DATA 621: BUSINESS ANALYTICS AND DATA MINING FINA PROJECT

Group 2 - Gabriel Campos, Melissa Bowman, Alexander Khaykin, & Jennifer Abinette

Last edited December 14, 2023

Load the data

hate_crime.csv

```
df_hate_crime <-
  read.csv(paste0(git_url,"hate_crime.csv"))
head(df_hate_crime,n=3)
```

```
##   incident_id data_year       ori   pug_agency_name pub_agency_unit
## 1          43      1991 AR0350100        Pine Bluff
## 2          44      1991 AR0350100        Pine Bluff
## 3          45      1991 AR0600300 North Little Rock
##   agency_type_name state_abbr state_name      division_name region_name
## 1             City         AR   Arkansas West South Central       South
## 2             City         AR   Arkansas West South Central       South
## 3             City         AR   Arkansas West South Central       South
##   population_group_code   population_group_description incident_date
## 1                     3 Cities from 50,000 thru 99,999      7/4/1991
## 2                     3 Cities from 50,000 thru 99,999    12/24/1991
## 3                     3 Cities from 50,000 thru 99,999     7/10/1991
##   adult_victim_count juvenile_victim_count total_offender_count
## 1                 NA                    NA                    1
## 2                 NA                    NA                    1
## 3                 NA                    NA                    1
##   adult_offender_count juvenile_offender_count            offender_race
## 1                   NA                      NA Black or African American
## 2                   NA                      NA Black or African American
## 3                   NA                      NA Black or African American
##   offender_ethnicity victim_count
## 1      Not Specified            1
## 2      Not Specified            2
## 3      Not Specified            2
##                                                offense_name
## 1                                        Aggravated Assault
## 2 Aggravated Assault;Destruction/Damage/Vandalism of Property
## 3     Aggravated Assault;Murder and Nonnegligent Manslaughter
##   total_individual_victims                      location_name
```

```
## 1                         1                     Residence/Home
## 2                         1 Highway/Road/Alley/Street/Sidewalk
## 3                         2                     Residence/Home
##                       bias_desc victim_types multiple_offense multiple_bias
## 1 Anti-Black or African American   Individual                S             S
## 2                     Anti-White   Individual                M             S
## 3                     Anti-White   Individual                M             S
```

Summary data

df_hate_crime

Summary

```
##   incident_id        data_year        ori            pug_agency_name
##  Min.   :      2   Min.   :1991   Length:241663      Length:241663
##  1st Qu.:  60446   1st Qu.:1999   Class :character   Class :character
##  Median : 120873   Median :2006   Mode  :character   Mode  :character
##  Mean   : 349025   Mean   :2007
##  3rd Qu.: 181301   3rd Qu.:2016
##  Max.   :1494167   Max.   :2022
##
##  pub_agency_unit    agency_type_name    state_abbr         state_name
##  Length:241663      Length:241663      Length:241663      Length:241663
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  division_name      region_name        population_group_code
##  Length:241663      Length:241663      Length:241663
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  population_group_description incident_date      adult_victim_count
##  Length:241663                Length:241663      Min.   :  0.00
##  Class :character             Class :character   1st Qu.:  0.00
##  Mode  :character             Mode  :character   Median :  1.00
##                                                  Mean   :  0.73
##                                                  3rd Qu.:  1.00
##                                                  Max.   :146.00
##                                                  NA's   :170538
##  juvenile_victim_count total_offender_count adult_offender_count
##  Min.   : 0.0          Min.   : 0.0000      Min.   : 0.00
##  1st Qu.: 0.0          1st Qu.: 0.0000      1st Qu.: 0.00
##  Median : 0.0          Median : 1.0000      Median : 0.00
##  Mean   : 0.1          Mean   : 0.9559      Mean   : 0.61
##  3rd Qu.: 0.0          3rd Qu.: 1.0000      3rd Qu.: 1.00
```

```
##  Max.   :60.0          Max.   :99.0000      Max.   :60.00
##  NA's   :172978                              NA's   :177148
##  juvenile_offender_count offender_race      offender_ethnicity
##  Min.   : 0.00           Length:241663      Length:241663
##  1st Qu.: 0.00           Class :character   Class :character
##  Median : 0.00           Mode  :character   Mode  :character
##  Mean   : 0.12
##  3rd Qu.: 0.00
##  Max.   :20.00
##  NA's   :177155
##   victim_count     offense_name       total_individual_victims
##  Min.   :  1.000   Length:241663      Min.   :  0.000
##  1st Qu.:  1.000   Class :character   1st Qu.:  1.000
##  Median :  1.000   Mode  :character   Median :  1.000
##  Mean   :  1.242                      Mean   :  0.989
##  3rd Qu.:  1.000                      3rd Qu.:  1.000
##  Max.   :900.000                      Max.   :147.000
##                                       NA's   :4859
##  location_name       bias_desc          victim_types       multiple_offense
##  Length:241663      Length:241663      Length:241663      Length:241663
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  multiple_bias
##  Length:241663
##  Class :character
##  Mode  :character
##
##
##
##
```

skim

```
## -- Data Summary ------------------------
##                            Values
## Name                       df_hate_crime
## Number of rows             241663
## Number of columns          28
## _______________________
## Column type frequency:
##   character                19
##   numeric                  9
## _______________________
## Group variables            None
##
## -- Variable type: character -----------------------------------------------------
##    skim_variable               n_missing complete_rate min max  empty n_unique
##  1 ori                                 0             1   9   9      0    10409
##  2 pug_agency_name                     0             1   3  67      0     6942
```

```
##  3 pub_agency_unit                    0             1   0  37 234474      690
##  4 agency_type_name                   0             1   4  21      0        8
##  5 state_abbr                         0             1   2   2      0       53
##  6 state_name                         0             1   4  20      0       53
##  7 division_name                      0             1   5  18      0       11
##  8 region_name                        0             1   4  16      0        6
##  9 population_group_code              0             1   0   2    555       21
## 10 population_group_description       0             1   0  67    555       21
## 11 incident_date                      0             1   8  10      0    11688
## 12 offender_race                      0             1   5  41      0        8
## 13 offender_ethnicity                 0             1   7  22      0        5
## 14 offense_name                       0             1   4 125      0      399
## 15 location_name                      0             1   9  81      0      149
## 16 bias_desc                          0             1   9 163      0      373
## 17 victim_types                       0             1   5  53      0       55
## 18 multiple_offense                   0             1   1   1      0        2
## 19 multiple_bias                      0             1   1   1      0        2
##    whitespace
##  1          0
##  2          0
##  3          0
##  4          0
##  5          0
##  6          0
##  7          0
##  8          0
##  9          0
## 10          0
## 11          0
## 12          0
## 13          0
## 14          0
## 15          0
## 16          0
## 17          0
## 18          0
## 19          0
##
## -- Variable type: numeric -------------------------------------------------------
##   skim_variable            n_missing complete_rate       mean         sd   p0
## 1 incident_id                      0             1     349025.    517581.       2
## 2 data_year                        0             1       2007.         9.41  1991
## 3 adult_victim_count          170538             0.294      0.729      1.13     0
## 4 juvenile_victim_count       172978             0.284      0.101      0.502    0
## 5 total_offender_count             0             1          0.956      1.32     0
## 6 adult_offender_count        177148             0.267      0.609      0.820    0
## 7 juvenile_offender_count     177155             0.267      0.118      0.516    0
## 8 victim_count                     0             1          1.24       2.18     1
## 9 total_individual_victims      4859             0.980      0.989      1.15     0
##      p25    p50     p75    p100 hist
## 1 60446. 120873 181300. 1494167
## 2  1999    2006   2016     2022
## 3     0       1      1      146
## 4     0       0      0       60
```

```
## 5     0       1       1       99
## 6     0       0       1       60
## 7     0       0       0       20
## 8     1       1       1      900
## 9     1       1       1      147
```

Binary Column and Count

```
## Number of Anti_semitic column reports in hatecrime.csv is: 30469  out of  241663
```

Transform to Numeric | Test/Training Dataframes

```r
df_hate_corr <- df_hate_crime
# frequency_map <- table(df_hate_crime$incident_id)
# df_hate_crime$incident_id_freq <- frequency_map[df_hate_crime$incident_id]

add_freq <- function(data, column_name) {
  # Compute frequencies, including NAs
  frequency_map <- table(data[[column_name]], useNA = "always")

  # Create a new column with frequency encoding (including NAs)
  new_col_name <- paste0(column_name, "_freq")
  data[[new_col_name]] <- frequency_map[match(data[[column_name]], names(frequency_map))]

  return(data)
}

# Loop through all columns and add frequency encoding columns (including NAs)
for (col in names(df_hate_corr)) {
  df_hate_corr <- add_freq(df_hate_corr, col)
}

# incase dropping categorical data is needed, method to remove is below
```

```r
df_hate_corr<-df_hate_corr%>%
  dplyr::select(matches("_freq"), matches("^Anti"))
```

```r
set.seed(123)

train_indices <- sample(seq_len(nrow(df_hate_corr)), 0.8 * nrow(df_hate_corr))

# Create training dataset
df_hate_train <- df_hate_corr[train_indices, ]

# Create test dataset
df_hate_test <- df_hate_corr[-train_indices, ]
```

```
rm(git_url,col,total_hate,train_indices,cnt_hate, add_freq)
```

Histograms

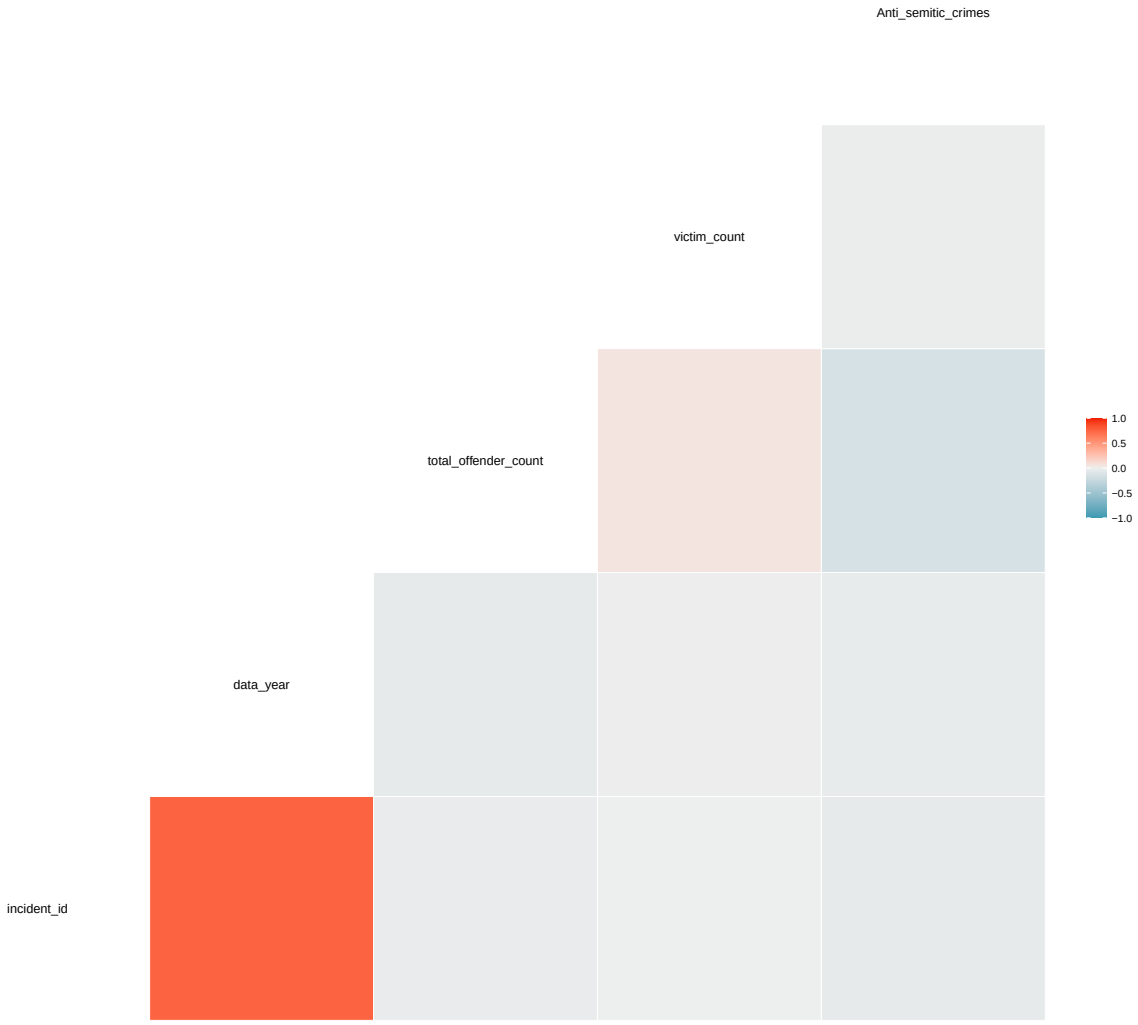
```
df_hate_train %>%
  dplyr::select( ends_with("_freq"))%>%
  gather() %>%
  ggplot(aes(x = value)) +
  geom_histogram(fill = "cadetblue") +
  facet_wrap(~key, scales = "free")
```

```
## Don't know how to automatically pick scale for object of type <table>.
## Defaulting to continuous.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


agency_type_name_freq
Anti_semitic_crimes_freq
bias_desc_freq
data_year_freq
division_name_freq
incident_id_freq
location_name_freq
multiple_bias_freq
multiple_offense_freq
offender_ethnicity_freq
offender_race_freq
offense_name_freq
ori_freq
population_group_code_freq
population_group_description_freq
pub_agency_unit_freq
pug_agency_name_freq
region_name_freq
state_abbr_freq
state_name_freq
total_offender_count_freq
victim_count_freq
victim_types_freq
count
value

Correlation plots

df_hate_crime




```
##                                incident_id_freq data_year_freq     ori_freq
## incident_id_freq                              1             NA           NA
## data_year_freq                               NA    1.000000000 -0.013871933
## ori_freq                                     NA   -0.013871933  1.000000000
## pug_agency_name_freq                         NA   -0.009026709  0.978751858
## pub_agency_unit_freq                         NA   -0.008593493  0.083213106
## agency_type_name_freq                        NA   -0.008639880  0.174126507
## state_abbr_freq                              NA    0.031645826  0.304742487
## state_name_freq                              NA    0.031645826  0.304742487
```

```
## division_name_freq                              NA    0.015078796  0.282266363
## region_name_freq                                NA   -0.003347976  0.269982508
## population_group_code_freq                      NA   -0.012610023  0.519418276
## population_group_description_freq               NA   -0.012610023  0.519418276
## total_offender_count_freq                       NA    0.054281506  0.002937708
## offender_race_freq                              NA   -0.149113424 -0.002460414
## offender_ethnicity_freq                         NA   -0.341244470  0.052554142
## victim_count_freq                               NA    0.019239989  0.050706014
## offense_name_freq                               NA   -0.012883781  0.034001332
## location_name_freq                              NA   -0.037712625  0.019486068
## bias_desc_freq                                  NA   -0.028375268 -0.067555490
## victim_types_freq                               NA    0.001542626 -0.046663692
## multiple_offense_freq                           NA   -0.008431119  0.052586233
## multiple_bias_freq                              NA   -0.088979373  0.024194291
## Anti_semitic_crimes_freq                        NA    0.032244528 -0.175632995
## Anti_semitic_crimes                             NA   -0.032244528  0.175632995
##                                  pug_agency_name_freq pub_agency_unit_freq
## incident_id_freq                                   NA                   NA
## data_year_freq                         -0.0090267094        -0.0085934926
## ori_freq                                0.9787518580         0.0832131065
## pug_agency_name_freq                    1.0000000000         0.0244091045
## pub_agency_unit_freq                    0.0244091045         1.0000000000
## agency_type_name_freq                   0.1253116828         0.4026528884
## state_abbr_freq                         0.3341394016         0.0184870773
## state_name_freq                         0.3341394016         0.0184870773
## division_name_freq                      0.3015699161         0.0089649710
## region_name_freq                        0.2802497543         0.0680044189
## population_group_code_freq              0.4890648899         0.3183436799
## population_group_description_freq       0.4890648899         0.3183436799
## total_offender_count_freq               0.0003490475        -0.0086457578
## offender_race_freq                      0.0005322316        -0.0271880690
## offender_ethnicity_freq                 0.0468513847         0.0026724826
## victim_count_freq                       0.0494492587        -0.0066972613
## offense_name_freq                       0.0273213134        -0.0196243331
## location_name_freq                      0.0193409017         0.0481590596
## bias_desc_freq                         -0.0645411554        -0.0113326741
## victim_types_freq                      -0.0406288469         0.0726715121
## multiple_offense_freq                   0.0493994452        -0.0009951712
## multiple_bias_freq                      0.0258286043         0.0038363575
## Anti_semitic_crimes_freq               -0.1657299726         0.0100349347
## Anti_semitic_crimes                     0.1657299726        -0.0100349347
##                                  agency_type_name_freq state_abbr_freq
## incident_id_freq                                    NA              NA
## data_year_freq                          -0.0086398804     0.031645826
## ori_freq                                 0.1741265071     0.304742487
## pug_agency_name_freq                     0.1253116828     0.334139402
## pub_agency_unit_freq                     0.4026528884     0.018487077
## agency_type_name_freq                    1.0000000000     0.114402324
## state_abbr_freq                          0.1144023242     1.000000000
## state_name_freq                          0.1144023242     1.000000000
## division_name_freq                       0.0712117000     0.750876436
## region_name_freq                         0.1659341904     0.565542814
## population_group_code_freq               0.5819465044     0.270791553
## population_group_description_freq        0.5819465044     0.270791553
```

```
## total_offender_count_freq                -0.0159341084    -0.016465433
## offender_race_freq                       -0.0337024353     0.089461419
## offender_ethnicity_freq                  -0.0072585035     0.005548728
## victim_count_freq                         0.0008185340     0.038803416
## offense_name_freq                        -0.0139839913     0.085988917
## location_name_freq                        0.0481382017    -0.010331819
## bias_desc_freq                           -0.0160700528    -0.022015592
## victim_types_freq                         0.1004581121    -0.015155810
## multiple_offense_freq                     0.0004102563     0.016582208
## multiple_bias_freq                       -0.0020385990     0.002446150
## Anti_semitic_crimes_freq                  0.0062762690    -0.112879840
## Anti_semitic_crimes                      -0.0062762690     0.112879840
##                                  state_name_freq division_name_freq
## incident_id_freq                              NA                 NA
## data_year_freq                       0.031645826        0.015078796
## ori_freq                             0.304742487        0.282266363
## pug_agency_name_freq                 0.334139402        0.301569916
## pub_agency_unit_freq                 0.018487077        0.008964971
## agency_type_name_freq                0.114402324        0.071211700
## state_abbr_freq                      1.000000000        0.750876436
## state_name_freq                      1.000000000        0.750876436
## division_name_freq                   0.750876436        1.000000000
## region_name_freq                     0.565542814        0.595152902
## population_group_code_freq           0.270791553        0.239753886
## population_group_description_freq    0.270791553        0.239753886
## total_offender_count_freq           -0.016465433       -0.005135586
## offender_race_freq                   0.089461419        0.073385878
## offender_ethnicity_freq              0.005548728        0.030195645
## victim_count_freq                    0.038803416        0.042751853
## offense_name_freq                    0.085988917        0.110222443
## location_name_freq                  -0.010331819       -0.015593313
## bias_desc_freq                      -0.022015592       -0.016128311
## victim_types_freq                   -0.015155810       -0.050110925
## multiple_offense_freq                0.016582208        0.020505843
## multiple_bias_freq                   0.002446150       -0.002353156
## Anti_semitic_crimes_freq            -0.112879840       -0.142115693
## Anti_semitic_crimes                  0.112879840        0.142115693
##                                  region_name_freq population_group_code_freq
## incident_id_freq                               NA                         NA
## data_year_freq                       -0.003347976               -0.012610023
## ori_freq                              0.269982508                0.519418276
## pug_agency_name_freq                  0.280249754                0.489064890
## pub_agency_unit_freq                  0.068004419                0.318343680
## agency_type_name_freq                 0.165934190                0.581946504
## state_abbr_freq                       0.565542814                0.270791553
## state_name_freq                       0.565542814                0.270791553
## division_name_freq                    0.595152902                0.239753886
## region_name_freq                      1.000000000                0.285205125
## population_group_code_freq            0.285205125                1.000000000
## population_group_description_freq     0.285205125                1.000000000
## total_offender_count_freq            -0.013990120               -0.025844677
## offender_race_freq                    0.094519873                0.008164061
## offender_ethnicity_freq               0.015924135                0.027683017
## victim_count_freq                     0.002501637                0.027910392
```

```
## offense_name_freq                    0.102410551               0.065311800
## location_name_freq                  -0.024821945               0.027680903
## bias_desc_freq                      -0.051355982              -0.020118975
## victim_types_freq                   -0.027886958               0.022821710
## multiple_offense_freq               -0.001767274               0.034804198
## multiple_bias_freq                  -0.015265828               0.008339386
## Anti_semitic_crimes_freq            -0.125811774              -0.118374892
## Anti_semitic_crimes                  0.125811774               0.118374892
##                                 population_group_description_freq
## incident_id_freq                                               NA
## data_year_freq                                        -0.012610023
## ori_freq                                               0.519418276
## pug_agency_name_freq                                   0.489064890
## pub_agency_unit_freq                                   0.318343680
## agency_type_name_freq                                  0.581946504
## state_abbr_freq                                        0.270791553
## state_name_freq                                        0.270791553
## division_name_freq                                     0.239753886
## region_name_freq                                       0.285205125
## population_group_code_freq                             1.000000000
## population_group_description_freq                      1.000000000
## total_offender_count_freq                             -0.025844677
## offender_race_freq                                     0.008164061
## offender_ethnicity_freq                                0.027683017
## victim_count_freq                                      0.027910392
## offense_name_freq                                      0.065311800
## location_name_freq                                     0.027680903
## bias_desc_freq                                        -0.020118975
## victim_types_freq                                      0.022821710
## multiple_offense_freq                                  0.034804198
## multiple_bias_freq                                     0.008339386
## Anti_semitic_crimes_freq                              -0.118374892
## Anti_semitic_crimes                                    0.118374892
##                                 total_offender_count_freq offender_race_freq
## incident_id_freq                                       NA                NA
## data_year_freq                                0.0542815062      -0.1491134244
## ori_freq                                      0.0029377084      -0.0024604135
## pug_agency_name_freq                          0.0003490475       0.0005322316
## pub_agency_unit_freq                         -0.0086457578      -0.0271880690
## agency_type_name_freq                        -0.0159341084      -0.0337024353
## state_abbr_freq                              -0.0164654335       0.0894614189
## state_name_freq                              -0.0164654335       0.0894614189
## division_name_freq                           -0.0051355855       0.0733858775
## region_name_freq                             -0.0139901201       0.0945198725
## population_group_code_freq                   -0.0258446766       0.0081640605
## population_group_description_freq            -0.0258446766       0.0081640605
## total_offender_count_freq                     1.0000000000       0.1186348207
## offender_race_freq                            0.1186348207       1.0000000000
## offender_ethnicity_freq                      -0.0933032695       0.0222564667
## victim_count_freq                             0.1415207359       0.0301469565
## offense_name_freq                             0.1080503828       0.1831251216
## location_name_freq                            0.0029731480       0.0503495808
## bias_desc_freq                                0.0111139032       0.1588410907
## victim_types_freq                             0.0054476705      -0.0997291939
```

```
## multiple_offense_freq                          0.0577077475       0.0141474074
## multiple_bias_freq                             0.0034199513       0.0229036288
## Anti_semitic_crimes_freq                      -0.0141950947      -0.1132929881
## Anti_semitic_crimes                            0.0141950947       0.1132929881
##                                  offender_ethnicity_freq victim_count_freq
## incident_id_freq                                      NA                NA
## data_year_freq                              -0.341244470       0.019239989
## ori_freq                                     0.052554142       0.050706014
## pug_agency_name_freq                         0.046851385       0.049449259
## pub_agency_unit_freq                         0.002672483      -0.006697261
## agency_type_name_freq                       -0.007258504       0.000818534
## state_abbr_freq                              0.005548728       0.038803416
## state_name_freq                              0.005548728       0.038803416
## division_name_freq                           0.030195645       0.042751853
## region_name_freq                             0.015924135       0.002501637
## population_group_code_freq                   0.027683017       0.027910392
## population_group_description_freq            0.027683017       0.027910392
## total_offender_count_freq                   -0.093303269       0.141520736
## offender_race_freq                           0.022256467       0.030146957
## offender_ethnicity_freq                      1.000000000      -0.008891950
## victim_count_freq                           -0.008891950       1.000000000
## offense_name_freq                            0.060216092       0.206616730
## location_name_freq                           0.065563508      -0.054328812
## bias_desc_freq                               0.049036420      -0.007926248
## victim_types_freq                           -0.048031871      -0.068302188
## multiple_offense_freq                        0.010628414       0.449643477
## multiple_bias_freq                           0.127879302       0.022120982
## Anti_semitic_crimes_freq                    -0.053597504      -0.062161226
## Anti_semitic_crimes                          0.053597504       0.062161226
##                                  offense_name_freq location_name_freq
## incident_id_freq                                NA                NA
## data_year_freq                      -0.0128837811       -0.037712625
## ori_freq                             0.0340013323        0.019486068
## pug_agency_name_freq                 0.0273213134        0.019340902
## pub_agency_unit_freq                -0.0196243331        0.048159060
## agency_type_name_freq               -0.0139839913        0.048138202
## state_abbr_freq                      0.0859889170       -0.010331819
## state_name_freq                      0.0859889170       -0.010331819
## division_name_freq                   0.1102224429       -0.015593313
## region_name_freq                     0.1024105512       -0.024821945
## population_group_code_freq           0.0653118003        0.027680903
## population_group_description_freq    0.0653118003        0.027680903
## total_offender_count_freq            0.1080503828        0.002973148
## offender_race_freq                   0.1831251216        0.050349581
## offender_ethnicity_freq              0.0602160919        0.065563508
## victim_count_freq                    0.2066167297       -0.054328812
## offense_name_freq                    1.0000000000        0.005208577
## location_name_freq                   0.0052085766        1.000000000
## bias_desc_freq                       0.1002329638        0.050949956
## victim_types_freq                   -0.1333142761        0.203239938
## multiple_offense_freq                0.4127595820       -0.013624832
## multiple_bias_freq                  -0.0008832887        0.030393131
## Anti_semitic_crimes_freq            -0.2021127092        0.022833644
## Anti_semitic_crimes                  0.2021127092       -0.022833644
```

```
##                                  bias_desc_freq victim_types_freq
## incident_id_freq                             NA                NA
## data_year_freq                    -0.028375268       0.001542626
## ori_freq                          -0.067555490      -0.046663692
## pug_agency_name_freq              -0.064541155      -0.040628847
## pub_agency_unit_freq              -0.011332674       0.072671512
## agency_type_name_freq             -0.016070053       0.100458112
## state_abbr_freq                   -0.022015592      -0.015155810
## state_name_freq                   -0.022015592      -0.015155810
## division_name_freq                -0.016128311      -0.050110925
## region_name_freq                  -0.051355982      -0.027886958
## population_group_code_freq        -0.020118975       0.022821710
## population_group_description_freq -0.020118975       0.022821710
## total_offender_count_freq          0.011113903       0.005447671
## offender_race_freq                 0.158841091      -0.099729194
## offender_ethnicity_freq            0.049036420      -0.048031871
## victim_count_freq                 -0.007926248      -0.068302188
## offense_name_freq                  0.100232964      -0.133314276
## location_name_freq                 0.050949956       0.203239938
## bias_desc_freq                     1.000000000       0.048423393
## victim_types_freq                  0.048423393       1.000000000
## multiple_offense_freq              0.002496918       0.019994019
## multiple_bias_freq                 0.103104372       0.020055042
## Anti_semitic_crimes_freq           0.115969336       0.283836422
## Anti_semitic_crimes               -0.115969336      -0.283836422
##                                  multiple_offense_freq multiple_bias_freq
## incident_id_freq                                    NA                NA
## data_year_freq                          -0.0084311191      -0.0889793733
## ori_freq                                 0.0525862334       0.0241942914
## pug_agency_name_freq                     0.0493994452       0.0258286043
## pub_agency_unit_freq                    -0.0009951712       0.0038363575
## agency_type_name_freq                    0.0004102563      -0.0020385990
## state_abbr_freq                          0.0165822083       0.0024461505
## state_name_freq                          0.0165822083       0.0024461505
## division_name_freq                       0.0205058429      -0.0023531565
## region_name_freq                        -0.0017672738      -0.0152658280
## population_group_code_freq               0.0348041975       0.0083393856
## population_group_description_freq        0.0348041975       0.0083393856
## total_offender_count_freq                0.0577077475       0.0034199513
## offender_race_freq                       0.0141474074       0.0229036288
## offender_ethnicity_freq                  0.0106284137       0.1278793023
## victim_count_freq                        0.4496434775       0.0221209819
## offense_name_freq                        0.4127595820      -0.0008832887
## location_name_freq                      -0.0136248317       0.0303931307
## bias_desc_freq                           0.0024969182       0.1031043720
## victim_types_freq                        0.0199940192       0.0200550420
## multiple_offense_freq                    1.0000000000       0.0582026971
## multiple_bias_freq                       0.0582026971       1.0000000000
## Anti_semitic_crimes_freq                -0.0390344463       0.0439969352
## Anti_semitic_crimes                      0.0390344463      -0.0439969352
##                                  Anti_semitic_crimes_freq Anti_semitic_crimes
## incident_id_freq                                       NA                  NA
## data_year_freq                                0.032244528        -0.032244528
## ori_freq                                     -0.175632995         0.175632995
```

```
## pug_agency_name_freq                -0.165729973         0.165729973
## pub_agency_unit_freq                 0.010034935        -0.010034935
## agency_type_name_freq                0.006276269        -0.006276269
## state_abbr_freq                     -0.112879840         0.112879840
## state_name_freq                     -0.112879840         0.112879840
## division_name_freq                  -0.142115693         0.142115693
## region_name_freq                    -0.125811774         0.125811774
## population_group_code_freq          -0.118374892         0.118374892
## population_group_description_freq   -0.118374892         0.118374892
## total_offender_count_freq           -0.014195095         0.014195095
## offender_race_freq                  -0.113292988         0.113292988
## offender_ethnicity_freq             -0.053597504         0.053597504
## victim_count_freq                   -0.062161226         0.062161226
## offense_name_freq                   -0.202112709         0.202112709
## location_name_freq                   0.022833644        -0.022833644
## bias_desc_freq                       0.115969336        -0.115969336
## victim_types_freq                    0.283836422        -0.283836422
## multiple_offense_freq               -0.039034446         0.039034446
## multiple_bias_freq                   0.043996935        -0.043996935
## Anti_semitic_crimes_freq             1.000000000        -1.000000000
## Anti_semitic_crimes                 -1.000000000         1.000000000
```

```
# df_hate_train<-table(df_hate_train)
```

```
base::summary(df_hate_train)
```

```
##  incident_id_freq data_year_freq          ori_freq         pug_agency_name_freq
##  n.vars :1        n.vars :1            n.vars :1           n.vars :1
##  n.cases:193330   n.cases:1515270938   n.cases:282226808   n.cases:311246346
##
##
##
##
##  pub_agency_unit_freq agency_type_name_freq state_abbr_freq
##  n.vars :1             n.vars :1             n.vars :1
##  n.cases:4.398e+10     n.cases:3.079e+10     n.cases:2.882e+09
##
##
##
##
##  state_name_freq     division_name_freq region_name_freq
##  n.vars :1           n.vars :1          n.vars :1
##  n.cases:2.882e+09   n.cases:6.85e+09   n.cases:1.202e+10
##
##
##
##
##  population_group_code_freq population_group_description_freq
##  n.vars :1                  n.vars :1
##  n.cases:4.4e+09            n.cases:4.4e+09
##
##
##
```

```
##
##  total_offender_count_freq offender_race_freq  offender_ethnicity_freq
##  n.vars :1                 n.vars :1           n.vars :1
##  n.cases:1.745e+10         n.cases:1.561e+10   n.cases:3.38e+10
##
##
##
##
##  victim_count_freq   offense_name_freq   location_name_freq
##  n.vars :1           n.vars :1           n.vars :1
##  n.cases:3.441e+10   n.cases:1.041e+10   n.cases:7.081e+09
##
##
##
##
##  bias_desc_freq      victim_types_freq   multiple_offense_freq
##  n.vars :1           n.vars :1           n.vars :1
##  n.cases:7.557e+09   n.cases:2.968e+10   n.cases:4.366e+10
##
##
##
##
##  multiple_bias_freq  Anti_semitic_crimes_freq Anti_semitic_crimes
##  n.vars :1           n.vars :1                Min.   :0.0000
##  n.cases:4.608e+10   n.cases:3.644e+10        1st Qu.:0.0000
##                                               Median :0.0000
##                                               Mean   :0.1256
##                                               3rd Qu.:0.0000
##                                               Max.   :1.0000
```

```
str(df_hate_train)
```

```
## 'data.frame':    193330 obs. of  24 variables:
##  $ incident_id_freq                : 'table' int [1:193330(1d)] 1 1 1 1 1 1 1 1 1 1 ...
##  $ data_year_freq                  : 'table' int [1:193330(1d)] 6270 7327 8039 7623 6592 10889 7684
##  $ ori_freq                        : 'table' int [1:193330(1d)] 20 111 34 4804 2631 11822 103 6 100
##  $ pug_agency_name_freq            : 'table' int [1:193330(1d)] 96 113 34 4804 2631 11822 103 6 101
##  $ pub_agency_unit_freq            : 'table' int [1:193330(1d)] 234474 234474 234474 234474 234474 1
##  $ agency_type_name_freq           : 'table' int [1:193330(1d)] 34713 192876 192876 192876 34713 191
##  $ state_abbr_freq                 : 'table' int [1:193330(1d)] 779 2431 22841 12012 21748 21748 228
##  $ state_name_freq                 : 'table' int [1:193330(1d)] 779 2431 22841 12012 21748 21748 228
##  $ division_name_freq              : 'table' int [1:193330(1d)] 13089 33308 48736 19613 48736 48736
##  $ region_name_freq                : 'table' int [1:193330(1d)] 50683 47609 68349 68349 68349 68349
##  $ population_group_code_freq      : 'table' int [1:193330(1d)] 6389 23046 23597 23703 22143 34651 1
##  $ population_group_description_freq: 'table' int [1:193330(1d)] 6389 23046 23597 23703 22143 34651 1
##  $ total_offender_count_freq       : 'table' int [1:193330(1d)] 118196 85378 118196 21622 85378 118
##  $ offender_race_freq              : 'table' int [1:193330(1d)] 32022 96520 95169 32022 96520 95169
##  $ offender_ethnicity_freq         : 'table' int [1:193330(1d)] 13513 19877 204104 204104 204104 204
##  $ victim_count_freq               : 'table' int [1:193330(1d)] 205573 205573 205573 205573 205573 1
##  $ offense_name_freq               : 'table' int [1:193330(1d)] 46620 70391 70391 70391 72045 25430
##  $ location_name_freq              : 'table' int [1:193330(1d)] 45164 31532 13827 45164 31532 71139
##  $ bias_desc_freq                  : 'table' int [1:193330(1d)] 5643 81440 81440 27115 29958 23860 8
##  $ victim_types_freq               : 'table' int [1:193330(1d)] 191377 191377 191377 191377 1600 191
##  $ multiple_offense_freq           : 'table' int [1:193330(1d)] 233460 233460 233460 233460 233460 1
```

```
##  $ multiple_bias_freq               : 'table' int [1:193330(1d)] 239987 239987 239987 239987 239987 1
##  $ Anti_semitic_crimes_freq         : 'table' int [1:193330(1d)] 211194 211194 211194 211194 30469 21
##  $ Anti_semitic_crimes              : num  0 0 0 0 1 0 0 0 0 0 ...
```

```
df_hate_train<-as.data.frame.matrix(df_hate_train)
```

```
str(df_hate_train)
```

```
## 'data.frame':    193330 obs. of  24 variables:
##  $ incident_id_freq                 : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ data_year_freq                   : int  6270 7327 8039 7623 6592 10889 7684 7623 7179 7327 ...
##  $ ori_freq                         : int  20 111 34 4804 2631 11822 103 6 100 96 ...
##  $ pug_agency_name_freq             : int  96 113 34 4804 2631 11822 103 6 101 630 ...
##  $ pub_agency_unit_freq             : int  234474 234474 234474 234474 234474 234474 234474 234474 10
##  $ agency_type_name_freq            : int  34713 192876 192876 192876 34713 192876 192876 192876 8038
##  $ state_abbr_freq                  : int  779 2431 22841 12012 21748 21748 22841 13052 10022 9570 .
##  $ state_name_freq                  : int  779 2431 22841 12012 21748 21748 22841 13052 10022 9570 .
##  $ division_name_freq               : int  13089 33308 48736 19613 48736 48736 48736 33308 33308 5525
##  $ region_name_freq                 : int  50683 47609 68349 68349 68349 68349 68349 47609 47609 7429
##  $ population_group_code_freq       : int  6389 23046 23597 23703 22143 34651 26734 14542 12623 22143
##  $ population_group_description_freq: int  6389 23046 23597 23703 22143 34651 26734 14542 12623 22143
##  $ total_offender_count_freq        : int  118196 85378 118196 21622 85378 118196 85378 85378 85378 1
##  $ offender_race_freq               : int  32022 96520 95169 32022 96520 95169 96520 96520 9546 95169
##  $ offender_ethnicity_freq          : int  13513 19877 204104 204104 204104 204104 204104 204104 2041
##  $ victim_count_freq                : int  205573 205573 205573 205573 205573 205573 205573 205573 20
##  $ offense_name_freq                : int  46620 70391 70391 70391 72045 25430 72045 1288 580 72045
##  $ location_name_freq               : int  45164 31532 13827 45164 31532 71139 31532 71139 1982 4003
##  $ bias_desc_freq                   : int  5643 81440 81440 27115 29958 23860 81440 81440 592 15442
##  $ victim_types_freq                : int  191377 191377 191377 191377 1600 191377 14563 191377 1913
##  $ multiple_offense_freq            : int  233460 233460 233460 233460 233460 233460 233460 233460 23
##  $ multiple_bias_freq               : int  239987 239987 239987 239987 239987 239987 239987 239987 23
##  $ Anti_semitic_crimes_freq         : int  211194 211194 211194 211194 30469 211194 211194 211194 211
##  $ Anti_semitic_crimes              : num  0 0 0 0 1 0 0 0 0 0 ...
```

```
df_hate_train<-as.data.frame(df_hate_train)
```

```
str(df_hate_train)
```

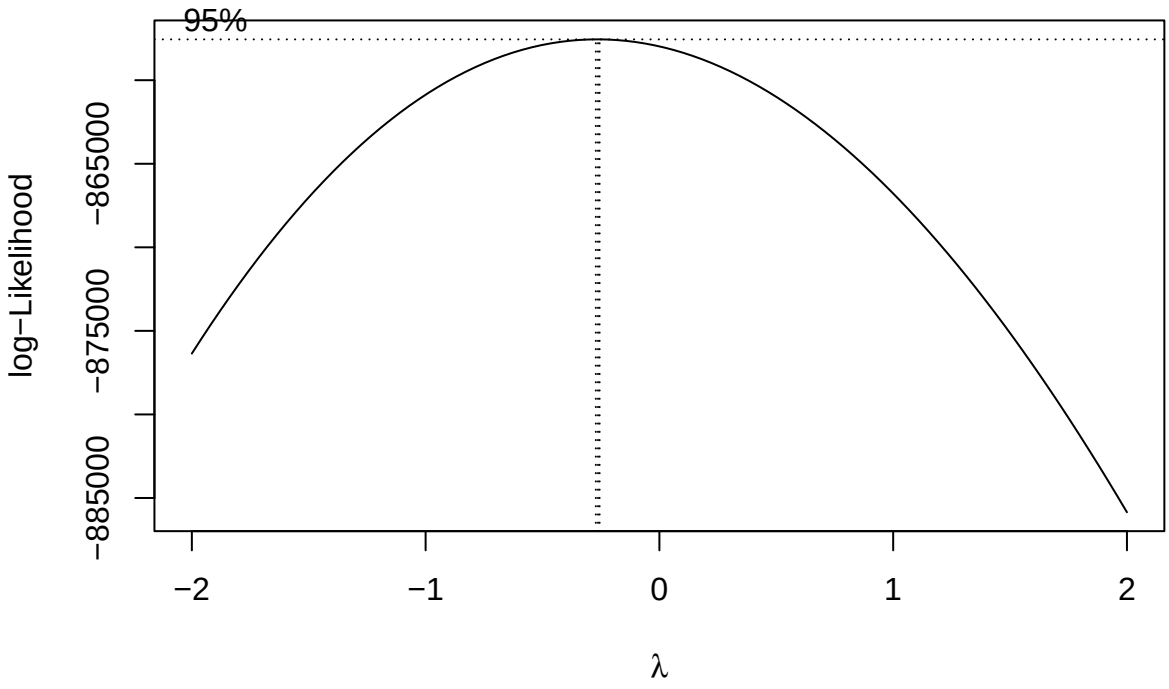
```
## 'data.frame':    193330 obs. of  24 variables:
##  $ incident_id_freq                 : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ data_year_freq                   : int  6270 7327 8039 7623 6592 10889 7684 7623 7179 7327 ...
##  $ ori_freq                         : int  20 111 34 4804 2631 11822 103 6 100 96 ...
##  $ pug_agency_name_freq             : int  96 113 34 4804 2631 11822 103 6 101 630 ...
##  $ pub_agency_unit_freq             : int  234474 234474 234474 234474 234474 234474 234474 234474 10
##  $ agency_type_name_freq            : int  34713 192876 192876 192876 34713 192876 192876 192876 8038
##  $ state_abbr_freq                  : int  779 2431 22841 12012 21748 21748 22841 13052 10022 9570 .
##  $ state_name_freq                  : int  779 2431 22841 12012 21748 21748 22841 13052 10022 9570 .
##  $ division_name_freq               : int  13089 33308 48736 19613 48736 48736 48736 33308 33308 5525
##  $ region_name_freq                 : int  50683 47609 68349 68349 68349 68349 68349 47609 47609 7429
##  $ population_group_code_freq       : int  6389 23046 23597 23703 22143 34651 26734 14542 12623 22143
##  $ population_group_description_freq: int  6389 23046 23597 23703 22143 34651 26734 14542 12623 22143
##  $ total_offender_count_freq        : int  118196 85378 118196 21622 85378 118196 85378 85378 85378 1
```

```
##  $ offender_race_freq              : int  32022 96520 95169 32022 96520 95169 96520 96520 9546 9516
##  $ offender_ethnicity_freq         : int  13513 19877 204104 204104 204104 204104 204104 204104 204
##  $ victim_count_freq               : int  205573 205573 205573 205573 205573 205573 205573 205573 2
##  $ offense_name_freq               : int  46620 70391 70391 70391 72045 25430 72045 1288 580 72045
##  $ location_name_freq              : int  45164 31532 13827 45164 31532 71139 31532 71139 1982 4003
##  $ bias_desc_freq                  : int  5643 81440 81440 27115 29958 23860 81440 81440 592 15442
##  $ victim_types_freq               : int  191377 191377 191377 191377 1600 191377 14563 191377 1913
##  $ multiple_offense_freq           : int  233460 233460 233460 233460 233460 233460 233460 233460 23
##  $ multiple_bias_freq              : int  239987 239987 239987 239987 239987 239987 239987 239987 23
##  $ Anti_semitic_crimes_freq        : int  211194 211194 211194 211194 30469 211194 211194 211194 21
##  $ Anti_semitic_crimes             : num  0 0 0 0 1 0 0 0 0 0 ...
```

#Checking to see if we need to weight anything #Checking to see if we need to weight anything

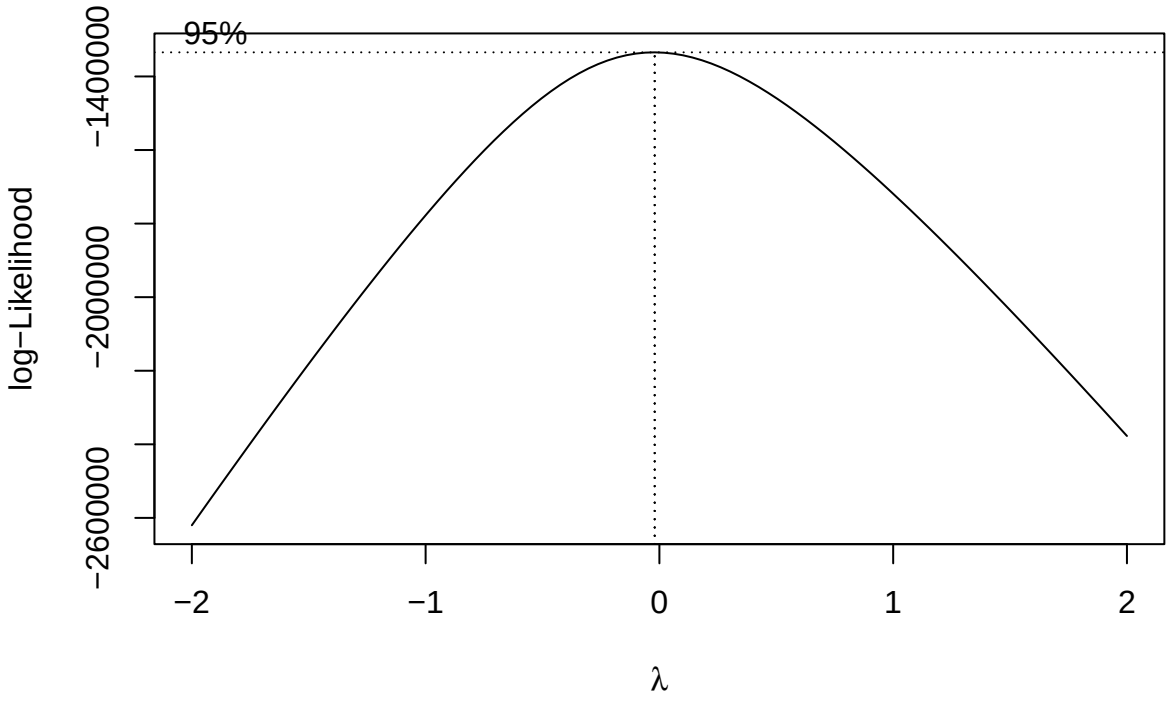
```
# # Fit a linear regression model with all variables
# lm_model <- lm(Anti_semitic_crimes ~ ., data = df_hate_train)
#
# # Extract residuals
# residuals_all <- residuals(lm_model)
#
# # Plot residuals for each predictor variable
# par(mfrow = c(2, 2))  # Set up a 2x2 grid for subplots
#
# for (variable in names(df_hate_train)) {
#   if (variable != "Anti_semitic_crimes") {  # Exclude the response variable
#     plot(df_hate_train[[variable]], residuals_all,
#          main = paste("Residuals vs", variable),
#          xlab = variable, ylab = "Residuals")
#   }
# }
#
# # Reset the plotting parameters
# par(mfrow = c(1, 1))
```

```
# Convert a DataFrame column to a list
freq_list <- as.numeric(as.list(df_hate_train$data_year_freq))

#find optimal lambda for Box-Cox transformation
bc <- boxcox(freq_list~ 1, lambda = seq(-2,2,0.1))
```



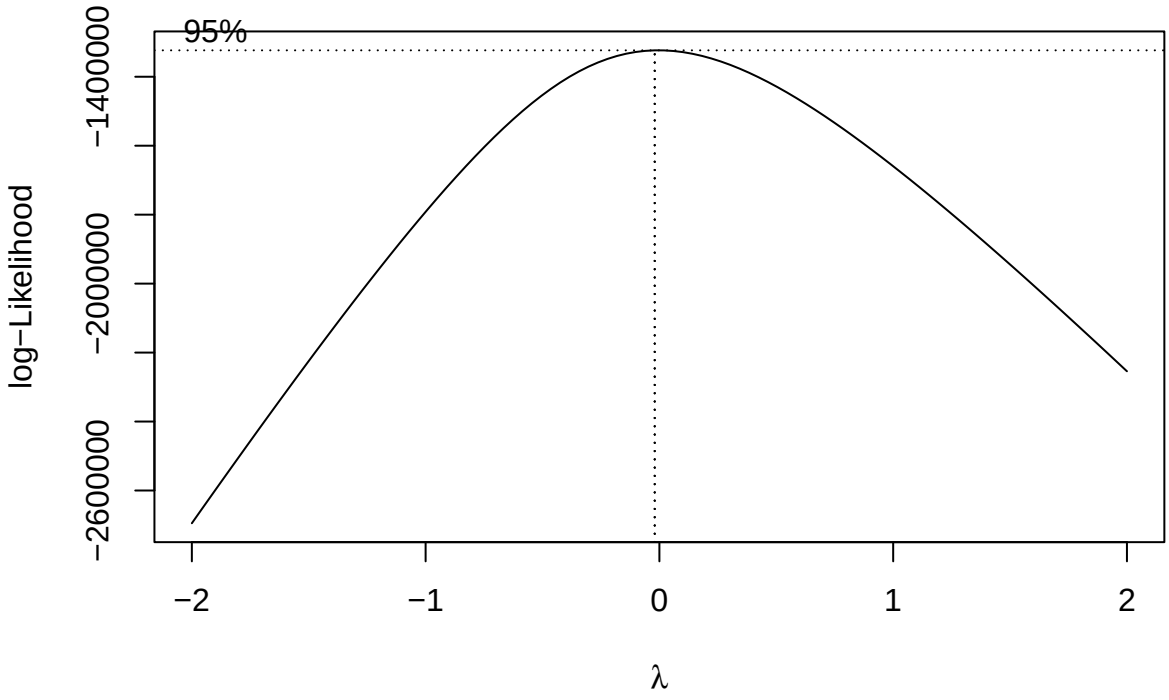

```
lambda_data_year <- bc$x[which.max(bc$y)]

# Apply the Box-Cox transformation
norm_data_year_freq <- (freq_list^lambda_data_year-1)/lambda_data_year

rm(bc,freq_list)
# hist(data_year_freq_norm )
# hist(df_hate_train$data_year_freq)
```

```
# Convert a DataFrame column to a list
freq_list <- as.numeric(as.list(df_hate_train$ori_freq))

#find optimal lambda for Box-Cox transformation
bc <- boxcox(freq_list~ 1, lambda = seq(-2,2,0.1))
```



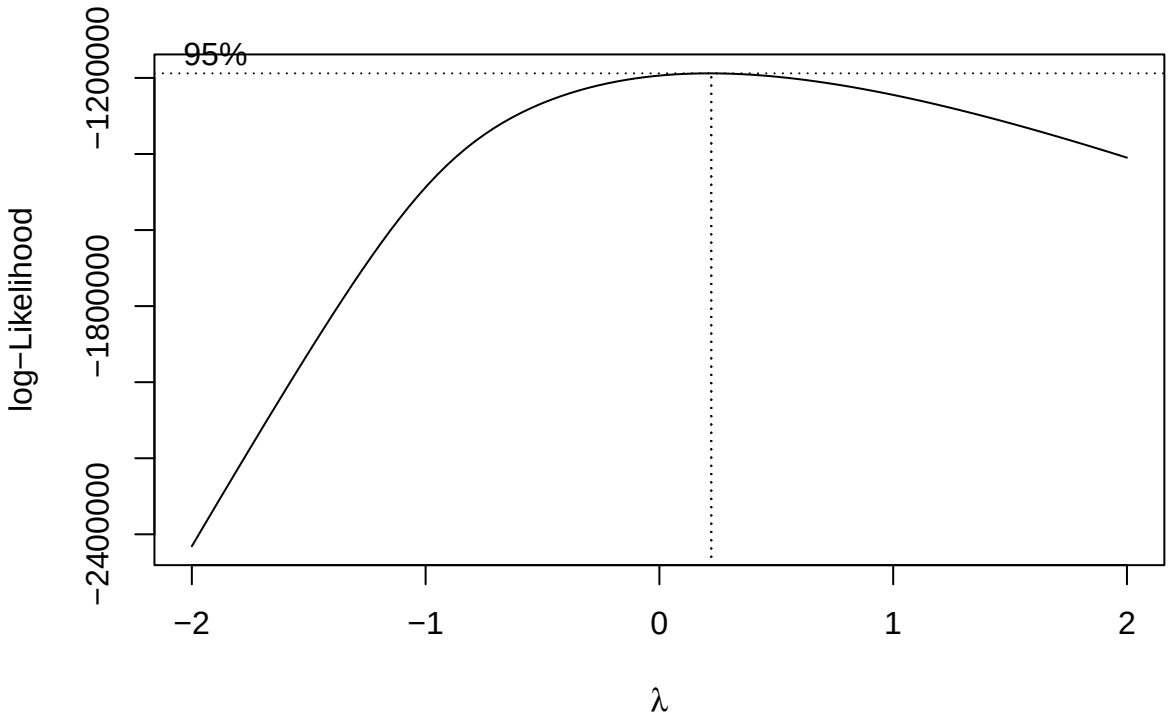

```
lambda_ori <- bc$x[which.max(bc$y)]

# Apply the Box-Cox transformation
norm_ori_freq <- (freq_list^lambda_ori-1)/lambda_ori

rm(bc,freq_list)
# hist(data_year_freq_norm )
# hist(df_hate_train$data_year_freq)
```

```
# Convert a DataFrame column to a list
freq_list <- as.numeric(as.list(df_hate_train$pug_agency_name_freq))

#find optimal lambda for Box-Cox transformation
bc <- boxcox(freq_list~ 1, lambda = seq(-2,2,0.1))
```



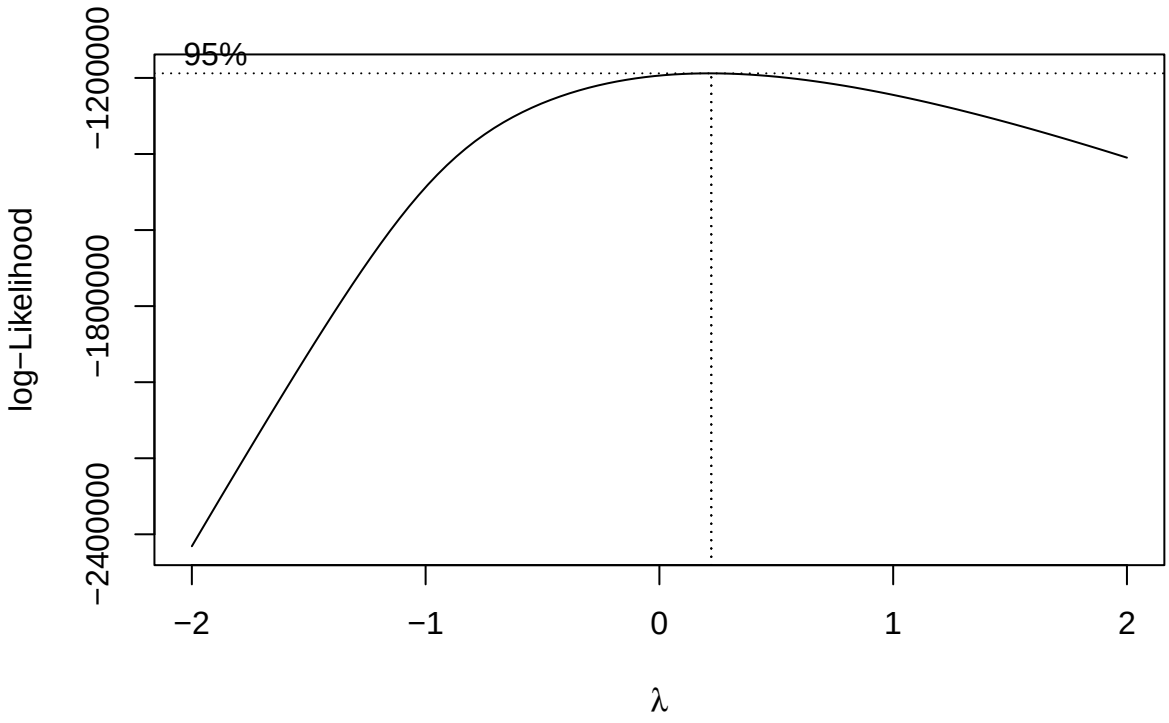

```
lambda_pug_agency_name <- bc$x[which.max(bc$y)]

# Apply the Box-Cox transformation
norm_pug_agency_name_freq <- (freq_list^lambda_pug_agency_name-1)/lambda_pug_agency_name

rm(bc,freq_list)
# hist(data_year_freq_norm )
# hist(df_hate_train$pug_agency_name_freq)
```

```
# Convert a DataFrame column to a list
freq_list <- as.numeric(as.list(df_hate_train$state_abbr_freq))

#find optimal lambda for Box-Cox transformation
bc <- boxcox(freq_list~ 1, lambda = seq(-2,2,0.1))
```



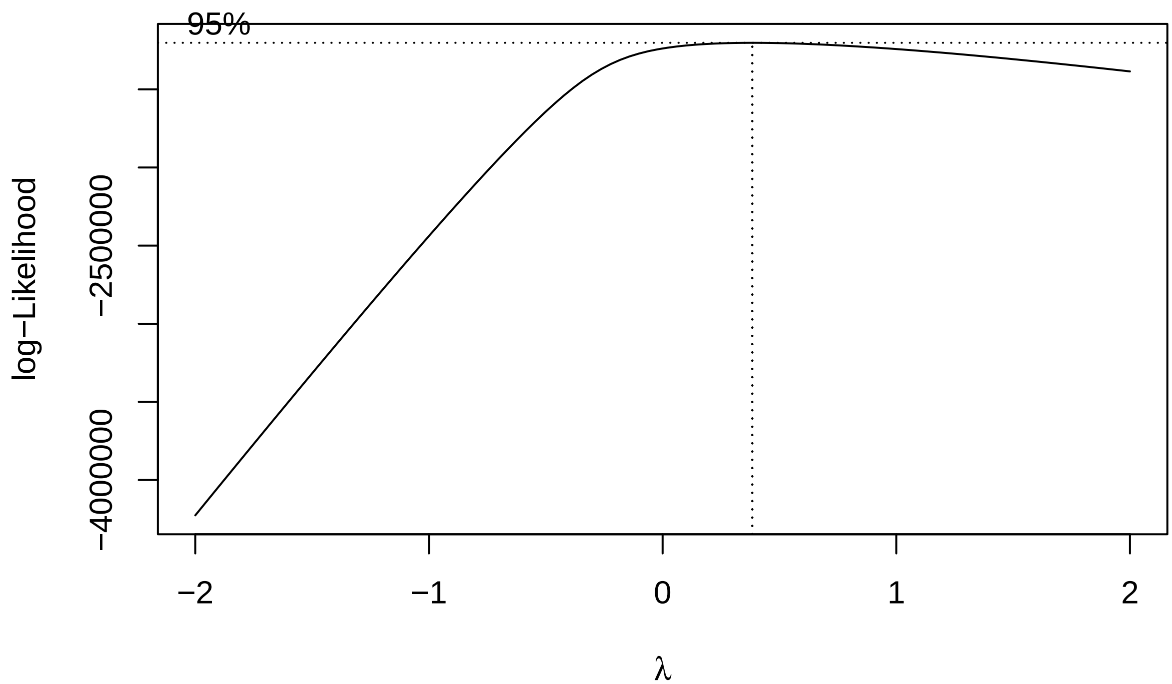

```
lambda_state_abbr <- bc$x[which.max(bc$y)]

# Apply the Box-Cox transformation
norm_state_abbr_freq <- (freq_list^lambda_state_abbr-1)/lambda_state_abbr

rm(bc,freq_list)
# hist(data_year_freq_norm )
# hist(df_hate_train$state_abbr_freq)
```

```
# Convert a DataFrame column to a list
freq_list <- as.numeric(as.list(df_hate_train$state_name_freq))

#find optimal lambda for Box-Cox transformation
bc <- boxcox(freq_list~ 1, lambda = seq(-2,2,0.1))
```




```
lambda_state_name <- bc$x[which.max(bc$y)]

# Apply the Box-Cox transformation
norm_state_name_freq <- (freq_list^lambda_state_name-1)/lambda_state_name

rm(bc,freq_list)
# hist(data_year_freq_norm )
# hist(df_hate_train$state_name_freq)
```

```
# Convert a DataFrame column to a list
freq_list <- as.numeric(as.list(df_hate_train$bias_desc_freq))

#find optimal lambda for Box-Cox transformation
bc <- boxcox(freq_list~ 1, lambda = seq(-2,2,0.1))
```



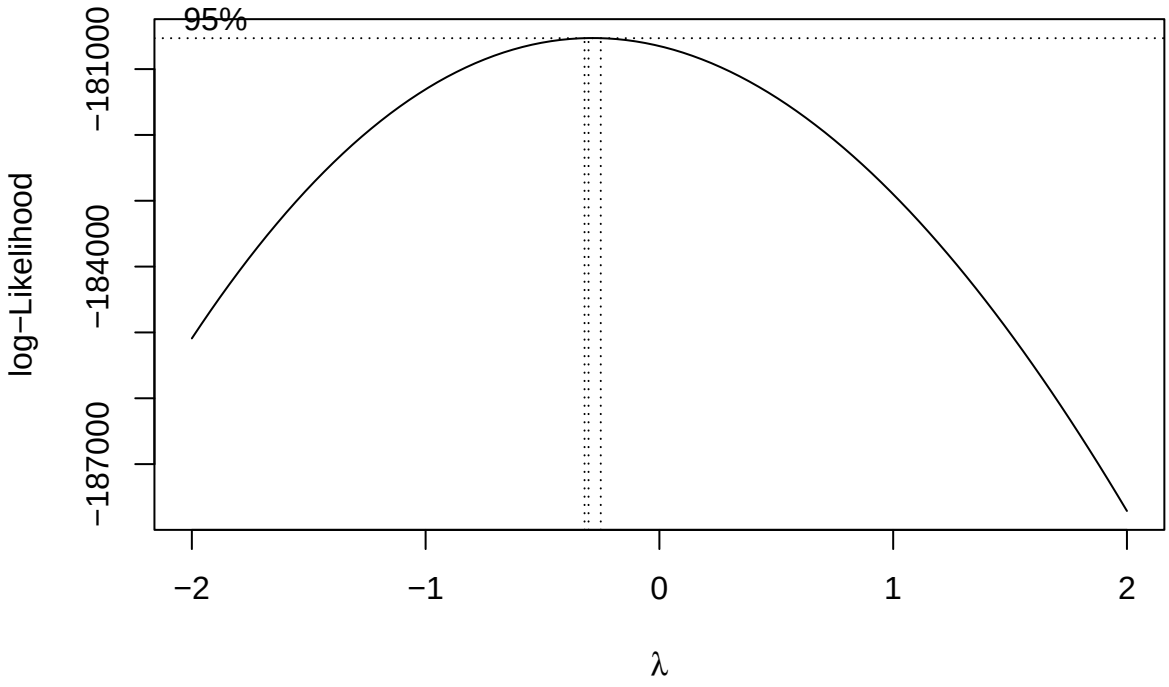

```
lambda_bias_desc <- bc$x[which.max(bc$y)]

# Apply the Box-Cox transformation
norm_bias_desc_freq <- (freq_list^lambda_bias_desc-1)/lambda_bias_desc

rm(bc,freq_list)
# hist(data_year_freq_norm )
# hist(df_hate_train$bias_desc_freq)
```

```
df_train_model <- cbind(df_hate_train,
                 as.data.frame(norm_bias_desc_freq),
                 as.data.frame(norm_data_year_freq),
                 as.data.frame(norm_ori_freq),
                 as.data.frame(norm_pug_agency_name_freq),
                 as.data.frame(norm_state_abbr_freq),
                 as.data.frame(norm_state_name_freq))
```

@@@ @@@@@@@@@@@@@@@@@@@@@@@@@@@@
@@@ @@@@@@@@@@@@@@@@@@@@@@@@@@@@
@@@ @@@@@@@@@@@@@@@@@@@@@@@@@@@@
@@@ @@@@@@@@@@@@@@@@@@@@@@@@@@@@

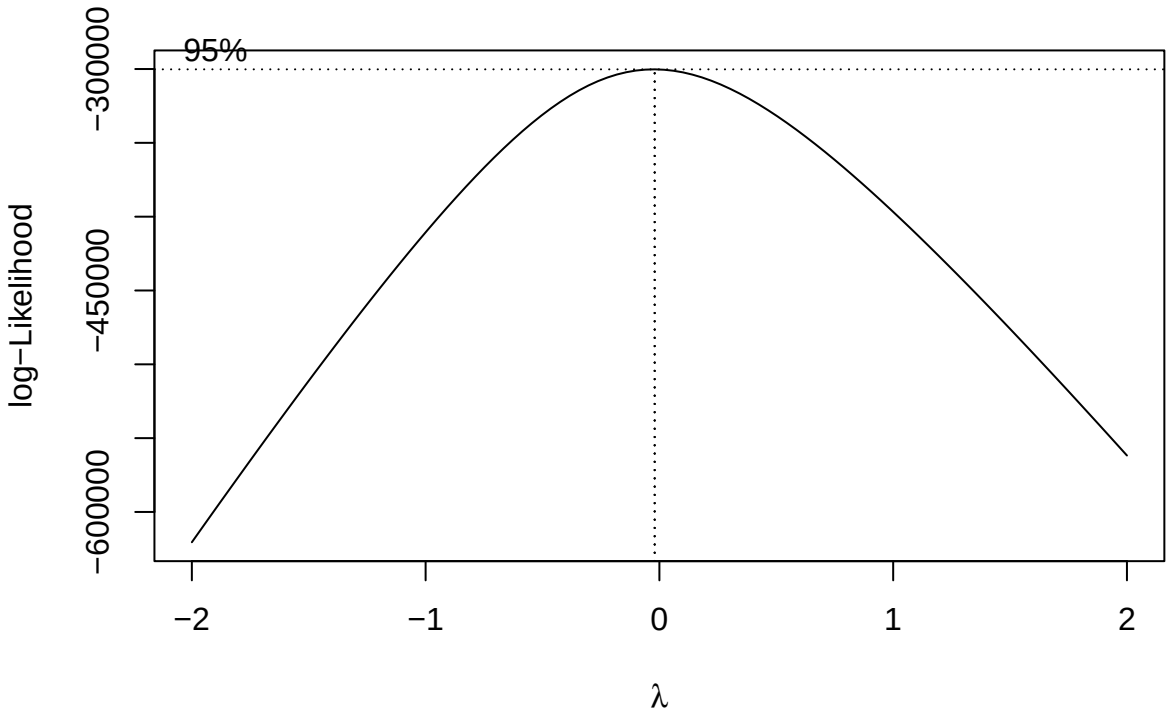
```
# Convert a DataFrame column to a list
freq_list <- as.numeric(as.list(df_hate_test$data_year_freq))
```

```r
#find optimal lambda for Box-Cox transformation
bc <- boxcox(freq_list~ 1, lambda = seq(-2,2,0.1))
```



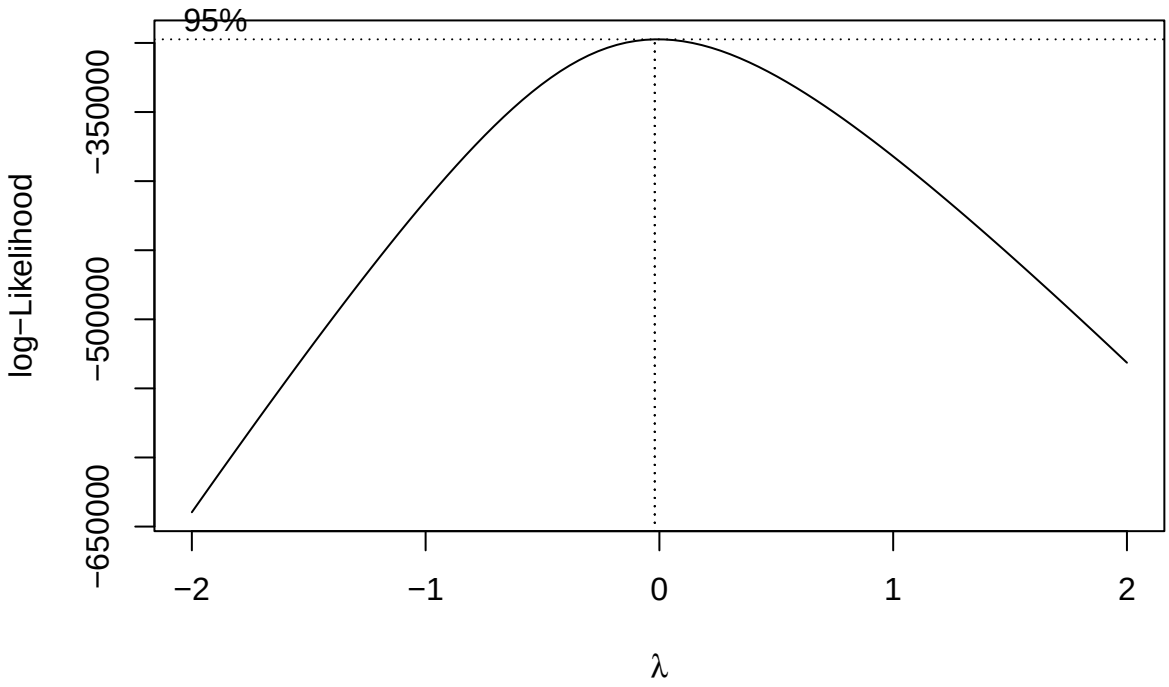

```r
lambda_data_year <- bc$x[which.max(bc$y)]

# Apply the Box-Cox transformation
norm_data_year_freq <- (freq_list^lambda_data_year-1)/lambda_data_year

rm(bc,freq_list)
# hist(data_year_freq_norm )
# hist(df_hate_test$data_year_freq)
```

```r
# Convert a DataFrame column to a list
freq_list <- as.numeric(as.list(df_hate_test$ori_freq))

#find optimal lambda for Box-Cox transformation
bc <- boxcox(freq_list~ 1, lambda = seq(-2,2,0.1))
```



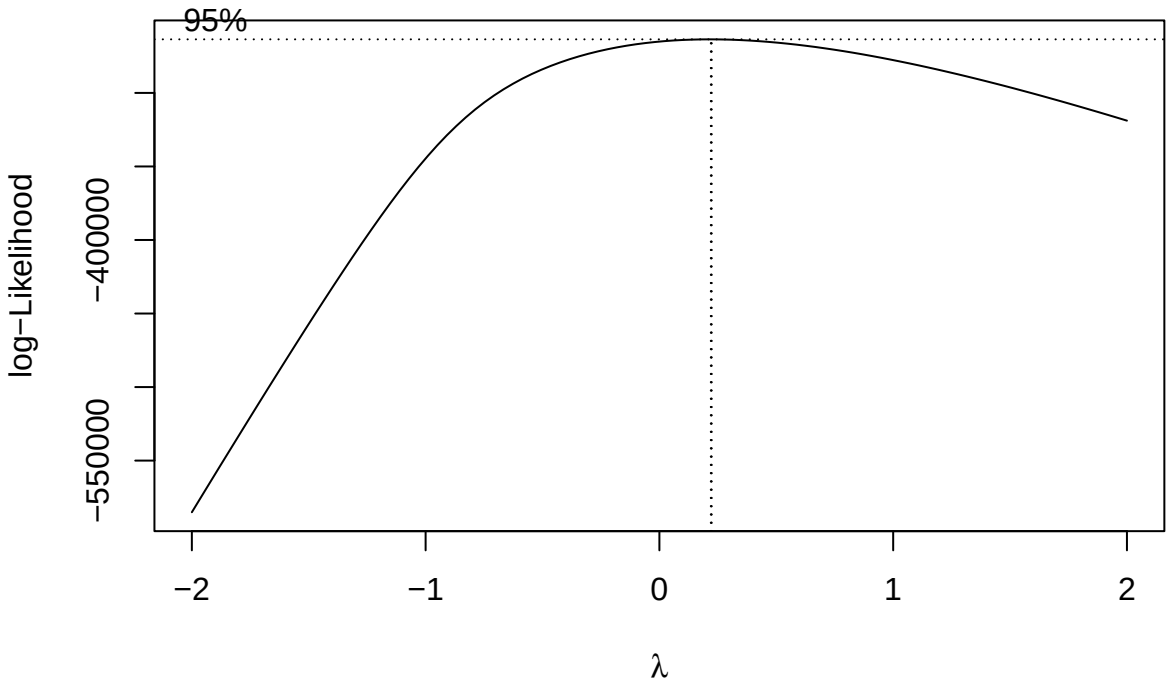

```
lambda_ori <- bc$x[which.max(bc$y)]

# Apply the Box-Cox transformation
norm_ori_freq <- (freq_list^lambda_ori-1)/lambda_ori

rm(bc,freq_list)
# hist(data_year_freq_norm )
# hist(df_hate_test$data_year_freq)
```

```
# Convert a DataFrame column to a list
freq_list <- as.numeric(as.list(df_hate_test$pug_agency_name_freq))

#find optimal lambda for Box-Cox transformation
bc <- boxcox(freq_list~ 1, lambda = seq(-2,2,0.1))
```



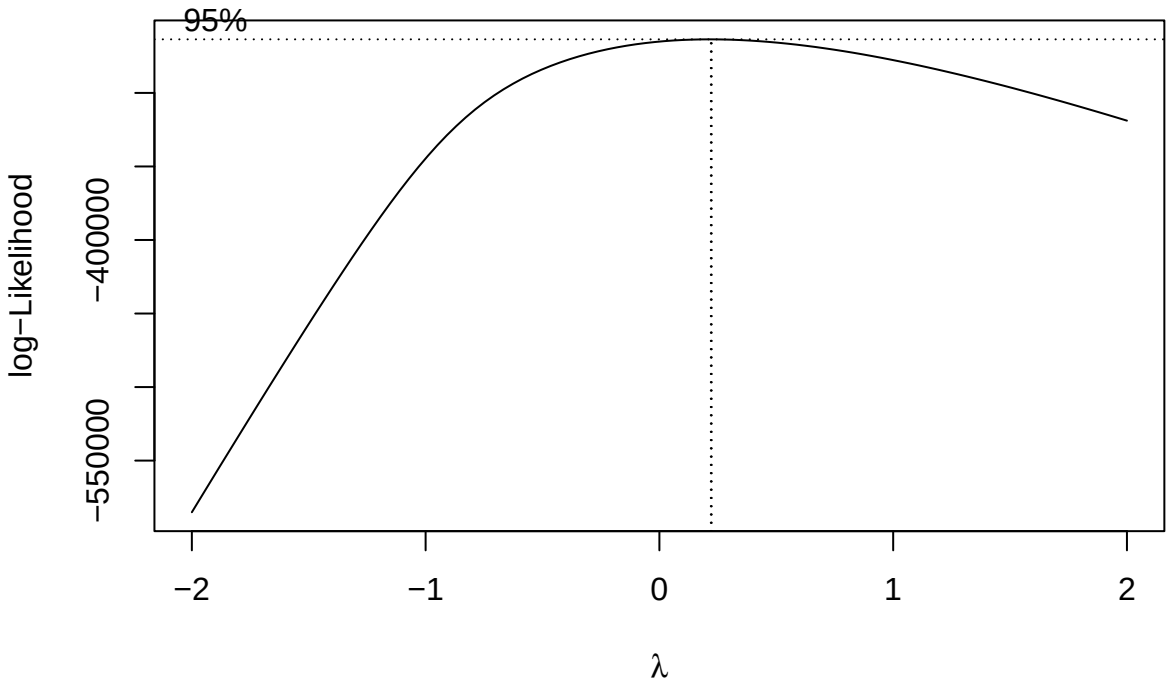

```
lambda_pug_agency_name <- bc$x[which.max(bc$y)]

# Apply the Box-Cox transformation
norm_pug_agency_name_freq <- (freq_list^lambda_pug_agency_name-1)/lambda_pug_agency_name

rm(bc,freq_list)
# hist(data_year_freq_norm )
# hist(df_hate_test$pug_agency_name_freq)
```

```
# Convert a DataFrame column to a list
freq_list <- as.numeric(as.list(df_hate_test$state_abbr_freq))

#find optimal lambda for Box-Cox transformation
bc <- boxcox(freq_list~ 1, lambda = seq(-2,2,0.1))
```



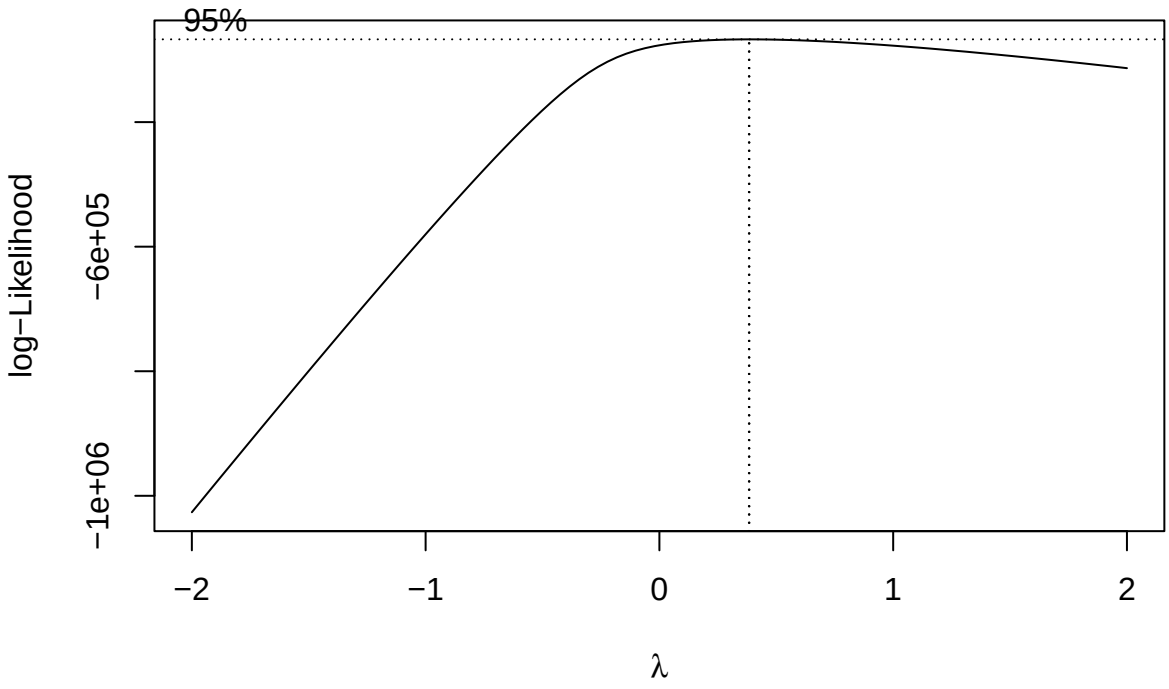

```
lambda_state_abbr <- bc$x[which.max(bc$y)]

# Apply the Box-Cox transformation
norm_state_abbr_freq <- (freq_list^lambda_state_abbr-1)/lambda_state_abbr

rm(bc,freq_list)
# hist(data_year_freq_norm )
# hist(df_hate_test$state_abbr_freq)
```

```
# Convert a DataFrame column to a list
freq_list <- as.numeric(as.list(df_hate_test$state_name_freq))

#find optimal lambda for Box-Cox transformation
bc <- boxcox(freq_list~ 1, lambda = seq(-2,2,0.1))
```




```
lambda_state_name <- bc$x[which.max(bc$y)]

# Apply the Box-Cox transformation
norm_state_name_freq <- (freq_list^lambda_state_name-1)/lambda_state_name

rm(bc,freq_list)
# hist(data_year_freq_norm )
# hist(df_hate_test$state_name_freq)
```

```
# Convert a DataFrame column to a list
freq_list <- as.numeric(as.list(df_hate_test$bias_desc_freq))

#find optimal lambda for Box-Cox transformation
bc <- boxcox(freq_list~ 1, lambda = seq(-2,2,0.1))
```




```
lambda_bias_desc <- bc$x[which.max(bc$y)]

# Apply the Box-Cox transformation
norm_bias_desc_freq <- (freq_list^lambda_bias_desc-1)/lambda_bias_desc

rm(bc,freq_list)
# hist(data_year_freq_norm )
# hist(df_hate_test$bias_desc_freq)
```

```
df_test_model <- cbind(df_hate_test,
                  as.data.frame(norm_bias_desc_freq),
                  as.data.frame(norm_data_year_freq),
                  as.data.frame(norm_ori_freq),
                  as.data.frame(norm_pug_agency_name_freq),
                  as.data.frame(norm_state_abbr_freq),
                  as.data.frame(norm_state_name_freq))
```