Investigating Romantic Relationship Quality and the Roles Race, Education, and Income Play

Sanjana Gupta, Natasha Halfin, Andreia Sodre Nichols, Karl Stavem, and Maggie Weatherly

Abstract

Since the beginning of recorded time, romantic relationships have played a critical role in society. Works of art, literature, and music are devoted to this theme; romance has even instigated wars (recall Helen of Troy served as the catalyst for the Trojan War). However, what factors actually contribute to the success or quality of romantic relationships? We sought to answer this question by investigating the survey "How Couples Meet and Stay Together" (HCMST), which contains a variety of demographic data points corresponding to surveyed individuals and their romantic partners. We focused specifically on the relationships race, education, and income individually have with relationship quality (as measured by the survey). To assess these relationships, we applied a variety of statistical methods including linear regression, ANOVA, and t-tests. Our analysis ultimately revealed the following high-level conclusions: 1) couples of the same race appear to report greater relationship quality than mixed-race couples; 2) education disparity alone does not seem to harm relationship quality; and 3) despite research that shows a marginal utility for income as it pertains to happiness, increased income does correlate to greater relationship quality. We provide further insight into our methods, findings, key assumptions, limitations, and additional opportunities in the body of this analysis.

Introduction

Romantic relationships are ubiquitous in human life. In Western culture, monogamous relationships and marriage are seen as the norm. However, a commonly shared statistic is that 50% of marriages end in divorce. Hence, what makes relationships successful? Which features are highly correlated with high quality relationships? We decided to investigate the following themes related to this topic:

- How does race affect relationship quality? Do mixed-race couples report higher relationship quality than same-race couples or vice versa?
- How does disparity in education level (or lack thereof) between partners affect relationship quality?
- How does combined income, as well as income disparity between partners, affect relationship quality?

Dataset Description

We leveraged the public use portion of the dataset from the study "How Couples Meet and Stay Together" (HCMST). The dataset was collected from 2009 to 2015 by surveying 4,002 American respondents over five waves. Response to Wave 1 (main survey), in 2009, was 71 percent, and response

rate fluctuated over the course of the remaining waves (i.e. not everyone responded during each wave). In total, the dataset has 534 features that include information such as marital status, sexuality, education level, income, and geographic location. Each record represents a survey response from one individual on behalf of themselves and their romantic partner (if they are coupled). In each wave, relationship status and marital status were recorded. Both heterosexual and LGBTQ individuals were surveyed. Additionally, coupled, uncoupled, married, and unmarried individuals responded to the survey.

Assumptions

Survey respondents were recruited by random digit dial phone survey; therefore, the survey population was randomized. However, we cannot say with confidence that the respondents are truly representative of the broader US population. In fact, the study oversampled LGBTQ populations to attempt to balance the study population. Furthermore, we assume that survey responses are truthful, although the data is self-reported and may not be entirely accurate.

Data Cleaning

To narrow the scope of our analysis, we focused only on the Wave 1 portion of our dataset (the original survey). Additionally, we dropped any rows associated with uncoupled individuals, which reduced our population by about 25%. We also dropped any columns that were not relevant to our research questions. Furthermore, the outcome variable of our analysis, relationship quality, was originally recorded as a categorical variable, with five options ranging from Poor to Excellent. We decided to create a new field that mapped the relationship quality categories to numbers (i.e. Poor =1, Excellent = 5). We retained the following columns for our final project dataset, which include some additional calculated fields:

- **RESPONDENT_RACE**: categorical field indicating respondent's race
- **PARTNER_RACE:** categorical field indicating partner's race
- **RESPONDENT_YEARSED:** numeric field indicating approximate number of years of education respondent has
- **PARTNER_YEARSED:** numeric field indicating approximate number of years of education partner has
- **Q23:** categorical field indicating who between the couple earned more
- **PPINCIMP**: categorical field; total combined household income by bin ranges
- **Diff_in_race:** Boolean field; 0 if RESPONDENT_RACE and PARTNER_RACE is equal, 1 otherwise (calculated field)
- **Diff_in_education:** absolute value of RESPONDENT_YEARSED - PARTNER_YEARSED (calculated field)
- **Party_with_most_education:** categorical field; 'partner' if PARTNER_YEARSED > RESPONDENT_YEARSED, 'respondent' if RESPONDENT_YEARSED > PARTNER_YEARSED, 'same' if RESPONDENT_YEARSED = PARTNER_YEARSED (calculated field)

- **Diff_in_income:** Boolean field; 0 if Q23 said that partners earned about the same amount, 1 otherwise (calculated field)
- **RELATIONSHIP_QUALITY:** categorical field that rates relationship quality on a scale of 1-5; 1 = Poor; 5 = Excellent (calculated field)
- **PPWORK:** categorical field indicating employment status
- **QFLAG:** Boolean field indicating if the respondent has a partner
- **PAPGLB_STATUS:** categorial field indicating if respondent identifies as gay, lesbian, or bisexual
- **PPMARIT:** categorical field indicating if respondent is married, widowed, divorced, separated, never married, or living with a partner
- **PPGENDER:** categorical field indicating if respondent is male or female
- **Q4:** categorial field indicating gender of partner
- **Q5:** Boolean field indicating if partner is same gender as respondent

Exploratory Data Analysis

Before we attempted to answer our specific research questions, we wanted to understand the underlying distributions of our project dataset. Below are our findings across multiple demographic dimensions.

Relationship Quality: the largest proportion of respondents were those who rated their relationship quality as excellent. This could potentially skew our findings.
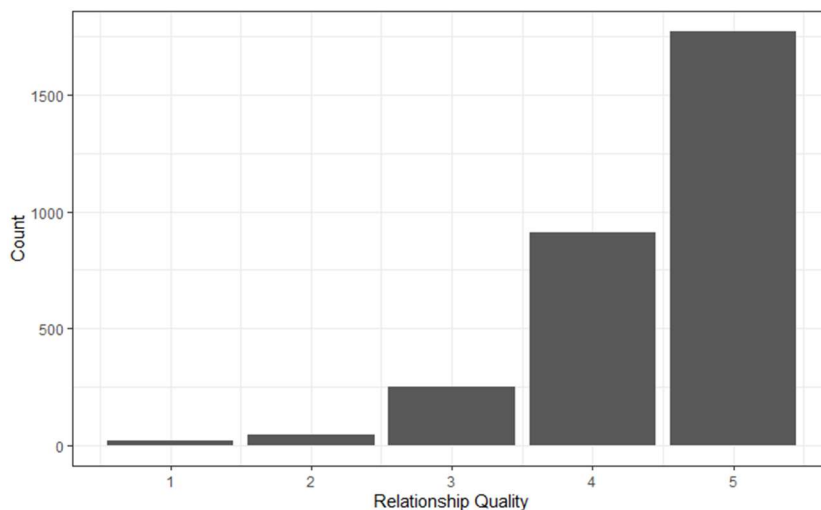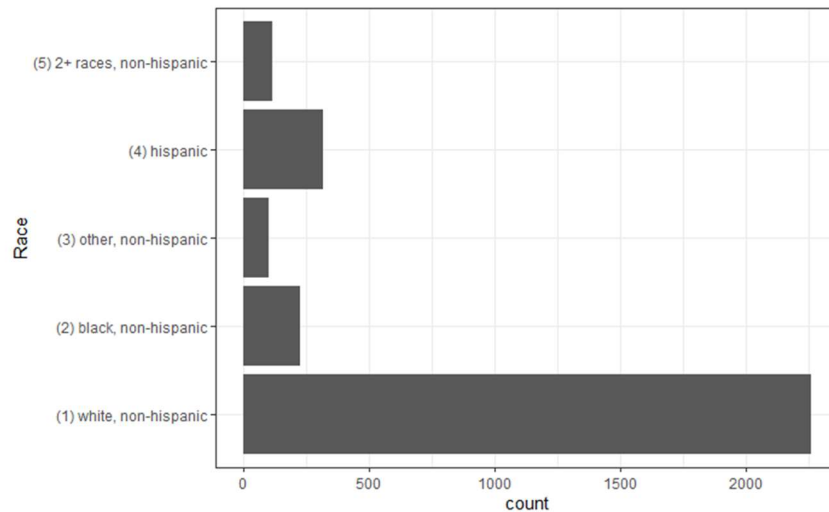
LGBT Status: a small part of the population (less than 1/5th) identified as LGBTQ and only 460 of them were part of a same sex couple.
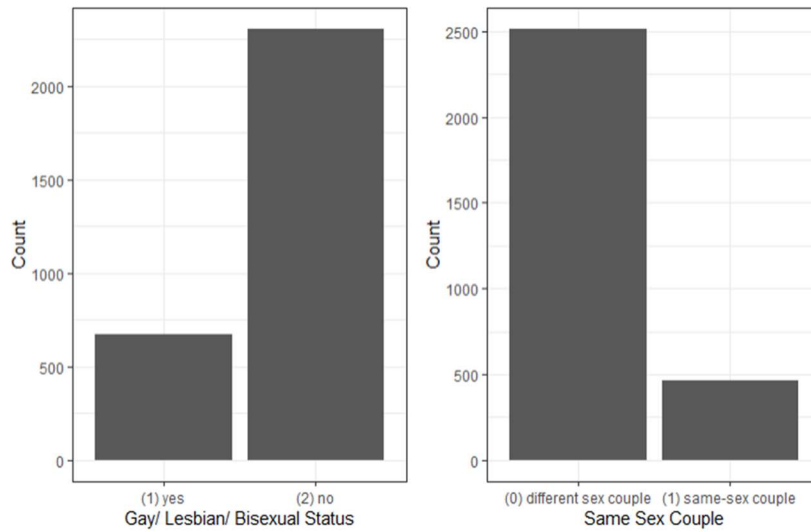

Ethnicity / Race: we found a significantly higher proportion of white subjects compared to other sampled races - over 2250 to be exact, which shows that they are likely over-represented.

**Education Level:** respondents with a bachelor's degree or higher have the greatest representation in the survey population; over 1000 respondents reported having a bachelor's degree or higher.
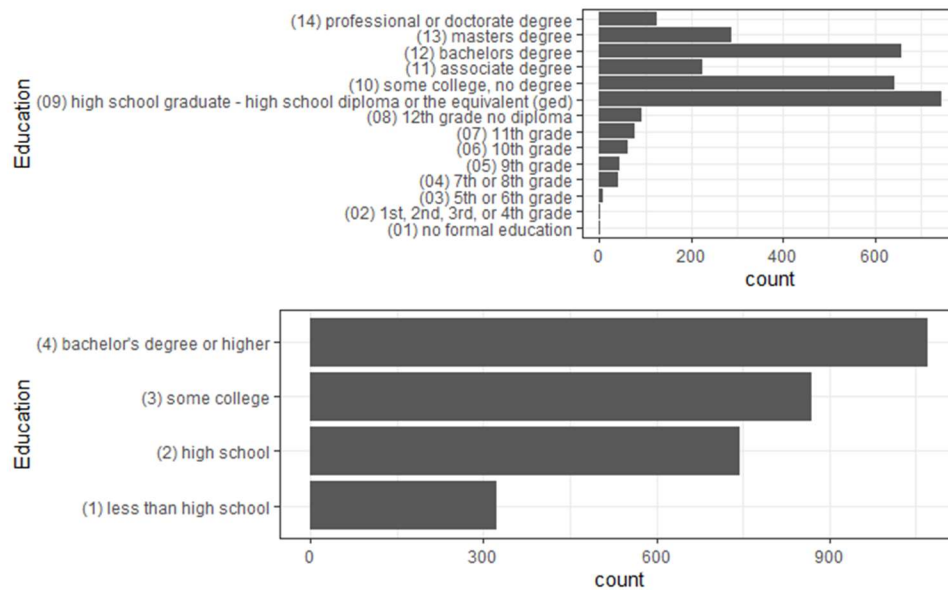
Income: Income was normally distributed around the $60-75K bin with mean household income around $67,000. Additionally, women reported earning less than their partners more often than males did.

Earnings among Partners in the US
Women were more likely to report making less than their partner
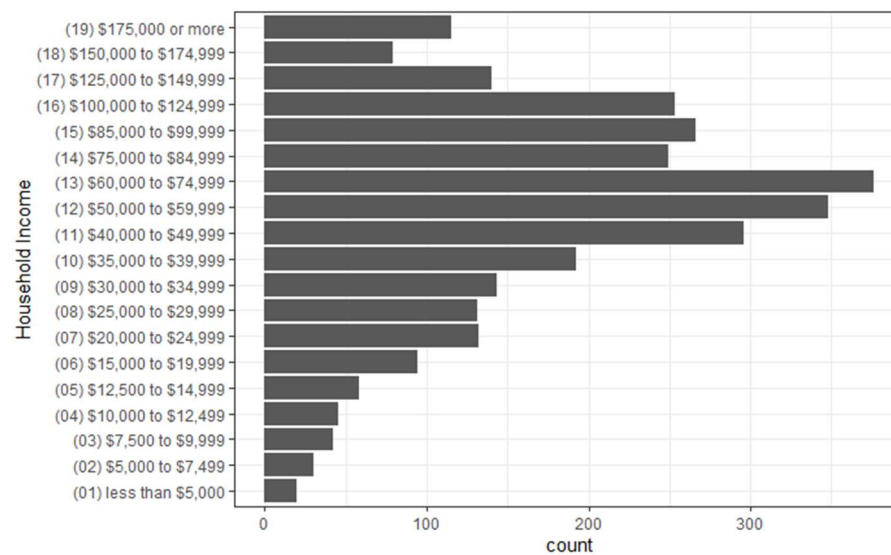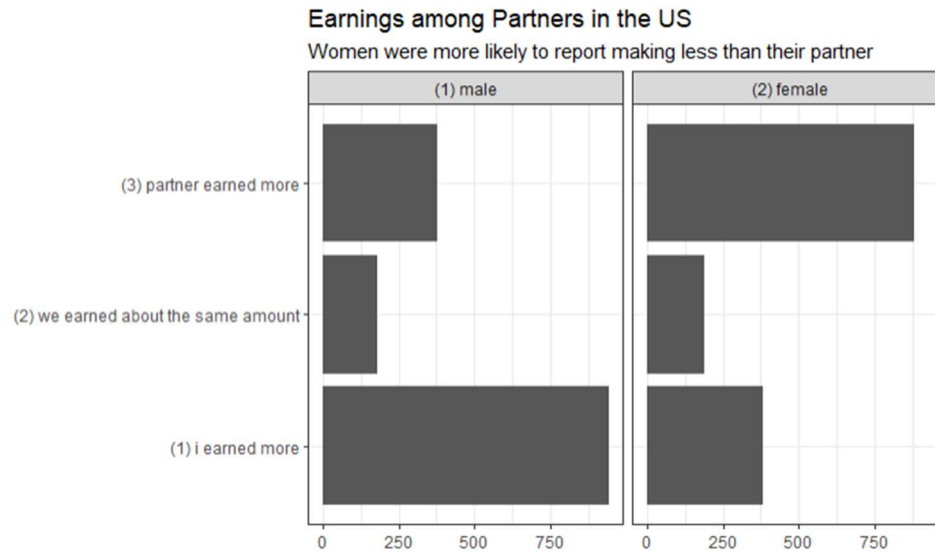
Results

We have divided our results into three sections corresponding to each of the demographic characteristics we examined: race, education, and income. Within each of these sections, we discuss more specific research questions and corresponding models, in which we provide an overview of the statistical methods utilized, how our data meets underlying assumptions, and finally, our results.

To answer our research questions, we generally applied a combination of analysis methods: t-tests, linear regression, and ANOVA. Our primary outcome variable and metric of relationship success was relationship quality. Below are the key assumptions we used for our analysis methods:

- **T-test:** for equal variances use the equal variance t-test; for unequal variances, use the Welch t-test; independence; and large sample size, or normality for small sample sizes
- **Linear regression:** normality or large sample size, constant variance, and linearity
- **ANOVA:** independence of samples, large sample size, and equal variance

Race

We decided to explore race to verify if it has any effect on mean relationship quality score reported by couples. The survey respondents were asked to identify with one of six race categories (White, Hispanic, Black, Asian Pacific Islander, American Indian, or Other). The respondents also provided information about their partners' race. Rows without an answer for either the respondent or the partner were removed.

Race Model One: T-test

Statistical Methods Model One:
The first question we wanted to answer was "*Do same-race couple report higher relationship quality than mixed-race (interracial) couples?*" We separated our dataset into two groups: couples where both partners are the same race, and couples where partners are different races. We utilized a t-test and our hypothesis testing is framed as follows:

Null Hypothesis: Same-race couples report equal or lower mean relationship quality than mixed-race (interracial) couples.

Alternative Hypothesis: Same-race couples report higher mean relationship quality than mixed-race (interracial) couples.

We met all t-test assumptions, but erred on the side of caution for variance and applied an unequal variance (Welch) t-test. Our population is summarized below:

Group	sample size	variance	mean
same-race	2476	0.56	4.48
mixed-race	509	0.67	4.35

When we calculated the power of our test, we realized it was below our desired power. We decided to oversample our mixed-race group with replacement in order to achieve power above 80%. Our adjusted population is displayed below:

Group	sample size	variance	mean
same-race	2476	0.56	4.48
mixed-race	**2476**	**0.62**	**4.37**

Results Model One:
The Welch t-test resulted in a p-value of nearly 0. Therefore, we have sufficient evidence to reject the null hypothesis that same-race couples have equal or lower mean relationship score than mixed-race couples.

Group 1	Group 2	p-value
same-race	mixed-race	1.317221e-07

Race Model Two: ANOVA

Statistical Methods Model Two:
Since we found out that same-race couples may report a higher mean relationship score than mixed-race couples, we wanted to answer the following: *"Are all same-race couples equally happy?"* We thus separated our dataset into six groups: one group representing each same-race couple. We decided to utilize an ANOVA test and our hypothesis testing is framed as follows:

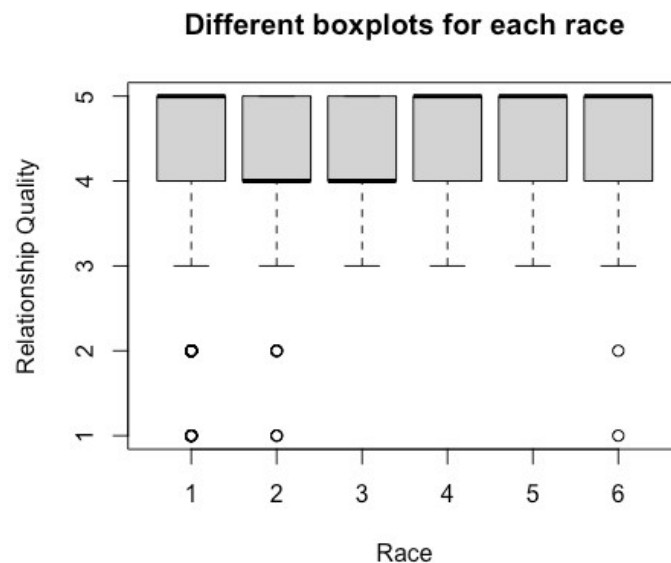**Null Hypothesis**: all types of same-race couples report the same mean relationship quality.

Alternative Hypothesis: at least one type of same-race couple does not report the same mean relationship quality.

We checked for the ANOVA assumptions of independence, large sample size or normal distribution, and equal variance. We already know that our samples meet the independence assumption. Our sample sizes for American Indian, Asian Pacific Islander, and Other are too small to hold the assumption of large sample size and to give us enough power. Therefore, after calculating the necessary size to achieve 80% power, we oversampled those populations to 56 samples.

Group	sample size	variance	mean	oversample sample size	oversample variance	oversample mean
(1) NH white	2150	0.53	4.51	2150	0.53	4.51
(2) NH black	176	0.80	4.21	176	0.80	4.21
(3) NH Amer Indian	5	0.70	4.20	**56**	**0.49**	**4.21**
(4) NH Asian Pac Islander	24	0.34	4.58	**56**	**0.38**	**4.62**
(5) NH Other	4	1.00	4.50	**56**	**0.76**	**4.50**
(6) Hispanic	117	0.62	4.42	117	0.62	4.42

Next, we tested our equal variance assumption. We interpreted the boxplot below as indicating no evidence against equality of variance.

Different boxplots for each race

Results Model Two:
We performed the ANOVA which resulted in an extremely small p-value. Therefore, we have sufficient evidence to reject the null hypothesis that all same-race couples have the same mean relationship quality score.

Predictor Categorical Variables	P-Value
White, black, Hispanic, Asian Pacific Islander, American Indian, Other	9.5e-07 ***

Race Model - Three: T-test

Statistical Methods Model Three:
In the previous question, we discovered that not all same-race couples report the same mean relationship quality score. Our follow-up question attempted to answer *"Which same-race couple reports higher relationship quality?"* We separated our dataset into six groups: one group representing each same-race couple. Our hypothesis testing is framed as follows:

Null Hypothesis: Same-race couples of race X report equal and lower mean relationship qualities than all other same-race couples combined.

Alternative Hypothesis: Same-race couples of race X report higher mean relationship quality than all other same-race couples combined.

In this model, we tested if one of the same-race couples has a statistically significantly higher mean compared to the overall mean of all of the other same-race couples. For example, we are comparing the mean relationship quality of same-race white couples with the mean same-race non-white couples.

We performed many Welch t-tests to compare the means of all same-race couples with its counterpart same-race couples. We decided to remove Asian Pacific Islander, American Indian, and Other races from Group 1 in this test due to their small sample sizes. We did not utilize oversampling here as it would require many more samples in order to achieve the required Power. However, these populations are still included in Group 2.

Results Model Three:
We only found sufficient evidence to reject the null hypothesis for the same-race white couple test, suggesting that same-race white couples report a higher mean relationship quality score than other same-race couples.

Group 1	Group 2	p-value
NH white	Non-white	5.25E-05
Hispanic	Non-Hispanic	0.817317
NH black	Non-black	0.9999844

Education

Education Model One: T-Test

Statistical Methods Model One:
Our first education model attempted to answer the question, "*Does a difference in education between partners in a relationship have an effect on relationship quality?*" We decided to use a t-test to answer this question. We divided our data into two groups: couples who have the same education level between partners, and couples who do not have the same education level between partners. The sample sizes of these two groups are 1022 and 1967, respectively, and therefore meet the t-test's large sample size assumption. We then calculated the variance of relationship quality of these two groups to determine if we should use an equal-variance t-test or an unequal-variance Welch t-test. Our variances were 0.539 for couples with the same education and 0.599 for couples with different education.

Although these values are close, we chose to err on the side of caution and to continue with a Welch t-test. We defined our Null and Alternative Hypotheses as follows:

Null Hypothesis: The mean relationship quality is the same for couples that have the same education level versus couples that have different education levels.

Alternative Hypothesis: The mean relationship quality is different for couples that have the same education levels versus couples that have different education levels.

Results Model One:
The Welch t-test resulted in a p-value of 0.3576, indicating there is insufficient evidence to reject the null hypothesis. Therefore, there is no evidence to suggest that education has an effect on relationship quality.

Group 1	Group 2	p-value
Different Education	Same Education	0.3576
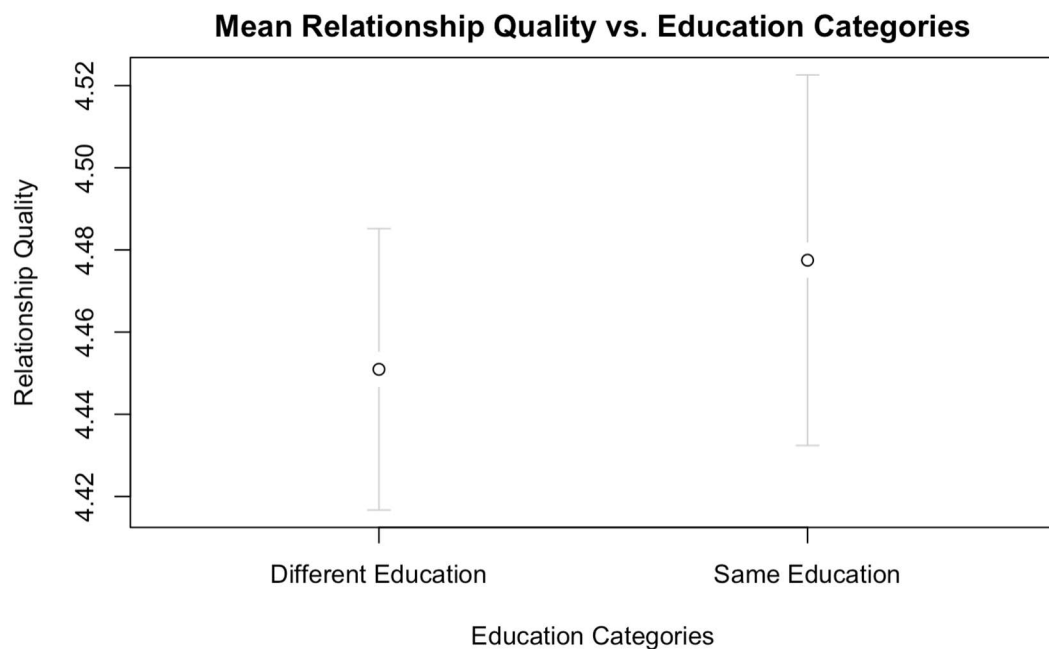
Education Model Two: Linear Regression

Statistical Methods Model Two:
Our second education model attempted to answer the question, "*Does education disparity in years between partners have an effect on relationship quality?*" To answer this, we used linear regression on a numeric calculated field, education disparity, which is calculated using two fields in our dataset, RESPONDENT_YRSED and PARTNER_YRSED, which represent the number of years of education for the Respondent and Partner, respectively. We calculated Education Disparity as the absolute value of RESPONDENT_YRSED - PARTNER_YRSED.

When considering the assumptions of linear regression models, we have a sufficiently large sample size of nearly 3000 samples; therefore, normality is not an issue. To assess linearity and constant variance, we plotted the residuals versus fitted values of the model. We address this in the section Results Model Two below. We defined our Null and Alternative Hypotheses as follows:

Null Hypothesis: The absolute value of the number of years of education disparity between partners in a relationship does not have an effect on mean relationship quality.

Alternative Hypothesis: The absolute value of the number of years of education disparity between partners in a relationship does have an effect on mean relationship quality.

Results Model Two:
Running a linear regression model with the calculated education disparity field as the predictor and relationship quality as the target variable yielded a coefficient of -0.0064 and p-value of 0.3734. The interpretation of the coefficient is that for every unit increase in education, there is a 0.0064 mean unit decrease in relationship quality. The p-value indicates that we do not have sufficient evidence to suggest that the absolute value of the number of years of education disparity has an effect on relationship quality.

Predictor Variable	Coefficient	P-Value
Difference in Education	-0.0064	0.3734

As previously noted, to address the linearity and constant variance assumption, we plotted the Residuals versus Fitted values of our linear regression model. We can see in the figure below that the data appears to hold the assumption of constant variance and linearity.


Residuals vs Fitted
lm(relationship_val ~ abs_diff_in_education)
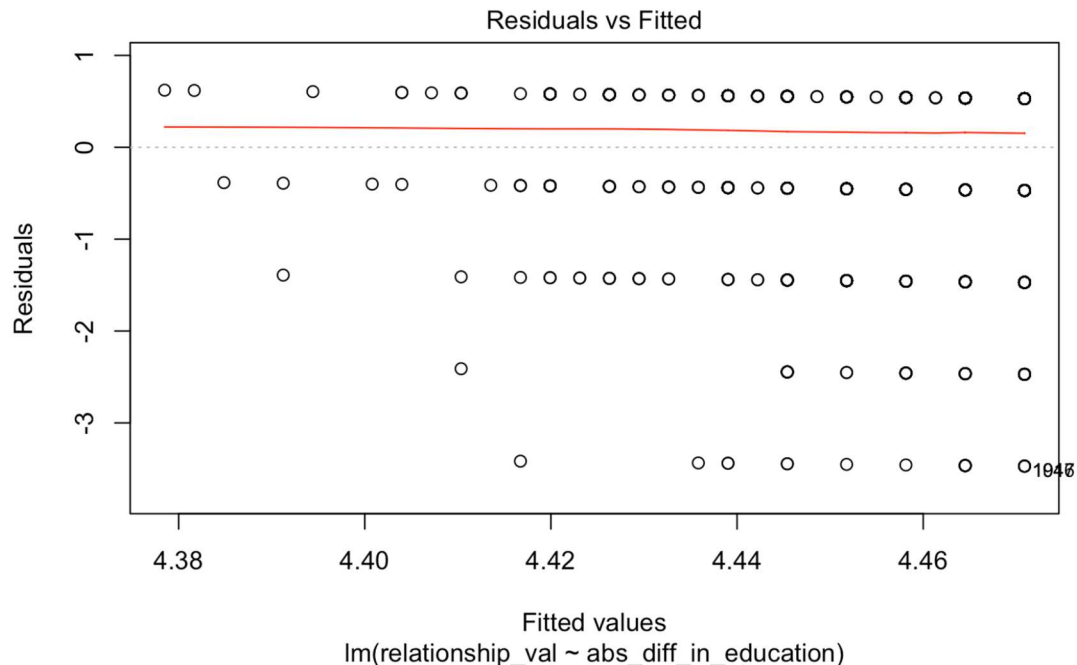
Education Model Three: ANOVA

Statistical Methods Model Three:
Our third education model looked into the question, *"In regard to relationship quality, does it matter who in the relationship has more education?"* We used ANOVA to investigate this question. For this model, our predictor variable was a calculated categorical variable indicating which partner in a relationship has more education. We created this by applying the following logic: if the difference in education variables is NEGATIVE, then the Partner has more education. However, If the difference in education variable is POSITIVE, then the Respondent has more education. And finally, if the difference in education variable is 0, then the Partner and Respondent have the same education.

With respect to model assumptions, the data meets the assumption of independence. The size of each of our three groups is 1022 for relationships with the same level of education, 1043 for relationships where the partner has more education, and 924 for relationships where the respondent has more education. These sample sizes meet the large sample size requirement for ANOVA. As for equal variance, we can use residual diagnostics in the form of a boxplot to check this. This plot, included below, appears to indicate equal variances.
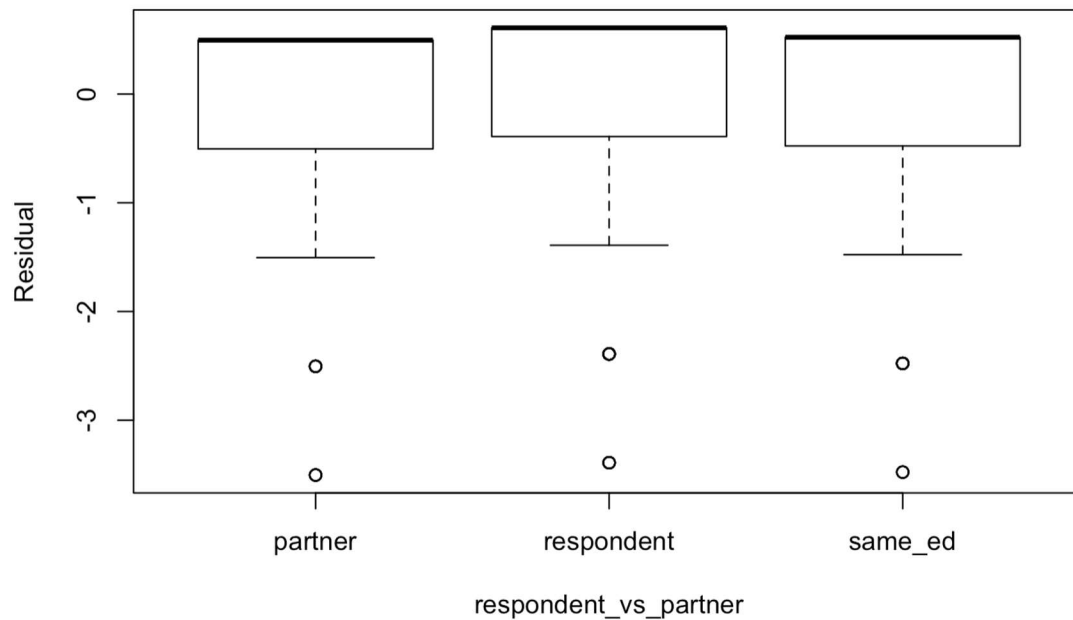
We defined our Null and Alternative Hypotheses as follows:

Null Hypothesis: The calculated categorical variable of which partner has more education (partner, respondent, or same) does not have an effect on mean relationship quality.

Alternative Hypothesis: The calculated categorical variable of which partner has more education (partner, respondent, or same) does have an effect on mean relationship quality.

Results Model Three:
Our ANOVA model produced a significant p-value of 0.0028. This result provides evidence to reject the Null Hypothesis, and therefore is evidence to suggest that which partner has higher education has an effect on relationship quality.

| Predictor Categorical Variables	P-Value
partner, respondent, same_ed	0.0028

Mean Relationship Quality vs. Education Categories
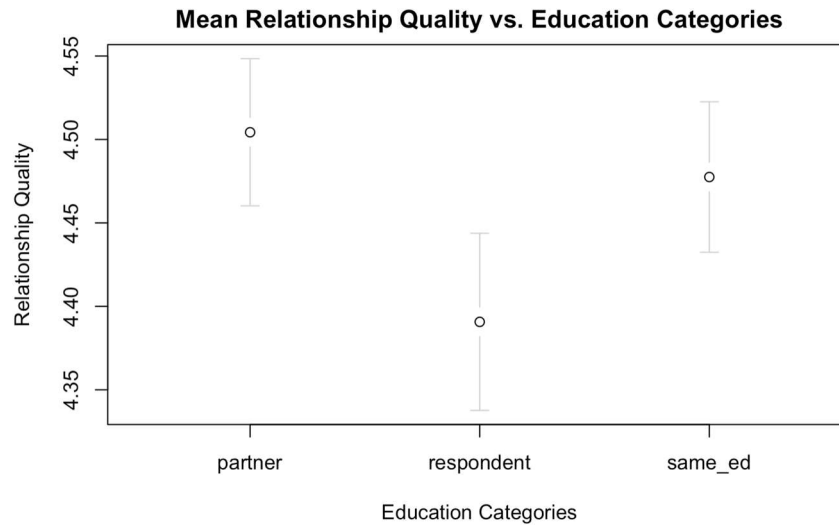
Income

Income Model One: ANOVA

Our first income model attempted to answer the question, *"Does total household income play a significant role in the relationship quality that couples report?"* The mean relationship quality for each of these income ranges appeared to suggest that increased income was associated with an increased relationship score.

Mean Relationship Quality vs. Income Level
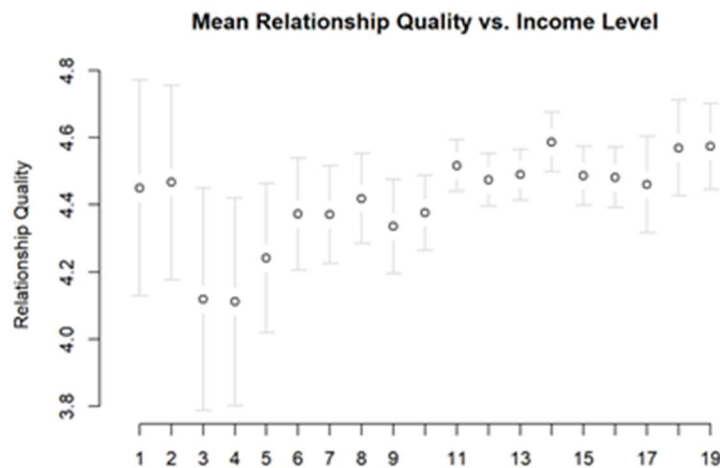
Statistical Methods Model One:
In order to test this relationship, we created an ANOVA model with total household income as the predictor variable and relationship quality as the outcome variable. The ANOVA model was a logical choice for this dataset given its approximate normality and nearly constant variance. Within our model,

income was treated as a factor with a separate mean relationship quality score fitted to each of the income groups. Our Null and Alternative Hypotheses are as follows:

Null Hypothesis: mean relationship quality is the same across all household income groups.

Alternative Hypothesis: mean relationship quality is not equal among all income groups.

Results Model One:
After running the model, it was not entirely surprising to see a very small p-value of 0.000271. This showed strong evidence against the null hypothesis and would seem to suggest that total household income is a significant factor in relationship quality.

| Predictor Variable	F-Statistic	P-Value
Total Household Income	2.583	0.000271

Income Model Two: T-test

With our second income model we attempted to answer the question, *"Do couples with income above a certain threshold report higher relationship quality than those that fall below?"* The rationale behind this question came from a Huffington post article which cited a study that discovered a marginal utility, or cap, at which increased income no longer corresponds to increased happiness[1]. The "happiness threshold" cited in the article was roughly $75,000 per year. We set our threshold to this value to examine the article's theory.

Statistical Methods Model Two:
We divided our respondents into two disjoint subgroups: couples that earned $75,000 and above per year and those that earned below $75,000. After isolating our respondents, we saw approximately 1,000 couples in the $75K and above group and approximately 1,900 in the below $75K group.

---

[1] https://www.huffpost.com/entry/map-happiness-benchmark_n_5592194
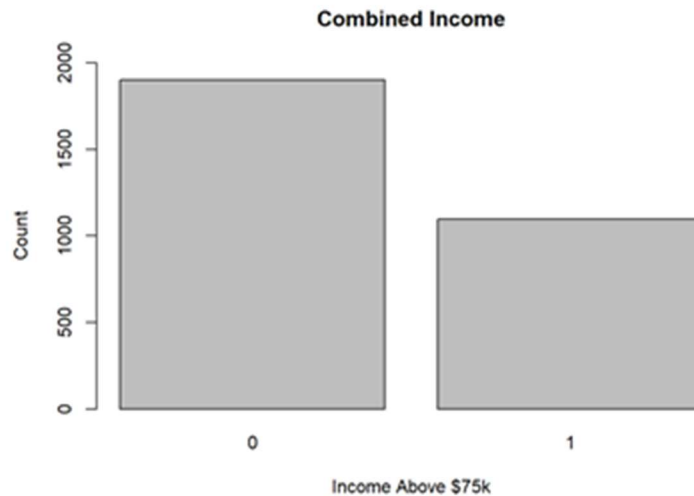
When we analyzed the data, we noted that the sample mean for those above $75,000 was slightly higher than those below.

$75K and above - Mean Relationship Quality	Below $75K - Mean Relationship Quality
4.52	4.42

Because of the large sample size in each group, we examined the relationship between these two means using a one-sided Welch t-test. Our Null and Alternative Hypotheses are as follows:

Null Hypothesis: mean relationship quality is the same in the below $75K group as the $75K and above group.

Alternative Hypothesis: mean relationship quality is greater for those in the higher income group (we are assuming the $75K and above group have reached the maximum possible happiness and relationship quality).

Results Model Two:
After running the model, we noted a small p-value of 0.0003428 demonstrating sufficient evidence to reject the null hypothesis. Evidence seems to suggest that mean relationship quality is greater for those earning $75,000 and above.

Predictor Variable	T-Statistic	P-Value
Total Household Income	3.3997	0.0003428

Income Model Three: Logistic Regression

In our third model, we studied whether income disparity affects the quality of the relationship.

Statistical Methods Model Three:
Within the survey, respondents were asked to answer the following question: *"Between you and [partner_name], who earned more income in 2008?"* Respondents answered with one of three options: 1) I earned more; 2) We earned about the same; or 3) Partner earned more. It is important to note that answers were based on feelings about income, rather than specific numbers. Using these subjective responses as the basis for this model, we divided the sample into two disjoint groups:

- No Income Disparity: *We earned about the same amount.*
- Income Disparity: *I earned more OR Partner earned more.*

Since all respondents fell naturally into one of these distinct groups, we created a simple binary variable to store this information (1 = Income Disparity, 0 = No Income Disparity). This binary parameter led to using a logistic regression model; however, it required that we format our research question in a particular way: *"Does relationship quality tell us anything about the odds of having an income disparity between partners?"*

Additionally, logistic regression models hold three convenient properties which were necessary given the form of this dataset:

- Logistic regression does not require a linear relationship
- The residuals of a logistic model do not need to be normally distributed
- Homoscedasticity is not required

Given our data adheres to these properties, our Null and Alternative Hypotheses are as follows:

Null Hypothesis: there is no relationship between relationship quality and income disparity.

Alternative Hypothesis: a relationship exists between relationship quality and income disparity.

Results Model Three:
After running the model, we arrived at the following results:

Predictor Variable	Coefficient	Exponentiated Coefficient	P-Value
Relationship Quality	-0.1269	0.881	0.09859

 We calculated an exponentiated coefficient of 0.881, which suggests that for each unit increase in relationship quality, the odds of having an income disparity in the relationship decrease by roughly 12%. However, the p-value in this model was quite high and suggested that this likely is not a highly significant result.

Conclusions and Additional Opportunities

In summary, we discovered that while same-race couples do appear to report higher relationship quality than mixed-race couples, the advantage appears to primarily benefit same-race white couples. Additionally, while education disparity alone may not affect relationship quality, which specific partner has more education may play a role in reported quality. Finally, greater combined income appears to have a positive correlation with relationship quality, even past the $75,000 "happiness threshold". Our results are not entirely surprising; race, education, and income serve as some of the most distinguishing factors of socio-economic class in our world today. While society has evolved from the strictly-defined social classes of centuries ago, we might infer from our analysis that people tend to be happier with partners most similar to them in socio-economic background. Clearly, our demographic characteristics may play a bigger role in the success of romantic relationships than we realize.

However, we must recognize the limitations within our dataset. First, the representation of race within our population was heavily skewed towards white respondents, which may have affected our results. Additionally, we found in our exploratory data analysis that the distribution of relationship quality ranking was heavier towards the 5 ranking, or Excellent score. As we discussed in our assumptions, this is a self-reported metric, and ultimately a subjective one, which could skew results. Furthermore,

because we focused our analysis on the Wave 1 data, our results are based on a point-in-time capture and do not reflect the potential for relationship quality to change over time.

There are several additional analysis opportunities within our dataset. We could expand our research questions to understand how race, education, and income affect relationship quality over time and leverage the longitudinal data points in Waves 2-5. We could also examine interaction between race, education, and income as they pertain to relationship quality, since our analysis focused on each variable individually. Additionally, we could apply survival analysis methods in order to estimate how long relationships may last given race, education, and income characteristics. In summary, the HCMST dataset is incredibly rich, and allows for a myriad of analysis opportunities with respect to the role demographics play in romantic relationships.

References

1. Rosenfeld, Michael J., Thomas, Reuben J., and Falcon, Maja. How Couples Meet and Stay Together (HCMST), Wave 1 2009, Wave 2 2010, Wave 3 2011, Wave 4 2013, Wave 5 2015, United States. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2016-03-18. https://doi.org/10.3886/ICPSR30103.v8
2. https://www.bgsu.edu/arts-and-sciences/center-for-family-demographic-research/help-resources-tools/data/data-holdings/family-and-fertility/how-couples-meet-and-stay-together
3. https://www.huffpost.com/entry/map-happiness-benchmark_n_5592194

Appendices

Code

The R and RMarkdown code used for the statistical analysis is available on our project GitHub. There are also additional analyses that we did not have a chance to cover in this report: https://github.com/sanjanagupta16/Data557_Statistical_Modeling_Project