Project Paper 1 - Revised

Rudi Herrig and Jordan Kim

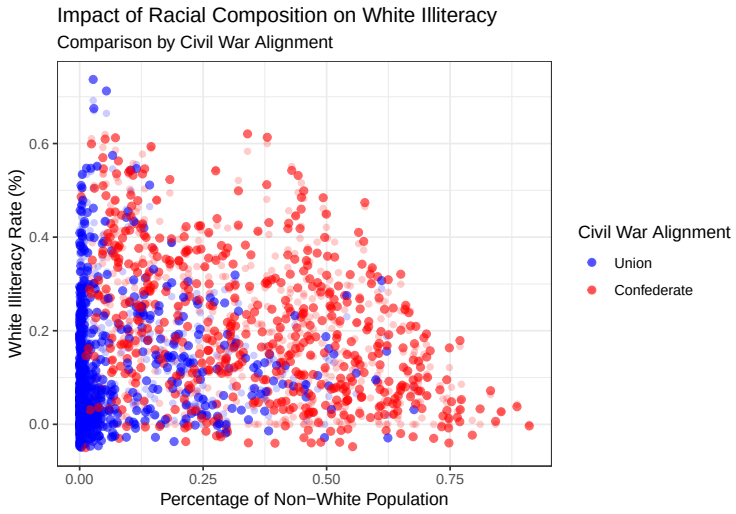
4 May 2024

I wanted to try creating a binary response variable indicating whether a county joined the South because the state in which it is located fought against the Union. I will use the following states: Alabama, Arkansas, Florida, Georgia, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, Texas, and Virginia. I will create a new variable, Confederate, which is equal to 1 if the county is in one of these states, and 0 if the county is not in one of these states. I will use the mutate function to create this new variable.

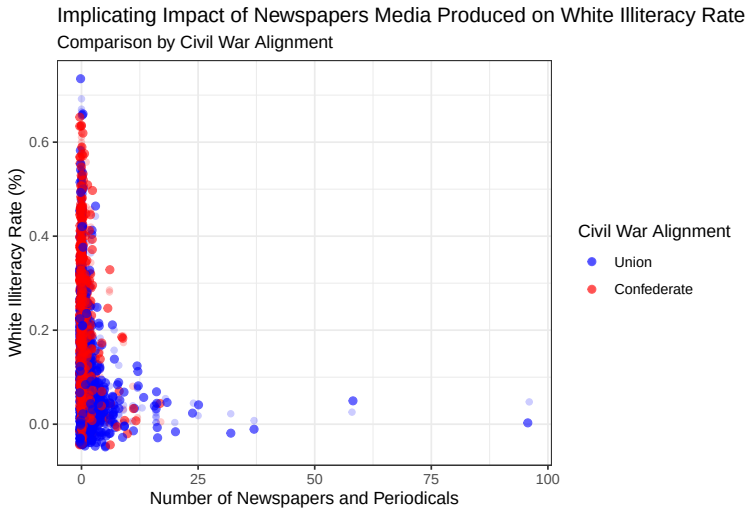
Filtering for only the columns we're interested in and creating a new list `WhiteIlliterate`

Visualizing data we're interested in to learn more about relationships among data set with Blue representing Union counties and Red representing Confederate counties

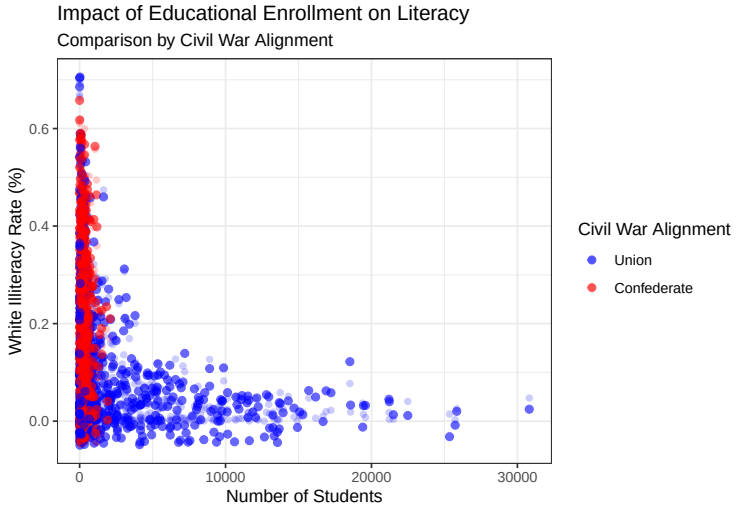
The visualization of the response variable, White Illiteracy, with Percentage of Non-White Population as potential predictor



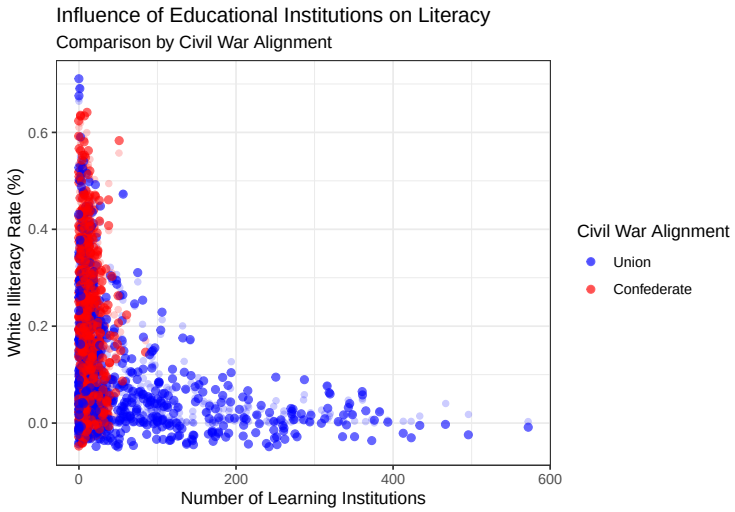
Visualizing the response variable, White Illiteracy, with Number of Newspapers and Periodicals produced as potential predictor



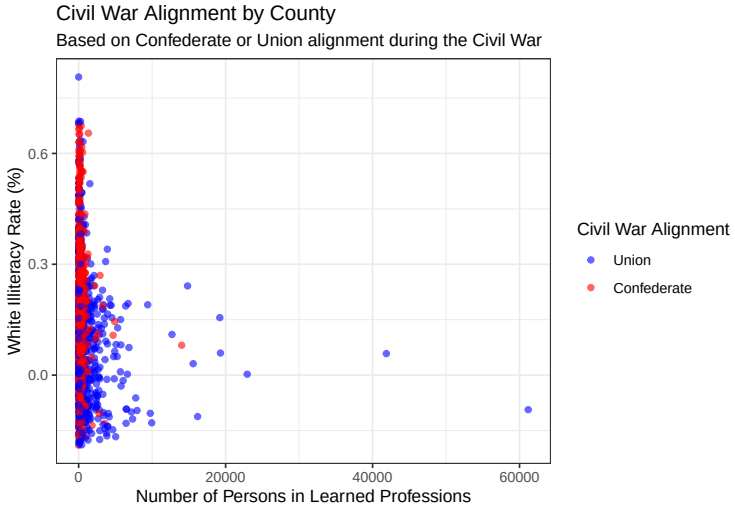
Visualizing the response variable, White Illiteracy, with Number of Students as potential predictor



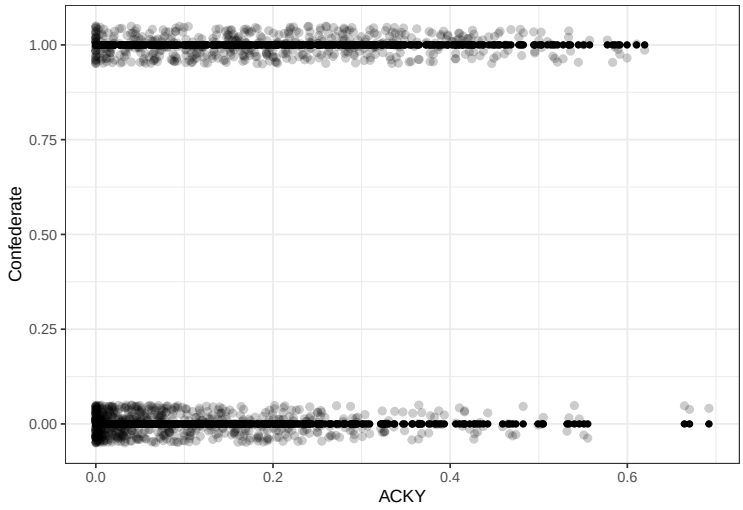
Visualizing the relationship between the Number of Learning Institutions and White Illiteracy



Visualizing the response variable, White Illiteracy, with Number of Persons in Learned Professions Requiring Literacy as potential predictor



Visualizing new binary variable, `Confederate`, as potential response, with White Illiteracy as potential predictor



Splitting Data into Training and Testing Sets

The Response Variable and potential Predictors

In this revision, we also created a new response variable ACKY, which is the percentage of white illiteracy in a county. It is calculated by dividing the Total Number of Whites Over 20 Who Cannot Read or Write by the Total White Population. We will use this as our response variable in our logistic regression model. We also created a binary variable, Confederate, which is equal to 1 if the county was in the South and joined the Confederacy and fought against the United States, and 0 if the county defended our nation and freedom against the evils of Slavery and White Supremacists. We will potentially use the following predictor variables: STATE - "The State (as factors)", ACD001 - "Total Population", ACN001 - "Total Urban Population", ACK001 - "Total White Persons Over 20 Who Cannot Read or Write", ACY - "Total White Persons Over 20", ACZYZ - "Percentage of Colored Population", ACH - "Total Learning Institutions", ACI - "Total Number of Students", ACL - "Number of Printing and Binding Services", ACM - "Number of Newspapers and Periodicals", AC1 - "Total Employed in Learned Professions", ACE - "White Deaf, Dumb, and Blind Persons Over 20", ACS - "Total Persons", ACSW - "Total White Persons", ACSNW - "Total Non-White Persons", ACSP - "Percentage of Non-White Population", ACSWP - "Percentage of White Population". We will use the glm function, Gaussian family, and ACKY~STATE+ACD001+ACN001+ACK001+ACY+ACZYZ+ACH+ACI+ACL+ACM+AC1+ACE+ACS+ACSW+ACSN formula to build a multiple linear regression model. As seen above, we have split our data into training and testing sets, for testing our model's performance.

Null Model

```
## [1] 0.1479186
```

We'll have a null model using the training set data, which is the mean of the response variable, ACKY, to compare our model's predictions on the testing set data performance to. The training set data mean White Illiteracy Rate for counties is 0.1479186, or 14.79% We will use the RMSE to evaluate our model's

performance. We will also look at the distributions of the predictor variables and the correlation between the predictor variables and the response variable.

We will also utilize various techniques to try to find the best model, from shrinkage, dimension reduction, model selection and variable selection. We will also look at the distributions of the predictor variables and the correlation between the predictor variables and the response variable.

The Predictor Variables

We have chosen to use the following predictor variables to predict white illiteracy in counties: STATE - "The State (as factors)", ACD001 - "Total Population", ACN001 - "Total Urban Population", ACK001 - "Total White Persons Over 20 Who Cannot Read or Write", ACY - "Total White Persons Over 20", ACZYZ - "Percentage of Colored Population", ACH - "Total Learning Institutions", ACI - "Total Number of Students", ACL - "Number of Printing and Binding Services", ACM - "Number of Newspapers and Periodicals", AC1 - "Total Employed in Learned Professions", ACE - "White Deaf, Dumb, and Blind Persons Over 20", ACS - "Total Persons", ACSW - "Total White Persons", ACSNW - "Total Non-White Persons", ACSP - "Percentage of Non-White Population", ACSWP - "Percentage of White Population", and Confederate - "Joined Confederacy and Raised Arms Against The United States = 1". We will use the glm function, Gaussian family, and ACKY~STATE+ACD001+ACN001+ACK001+ACY+ACZYZ+ACH+ACI+ACL+ACM+AC1+ACE+ACS+ACSW+ACSNW+ACSP+ACSWP+Confederate formula to build a multiple linear regression model. As seen above, we have split our data into training and testing sets, for testing our model's performance. We will also look at the distributions of the predictor variables and the correlation between the predictor variables and the response variable using the `corrr` package.

building multiple logistic regression model using training dataset

```
##
## Call:
## glm(formula = ACKY ~ STATE + ACD001 + ACN001 + ACK001 + ACY +
##     ACZYZ + ACH + ACI + ACL + ACM + AC1 + ACE + ACS + ACSW +
##     ACSNW + ACSP + ACSWP + Confederate + ACD001:Confederate +
##     ACN001:Confederate + ACK001:Confederate + ACY:Confederate +
##     ACZYZ:Confederate + ACH:Confederate + ACI:Confederate + ACL:Confederate +
##     ACM:Confederate + AC1:Confederate + ACE:Confederate + ACS:Confederate +
##     ACSW:Confederate + ACSNW:Confederate + ACSP:Confederate +
##     ACSWP:Confederate, family = "binomial", data = train)
##
## Coefficients: (7 not defined because of singularities)
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -9.500e-01  8.296e-01  -1.145    0.252
## STATEArkansas              -1.785e-01  7.744e-01  -0.231    0.818
## STATEConnecticut           -2.987e+00  9.573e+00  -0.312    0.755
## STATEDelaware              -1.735e+00  2.679e+00  -0.648    0.517
## STATEDistrict Of Columbia -8.084e-01  4.422e+00  -0.183    0.855
## STATEFlorida Territory     -3.350e-01  1.606e+00  -0.209    0.835
## STATEGeorgia                5.113e-02  6.068e-01   0.084    0.933
## STATEIllinois              -8.469e-01  9.227e-01  -0.918    0.359
## STATEIndiana               -8.724e-01  9.598e-01  -0.909    0.363
## STATEIowa Territory        -1.832e+00  1.527e+00  -1.200    0.230
## STATEKentucky              -2.224e-01  1.141e+00  -0.195    0.845
## STATELouisiana             -4.114e-01  9.377e-01  -0.439    0.661
```

```
## STATEMaine                -1.572e+00  2.282e+00  -0.689    0.491
## STATEMaryland             -1.085e-01  1.530e+00  -0.071    0.943
## STATEMassachusetts        -2.291e+00  3.388e+00  -0.676    0.499
## STATEMichigan             -2.420e+00  1.763e+00  -1.373    0.170
## STATEMississippi          -1.935e-01  6.890e-01  -0.281    0.779
## STATEMissouri             -3.810e-01  1.142e+00  -0.334    0.739
## STATENew Hampshire        -2.558e+00  6.709e+00  -0.381    0.703
## STATENew Jersey           -1.470e+00  1.614e+00  -0.911    0.362
## STATENew York             -1.598e+00  1.693e+00  -0.944    0.345
## STATENorth Carolina        4.391e-02  7.045e-01   0.062    0.950
## STATEOhio                 -1.422e+00  1.137e+00  -1.251    0.211
## STATEPennsylvania         -2.008e+00  1.484e+00  -1.353    0.176
## STATERhode Island         -2.354e+00  4.369e+00  -0.539    0.590
## STATESouth Carolina       -1.336e-02  8.148e-01  -0.016    0.987
## STATETennessee            -3.818e-02  6.451e-01  -0.059    0.953
## STATEVermont              -1.795e+00  2.143e+00  -0.838    0.402
## STATEVirginia             -3.447e-02  6.730e-01  -0.051    0.959
## STATEWisconsin Territory  -1.175e+00  1.161e+00  -1.012    0.312
## ACD001                     2.419e-05  2.828e-04   0.086    0.932
## ACN001                     3.747e-05  1.077e-04   0.348    0.728
## ACK001                     1.962e-03  4.986e-04   3.935 8.32e-05 ***
## ACY                        4.510e-05  6.292e-04   0.072    0.943
## ACZYZ                     -4.623e-01  7.315e+00  -0.063    0.950
## ACH                       -1.953e-03  1.631e-02  -0.120    0.905
## ACI                        2.110e-04  3.281e-04   0.643    0.520
## ACL                        3.542e-02  1.986e-01   0.178    0.858
## ACM                       -8.947e-02  1.998e-01  -0.448    0.654
## AC1                        6.749e-05  5.739e-04   0.118    0.906
## ACE                       -5.364e-03  3.962e-02  -0.135    0.892
## ACS                              NA         NA      NA       NA
## ACSW                      -1.619e-04  4.059e-04  -0.399    0.690
## ACSNW                            NA         NA      NA       NA
## ACSP                      -1.862e+00  1.149e+01  -0.162    0.871
## ACSWP                            NA         NA      NA       NA
## Confederate                      NA         NA      NA       NA
## ACD001:Confederate         4.371e-05  2.937e-04   0.149    0.882
## ACN001:Confederate        -2.853e-05  1.742e-04  -0.164    0.870
## ACK001:Confederate         9.058e-05  6.720e-04   0.135    0.893
## ACY:Confederate           -1.424e-04  1.233e-03  -0.116    0.908
## ACZYZ:Confederate          3.647e+00  8.059e+00   0.452    0.651
## ACH:Confederate            9.093e-03  3.820e-02   0.238    0.812
## ACI:Confederate           -3.633e-04  1.309e-03  -0.278    0.781
## ACL:Confederate            2.141e-02  4.280e-01   0.050    0.960
## ACM:Confederate            5.362e-02  3.847e-01   0.139    0.889
## AC1:Confederate            1.263e-05  7.325e-04   0.017    0.986
## ACE:Confederate            1.615e-02  6.632e-02   0.243    0.808
## ACS:Confederate                  NA         NA      NA       NA
## ACSW:Confederate          -1.499e-04  5.919e-04  -0.253    0.800
## ACSNW:Confederate                NA         NA      NA       NA
## ACSP:Confederate          -3.709e+00  1.227e+01  -0.302    0.762
## ACSWP:Confederate                NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

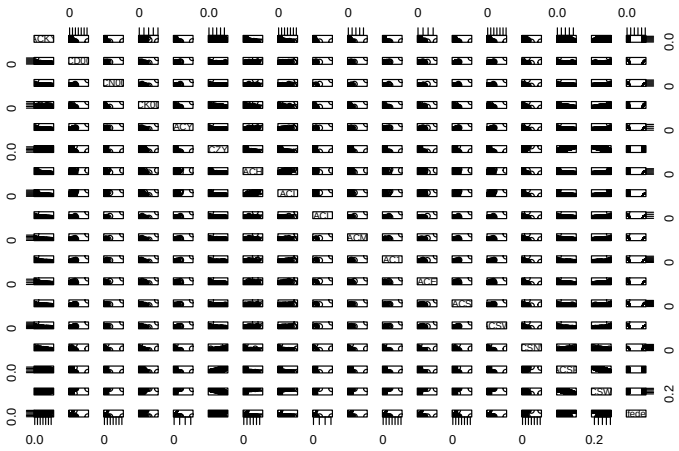
```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 155.369  on 915  degrees of freedom
## Residual deviance:  39.785  on 860  degrees of freedom
## AIC: 437.5
##
## Number of Fisher Scoring iterations: 8
```

The only significant predictor variable is the interaction between the percentage of white population and the Confederate variable. This is not a good sign for our model, as we want to see more significant predictor variables. This may be because of the distribution of our response, or the simple variability between counties.

Association between Predictor Variables and Response Variable

```
##                    ACKY      ACD001      ACN001      ACK001         ACY
## ACKY         1.00000000 -0.25764267 -0.07753955  0.51138311 -0.25278476
## ACD001      -0.25764267  1.00000000  0.74248606  0.44067531  0.97057627
## ACN001      -0.07753955  0.74248606  1.00000000  0.44416038  0.77225406
## ACK001       0.51138311  0.44067531  0.44416038  1.00000000  0.41913916
## ACY         -0.25278476  0.97057627  0.77225406  0.41913916  1.00000000
## ACZYZ        0.12697627 -0.13265099 -0.05057674 -0.04277366 -0.27112328
## ACH         -0.33353130  0.73837179  0.27839977  0.12354185  0.76517440
## ACI         -0.31356932  0.80532172  0.43993953  0.18293357  0.84486526
## ACL         -0.16548764  0.83966745  0.94829722  0.43423565  0.86977206
## ACM         -0.18251260  0.83272074  0.90742258  0.42513854  0.84604892
## AC1         -0.17467064  0.88155247  0.90895697  0.42157365  0.92074839
## ACE         -0.15827469  0.85155623  0.67588847  0.46359274  0.87589298
## ACS         -0.25764267  1.00000000  0.74248606  0.44067531  0.97057627
## ACSW        -0.25792832  0.97276132  0.72523468  0.43089226  0.99377613
## ACSNW       -0.03303765  0.24673732  0.17077049  0.09944876  0.03196855
## ACSP         0.05877044 -0.09718914 -0.04023403 -0.07309328 -0.24319392
## ACSWP       -0.05877044  0.09718914  0.04023403  0.07309328  0.24319392
## Confederate  0.34459402 -0.17918216 -0.06376334  0.09669169 -0.26861763
##                   ACZYZ         ACH         ACI         ACL         ACM
## ACKY         0.12697627 -0.33353130 -0.31356932 -0.1654876 -0.18251260
## ACD001      -0.13265099  0.73837179  0.80532172  0.8396675  0.83272074
## ACN001      -0.05057674  0.27839977  0.43993953  0.9482972  0.90742258
## ACK001      -0.04277366  0.12354185  0.18293357  0.4342357  0.42513854
## ACY         -0.27112328  0.76517440  0.84486526  0.8697721  0.84604892
## ACZYZ        1.00000000 -0.34524290 -0.33262995 -0.1305540 -0.11586078
## ACH         -0.34524290  1.00000000  0.96095641  0.4446565  0.44166444
## ACI         -0.33262995  0.96095641  1.00000000  0.5803914  0.56326185
## ACL         -0.13055399  0.44465651  0.58039138  1.0000000  0.96036028
## ACM         -0.11586078  0.44166444  0.56326185  0.9603603  1.00000000
## AC1         -0.14733957  0.55793248  0.68890884  0.9286378  0.89540501
## ACE         -0.20190504  0.68271380  0.75207280  0.7748669  0.76089307
## ACS         -0.13265099  0.73837179  0.80532172  0.8396675  0.83272074
## ACSW        -0.29803236  0.78507274  0.85187414  0.8350930  0.81733343
## ACSNW        0.67375548 -0.09713018 -0.08761389  0.1306865  0.17496678
## ACSP         0.97967366 -0.30052114 -0.29145867 -0.1096502 -0.09317359
## ACSWP       -0.97967366  0.30052114  0.29145867  0.1096502  0.09317359
## Confederate  0.70443541 -0.33870719 -0.32417558 -0.1481079 -0.14408170
```

```
##                    AC1         ACE         ACS        ACSW       ACSNW
## ACKY        -0.1746706 -0.15827469 -0.25764267 -0.25792832 -0.03303765
## ACD001       0.8815525  0.85155623  1.00000000  0.97276132  0.24673732
## ACN001       0.9089570  0.67588847  0.74248606  0.72523468  0.17077049
## ACK001       0.4215737  0.46359274  0.44067531  0.43089226  0.09944876
## ACY          0.9207484  0.87589298  0.97057627  0.99377613  0.03196855
## ACZYZ       -0.1473396 -0.20190504 -0.13265099 -0.29803236  0.67375548
## ACH          0.5579325  0.68271380  0.73837179  0.78507274 -0.09713018
## ACI          0.6889088  0.75207280  0.80532172  0.85187414 -0.08761389
## ACL          0.9286378  0.77486689  0.83966745  0.83509297  0.13068653
## ACM          0.8954050  0.76089307  0.83272074  0.81733343  0.17496678
## AC1          1.0000000  0.80894961  0.88155247  0.88451061  0.10476128
## ACE          0.8089496  1.00000000  0.85155623  0.86377622  0.06205648
## ACS          0.8815525  0.85155623  1.00000000  0.97276132  0.24673732
## ACSW         0.8845106  0.86377622  0.97276132  1.00000000  0.01537449
## ACSNW        0.1047613  0.06205648  0.24673732  0.01537449  1.00000000
## ACSP        -0.1241502 -0.17978675 -0.09718914 -0.27153566  0.71594617
## ACSWP        0.1241502  0.17978675  0.09718914  0.27153566 -0.71594617
## Confederate -0.1661241 -0.19935572 -0.17918216 -0.28593900  0.42249206
##                    ACSP       ACSWP Confederate
## ACKY         0.05877044 -0.05877044  0.34459402
## ACD001      -0.09718914  0.09718914 -0.17918216
## ACN001      -0.04023403  0.04023403 -0.06376334
## ACK001      -0.07309328  0.07309328  0.09669169
## ACY         -0.24319392  0.24319392 -0.26861763
## ACZYZ        0.97967366 -0.97967366  0.70443541
## ACH         -0.30052114  0.30052114 -0.33870719
## ACI         -0.29145867  0.29145867 -0.32417558
## ACL         -0.10965020  0.10965020 -0.14810789
## ACM         -0.09317359  0.09317359 -0.14408170
## AC1         -0.12415015  0.12415015 -0.16612415
## ACE         -0.17978675  0.17978675 -0.19935572
## ACS         -0.09718914  0.09718914 -0.17918216
## ACSW        -0.27153566  0.27153566 -0.28593900
## ACSNW        0.71594617 -0.71594617  0.42249206
## ACSP         1.00000000 -1.00000000  0.65659605
## ACSWP       -1.00000000  1.00000000 -0.65659605
## Confederate  0.65659605 -0.65659605  1.00000000
```



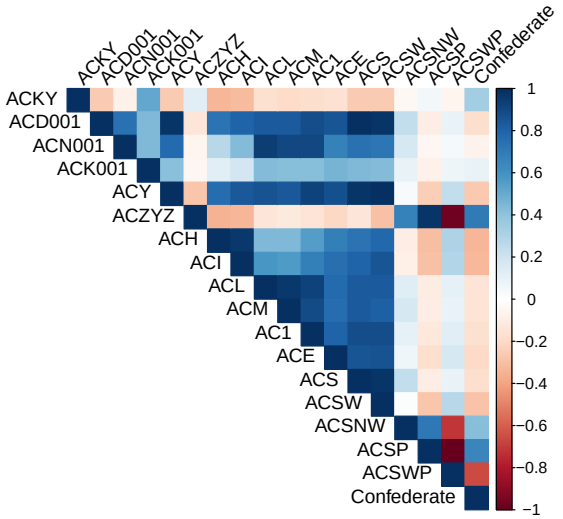
```
## pdf
##   2

## pdf
##   2

## pdf
##   2

## pdf
##   2
```

The correlation between the response variable and the predictor variables is varied, with the highest correlation being 0.51138311 with `ACK001`, and the second highest being 0.34459402 with `Confederate`. Weaker correlations and negative weaker correlations with many variables. `ACH` and `ACI` had negative correlation coefficients of -0.33353130 and -0.31356932. This is not a good sign for our model, as we want to see a strong correlation between the predictor variables and the response variable. This may be because of the distribution of our response, or the simple variability between counties. Let's create a color-coded correlation matrix plot for a better visualization.



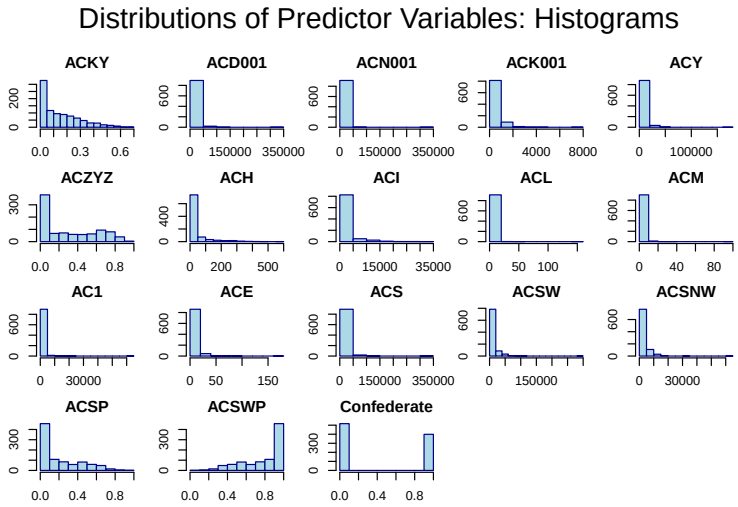
```
## pdf
##   2
```

```
## pdf
##   2
```

```
## pdf
##   2
```

The correlation matrix and the pairwise scatterplots confirm that there is no strong correlation between the predictor variables and the response variable. Again, this is not a good sign for our model, but this may be due to the distribution of our response, or the simple variability between counties.

Distributions of the Predictor Variables




```
## pdf
##   2
```

The Null Model

The null model is the mean of the response variable, ACKY, which is 0.1479186, or 14.79%. We will use the RMSE to evaluate our model's performance. We will also look at the distributions of the predictor variables and the correlation between the predictor variables and the response variable.

Model Evaluation

We will use the RMSE to evaluate our model's performance. We will also look at the distributions of the predictor variables and the correlation between the predictor variables and the response variable.

```
## [1] 0.07180792
```

Our model's RMSE is 0.07180792, or 7.18%. This is a good sign for our model, as we want to see a low RMSE. This means that our model's predictions are close to the actual values. We will also look at the distributions of the predictor variables and the correlation between the predictor variables and the response variable.