Analysis on Factors that Influence Voting Intention

By Group 19: Yifei Cao, Zhong Cheng, Ruoxi Lan and Yahan Wang

01

PROJECT OVERVIEW

Project Objectives

- **Background**

  Since 2012, the voter turnout rate has been rising and attained the highest at 67% in 2020. We are interested in the underlying reasons that might have caused the spike. We are also wondering why the remaining 33% did not vote. The non-voters dataset comes from polling done by Ipsos for FiveThirtyEight. The poll was conducted in 2020 among a sample of U.S. citizens that included 5836 respondents from different racial and ethnic groups.

- **Project Objectives**

  The most interesting questions about elections are not who won, but why people voted the way they did, or what the results implicated. We are concerned about major reasons that affect electoral behavior to better understand the election outcome. We thus will discuss the statistical and modeling methods that we learned in this data mining course to analyze the voting behavior based on the poll results.

- **Data Source**

  Non-Voters from FiveThirtyEight https://github.com/fivethirtyeight/data/tree/master/non-voters

Dataset Overview

- **Initial Overview:**
  Data looks not very clear
  - 114 numeric variables
  - 5 character variables

- **Adjustments:**
  - Set an outcome: Voting_Intention.
  - Data contains too many coded numeric variables which should be transformed to categories
    For example, for the CANDIDATE variable:

    Q23. Candidate that you are planning to support?

    1. Donald Trump
    2. Joe Biden
    3. Unsure

  - Deleted unrelated or duplicate attributes.
  - Re-grouped values: For instance, one variable is a question that has answers from 1 (Strongly agree) to 4 (Strongly disagree), we combined them into two values: “Agree” and “Disagree”.
  - **RespId as a unique identifier.**

Data Summary

Output

Voting_Intention

Yes / No

Numeric (4)

PPAGE: Age of respondent

Weight: Weight of respondent

Voting Times:

Number of voting in the past 6 elections

Number of Confident Ways:

How many ways of voting does respondent feel secure and safe from fraud

Category (17)

GENDER

EDU

College,
High school or less

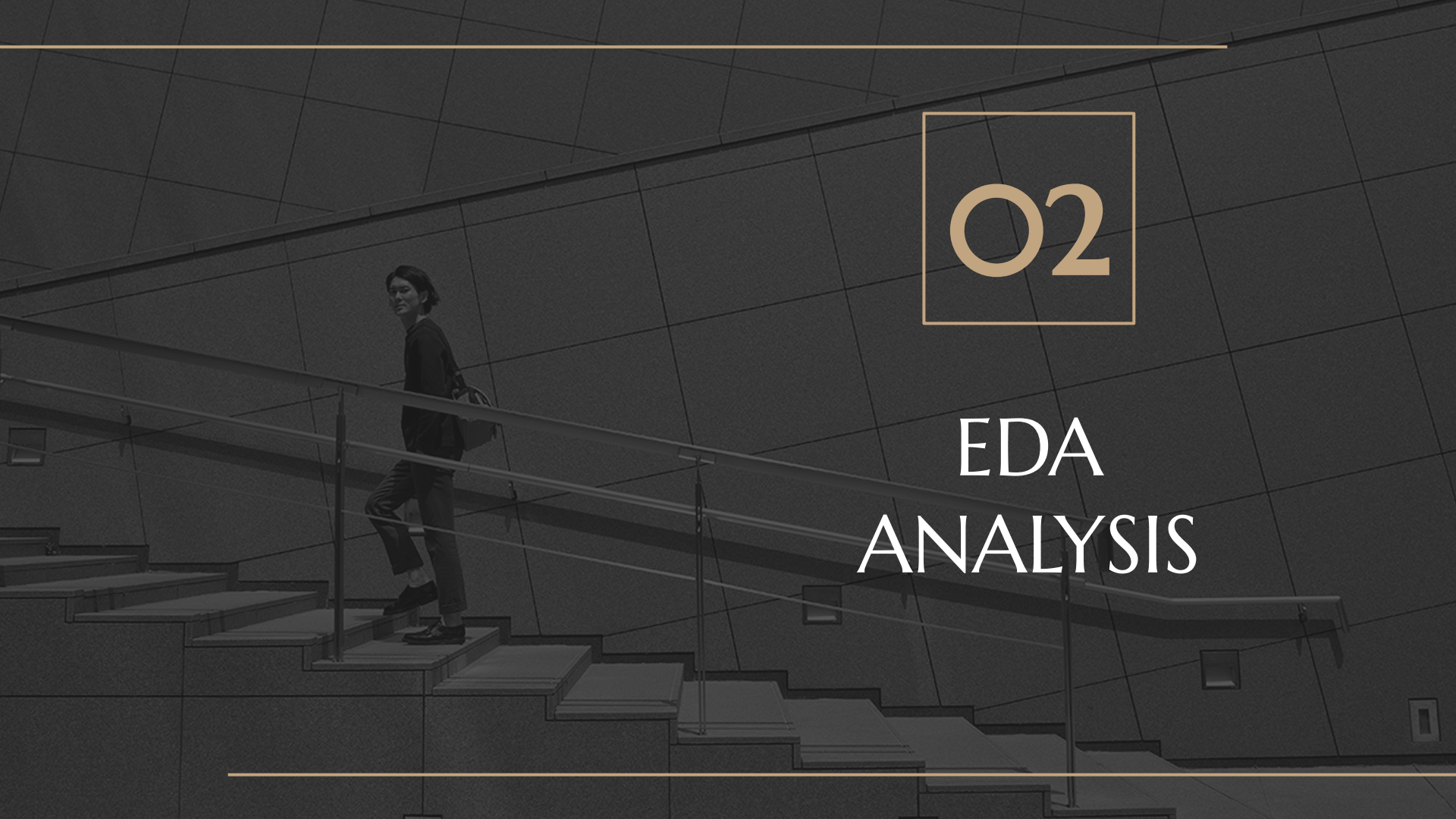
RACE

White, Black, Hispanic, Other/Mixed

INCOME_CAT

Less than $40k, $40-75k, $75-125k, $125k or more

.....



02

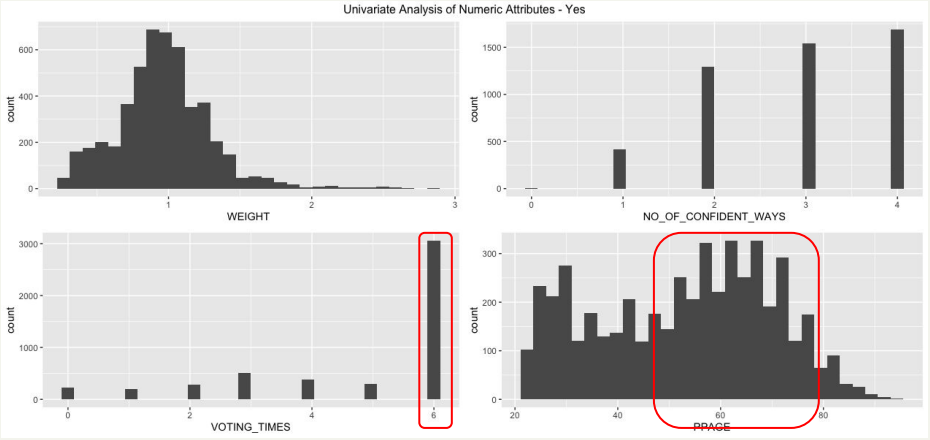
EDA ANALYSIS

Univariate Analysis

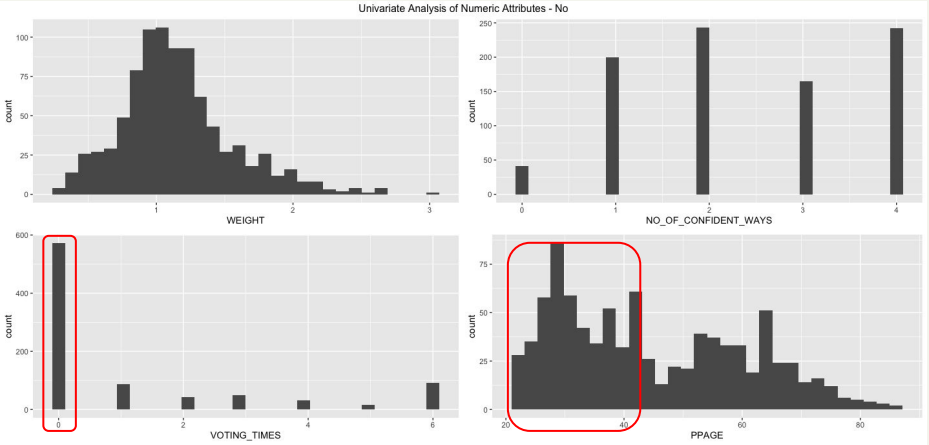
This part will be the univariate analysis of numeric and factors attributes by comparing different VOTING_INTENTION

Univariate Analysis - Numeric Attributes

VOTING_INTENTION = Yes



VOTING_INTENTION = No



Univariate Analysis - Factors Attributes

Logical Groups

Subjective Factors

ATTITUDE, TRUST_FOR_PRESIDENCY, DIFFICULTY_TO_VOTE, REGISTERED, PREFERRED_METHOD_OF_VOTING, CANDIDATE

Objective Factors

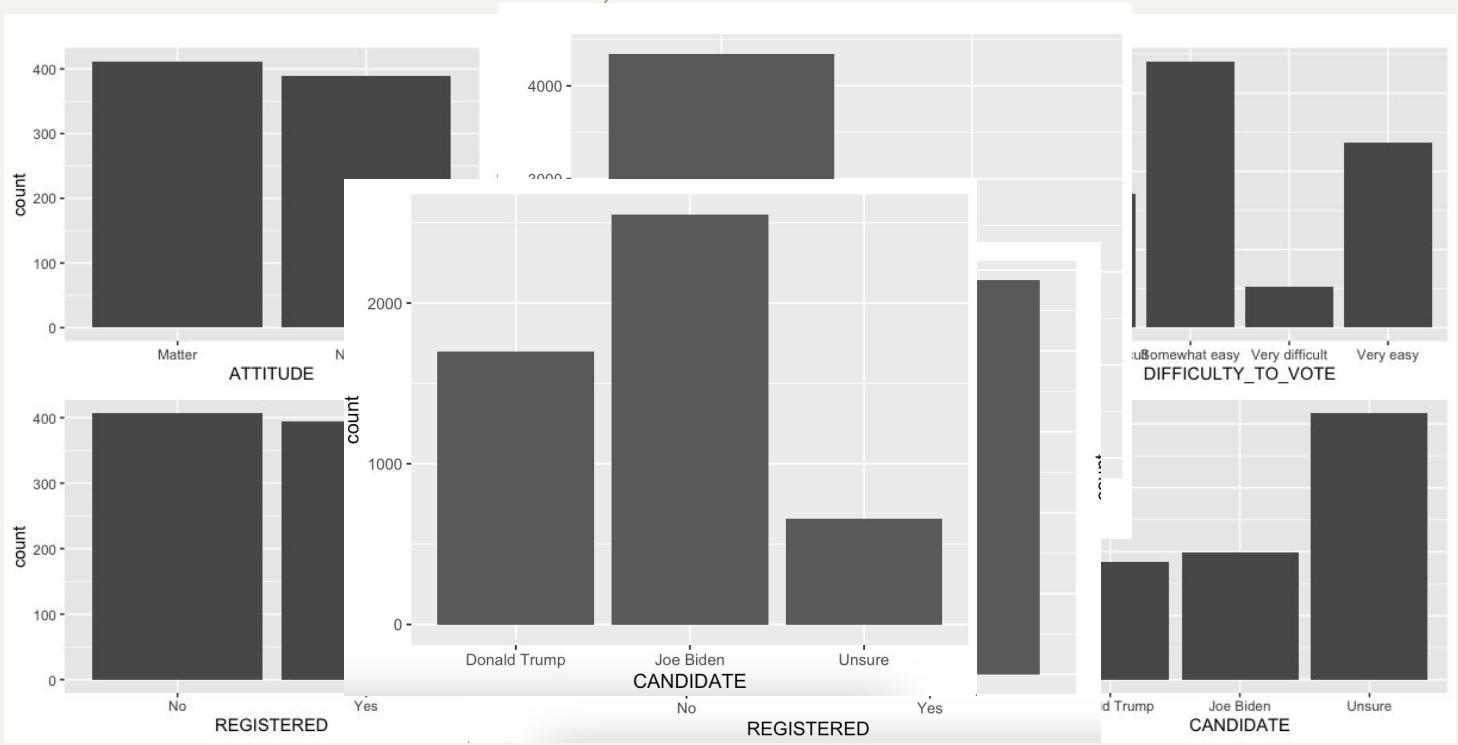
PERSONAL_REASON, COVID_REASON, FAMILY_REASON

Personal Information

PARTY, EDUC, RACE, GENDER, INCOME_CAT, VOTER_CATEGORY

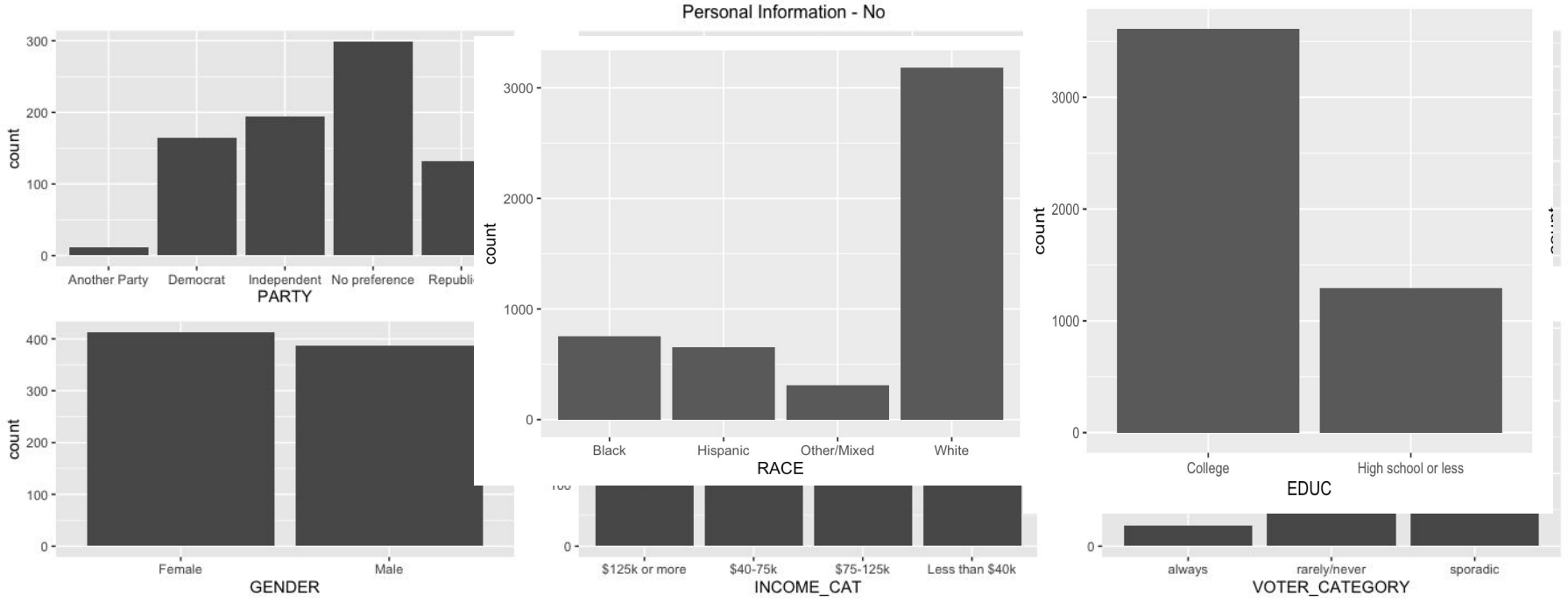
Univariate Analysis - Factors Attributes

Subjective Factors



Univariate Analysis - Factors Attributes

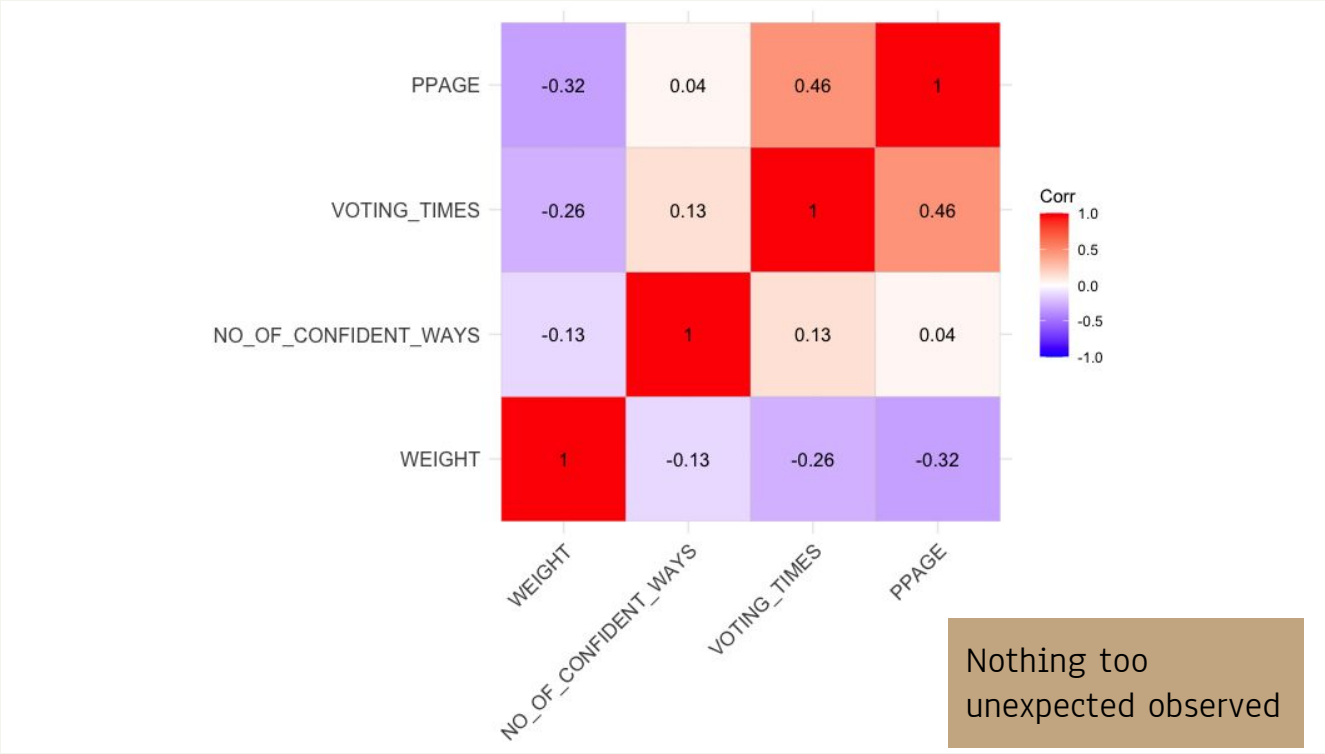
Personal Information



Bivariate Analysis

This part will be the bivariate analysis between Categories (factors) attributes and Measures (numerics) attributes.

Bivariate Analysis - Measures vs Measures



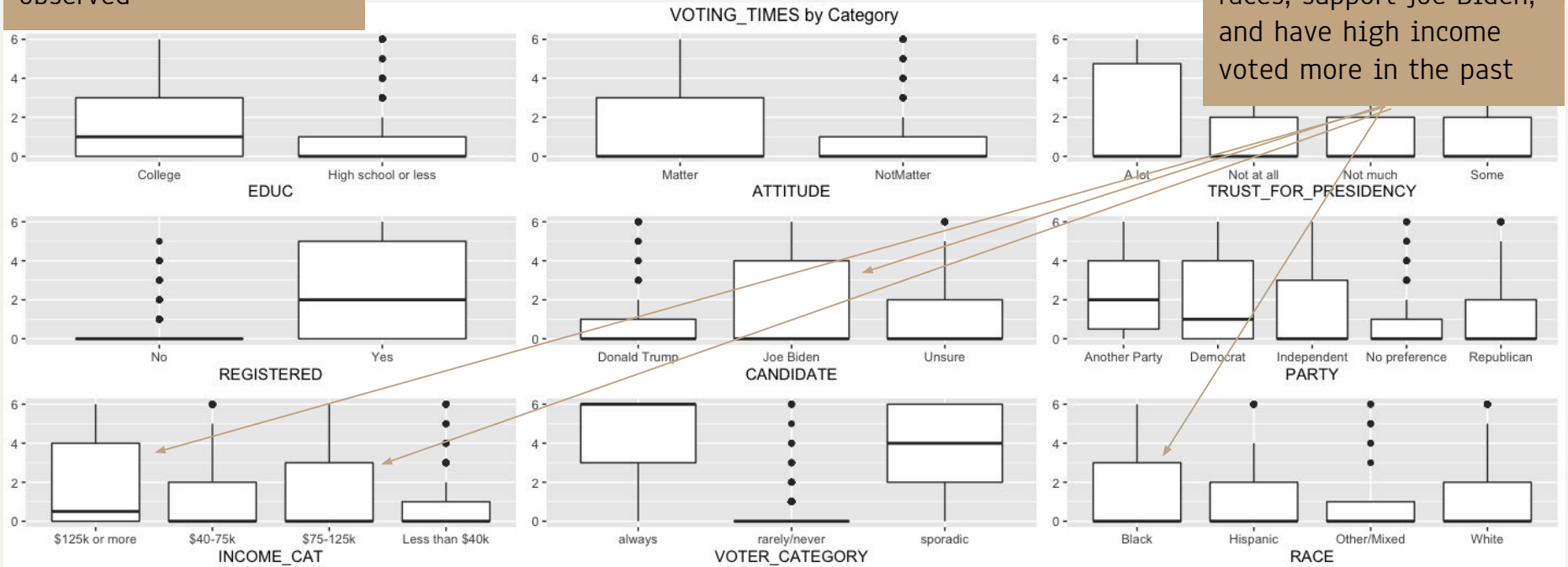
Nothing too unexpected observed

Bivariate Analysis - Categories vs Measures

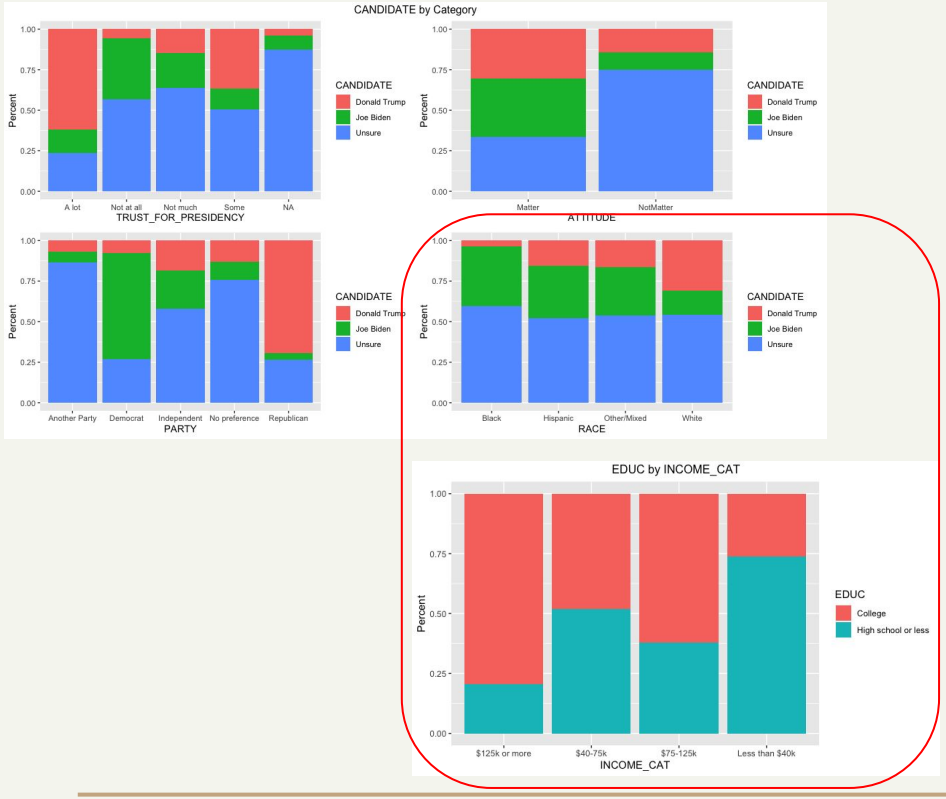
VOTING_TIMES by Selected Categories

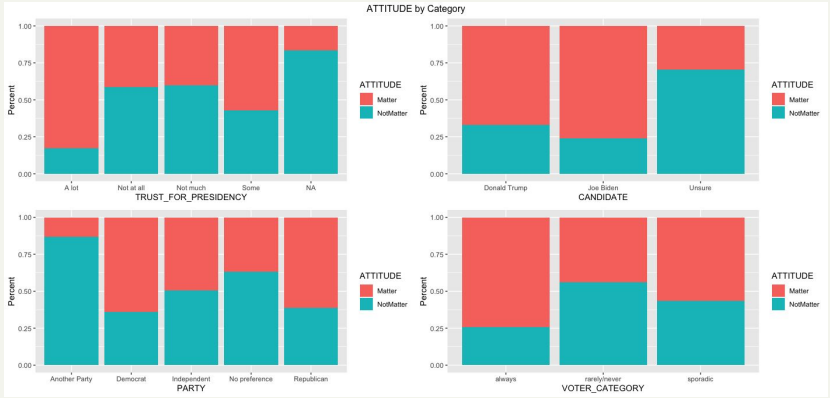
Nothing too unexpected observed

People who are black races, support Joe Biden, and have high income voted more in the past



Bivariate Analysis - Categories vs Categories





03

Statistical Analysis

STATISTICAL ANALYSIS

Anova Test

Examining the statistical significance of categorical attributes, and then conduct multiple comparisons of means between groups of each attributes by using the function TukeyHSD()

T.test

Comparing the means of each groups of numerical variables

Chi-Square Test

Check whether attributes are significant or insignificant

STATISTICAL ANALYSIS

Anova Test Result

```
             Df Sum Sq Mean Sq F value Pr(>F)
ATTITUDE      2  109.8   54.90   496.4 <2e-16 ***
Residuals   5833  645.2    0.11
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = VOTING_INTENTION ~ ATTITUDE, data = nonvoters_3)

$ATTITUDE
                                                                    diff        lwr       upr    p adj
Things will be pretty much the same--1                         0.2372183 0.09502326 0.3794134 0.000274
Who wins the election really matters--1                        0.5876065 0.44712550 0.7280874 0.000000
Who wins the election really matters-Things will be pretty much the same 0.3503881 0.32323496 0.3775413 0.000000
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
CANDIDATE     2  106.3   53.13   477.8 <2e-16 ***
Residuals   5833  648.7    0.11
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = VOTING_INTENTION ~ CANDIDATE, data = nonvoters_3)

$CANDIDATE
                           diff          lwr          upr     p adj
Joe Biden-Donald Trump  0.02823836  0.004987698  0.05148903 0.0123074
Unsure-Donald Trump    -0.32083188 -0.350004694 -0.29165906 0.0000000
Unsure-Joe Biden       -0.34907024 -0.376468126 -0.32167235 0.0000000
```

```
                      Df Sum Sq Mean Sq F value   Pr(>F)
TRUST_FOR_PRESIDENCY   1    1.5  1.4924   11.74 0.000616 ***
Residuals           5787  735.7  0.1271
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
因为不存在，47个观察量被删除了
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = VOTING_INTENTION ~ TRUST_FOR_PRESIDENCY, data = nonvoters_3)

$TRUST_FOR_PRESIDENCY
             diff        lwr        upr     p adj
Yes-No 0.03239766 0.01386129 0.05093403 0.0006161
```

ATTITUDE: People who think “Things will be pretty much the same” are less likely to vote.

TRUST FOR PRESIDENCY: People with answer of No are unlikely to vote.

CANDIDATE: People who are unsure who they support are less likely to vote.

STATISTICAL ANALYSIS

Anova Test Result

```
             Df Sum Sq Mean Sq F value Pr(>F)
PARTY         4  104.7  26.176   234.7 <2e-16 ***
Residuals  5831  650.3   0.112
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = VOTING_INTENTION ~ PARTY, data = nonvoters_3)

$PARTY
                                  diff         lwr         upr     p adj
Democrat-Another Party       0.066600067 -0.02721681  0.16041694 0.2977305
Independent-Another Party    0.008357257 -0.08635039  0.10306490 0.9992585
No preference-Another Party -0.352663121 -0.45035282 -0.25497342 0.0000000
Republican-Another Party     0.063575453 -0.03080895  0.15795986 0.3517044
Independent-Democrat        -0.058242810 -0.08982448 -0.02666114 0.0000050
No preference-Democrat      -0.419263188 -0.45890083 -0.37962555 0.0000000
Republican-Democrat         -0.003024614 -0.03362332  0.02757409 0.9988405
No preference-Independent   -0.361020378 -0.40272258 -0.31931818 0.0000000
Republican-Independent       0.055218196  0.02198851  0.08844788 0.0000578
Republican-No preference     0.416238574  0.37527576  0.45720139 0.0000000
```

```
                Df Sum Sq Mean Sq F value Pr(>F)
VOTER_CATEGORY   2  176.6   88.31   890.6 <2e-16 ***
Residuals     5833  578.4    0.10
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = VOTING_INTENTION ~ VOTER_CATEGORY, data = nonvoters_3)

$VOTER_CATEGORY
                           diff         lwr          upr     p adj
rarely/never-always   -0.41993431 -0.44594264 -0.393925982 0.0000000
sporadic-always       -0.03180984 -0.05445027 -0.009169407 0.0028581
sporadic-rarely/never  0.38812447  0.36389141  0.412357537 0.0000000
```

PARTY: People have no party preference are less likely to vote.

VOTER CATEGORY: People who rarely or never attend the vote are less likely to vote.

STATISTICAL ANALYSIS

T-Test Result

```
        Welch Two Sample t-test

data:  WEIGHT by VOTING_INTENTION
t = 12.817, df = 1093, p-value < 2.2e-16
alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
95 percent confidence interval:
 0.1587043 0.2160778
sample estimates:
 mean in group No mean in group Yes
        1.1498048         0.9624138


        Welch Two Sample t-test

data:  VOTING_TIMES by VOTING_INTENTION
t = -49.352, df = 1167, p-value < 2.2e-16
alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
95 percent confidence interval:
 -3.714862 -3.430784
sample estimates:
 mean in group No mean in group Yes
         1.210999          4.783822


        Welch Two Sample t-test

data:  PPAGE by VOTING_INTENTION
t = -14.085, df = 1292.3, p-value < 2.2e-16
alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
95 percent confidence interval:
 -9.317099 -7.039014
sample estimates:
 mean in group No mean in group Yes
         44.76431          52.94237
```

Conclusion:

- WEIGHT: People with higher weight are less likely to vote.
- AGE: Younger people are less likely to vote.
- VOTING TIMES: People with less voting time are less likely to vote.

STATISTICAL ANALYSIS

Chi Square Test Result

Attribute	P-Value < 0.05
Attitude	Yes
Trust for Presidency	Yes
Personal Reasons	Yes
Covid Reasons	Yes
Family Reasons	Yes
Difficulty to Vote	Yes
Registered	Yes
Candidate	Yes

Attribute	P-Value < 0.05
Preferred Method	Yes
Voting Behavior	Yes
Party	Yes
Education	Yes
Race	Yes
Gender	No
Income Category	Yes
Voting Category	Yes

Insignificant Attribute

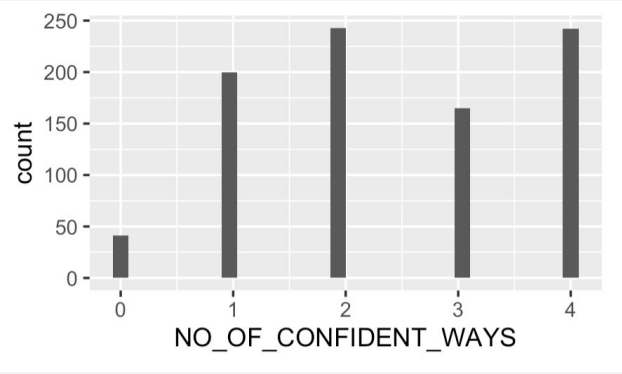
04

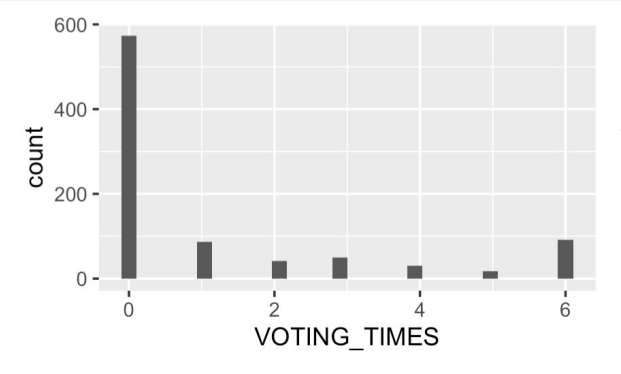
Modeling

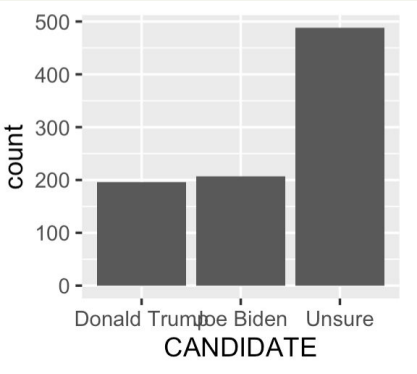
Logistic Regression & Decision Tree

Logistic Regression

Step 1: Data Preparation







- For nonvoters, cut-off point of above three categories are
  NO_OF_CONFIDENT_WAYS = 0
  VOTING_TIMES = 0
  CANDIDATE: have preference or not

- Remodel the category to:
  NO_OF_CONFIDENT_WAYS = { 0, morethan0 }
  VOTING_TIMES = { 0, morethan0 }
  CANDIDATE = {Yes, No}

```
nonvoters = nonvoters %>% mutate(VOTING_TIMES = as.factor(
  case_when(
    VOTING_TIMES == 0 ~ '0',
    VOTING_TIMES == 1 ~ 'morethan0',
    VOTING_TIMES == 2 ~ 'morethan0',
    VOTING_TIMES == 3 ~ 'morethan0',
    VOTING_TIMES == 4 ~ 'morethan0',
    VOTING_TIMES == 5 ~ 'morethan0',
    VOTING_TIMES == 6 ~ 'morethan0'
    )
)
)
```

Logistic Regression

Step 2: Partition

randomly splitting the data into
training set: 80% for building a predictive model
test set: 20% for evaluating the model

```
set.seed(123)
training_samples <- mod1$VOTING_INTENTION %>%
  createDataPartition(p = 0.8, list = FALSE)
train1  <- mod1[training_samples, ]
test1 <- mod1[-training_samples, ]
```

Step 3: Create model

Significant factors:	Insignificant factors:
Attitude	Party
Difficulty to vote	Gender
Candidate	Trust for residency
Voting time	Number of confident ways
Education	

```
Call:
glm(formula = VOTING_INTENTION ~ ., family = "binomial", data = train1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9721   0.1960   0.2485   0.3343   2.8105

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -1.85413    1.31602  -1.409 0.158869
ATTITUDENotMatter               -1.27161    0.13473  -9.438  < 2e-16 ***
TRUST_FOR_PRESIDENCYYes          0.01968    0.12999   0.151 0.879677
PERSONAL_REASONSYes              0.10388    0.12629   0.823 0.410755
COVID_REASONSYes                 0.28601    0.12204   2.344 0.019097 *
FAMILY_REASONSYes               -0.36325    0.13757  -2.640 0.008280 **
DIFFICULTY_TO_VOTEEasy           0.51824    0.14585   3.553 0.000381 ***
NO_OF_CONFIDENT_WAYSmorethan0    0.70290    1.21409   0.579 0.562622
CANDIDATEYes                     0.92393    0.14054   6.574 4.89e-11 ***
VOTING_TIMESmorethan0            3.03316    0.12844  23.615  < 2e-16 ***
PARTYDemocrat                    0.23856    0.43921   0.543 0.587018
PARTYIndependent                 0.10394    0.43284   0.240 0.810218
PARTYNo preference              -0.44705    0.43875  -1.019 0.308248
PARTYRepublican                  0.27177    0.44420   0.612 0.540668
EDUCHigh school or less         -0.53948    0.12658  -4.262 2.03e-05 ***
RACEHispanic                     0.42182    0.20156   2.093 0.036367 *
RACEOther/Mixed                  0.22857    0.25903   0.882 0.377560
RACEWhite                        0.32701    0.16752   1.952 0.050934 .
GENDERMale                       0.01062    0.11789   0.090 0.928232
INCOME_CAT$40-75k               -0.18672    0.18850  -0.991 0.321918
INCOME_CAT$75-125k              -0.38226    0.18360  -2.082 0.037341 *
INCOME_CATLess than $40k        -0.48988    0.19123  -2.562 0.010413 *
```

Logistic Regression

```
test1$model_prob <- predict(logistic_model, test1,
type = "response")
test1<- test1  %>% mutate(model_pred =
1*(model_prob > .5) + 0,
                          VOTING_INTENTION_binary
= 1*(VOTING_INTENTION == "Yes") + 0)
test1 <- na.omit(test1)
test1 <- test1 %>% mutate(accurate = 1*(model_pred
== VOTING_INTENTION_binary))
sum(test1$accurate)/nrow(test1)
```

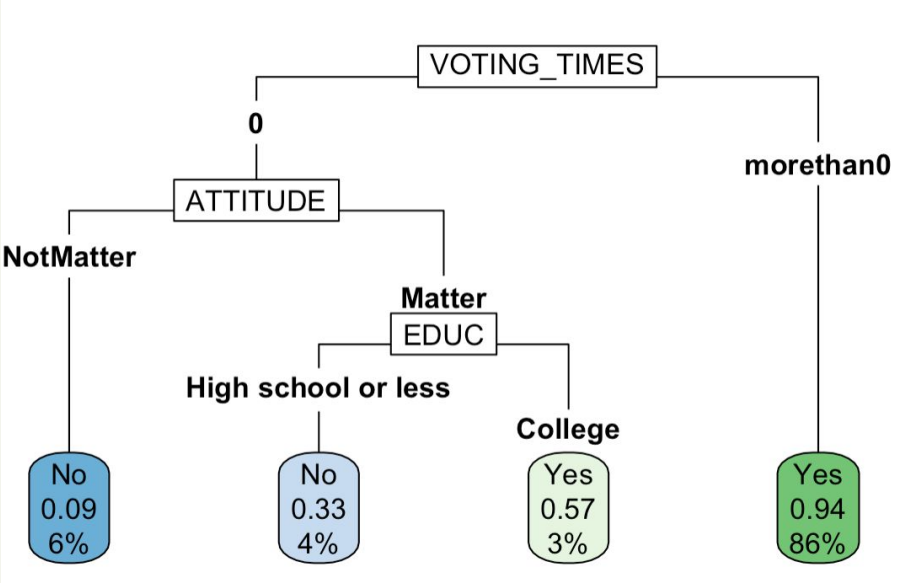
```
[1] 0.9095652
```

Step 4: Quantify the Results

The function predict, applied to the model and the training set will give us a probability for every observation. We save those probabilities to the testing set under the variable “model_pred”.

The accuracy rate is 0.9095652

Decision Tree



- Using 5 significant factors from logistic regression model
- Misclassification rate is 0.09304348
- Confusion Matrix

	Actual	
Predicted	No	Yes
No	85	26
Yes	81	958

```
Variables actually used in tree construction:
[1] ATTITUDE     CANDIDATE    EDUC         VOTING_TIMES
```

```
VOTING_TIMES     ATTITUDE    CANDIDATE         EDUC
  531.729540    41.656055    27.761153     5.944306
```

Decision Tree

- Using all factors from logistic regression model
- Misclassification rate is 0.09043478

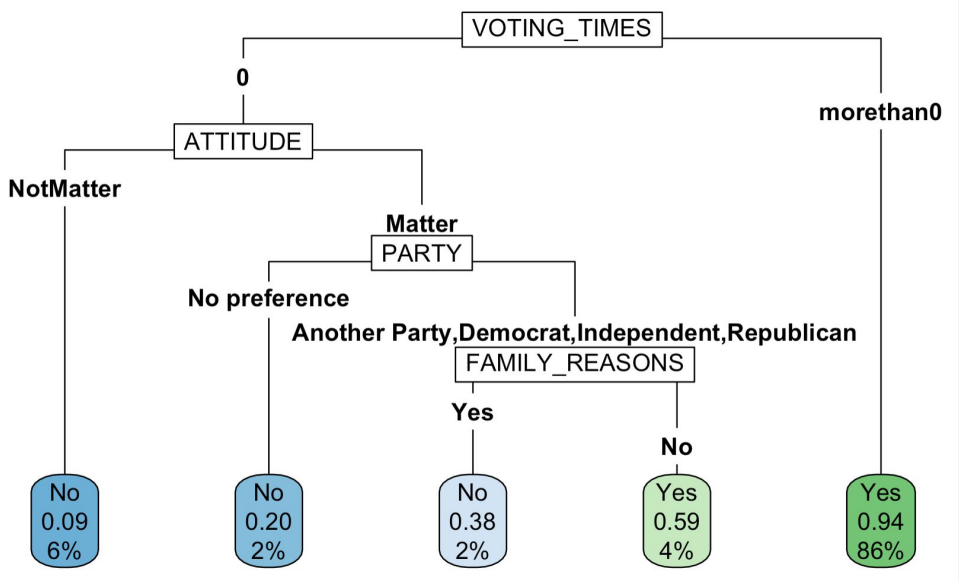
	Actual	
Predicted	No	Yes
No	85	23
Yes	81	961

- Variables actually in use:

ATTITUDE EDUC INCOME_CAT PARTY VOTING_TIMES

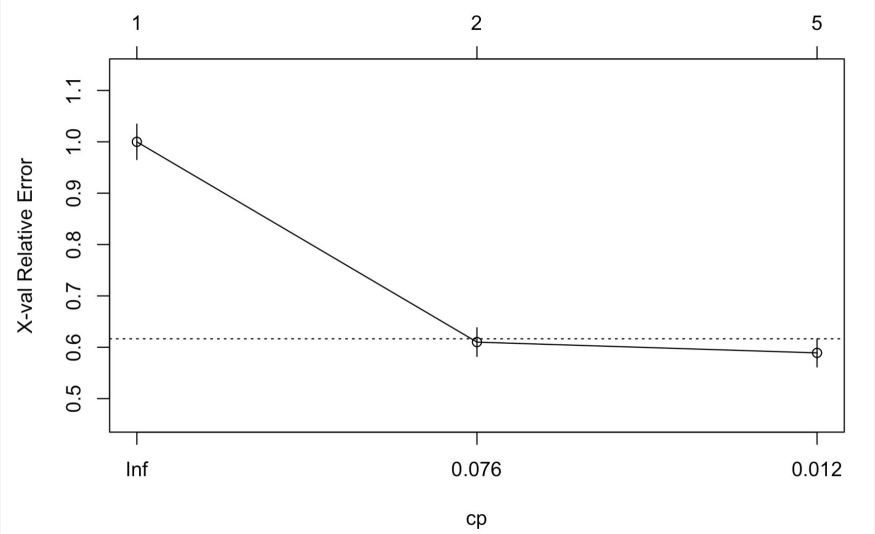
- Variable Importance

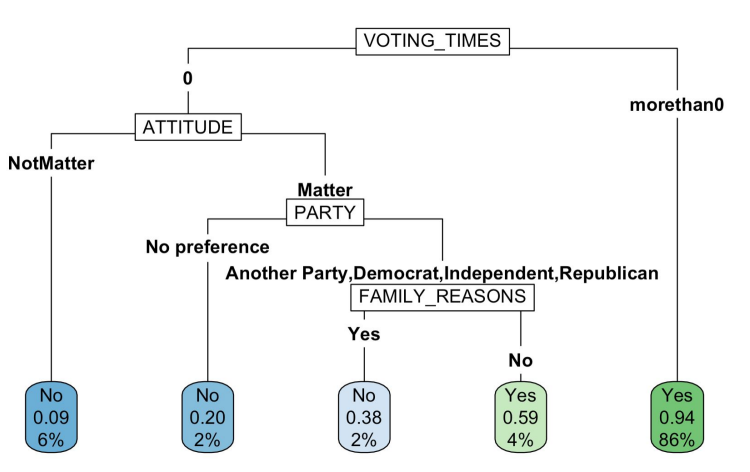
VOTING_TIMES	ATTITUDE	PARTY	CANDIDATE
430.2169338	32.8586969	17.0438024	13.1642937
NO_OF_CONFIDENT_WAYS	EDUC	INCOME_CAT	TRUST_FOR_PRESIDENCY
6.7772981	4.3353323	3.4728895	0.5114194
RACE	PERSONAL_REASONS		
0.1970606	0.1313737		



Decision Tree - Pruning Tree

- Prevent overfitting
- Optimal CP: minsplit = 5, cp = 0.012
- Same misclassification rate : 0.09043478





Conclusion

	1st	2nd	3th	4th	Accuracy
Logistics Regression	Voting Times	Number of Confident ways	Attitude	Candidate	90.95%
Decision tree1	Voting Times	Attitude	Candidate	Education	90.69%
Decision tree2	Voting Times	Attitude	Party	Candidate	90.96%

05

PROJECT SUMMARY

Conclusions

- The modeling results indicate that voting time is the most important factor influencing people's voting behavior. People who have voting experiences in the past are more likely to vote in the future. Thus, if the government wants more people to engage in voting, they can offer rewards to those who have never voted before.

- Common Significant Factors:
  Attitude: People who believe that their vote won't change the current situation are less likely to vote
  Candidate: People who have a clear candidate are more likely to vote
  Party: No party preference leads to less possibility to vote
  Education: People who have received higher education are more likely to vote

- Gender is not a significant factor that decides whether the respondent will vote

Work Allocation

- Data Cleaning, project plan, project overview and summary - Yifei Cao
- EDA Analysis - Cheng Zhong
- Statistical Analysis - Yahan Wang
- Modeling - Ruoxi Lan

06

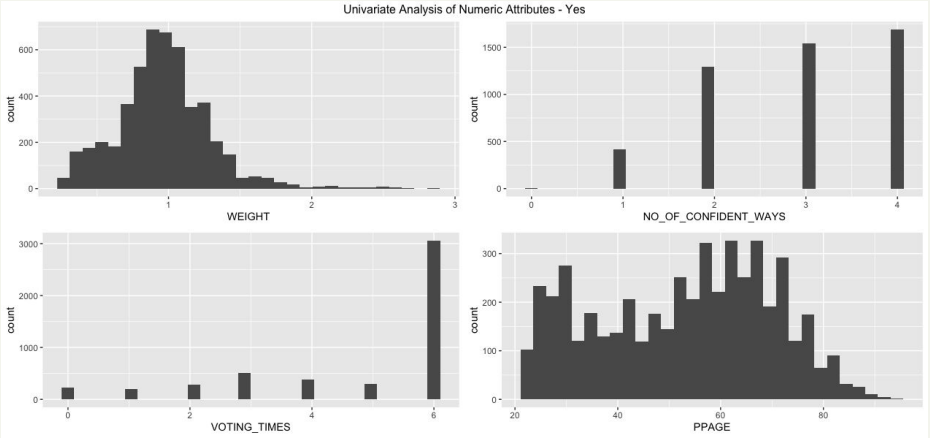
Full EDA Report

Univariate Analysis

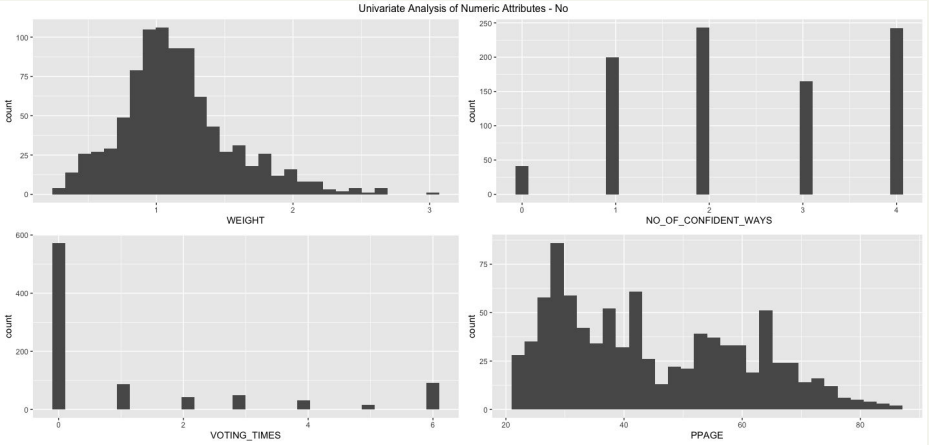
This part will be the univariate analysis of numeric and factors attributes by comparing different VOTING_INTENTION

Univariate Analysis - Numeric Attributes

VOTING_INTENTION = Yes

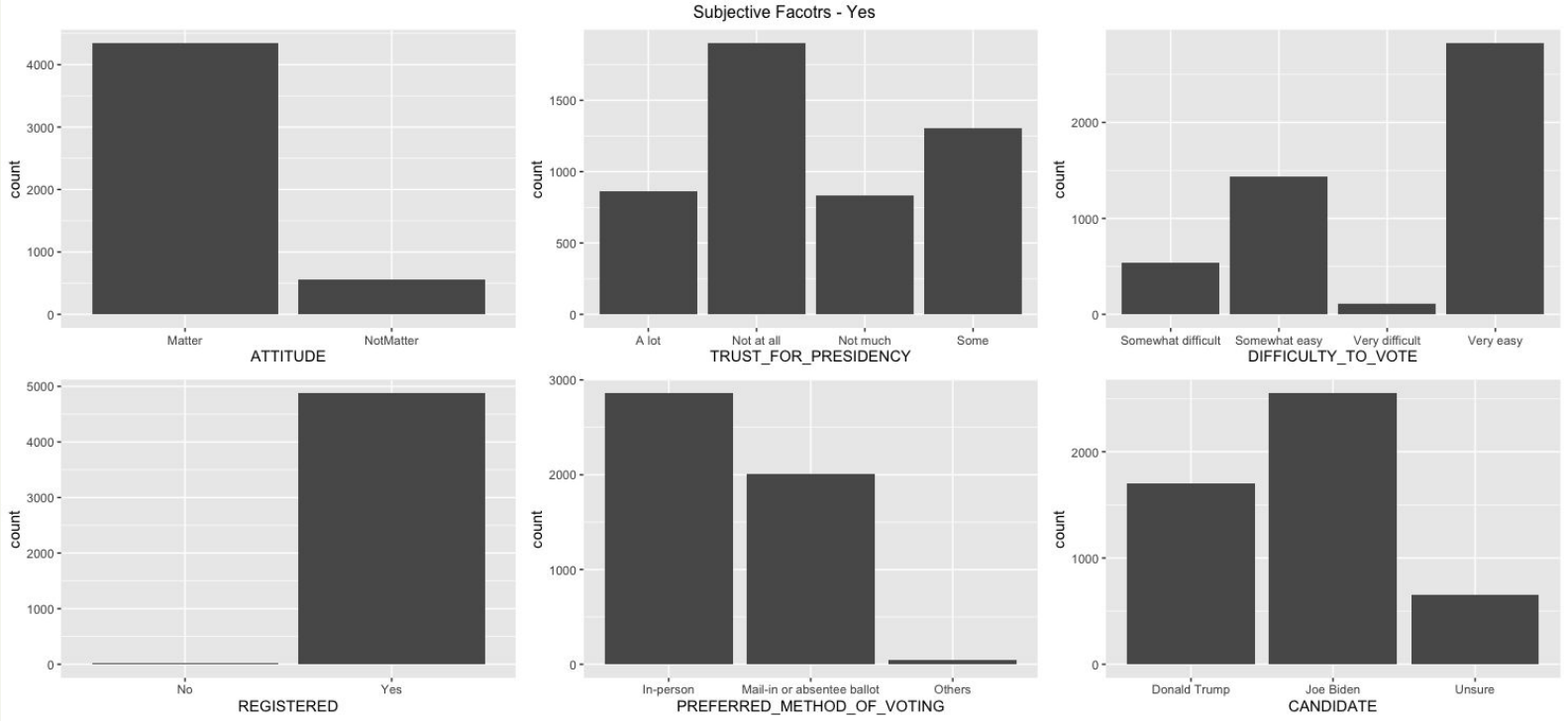


VOTING_INTENTION = No



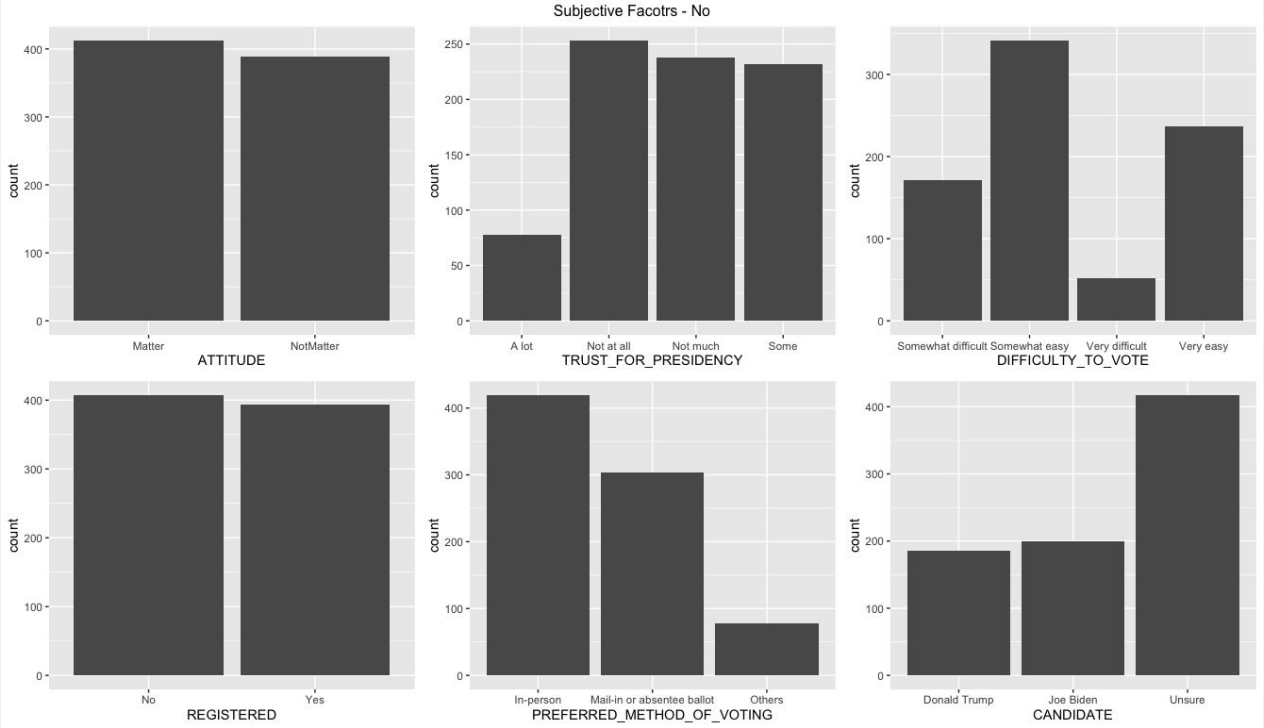
Univariate Analysis - Factors Attributes

Subjective Factors - VOTING_INTENTION = Yes



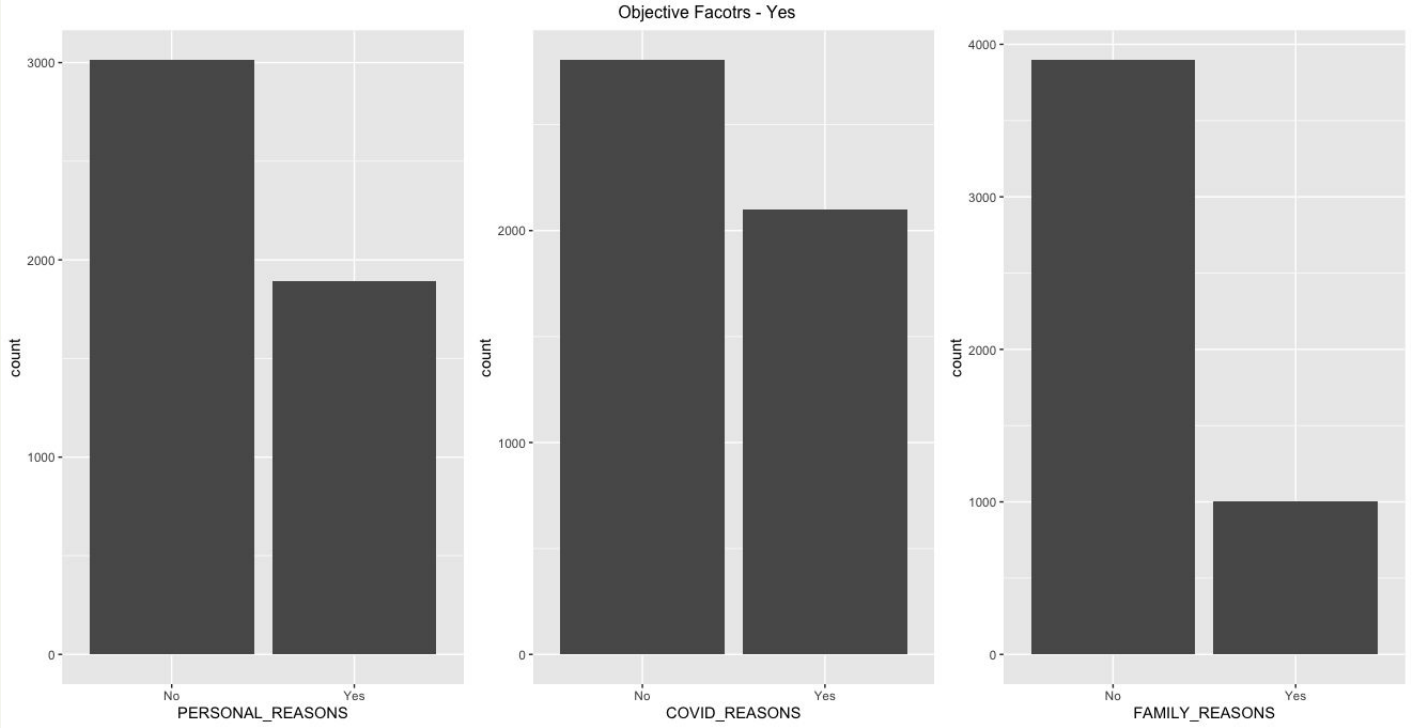
Univariate Analysis - Factors Attributes

Subjective Factors - VOTING_INTENTION = No



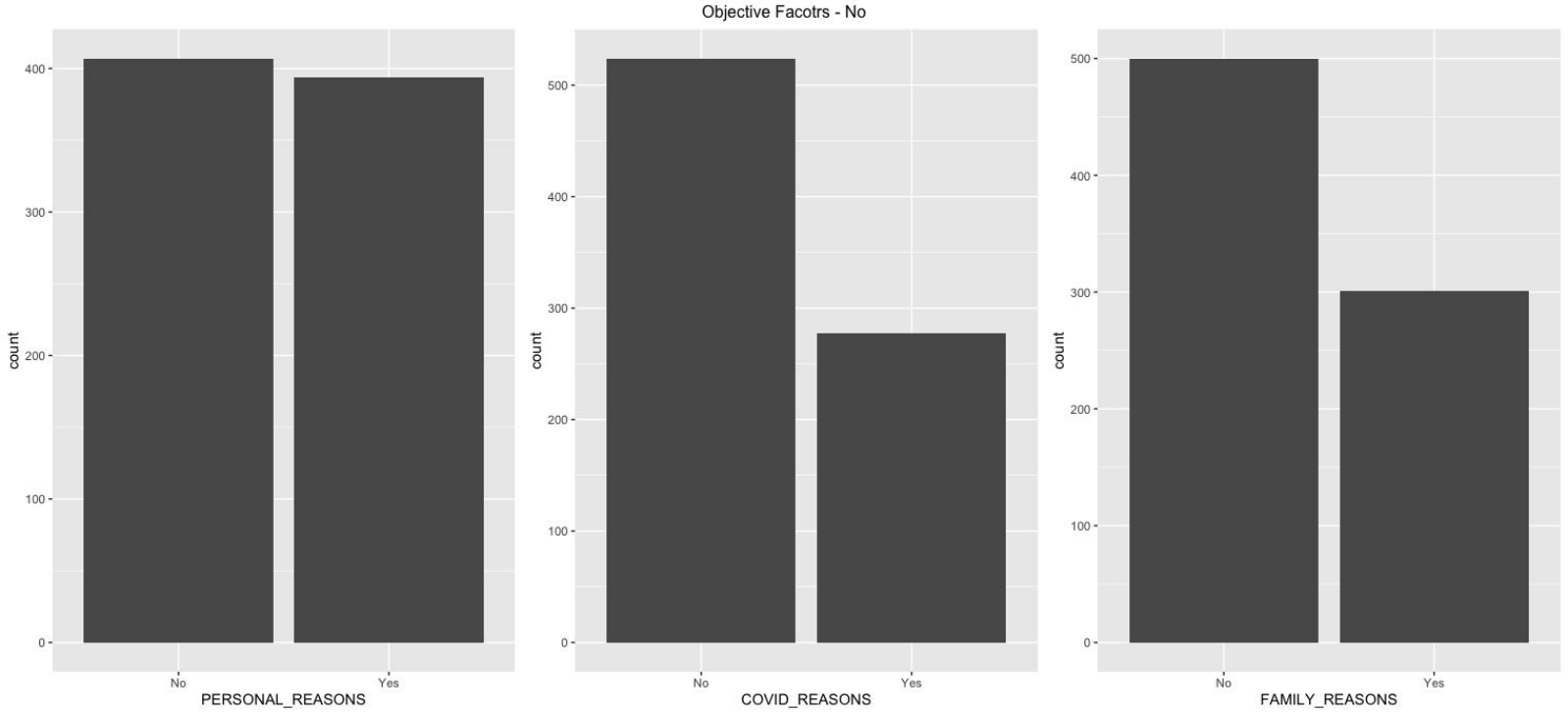
Univariate Analysis - Factors Attributes

Objective Factors - VOTING_INTENTION = Yes



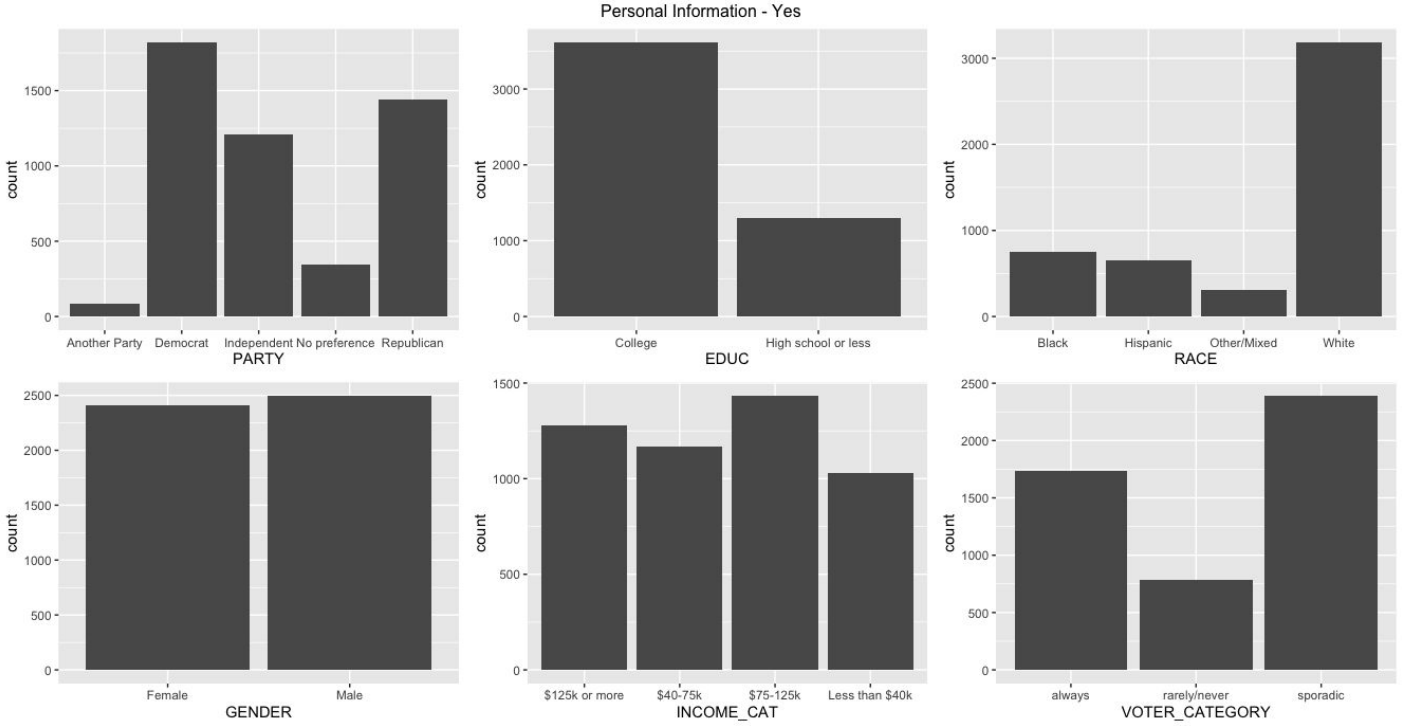
Univariate Analysis - Factors Attributes

Objective Factors - VOTING_INTENTION = No



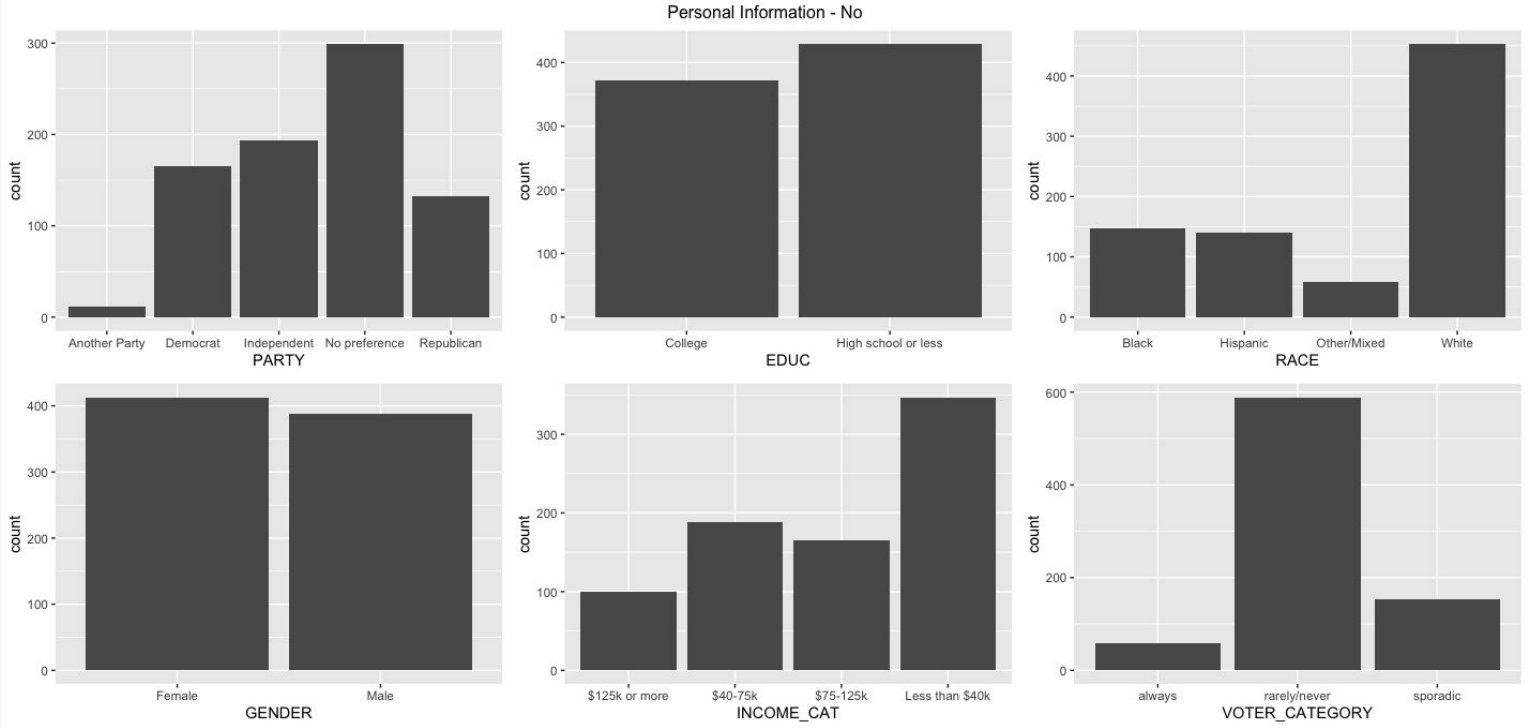
Univariate Analysis - Factors Attributes

Personal Information - VOTING_INTENTION = Yes



Univariate Analysis - Factors Attributes

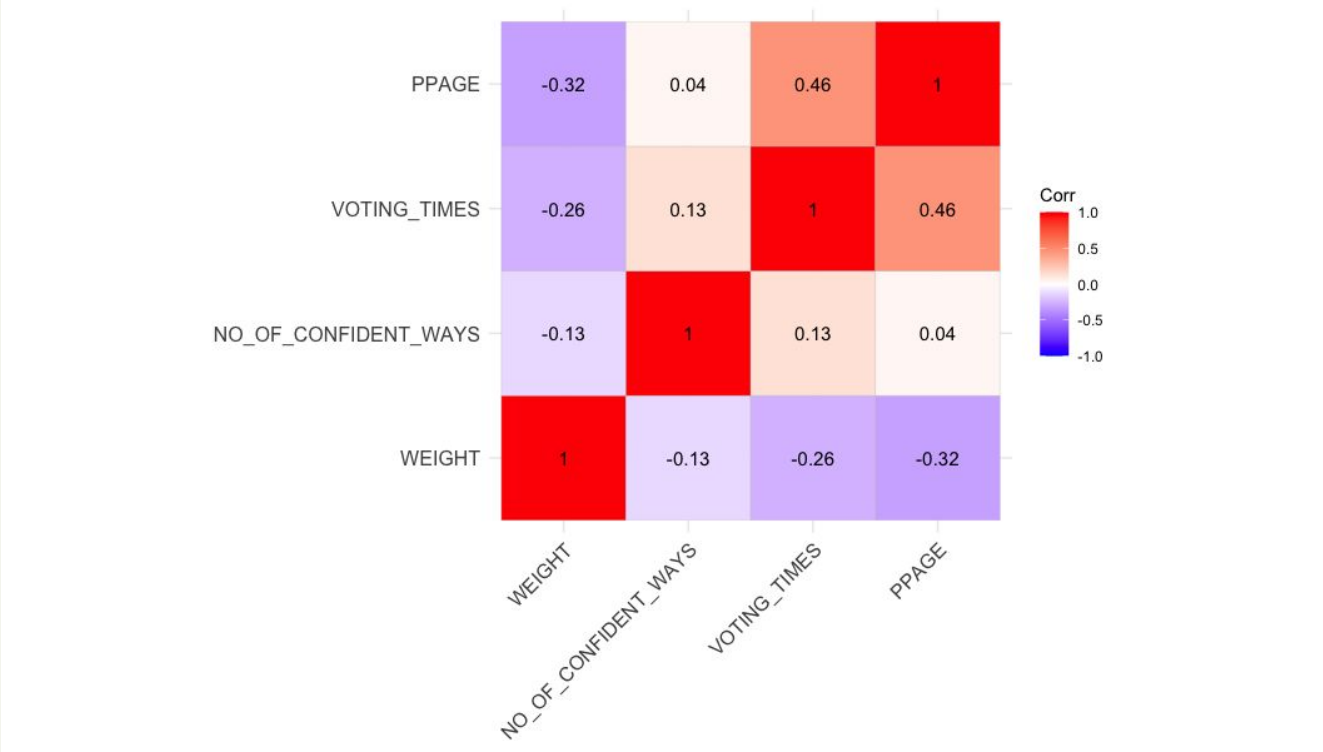
Personal Information - VOTING_INTENTION = No



Bivariate Analysis

This part will be the bivariate analysis between Categories (factors) attributes and Measures (numerics) attributes.

Bivariate Analysis - Measures vs Measures

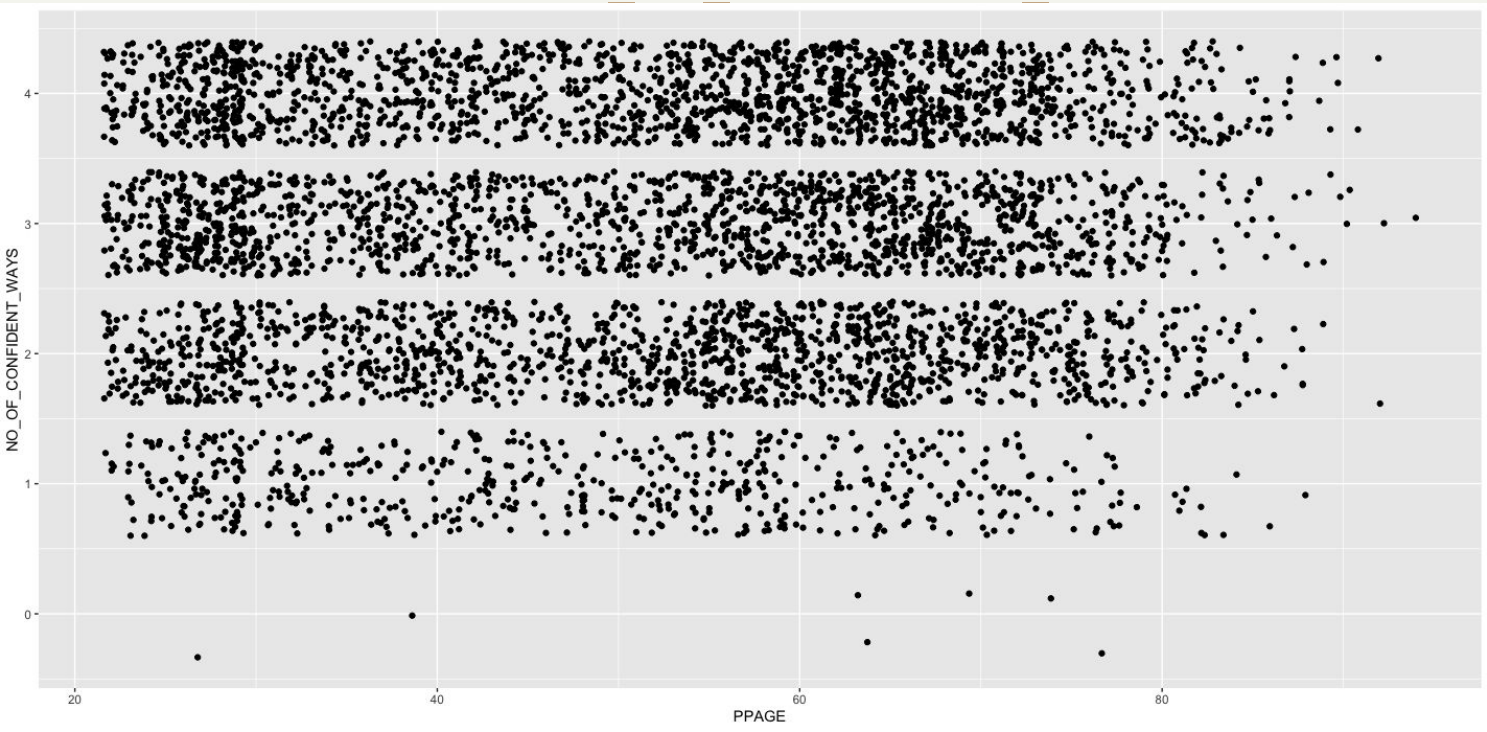


Bivariate Analysis - Measures vs Measures

PPAGE vs VOTING_TIMES

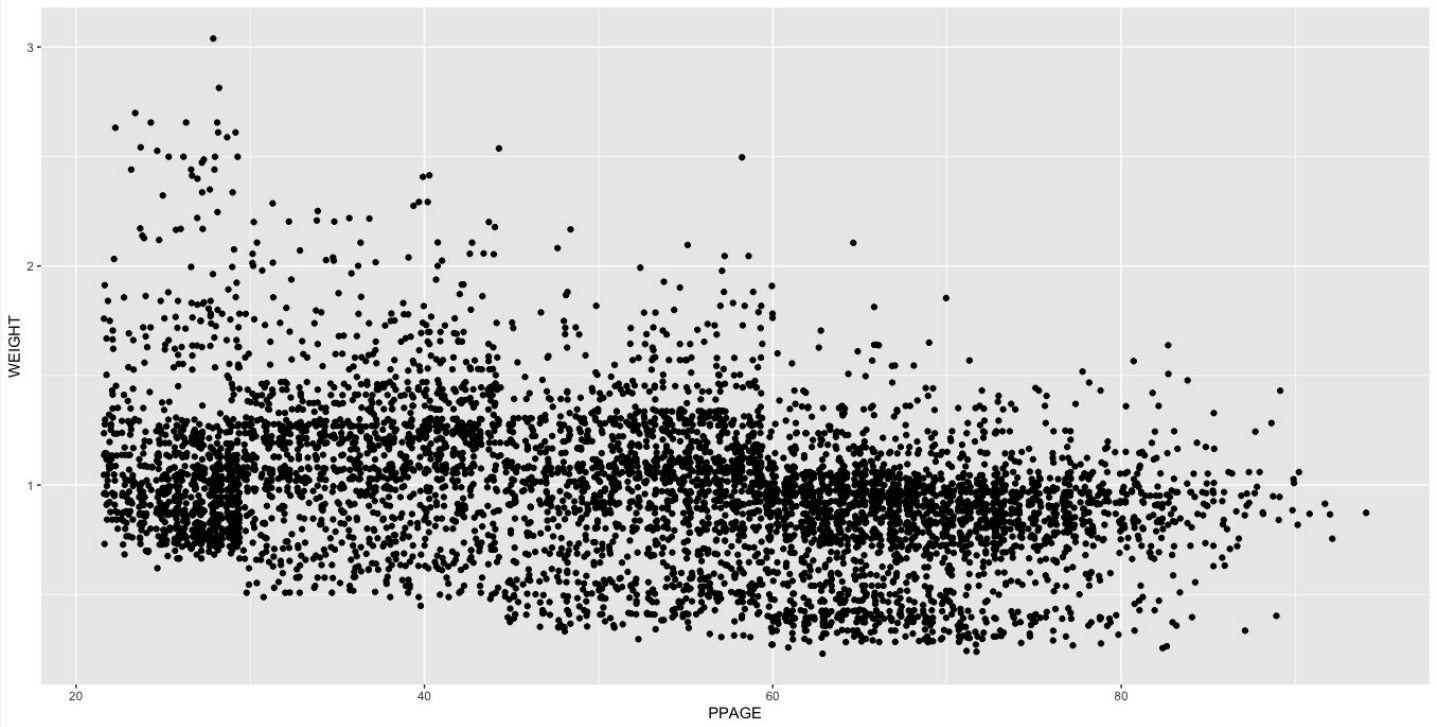
Bivariate Analysis - Measures vs Measures

PPAGE vs NO_OF_CONFIDENT_WAYS



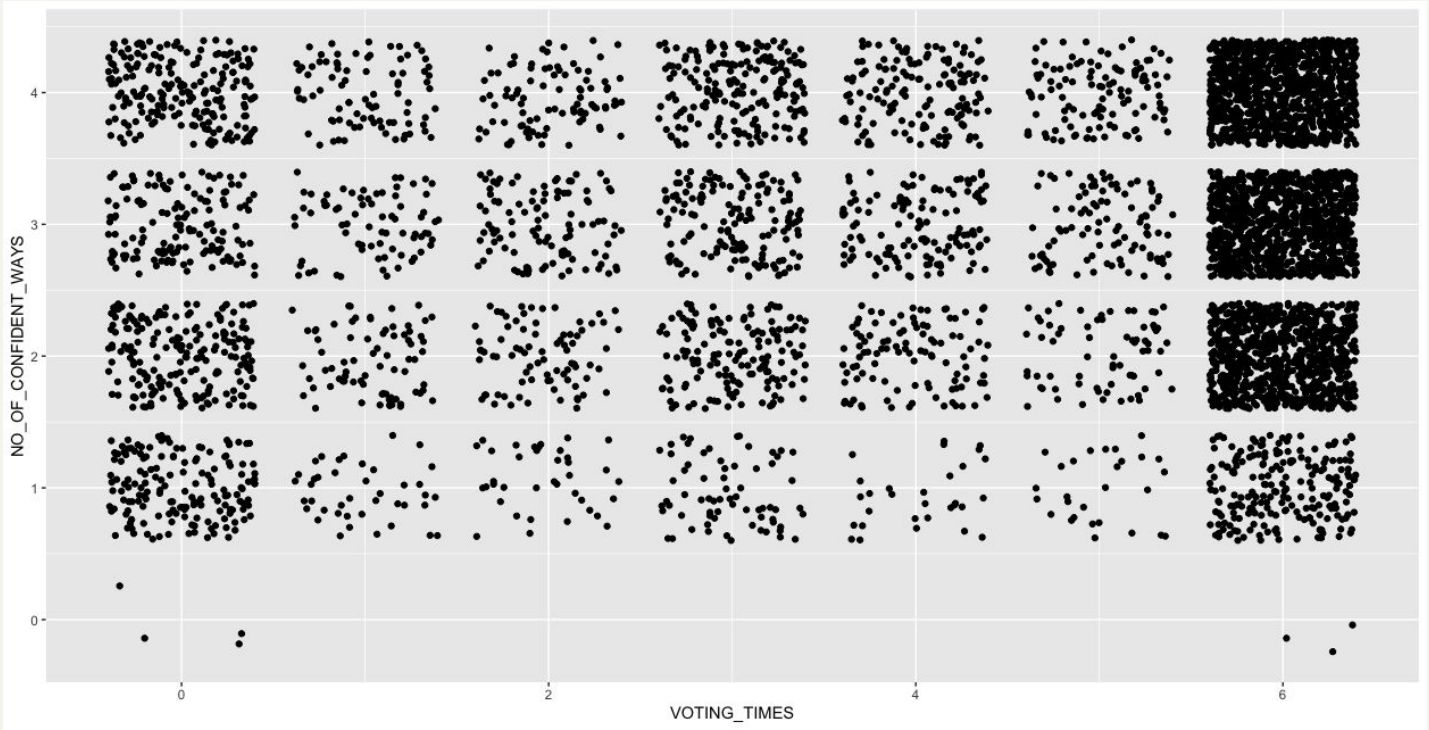
Bivariate Analysis - Measures vs Measures

PPAGE vs WEIGHT



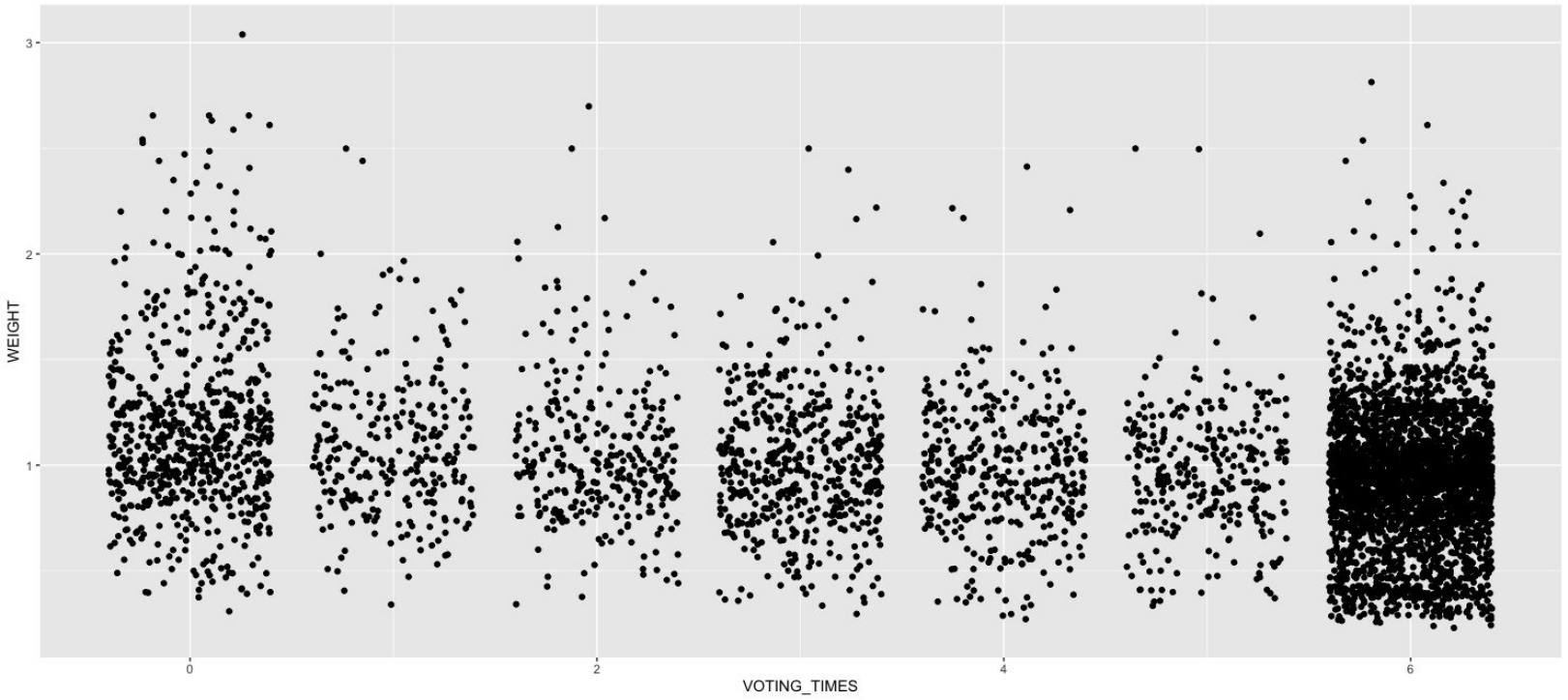
Bivariate Analysis - Measures vs Measures

VOTING_TIMES vs NO_OF_CONFIDENT_WAYS



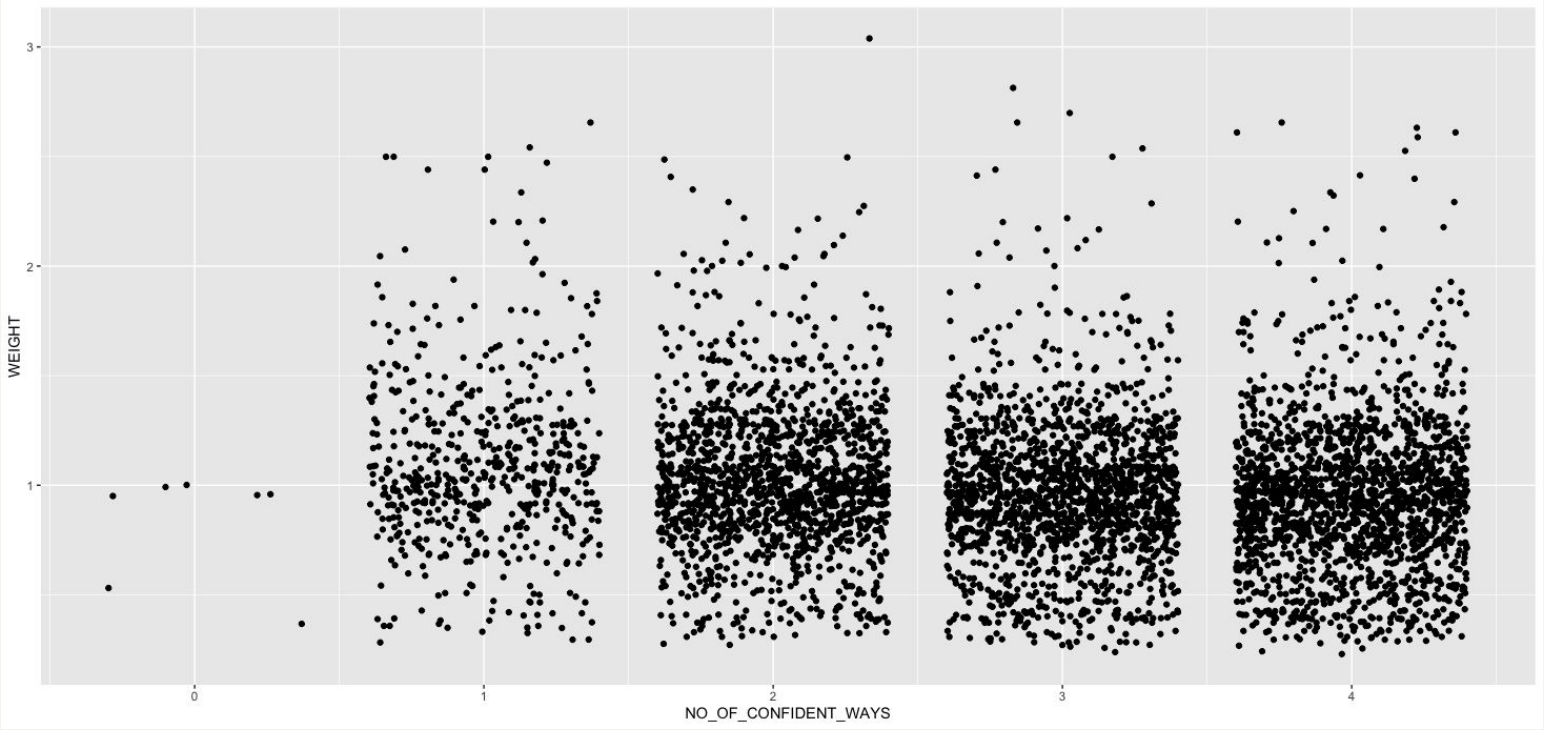
Bivariate Analysis - Measures vs Measures

VOTING_TIMES vs WEIGHT



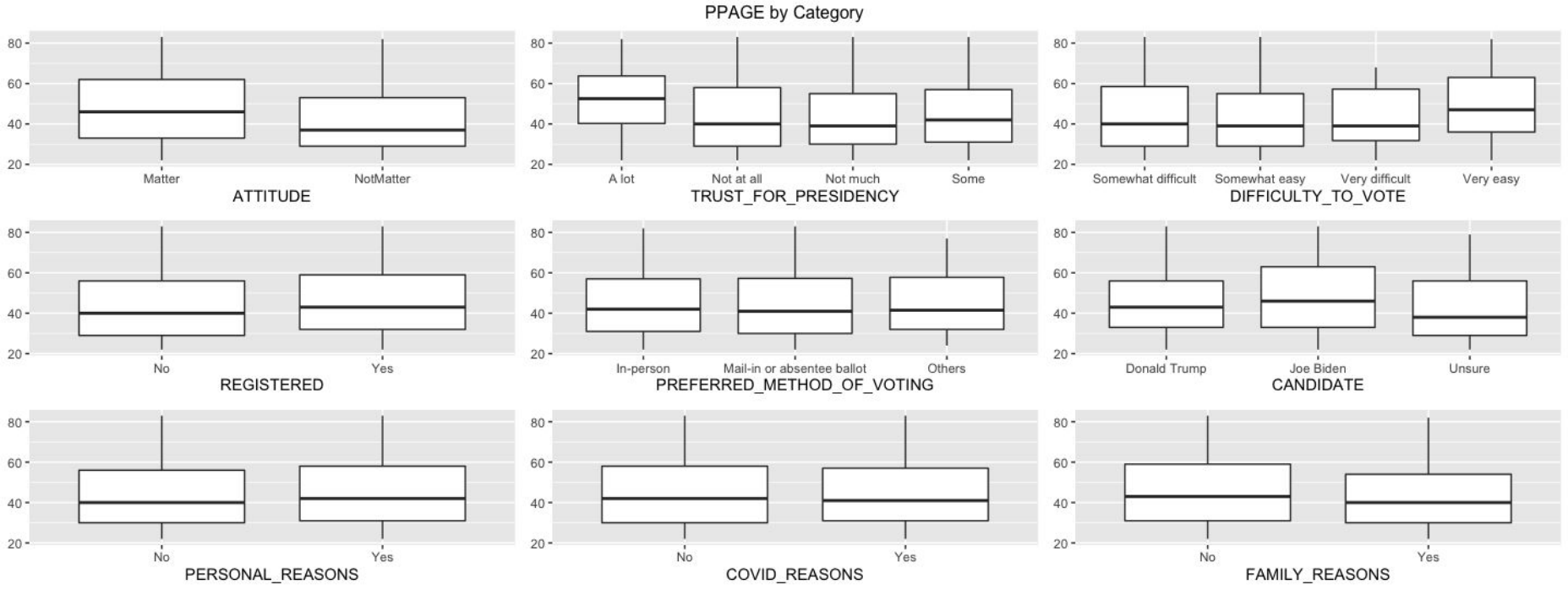
Bivariate Analysis - Measures vs Measures

NO_OF_CONFIDENT_WAYS vs WEIGHT



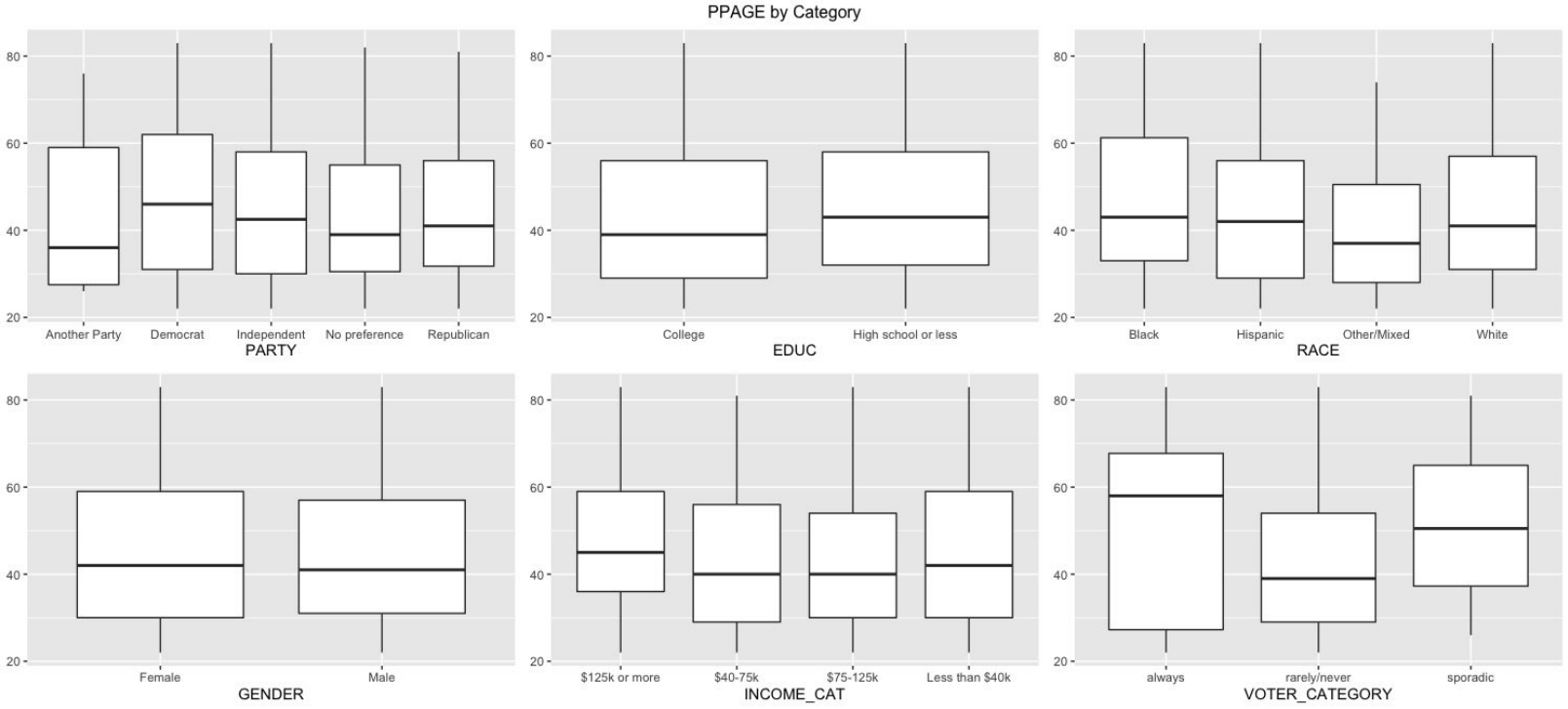
Bivariate Analysis - Categories vs Measures

PPAGE by Categories



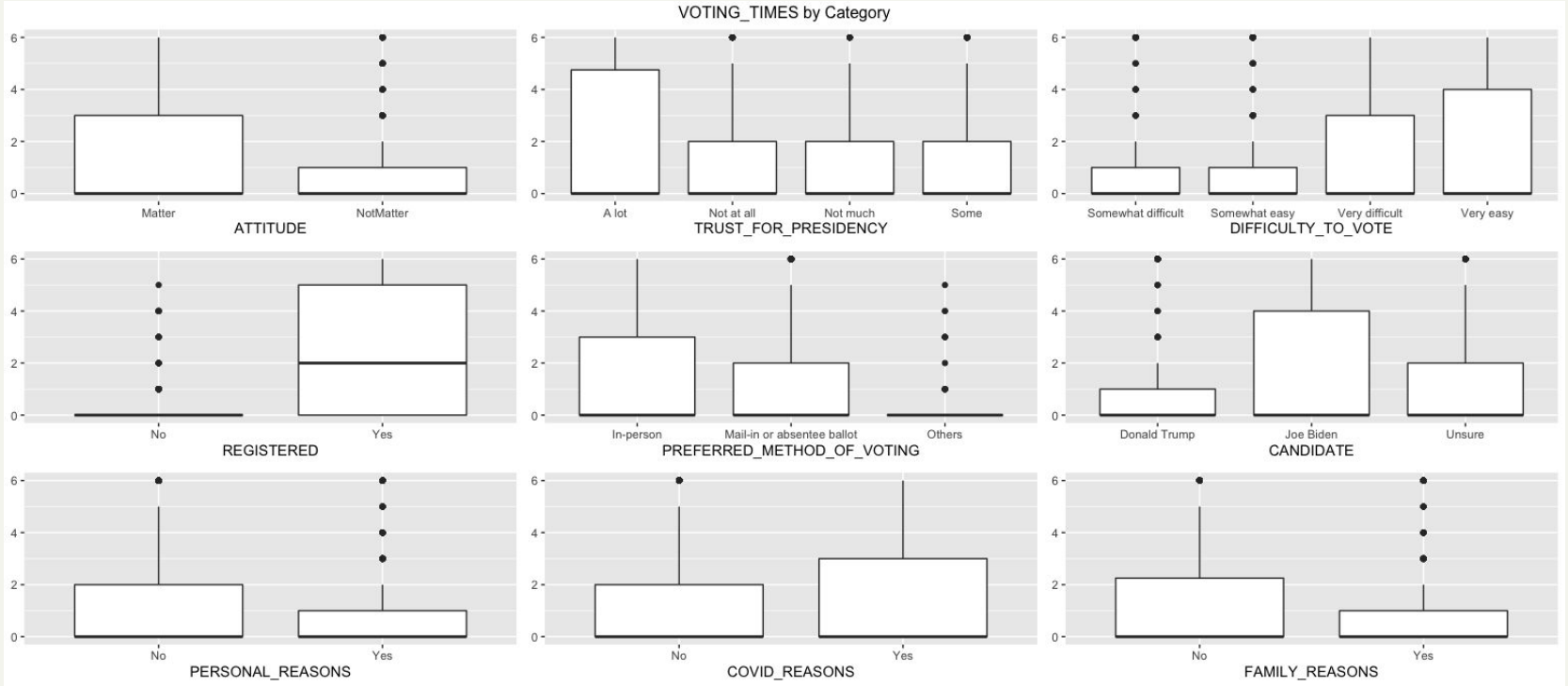
Bivariate Analysis - Categories vs Measures

PPAGE by Categories



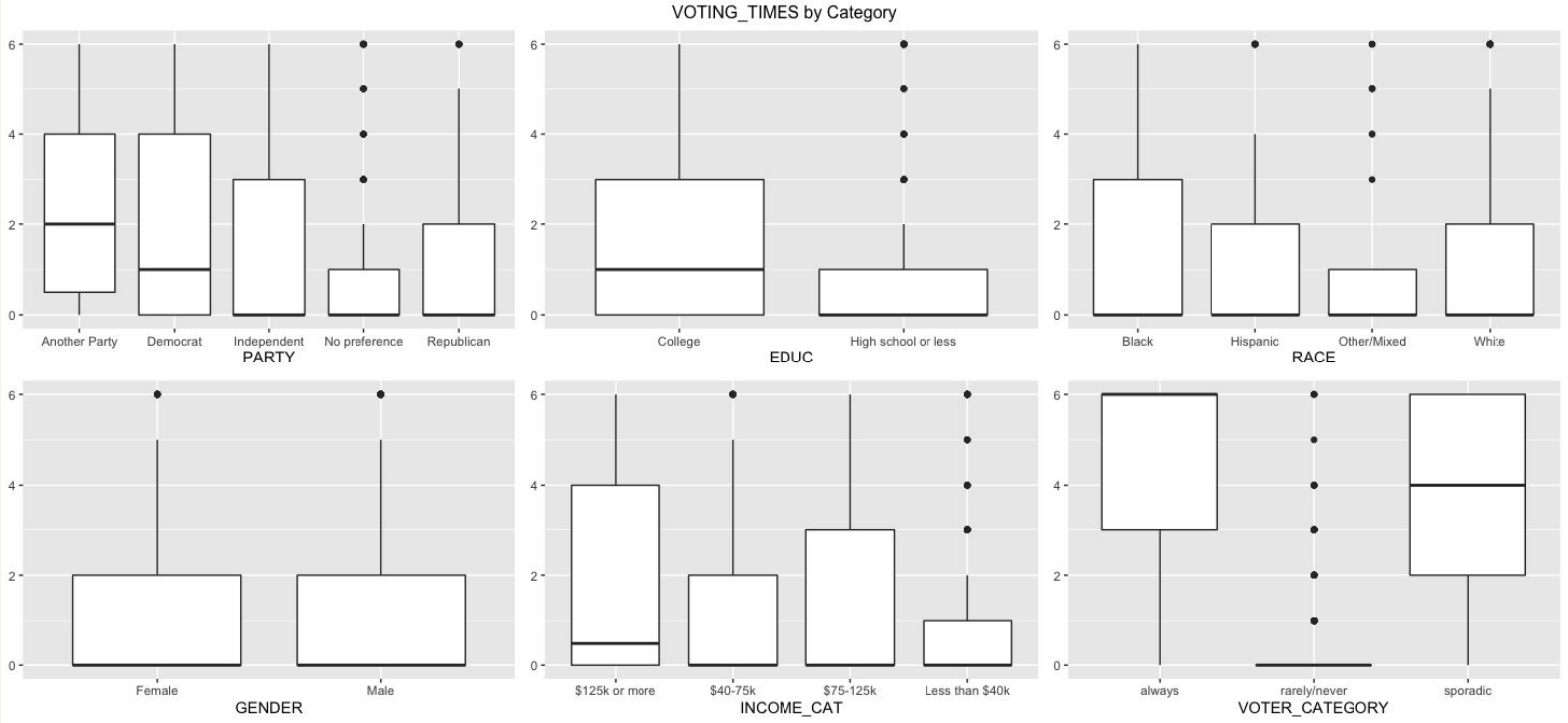
Bivariate Analysis - Categories vs Measures

VOTING_TIMES by Categories



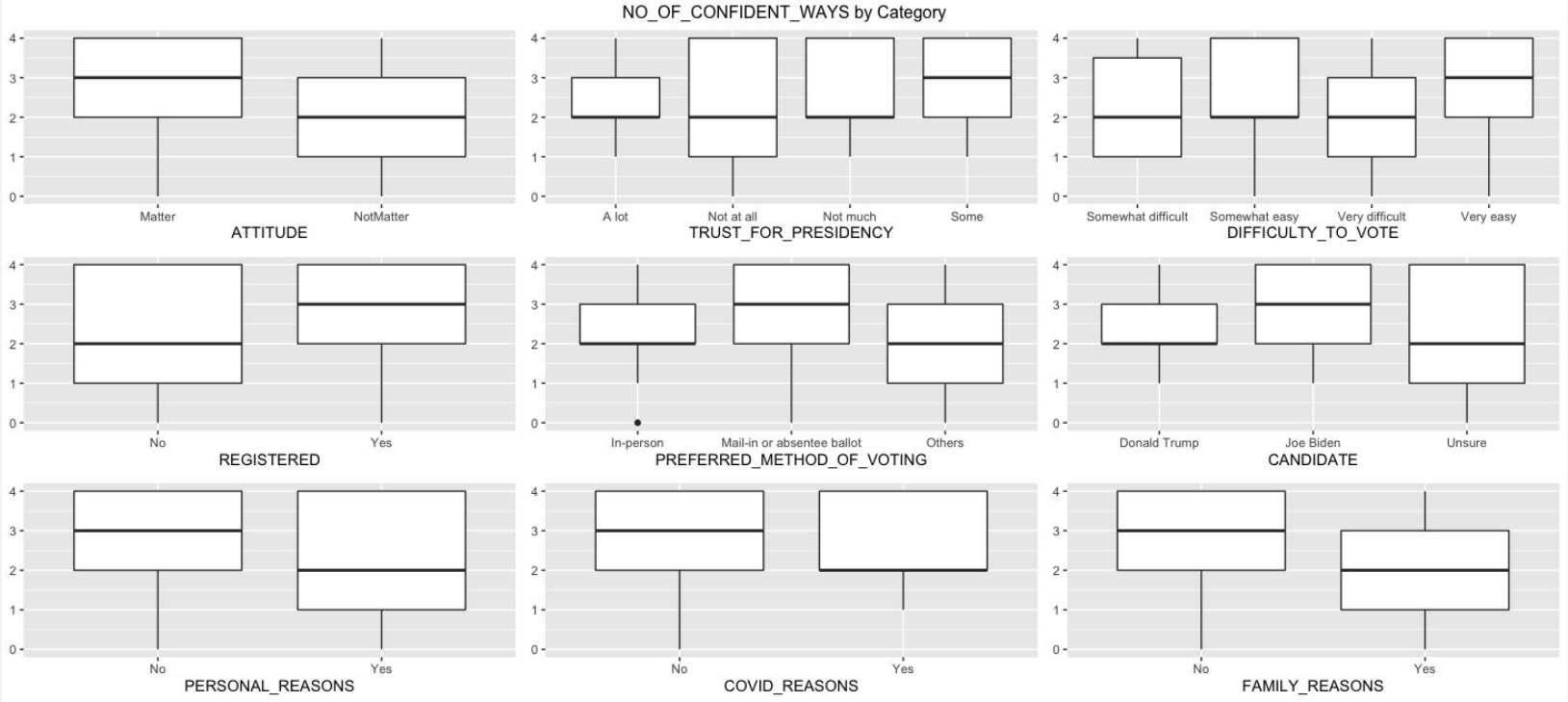
Bivariate Analysis - Categories vs Measures

VOTING_TIMES by Categories



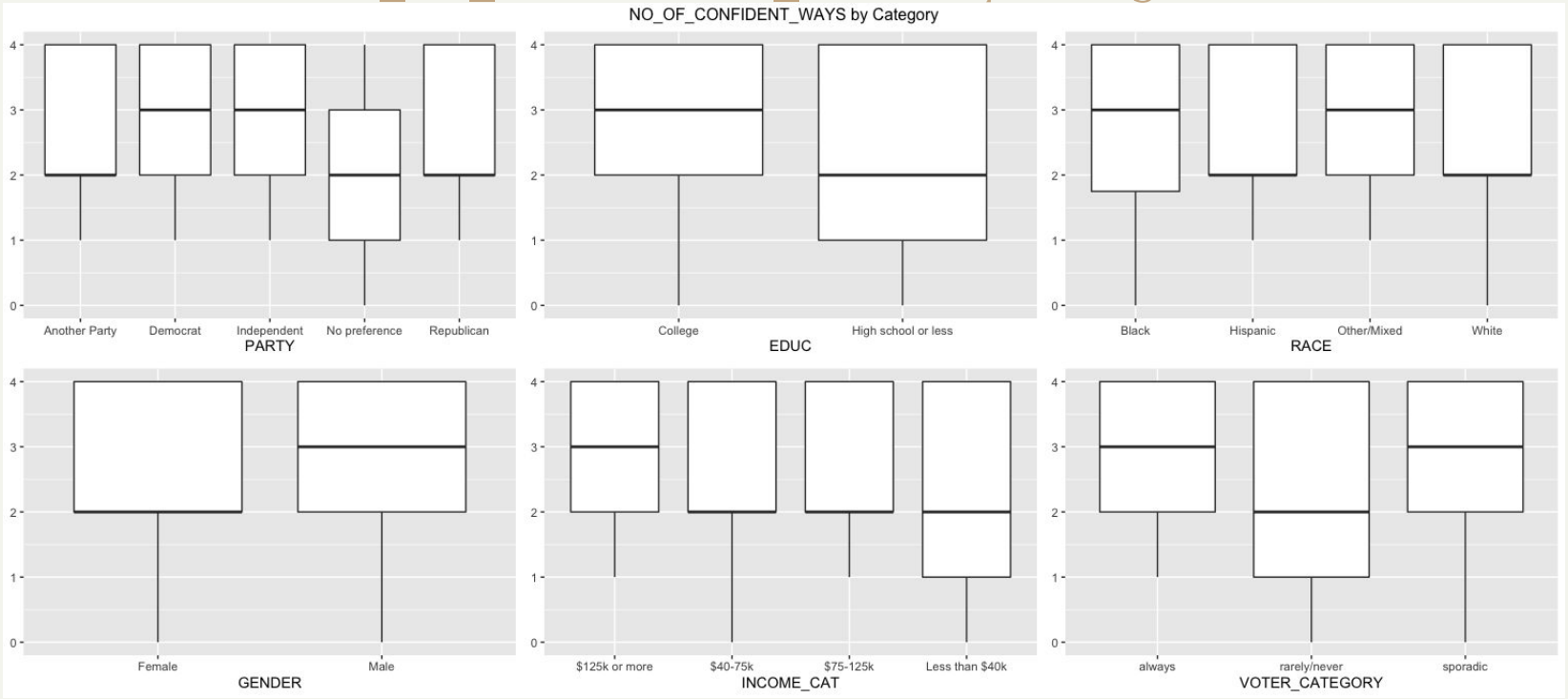
Bivariate Analysis - Categories vs Measures

NO_OF_CONFIDENT_WAYS by Categories



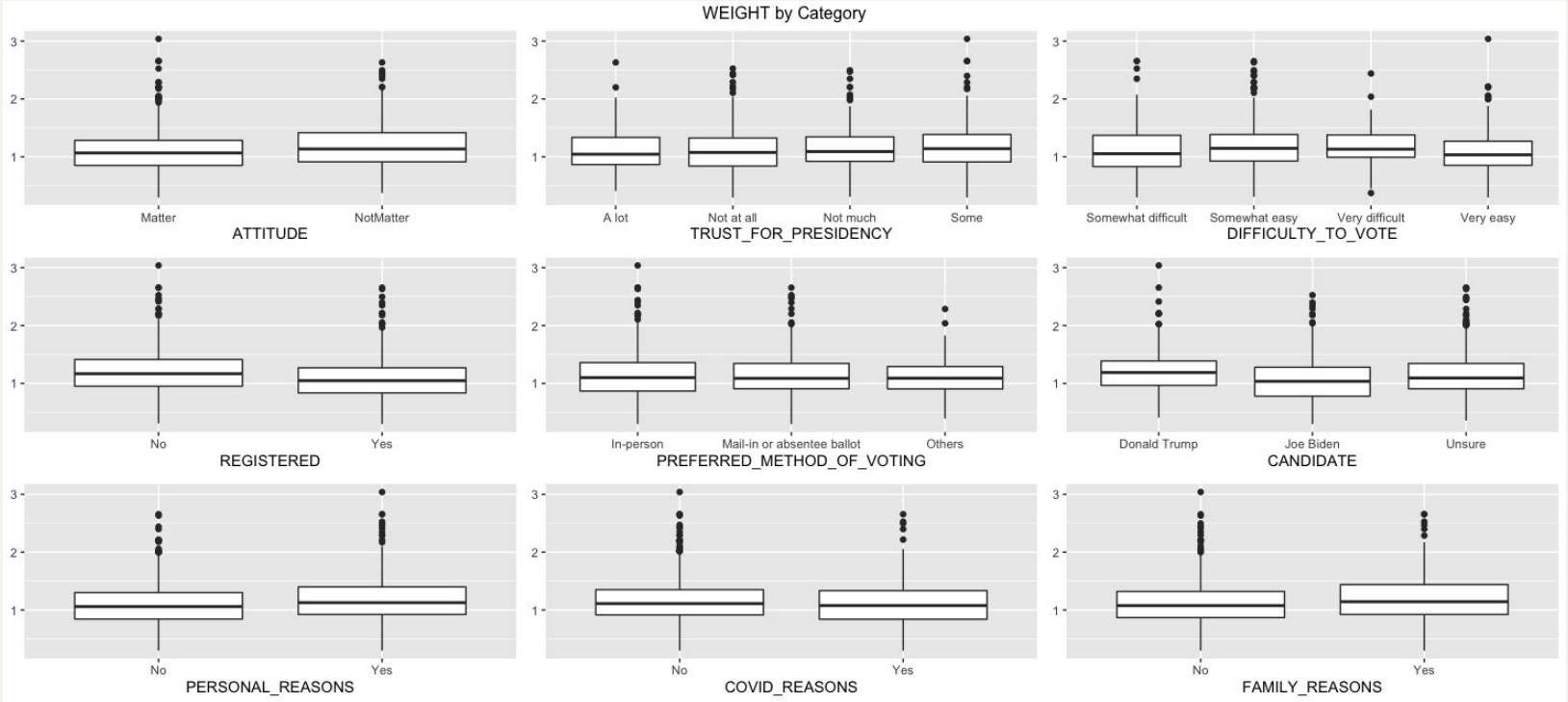
Bivariate Analysis - Categories vs Measures

NO_OF_CONFIDENT_WAYS by Categories



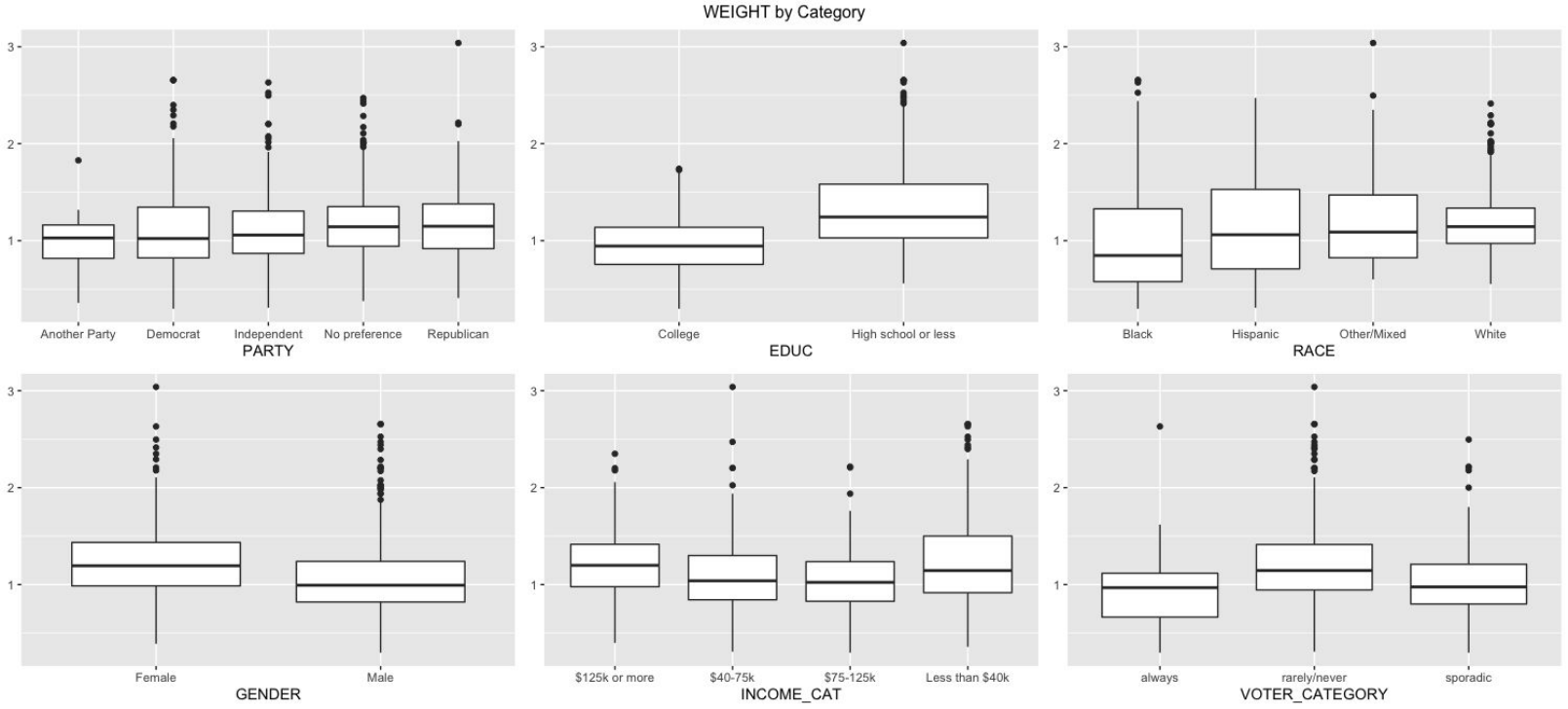
Bivariate Analysis - Categories vs Measures

WEIGHT by Categories

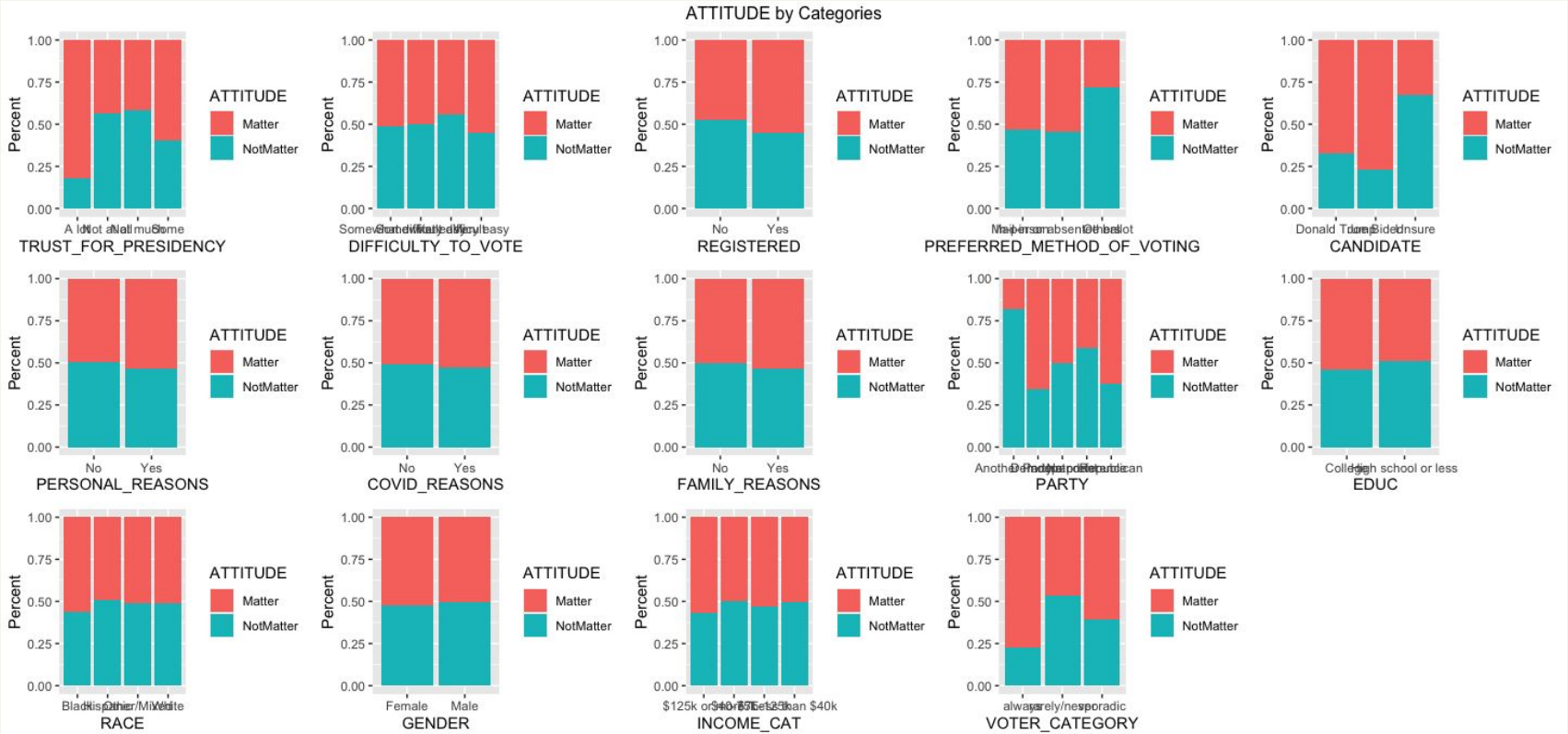


Bivariate Analysis - Categories vs Measures

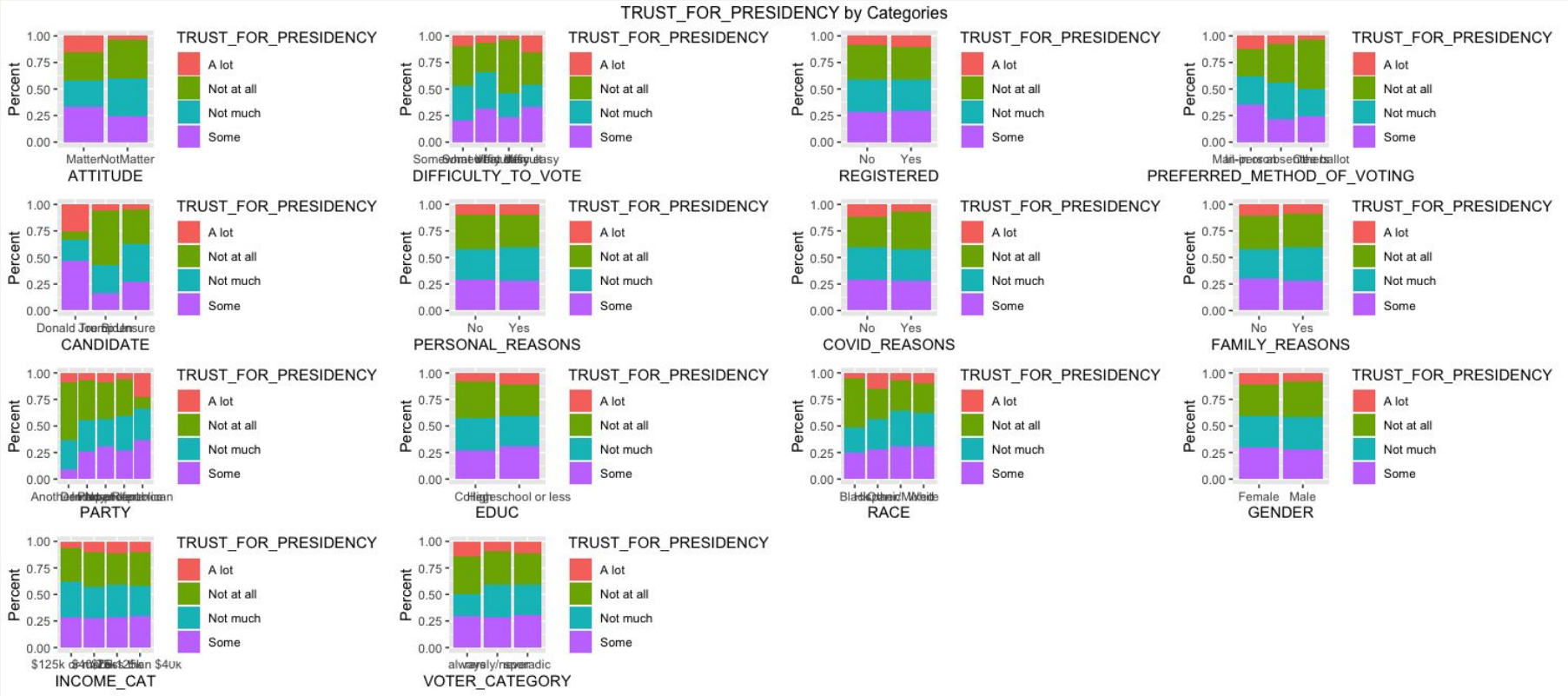
WEIGHT by Categories



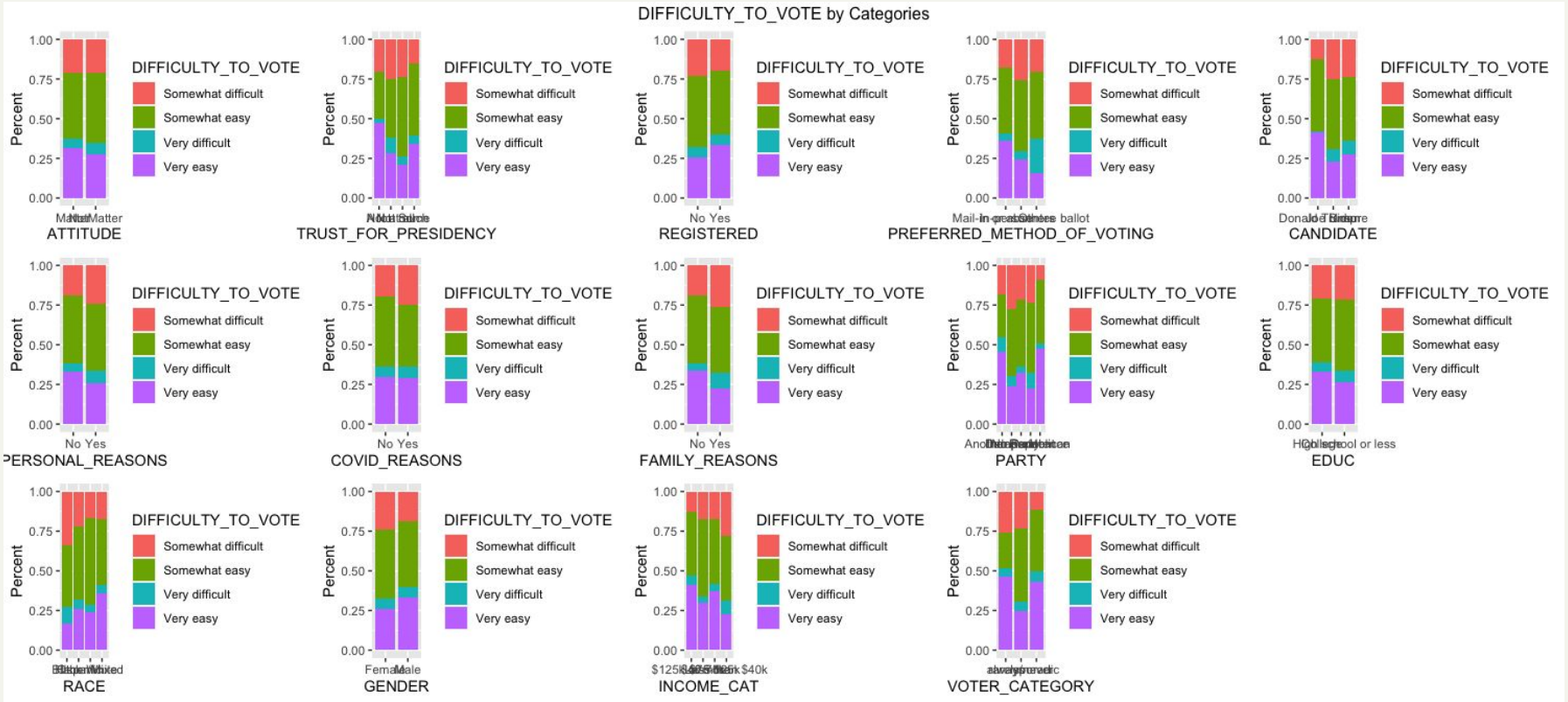
Bivariate Analysis - Categories vs Categories



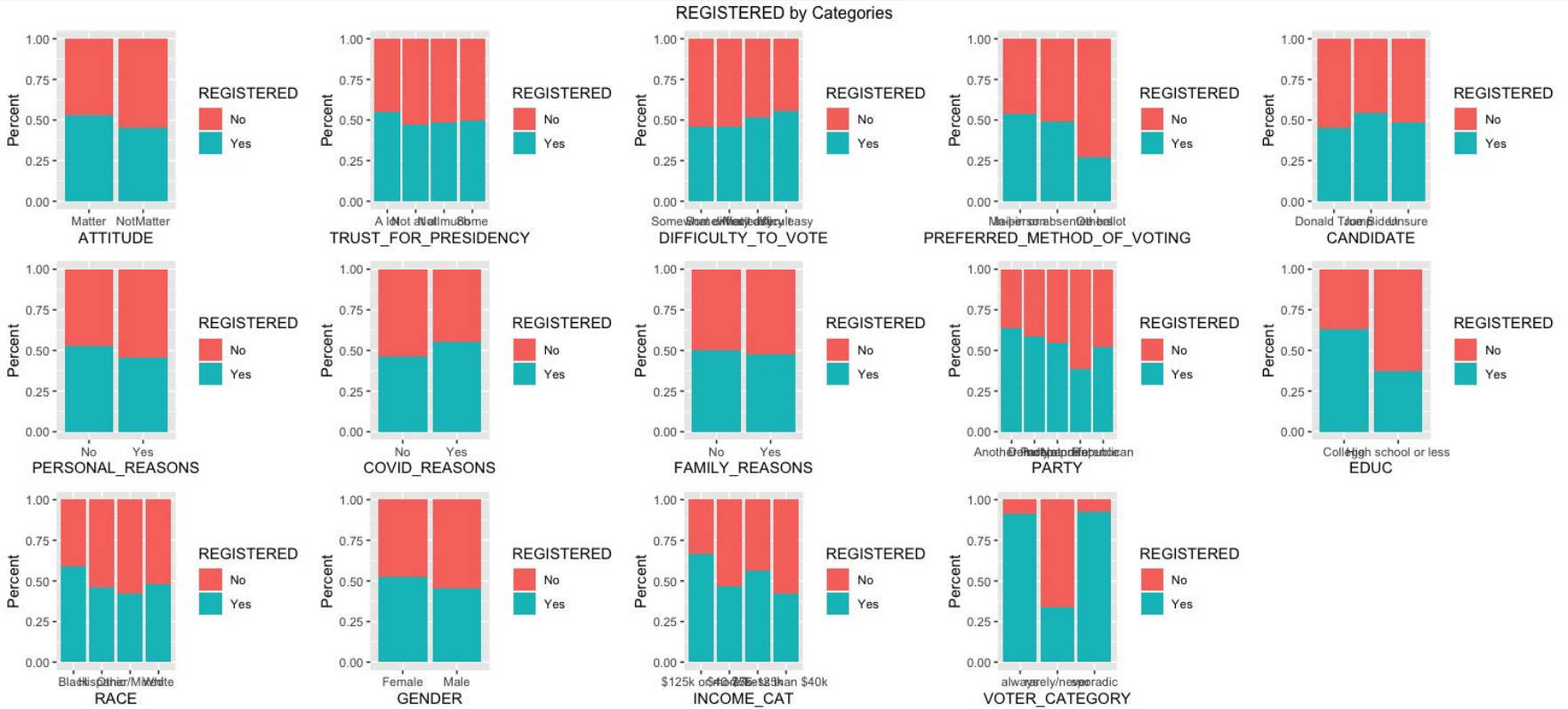
Bivariate Analysis - Categories vs Categories



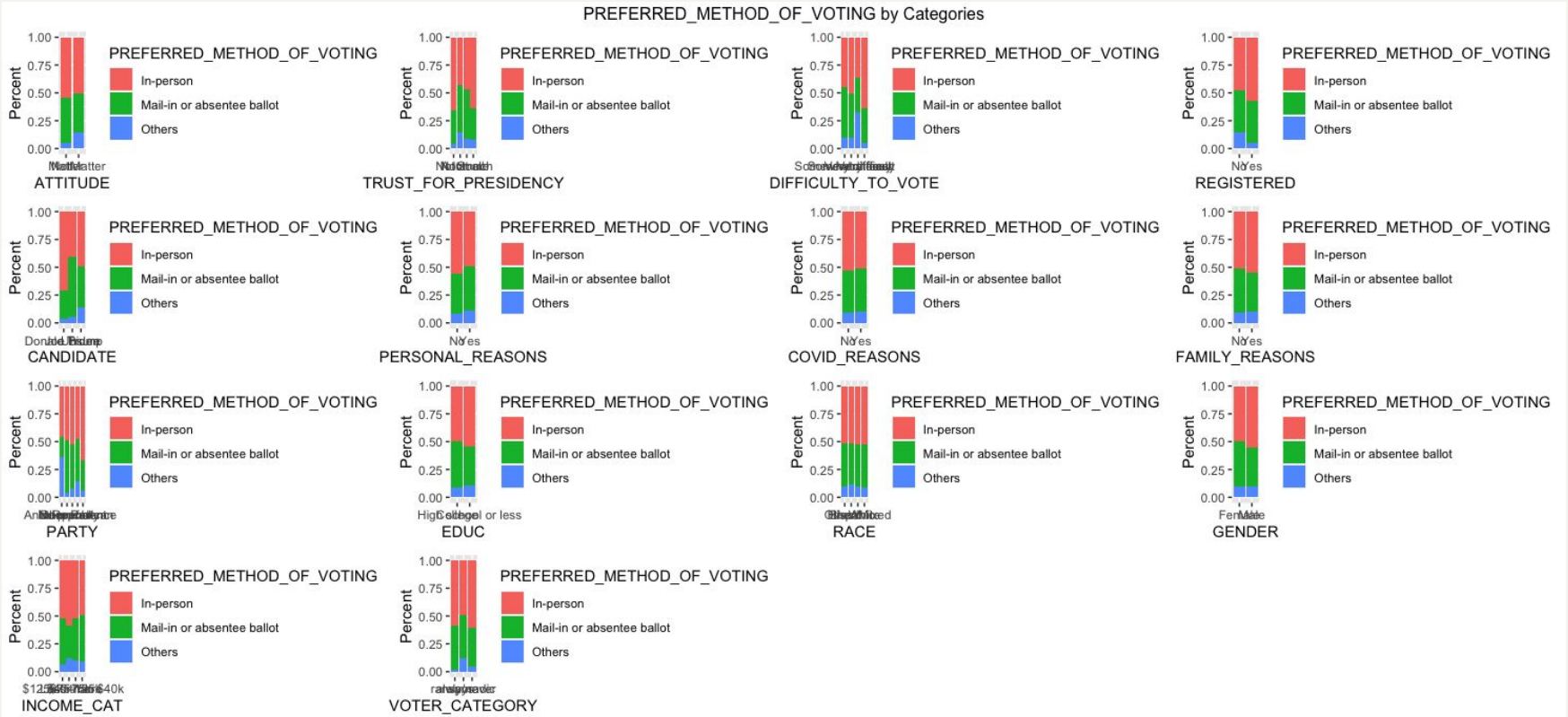
Bivariate Analysis - Categories vs Categories



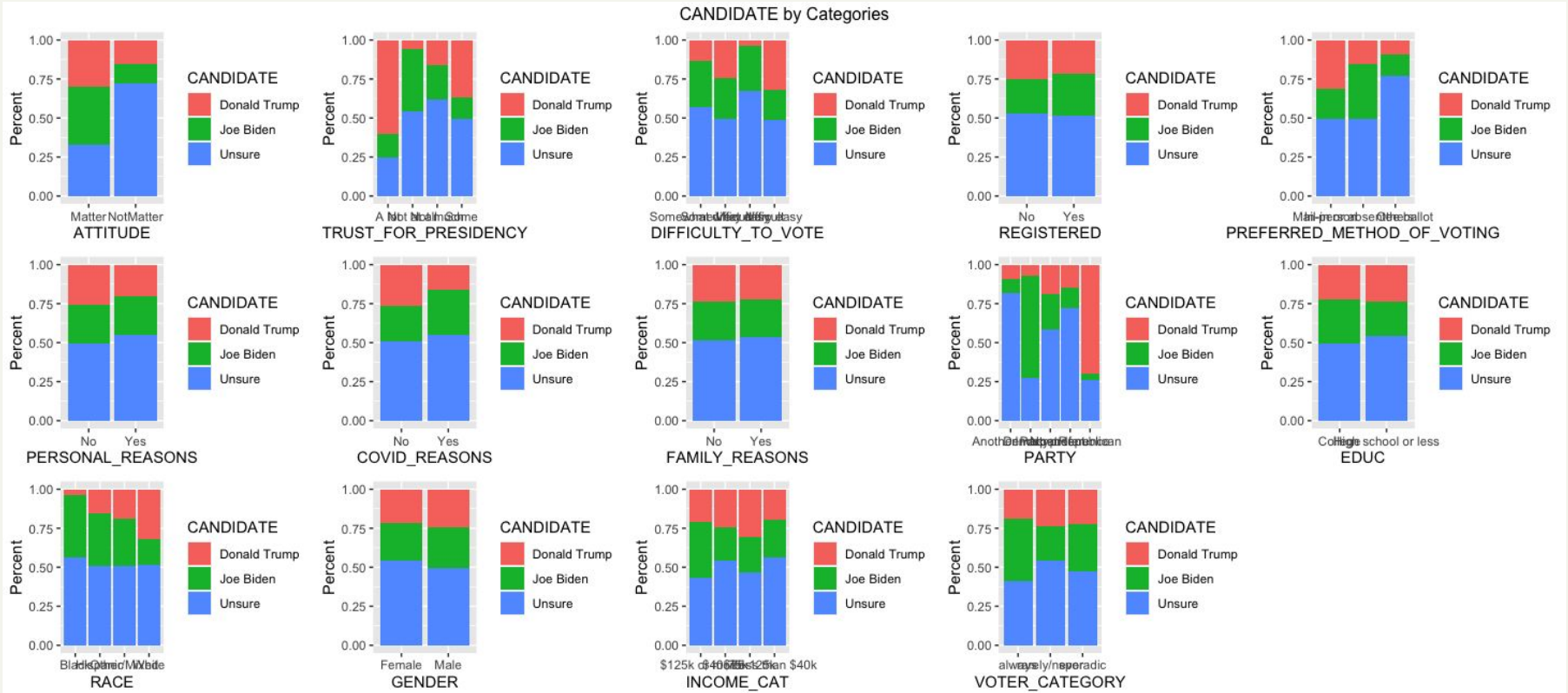
Bivariate Analysis - Categories vs Categories



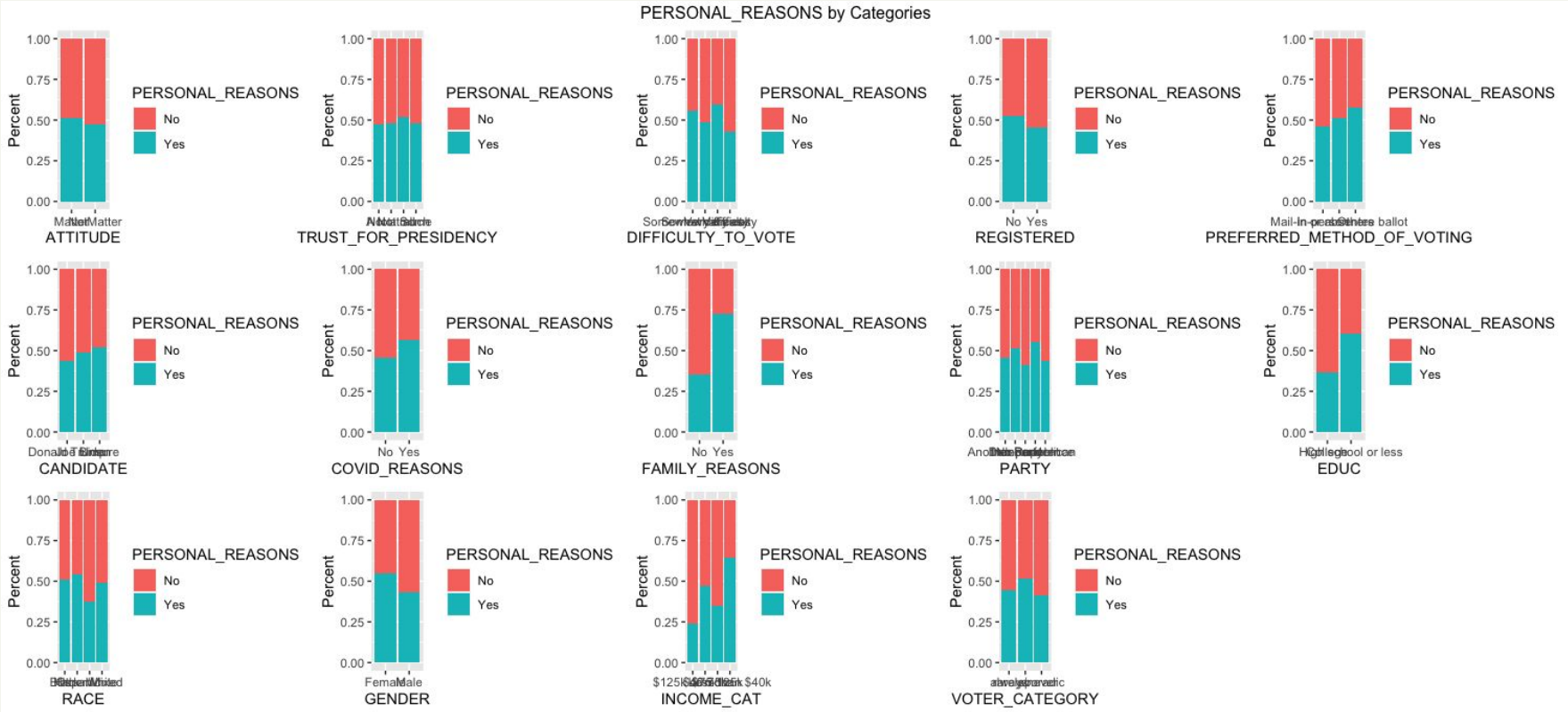
Bivariate Analysis - Categories vs Categories



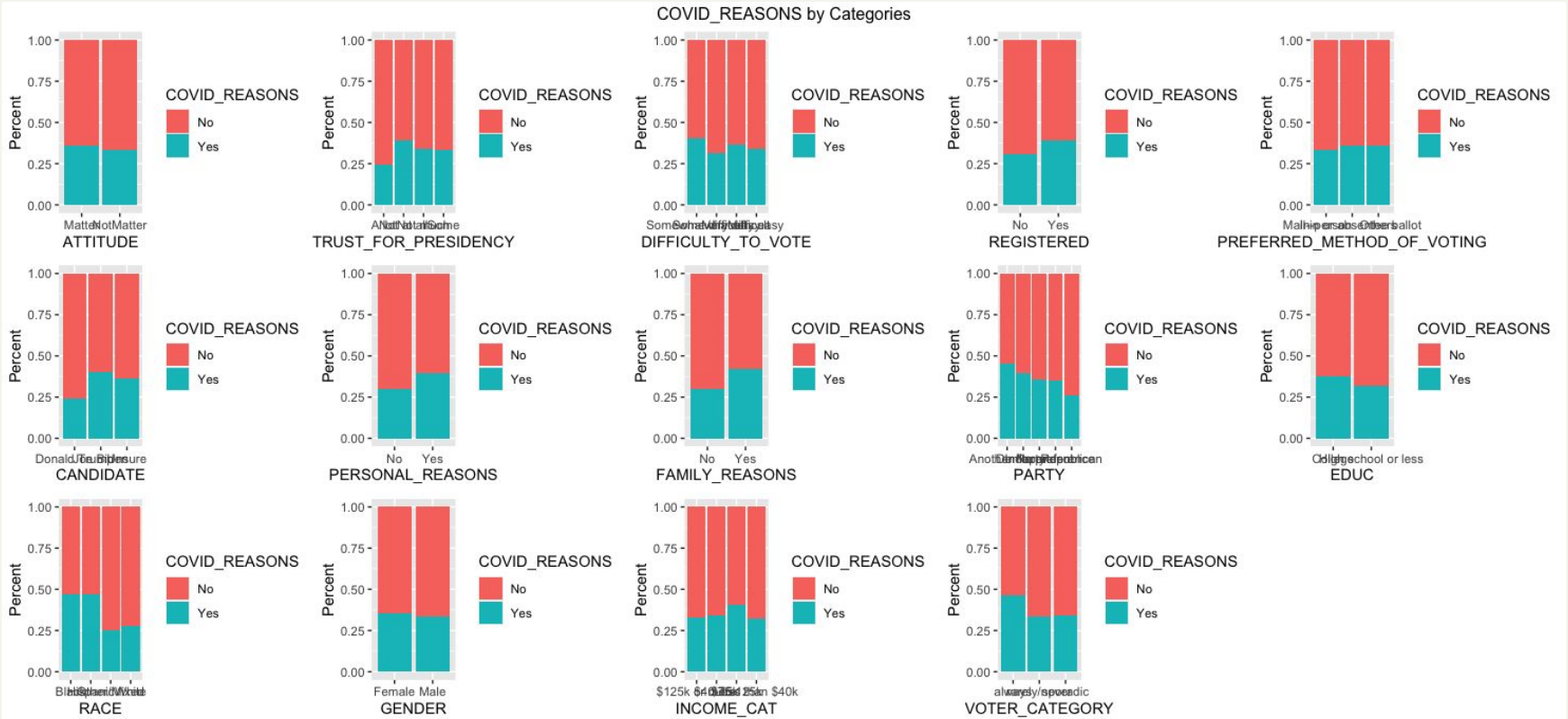
Bivariate Analysis - Categories vs Categories



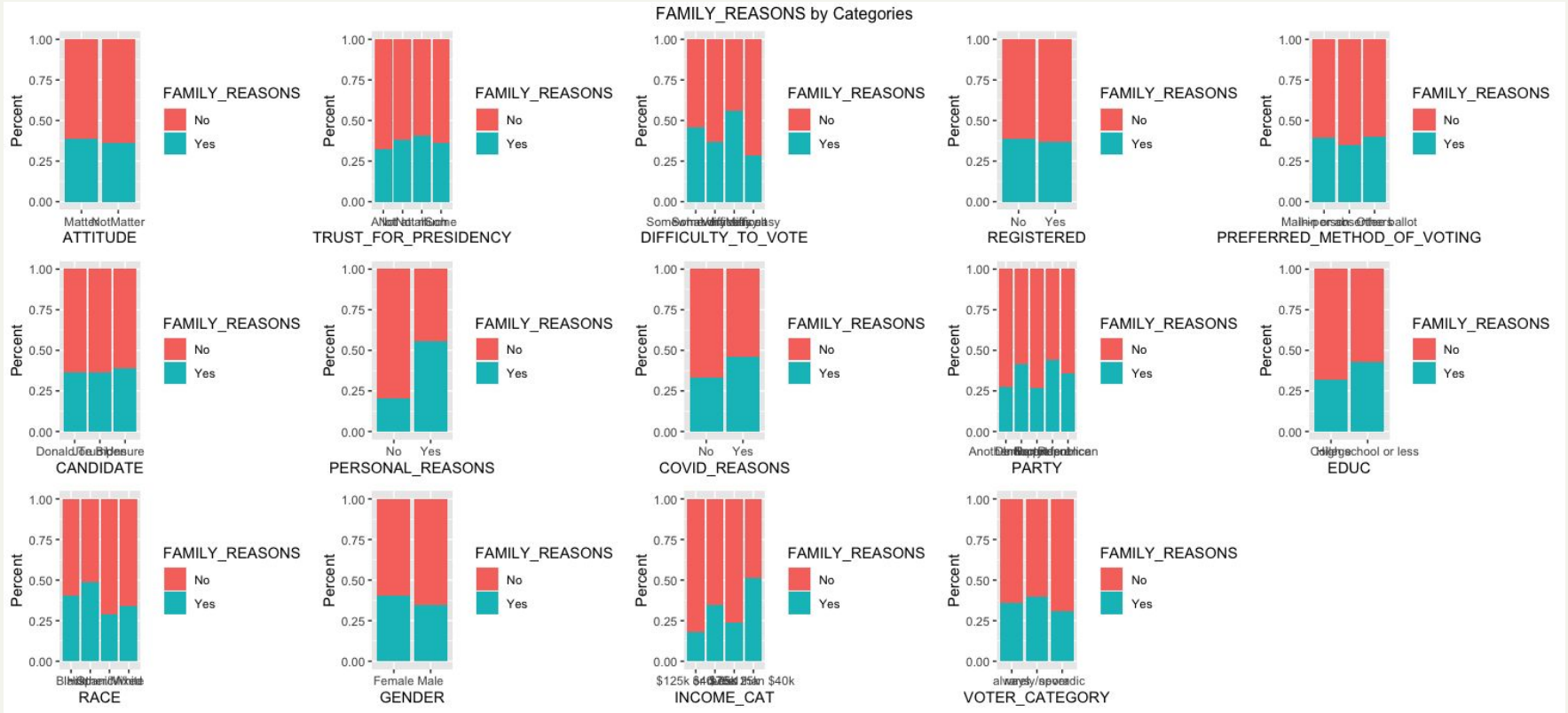
Bivariate Analysis - Categories vs Categories



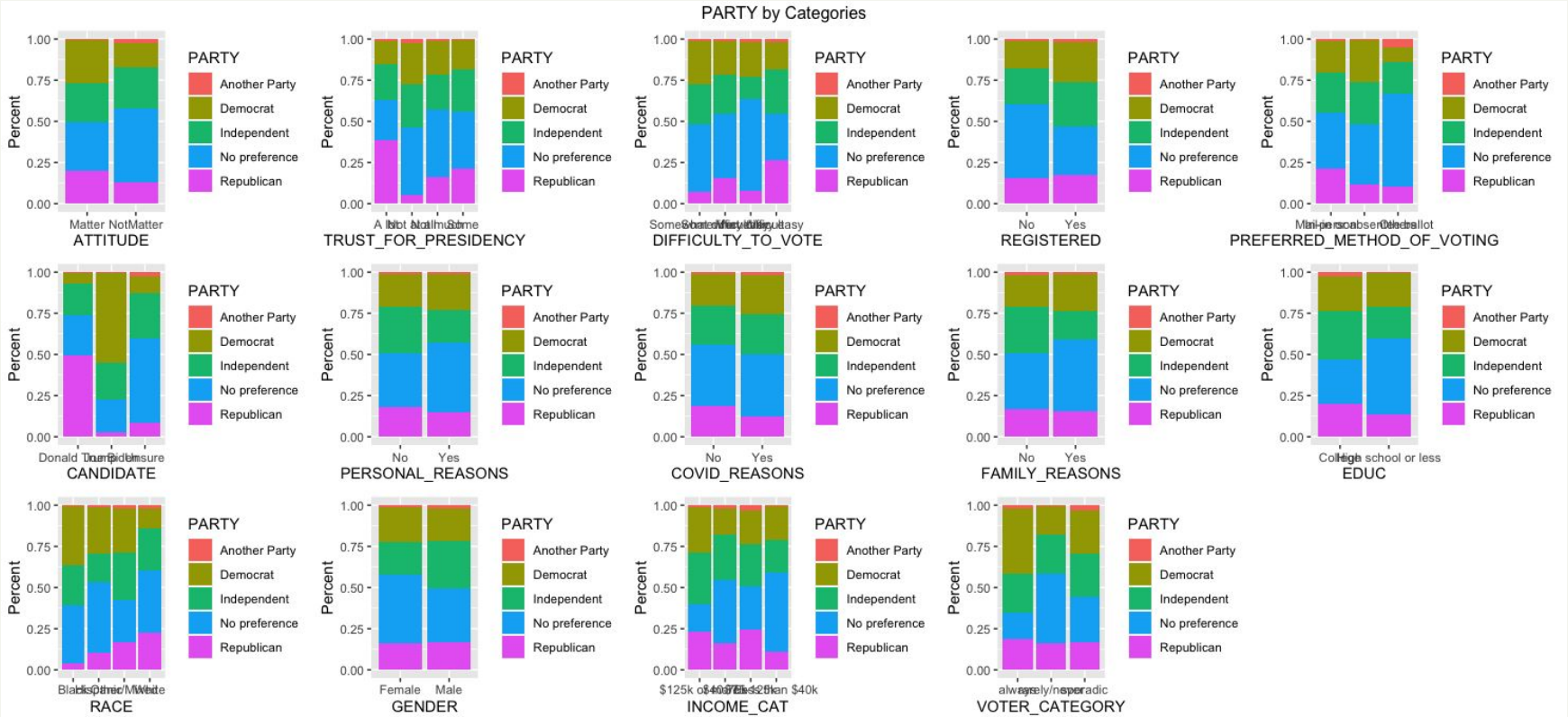
Bivariate Analysis - Categories vs Categories



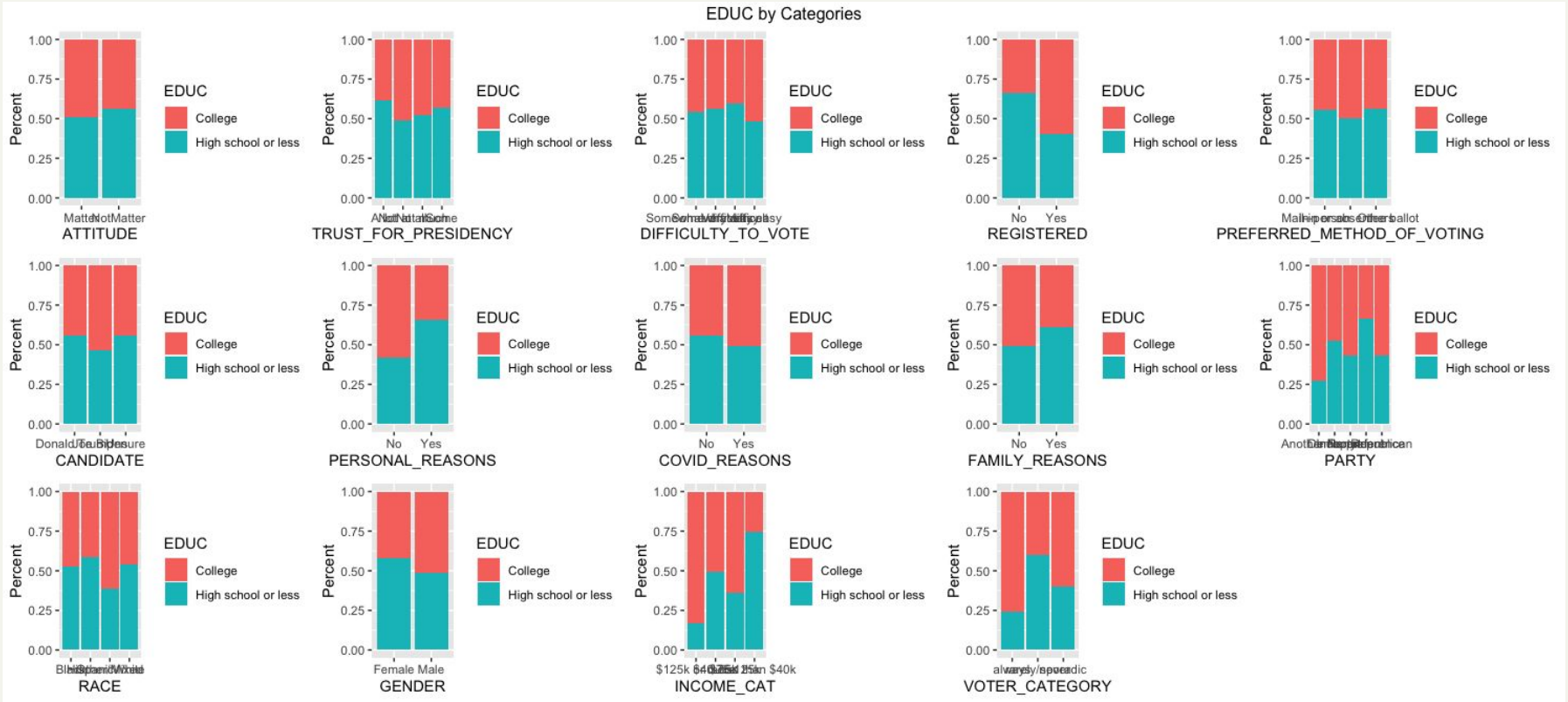
Bivariate Analysis - Categories vs Categories



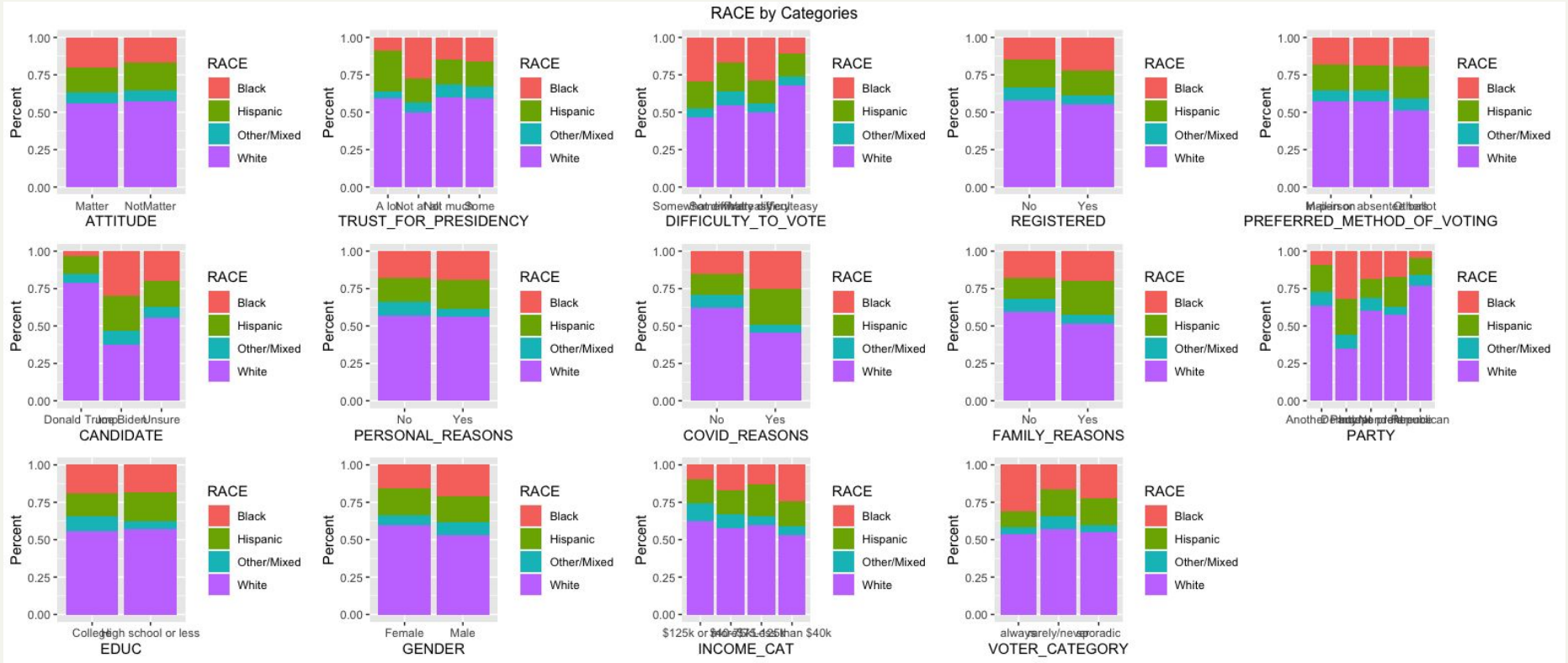
Bivariate Analysis - Categories vs Categories



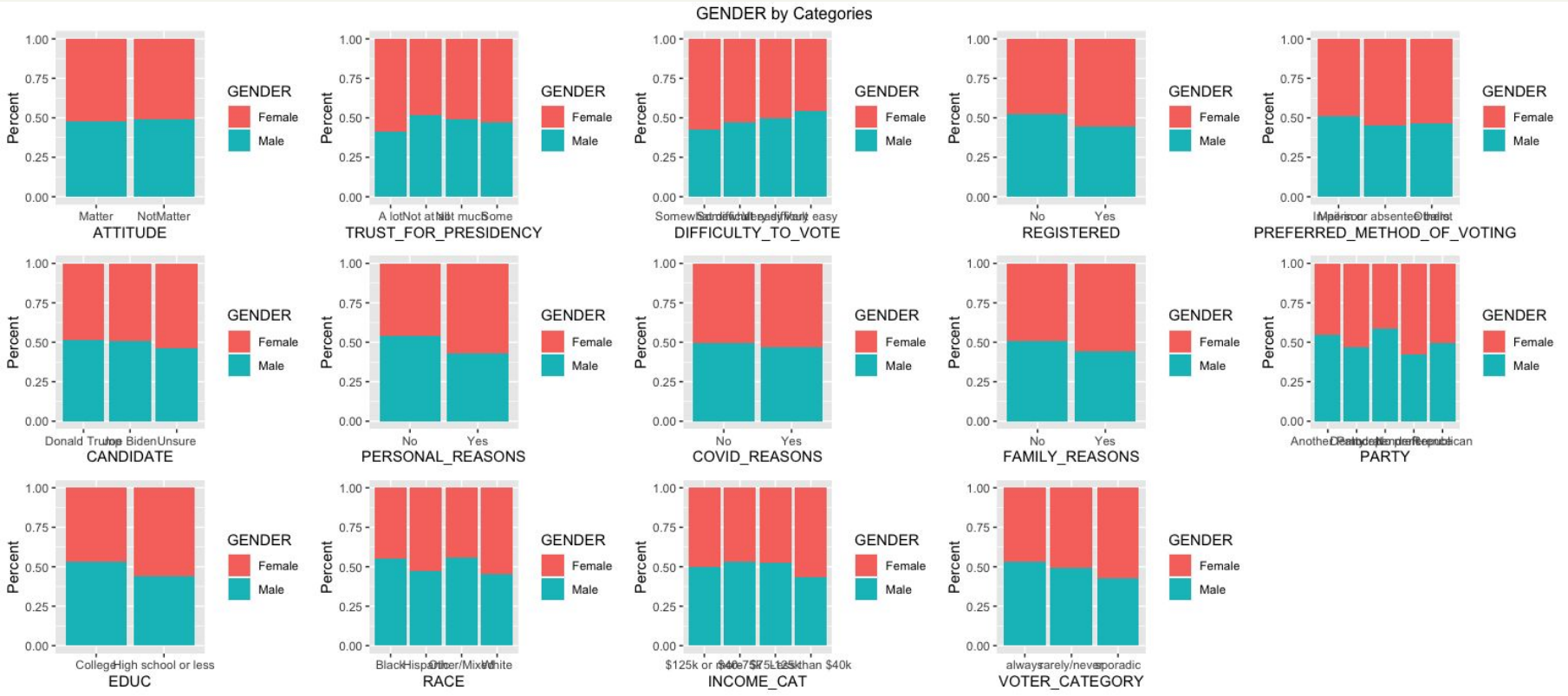
Bivariate Analysis - Categories vs Categories



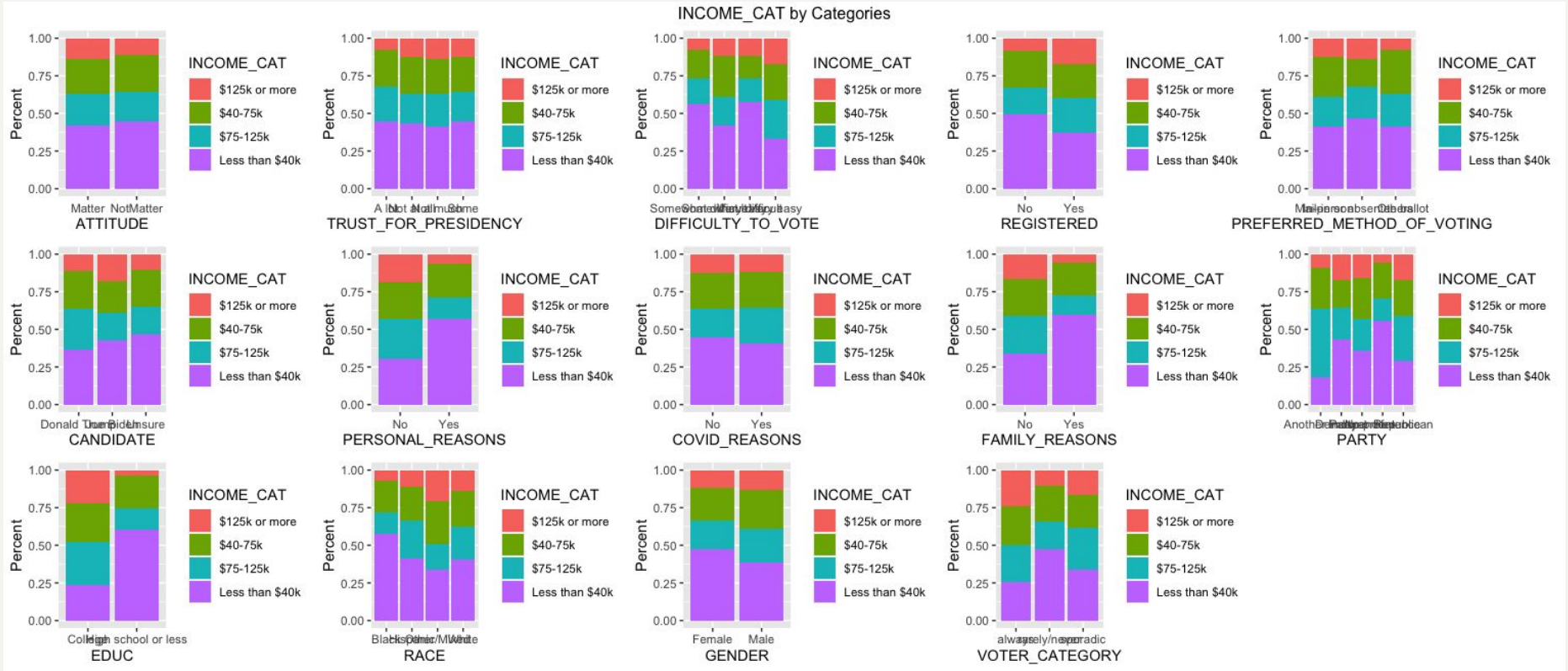
Bivariate Analysis - Categories vs Categories



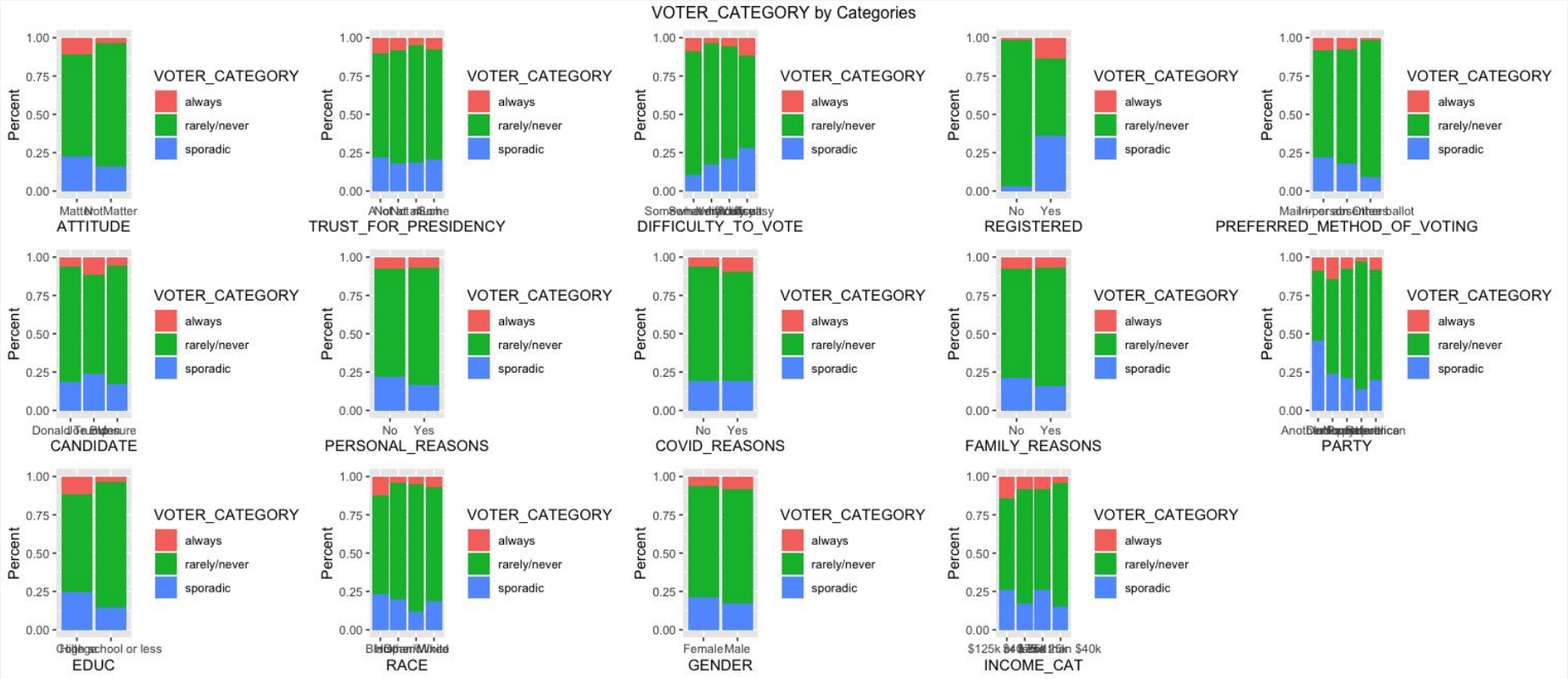
Bivariate Analysis - Categories vs Categories



Bivariate Analysis - Categories vs Categories



Bivariate Analysis - Categories vs Categories



Work Allocation

- Data Cleaning, project plan, project overview and summary - Yifei Cao
- EDA Analysis - Cheng Zhong
- Statistical Analysis - Yahan Wang
- Modeling - Ruoxi Lan

THANK YOU!

Group 19

Yifei Cao
Zhong Cheng
Ruoxi Lan
Yahan Wang

April 28th, 2022