M9 - Capstone - Voting Party Predictor and Influencer Identification

Tim Dimacchia

2020-09-21

Contents

1 Preface - About the Author


```
  Author: Tim DiMacchia
    Date: September 20, 2020
   Email: tdimacch@mac.com
[LinkedIn:](www.linkedin.com/in/timdimacchia)
[More About The Author - click for URL](https://docs.google.com/presentation/d/1q9buIIOHkAzYgvzcK
```


2 Project Background

The following project is required as part of the HarvardX (edX) PH125.9 Data Science: Capstone course. It is required to complete the 9 module series to receive a Professional Certificate in DataScience.

As part of this project, we were strongly discouraged from using well-known datasets, particularly ones that have been used as examples in previous courses or are similar to them (such as the iris, titanic, mnist, or movielens datasets, among others). Providing an opoportunity to learn and use new datasets.

Recommended data source locations were the UCI Machine Learning Repository and Kaggle. The required dataset must be automatically downloaded in your code or included with your submission.

The foundation for this project came from the Kaggle competion - "Can we predict voting outcomes?" This competition may be found at: Kaggle Competition - Predicting Election Results. Project competition was only open to students of 15.07x - The Analytics Edge

Data for this project was provided by Show of Hands, an informal voting polling platform for use on mobile devices and the web. Show of Hands specilizes in seeing what aspects and characteristics of people's lives predict how they will be voting for presidential elections.

Show of Hands has been downloaded over 300,000 times across Apple and Android app stores, and users have cast more than 75 million votes.

Other project ideas that were considered were:

- Anayizing Rider Share data for a particular city (San Francisco, Reno, Las Vegas, etc.)
- Job Offer Salary Prediction for a canidate
- City Housing Prices (California, Reno, etc.)
- Olympic Race Walker Result prediction
- Predicting Fake News
- Predicting Car Prices

I have choosen this project, because we are approximately 50 days from our next presidential election and building a predictor was considered a 'fun' exercise to compare against the actual results.

The following appendixes have been created:

- Appendix-A Package Installations
- Appendix-B Dataset Inspections
- Appendix-C Demographic Charts
- Appendix-D Correlation Matrixes and Heat Maps
- Appendix-E References
- Appendix-F Peer Assignment Grading Requirements
- Appendix-G Survey Questions
- Appendix-H List of Tables
- Appendix-J List of Figures

3 Project Goal

The goal of this project is to apply machine learning techniques that go beyond standard linear regression modeling using a publicly available dataset of our choice. As mentioned in the Project Background section, we will be using the election survey results provided by Show of Hands which consists of thousands of users and one hundred different questions to see which responses predict voting outcomes.

We have created two goals for this project which are modifications of the Kaggle compeition goal:

- Goal #1: Predict which candidate platform will win the election
- Goal #2: Identify which questions have the largest influence on Goal #1.

Special Note: This class is considered to be an **introductory class** in both Machine Language and the programming language R. Conversly the Kaggle competition was for an **advanced class**.

4 Project Grading

Appendix - F (Peer Assignment Grading Requirements) contains the requirements and point allocation for each requirements of this project.

Project submission must include:

- A report in the form of a PDF document
- The resulting Rmd file
- The R source code / script that performs the machine learning task.

Addiitionally, access to the datasete must also be made either through automatic download or inclusion in a GitHub repository.

We will be providing our datasets via project submission attachment in the form of the .zip file. * Note: The user must expand this .zip file into their current working directory.*

5 Dataset Acquisition - The Original Datasets

For this project, 3 files were provided as part of the Kaggle competition. Those files are:

- Train.csv : Contains the dataset to be used for training the Machine Learning Model
- Test.csv : Contains the dataset to test the Machine Learning Model
- Questions.csv : The questions which were given to the participants.

Before proceding we will inspect the datasets to ensure all data has been properly wrangled.

Data Fields for the Train Dataset:

- USER_ID : an anonymous id unique to a given user
- YOB : the year of birth of the user
- Gender : the gender of the user, either Male or Female
- Income : the household income of the user. Either not provided, or one of:

```
- "under $25,000"
- "$25,001 - $50,000"
- "$50,000 - $74,999"
- "$75,000 - $100,000"
- "$100,001 - $150,000"
- or "over $150,000"
```

- HouseholdStatus : the household status of the user. Either not provided, or one of:

```
- "Domestic Partners (no kids)"
- "Domestic Partners (w/kids)"
- "Married (no kids)"
- "Married (w/kids)"
- "Single (no kids)"
- or "Single (w/kids)"
```

- EducationalLevel : the education level of the user. Either not provided, or one of:

```
- "Current K-12"
- "High School Diploma"
- "Current Undergraduate"
- "Associate's Degree"
- "Bachelor's Degree"
- "Master's Degree"
- or "Doctoral Degree".
```

- Party : the political party for whom the user intends to vote for. Either "Democrat" or "Republican
- Q124742, Q124122, . . . , Q96024 101 different questions that the users were asked on Show of Hands. If the user didn't answer the question, there is a blank. For information about the question text and possible answers, see the file Questions.pdf.

Data fields for the Test Dataset are the same as the Train Dataset excluding the Party field.

5.1 Data Wrangling & Data Subset Creation

After reading the original files in, they were separated into the following:

- Kaggle_train : Original data read in from Train.csv
- Kaggle_test : Original data read in from Test.csv
- Kaggle_questions : Original data read in from Questions.csv
- train_subset : Removing the questions from the Train dataset
- test_subset : Removing the questions from the Test dataset
- train_subset_questions : Adding *key* questions back into the Train dataset
- test_subset_questions : Adding *key* question back into the Test dataset
- train_sample : Machine Learning ratio split of dataset for training
- test_sample : Machine Learning ratio split of the dataset for testing
- train_sample_questions : Train sample wih *key* questions added back in
- test_sample_questions : Test sample with *key* questions added back in
- train_sample_no_na : Removing NA's from the Train Sample
- train_sample_questions_no_na : Removing NA's from the Train Sample with *key* questions
- test_sample_no_na : Removing NA's from the Test Sample

Worth noting, even though we checked for "NA" upon reading the training dataset, test dataset and questions, we can see from Appendix B - Data Inspection summaries that some fields contain NA's within their vector. Hence we wrangled the datasets to remove NA's and created new datasets.

Later in this project we will be using the survey questions to tune our model predictions.
While subjective, we have reduced the 100 questions down to a few. Later we will reduce this list even further based upon their correlation influence.

For now the *key* questions which were identified were:

- 100010,Do you watch some amount of TV most days?,"Yes,No"
- 100562,Do you think your life will be better five years from now than it is today?,"Yes,No"
- 102674, Do you have any credit card debt that is more than one month old?,"Yes,No"
- 106042,Are you taking any prescription medications?,"Yes,No"
- 106388,Do you work 50+ hours per week?,"Yes,No"
- 108343,Do you feel like you have too much personal financial debt?,"Yes,No"
- 108617,Do you live in a single-parent household?,"Yes,No"
- 109244,Are you a feminist?,"Yes,No"
- 112512,Are you naturally skeptical?,"Yes,No"
- 113992,Do you gamble?,"Yes,No"
- 115899,Would you say most of the hardship in your life has been the result of circumstances beyond your own control
- 123464,Do you currently have a job that pays minimum wage?,"Yes,No"
- 123621,Are you currently employed in a full-time job?,"Yes,No"

5.2 Splitting the Training and Test Datasets

We have learned that when splitting the datasets, there are two competing concerns.

- Having too small training data, our parameters estimates will have greater variance.
- Having too small testing data, our performance statistics will have greater variance.

A training dataset is defined to be the data used to fit or train the model. Conversely, a testing dataset is the sample of the data used to provide an unbiased evaluation of the final model which was determined from the training dataset.

Optimal performance of our machine learning model is achieved by identifing the best split ratio between the training and testing dataset. The larger the original dataset the more appropriate it is to identify an optimial split ratio thereby improving the effectiviness of both teaching and testing the model.

Common split percentages vary from:

- Train: 80%, Test: 20%
- Train: 70%, Test: 30%
- Train: 67%, Test: 33%
- Train: 50%, Test: 50%

The most common used ratio is the 80:20 split, refered to as the Pareto Principle, which states that roughly 80% of the effects come from 20% of the causes.

According to the following research done at AT&T Bell Laboratories, A Scaling law for validation-set training-set size ratio the optimal ratio is achieved through the following formula: Test Set(v) : training set(t) = v/t, scales like ln(N/h-max), where N is the number of data families and h-max is the largest complexity of these families.

Ratio selection is also influenced by which type of modeling technique is being used. Since we are using multiple modeling techniques, we will choose a less signitific method for selecting our final split ratio. We will run our modeling project with the for ratios above and select the ratio which offers the best accuracy results.

While changing the ratio didn't significantly change our accuracy results, the best performing ratio was an 70/30 split. Therefore, we will use the 80/20 split for the final version of this project.

Table 1: Train and Test Ratio Splits - Comparing Results

Train	Test	Optimal.Algorithm	Optimal.Accuracy
0.80	0.20	Leave-1-Out	0.59110
0.70	0.30	Leave-1-Out	0.59443
0.67	0.33	Leave-1-Out	0.58898
0.60	0.40	Leave-1-Out	0.58546
0.50	0.50	Leave-1-Out	0.56955

ModelType	Accuracy	Precision	Sensitivity	Specifity	Winner
Classification Model: CART	0.54717	0.60386	0.42373	0.42373	Democrat
Tree Based Model: Random Forest (RFM)	0.55599	0.56379	0.67990	0.67990	Democrat
Conditional Probability Model - Naive Bayes	0.55256	0.56706	0.65932	0.65932	Democrat
Logistic Regression Model (LRM) - Stepwise	0.55208	0.56159	0.66749	0.66749	Democrat
Logistic Regression Moodel (LRM) - BLR	0.55208	0.56159	0.66749	0.66749	Democrat
Logistic Regression Model (LRM) - LDA	0.55208	0.56159	0.66749	0.66749	Democrat
Logistic Regression Model (LRM) - QDA	0.54688	0.58017	0.49380	0.49380	Democrat
* TUNED LRM - LDA Model	0.56552	0.55855	0.80116	0.80116	Democrat
* TUNED LRM - QDA Model	0.50302	0.60163	0.14286	0.14286	Democrat
* TUNED LRM - Stepwise	0.56956	0.56075	0.81081	0.81081	Democrat
* TUNED - Naive Bayes	0.56604	0.56646	0.77288	0.77288	Democrat
* TUNED - LRM - BLR	0.51714	0.52759	0.75385	0.75385	Democrat
* TUNED - Random Forest	0.56754	0.56815	0.71622	0.71622	Democrat
* TUNED - CART	0.53010	0.53010	1.00000	1.00000	Democrat
Cross Validation Model - Leave-One-Out	0.59711	0.59958	0.70574	0.70574	Democrat
Cross Validation Model - k-Fold	0.54886	0.59861	0.44560	0.44560	Democrat

ModelType	Accuracy	Precision	Sensitivity	Specifity	Winner
Classification Model: CART	0.53713	0.55785	0.61017	0.61017	Democrat
Tree Based Model: Random Forest (RFM)	0.54530	0.55949	0.64648	0.64648	Democrat
Conditional Probability Model - Naive Bayes	0.54970	0.56413	0.66102	0.66102	Democrat
Logistic Regression Model (LRM) - Stepwise	0.55306	0.56467	0.66448	0.66448	Democrat
Logistic Regression Moodel (LRM) - BLR	0.55306	0.56467	0.66448	0.66448	Democrat
Logistic Regression Model (LRM) - LDA	0.55306	0.56467	0.66448	0.66448	Democrat
Logistic Regression Model (LRM) - QDA	0.52459	0.56303	0.43863	0.43863	Democrat
* TUNED LRM - LDA Model	0.56091	0.55527	0.82015	0.82015	Democrat
* TUNED LRM - Stepwise	0.55823	0.55506	0.79719	0.79719	Democrat
* TUNED - Naive Bayes	0.56407	0.56569	0.76384	0.76384	Democrat
* TUNED - LRM - BLR	0.50803	0.52504	0.76671	0.76671	Democrat
* TUNED - Random Forest	0.56560	0.56153	0.78571	0.78571	Democrat
* TUNED - CART	0.52994	0.52994	1.00000	1.00000	Democrat
Cross Validation Model - Leave-One-Out	0.59443	0.59437	0.70333	0.70333	Democrat
Cross Validation Model - k-Fold	0.55122	0.60283	0.44291	0.44291	Democrat

ModelType	Accuracy	Precision	Sensitivity	Specifity	Winner
Classification Model: CART	0.54516	0.56485	0.61704	0.61704	Democrat
Tree Based Model: Random Forest (RFM)	0.54917	0.56450	0.65579	0.65579	Democrat
Conditional Probability Model - Naive Bayes	0.55767	0.56851	0.68583	0.68583	Democrat
Logistic Regression Model (LRM) - Stepwise	0.56255	0.57412	0.67804	0.67804	Democrat
Logistic Regression Moodel (LRM) - BLR	0.56255	0.57412	0.67804	0.67804	Democrat
Logistic Regression Model (LRM) - LDA	0.56255	0.57412	0.67804	0.67804	Democrat
Logistic Regression Model (LRM) - QDA	0.53501	0.57815	0.45549	0.45549	Democrat
* TUNED LRM - LDA Model	0.56197	0.55723	0.81040	0.81040	Democrat
* TUNED LRM - QDA Model	0.50911	0.64322	0.14798	0.14798	Democrat
* TUNED LRM - Stepwise	0.55650	0.55261	0.81965	0.81965	Democrat
* TUNED - Naive Bayes	0.56094	0.56379	0.75770	0.75770	Democrat
* TUNED - LRM - BLR	0.52309	0.53676	0.77335	0.77335	Democrat
* TUNED - Random Forest	0.56622	0.56481	0.76069	0.76069	Democrat
* TUNED - CART	0.52992	0.52992	1.00000	1.00000	Democrat
Cross Validation Model - Leave-One-Out	0.58898	0.59027	0.69532	0.69532	Democrat
Cross Validation Model - k-Fold	0.55205	0.60340	0.44607	0.44607	Democrat

ModelType	Accuracy	Precision	Sensitivity	Specifity	Winner
Classification Model: CART	0.54019	0.55299	0.68983	0.68983	Democrat
Tree Based Model: Random Forest (RFM)	0.55491	0.57692	0.65139	0.65139	Democrat
Conditional Probability Model - Naive Bayes	0.55366	0.56586	0.67712	0.67712	Democrat
Logistic Regression Model (LRM) - Stepwise	0.55880	0.58099	0.64897	0.64897	Democrat
Logistic Regression Moodel (LRM) - BLR	0.55880	0.58099	0.64897	0.64897	Democrat
Logistic Regression Model (LRM) - LDA	0.55880	0.58099	0.64897	0.64897	Democrat
Logistic Regression Model (LRM) - QDA	0.53021	0.58230	0.45235	0.45235	Democrat
* TUNED LRM - LDA Model	0.55924	0.55686	0.80989	0.80989	Democrat
* TUNED LRM - QDA Model	0.50251	0.62348	0.14639	0.14639	Democrat
* TUNED LRM - Stepwise	0.54669	0.55871	0.67395	0.67395	Democrat
* TUNED - Naive Bayes	0.55860	0.56222	0.75424	0.75424	Democrat
* TUNED - LRM - BLR	0.52410	0.53237	0.77746	0.77746	Democrat
* TUNED - Random Forest	0.56225	0.56686	0.72529	0.72529	Democrat
* TUNED - CART	0.52986	0.52986	1.00000	1.00000	Democrat
Cross Validation Model - Leave-One-Out	0.58546	0.58684	0.68514	0.68514	Democrat
Cross Validation Model - k-Fold	0.54879	0.59953	0.44550	0.44550	Democrat

ModelType	Accuracy	Precision	Sensitivity	Specifity	Winner
Classification Model: CART	0.55316	0.56315	0.69831	0.69831	Democrat
Tree Based Model: Random Forest (RFM)	0.55225	0.57168	0.64384	0.64384	Democrat
Conditional Probability Model - Naive Bayes	0.55675	0.56827	0.68000	0.68000	Democrat
Logistic Regression Model (LRM) - Stepwise	0.55643	0.57555	0.64481	0.64481	Democrat
Logistic Regression Moodel (LRM) - BLR	0.55643	0.57555	0.64481	0.64481	Democrat
Logistic Regression Model (LRM) - LDA	0.55643	0.57555	0.64481	0.64481	Democrat
Logistic Regression Model (LRM) - QDA	0.53657	0.58323	0.46282	0.46282	Democrat
* TUNED LRM - LDA Model	0.56089	0.55913	0.81853	0.81853	Democrat
* TUNED LRM - QDA Model	0.49677	0.60681	0.14882	0.14882	Democrat
* TUNED LRM - Stepwise	0.53669	0.56656	0.54290	0.54290	Democrat
* TUNED - Naive Bayes	0.55963	0.55425	0.86237	0.86237	Democrat
* TUNED - LRM - BLR	0.51411	0.53268	0.77160	0.77160	Democrat
* TUNED - Random Forest	0.55645	0.56620	0.70463	0.70463	Democrat
* TUNED - CART	0.52981	0.52981	1.00000	1.00000	Democrat
Cross Validation Model - Leave-One-Out	0.56955	0.58011	0.65768	0.65768	Democrat
Cross Validation Model - k-Fold	0.54929	0.60189	0.44122	0.44122	Democrat

Figure 1: Train and Test Ratio Splits - Actual Results

5.3 Data Structure Analysis - The Datasets

A complete inspection of the datasets have been provided in Appendix - B - Dataset Inspection.

5.4 Visualizing the Data / Initial Observations

As an initial method to help understand the survey results collected from the participants, we have generated some simple distribution graphs. These graphs can be found in Appendix - C: Demographic Figures.

Summarizing the demographics:

- Dataset by Party : Results show 53% Democrat, 47% Republican
- Dataset by Gender : Results show 39.1% Female, 60.1% Male
- Dataset by Income Bands : Top 2 largest Income bands are - 50k-74,999 @ 18.4%, 100,001-150k @ 17.5%
- Dataset by Household Status : Top 2 largest Household Status bands are: Single (no kids) @ 46.7%, Married (w/kids) @ 31.6%
- Dataset by Educaiton Levels : Top 2 largest Education Level bands are: Bachelor's degree @ 25.9%, Current K-12 @ 17.8%
- Dataset by Age : Average age was 48.16 years.

6 Correlation Matrixes and HeatMaps

Appendix - F (Correlation Matrixes and Heat Maps) helps identify those fields which are highly correlated.

A subset of these visualizations have been provided below. Positive correlations are displayed in blue and negative correlations in red color. Color intensity and the size of the circle are proportional to the correlation coefficients. In the right side of the correlogram, the legend color shows the correlation coefficients and the corresponding colors.

We have trimmed the correlation matrix results to only show the up quadrant. From these plots, we can identify high correlations to be:

- Dataset without Questions: The top 4 coorelations are HouseholdStatus, Income, YOB, and Party
- Dataset with Questions : The top 4 coorelations are: Q106388, Q102674, Q100562, and Q1023621

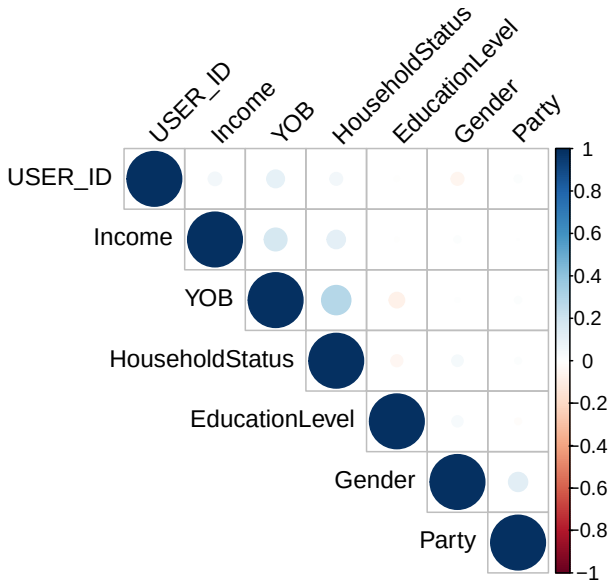


Figure 2: Correlation Matrixes

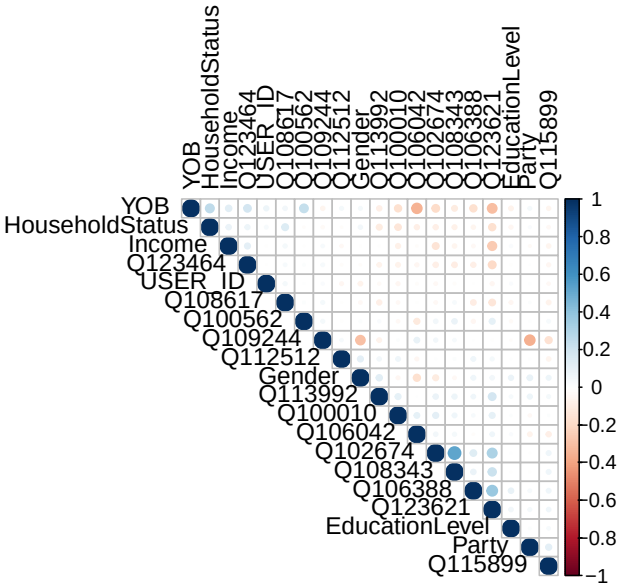


Figure 3: Correlation Matrixes

7 The Modeling Approach: Simple and More

7.1 Modeling Background

Machine Learning can be summarized as a learning function (f) that maps input variables (X) to output variables (Y).

$$Y = f(x)$$

. This basic function takes on different forms, and is for all general purposes unknown. The model's learning aspect is accomplished through training data.

Different algorithms make different assumptions or biases about the form of the function and how it can be learned.

On the preceeding page we have provided an overiview of just a few of these different methods respective to their machine learning disciplines.

For the purpose of this project, we will focuse on the following model methods:

- `Classification Trees`
- Conditional Probability
- `Linear Regression`
- `Logistic Regression`
- `Tree Based Model`
- Cross Validation Model

7.2 Machine Learning Modeling Methods

Figure 4: Machine Learning Algorithms

7.3 Models Choosen

During our project endevour, we will analyize the results of 10 different models across these 6 modeling categories. The majority of our models used are within the logistic regression discipline.

- Classification Trees : Classification and Regression Tree (CART)
- Conditional Probability : Naive Bayes (NB)
- Linear Regression : Binary Linear Regression (BLR)
- Logistic Regression
  - Stepwise (Simple)
  - Binary Logistic Regression (BLR)
  - Linear Discriminant Analysis (LDA)
  - Quadratic Discriminant Anlaysis (QDA)
- Tree-Based : Random Forest
- Cross-Validation
  - k-Fold
  - Leave-One-Out

7.4 Strengths and Weaknesses of the Modeling Categories

Strengths and weaknesses of each of the selected model categories has been summarized in the tabel below.

Machine Learning Algorithm	Models Used	Model Strengths	Model Weaknes
Linear Regression	**Bayesian Linear Regression (BLR)**	Linear regression is straightforward and a relatively simple method. It can be regularized to avoid overfitting. Llinear models can be updated easily with new data. Linear regression models are relatively easy to implement.	Linear regression performs poorly when there are non-linear relationships. LR Models are not naturally flexible enough to capture more complex patterns. Adding the right interaction terms or polynomials can be tricky and time-consuming. LR assumes linear relationship between dependent and independent variables, which is incorrect in most cases. It is sensitive to outliers. If the number of observations are less, it leads to over
Logistic Regression	**Linear Discriminant Analysis (LDA)** **Quadratic Discriminant Analysis (QDA)**	Outputs have a nice probabilistic interpretation, and the algorithm can be regularized to avoid overfitting. Logistic models can be updated easily with new data. Logistic Regression Models do not assume linear relationship between independent and dependent variables. Dependent variables does not need to be normally distributed.	Logistic regression tends to underperform when there are multiple or non-linear decision boundaries. Logistic Regression Models are not flexible enough to naturally capture more complex relationships. Requires more data to achieve stability. Effective mostly on linearly separable.
Classification Trees	**Classification and Regression Tree (CART)**	Classification tree ensembles perform very well in practice. They are robust to outliers, scalable, and able to naturally model non-linear decision boundaries.	Unconstrained, individual trees are prone to overfitting.
Conditional Probability	**Naïve Bayes (NB)**	NB models actually perform surprisingly well in practice. They are easy to implement and can scale with your dataset	Due to their sheer simplicity, NB models are often beaten by models properly trained and tuned using the previous algorithms listed.
Tree Based	**Random Forest (RF)**	One third of data is not used for training, hence it can be used for testing. Tree based models have are high performaning and accurate. Provides feature importance estimate. Can automatically handle missing values.	Less interpret-ability. Can over fit the data. Requires more computational resources Prediction time is high
Cross Validation	**k-Fold** **Leave-One-Out**	Reduces Overfitting The model attains their generalization capabilities. Provides Hyperparameter Tuning to increase the efficiency of the algorithm. It can balance out the predicted features' classes if there are unbalanced datasets. Calculate differences and standard deviation.	Increases Training Time. Cross Validation requires training the model on multiple training sets. Needs Expensive Computation: Cross Validation is computationally very expensive in terms of processing power required.

Figure 5: Model Comparisons

7.5 Model Tuning

Initial analysis was performed on the datasets excluding the original survey questions. These models were later tuned by including the subset of the *key* questions.

From the datasets, the field USER_ID was irreleveant. For the fields Gender, HouseholdStatus and EducationLevel the data was analyized via buckets.

- Gender (2 Buckets)

```
- Male
- Female
```

- Income (6 Buckets)

```
- "under $25,000"
- "$25,001 - $50,000"
- "$50,000 - $74,999"
- "$75,000 - $100,000"
- "$100,001 - $150,000"
- or "over $150,000"
```

- HouseholdStatus (6 Buckets)

```
- "Domestic Partners (no kids)"
- "Domestic Partners (w/kids)"
- "Married (no kids)"
- "Married (w/kids)"
- "Single (no kids)"
- or "Single (w/kids)"
```

- EducationalLevel (7 Buckets)

```
- "Current K-12"
- "High School Diploma"
- "Current Undergraduate"
- "Associate's Degree"
- "Bachelor's Degree"
- "Master's Degree"
- or "Doctoral Degree".
```

Worth noting, these buckets were already created as part of the original data wrangling.

7.6 Initial Model Results

We now begin our model analysis. As described in the previous section, we will be analyzing the following machine learning algorithms.

- Classification Trees : Classification and Regression Tree (CART)
- Conditional Probability : Naive Bayes (NB)
- Linear Regression : Binary Linear Regression (BLR)
- Logistic Regression : Stepwise (Simple) : Binary Logistic Regression (BLR) : Linear Discriminant Analysis (LDA) : Quadratic Discriminant Anlaysis (QDA)
- Tree-Based : Random Forest
- Cross-Validation : k-Fold : Leave-One-Out

For each algorithm, we will summarize the results in two tables. A confusion matrix table and a combined results table highlighting the best and worst performs up to that point, starting wth classifications trees.

7.6.1 Classification Tree Model - Classification and Regression Tree (CART)

A Classification And Regression Tree (CART), is a predictive model, which explains how an outcome's values can be predicted based on other values. A CART output is a decision tree where each fork is a split in a predictor variable and each end node contains a prediction for the outcome variable.

In addtion to our tables, we will also present a decision tree graphic to illustrate the different forks and predictors.

Decisons tree based for the Classificaiton Model - CART

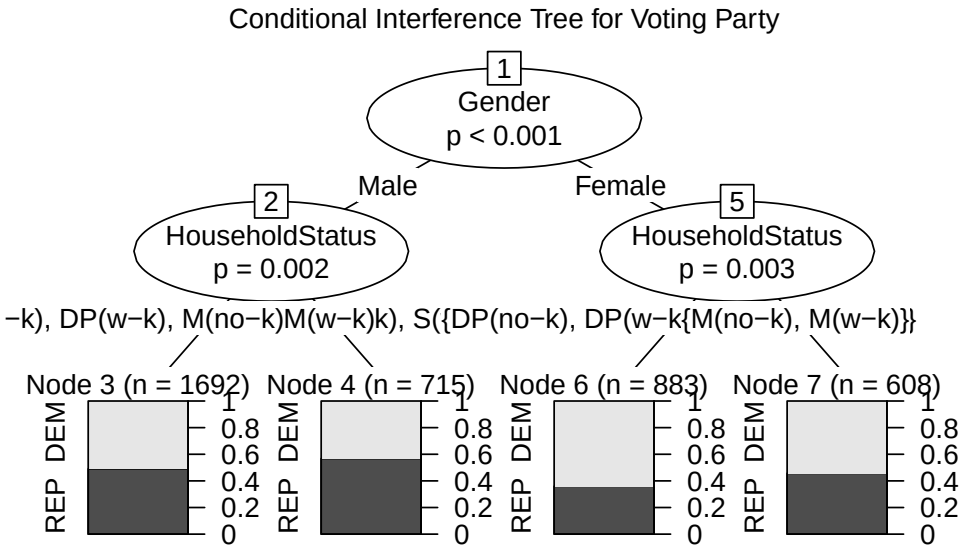


Figure 6: Decision Tree

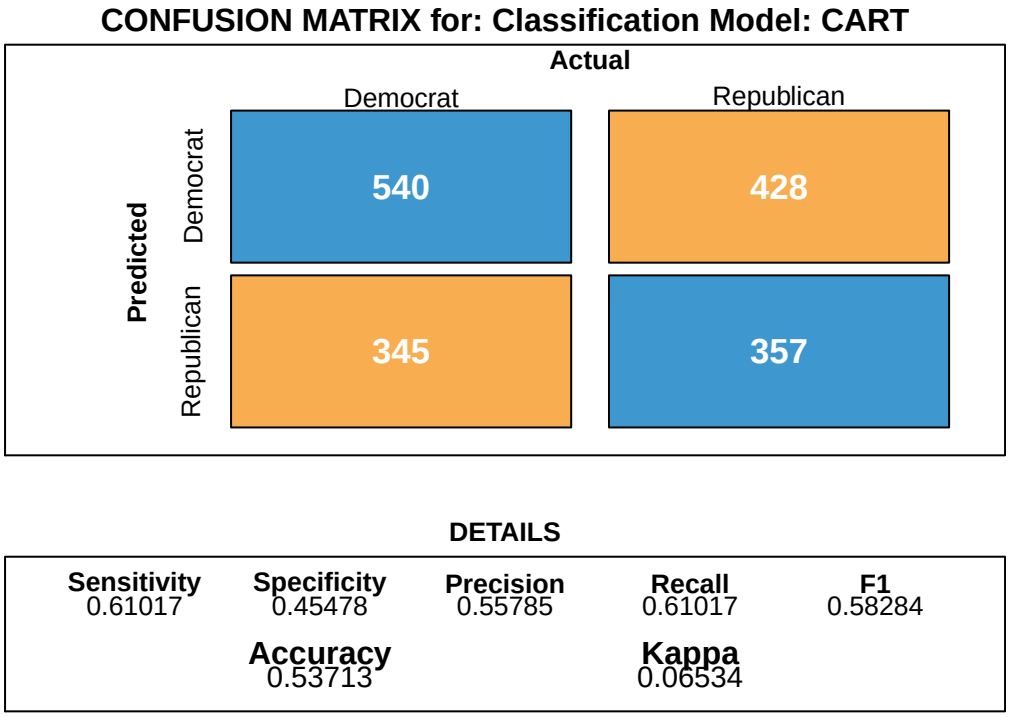


Figure 7: Confustion Matrix - CART

Table 2: Prediction Results: Classification Model: CART Model - Added

ModelType	Accuracy	Precision	Sensitivity	Specifity	Winner
Classification Model: CART	0.53713	0.55785	0.61017	0.61017	Democrat

7.6.2 Tree Based Model - Random Forest Model (RFM)

Continuing our tree based models, we will look at the performace of a Random Forest Model (RFM). A Random Forest Tree is a learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at the training time and outputting classes (classification) or mean prediction (regression) of the individual trees.

Our results are as follows:

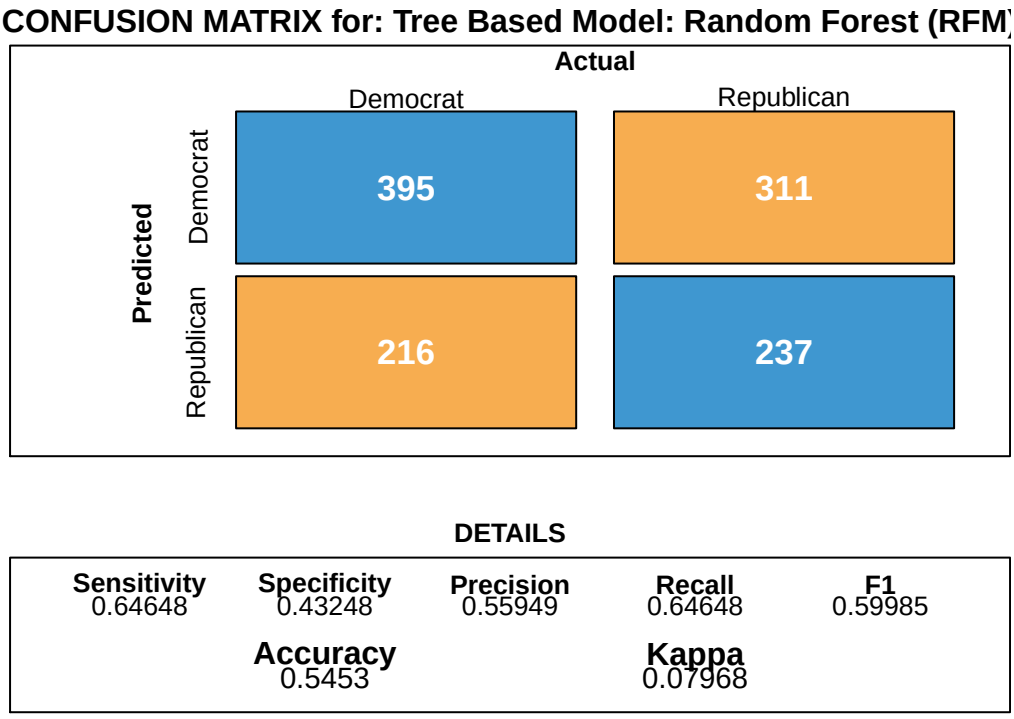


Figure 8: Confusion Matrix - RFM

Table 3: Prediction Results: Tree Based Model: Random Forest (RFM) Model - Added

ModelType	Accuracy	Precision	Sensitivity	Specifity	Winner
Classification Model: CART	0.53713	0.55785	0.61017	0.61017	Democrat
Tree Based Model: Random Forest (RFM)	0.54530	0.55949	0.64648	0.64648	Democrat

7.6.3 Conditional Probablity Model - Naive Bayes (NB)

Continuing our learning model momentum, we move on to a different class of models. Specifically, conditional probability modeling and the Naive Bayes (NB) algorithm.

Naïve Bayes is a classification method based on Bayes' theorem that derives the probability of the given feature vector being associated with a label. Naïve Bayes has a naive assumption of conditional independence for every feature, which means that the algorithm expects the features to be independent which may not always be the case.

Naïve Bayes assumes all the features to be conditionally independent. So, if some of the features are in fact dependent on each other (in case of a large feature space), the prediction might be poor.

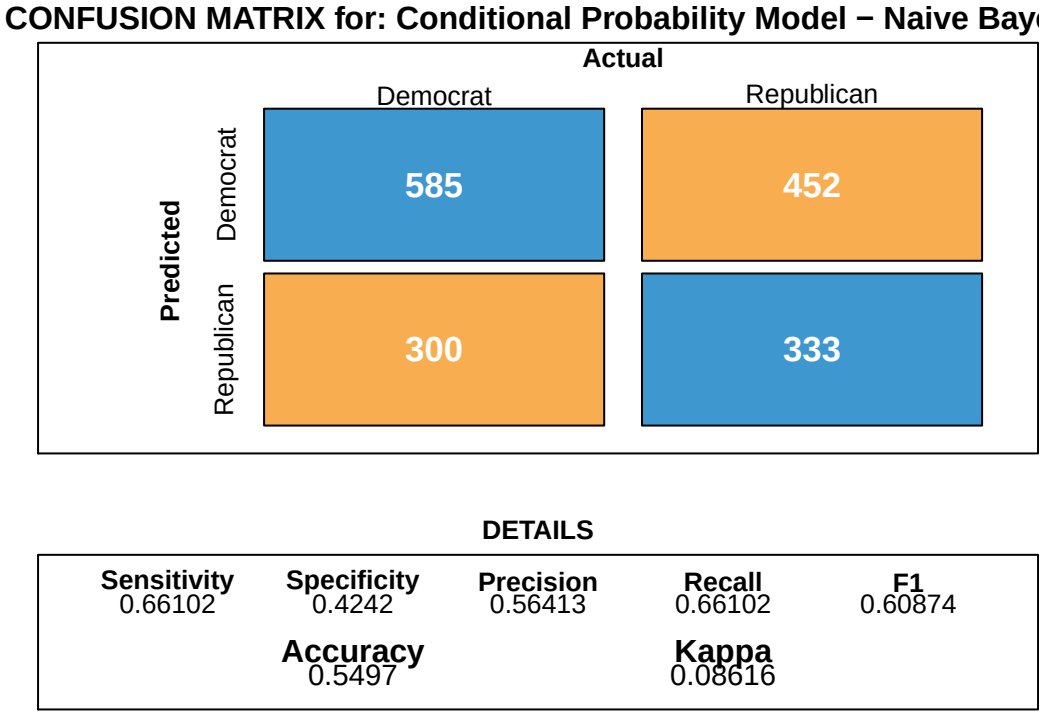


Figure 9: Confusion Matrix - Naive Bayes

Table 4: Prediction Results: Conditional Probability Model - Naive Bayes Model - Added

ModelType	Accuracy	Precision	Sensitivity	Specifity	Winner
Classification Model: CART	0.53713	0.55785	0.61017	0.61017	Democrat
Tree Based Model: Random Forest (RFM)	0.54530	0.55949	0.64648	0.64648	Democrat
Conditional Probability Model - Naive Bayes	0.54970	0.56413	0.66102	0.66102	Democrat

7.6.4 Logistic Regression Model (LRM) - Step Wise

Moving on to a different type of regression models (linear and logistic), We will start with logistic regression algorithms. The majority of our models will be from this discipline. Logistic regression is an algorithm used to predict the probability of a target variable. The nature of the target or dependent variable is dichotomous, which means there would be only two possible classes.

It is the go-to method for binary classification problems (problems with two class values).

Starting off the logistic regressions series of models is Stepwise Regression. Stepwise Regression is a method of fitting regression models in which the choice of predictive variables is carried out by a procedure where each step utilizes a variable considered for addition to or subtraction from the set of explanatory variables based on some prespecified criterion.

The results from our Stepwise mode are listed below.

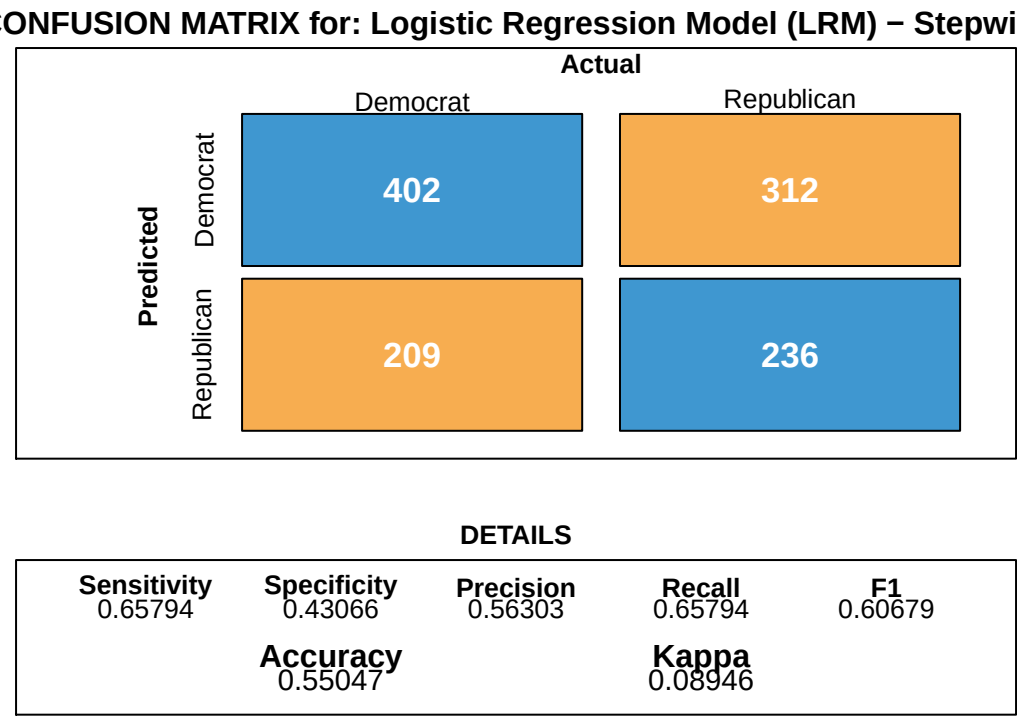


Figure 10: Confusion Matrix - LRM

Table 5: Prediction Results: Logistic Regression Model (LRM) - Stepwise Model - Added

ModelType	Accuracy	Precision	Sensitivity	Specifity	Winner
Classification Model: CART	0.53713	0.55785	0.61017	0.61017	Democrat
Tree Based Model: Random Forest (RFM)	0.54530	0.55949	0.64648	0.64648	Democrat
Conditional Probability Model - Naive Bayes	0.54970	0.56413	0.66102	0.66102	Democrat
Logistic Regression Model (LRM) - Stepwise	0.55047	0.56303	0.65794	0.65794	Democrat

7.6.5 Logistic Regression (BLR) - Binary

Binary logistic regression (BLR) is the simplest form of logistic regression in which the target or dependent variable can have only 2 possible types either 1 or 0.

Results from our binary logistic regression model can be found below.

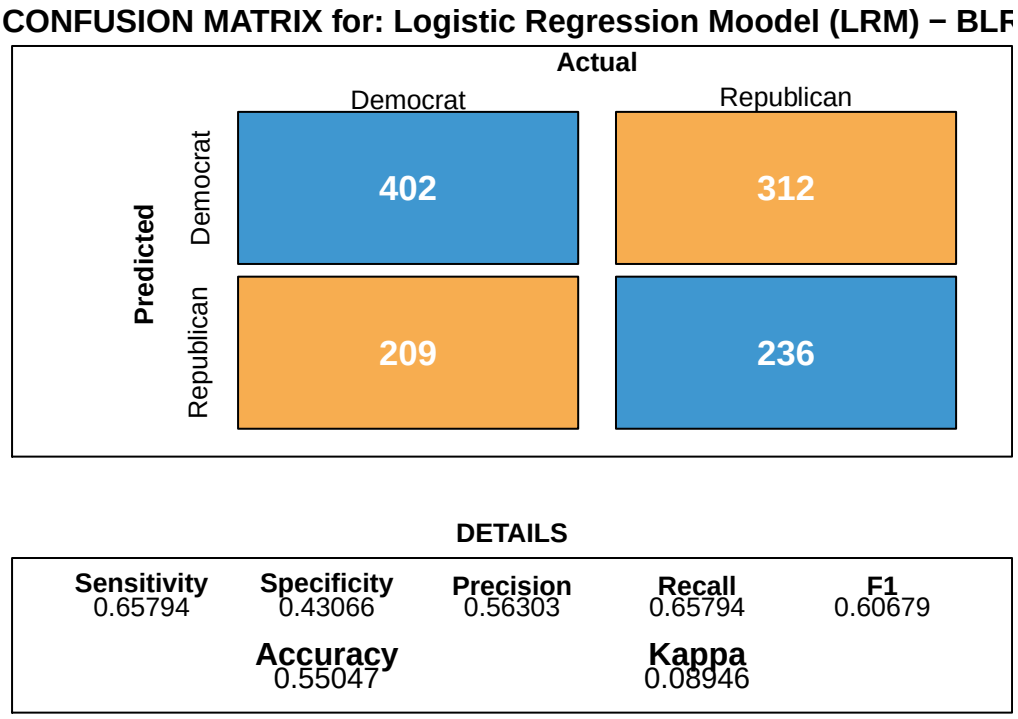


Figure 11: Confusion Matrix - BLR

Table 6: Prediction Results: Logistic Regression Moodel (LRM) - BLR Model - Added

ModelType	Accuracy	Precision	Sensitivity	Specifity	Winner
Classification Model: CART	0.53713	0.55785	0.61017	0.61017	Democrat
Tree Based Model: Random Forest (RFM)	0.54530	0.55949	0.64648	0.64648	Democrat
Conditional Probability Model - Naive Bayes	0.54970	0.56413	0.66102	0.66102	Democrat
Logistic Regression Model (LRM) - Stepwise	0.55047	0.56303	0.65794	0.65794	Democrat
Logistic Regression Moodel (LRM) - BLR	0.55047	0.56303	0.65794	0.65794	Democrat

7.6.6 Logistic Regression - Latent Dirichlet Allocation (LDA)

Linear Dirichlet Allocation (LDA) logistic regression algorithems. LDA algorithms are a generalization of Fisher's linear discriminant, a method used in statistics, pattern recognition, and machine learning to find a linear combination of features that characterizes or separates two or more classes of objects or events. The resulting combination may be used as a linear classifier, or, more commonly, for dimensionality reduction before later classification

Below are our observed resutls using LDA techniques.

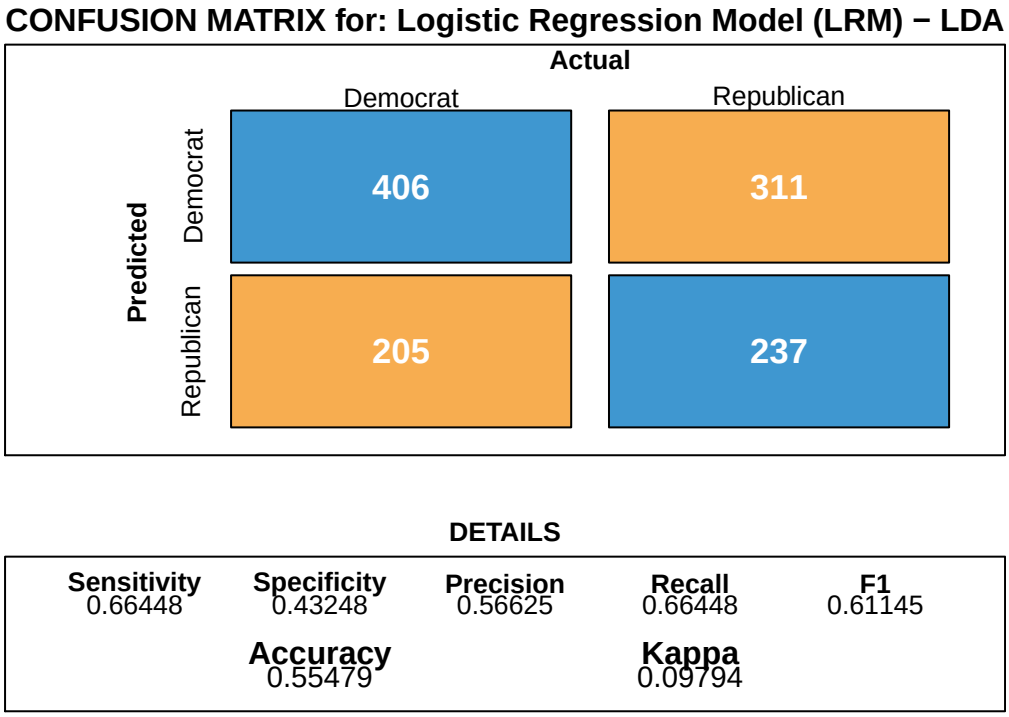


Figure 12: Confusion Matrix - LDA

Table 7: Prediction Results: Logistic Regression Model (LRM) - LDA Model - Added

ModelType	Accuracy	Precision	Sensitivity	Specifity	Winner
Classification Model: CART	0.53713	0.55785	0.61017	0.61017	Democrat
Tree Based Model: Random Forest (RFM)	0.54530	0.55949	0.64648	0.64648	Democrat
Conditional Probability Model - Naive Bayes	0.54970	0.56413	0.66102	0.66102	Democrat
Logistic Regression Model (LRM) - Stepwise	0.55047	0.56303	0.65794	0.65794	Democrat
Logistic Regression Moodel (LRM) - BLR	0.55047	0.56303	0.65794	0.65794	Democrat
Logistic Regression Model (LRM) - LDA	0.55479	0.56625	0.66448	0.66448	Democrat

7.6.7 Logistic Regression - Quadratic Discrimination Analysis (QDA)

Like LDA, the QDA classifier assumes that for the observations each identified class is drawn from a Gaussian distribution. However, unlike LDA, QDA assumes that each class has its own covariance matrix. In other words, the predictor variables are not assumed to have common variance across each of their associated levels.

QDA performance is . . .

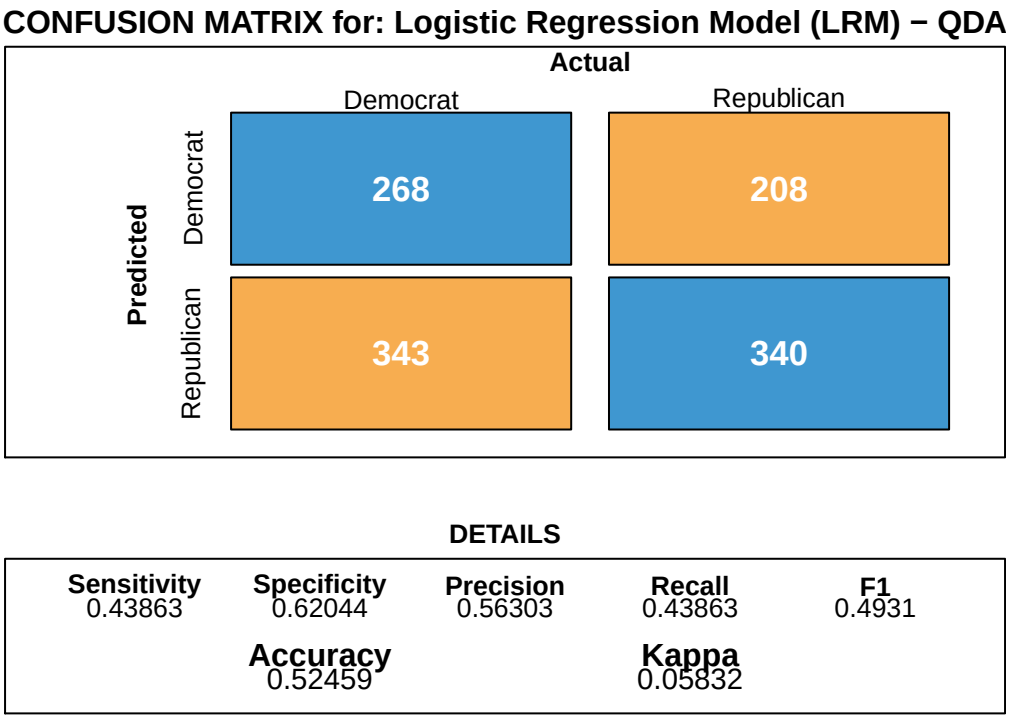


Figure 13: Confusion Matrix - QDA

Table 8: Prediction Results: Logistic Regression Model (LRM) - QDA Model - Added

ModelType	Accuracy	Precision	Sensitivity	Specifity	Winner
Classification Model: CART	0.53713	0.55785	0.61017	0.61017	Democrat
Tree Based Model: Random Forest (RFM)	0.54530	0.55949	0.64648	0.64648	Democrat
Conditional Probability Model - Naive Bayes	0.54970	0.56413	0.66102	0.66102	Democrat
Logistic Regression Model (LRM) - Stepwise	0.55047	0.56303	0.65794	0.65794	Democrat
Logistic Regression Moodel (LRM) - BLR	0.55047	0.56303	0.65794	0.65794	Democrat
Logistic Regression Model (LRM) - LDA	0.55479	0.56625	0.66448	0.66448	Democrat
Logistic Regression Model (LRM) - QDA	0.52459	0.56303	0.43863	0.43863	Democrat

7.7 Model Tuning Results

In the previous sections, we explored 7 different predictor algorithms without considering any tuning opportunities. We will now apply a generalized tuning across these 7 different algorithms to see if our prediction results experience any improvements.

Our tuning approach will include additonal predictors which were based upon the survey questionaire.

While the questionaire included 100 questions, we used our correlation matrixes and heatmaps to narrow our tuning dataset to include only those questions that had the highest correlation.

In the interest of conserving space, we will only display that final table of results vs. individual model results which was done in the previous sections.

Table 9: Prediction Results: * TUNED - CART Model - Added

ModelType	**Accuracy**	**Precision**	**Sensitivity**	**Specifity**	**Winner**
Classification Model: CART	0.53713	0.55785	0.61017	0.61017	Democrat
Tree Based Model: Random Forest (RFM)	0.54530	0.55949	0.64648	0.64648	Democrat
Conditional Probability Model - Naive Bayes	0.54970	0.56413	0.66102	0.66102	Democrat
Logistic Regression Model (LRM) - Stepwise	0.55047	0.56303	0.65794	0.65794	Democrat
Logistic Regression Moodel (LRM) - BLR	0.55047	0.56303	0.65794	0.65794	Democrat
Logistic Regression Model (LRM) - LDA	0.55479	0.56625	0.66448	0.66448	Democrat
Logistic Regression Model (LRM) - QDA	0.52459	0.56303	0.43863	0.43863	Democrat
* TUNED LRM - LDA Model	0.56091	0.55527	0.82015	0.82015	Democrat
* TUNED LRM - QDA Model	0.50937	0.64407	0.14541	0.14541	Democrat
* TUNED LRM - Stepwise	0.55823	0.55506	0.79719	0.79719	Democrat
* TUNED - Naive Bayes	0.56407	0.56569	0.76384	0.76384	Democrat
* TUNED - LRM - BLR	0.50803	0.52504	0.76671	0.76671	Democrat
* TUNED - Random Forest	0.56560	0.56153	0.78571	0.78571	Democrat
* TUNED - CART	0.52994	0.52994	1.00000	1.00000	Democrat

7.7.1 Model Resampling: Cross-Validatioan - Leave-One-Out

Venturing out a bit farther in the relm of machine learning algorithms, we will be extending our modeling analysis to include another class of models called cross-validation. Cross-validation is a procedure that has a single parameter called k which refers to the number of groups or folds that a given data sample is to be split into. Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. It uses a limited sample in order to estimate how the model is expected to perform.

It is a popular method because it is simple to understand and because it generally results in a less biased or less optimistic estimate of the model skill over other methods, such as a simple train/test split.

The first (of our two) cross-validation algorithms will be the Leave-One-Out. Leave-One-Out models number of folds equals the number of instances in the data set. Thus, the learning algorithm is applied once for each instance, using all other instances as a training set and using the selected instance as a single-item test set. This process is closely related to the statistical method of jack-knife estimation.

Our Leave-One-Out performance results are listed below.

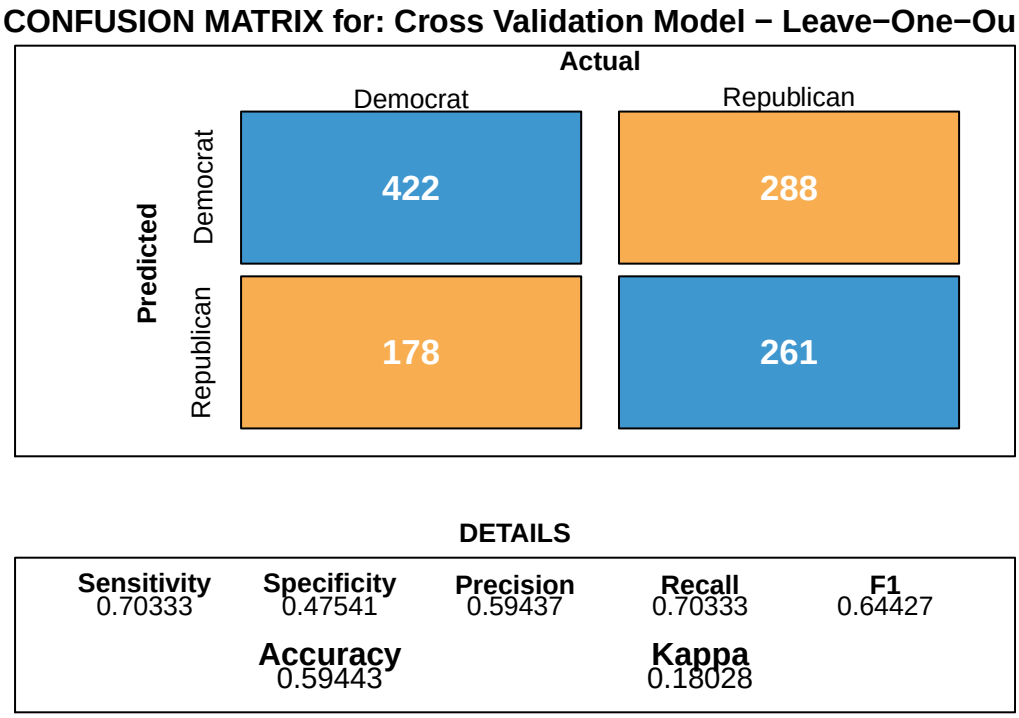


Figure 14: Confusion Matrix - X-VAL - Leave-One-Out

```
##
## Call:
## glm(formula = Party ~ ., family = "binomial", data = train_sample_no_na)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.603  -1.109  -0.868   1.172   1.804
##
## Coefficients:
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                             -1.68e+01   7.30e+00   -2.30  0.02166
## USER_ID                                  2.72e-05   2.04e-05    1.33  0.18267
## YOB                                      8.02e-03   3.70e-03    2.17  0.03012
## GenderMale                               5.07e-01   8.40e-02    6.04  1.5e-09
## Income$25,001 - $50,000                  6.44e-02   1.46e-01    0.44  0.65941
## Income$50,000 - $74,999                 -5.58e-02   1.35e-01   -0.41  0.67884
## Income$75,000 - $100,000                -1.07e-01   1.40e-01   -0.77  0.44347
## Incomeover $150,000                      2.20e-01   1.41e-01    1.56  0.11953
## Incomeunder $25,000                     -3.34e-02   1.50e-01   -0.22  0.82374
## HouseholdStatusDomestic Partners (w/kids)  2.20e-01   4.20e-01    0.52  0.60125
## HouseholdStatusMarried (no kids)         7.15e-01   2.56e-01    2.79  0.00523
## HouseholdStatusMarried (w/kids)          8.92e-01   2.43e-01    3.67  0.00024
## HouseholdStatusSingle (no kids)          4.30e-01   2.42e-01    1.78  0.07520
## HouseholdStatusSingle (w/kids)           5.47e-01   3.00e-01    1.82  0.06846
## EducationLevelBachelor's Degree         -3.34e-01   1.54e-01   -2.17  0.03006
## EducationLevelCurrent K-12              -1.03e-01   1.95e-01   -0.53  0.59670
## EducationLevelCurrent Undergraduate     -3.98e-01   1.78e-01   -2.23  0.02555
## EducationLevelDoctoral Degree           -4.50e-01   2.40e-01   -1.87  0.06118
## EducationLevelHigh School Diploma       -1.33e-01   1.68e-01   -0.79  0.42955
## EducationLevelMaster's Degree           -4.43e-01   1.72e-01   -2.57  0.01010
##
## (Intercept)                             *
## USER_ID
## YOB                                     *
## GenderMale                              ***
## Income$25,001 - $50,000
## Income$50,000 - $74,999
## Income$75,000 - $100,000
## Incomeover $150,000
## Incomeunder $25,000
## HouseholdStatusDomestic Partners (w/kids)
## HouseholdStatusMarried (no kids)        **
## HouseholdStatusMarried (w/kids)         ***
## HouseholdStatusSingle (no kids)         .
## HouseholdStatusSingle (w/kids)          .
## EducationLevelBachelor's Degree         *
## EducationLevelCurrent K-12
## EducationLevelCurrent Undergraduate     *
## EducationLevelDoctoral Degree           .
## EducationLevelHigh School Diploma
## EducationLevelMaster's Degree           *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3616.1  on 2617  degrees of freedom
## Residual deviance: 3529.9  on 2598  degrees of freedom
## AIC: 3570
##
## Number of Fisher Scoring iterations: 4
```

7.7.2 Model Resampling: K-Fold

Our final category for both cross-validation and this project is the k-Fold algorithm. A K-Fold cross-validation algortihm takes a given data set, splits it into K number of sections/folds where each fold is used as a testing set at some point.

Using a 5-Fold example, the first iteration - representing the first fold - is used to test the model and the rest are used to train the model. In the second iteration - the 2nd fold is used as the testing set while the rest serve as the training set. This process is repeated until each fold of the 5 folds have been used as the testing set.

This model is highly dependent on a carefully choosen k value.

A poorly chosen value for k may result in mis-representative, such as a score with a high variance (that may change a lot based on the data used to fit the model), or a high bias, (such as an overestimate of the skill of the model).

Three common tactics for choosing a value for k are as follows:

- Representative: The value for k is chosen such that each train/test group of data samples is large enough to be statistically representative of the broader dataset.
- k=10: The value for k is fixed to 10, a value that has been found through experimentation to generally result in a model skill estimate with low bias a modest variance.
- k=n: The value for k is fixed to n, where n is the size of the dataset to give each test sample an opportunity to be used in the hold out dataset. This approach is called leave-one-out cross-validation.

For our analysis we will be using the k=10 approach with results displayed below.

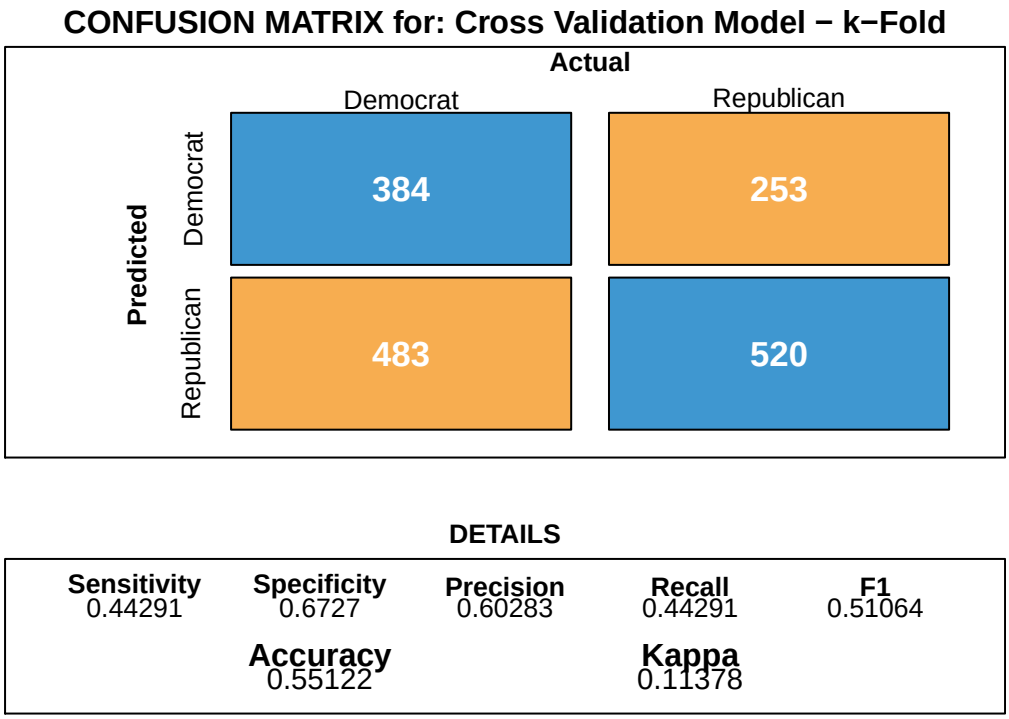


Figure 15: Confusion Matrix - X-VAL - k-Fold

8 Conclusion

We started this project with the desire to achieve two goals. Recalling from Section 3 - Project Goal, these goals were:

- Goal #1: Predict which candidate platform will win the election
- Goal #2: Identify which questions have the largest influence on Goal #1.

In reviewing the results from our corelation matrixes and heatmaps (Section 6), we achieved Goal #2 by identify the questions having the highest correlation. They were: Q106388, Q102674, Q100562 and Q1023621. As an anxillary benefit, we were also able to identify the fields having the highest correlation. The top 4 fields were: HouseholdStatus, Income, YOB and Party.

Based upon our dataset, Goal #1 was uninamiously identified to be the Democratic party. These results were achieved and validated using 10 different modeling techniques spanning 6 different modeling categories.

The following table summarizes our results achieved.

Table 10: Prediction Results: Cross Validation Model - k-Fold Model - Added

ModelType	**Accuracy**	**Precision**	**Sensitivity**	**Specifity**	**Winner**
Classification Model: CART	0.53713	0.55785	0.61017	0.61017	Democrat
Tree Based Model: Random Forest (RFM)	0.54530	0.55949	0.64648	0.64648	Democrat
Conditional Probability Model - Naive Bayes	0.54970	0.56413	0.66102	0.66102	Democrat
Logistic Regression Model (LRM) - Stepwise	0.55047	0.56303	0.65794	0.65794	Democrat
Logistic Regression Moodel (LRM) - BLR	0.55047	0.56303	0.65794	0.65794	Democrat
Logistic Regression Model (LRM) - LDA	0.55479	0.56625	0.66448	0.66448	Democrat
Logistic Regression Model (LRM) - QDA	0.52459	0.56303	0.43863	0.43863	Democrat
* TUNED LRM - LDA Model	0.56091	0.55527	0.82015	0.82015	Democrat
* TUNED LRM - QDA Model	0.50937	0.64407	0.14541	0.14541	Democrat
* TUNED LRM - Stepwise	0.55823	0.55506	0.79719	0.79719	Democrat
* TUNED - Naive Bayes	0.56407	0.56569	0.76384	0.76384	Democrat
* TUNED - LRM - BLR	0.50803	0.52504	0.76671	0.76671	Democrat
* TUNED - Random Forest	0.56560	0.56153	0.78571	0.78571	Democrat
* TUNED - CART	0.52994	0.52994	1.00000	1.00000	Democrat
Cross Validation Model - Leave-One-Out	0.59443	0.59437	0.70333	0.70333	Democrat
Cross Validation Model - k-Fold	0.55122	0.60283	0.44291	0.44291	Democrat

From these results, the top two performing techniques were:

- Cross Validation Leave-One-Out Model with an accuracy of .59443
- Tree Based Random Forest Tuned Model with an accuracy of .56560

Our worst peforming technique was the Binary Logistic Regression model with an accuracy of .50803

Another conclusion that can be extraopolated from our final results envolves the performance of the Random Forest model. Despite having the 2nd best accuracy results, when comparing the top performers for the tuned and non-tuned models the Random Forest model had the biggest gain in accuracy improvement (with a gain of .0203).

Finally, exploring our top performing technique further, we can observe additional information from the top performing model. This information was found in Section 8.7.1-Model Resampling: Cross-Validation - Leave-One_Out which tells us that:

For continuous variables, our top performing interpretations are as follows:

- For every one unit increase in Gender=Male, the log odds of being a Democrat (vs. Republican) increases by .507
- For every one unit increase in Married (w/kids), the log odds of being a Democrat increases by .892
- For every one unit increase in Married (no kids), the log odds of being a Democrat increases by .715

For categorical variables, our top performing technique further, we can observe that:

- Gender: Being Male, changes the log odds of being Democrat by .507
- Income: Being in the Income bracket of $25,001-$50,000 changes the log odds of being Democrate by .644
- Status: Being Married (w/kids) changes the log odds of being Democrat by .892 and Married (w/o kids) changes by .715
- Education: Having a Masters degree changes the log odds of being Democrat by -.443 (in favor of Republican)

Note: National poverty is classified as anyone making less than $32,000 per year for single person household, $43,000 for a household with two persons and $54,300 for a 3 person household. 2020 Health & Human Services Poverty Guidelines / Federal Poverty Levels

9 Future Work

9.1 Data Acquisition Improvements

9.1.1 Voter Turnout

Since the Voting Rights Act of 1965, there has been a long term increase in the ability of individuals to paricipate in elections. Especially here within the United States. Conversely the effects of other legislation intended to increase voter turnout, such as the National Voter Registration Act, have been more limited on their improved performance.

Many believe that voter turnout has a strong correlation to a thriving democracy. Hence, policymakers and citizens often support electorial reform measures based on whether they will or will not increase voter turnout. Despite these political debates, academic research suggests that in most cases, policy changes usually has little or not effct on voter turnout.

However, according to What Affects Voter Turnout Rates. There are 5 major categories that influence voter turnout. They are:

- Electorial Competitiveness
- Election Type
- Voting Laws
- Demographics

Additional research also suggest another category - years with presidential elections. This is defined to be either "on years" representing presidential election years or "off years" those years that are not part of presidential election years. Elections that occur in odd-numbered years and at times other than November typically have significant lower turnout rates.

From the above additions, the only variable which was included in our project were the demographics category. Extending our general model to include these additional categories would be another opportunity.

However, it should be noted that the chief difficulty in using public opinion surveys to determine individual turnout and thus predict election results, is the problem of social-desirability bias. This unique phenonama can be defined as someone voting because of being a 'good citizen'.

Our recommendation would be to avoid using this modeling effect. The other categories mentioned above should be sufficient.

9.1.2 13 keys to the White House

Based on the research devised by the American historian Allan Lichtman and Russian scientist Vladimir Keilis-Borok (Lichtman et al., 1981), the authors were able to create a data model method that included 13 key factors. The authors believed these 13 key factors would more accruately predict presidential election outcomes.

Depending on how many questions were answered in a certain way, their model predicted an outcome of the election. It is believed that this method has been found to be quite predictive of the election results: it has been predictive of every election since the method was devised in 1981.

It would be an interesting exercise to run our model based upon their most recent dataset. Extending predictions even further could be done by adding this data to other predictive datasources.

9.1.3 More Data / Datasources

While the datasets contained a good sample size, it would be interesting to investigate other datas sources espcially those that focus on polling, voter census, and demographic data. Obtaining additional sources will also help minimize any biases that may have occurred within our dataset.

Datasets could also contain such potential influencers as holidays, events, airport travel demographics, and most recently COVID related cases and volume for potentially each state or its entirity.

9.1.4 COVID Demographics

With the recent COVID pandemic, it is anticipated that we will experience some changes in voting behavior. The actual influence at this time is unclear and will be studied for years to come.

9.2 Modeling Improvements

9.2.1 Ensemble Methods - Bagging, Boosting and Stacking

Machine Learning using Ensemble Methods, help improve results by combining several methods improving predictive performance compared to singular models. It should be noted, that the use of Ensemble Methods have placed first in many competitions such as: the Netflix Competition, KDD 2009, and Kaggle.

Bagging stands for bootstrap aggregation and is a way to decrease the variance in the prediction by generating additional data for training from datasets using combinations with repetitions to produce multi-sets of the original data.

Boosting is an iterative technique which adjusts the weight of an observation based on the last classification adding more weight to data which was misclassified by earlier evaluation rounds. It is used to convert weak learning algorithms into strong learning algorithms.

Stacking is a learning technique that combines multiple classification or regression models via a meta-classifier or meta-regressor where the meta-model is trained on the output of the base level model as features.

Since Ensemble Methods have had huge compeition success, it would be worth an exploration on how they perform for our election predictions.

9.2.2 Random Forest Tuning with Boruta

An interesting observation in comparing the final results between our tuned and non-tuned models was that the Random Forest model had the biggest gain in improvement. Despite having the 2nd best accuracy results, there may be an opportunity to achieve better results using Boruta Tuning Machine algorithms.

A Boruta algorithm is a wrapper built around the random forest classification algorithm implemented in the R package randomForest (Liaw and Wiener 2002). It uses a concept of of Feature Importance. Feature Importance is a class of techniques for assigning scores to the input features as a predictive model that indicates the relative importance of each feature when making a prediction.

Through this technique, Boruta modeling tries to capture all the important/interesting features you might have in the dataset with respect to an outcome variable.

With this tuning, it is anticipated that our accuruacy results would be further improved and may even become the best perfomring technique.

9.2.3 Neural Learning Model

The methods used in this project were, for the most part, based upon some type of classification problem. Another opportunity would be to build a Neural Network Learning model based on predicting voter's political preferences.

9.3 Code Optimization

Code opitimization opportunities include such areas as: Libriary and Package Optimization, R performance tuning, and general organizational changes.

9.4 Report Optimizations

Due to the idiosynchroncies of YAML, Rmarkdown, Kniter and LaTex generation, we have a wealth of conversion improvements rich within the area of output generation specific to cross-referencing, formatting, figure and table placement, color control and font formatting for emphasis.

Additionally, the table formating function could be updated to:

- format a particular cell versus the entire row
- evaluating other criteria when accuracy results are identical (i.e. multiple criteria for ordered ranking)

For those that would be interested in having more background on the models, mathimatical formulas could also be added.

9.5 Data Wrangling

Optimization for wrangling the requried datasets would also be recommended. Checks and clean up should be consolidated in the intial beginning in a format that can be used throughout the entire project. Currenty, the program has had to perform additional conditional checks and conversions for the modules to properly execute.

9.6 Results

While our project has produced results, it is unclear if these results are close to the antipiated results. On the surface, it appears they are lower than expected. If this is the case, guideance will be needed to understand where mistakes were made.

10 Appendixes

10.1 Appendix - A: Package Installations

The following packages were loaded for this project:

```
##  [1] "naivebayes"          "glmnet"              "Matrix"
##  [4] "Boruta"              "e1071"               "PerformanceAnalytics"
##  [7] "xts"                 "corrplot"            "class"
## [10] "ggthemes"            "MASS"                "rpart.plot"
## [13] "rpart"               "party"               "strucchange"
## [16] "sandwich"            "zoo"                 "modeltools"
## [19] "stats4"              "mvtnorm"             "grid"
## [22] "pROC"                "DescTools"           "tinytex"
## [25] "recosystem"          "lubridate"           "caret"
## [28] "lattice"             "kableExtra"          "scales"
## [31] "randomForest"        "boot"                "knitr"
## [34] "caTools"             "forcats"             "stringr"
## [37] "dplyr"               "purrr"               "readr"
## [40] "tidyr"               "tibble"              "tidyverse"
## [43] "data.table"          "pacman"              "ggplot2"
## [46] "stats"               "graphics"            "grDevices"
## [49] "utils"               "datasets"            "methods"
## [52] "base"
```

10.2 Appendix - B - Dataset Inspection

10.2.1 Dataset - Train

```
## Data Stucture: Original Dataset = Train Dataset

## 'data.frame':    5568 obs. of  108 variables:
##  $ USER_ID        : int  1 4 5 8 9 10 11 12 13 15 ...
##  $ YOB            : int  1938 1970 1997 1983 1984 1997 1983 1996 NA 1981 ...
##  $ Gender         : Factor w/ 2 levels "Female","Male": 2 1 2 2 1 1 2 2 2 1 ...
##  $ Income         : Factor w/ 6 levels "$100,001 - $150,000",..: NA 5 4 1 3 5 2 4 NA 3 ...
##  $ HouseholdStatus: Factor w/ 6 levels "Domestic Partners (no kids)",..: 4 2 5 4 4 5 3 5 5 4 ...
##  $ EducationLevel : Factor w/ 7 levels "Associate's Degree",..: NA 2 6 2 6 3 4 3 3 NA ...
##  $ Party          : Factor w/ 2 levels "Democrat","Republican": 1 1 2 1 2 1 1 2 2 2 ...
##  $ Q124742        : Factor w/ 2 levels "No","Yes": 1 NA NA 1 1 NA NA 2 1 1 ...
##  $ Q124122        : Factor w/ 2 levels "No","Yes": NA 2 2 2 2 NA NA 2 NA 1 ...
##  $ Q123464        : Factor w/ 2 levels "No","Yes": 1 1 2 1 1 NA NA 1 2 1 ...
##  $ Q123621        : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 NA NA 1 1 1 ...
##  $ Q122769        : Factor w/ 2 levels "No","Yes": 1 1 NA 1 1 1 NA 1 NA 2 ...
##  $ Q122770        : Factor w/ 2 levels "No","Yes": 2 2 2 1 2 2 NA 2 2 1 ...
##  $ Q122771        : Factor w/ 2 levels "Private","Public": 2 2 1 2 2 2 NA 1 2 2 ...
##  $ Q122120        : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 NA 2 2 1 ...
##  $ Q121699        : Factor w/ 2 levels "No","Yes": 2 2 1 2 2 2 1 NA 1 1 2 ...
##  $ Q121700        : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 NA 1 1 1 ...
##  $ Q120978        : Factor w/ 2 levels "No","Yes": NA 2 2 2 2 2 NA 1 2 2 ...
##  $ Q121011        : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 NA 1 1 2 ...
##  $ Q120379        : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 NA 2 2 1 ...
##  $ Q120650        : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 NA 2 NA 2 ...
##  $ Q120472        : Factor w/ 2 levels "Art","Science": NA 2 2 2 1 2 NA 2 2 1 ...
##  $ Q120194        : Factor w/ 2 levels "Study first",..: 2 1 1 2 2 2 NA 2 1 2 ...
##  $ Q120012        : Factor w/ 2 levels "No","Yes": 1 2 NA 1 2 2 NA 1 1 2 ...
##  $ Q120014        : Factor w/ 2 levels "No","Yes": 1 2 2 2 1 2 NA 2 1 1 ...
##  $ Q119334        : Factor w/ 2 levels "No","Yes": NA 1 1 2 1 1 1 1 NA 1 ...
##  $ Q119851        : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 2 NA 1 1 2 ...
##  $ Q119650        : Factor w/ 2 levels "Giving","Receiving": NA 2 2 1 1 2 NA 2 2 1 ...
##  $ Q118892        : Factor w/ 2 levels "No","Yes": 2 1 1 2 1 1 1 2 NA 1 ...
##  $ Q118117        : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 1 1 2 2 ...
##  $ Q118232        : Factor w/ 2 levels "Idealist","Pragmatist": 1 2 2 1 1 2 2 1 NA 1 ...
##  $ Q118233        : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 1 1 1 2 ...
##  $ Q118237        : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 1 2 2 2 2 ...
##  $ Q117186        : Factor w/ 2 levels "Cool headed",..: NA 1 1 1 2 NA 1 1 1 1 ...
##  $ Q117193        : Factor w/ 2 levels "Odd hours","Standard hours": NA 2 1 2 2 2 1 2 NA 2 ...
##  $ Q116797        : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 NA 1 1 1 2 ...
##  $ Q116881        : Factor w/ 2 levels "Happy","Right": 1 1 2 1 1 NA 1 2 1 1 ...
##  $ Q116953        : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 NA 1 1 2 1 ...
##  $ Q116601        : Factor w/ 2 levels "No","Yes": 2 2 1 2 2 2 NA 2 1 1 2 ...
##  $ Q116441        : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 NA 1 1 1 2 ...
##  $ Q116448        : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 NA 2 1 1 2 ...
##  $ Q116197        : Factor w/ 2 levels "A.M.","P.M.": 2 1 1 1 2 NA 2 2 2 2 ...
##  $ Q115602        : Factor w/ 2 levels "No","Yes": 2 1 2 2 1 NA 2 2 2 2 ...
##  $ Q115777        : Factor w/ 2 levels "End","Start": 2 1 2 2 1 NA 1 2 2 2 ...
##  $ Q115610        : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 1 NA 2 2 2 2 ...
##  $ Q115611        : Factor w/ 2 levels "No","Yes": 1 1 2 1 1 NA 1 2 1 2 ...
##  $ Q115899        : Factor w/ 2 levels "Circumstances",..: 1 2 1 1 2 NA 2 1 2 2 ...
```

```
##  $ Q115390       : Factor w/ 2 levels "No","Yes": 2 2 1 2 1 NA 2 1 2 2 ...
##  $ Q114961       : Factor w/ 2 levels "No","Yes": 2 2 2 1 2 NA 2 1 2 1 ...
##  $ Q114748       : Factor w/ 2 levels "No","Yes": 2 1 1 1 2 NA 1 2 2 1 ...
##  $ Q115195       : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 NA 2 1 NA 1 ...
##  $ Q114517       : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 NA NA 1 NA 1 ...
##  $ Q114386       : Factor w/ 2 levels "Mysterious","TMI": NA 1 1 2 2 NA NA 1 NA 2 ...
##  $ Q113992       : Factor w/ 2 levels "No","Yes": 2 1 1 1 2 NA NA 1 NA 1 ...
##  $ Q114152       : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 NA NA 1 NA 2 ...
##  $ Q113583       : Factor w/ 2 levels "Talk","Tunes": 1 NA 2 1 2 NA NA 2 2 2 ...
##  $ Q113584       : Factor w/ 2 levels "People","Technology": 2 NA 2 1 1 NA NA 1 2 2 ...
##  $ Q113181       : Factor w/ 2 levels "No","Yes": 1 NA 2 1 1 NA NA 2 NA 2 ...
##  $ Q112478       : Factor w/ 2 levels "No","Yes": 1 NA 2 2 1 NA NA 1 NA 1 ...
##  $ Q112512       : Factor w/ 2 levels "No","Yes": 2 NA 2 2 2 NA NA 2 2 2 ...
##  $ Q112270       : Factor w/ 2 levels "No","Yes": NA NA 2 2 1 NA NA 2 1 2 ...
##  $ Q111848       : Factor w/ 2 levels "No","Yes": 1 NA 1 2 1 2 NA 2 2 1 ...
##  $ Q111580       : Factor w/ 2 levels "Demanding","Supportive": 1 NA 2 2 1 2 2 1 1 1 ...
##  $ Q111220       : Factor w/ 2 levels "No","Yes": 1 NA 1 1 2 1 1 1 1 2 ...
##  $ Q110740       : Factor w/ 2 levels "Mac","PC": NA 1 2 1 2 2 2 2 2 2 ...
##  $ Q109367       : Factor w/ 2 levels "No","Yes": 1 2 1 2 2 NA 2 1 1 2 ...
##  $ Q108950       : Factor w/ 2 levels "Cautious","Risk-friendly": 1 1 1 2 1 NA 1 1 NA 1 ...
##  $ Q109244       : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 NA 2 1 1 1 ...
##  $ Q108855       : Factor w/ 2 levels "Umm...","Yes!": 2 1 1 1 2 NA 1 2 NA 2 ...
##  $ Q108617       : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 NA 1 1 1 1 ...
##  $ Q108856       : Factor w/ 2 levels "Socialize","Space": 2 2 2 1 1 NA NA 1 2 1 ...
##  $ Q108754       : Factor w/ 2 levels "No","Yes": 1 2 1 2 1 NA 1 1 1 1 ...
##  $ Q108342       : Factor w/ 2 levels "In-person","Online": 1 1 1 2 2 1 2 1 NA 2 ...
##  $ Q108343       : Factor w/ 2 levels "No","Yes": NA 1 1 1 1 1 2 1 NA 2 ...
##  $ Q107869       : Factor w/ 2 levels "No","Yes": 2 2 1 2 1 1 NA 1 NA 2 ...
##  $ Q107491       : Factor w/ 2 levels "No","Yes": 1 2 2 1 2 2 2 2 2 2 ...
##  $ Q106993       : Factor w/ 2 levels "No","Yes": 2 1 2 2 2 2 1 2 2 2 ...
##  $ Q106997       : Factor w/ 2 levels "Grrr people",..: 2 2 1 1 2 1 1 1 1 2 ...
##  $ Q106272       : Factor w/ 2 levels "No","Yes": 2 2 2 1 2 2 2 2 NA 1 ...
##  $ Q106388       : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 1 1 1 1 1 ...
##  $ Q106389       : Factor w/ 2 levels "No","Yes": 2 2 1 2 2 2 2 2 2 1 ...
##  $ Q106042       : Factor w/ 2 levels "No","Yes": 2 2 1 2 2 2 NA 1 1 2 ...
##  $ Q105840       : Factor w/ 2 levels "No","Yes": NA 2 1 1 2 1 2 1 1 2 ...
##  $ Q105655       : Factor w/ 2 levels "No","Yes": 1 1 1 2 2 1 1 2 2 2 ...
##  $ Q104996       : Factor w/ 2 levels "No","Yes": 2 2 1 2 1 2 1 1 1 2 ...
##  $ Q103293       : Factor w/ 2 levels "No","Yes": 1 NA 2 1 1 2 2 2 2 2 ...
##  $ Q102906       : Factor w/ 2 levels "No","Yes": 1 NA 1 1 1 2 2 1 1 2 ...
##  $ Q102674       : Factor w/ 2 levels "No","Yes": 1 NA 1 1 2 1 1 1 NA 1 ...
##  $ Q102687       : Factor w/ 2 levels "No","Yes": 2 NA 2 2 1 2 1 2 2 1 ...
##  $ Q102289       : Factor w/ 2 levels "No","Yes": 1 NA 1 2 1 NA NA 1 1 1 ...
##  $ Q102089       : Factor w/ 2 levels "Own","Rent": 1 NA 1 1 1 NA 1 1 NA 1 ...
##  $ Q101162       : Factor w/ 2 levels "Optimist","Pessimist": 1 NA 2 1 1 NA 2 2 1 1 ...
##  $ Q101163       : Factor w/ 2 levels "Dad","Mom": NA NA 2 2 2 NA NA 2 1 2 ...
##  $ Q101596       : Factor w/ 2 levels "No","Yes": 2 NA 1 1 1 NA NA 2 1 1 ...
##  $ Q100689       : Factor w/ 2 levels "No","Yes": 2 NA 1 1 2 NA 2 2 1 2 ...
##  $ Q100680       : Factor w/ 2 levels "No","Yes": 1 NA 1 1 2 NA 2 1 2 2 ...
##  $ Q100562       : Factor w/ 2 levels "No","Yes": 1 NA 1 2 2 NA 2 2 1 2 ...
##  $ Q99982        : Factor w/ 2 levels "Check!","Nope": 2 NA 2 1 2 NA NA 2 2 1 ...
##  $ Q100010       : Factor w/ 2 levels "No","Yes": 2 NA 2 1 2 NA 2 2 2 2 ...
##   [list output truncated]
```

```
##
##  Inspecting Initial Rows: Original Dataset = Train Dataset

##   USER_ID  YOB Gender              Income            HouseholdStatus
## 1       1 1938   Male                <NA>           Married (w/kids)
## 2       4 1970 Female       over $150,000 Domestic Partners (w/kids)
## 3       5 1997   Male  $75,000 - $100,000           Single (no kids)
## 4       8 1983   Male $100,001 - $150,000           Married (w/kids)
## 5       9 1984 Female   $50,000 - $74,999           Married (w/kids)
## 6      10 1997 Female       over $150,000           Single (no kids)
##        EducationLevel      Party Q124742 Q124122 Q123464 Q123621 Q122769
## 1                <NA>   Democrat      No    <NA>      No      No      No
## 2   Bachelor's Degree   Democrat    <NA>     Yes      No      No      No
## 3 High School Diploma Republican    <NA>     Yes     Yes      No    <NA>
## 4   Bachelor's Degree   Democrat      No     Yes      No     Yes      No
## 5 High School Diploma Republican      No     Yes      No      No      No
## 6        Current K-12   Democrat    <NA>    <NA>    <NA>    <NA>      No
##   Q122770 Q122771 Q122120 Q121699 Q121700 Q120978 Q121011 Q120379 Q120650
## 1     Yes  Public      No     Yes      No    <NA>      No      No     Yes
## 2     Yes  Public      No     Yes      No     Yes      No      No     Yes
## 3     Yes Private      No      No      No     Yes      No      No     Yes
## 4      No  Public      No     Yes      No     Yes      No      No     Yes
## 5     Yes  Public      No     Yes      No     Yes     Yes      No     Yes
## 6     Yes  Public      No      No      No     Yes      No     Yes     Yes
##   Q120472     Q120194 Q120012 Q120014 Q119334 Q119851   Q119650 Q118892 Q118117
## 1    <NA>   Try first      No      No    <NA>     Yes      <NA>     Yes     Yes
## 2 Science Study first     Yes     Yes      No      No Receiving      No      No
## 3 Science Study first    <NA>     Yes      No     Yes Receiving      No     Yes
## 4 Science   Try first      No     Yes     Yes      No    Giving     Yes      No
## 5     Art   Try first     Yes      No      No      No    Giving      No      No
## 6 Science   Try first     Yes     Yes      No     Yes Receiving      No      No
##      Q118232 Q118233 Q118237     Q117186        Q117193 Q116797 Q116881 Q116953
## 1   Idealist      No      No        <NA>           <NA>     Yes   Happy     Yes
## 2 Pragmatist      No      No Cool headed Standard hours      No   Happy     Yes
## 3 Pragmatist      No     Yes Cool headed      Odd hours      No   Right     Yes
## 4   Idealist      No      No Cool headed Standard hours      No   Happy     Yes
## 5   Idealist     Yes     Yes  Hot headed Standard hours      No   Happy     Yes
## 6 Pragmatist      No      No        <NA> Standard hours    <NA>    <NA>    <NA>
##   Q116601 Q116441 Q116448 Q116197 Q115602 Q115777 Q115610 Q115611       Q115899
## 1     Yes      No      No    P.M.     Yes   Start     Yes      No Circumstances
## 2     Yes     Yes      No    A.M.      No     End     Yes      No            Me
## 3      No      No     Yes    A.M.     Yes   Start     Yes     Yes Circumstances
## 4     Yes      No      No    A.M.     Yes   Start     Yes      No Circumstances
## 5     Yes      No     Yes    P.M.      No     End      No      No            Me
## 6    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>          <NA>
##   Q115390 Q114961 Q114748 Q115195 Q114517    Q114386 Q113992 Q114152 Q113583
## 1     Yes     Yes     Yes     Yes      No       <NA>     Yes     Yes    Talk
## 2     Yes     Yes      No     Yes      No Mysterious      No      No    <NA>
## 3      No     Yes      No     Yes     Yes Mysterious      No      No   Tunes
## 4     Yes      No      No     Yes      No        TMI      No      No    Talk
## 5      No     Yes     Yes     Yes     Yes        TMI     Yes      No   Tunes
## 6    <NA>    <NA>    <NA>    <NA>    <NA>       <NA>    <NA>    <NA>    <NA>
##      Q113584 Q113181 Q112478 Q112512 Q112270 Q111848    Q111580 Q111220 Q110740
## 1 Technology      No      No     Yes    <NA>      No  Demanding      No    <NA>
```

```
## 2       <NA>    <NA>    <NA>    <NA>    <NA>    <NA>       <NA>    <NA>     Mac
## 3 Technology     Yes     Yes     Yes     Yes      No Supportive      No      PC
## 4     People      No     Yes     Yes     Yes     Yes Supportive      No     Mac
## 5     People      No      No     Yes      No      No  Demanding     Yes      PC
## 6       <NA>    <NA>    <NA>    <NA>    <NA>     Yes Supportive      No      PC
##   Q109367       Q108950 Q109244 Q108855 Q108617   Q108856 Q108754   Q108342
## 1      No      Cautious      No    Yes!      No     Space      No In-person
## 2     Yes      Cautious      No  Umm...      No     Space     Yes In-person
## 3      No      Cautious      No  Umm...      No     Space      No In-person
## 4     Yes Risk-friendly      No  Umm...      No Socialize     Yes    Online
## 5     Yes      Cautious      No    Yes!      No Socialize      No    Online
## 6    <NA>          <NA>    <NA>    <NA>    <NA>      <NA>    <NA> In-person
##   Q108343 Q107869 Q107491 Q106993     Q106997 Q106272 Q106388 Q106389 Q106042
## 1    <NA>     Yes      No     Yes Yay people!     Yes      No     Yes     Yes
## 2      No     Yes     Yes      No Yay people!     Yes     Yes     Yes     Yes
## 3      No      No     Yes     Yes Grrr people     Yes      No      No      No
## 4      No     Yes      No     Yes Grrr people      No      No     Yes     Yes
## 5      No      No     Yes     Yes Yay people!     Yes      No     Yes     Yes
## 6      No      No     Yes     Yes Grrr people     Yes      No     Yes     Yes
##   Q105840 Q105655 Q104996 Q103293 Q102906 Q102674 Q102687 Q102289 Q102089
## 1    <NA>      No     Yes      No      No      No     Yes      No     Own
## 2     Yes      No     Yes    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>
## 3      No      No      No     Yes      No      No     Yes      No     Own
## 4      No     Yes     Yes      No      No      No     Yes     Yes     Own
## 5     Yes     Yes      No      No      No     Yes      No      No     Own
## 6      No      No     Yes     Yes     Yes      No     Yes    <NA>    <NA>
##     Q101162 Q101163 Q101596 Q100689 Q100680 Q100562 Q99982 Q100010 Q99716
## 1  Optimist    <NA>     Yes     Yes      No      No   Nope     Yes     No
## 2      <NA>    <NA>    <NA>    <NA>    <NA>    <NA>   <NA>    <NA>   <NA>
## 3 Pessimist     Mom      No      No      No      No   Nope     Yes     No
## 4  Optimist     Mom      No      No      No     Yes Check!      No     No
## 5  Optimist     Mom      No     Yes     Yes     Yes   Nope     Yes     No
## 6      <NA>    <NA>    <NA>    <NA>    <NA>    <NA>   <NA>    <NA>   <NA>
##   Q99581 Q99480 Q98869 Q98578     Q98059 Q98078 Q98197 Q96024
## 1     No   <NA>     No   <NA> Only-child     No     No    Yes
## 2   <NA>     No     No     No Only-child    Yes     No     No
## 3     No     No    Yes     No        Yes     No    Yes     No
## 4     No    Yes    Yes     No        Yes     No     No    Yes
## 5     No    Yes     No     No        Yes     No     No    Yes
## 6   <NA>   <NA>   <NA>   <NA>       <NA>   <NA>   <NA>   <NA>

##
##  Statistical Summary: Original Dataset = Train Dataset

##     USER_ID          YOB          Gender                      Income
##  Min.   :   1   Min.   :1880   Female:2130   $100,001 - $150,000: 768
##  1st Qu.:1732   1st Qu.:1970   Male  :3325   $25,001 - $50,000  : 708
##  Median :3460   Median :1983   NA's  : 113   $50,000 - $74,999  : 818
##  Mean   :3470   Mean   :1980                 $75,000 - $100,000 : 740
##  3rd Qu.:5210   3rd Qu.:1993                 over $150,000      : 738
##  Max.   :6960   Max.   :2039                 under $25,000      : 768
##                 NA's   :333                  NA's               :1028
##                     HouseholdStatus              EducationLevel
##  Domestic Partners (no kids): 180   Bachelor's Degree    :1206
```

```
##  Domestic Partners (w/kids) :  61   Current K-12         : 831
##  Married (no kids)          : 652   Current Undergraduate: 767
##  Married (w/kids)           :1594   High School Diploma  : 681
##  Single (no kids)           :2431   Master's Degree      : 639
##  Single (w/kids)            : 200   (Other)              : 578
##  NA's                       : 450   NA's                 : 866
##        Party      Q124742     Q124122     Q123464     Q123621     Q122769
##  Democrat  :2951   No  :1346   No  :1348   No  :3058   No  :1533   No  :2079
##  Republican:2617   Yes : 769   Yes :1749   Yes : 205   Yes :1636   Yes :1284
##                    NA's:3453   NA's:2471   NA's:2305   NA's:2399   NA's:2205
##
##
##
##
##  Q122770        Q122771     Q122120     Q121699     Q121700     Q120978
##  No  :1460   Private: 625   No  :2639   No  : 995   No  :3179   No  :1626
##  Yes :2045   Public :2896   Yes : 895   Yes :2749   Yes : 528   Yes :2093
##  NA's:2063   NA's   :2047   NA's:2034   NA's:1824   NA's:1861   NA's:1849
##
##
##
##
##  Q121011     Q120379     Q120650        Q120472            Q120194
##  No  :1672   No  :1965   No  : 320   Art    :1118   Study first:2034
##  Yes :2087   Yes :1725   Yes :3436   Science:2509   Try first  :1466
##  NA's:1809   NA's:1878   NA's:1812   NA's   :1941   NA's       :2068
##
##
##
##
##  Q120012     Q120014     Q119334     Q119851          Q119650     Q118892
##  No  :1953   No  :1382   No  :1816   No  :2214   Giving   :2777   No  :1446
##  Yes :1750   Yes :2139   Yes :1788   Yes :1575   Receiving: 885   Yes :2366
##  NA's:1865   NA's:2047   NA's:1964   NA's:1779   NA's     :1906   NA's:1756
##
##
##
##
##  Q118117           Q118232     Q118233     Q118237            Q117186
##  No  :2186   Idealist  :1416   No  :2517   No  :1886   Cool headed:2153
##  Yes :1520   Pragmatist:1755   Yes : 952   Yes :1630   Hot headed :1146
##  NA's:1862   NA's      :2397   NA's:2099   NA's:2052   NA's       :2269
##
##
##
##
##           Q117193     Q116797      Q116881     Q116953     Q116601
##  Odd hours     :1383   No  :2190   Happy:2285   No  :1092   No  : 577
##  Standard hours:1955   Yes :1167   Right: 975   Yes :2195   Yes :2907
##  NA's          :2230   NA's:2211   NA's :2308   NA's:2281   NA's:2084
##
##
##
##
```

```
##  Q116441     Q116448     Q116197     Q115602      Q115777     Q115610
##  No  :2180   No  :1863   A.M.:1155   No  : 726   End  :1403   No  : 615
##  Yes :1250   Yes :1542   P.M.:2285   Yes :2757   Start:1946   Yes :2854
##  NA's:2138   NA's:2163   NA's:2128   NA's:2085   NA's :2219   NA's:2099
##
##
##
##
##  Q115611              Q115899     Q115390     Q114961     Q114748
##  No  :2324   Circumstances:1480   No  :1310   No  :1721   No  :1556
##  Yes :1299   Me            :1860   Yes :1980   Yes :1695   Yes :2056
##  NA's:1945   NA's          :2228   NA's:2278   NA's:2152   NA's:1956
##
##
##
##
##  Q115195     Q114517           Q114386     Q113992     Q114152      Q113583
##  No  :1216   No  :2420   Mysterious:1955   No  :2507   No  :2307   Talk :1138
##  Yes :2245   Yes :1103   TMI       :1478   Yes :1083   Yes :1016   Tunes:2323
##  NA's:2107   NA's:2045   NA's      :2135   NA's:1978   NA's:2245   NA's :2107
##
##
##
##
##        Q113584     Q113181     Q112478     Q112512     Q112270     Q111848
##  People    :1705   No  :2046   No  :1309   No  : 643   No  :1809   No  :1419
##  Technology:1746   Yes :1463   Yes :2015   Yes :2783   Yes :1490   Yes :2186
##  NA's      :2117   NA's:2059   NA's:2244   NA's:2142   NA's:2269   NA's:1963
##
##
##
##
##        Q111580     Q111220     Q110740     Q109367              Q108950
##  Demanding :1190   No  :2580   Mac :1489   No  :1350   Cautious     :2338
##  Supportive:2216   Yes : 945   PC  :2090   Yes :2118   Risk-friendly:1112
##  NA's      :2162   NA's:2043   NA's:1989   NA's:2100   NA's         :2118
##
##
##
##
##  Q109244       Q108855     Q108617          Q108856     Q108754
##  No  :2459   Umm...:1277   No  :2994   Socialize: 924   No  :2243
##  Yes : 925   Yes!  :1887   Yes : 421   Space    :2232   Yes :1100
##  NA's:2184   NA's  :2404   NA's:2153   NA's     :2412   NA's:2225
##
##
##
##
##       Q108342     Q108343     Q107869     Q107491     Q106993
##  In-person:2301   No  :2049   No  :1566   No  : 430   No  : 606
##  Online   :1072   Yes :1337   Yes :1810   Yes :3014   Yes :2830
##  NA's     :2195   NA's:2182   NA's:2192   NA's:2124   NA's:2132
##
##
```

```
##
##
##         Q106997     Q106272     Q106388     Q106389     Q106042     Q105840
##  Grrr people:1824   No  : 973   No  :2391   No  :1714   No  :1768   No  :1794
##  Yay people!:1596   Yes :2424   Yes : 932   Yes :1555   Yes :1594   Yes :1486
##  NA's       :2148   NA's:2171   NA's:2245   NA's:2299   NA's:2206   NA's:2288
##
##
##
##
##  Q105655     Q104996     Q103293     Q102906     Q102674     Q102687
##  No  :1557   No  :1679   No  :1889   No  :2135   No  :2125   No  :1689
##  Yes :1929   Yes :1789   Yes :1559   Yes :1184   Yes :1162   Yes :1719
##  NA's:2082   NA's:2100   NA's:2120   NA's:2249   NA's:2281   NA's:2160
##
##
##
##
##  Q102289     Q102089          Q101162     Q101163     Q101596     Q100689
##  No  :2315   Own :2302   Optimist :2044   Dad :1796   No  :2160   No  :1413
##  Yes :1013   Rent:1086   Pessimist:1269   Mom :1388   Yes :1158   Yes :2108
##  NA's:2240   NA's:2180   NA's     :2255   NA's:2384   NA's:2250   NA's:2047
##
##
##
##
##  Q100680     Q100562        Q99982     Q100010      Q99716      Q99581
##  No  :1338   No  : 648   Check!:1709   No  : 696   No  :2959   No  :2959
##  Yes :2012   Yes :2701   Nope  :1570   Yes :2724   Yes : 382   Yes : 457
##  NA's:2218   NA's:2219   NA's  :2289   NA's:2148   NA's:2227   NA's:2152
##
##
##
##
##   Q99480      Q98869      Q98578            Q98059      Q98078      Q98197
##  No  : 750   No  : 730   No  :2095   Only-child: 331   No  :1853   No  :2002
##  Yes :2653   Yes :2513   Yes :1182   Yes       :3137   Yes :1368   Yes :1305
##  NA's:2165   NA's:2325   NA's:2291   NA's      :2100   NA's:2347   NA's:2261
##
##
##
##
##   Q96024
##  No  :1277
##  Yes :2001
##  NA's:2290
##
##
##
##
```

10.2.2 Dataset - Test

```
## Data Stucture: Original Dataset = Test Dataset

## 'data.frame':    1392 obs. of  107 variables:
##  $ USER_ID        : int  2 3 6 7 14 28 29 37 44 56 ...
##  $ YOB            : int  1985 1983 1995 1980 1980 1973 1968 1961 1989 1975 ...
##  $ Gender         : Factor w/ 2 levels "Female","Male": 1 2 2 1 1 2 1 2 1 2 ...
##  $ Income         : Factor w/ 6 levels "$100,001 - $150,000",..: 2 3 4 3 NA 5 3 5 6 4 ...
##  $ HouseholdStatus: Factor w/ 6 levels "Domestic Partners (no kids)",..: 5 4 5 5 3 3 5 1 5 4 ...
##  $ EducationLevel : Factor w/ 7 levels "Associate's Degree",..: 7 4 3 7 4 7 2 6 6 2 ...
##  $ Q124742        : Factor w/ 2 levels "No","Yes": NA NA NA 2 NA 1 NA NA NA 1 ...
##  $ Q124122        : Factor w/ 2 levels "No","Yes": 2 NA NA 2 2 2 NA 2 2 NA ...
##  $ Q123464        : Factor w/ 2 levels "No","Yes": 1 1 NA 1 1 1 NA 1 1 1 ...
##  $ Q123621        : Factor w/ 2 levels "No","Yes": 2 NA NA 2 2 2 NA 2 2 2 ...
##  $ Q122769        : Factor w/ 2 levels "No","Yes": 1 2 NA 2 1 1 NA NA 1 2 ...
##  $ Q122770        : Factor w/ 2 levels "No","Yes": 1 2 NA 2 1 1 NA NA 2 1 ...
##  $ Q122771        : Factor w/ 2 levels "Private","Public": 2 2 NA 2 2 2 NA NA 2 2 ...
##  $ Q122120        : Factor w/ 2 levels "No","Yes": 1 1 NA 1 2 1 NA NA 1 1 ...
##  $ Q121699        : Factor w/ 2 levels "No","Yes": 2 2 1 2 2 2 2 2 2 2 ...
##  $ Q121700        : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 1 1 1 ...
##  $ Q120978        : Factor w/ 2 levels "No","Yes": 2 NA 1 2 2 2 2 1 1 2 ...
##  $ Q121011        : Factor w/ 2 levels "No","Yes": 1 NA 2 1 2 2 2 1 2 2 ...
##  $ Q120379        : Factor w/ 2 levels "No","Yes": 2 NA 1 2 1 2 NA 1 2 1 ...
##  $ Q120650        : Factor w/ 2 levels "No","Yes": 2 NA 2 2 2 2 NA 2 1 2 ...
##  $ Q120472        : Factor w/ 2 levels "Art","Science": 2 NA 2 2 1 2 NA 2 1 1 ...
##  $ Q120194        : Factor w/ 2 levels "Study first",..: 1 1 2 2 2 2 NA NA 1 2 ...
##  $ Q120012        : Factor w/ 2 levels "No","Yes": 2 1 1 2 2 2 NA NA 2 2 ...
##  $ Q120014        : Factor w/ 2 levels "No","Yes": 2 2 2 1 2 2 NA NA 2 1 ...
##  $ Q119334        : Factor w/ 2 levels "No","Yes": 2 NA 1 1 2 1 NA NA NA 1 ...
##  $ Q119851        : Factor w/ 2 levels "No","Yes": 1 1 2 2 2 1 NA NA 2 2 ...
##  $ Q119650        : Factor w/ 2 levels "Giving","Receiving": 1 NA 1 1 1 1 NA NA 2 1 ...
##  $ Q118892        : Factor w/ 2 levels "No","Yes": 2 NA NA 2 1 1 NA NA 1 2 ...
##  $ Q118117        : Factor w/ 2 levels "No","Yes": 1 NA NA 2 1 2 NA 1 1 1 ...
##  $ Q118232        : Factor w/ 2 levels "Idealist","Pragmatist": 1 NA NA 1 1 2 NA NA NA 1 ...
##  $ Q118233        : Factor w/ 2 levels "No","Yes": 1 NA NA 1 1 2 NA 1 NA 1 ...
##  $ Q118237        : Factor w/ 2 levels "No","Yes": 2 NA NA 1 2 1 NA 1 1 2 ...
##  $ Q117186        : Factor w/ 2 levels "Cool headed",..: 1 NA NA 1 2 2 NA 1 NA NA ...
##  $ Q117193        : Factor w/ 2 levels "Odd hours","Standard hours": 1 NA NA 2 2 1 NA 1 2 2 ...
##  $ Q116797        : Factor w/ 2 levels "No","Yes": 2 NA NA 1 2 2 NA NA 1 1 ...
##  $ Q116881        : Factor w/ 2 levels "Happy","Right": 1 NA NA 1 1 2 NA 1 NA 1 ...
##  $ Q116953        : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 NA 1 1 1 ...
##  $ Q116601        : Factor w/ 2 levels "No","Yes": 2 2 NA 1 2 2 NA NA 2 2 ...
##  $ Q116441        : Factor w/ 2 levels "No","Yes": 1 NA NA 1 2 2 NA NA 1 2 ...
##  $ Q116448        : Factor w/ 2 levels "No","Yes": 2 NA NA 2 1 2 NA NA 1 2 ...
##  $ Q116197        : Factor w/ 2 levels "A.M.","P.M.": 1 2 NA 1 2 2 NA NA 1 2 ...
##  $ Q115602        : Factor w/ 2 levels "No","Yes": 2 NA NA 2 2 NA NA 2 2 1 ...
##  $ Q115777        : Factor w/ 2 levels "End","Start": 1 NA NA 2 1 1 2 NA 2 1 ...
##  $ Q115610        : Factor w/ 2 levels "No","Yes": 2 NA NA 2 1 2 NA 2 2 2 ...
##  $ Q115611        : Factor w/ 2 levels "No","Yes": 1 NA NA 1 1 2 NA 2 1 1 ...
##  $ Q115899        : Factor w/ 2 levels "Circumstances",..: 2 NA NA 2 2 1 NA NA 1 2 ...
##  $ Q115390        : Factor w/ 2 levels "No","Yes": 1 NA 2 2 1 1 2 1 1 2 ...
##  $ Q114961        : Factor w/ 2 levels "No","Yes": 2 NA 1 1 1 2 NA 2 2 2 ...
##  $ Q114748        : Factor w/ 2 levels "No","Yes": 1 NA 2 2 1 1 NA 2 2 1 ...
```

```
##  $ Q115195       : Factor w/ 2 levels "No","Yes": 2 1 2 2 2 2 2 2 2 1 ...
##  $ Q114517       : Factor w/ 2 levels "No","Yes": 2 NA 1 2 1 1 NA NA 1 1 ...
##  $ Q114386       : Factor w/ 2 levels "Mysterious","TMI": 2 NA 2 2 2 2 NA 2 2 2 ...
##  $ Q113992       : Factor w/ 2 levels "No","Yes": NA 2 1 1 1 2 2 2 1 1 ...
##  $ Q114152       : Factor w/ 2 levels "No","Yes": 1 1 1 2 NA 1 NA 1 2 1 ...
##  $ Q113583       : Factor w/ 2 levels "Talk","Tunes": 2 NA 2 1 2 1 1 NA 2 1 ...
##  $ Q113584       : Factor w/ 2 levels "People","Technology": 1 NA 2 1 2 2 2 NA 2 2 ...
##  $ Q113181       : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 NA 1 1 1 ...
##  $ Q112478       : Factor w/ 2 levels "No","Yes": 2 NA 1 1 2 2 NA 2 1 2 ...
##  $ Q112512       : Factor w/ 2 levels "No","Yes": 1 NA 2 2 2 2 NA NA 1 2 ...
##  $ Q112270       : Factor w/ 2 levels "No","Yes": 2 1 1 1 NA 1 NA NA 1 1 ...
##  $ Q111848       : Factor w/ 2 levels "No","Yes": 2 2 NA 2 2 2 NA NA 1 1 ...
##  $ Q111580       : Factor w/ 2 levels "Demanding","Supportive": 2 NA NA 2 2 1 NA NA 1 2 ...
##  $ Q111220       : Factor w/ 2 levels "No","Yes": 1 1 NA 1 2 1 2 1 1 1 ...
##  $ Q110740       : Factor w/ 2 levels "Mac","PC": NA NA NA 2 1 2 2 NA 2 2 ...
##  $ Q109367       : Factor w/ 2 levels "No","Yes": 2 2 1 1 2 2 NA 1 2 2 ...
##  $ Q108950       : Factor w/ 2 levels "Cautious","Risk-friendly": 1 1 NA 1 1 1 NA 1 1 2 ...
##  $ Q109244       : Factor w/ 2 levels "No","Yes": 2 1 1 2 1 1 NA 1 2 1 ...
##  $ Q108855       : Factor w/ 2 levels "Umm...","Yes!": 2 2 NA 2 2 1 NA NA 2 1 ...
##  $ Q108617       : Factor w/ 2 levels "No","Yes": NA 1 1 1 1 1 NA NA 1 1 ...
##  $ Q108856       : Factor w/ 2 levels "Socialize","Space": NA 2 NA 2 2 2 NA NA 1 2 ...
##  $ Q108754       : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 NA NA 1 1 ...
##  $ Q108342       : Factor w/ 2 levels "In-person","Online": 1 NA 1 2 1 1 1 NA 1 1 ...
##  $ Q108343       : Factor w/ 2 levels "No","Yes": 2 NA 1 1 1 2 2 NA 1 2 ...
##  $ Q107869       : Factor w/ 2 levels "No","Yes": NA 2 1 1 1 NA NA 2 2 2 ...
##  $ Q107491       : Factor w/ 2 levels "No","Yes": NA 2 2 2 2 2 2 2 NA 2 2 ...
##  $ Q106993       : Factor w/ 2 levels "No","Yes": NA 2 2 2 1 2 NA 2 2 2 ...
##  $ Q106997       : Factor w/ 2 levels "Grrr people",..: NA 1 2 2 1 1 NA 2 2 1 ...
##  $ Q106272       : Factor w/ 2 levels "No","Yes": NA 2 2 1 1 2 2 2 1 2 ...
##  $ Q106388       : Factor w/ 2 levels "No","Yes": NA 1 1 1 1 1 1 NA 1 1 ...
##  $ Q106389       : Factor w/ 2 levels "No","Yes": NA 1 2 1 1 2 NA NA 1 1 ...
##  $ Q106042       : Factor w/ 2 levels "No","Yes": NA 2 1 1 2 2 2 2 2 2 ...
##  $ Q105840       : Factor w/ 2 levels "No","Yes": NA 1 1 1 2 1 NA NA 1 1 ...
##  $ Q105655       : Factor w/ 2 levels "No","Yes": NA 2 2 2 2 1 NA NA 2 2 ...
##  $ Q104996       : Factor w/ 2 levels "No","Yes": NA 1 2 2 2 1 2 1 1 1 ...
##  $ Q103293       : Factor w/ 2 levels "No","Yes": NA 1 1 1 2 2 NA 1 1 1 ...
##  $ Q102906       : Factor w/ 2 levels "No","Yes": NA NA 1 1 1 2 2 NA 1 1 ...
##  $ Q102674       : Factor w/ 2 levels "No","Yes": NA NA 1 1 2 2 2 NA 1 2 ...
##  $ Q102687       : Factor w/ 2 levels "No","Yes": NA NA 1 1 1 2 2 NA 2 1 ...
##  $ Q102289       : Factor w/ 2 levels "No","Yes": NA NA 2 1 1 1 1 1 1 1 ...
##  $ Q102089       : Factor w/ 2 levels "Own","Rent": NA 2 1 1 1 1 NA 1 2 1 ...
##  $ Q101162       : Factor w/ 2 levels "Optimist","Pessimist": NA 2 1 1 2 2 NA 1 1 2 ...
##  $ Q101163       : Factor w/ 2 levels "Dad","Mom": NA 1 2 1 2 2 NA 1 2 2 ...
##  $ Q101596       : Factor w/ 2 levels "No","Yes": NA NA 1 1 1 1 2 NA 1 1 ...
##  $ Q100689       : Factor w/ 2 levels "No","Yes": 1 NA 1 1 2 2 2 1 1 2 ...
##  $ Q100680       : Factor w/ 2 levels "No","Yes": 2 NA 2 2 2 2 2 1 2 1 ...
##  $ Q100562       : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Q99982        : Factor w/ 2 levels "Check!","Nope": NA NA 2 2 2 2 2 NA 2 2 ...
##  $ Q100010       : Factor w/ 2 levels "No","Yes": NA NA 1 2 2 2 2 2 1 2 ...
##  $ Q99716        : Factor w/ 2 levels "No","Yes": NA NA 1 1 1 1 NA 1 2 1 ...
##   [list output truncated]

##
##  Inspecting Initial Rows: Original Dataset = Test Dataset
```

```
##   USER_ID  YOB Gender             Income   HouseholdStatus
## 1       2 1985 Female  $25,001 - $50,000  Single (no kids)
## 2       3 1983   Male  $50,000 - $74,999  Married (w/kids)
## 3       6 1995   Male $75,000 - $100,000  Single (no kids)
## 4       7 1980 Female  $50,000 - $74,999  Single (no kids)
## 5      14 1980 Female              <NA> Married (no kids)
## 6      28 1973   Male      over $150,000 Married (no kids)
##          EducationLevel Q124742 Q124122 Q123464 Q123621 Q122769 Q122770 Q122771
## 1       Master's Degree    <NA>     Yes      No     Yes      No      No  Public
## 2 Current Undergraduate    <NA>    <NA>      No    <NA>     Yes     Yes  Public
## 3          Current K-12    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>
## 4       Master's Degree     Yes     Yes      No     Yes     Yes     Yes  Public
## 5 Current Undergraduate    <NA>     Yes      No     Yes      No      No  Public
## 6       Master's Degree      No     Yes      No     Yes      No      No  Public
##   Q122120 Q121699 Q121700 Q120978 Q121011 Q120379 Q120650 Q120472     Q120194
## 1      No     Yes     Yes     Yes      No     Yes     Yes Science Study first
## 2      No     Yes      No    <NA>    <NA>    <NA>    <NA>    <NA> Study first
## 3    <NA>      No      No      No     Yes      No     Yes Science   Try first
## 4      No     Yes      No     Yes      No     Yes     Yes Science   Try first
## 5     Yes     Yes      No     Yes     Yes      No     Yes     Art   Try first
## 6      No     Yes      No     Yes     Yes     Yes     Yes Science   Try first
##   Q120012 Q120014 Q119334 Q119851 Q119650 Q118892 Q118117    Q118232 Q118233
## 1     Yes     Yes     Yes      No  Giving     Yes      No   Idealist      No
## 2      No     Yes    <NA>      No    <NA>    <NA>    <NA>       <NA>    <NA>
## 3      No     Yes      No     Yes  Giving    <NA>    <NA>       <NA>    <NA>
## 4     Yes      No      No     Yes  Giving     Yes     Yes   Idealist      No
## 5     Yes     Yes     Yes     Yes  Giving      No      No   Idealist      No
## 6     Yes     Yes      No      No  Giving      No     Yes Pragmatist     Yes
##   Q118237     Q117186        Q117193 Q116797 Q116881 Q116953 Q116601 Q116441
## 1     Yes Cool headed      Odd hours     Yes   Happy     Yes     Yes      No
## 2    <NA>        <NA>           <NA>    <NA>    <NA>     Yes     Yes    <NA>
## 3    <NA>        <NA>           <NA>    <NA>    <NA>     Yes    <NA>    <NA>
## 4      No Cool headed Standard hours      No   Happy     Yes      No      No
## 5     Yes  Hot headed Standard hours     Yes   Happy     Yes     Yes     Yes
## 6      No  Hot headed      Odd hours     Yes   Right     Yes     Yes     Yes
##   Q116448 Q116197 Q115602 Q115777 Q115610 Q115611       Q115899 Q115390 Q114961
## 1     Yes    A.M.     Yes     End     Yes      No            Me      No     Yes
## 2    <NA>    P.M.    <NA>    <NA>    <NA>    <NA>          <NA>    <NA>    <NA>
## 3    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>          <NA>     Yes      No
## 4     Yes    A.M.     Yes   Start     Yes      No            Me     Yes      No
## 5      No    P.M.     Yes     End      No      No            Me      No      No
## 6     Yes    P.M.    <NA>     End     Yes     Yes Circumstances      No     Yes
##   Q114748 Q115195 Q114517 Q114386 Q113992 Q114152 Q113583    Q113584 Q113181
## 1      No     Yes     Yes     TMI    <NA>      No   Tunes     People     Yes
## 2    <NA>      No    <NA>    <NA>     Yes      No    <NA>       <NA>      No
## 3     Yes     Yes      No     TMI      No      No   Tunes Technology     Yes
## 4     Yes     Yes     Yes     TMI      No     Yes    Talk     People      No
## 5      No     Yes      No     TMI      No    <NA>   Tunes Technology      No
## 6      No     Yes      No     TMI     Yes      No    Talk Technology      No
##   Q112478 Q112512 Q112270 Q111848    Q111580 Q111220 Q110740 Q109367  Q108950
## 1     Yes      No     Yes     Yes Supportive      No    <NA>     Yes Cautious
## 2    <NA>    <NA>      No     Yes       <NA>      No    <NA>     Yes Cautious
## 3      No     Yes      No    <NA>       <NA>    <NA>    <NA>      No     <NA>
## 4      No     Yes      No     Yes Supportive      No      PC      No Cautious
```

```
## 5     Yes     Yes    <NA>     Yes Supportive     Yes     Mac     Yes Cautious
## 6     Yes     Yes      No     Yes  Demanding      No      PC     Yes Cautious
##   Q109244 Q108855 Q108617 Q108856 Q108754   Q108342 Q108343 Q107869 Q107491
## 1     Yes    Yes!    <NA>    <NA>     Yes In-person     Yes    <NA>    <NA>
## 2      No    Yes!      No   Space      No      <NA>    <NA>     Yes     Yes
## 3      No    <NA>      No    <NA>     Yes In-person      No      No     Yes
## 4     Yes    Yes!      No   Space      No    Online      No      No     Yes
## 5      No    Yes!      No   Space      No In-person      No      No     Yes
## 6      No  Umm...      No   Space      No In-person     Yes    <NA>     Yes
##   Q106993     Q106997 Q106272 Q106388 Q106389 Q106042 Q105840 Q105655 Q104996
## 1    <NA>        <NA>    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>
## 2     Yes Grrr people     Yes      No      No     Yes      No     Yes      No
## 3     Yes Yay people!     Yes      No     Yes      No      No     Yes     Yes
## 4     Yes Yay people!      No      No      No      No      No     Yes     Yes
## 5      No Grrr people      No      No      No     Yes     Yes     Yes     Yes
## 6     Yes Grrr people     Yes      No     Yes     Yes      No      No      No
##   Q103293 Q102906 Q102674 Q102687 Q102289 Q102089   Q101162 Q101163 Q101596
## 1    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>      <NA>    <NA>    <NA>
## 2      No    <NA>    <NA>    <NA>    <NA>    Rent Pessimist     Dad    <NA>
## 3      No      No      No      No     Yes     Own  Optimist     Mom      No
## 4      No      No      No      No      No     Own  Optimist     Dad      No
## 5     Yes      No     Yes      No      No     Own Pessimist     Mom      No
## 6     Yes     Yes     Yes     Yes      No     Own Pessimist     Mom      No
##   Q100689 Q100680 Q100562 Q99982 Q100010 Q99716 Q99581 Q99480 Q98869 Q98578
## 1      No     Yes     Yes   <NA>    <NA>   <NA>   <NA>   <NA>    Yes   <NA>
## 2    <NA>    <NA>     Yes   <NA>    <NA>   <NA>   <NA>   <NA>    Yes   <NA>
## 3      No     Yes     Yes   Nope      No     No     No    Yes    Yes     No
## 4      No     Yes     Yes   Nope     Yes     No     No     No    Yes     No
## 5     Yes     Yes     Yes   Nope     Yes     No     No    Yes     No     No
## 6     Yes     Yes     Yes   Nope     Yes     No     No    Yes     No     No
##   Q98059 Q98078 Q98197 Q96024
## 1   <NA>   <NA>   <NA>   <NA>
## 2    Yes    Yes     No    Yes
## 3    Yes     No    Yes    Yes
## 4    Yes     No     No    Yes
## 5    Yes     No     No     No
## 6    Yes     No     No    Yes

##
##  Statistical Summary: Original Dataset = Test Dataset

##     USER_ID          YOB          Gender                    Income
##  Min.   :   2   Min.   :1900   Female:525   $100,001 - $150,000:185
##  1st Qu.:1774   1st Qu.:1970   Male  :837   $25,001 - $50,000  :195
##  Median :3540   Median :1984   NA's  : 30   $50,000 - $74,999  :201
##  Mean   :3524   Mean   :1980                $75,000 - $100,000 :201
##  3rd Qu.:5264   3rd Qu.:1993                over $150,000      :184
##  Max.   :6947   Max.   :2003                under $25,000      :181
##                 NA's   :82                  NA's               :245
##                     HouseholdStatus               EducationLevel Q124742
##  Domestic Partners (no kids): 37    Bachelor's Degree    :318    No  :338
##  Domestic Partners (w/kids) : 10    Current K-12         :212    Yes :167
##  Married (no kids)          :169    Current Undergraduate:209    NA's:887
##  Married (w/kids)           :371    Master's Degree      :156
```

```
##  Single (no kids)           :638    High School Diploma  :150
##  Single (w/kids)            : 65    (Other)              :146
##  NA's                       :102    NA's                 :201
##  Q124122    Q123464    Q123621    Q122769    Q122770       Q122771
##  No  :282   No  :738   No  :371   No  :509   No  :380   Private:127
##  Yes :467   Yes : 47   Yes :402   Yes :310   Yes :478   Public :733
##  NA's:643   NA's:607   NA's:619   NA's:573   NA's:534   NA's   :532
##
##
##
##
##  Q122120    Q121699    Q121700    Q120978    Q121011    Q120379    Q120650
##  No  :645   No  :225   No  :794   No  :416   No  :412   No  :480   No  : 76
##  Yes :229   Yes :712   Yes :131   Yes :522   Yes :533   Yes :429   Yes :845
##  NA's:518   NA's:455   NA's:467   NA's:454   NA's:447   NA's:483   NA's:471
##
##
##
##
##     Q120472           Q120194    Q120012    Q120014    Q119334    Q119851
##  Art    :287   Study first:495   No  :477   No  :351   No  :460   No  :537
##  Science:613   Try first  :362   Yes :436   Yes :517   Yes :419   Yes :391
##  NA's   :492   NA's       :535   NA's:479   NA's:524   NA's:513   NA's:464
##
##
##
##
##       Q119650    Q118892    Q118117          Q118232    Q118233    Q118237
##  Giving   :685   No  :344   No  :528   Idealist  :329   No  :590   No  :465
##  Receiving:239   Yes :598   Yes :384   Pragmatist:442   Yes :242   Yes :387
##  NA's     :468   NA's:450   NA's:480   NA's      :621   NA's:560   NA's:540
##
##
##
##
##         Q117186              Q117193    Q116797     Q116881    Q116953
##  Cool headed:513   Odd hours     :335   No  :555   Happy:577   No  :282
##  Hot headed :303   Standard hours:488   Yes :277   Right:234   Yes :543
##  NA's       :576   NA's          :569   NA's:560   NA's :581   NA's:567
##
##
##
##
##  Q116601    Q116441    Q116448    Q116197    Q115602     Q115777    Q115610
##  No  :150   No  :513   No  :426   A.M.:284   No  :199   End  :341   No  :163
##  Yes :720   Yes :333   Yes :399   P.M.:579   Yes :659   Start:485   Yes :691
##  NA's:522   NA's:546   NA's:567   NA's:529   NA's:534   NA's :566   NA's:538
##
##
##
##
##  Q115611             Q115899    Q115390    Q114961    Q114748    Q115195
##  No  :550   Circumstances:369   No  :321   No  :422   No  :383   No  :282
##  Yes :344   Me           :462   Yes :489   Yes :435   Yes :503   Yes :570
```

```
##  NA's:498   NA's         :561   NA's:582   NA's:535   NA's:506   NA's:540
##
##
##
##
##  Q114517          Q114386    Q113992    Q114152     Q113583          Q113584
##  No  :561   Mysterious:500   No  :593   No  :554   Talk :301   People    :413
##  Yes :309   TMI       :341   Yes :275   Yes :254   Tunes:566   Technology:442
##  NA's:522   NA's      :551   NA's:524   NA's:584   NA's :525   NA's      :537
##
##
##
##
##  Q113181    Q112478    Q112512    Q112270    Q111848          Q111580
##  No  :528   No  :319   No  :162   No  :484   No  :361   Demanding :305
##  Yes :347   Yes :527   Yes :696   Yes :357   Yes :545   Supportive:563
##  NA's:517   NA's:546   NA's:534   NA's:551   NA's:486   NA's      :524
##
##
##
##
##  Q111220    Q110740    Q109367             Q108950    Q109244      Q108855
##  No  :636   Mac :382   No  :312   Cautious     :579   No  :622   Umm...:318
##  Yes :236   PC  :520   Yes :556   Risk-friendly:290   Yes :223   Yes!  :470
##  NA's:520   NA's:490   NA's:524   NA's         :523   NA's:547   NA's  :604
##
##
##
##
##  Q108617          Q108856    Q108754         Q108342    Q108343    Q107869
##  No  :733   Socialize:226   No  :552   In-person:591   No  :498   No  :352
##  Yes :116   Space    :571   Yes :295   Online   :236   Yes :340   Yes :470
##  NA's:543   NA's     :595   NA's:545   NA's     :565   NA's:554   NA's:570
##
##
##
##
##  Q107491    Q106993           Q106997    Q106272    Q106388    Q106389
##  No  :109   No  :138   Grrr people:468   No  :257   No  :603   No  :421
##  Yes :740   Yes :710   Yay people!:370   Yes :584   Yes :216   Yes :399
##  NA's:543   NA's:544   NA's       :554   NA's:551   NA's:573   NA's:572
##
##
##
##
##  Q106042    Q105840    Q105655    Q104996    Q103293    Q102906    Q102674
##  No  :433   No  :457   No  :368   No  :431   No  :435   No  :504   No  :509
##  Yes :403   Yes :347   Yes :494   Yes :441   Yes :403   Yes :297   Yes :300
##  NA's:556   NA's:588   NA's:530   NA's:520   NA's:554   NA's:591   NA's:583
##
##
##
##
##  Q102687    Q102289    Q102089         Q101162    Q101163    Q101596
```

```
##  No  :402   No  :604   Own :593   Optimist :523   Dad :453   No  :547
##  Yes :438   Yes :238   Rent:243   Pessimist:308   Mom :328   Yes :271
##  NA's:552   NA's:550   NA's:556   NA's     :561   NA's:611   NA's:574
##
##
##
##
##  Q100689    Q100680    Q100562       Q99982    Q100010     Q99716     Q99581
##  No  :351   No  :337   No  :152   Check!:423   No  :155   No  :713   No  :754
##  Yes :520   Yes :486   Yes :666   Nope  :387   Yes :697   Yes :116   Yes :100
##  NA's:521   NA's:569   NA's:574   NA's  :582   NA's:540   NA's:563   NA's:538
##
##
##
##
##   Q99480     Q98869     Q98578           Q98059     Q98078     Q98197
##  No  :179   No  :195   No  :498   Only-child: 89   No  :461   No  :495
##  Yes :678   Yes :616   Yes :318   Yes       :774   Yes :333   Yes :322
##  NA's:535   NA's:581   NA's:576   NA's      :529   NA's:598   NA's:575
##
##
##
##
##   Q96024
##  No  :300
##  Yes :524
##  NA's:568
##
##
##
##
```

10.2.3 Dataset - Train Subset

```
## Data Structure: Subset Dataset = Train Subset

## 'data.frame':    5568 obs. of  7 variables:
##  $ USER_ID        : int  1 4 5 8 9 10 11 12 13 15 ...
##  $ YOB            : int  1938 1970 1997 1983 1984 1997 1983 1996 NA 1981 ...
##  $ Gender         : Factor w/ 2 levels "Female","Male": 2 1 2 2 1 1 2 2 2 1 ...
##  $ Income         : Factor w/ 6 levels "$100,001 - $150,000",..: NA 5 4 1 3 5 2 4 NA 3 ...
##  $ HouseholdStatus: Factor w/ 6 levels "Domestic Partners (no kids)",..: 4 2 5 4 4 5 3 5 5 4 ...
##  $ EducationLevel : Factor w/ 7 levels "Associate's Degree",..: NA 2 6 2 6 3 4 3 3 NA ...
##  $ Party          : Factor w/ 2 levels "Democrat","Republican": 1 1 2 1 2 1 1 2 2 2 ...

##
##  Inspecting Initial Rows: Subset Dataset = Train Subset

##   USER_ID  YOB Gender              Income            HouseholdStatus
## 1       1 1938   Male                <NA>           Married (w/kids)
## 2       4 1970 Female       over $150,000 Domestic Partners (w/kids)
## 3       5 1997   Male  $75,000 - $100,000           Single (no kids)
## 4       8 1983   Male $100,001 - $150,000           Married (w/kids)
## 5       9 1984 Female   $50,000 - $74,999           Married (w/kids)
## 6      10 1997 Female       over $150,000           Single (no kids)
##        EducationLevel      Party
## 1                <NA>   Democrat
## 2   Bachelor's Degree   Democrat
## 3 High School Diploma Republican
## 4   Bachelor's Degree   Democrat
## 5 High School Diploma Republican
## 6        Current K-12   Democrat

##
##  Statistical Summary: Subset Dataset = Train Subset

##     USER_ID           YOB          Gender                      Income
##  Min.   :   1   Min.   :1880   Female:2130   $100,001 - $150,000: 768
##  1st Qu.:1732   1st Qu.:1970   Male  :3325   $25,001 - $50,000  : 708
##  Median :3460   Median :1983   NA's  : 113   $50,000 - $74,999  : 818
##  Mean   :3470   Mean   :1980                 $75,000 - $100,000 : 740
##  3rd Qu.:5210   3rd Qu.:1993                 over $150,000      : 738
##  Max.   :6960   Max.   :2039                 under $25,000      : 768
##                 NA's   :333                  NA's               :1028
##                     HouseholdStatus              EducationLevel
##  Domestic Partners (no kids): 180   Bachelor's Degree    :1206
##  Domestic Partners (w/kids) :  61   Current K-12         : 831
##  Married (no kids)          : 652   Current Undergraduate: 767
##  Married (w/kids)           :1594   High School Diploma  : 681
##  Single (no kids)           :2431   Master's Degree      : 639
##  Single (w/kids)            : 200   (Other)              : 578
##  NA's                       : 450   NA's                 : 866
##         Party
##  Democrat  :2951
##  Republican:2617
```

```
##
##
##
##
##
```

10.2.4 Dataset - Test Subset

```
## Data Structure: Subset Dataset = Test Subset

## 'data.frame':    1392 obs. of  6 variables:
##  $ USER_ID        : int  2 3 6 7 14 28 29 37 44 56 ...
##  $ YOB            : int  1985 1983 1995 1980 1980 1973 1968 1961 1989 1975 ...
##  $ Gender         : Factor w/ 2 levels "Female","Male": 1 2 2 1 1 2 1 2 1 2 ...
##  $ Income         : Factor w/ 6 levels "$100,001 - $150,000",..: 2 3 4 3 NA 5 3 5 6 4 ...
##  $ HouseholdStatus: Factor w/ 6 levels "Domestic Partners (no kids)",..: 5 4 5 5 3 3 5 1 5 4 ...
##  $ EducationLevel : Factor w/ 7 levels "Associate's Degree",..: 7 4 3 7 4 7 2 6 6 2 ...

##
##  Inspecting Initial Rows: Subset Dataset = Test Subset

##   USER_ID  YOB Gender             Income   HouseholdStatus
## 1       2 1985 Female  $25,001 - $50,000  Single (no kids)
## 2       3 1983   Male  $50,000 - $74,999  Married (w/kids)
## 3       6 1995   Male $75,000 - $100,000  Single (no kids)
## 4       7 1980 Female  $50,000 - $74,999  Single (no kids)
## 5      14 1980 Female              <NA> Married (no kids)
## 6      28 1973   Male      over $150,000 Married (no kids)
##          EducationLevel
## 1       Master's Degree
## 2 Current Undergraduate
## 3          Current K-12
## 4       Master's Degree
## 5 Current Undergraduate
## 6       Master's Degree

##
##  Statistical Summary: Subset Dataset = Test Subset

##     USER_ID          YOB          Gender                     Income
##  Min.   :   2   Min.   :1900   Female:525   $100,001 - $150,000:185
##  1st Qu.:1774   1st Qu.:1970   Male  :837   $25,001 - $50,000  :195
##  Median :3540   Median :1984   NA's  : 30   $50,000 - $74,999  :201
##  Mean   :3524   Mean   :1980                $75,000 - $100,000 :201
##  3rd Qu.:5264   3rd Qu.:1993                over $150,000      :184
##  Max.   :6947   Max.   :2003                under $25,000      :181
##                 NA's   :82                  NA's               :245
##                    HouseholdStatus              EducationLevel
##  Domestic Partners (no kids): 37    Bachelor's Degree    :318
##  Domestic Partners (w/kids) : 10    Current K-12         :212
##  Married (no kids)          :169    Current Undergraduate:209
##  Married (w/kids)           :371    Master's Degree      :156
##  Single (no kids)           :638    High School Diploma  :150
##  Single (w/kids)            : 65    (Other)              :146
##  NA's                       :102    NA's                 :201
```

10.2.5 Dataset - Train Sample (70% of Training Dataset)

```
## Data Structure: Train Sample 70% Dataset = Train Sample

## 'data.frame':    3898 obs. of  7 variables:
##  $ USER_ID        : int  1 4 5 8 11 12 13 16 17 19 ...
##  $ YOB            : int  1938 1970 1997 1983 1983 1996 NA 1971 1977 1996 ...
##  $ Gender         : Factor w/ 2 levels "Female","Male": 2 1 2 2 2 2 2 2 1 2 ...
##  $ Income         : Factor w/ 6 levels "$100,001 - $150,000",..: NA 5 4 1 2 4 NA 5 1 4 ...
##  $ HouseholdStatus: Factor w/ 6 levels "Domestic Partners (no kids)",..: 4 2 5 4 3 5 5 3 1 5 ...
##  $ EducationLevel : Factor w/ 7 levels "Associate's Degree",..: NA 2 6 2 4 3 3 2 2 3 ...
##  $ Party          : Factor w/ 2 levels "Democrat","Republican": 1 1 2 1 1 2 2 2 2 2 ...

##
##  Inspection Initial Rows: Train Sample 70% Dataset = Train Sample

##   USER_ID  YOB Gender              Income            HouseholdStatus
## 1       1 1938   Male                <NA>           Married (w/kids)
## 2       4 1970 Female       over $150,000 Domestic Partners (w/kids)
## 3       5 1997   Male  $75,000 - $100,000            Single (no kids)
## 4       8 1983   Male $100,001 - $150,000           Married (w/kids)
## 7      11 1983   Male   $25,001 - $50,000          Married (no kids)
## 8      12 1996   Male  $75,000 - $100,000            Single (no kids)
##          EducationLevel      Party
## 1                  <NA>   Democrat
## 2     Bachelor's Degree   Democrat
## 3   High School Diploma Republican
## 4     Bachelor's Degree   Democrat
## 7 Current Undergraduate   Democrat
## 8          Current K-12 Republican

##
##  Statistical Summary: Train Sample 70% Dataset = Train Sample

##     USER_ID          YOB          Gender                      Income
##  Min.   :   1   Min.   :1881   Female:1491   $100,001 - $150,000:551
##  1st Qu.:1736   1st Qu.:1970   Male  :2324   $25,001 - $50,000  :491
##  Median :3465   Median :1983   NA's  :  83   $50,000 - $74,999  :581
##  Mean   :3478   Mean   :1980                 $75,000 - $100,000 :517
##  3rd Qu.:5218   3rd Qu.:1993                 over $150,000      :508
##  Max.   :6960   Max.   :2039                 under $25,000      :507
##                 NA's   :236                  NA's               :743
##                     HouseholdStatus              EducationLevel
##  Domestic Partners (no kids): 118   Bachelor's Degree    :850
##  Domestic Partners (w/kids) :  47   Current K-12         :585
##  Married (no kids)          : 465   Current Undergraduate:532
##  Married (w/kids)           :1130   High School Diploma  :472
##  Single (no kids)           :1672   Master's Degree      :434
##  Single (w/kids)            : 149   (Other)              :407
##  NA's                       : 317   NA's                 :618
##         Party
##  Democrat  :2066
##  Republican:1832
```

```
##
##
##
##
##
```

10.2.6 Dataset - Test Sample (30% of Test Dataset)

```
## Data Structure: Test Sample 30% Dataset = Test Sample

## 'data.frame':    1670 obs. of  7 variables:
##  $ USER_ID        : int  9 10 15 18 22 34 36 38 40 42 ...
##  $ YOB            : int  1984 1997 1981 1971 1997 1950 1977 1979 1979 1950 ...
##  $ Gender         : Factor w/ 2 levels "Female","Male": 1 1 1 2 1 1 1 1 1 2 ...
##  $ Income         : Factor w/ 6 levels "$100,001 - $150,000",..: 3 5 3 3 4 NA 2 4 2 2 ...
##  $ HouseholdStatus: Factor w/ 6 levels "Domestic Partners (no kids)",..: 4 5 4 4 5 3 5 5 4 4 ...
##  $ EducationLevel : Factor w/ 7 levels "Associate's Degree",..: 6 3 NA 6 NA 1 2 6 2 2 ...
##  $ Party          : Factor w/ 2 levels "Democrat","Republican": 2 1 2 1 1 1 1 1 1 1 ...

##
##  Inspection Initial Rows: Test Sample 30% Dataset = Test Sample

##    USER_ID  YOB Gender             Income   HouseholdStatus      EducationLevel
## 5        9 1984 Female  $50,000 - $74,999  Married (w/kids) High School Diploma
## 6       10 1997 Female      over $150,000  Single (no kids)        Current K-12
## 10      15 1981 Female  $50,000 - $74,999  Married (w/kids)                <NA>
## 13      18 1971   Male  $50,000 - $74,999  Married (w/kids) High School Diploma
## 17      22 1997 Female $75,000 - $100,000  Single (no kids)                <NA>
## 27      34 1950 Female               <NA> Married (no kids)  Associate's Degree
##         Party
## 5  Republican
## 6    Democrat
## 10 Republican
## 13   Democrat
## 17   Democrat
## 27   Democrat

##
##  Statistical Summary: Test Sample 30% Dataset = Test Sample

##     USER_ID          YOB          Gender                      Income
##  Min.   :   9   Min.   :1880   Female: 639   $100,001 - $150,000:217
##  1st Qu.:1730   1st Qu.:1971   Male  :1001   $25,001 - $50,000  :217
##  Median :3448   Median :1983   NA's  :  30   $50,000 - $74,999  :237
##  Mean   :3451   Mean   :1980                 $75,000 - $100,000 :223
##  3rd Qu.:5201   3rd Qu.:1993                 over $150,000      :230
##  Max.   :6959   Max.   :2039                 under $25,000      :261
##                 NA's   :97                   NA's               :285
##                    HouseholdStatus              EducationLevel
##  Domestic Partners (no kids): 62    Bachelor's Degree    :356
##  Domestic Partners (w/kids) : 14    Current K-12         :246
##  Married (no kids)          :187    Current Undergraduate:235
##  Married (w/kids)           :464    High School Diploma  :209
##  Single (no kids)           :759    Master's Degree      :205
##  Single (w/kids)            : 51    (Other)              :171
##  NA's                       :133    NA's                 :248
##        Party
##  Democrat  :885
##  Republican:785
```

```
##
##
##
##
##
```

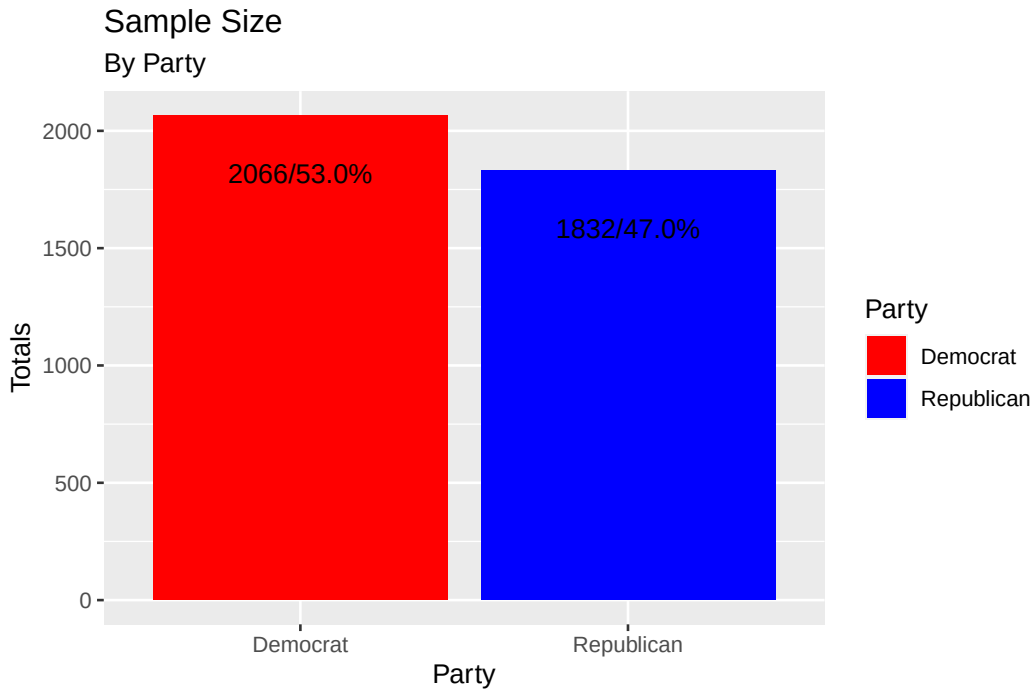
10.2.7 Dataset - Train Sample (70% of Training Dataset excluding NA Fields) with Key Questions

```
## Data Structure: Train Sample Dataset excluding NA Fields = Test Sample No NA

## 'data.frame':    2618 obs. of  7 variables:
##  $ USER_ID        : int  4 5 8 11 12 16 17 19 20 21 ...
##  $ YOB            : int  1970 1997 1983 1983 1996 1971 1977 1996 1970 1979 ...
##  $ Gender         : Factor w/ 2 levels "Female","Male": 1 2 2 2 2 2 1 2 2 2 ...
##  $ Income         : Factor w/ 6 levels "$100,001 - $150,000",..: 5 4 1 2 4 5 1 4 4 5 ...
##  $ HouseholdStatus: Factor w/ 6 levels "Domestic Partners (no kids)",..: 2 5 4 3 5 3 1 5 4 4 ...
##  $ EducationLevel : Factor w/ 7 levels "Associate's Degree",..: 2 6 2 4 3 2 2 3 2 2 ...
##  $ Party          : Factor w/ 2 levels "Democrat","Republican": 1 2 1 1 2 2 2 2 2 2 ...
##  - attr(*, "na.action")= 'omit' Named int  1 7 19 24 29 32 33 39 44 47 ...
##   ..- attr(*, "names")= chr  "1" "9" "24" "33" ...

##
##  Inspection Initial Rows: Train Sample Dataset excluding NA Fields - Test Sample No NA

##    USER_ID  YOB Gender              Income            HouseholdStatus
## 2        4 1970 Female       over $150,000 Domestic Partners (w/kids)
## 3        5 1997   Male  $75,000 - $100,000           Single (no kids)
## 4        8 1983   Male $100,001 - $150,000           Married (w/kids)
## 7       11 1983   Male   $25,001 - $50,000          Married (no kids)
## 8       12 1996   Male  $75,000 - $100,000           Single (no kids)
## 11      16 1971   Male       over $150,000          Married (no kids)
##           EducationLevel      Party
## 2      Bachelor's Degree   Democrat
## 3    High School Diploma Republican
## 4      Bachelor's Degree   Democrat
## 7  Current Undergraduate   Democrat
## 8            Current K-12 Republican
## 11     Bachelor's Degree Republican

##
##  Statistial Summary: Train Sample Dataset excluding NA Fields - Test Sample No NA

##     USER_ID          YOB          Gender                      Income
##  Min.   :   4   Min.   :1881   Female: 986   $100,001 - $150,000:451
##  1st Qu.:1485   1st Qu.:1970   Male  :1632   $25,001 - $50,000  :421
##  Median :3206   Median :1982                 $50,000 - $74,999  :489
##  Mean   :3244   Mean   :1980                 $75,000 - $100,000 :413
##  3rd Qu.:4882   3rd Qu.:1992                 over $150,000      :400
##  Max.   :6960   Max.   :2013                 under $25,000      :444
##
##                     HouseholdStatus              EducationLevel
##  Domestic Partners (no kids):  91   Associate's Degree   :236
##  Domestic Partners (w/kids) :  36   Bachelor's Degree    :733
##  Married (no kids)          : 341   Current K-12         :346
##  Married (w/kids)           : 828   Current Undergraduate:415
##  Single (no kids)           :1205   Doctoral Degree      :113
##  Single (w/kids)            : 117   High School Diploma  :395
##                                     Master's Degree      :380
```

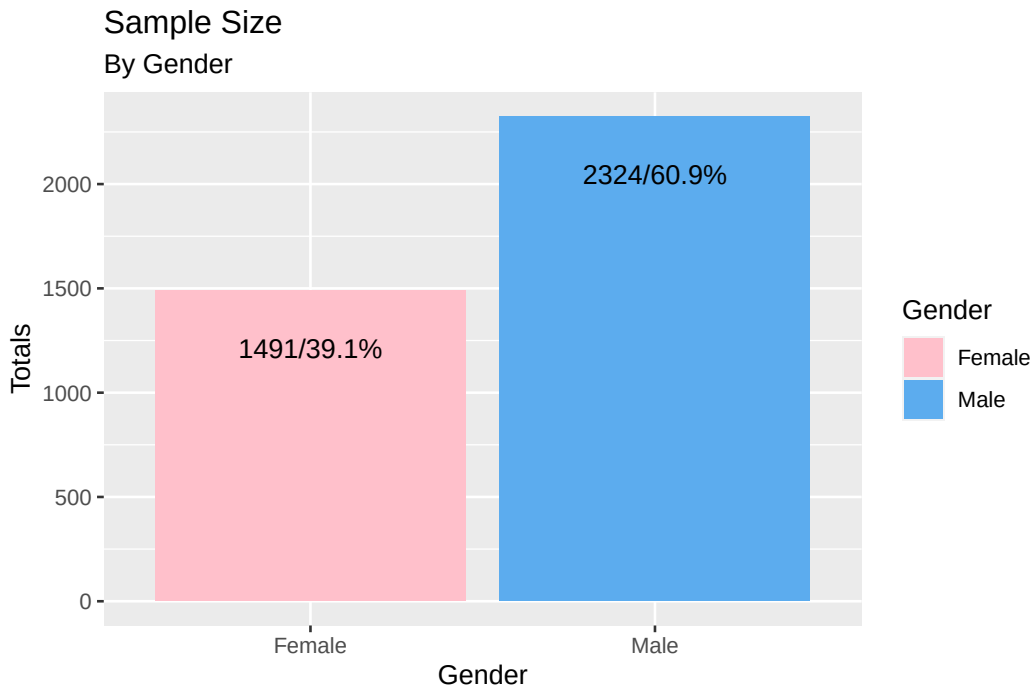
```
##         Party
##  Democrat  :1402
##  Republican:1216
##
##
##
##
##
```

10.3 Appendix - C: Demographic Figures

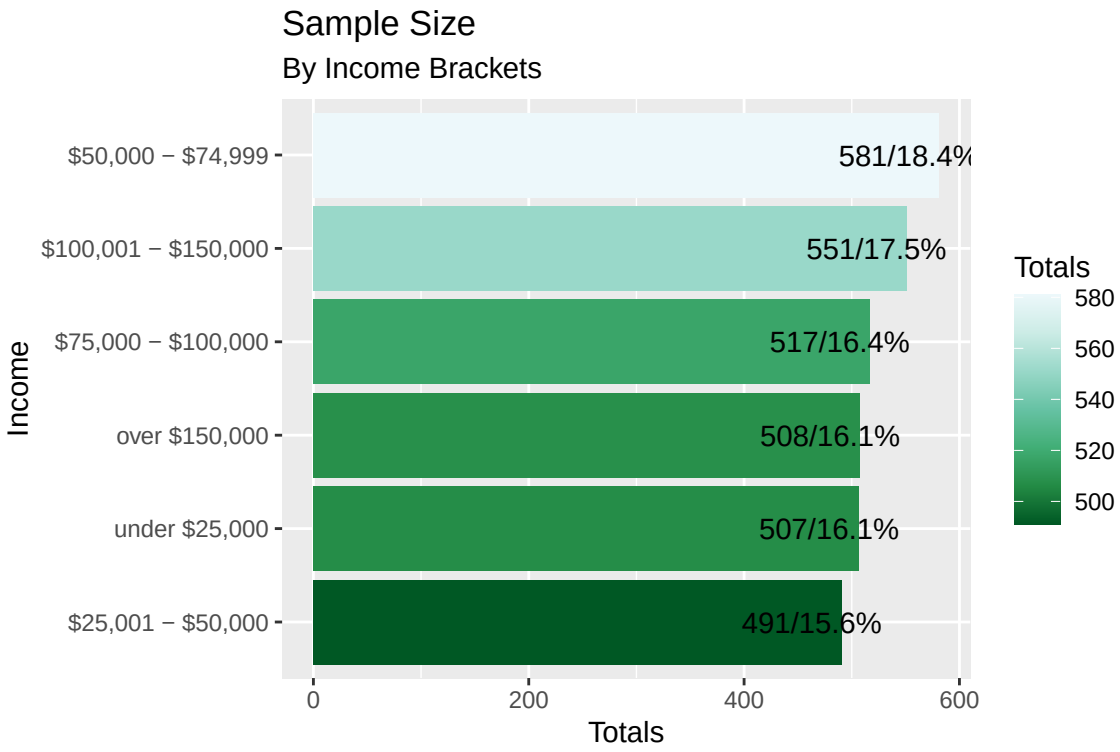
10.3.1 Plotting Dataset by Voting Party



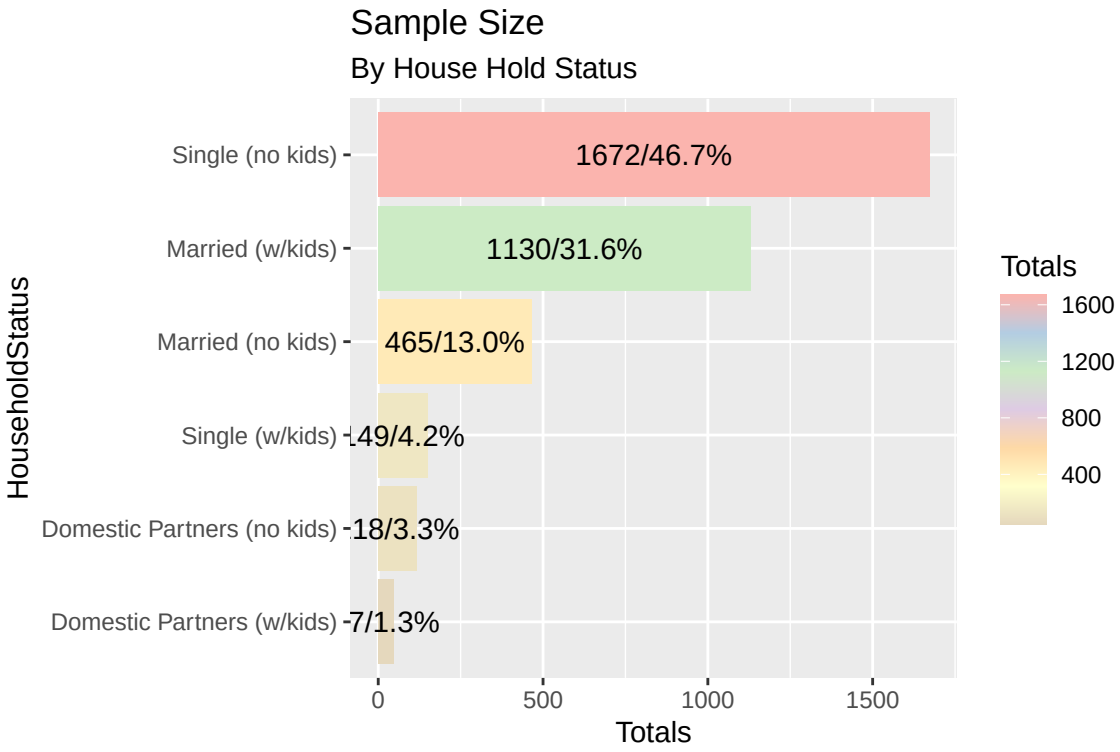
10.3.2 Plotting Dataset by Gender



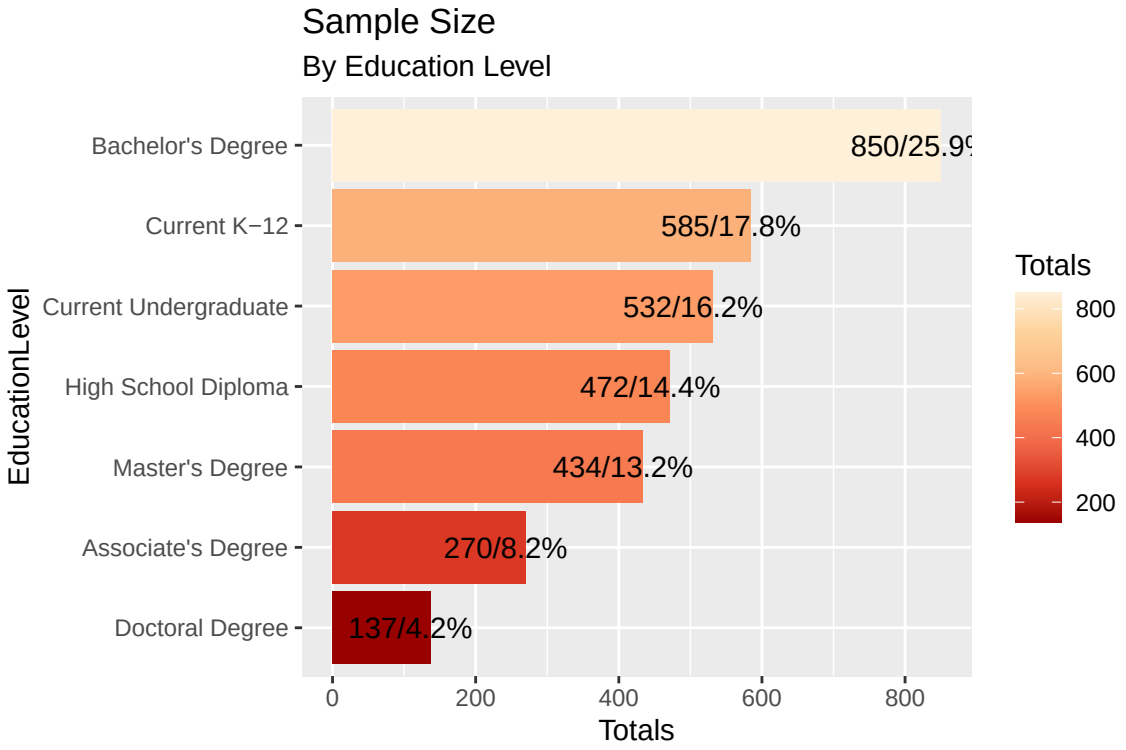
10.3.3 Plotting Dataset by Income Bands



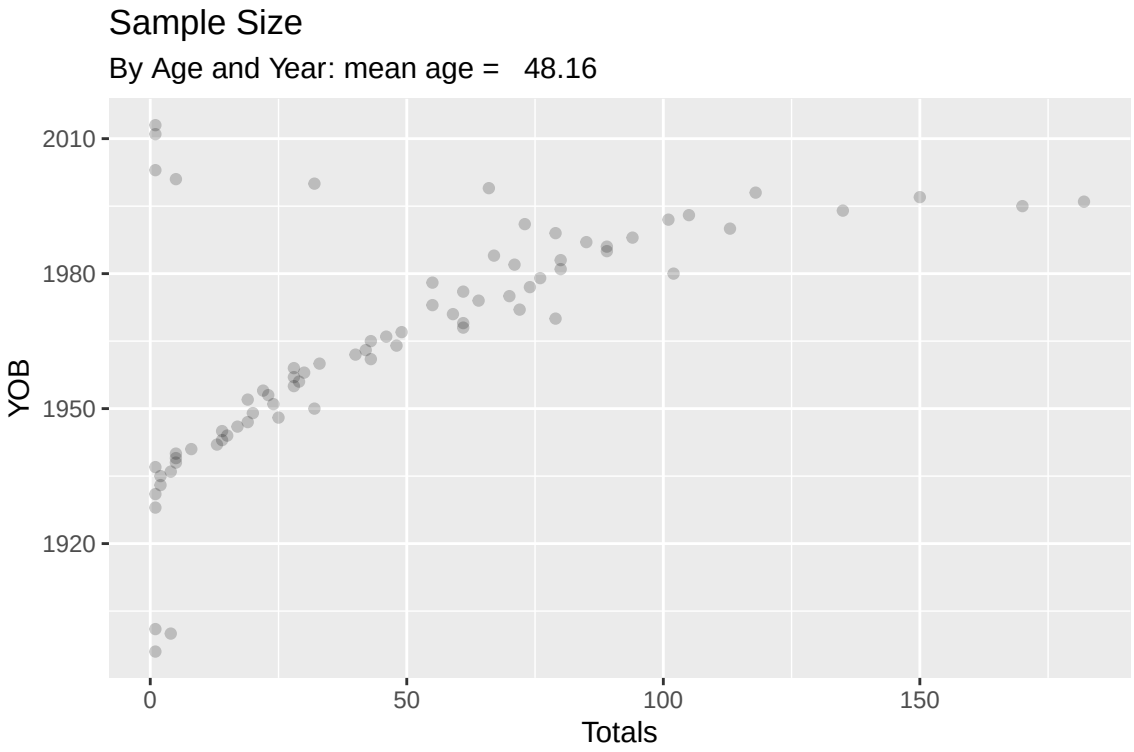
10.3.4 Plotting Dataset by Household Status



10.3.5 Plotting Dataset by Education Levels

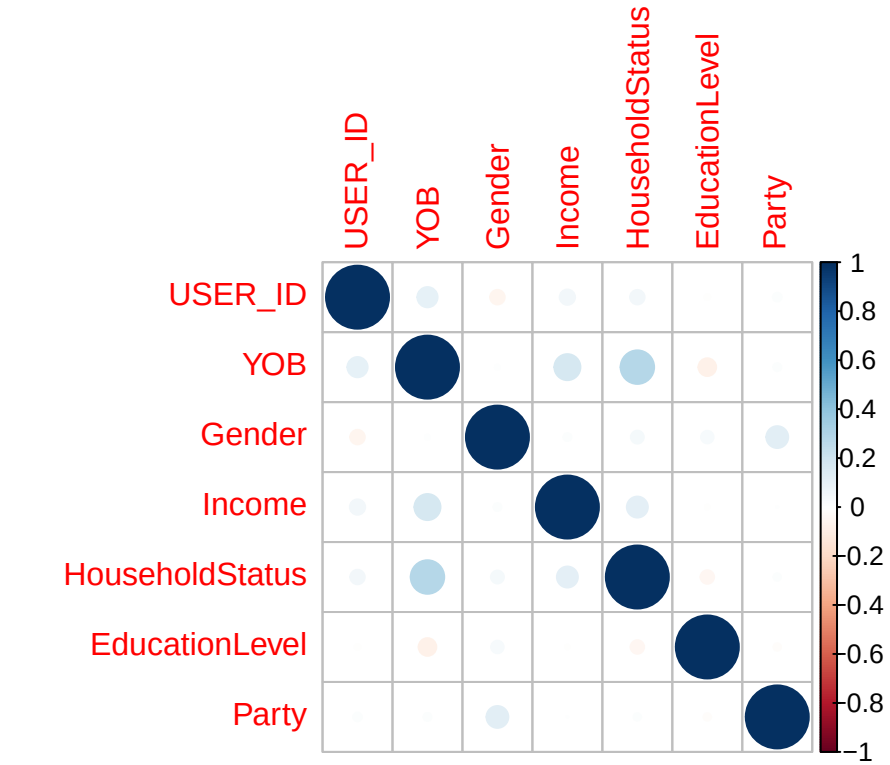


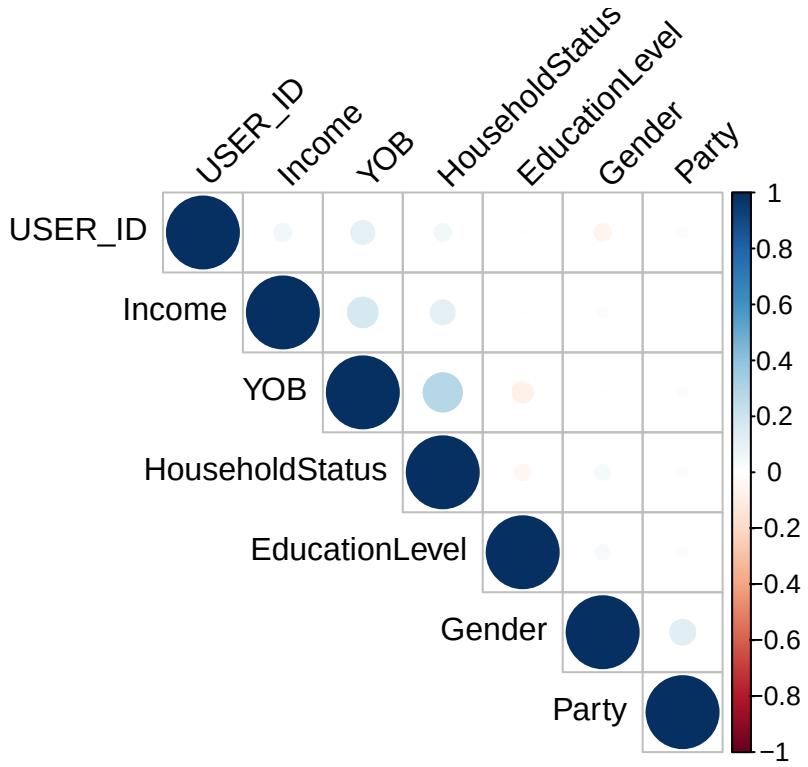
10.3.6 Plotting Dataset by Age Distribution



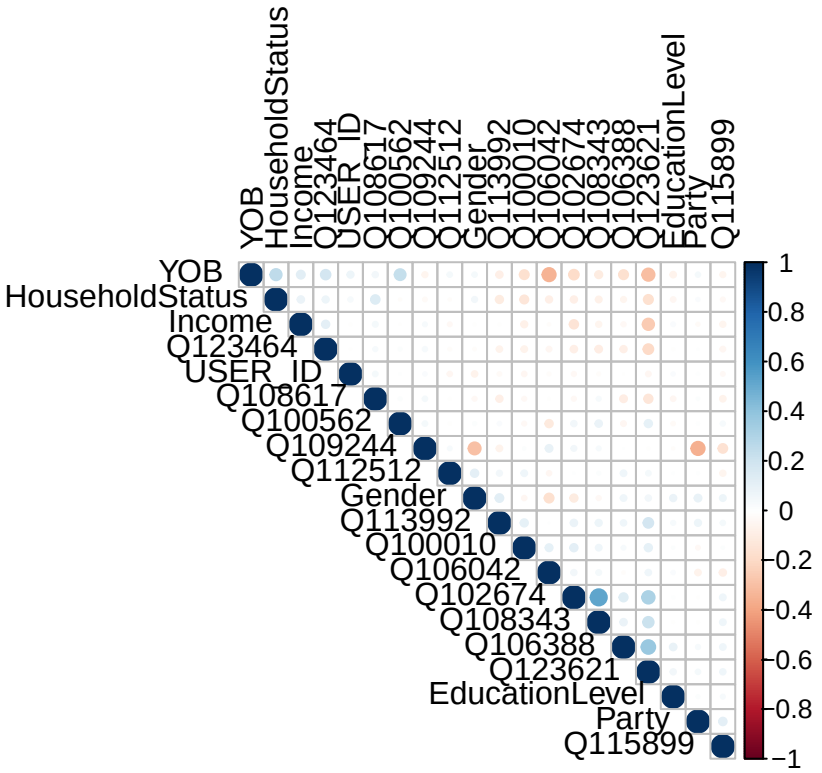
10.4 Appendix - D: Correlation Matrixes and Heat Maps

10.4.1 Correlation Matrix with Original Dataset

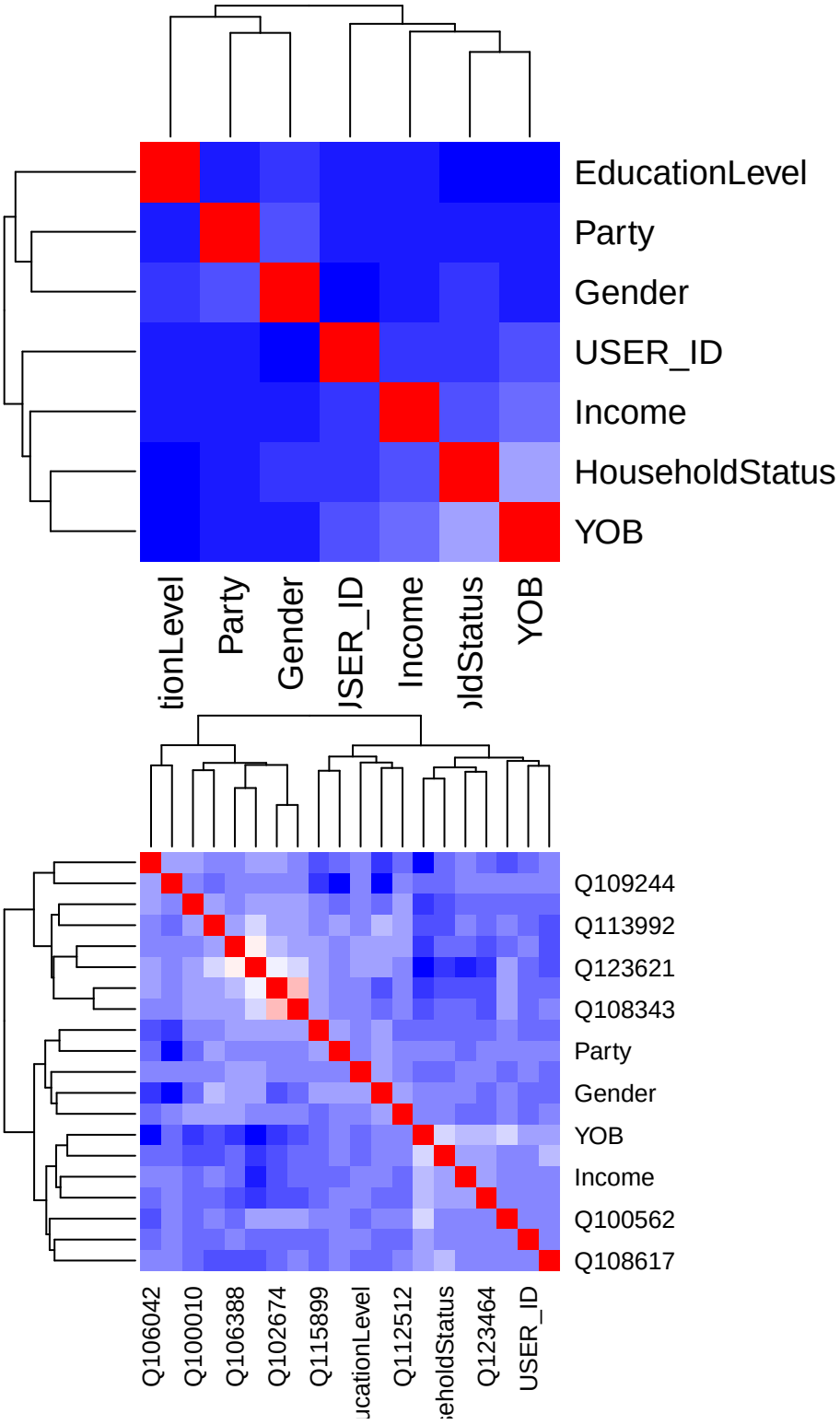




10.4.2 Correlation Matrix with Enhanced Dataset including Survey Questions



10.4.3 Heat Maps for both Datasets



10.5 Appendix - E: References

1. Mark Wickham (2018) Practical Jave Machine Learning: Projects with Google Cloud Platform and Amazon Web Services
2. Rafael A. Irizarry (2019), Introduction to Data Science: Data Analysis and Prediction Algorithms with R
3. Hie,Allaire,Grolemund (2020) R Markdown: The Definitive Guide
4. Yixuan Qiu (2017), recosystem: recommendation System Using Parallel Matrix Factorization
5. Tilman M. Davies (2016), The Book of R: A First Course In Programming and Statistics
6. Vries, Meys (2015), R For Dummies
7. Alvira Swalin (2018), Choosing the Right Metric for Evaluationg Machine Learning Models - Part 1
8. Shervin Minaee (2019), 20 Popular Machine Learning Metrics. Part 1: Classification & Regression Evaluation Metrics
9. Georgios Drakos (2018), How to select the Right Evalation Metri for Machine Learning Models: Part 2 Regression Metrics
10. Machine Learning: Classification Models
11. Predicting Voting Affiliation Using Machine Learning
12. Types of Classification Tasks in Machine Learning
13. 8 Proven Ways for Improving the "Accuracy" of a Machine Learning Model
14. What Affects Voter Turnout Rates
15. Voter Turnout Demographics
16. Voter Turnout
17. Parametric vs Nonparametric models?
18. Classification and Regression Trees for Machine Learning
19. Modern Machine Learning Algorithms: Strengths and Weaknesses
20. Pros and cons of common Machine Learning Algorithms
21. Machine Learning Algorithms Pros and Cons
22. Advantages and Disadvantages of Cross Validation in Machine Learning
23. Journa of Statistical Software: Feature Selection with the Boruta Package
24. Feature Selection in R with the Boruta R Package
25. bookdown: Authoring Books and Technical Documents with R Markdown
26. A Gentle Introduction to Threshold-Moving for Imbalanced Classification
27. Linear & Quadratic Discriminant Analysis
28. How to choose machine learning algorithms
29. How to Choose a Machine Learning Model - Some Guidelines
30. An easy guide to choose the right Machine Learning Algorithm
31. How to Read a Confusion Matrix
32. An Introduction to Statistical Learning
33. Using a Neural Network to Predict Voter Preferences
34. Ensemble Learning to Improve Machine Learning Results
35. Train-Test Split for Evaluating Machine Learning Algorithms
36. A Scaling law for validation-set training-set size ratio
37. Linear Discriminant Analysis vs Random Forests
38. Beginner's Guide to LDA Topic Modelling with R
39. Evaluation of Classification Model Accuracy: Essentials
40. Forecasting the 2015 General Election with Internet Big Data: An Application of the TRUST Framework

10.6 Appendix - F: Peer Assignment Grading Requirements

Grading (Rubric to be used by Peers)

Files (5 Points possible): Files Requirements 3 files: R script, RMD, and .pdf must be submitted.: - Files Points (5 points possible): Files Requirements:

```
- 0 points: No files provided AND/OR the files provided appear to violate the edX Honor Code.
- 3 points: One file is missing and/or not in the correct format.
- 5 points: All 3 files were submitted in the requested formats.
```

Report (25 points possible) : Report Requirements:

```
- Documents the analysis and presents findings, Contains supporting statistics and figures
- Written in English and must include the following (at a minimum sections)
    - Introduction/Overview/Executive Summary:
          Describes the dataset and summarizes the goal of the project and key steps performed
    -  Methods/Analysis
          Explains the process and techniquess used including: data cleaning,
          data exploration and visualization, insights gained, and modeling approach
    -  Results: Presents modeling results and discusses the model performance
    -  Conclusion: Brief summary of the report, its limitations and future work
```

- Report Points:
  - 0 points: The report is either not uploaded or contains very minimal information AND/OR the report appears to violate the edX Honor Code.
  - 5 points: Multiple required sections of the report are missing.
  - 10 points:The report includes all required sections, but the report is significantly difficult to follow or missing supporting detail in multiple sections.
  - 15 points: The report includes all required sections, but the report is difficult to follow or missing supporting detail in one section.
  - 20 points: The report includes all required sections and is easy to follow, but with minor flaws in one section.
  - 25 points: The report includes all required sections, is easy to follow with good supporting detail throughout, and is insightful and innovative.

Code (20 points): Code Requirements- Code should be well commented and easy to follow

- Code Points
  - 0 points: Code does not run and produces many errors or the code appears to violate the edX Honor Code.
  - 5 points: Code runs but does not produce output consistent with what is presented in the report OR there is overtraining (the test set is used for training steps).
  - 10 points:Code runs but is difficult to follow and/or may not produce output entirely consistent with what is presented in the report.
  - 15 points: Code runs, can be followed, is at least mostly consistent with the report, but is lacking (sufficient) comments and explanation OR uses absolute paths instead of relative paths OR does not automatically install missing packages OR does not provide easy access to the dataset (either via automatic download or inclusion in a GitHub repository).
  - 20 points: Code runs easily, is easy to follow, is consistent with the report, and is well-commented. All file paths are relative and missing packages are automatically installed with if(!require) statements.

10.7 Appendix - H: List of tables

10.8 Appendix - J: List of figures

10.9 Appendix - G: Survey Questions

Question ID	Question Text	Possible Answers
96024	Are you good at math?	Yes,No
98059	Do/did you have any siblings?	Yes,Only-child
98078	Do you have a "go-to" creative outlet?	Yes,No
98197	Do you pray or meditate on a regular basis?	Yes,No
98578	Do you exercise 3 or more times per week?	Yes,No
98869	Does life have a purpose?	Yes,No
99480	Did your parents spank you as a form of discipline/punishment?	Yes,No
99581	Are you left-handed?	Yes,No
99716	Do you live alone?	Yes,No
99982	Do you keep check-lists of tasks you need to accomplish?	Check!,Nope
100010	Do you watch some amount of TV most days?	Yes,No
100562	Do you think your life will be better five years from now than it is today?	Yes,No
100680	Have you cried in the past 60 days?	Yes,No
100689	Do you feel like you are currently overweight?	Yes,No
101162	Are you generally more of an optimist or a pessimist?	Optimist,Pessimist
101163	Which parent "wore the pants" in your household?	Mom,Dad
101596	As a kid, did you ever build (or help build) a tree-house?	Yes,No
102089	Do you rent or own your primary residence?	Rent,Own
102289	Does your life feel adventurous?	Yes,No
102674	Do you have any credit card debt that is more than one month old?	Yes,No
102687	Do you eat breakfast every day?	Yes,No
102906	Are you currently carrying a grudge against anyone in your personal life?	Yes,No
103293	Do you have more than one pet?	Yes,No
104996	Do you brush your teeth two or more times every day?	Yes,No
105655	Were you awakened by an alarm clock this morning?	Yes,No
105840	Do you ever treat yourself to "retail therapy"?	Yes,No
106042	Are you taking any prescription medications?	Yes,No
106272	Do you own any power tools? (power saws, drills, etc.)	Yes,No
106388	Do you work 50+ hours per week?	Yes,No
106389	Are you a good/effective liar?	Yes,No
106993	Do you like your given first name?	Yes,No
106997	Do you generally like people, or do most of them tend to get on your nerves pretty easily?	Yay people!,Grrr people
107491	Do you punctuate text messages?	Yes,No
107869	Do you feel like you're "normal"?	Yes,No
108342	Do you spend more time with friends online or in-person?	Online,In-person
108343	Do you feel like you have too much personal financial debt?	Yes,No
108617	Do you live in a single-parent household?	Yes,No
108754	Do both of your parents have college degrees?	Yes,No
108855	Do you enjoy getting together with your extended family?	Yes!,Umm...
108856	Lots of people are around! Are you more likely to be right in the middle of things, or looking for your own quieter space?	Socialize,Space
108950	Are you generally a cautious person, or are you comfortable taking risks?	Cautious,Risk-friendly
109244	Are you a feminist?	Yes,No
109367	Have you ever been poor (however you personally defined it at the time)?	Yes,No
110740	Mac or PC?	Mac,PC
111220	Is your alarm clock intentionally set to be a few minutes fast?	Yes,No
111580	As a teenager, do/did you have parents who were generally more supportive or demanding?	Supportive,Demanding
111848	Did you ever get a straight-A report card in high school or college?	Yes,No
112270	Are you better looking than your best friend?	Yes,No
112478	Do you have any phobias?	Yes,No
112512	Are you naturally skeptical?	Yes,No
113181	Do you meditate or pray on a regular basis?	Yes,No
113583	While driving: music or talk/news radio?	Tunes,Talk
113584	During your average day, do you spend more time interacting with people (face-to-face) or technology?	People,Technology
113992	Do you gamble?	Yes,No
114152	Do you support a particular charitable cause with a lot of your time and/or money?	Yes,No
114386	Are you more likely to over-share or under-share?	TMI,Mysterious
114517	Do you turn a TV on in the morning while getting ready for your day?	Yes,No
114748	Do you drink the unfiltered tap water in your home?	Yes,No
114961	Can money buy happiness?	Yes,No
115195	Do you live within 20 miles of a major metropolitan area?	Yes,No
115390	Has your personality changed much from what you were like as a child?	Yes,No
115602	Were you an obedient child?	Yes,No
115610	Does the "power of positive thinking" actually work?	Yes,No
115611	Do you personally own a gun?	Yes,No
115777	Do you find it easier to start and maintain a new good habit, or to permanently kick a bad habit?	Start,End

Figure 16: Survey - Election Questions - Page 1

115899	Would you say most of the hardship in your life has been the result of circumstances beyond your own control, or has it been mostly the result of your own decisions and actions?	Circumstances,Me
116197	Are you a morning person or a night person?	A.M.,P.M.
116441	Do you have a car payment?	Yes,No
116448	If you had to stop telling *any* lies for 6 months (even the smallest "little-white-lie" would immediately make you violently ill), would it change your life in any noticeable way?	Yes,No
116601	Have you ever traveled out of the U.S.?	Yes,No
116797	Do you take a daily multi-vitamin?	Yes,No
116881	Would you rather be happy or right?	Happy,Right
116953	Do you like rules?	Yes,No
117186	Do you have a quick temper?	Hot headed,Cool headed
117193	Do you work (or attend school) on a pretty standard "9-to-5ish" daytime schedule, or do you have to work unusual hours?	Standard hours,Odd hours
118117	Have you lived in the same state your whole life?	Yes,No
118232	Are you more of an idealist or a pragmatist?	Idealist,Pragmatist
118233	Have you ever had your life genuinely threatened by intentional violence (or the threat of it)?	Yes,No
118237	Do you feel like you are "in over-your-head" in any aspect of your life right now?	Yes,No
118892	Do you wear glasses or contact lenses?	Yes,No
119334	Did you accomplish anything exciting or inspiring in 2013? (comments from the 2012 poll are linked for inspiration)	Yes,No
119650	Which do you really enjoy more: giving or receiving?	Giving,Receiving
119851	Are you in the middle of reading a good book right now?	Yes,No
120012	Does the weather have a large effect on your mood?	Yes,No
120014	Are you more successful than most of your high-school friends?	Yes,No
120194	Your generally preferred approach to starting a new task: read up on everything you can before trying it out, or dive in with almost no knowledge and learn as you go?	Study first,Try first
120379	Do you have (or plan to pursue) a Masters or Doctoral degree?	Yes,No
120472	Science or Art?	Science,Art
120650	Were your parents married when you were born?	Yes,No
120978	As a kid, did you watch Sesame Street on a regular basis?	Yes,No
121011	Changing or losing a job, getting married or divorced, the death of a close relative, moving, a major health issue, bankruptcy...all are life events that can create high stress for people. Have you experienced any of these in 2013?	Yes,No
121699	2013: did you drink alcohol?	Yes,No
121700	2013: did you start a new romantic relationship?	Yes,No
122120	Your significant other takes an extra long look at a very attractive person (of your gender) walking past both of you. Are you upset?	Yes,No
122769	Do you collect anything (as a hobby)?	Yes,No
122770	Do you have more than $20 cash in your wallet or purse right now?	Yes,No
122771	Do/did you get most of your K-12 education in public school, or private school?	Public,Private
123464	Do you currently have a job that pays minimum wage?	Yes,No
123621	Are you currently employed in a full-time job?	Yes,No
124122	Did your parents fight in front of you?	Yes,No
124742	Do you have to personally interact with anyone that you really dislike on a daily basis?	Yes,No

Figure 17: Survey - Election Questions - Page 2