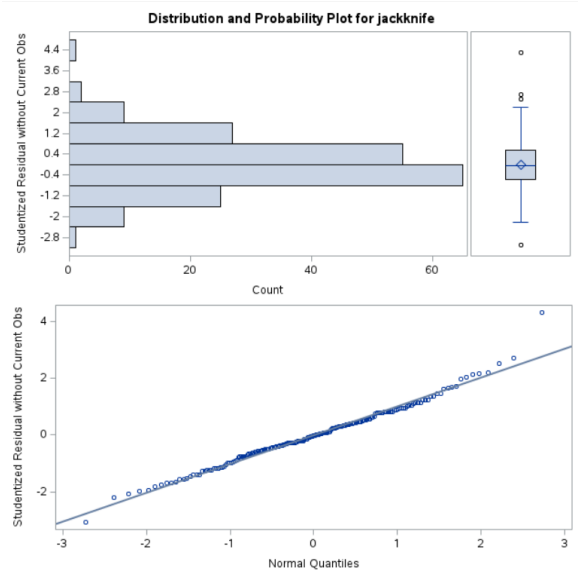
Female Like Rating (F) = (Attractiveness * Fun) + (Sincerity * Intelligence) + (Intelligence * Shared Interests) + Intercept

Through analysis of this model, it was found that all assumptions of regression were met (residuals normally distributed with a mean of 0 and constant variance with independence between variables), and various outliers were identified based on the approach utilized (Figure 3 below). All the tests carried out in the below Figure 3 underscore the efficacy of the model created to predict female participants' like ratings. Additionally, a collinearity test was conducted to ensure any variables in the model are not dependent on one another, which would result in inaccurate predictions. The variance indication factor (VIF) and condition index were calculated and found to be in passing ranges (VIF < 10 and Condition Index < 30), indicating no collinearity in the model. Shrinkage was again tested

on the female model to determine the reliability of the model for future predictions. The calculated shrinkage value of -0.0659 indicates the female model is reliable for predicting female like scores as the absolute value of shrinkage (0.0659) < 0.1.

**Figure 3: Female Model Residual Analysis**

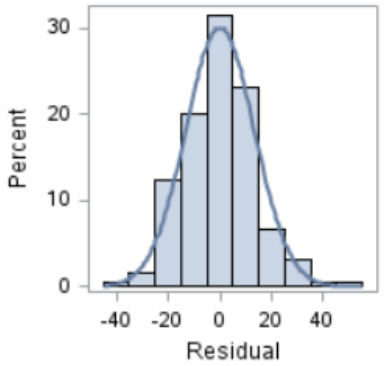
Normal Distribution Proof (Histogram, Box & Whisker Plot, & Q-Q Residual Plot)




Outlier Summary:

1. Jackknife = 13 (net -2 outliers)
2. Leverage = 10 (net -17 outliers)
3. Cooks distance = 0 outliers
4. Boxplot = 6 outliers

Proof of Normal Distribution




Residual distribution plot with normal curve overlaid indicates corrected skew and reduced kurtosis. Lower kurtosis value indicates fewer outliers and improved stability

Skew value = 0.3807024
Kurtosis = 1.45601307

Normal Distribution Tests

| Tests for Normality | | | | |
|---|---|---|---|---|
| Test | Statistic | | p Value | |
| Shapiro-Wilk | W | 0.984794 | Pr < W | 0.0343 |
| Kolmogorov-Smirnov | D | 0.041952 | Pr > D | >0.1500 |
| Cramer-von Mises | W-Sq | 0.074455 | Pr > W-Sq | 0.2462 |
| Anderson-Darling | A-Sq | 0.48107 | Pr > A-Sq | 0.2358 |

Kolmogorov-Smirnov, Cramer-von Mises, and Anderson-Darling tests p >0.05 indicating normal distribution

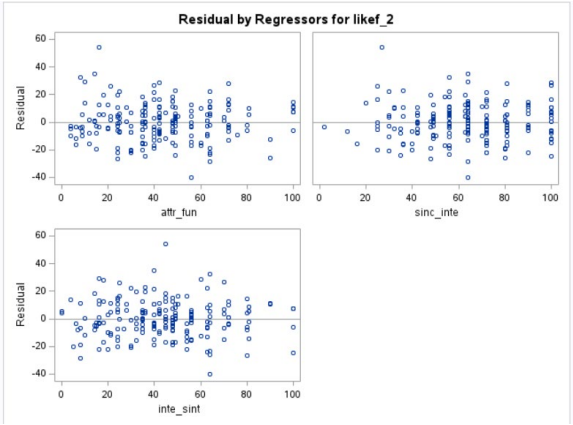
*(H0: Normal Distribution, p not <0.05 -> cannot reject H0)*

**Figure 3: Female Model Residual Analysis**

Residuals vs Predictions Test


Shows improved centering around y = 0, supports assumption mean of residuals = 0

Residuals vs Regressors Tests




Shows reduced funneling and improved homoscedasticity (assumption variance is constant, met)

**SUMMARY OF FINDINGS**

Two models were identified for use in predicting male and female participants' recorded like scores reliably and accurately. In comparison of the two prediction models, it was confirmed that male and female participants place value in differing personality characteristics when rating a potential partner with a "like" score. While both genders value fun, attractiveness, sincerity, & shared interests, each gender valued these attributes at different importance levels. Additionally, males valued attractiveness higher than females as the attractiveness rankings were included in two combinations in the model.

These models can be utilized to improve potential match creations for Pizzaz.com as well as allow for user customization in the app. This will allow users to filter for matches based on their predicted like values to curate their experience. Additionally, in order to improve future prediction models, it is recommended that Pizzaz ensure no blank values are permitted on user ratings or an imputation process is identified to fill in the data gaps.