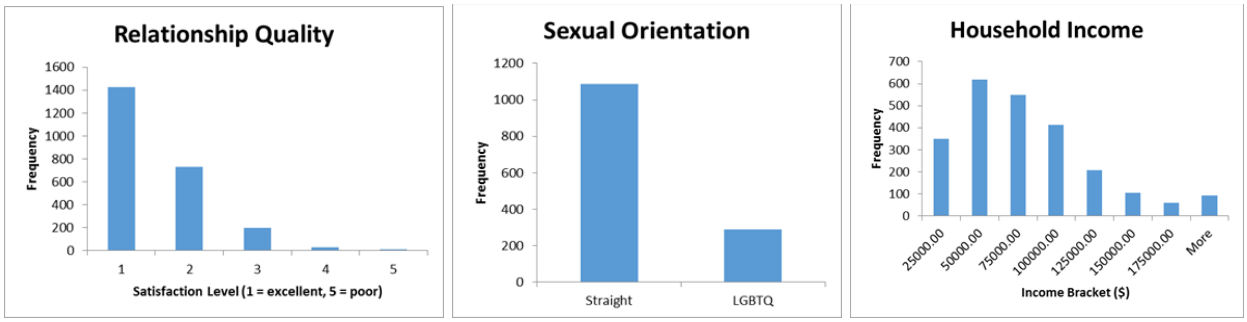
Midterm Report

Catherine Weldon, Angeline Pinilla, Josh Robbins

The Data

Our data set seeks to find factors that determine the success of a relationship/marriage. The data set has 5 waves (one survey/year for five years). The first survey has 35 questions -- many questions having multiple parts -- and focuses on demographic information such as income, race, age, sexual orientation, and various other questions we hope can provide insight into how successful a relationship will be. The metric we are using to measure the success of the relationship is the quality of the relationship where participants rate their relationship on a scale of 1 - 5. The other waves provide follow-up data for most of the participants, detailing if they broke up, and if so, why. We plan to use these follow-up waves to further analyze factors that contribute to break-ups.



**The above histograms refer to the train data set and includes features recorded from the first wave.*

For the first wave, out of the original 4,002 participants, only 2,996 had data for quality of relationship. For the second wave, only 2007 of those participants were able to be contacted for the follow-up survey, the third wave had 1,553 respondents, the fourth wave had 1,216 respondents, and the fifth and final wave only had 891 respondents. We can tell this data is missing because each wave has a field for how many people completed the survey. The reasons for this missing data is either that they either no longer wanted to participate in the study, or they were ineligible for follow-up waves because they weren't in their original relationship. There were also text responses in this survey that are missing from the data set so all the analysis will revolve around quantitative data. There doesn't appear to be any corrupted data as all the entries are in the range of possible responses as indicated on the survey. It is not clear if any of these values were every incorrectly inputted or modified. For these reasons, we will assume that the data has not been corrupted.

Preliminary Analysis

The feature used as our variable to be predicted is Q34 which rates the quality of the relationship on a scale of 1 - 5. There are a number of features that are being considered in making the model that predicts this variable. The data for the 2,996 participants who responded to Q34 was separated into a train and a test data set where the train set contained 80% of the data and the test set contained the other 20%.

Because our sample size is relatively small (4,002 participants) over-fitting is our main concern. To avoid over-fitting, we will not choose a model that is too flexible and we will look at the variance and bias to determine an appropriate tradeoff. If it still seems like over-fitting could be an issue, we will bootstrap to artificially create more data to analyze our model.

In our preliminary analyses, when using a classification model to predict the ordinal variable 'Q34', we will test the effectiveness of the learned model by evaluating an unbiased estimate of the accuracy on our unseen data (i.e. the test set we set aside). We can calculate the test error rate: the average number of incorrect predictions on the test set when compared to the true class labels of the test set. Additionally, we can draw a comparison between the train error rate and the test error rate to identify the model's underfitting/overfitting behavior.

In our preliminary models (see python script within our 'data-processing' subdirectory for implementation details), we aimed to classify our data on the response variable 'Q34': a partner's perceived measure of quality of his/her relationship, on a scale 1-5. Thus we used this scale as the class labels for our data. To perform this classification, we first used the following from scikit-learn's modules: a 'Multiclass learning model using One-vs-Rest' (train error-rate = 0.55, test error-rate = 0.56), and then implemented a 'RandomForestClassifier' (train error-rate = 0.01, test error-rate = 0.12) based on the previous model's low performance accuracy. In our first run of these models we used a total of 158 features after preprocessing, which required dropping feature columns which represented one level of a particular categorical variable (e.g. RaceWhite, RaceBlack, etc.).

Next Steps

At the next phase of our project, we plan to implement a more robust method of feature selection and/or transformations to fine tune our features (e.g. recode categorical variables, such as Race, that are represented in the data set by multiple columns) and in turn improve model accuracy and efficiency. In addition, with future models, we will implement a K-fold cross-validation to compute a performance measure such as accuracy from the resulting model's performance on the remaining part of the data not used in the k-1 folds marked for training.

After we feel we have been able to create an effective model to predict the quality of a relationship, we will potentially consider future waves to create a model that predicts whether the couple will break up within the next five years and what the reason for the break-up might be. As there is far less data for the follow-up waves (particularly the fifth), over-fitting will be even

more of an issue and the resulting model will have to greatly take that into consideration. To complete these goals by November 25th, we will aim to our model to predict relationship quality done by November 14th. If we feel this model is not sufficient to predict relationship success, we will then create our follow-up model that predicts break-ups and aim to complete this by November 20th. The next five days will be spent doing some final error analysis and writing our final report.