Data Science Language Analysis

EDA Report

Nazli Ozum Kafaee
Prash Medirattaa
Avinash Prabhakaran

2018-04-15

Contents


```
#Loading the required packages
suppressPackageStartupMessages(library(tidyverse))
```

```
#Reading in processed data.
responses <- read.csv(file = "../docs/survey_results_clean.csv")
```

```
#Binary encoding the response variable. Python -> 1; R -> 0
data <- responses %>% mutate(binary = if_else(preference == "Python", 1, 0))

#Releveling the reference task from Data Viz -> Machine Learning
data_relevel <- data
data_relevel$task <-relevel(data$task,ref="Machine Learning")

#Fitting a GLM without any confounding variables.
mod <- glm(binary ~ task, data = data)
summary(mod)
```

```
##
## Call:
## glm(formula = binary ~ task, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8542  -0.3044   0.1458   0.1458   0.6956
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            0.30435    0.08722   3.489 0.000782 ***
## taskData wrangling     0.19565    0.14180   1.380 0.171403
## taskMachine Learning   0.54982    0.10608   5.183 1.54e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1749845)
##
##     Null deviance: 19.412  on 84  degrees of freedom
## Residual deviance: 14.349  on 82  degrees of freedom
## AIC: 98.005
##
## Number of Fisher Scoring iterations: 2
```

```r
mod <- glm(binary ~ task, data = data_relevel)
summary(mod)
```

```
##
## Call:
## glm(formula = binary ~ task, data = data_relevel)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8542  -0.3044   0.1458   0.1458   0.6956
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              0.85417    0.06038  14.147  < 2e-16 ***
## taskData visualization -0.54982    0.10608  -5.183 1.54e-06 ***
## taskData wrangling     -0.35417    0.12706  -2.787   0.0066 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1749845)
##
##     Null deviance: 19.412  on 84  degrees of freedom
## Residual deviance: 14.349  on 82  degrees of freedom
## AIC: 98.005
##
## Number of Fisher Scoring iterations: 2
```

```r
#Fitting GLM with all the confounding variables.
responses %>% colnames()
```

```
## [1] "background" "experience" "attitude"   "first"      "preference"
## [6] "task"       "active"
```

```r
mod <- glm(binary ~ task + background + experience + attitude + first + active, data = data)
summary(mod)
```

```
##
## Call:
## glm(formula = binary ~ task + background + experience + attitude +
##     first + active, data = data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.70350  -0.23889   0.06545   0.24716   0.94279
##
## Coefficients:
##                                                  Estimate Std. Error
## (Intercept)                                       0.11169    0.36299
## taskData wrangling                                0.19022    0.15196
## taskMachine Learning                              0.35832    0.12377
## backgroundComputer Science / Computer Engineering 0.28719    0.17198
## backgroundEngineering                             0.01071    0.18007
## backgroundMathematics / Statistics                0.09489    0.15570
## backgroundOther                                   0.03331    0.15882
## experienceLess than 1                            -0.01563    0.10983
```

```
## experienceMore than 5                                -0.21754    0.16123
## attitudeNo                                           -0.12016    0.23107
## attitudeYes                                           0.13258    0.13971
## firstJava                                            -0.25169    0.17551
## firstMatlab                                          -0.02976    0.19603
## firstOther                                            0.07177    0.15472
## firstPython                                           0.23702    0.14471
## firstR                                               -0.26514    0.19120
## firstSAS                                             -0.19857    0.23772
## active2                                               0.04478    0.22965
## active3                                               0.27592    0.24604
## active4                                               0.14778    0.29197
## active5 or more                                       0.34911    0.28842
##                                                     t value Pr(>|t|)
## (Intercept)                                           0.308  0.75932
## taskData wrangling                                    1.252  0.21520
## taskMachine Learning                                  2.895  0.00518 **
## backgroundComputer Science / Computer Engineering     1.670  0.09983 .
## backgroundEngineering                                 0.059  0.95277
## backgroundMathematics / Statistics                    0.609  0.54439
## backgroundOther                                       0.210  0.83452
## experienceLess than 1                                -0.142  0.88727
## experienceMore than 5                                -1.349  0.18199
## attitudeNo                                           -0.520  0.60484
## attitudeYes                                           0.949  0.34620
## firstJava                                            -1.434  0.15643
## firstMatlab                                          -0.152  0.87979
## firstOther                                            0.464  0.64433
## firstPython                                           1.638  0.10635
## firstR                                               -1.387  0.17033
## firstSAS                                             -0.835  0.40666
## active2                                               0.195  0.84602
## active3                                               1.121  0.26628
## active4                                               0.506  0.61448
## active5 or more                                       1.210  0.23055
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1475587)
##
##     Null deviance: 19.4118  on 84  degrees of freedom
## Residual deviance:  9.4438  on 64  degrees of freedom
## AIC: 98.449
##
## Number of Fisher Scoring iterations: 2
```

```
mod <- glm(binary ~ task + background + experience + attitude + first + active, data = data_relevel)
summary(mod)
```

```
##
## Call:
## glm(formula = binary ~ task + background + experience + attitude +
##     first + active, data = data_relevel)
##
## Deviance Residuals:
```

```
##      Min        1Q    Median        3Q       Max
## -0.70350  -0.23889   0.06545   0.24716   0.94279
##
## Coefficients:
##                                                  Estimate Std. Error
## (Intercept)                                       0.47001    0.36064
## taskData visualization                           -0.35832    0.12377
## taskData wrangling                               -0.16810    0.13950
## backgroundComputer Science / Computer Engineering  0.28719    0.17198
## backgroundEngineering                             0.01071    0.18007
## backgroundMathematics / Statistics                0.09489    0.15570
## backgroundOther                                   0.03331    0.15882
## experienceLess than 1                            -0.01563    0.10983
## experienceMore than 5                            -0.21754    0.16123
## attitudeNo                                       -0.12016    0.23107
## attitudeYes                                       0.13258    0.13971
## firstJava                                        -0.25169    0.17551
## firstMatlab                                      -0.02976    0.19603
## firstOther                                        0.07177    0.15472
## firstPython                                       0.23702    0.14471
## firstR                                           -0.26514    0.19120
## firstSAS                                         -0.19857    0.23772
## active2                                           0.04478    0.22965
## active3                                           0.27592    0.24604
## active4                                           0.14778    0.29197
## active5 or more                                   0.34911    0.28842
##                                                  t value Pr(>|t|)
## (Intercept)                                        1.303  0.19716
## taskData visualization                            -2.895  0.00518 **
## taskData wrangling                                -1.205  0.23265
## backgroundComputer Science / Computer Engineering   1.670  0.09983 .
## backgroundEngineering                              0.059  0.95277
## backgroundMathematics / Statistics                 0.609  0.54439
## backgroundOther                                    0.210  0.83452
## experienceLess than 1                             -0.142  0.88727
## experienceMore than 5                             -1.349  0.18199
## attitudeNo                                        -0.520  0.60484
## attitudeYes                                        0.949  0.34620
## firstJava                                         -1.434  0.15643
## firstMatlab                                       -0.152  0.87979
## firstOther                                         0.464  0.64433
## firstPython                                        1.638  0.10635
## firstR                                            -1.387  0.17033
## firstSAS                                          -0.835  0.40666
## active2                                            0.195  0.84602
## active3                                            1.121  0.26628
## active4                                            0.506  0.61448
## active5 or more                                    1.210  0.23055
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1475587)
##
##     Null deviance: 19.4118  on 84  degrees of freedom
```

```
## Residual deviance:  9.4438  on 64  degrees of freedom
## AIC: 98.449
##
## Number of Fisher Scoring iterations: 2
```

```
#Removing Attitude as Confounder as change
mod <- glm(binary ~ task + background + experience + attitude + first + active, data = data)
summary(mod)
```

```
##
## Call:
## glm(formula = binary ~ task + background + experience + attitude +
##     first + active, data = data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.70350  -0.23889   0.06545   0.24716   0.94279
##
## Coefficients:
##                                                 Estimate Std. Error
## (Intercept)                                      0.11169    0.36299
## taskData wrangling                               0.19022    0.15196
## taskMachine Learning                             0.35832    0.12377
## backgroundComputer Science / Computer Engineering  0.28719    0.17198
## backgroundEngineering                            0.01071    0.18007
## backgroundMathematics / Statistics               0.09489    0.15570
## backgroundOther                                  0.03331    0.15882
## experienceLess than 1                           -0.01563    0.10983
## experienceMore than 5                           -0.21754    0.16123
## attitudeNo                                      -0.12016    0.23107
## attitudeYes                                      0.13258    0.13971
## firstJava                                       -0.25169    0.17551
## firstMatlab                                     -0.02976    0.19603
## firstOther                                       0.07177    0.15472
## firstPython                                      0.23702    0.14471
## firstR                                          -0.26514    0.19120
## firstSAS                                        -0.19857    0.23772
## active2                                          0.04478    0.22965
## active3                                          0.27592    0.24604
## active4                                          0.14778    0.29197
## active5 or more                                  0.34911    0.28842
##                                                 t value Pr(>|t|)
## (Intercept)                                       0.308  0.75932
## taskData wrangling                                1.252  0.21520
## taskMachine Learning                              2.895  0.00518 **
## backgroundComputer Science / Computer Engineering   1.670  0.09983 .
## backgroundEngineering                             0.059  0.95277
## backgroundMathematics / Statistics                0.609  0.54439
## backgroundOther                                   0.210  0.83452
## experienceLess than 1                            -0.142  0.88727
## experienceMore than 5                            -1.349  0.18199
## attitudeNo                                       -0.520  0.60484
## attitudeYes                                       0.949  0.34620
## firstJava                                        -1.434  0.15643
## firstMatlab                                      -0.152  0.87979
```

```
## firstOther                                              0.464  0.64433
## firstPython                                             1.638  0.10635
## firstR                                                 -1.387  0.17033
## firstSAS                                               -0.835  0.40666
## active2                                                 0.195  0.84602
## active3                                                 1.121  0.26628
## active4                                                 0.506  0.61448
## active5 or more                                         1.210  0.23055
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1475587)
##
##     Null deviance: 19.4118  on 84  degrees of freedom
## Residual deviance:  9.4438  on 64  degrees of freedom
## AIC: 98.449
##
## Number of Fisher Scoring iterations: 2
```

```
#Removing Attitude as Confounder as change
mod <- glm(binary ~ task + background + experience + first + active, data = data_relevel)
summary(mod)
```

```
##
## Call:
## glm(formula = binary ~ task + background + experience + first +
##     active, data = data_relevel)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.68805  -0.24950   0.05242   0.22798   0.97175
##
## Coefficients:
##                                                   Estimate Std. Error
## (Intercept)                                      0.5899898  0.3049358
## taskData visualization                          -0.4405692  0.1079008
## taskData wrangling                              -0.1999476  0.1367326
## backgroundComputer Science / Computer Engineering  0.3008693  0.1658952
## backgroundEngineering                            0.0009851  0.1736922
## backgroundMathematics / Statistics               0.0553966  0.1479245
## backgroundOther                                  0.0648374  0.1566880
## experienceLess than 1                           -0.0469194  0.1072424
## experienceMore than 5                           -0.1960933  0.1597792
## firstJava                                       -0.2244379  0.1731371
## firstMatlab                                     -0.0580671  0.1850675
## firstOther                                       0.1002980  0.1520268
## firstPython                                      0.2314974  0.1443899
## firstR                                          -0.2306867  0.1848307
## firstSAS                                        -0.1693769  0.2364815
## active2                                          0.0446813  0.2267958
## active3                                          0.2806987  0.2411687
## active4                                          0.1525118  0.2910528
## active5 or more                                  0.3280160  0.2842357
##                                                 t value Pr(>|t|)
## (Intercept)                                       1.935 0.057302 .
```

```
## taskData visualization                                -4.083 0.000122 ***
## taskData wrangling                                    -1.462 0.148398
## backgroundComputer Science / Computer Engineering      1.814 0.074284 .
## backgroundEngineering                                  0.006 0.995492
## backgroundMathematics / Statistics                     0.374 0.709239
## backgroundOther                                        0.414 0.680363
## experienceLess than 1                                 -0.438 0.663171
## experienceMore than 5                                 -1.227 0.224078
## firstJava                                             -1.296 0.199386
## firstMatlab                                           -0.314 0.754691
## firstOther                                             0.660 0.511717
## firstPython                                            1.603 0.113648
## firstR                                                -1.248 0.216405
## firstSAS                                              -0.716 0.476371
## active2                                                0.197 0.844424
## active3                                                1.164 0.248651
## active4                                                0.524 0.602034
## active5 or more                                        1.154 0.252650
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1472337)
##
##     Null deviance: 19.4118  on 84  degrees of freedom
## Residual deviance:  9.7174  on 66  degrees of freedom
## AIC: 96.877
##
## Number of Fisher Scoring iterations: 2
```

```r
#Removing Experience as Confounder
mod <- glm(binary ~ task + background + first + active, data = data)
summary(mod)
```

```
##
## Call:
## glm(formula = binary ~ task + background + first + active, data = data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.70057  -0.23149   0.05643   0.20568   0.98900
##
## Coefficients:
##                                                   Estimate Std. Error
## (Intercept)                                       0.123789   0.290677
## taskData wrangling                                0.231847   0.142430
## taskMachine Learning                              0.439318   0.106747
## backgroundComputer Science / Computer Engineering 0.297498   0.161287
## backgroundEngineering                             0.001356   0.170932
## backgroundMathematics / Statistics                0.051551   0.143090
## backgroundOther                                   0.055676   0.153693
## firstJava                                        -0.237718   0.169863
## firstMatlab                                      -0.054294   0.183910
## firstOther                                        0.081320   0.147488
## firstPython                                       0.226013   0.142746
## firstR                                           -0.224611   0.183648
```

```
## firstSAS                                              -0.217960   0.231030
## active2                                                0.056145   0.220587
## active3                                                0.291709   0.229980
## active4                                                0.115219   0.279899
## active5 or more                                        0.302607   0.270597
##                                                       t value Pr(>|t|)
## (Intercept)                                             0.426 0.671552
## taskData wrangling                                      1.628 0.108194
## taskMachine Learning                                    4.116 0.000107 ***
## backgroundComputer Science / Computer Engineering       1.845 0.069463 .
## backgroundEngineering                                   0.008 0.993696
## backgroundMathematics / Statistics                      0.360 0.719763
## backgroundOther                                         0.362 0.718288
## firstJava                                              -1.399 0.166218
## firstMatlab                                            -0.295 0.768724
## firstOther                                              0.551 0.583187
## firstPython                                             1.583 0.117987
## firstR                                                 -1.223 0.225532
## firstSAS                                               -0.943 0.348803
## active2                                                 0.255 0.799857
## active3                                                 1.268 0.208977
## active4                                                 0.412 0.681895
## active5 or more                                         1.118 0.267377
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1463496)
##
##     Null deviance: 19.4118  on 84  degrees of freedom
## Residual deviance:  9.9518  on 68  degrees of freedom
## AIC: 94.903
##
## Number of Fisher Scoring iterations: 2
```

```
#Removing Experience as Confounder
mod <- glm(binary ~ task + background + first + active, data = data_relevel)
summary(mod)
```

```
##
## Call:
## glm(formula = binary ~ task + background + first + active, data = data_relevel)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.70057  -0.23149   0.05643   0.20568   0.98900
##
## Coefficients:
##                                                       Estimate Std. Error
## (Intercept)                                           0.563107   0.280086
## taskData visualization                               -0.439318   0.106747
## taskData wrangling                                   -0.207470   0.128130
## backgroundComputer Science / Computer Engineering     0.297498   0.161287
## backgroundEngineering                                 0.001356   0.170932
## backgroundMathematics / Statistics                    0.051551   0.143090
## backgroundOther                                       0.055676   0.153693
```

```
## firstJava                                            -0.237718   0.169863
## firstMatlab                                          -0.054294   0.183910
## firstOther                                            0.081320   0.147488
## firstPython                                           0.226013   0.142746
## firstR                                               -0.224611   0.183648
## firstSAS                                             -0.217960   0.231030
## active2                                               0.056145   0.220587
## active3                                               0.291709   0.229980
## active4                                               0.115219   0.279899
## active5 or more                                       0.302607   0.270597
##                                                     t value Pr(>|t|)
## (Intercept)                                           2.010 0.048349 *
## taskData visualization                               -4.116 0.000107 ***
## taskData wrangling                                   -1.619 0.110029
## backgroundComputer Science / Computer Engineering     1.845 0.069463 .
## backgroundEngineering                                 0.008 0.993696
## backgroundMathematics / Statistics                    0.360 0.719763
## backgroundOther                                       0.362 0.718288
## firstJava                                            -1.399 0.166218
## firstMatlab                                          -0.295 0.768724
## firstOther                                            0.551 0.583187
## firstPython                                           1.583 0.117987
## firstR                                               -1.223 0.225532
## firstSAS                                             -0.943 0.348803
## active2                                               0.255 0.799857
## active3                                               1.268 0.208977
## active4                                               0.412 0.681895
## active5 or more                                       1.118 0.267377
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1463496)
##
##     Null deviance: 19.4118  on 84  degrees of freedom
## Residual deviance:  9.9518  on 68  degrees of freedom
## AIC: 94.903
##
## Number of Fisher Scoring iterations: 2
```

```r
#Removing active as Confounder
mod <- glm(binary ~ task + background + first, data = data)
summary(mod)
```

```
##
## Call:
## glm(formula = binary ~ task + background + first, data = data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.83143  -0.22886  -0.01263   0.23906   0.98530
##
## Coefficients:
##                                                   Estimate Std. Error
## (Intercept)                                        0.27336    0.19617
## taskData wrangling                                 0.21416    0.14134
```

```
## taskMachine Learning                                 0.45688    0.10742
## backgroundComputer Science / Computer Engineering  0.28239    0.15347
## backgroundEngineering                                0.03069    0.17190
## backgroundMathematics / Statistics                   0.04130    0.14440
## backgroundOther                                      0.07504    0.15537
## firstJava                                           -0.25070    0.17082
## firstMatlab                                         -0.05243    0.18254
## firstOther                                           0.10118    0.13909
## firstPython                                          0.16540    0.13925
## firstR                                              -0.29355    0.18121
## firstSAS                                            -0.25866    0.23433
##                                                     t value Pr(>|t|)
## (Intercept)                                           1.393   0.1678
## taskData wrangling                                    1.515   0.1341
## taskMachine Learning                                  4.253 6.24e-05 ***
## backgroundComputer Science / Computer Engineering   1.840   0.0699 .
## backgroundEngineering                                 0.179   0.8588
## backgroundMathematics / Statistics                    0.286   0.7757
## backgroundOther                                       0.483   0.6306
## firstJava                                            -1.468   0.1466
## firstMatlab                                          -0.287   0.7748
## firstOther                                            0.727   0.4693
## firstPython                                           1.188   0.2388
## firstR                                               -1.620   0.1096
## firstSAS                                             -1.104   0.2733
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1514782)
##
##     Null deviance: 19.412  on 84  degrees of freedom
## Residual deviance: 10.906  on 72  degrees of freedom
## AIC: 94.689
##
## Number of Fisher Scoring iterations: 2
```

```r
#Removing active as Confounder
mod <- glm(binary ~ task + background + first, data = data_relevel)
summary(mod)
```

```
##
## Call:
## glm(formula = binary ~ task + background + first, data = data_relevel)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.83143  -0.22886  -0.01263   0.23906   0.98530
##
## Coefficients:
##                                                     Estimate Std. Error
## (Intercept)                                          0.73024    0.16033
## taskData visualization                              -0.45688    0.10742
## taskData wrangling                                  -0.24272    0.12641
## backgroundComputer Science / Computer Engineering  0.28239    0.15347
## backgroundEngineering                                0.03069    0.17190
```

```
## backgroundMathematics / Statistics                 0.04130    0.14440
## backgroundOther                                    0.07504    0.15537
## firstJava                                         -0.25070    0.17082
## firstMatlab                                       -0.05243    0.18254
## firstOther                                         0.10118    0.13909
## firstPython                                        0.16540    0.13925
## firstR                                            -0.29355    0.18121
## firstSAS                                          -0.25866    0.23433
##                                                   t value Pr(>|t|)
## (Intercept)                                         4.555 2.09e-05 ***
## taskData visualization                             -4.253 6.24e-05 ***
## taskData wrangling                                 -1.920   0.0588 .
## backgroundComputer Science / Computer Engineering   1.840   0.0699 .
## backgroundEngineering                               0.179   0.8588
## backgroundMathematics / Statistics                  0.286   0.7757
## backgroundOther                                     0.483   0.6306
## firstJava                                          -1.468   0.1466
## firstMatlab                                        -0.287   0.7748
## firstOther                                          0.727   0.4693
## firstPython                                         1.188   0.2388
## firstR                                             -1.620   0.1096
## firstSAS                                           -1.104   0.2733
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1514782)
##
##     Null deviance: 19.412  on 84  degrees of freedom
## Residual deviance: 10.906  on 72  degrees of freedom
## AIC: 94.689
##
## Number of Fisher Scoring iterations: 2
```

```r
#Removing first as Confounder
mod <- glm(binary ~ task + background, data = data)
summary(mod)
```

```
##
## Call:
## glm(formula = binary ~ task + background, data = data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.82474  -0.28669  -0.04575   0.23942   0.85264
##
## Coefficients:
##                                                   Estimate Std. Error
## (Intercept)                                        0.14736    0.13300
## taskData wrangling                                 0.26757    0.14353
## taskMachine Learning                               0.53805    0.10607
## backgroundComputer Science / Computer Engineering  0.36034    0.13958
## backgroundEngineering                              0.07517    0.15309
## backgroundMathematics / Statistics                 0.10881    0.14353
## backgroundOther                                    0.13933    0.14412
##                                                   t value Pr(>|t|)
```

```
## (Intercept)                                          1.108   0.2713
## taskData wrangling                                   1.864   0.0660 .
## taskMachine Learning                                 5.073 2.59e-06 ***
## backgroundComputer Science / Computer Engineering    2.582   0.0117 *
## backgroundEngineering                                0.491   0.6248
## backgroundMathematics / Statistics                   0.758   0.4507
## backgroundOther                                      0.967   0.3366
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1670895)
##
##     Null deviance: 19.412  on 84  degrees of freedom
## Residual deviance: 13.033  on 78  degrees of freedom
## AIC: 97.83
##
## Number of Fisher Scoring iterations: 2
```

```
#Removing first as Confounder
mod <- glm(binary ~ task + background, data = data_relevel)
summary(mod)
```

```
##
## Call:
## glm(formula = binary ~ task + background, data = data_relevel)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.82474  -0.28669  -0.04575   0.23942   0.85264
##
## Coefficients:
##                                                  Estimate Std. Error
## (Intercept)                                       0.68541    0.11334
## taskData visualization                           -0.53805    0.10607
## taskData wrangling                               -0.27048    0.12789
## backgroundComputer Science / Computer Engineering  0.36034    0.13958
## backgroundEngineering                             0.07517    0.15309
## backgroundMathematics / Statistics                0.10881    0.14353
## backgroundOther                                   0.13933    0.14412
##                                                  t value Pr(>|t|)
## (Intercept)                                        6.047 4.78e-08 ***
## taskData visualization                            -5.073 2.59e-06 ***
## taskData wrangling                                -2.115   0.0376 *
## backgroundComputer Science / Computer Engineering   2.582   0.0117 *
## backgroundEngineering                              0.491   0.6248
## backgroundMathematics / Statistics                 0.758   0.4507
## backgroundOther                                    0.967   0.3366
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1670895)
##
##     Null deviance: 19.412  on 84  degrees of freedom
## Residual deviance: 13.033  on 78  degrees of freedom
## AIC: 97.83
```

```
##
## Number of Fisher Scoring iterations: 2
```

Not Removing first language as the AIC score of the model increases from 94.689 to 97.83.

```
#Removing background as Confounder
mod <- glm(binary ~ task + first, data = data)
summary(mod)
```

```
##
## Call:
## glm(formula = binary ~ task + first, data = data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.93373  -0.19996   0.06627   0.12049   0.96404
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            0.43006    0.12617   3.409  0.00105 **
## taskData wrangling     0.16400    0.13703   1.197  0.23512
## taskMachine Learning   0.44945    0.10556   4.258 5.84e-05 ***
## firstJava             -0.27979    0.16836  -1.662  0.10065
## firstMatlab           -0.09811    0.17530  -0.560  0.57733
## firstOther             0.05421    0.13368   0.406  0.68623
## firstPython            0.11091    0.12582   0.882  0.38081
## firstR                -0.39410    0.16286  -2.420  0.01791 *
## firstSAS              -0.38555    0.20386  -1.891  0.06240 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1546241)
##
##     Null deviance: 19.412  on 84  degrees of freedom
## Residual deviance: 11.751  on 76  degrees of freedom
## AIC: 93.032
##
## Number of Fisher Scoring iterations: 2
```

```
#Removing background as Confounder
mod <- glm(binary ~ task + first, data = data_relevel)
summary(mod)
```

```
##
## Call:
## glm(formula = binary ~ task + first, data = data_relevel)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.93373  -0.19996   0.06627   0.12049   0.96404
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              0.87951    0.08765  10.034 1.43e-15 ***
## taskData visualization -0.44945    0.10556  -4.258 5.84e-05 ***
## taskData wrangling     -0.28545    0.12592  -2.267   0.0262 *
```

```
## firstJava              -0.27979    0.16836  -1.662   0.1007
## firstMatlab            -0.09811    0.17530  -0.560   0.5773
## firstOther              0.05421    0.13368   0.406   0.6862
## firstPython             0.11091    0.12582   0.882   0.3808
## firstR                 -0.39410    0.16286  -2.420   0.0179 *
## firstSAS               -0.38555    0.20386  -1.891   0.0624 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1546241)
##
##     Null deviance: 19.412  on 84  degrees of freedom
## Residual deviance: 11.751  on 76  degrees of freedom
## AIC: 93.032
##
## Number of Fisher Scoring iterations: 2
```

Removing Background as confounder as the model with only first language as confounder gives the lowest AIC score

Final Model

```
#Model with first language and background
mod <- glm(binary ~ task + first, data = data)
summary(mod)
```

```
##
## Call:
## glm(formula = binary ~ task + first, data = data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.93373  -0.19996   0.06627   0.12049   0.96404
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            0.43006    0.12617   3.409  0.00105 **
## taskData wrangling     0.16400    0.13703   1.197  0.23512
## taskMachine Learning   0.44945    0.10556   4.258 5.84e-05 ***
## firstJava             -0.27979    0.16836  -1.662  0.10065
## firstMatlab           -0.09811    0.17530  -0.560  0.57733
## firstOther             0.05421    0.13368   0.406  0.68623
## firstPython            0.11091    0.12582   0.882  0.38081
## firstR                -0.39410    0.16286  -2.420  0.01791 *
## firstSAS              -0.38555    0.20386  -1.891  0.06240 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1546241)
##
##     Null deviance: 19.412  on 84  degrees of freedom
## Residual deviance: 11.751  on 76  degrees of freedom
## AIC: 93.032
```

```
##
## Number of Fisher Scoring iterations: 2
```

```
model <- glm(binary ~ task + first, data = data_relevel)
summary(model)
```

```
##
## Call:
## glm(formula = binary ~ task + first, data = data_relevel)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.93373  -0.19996   0.06627   0.12049   0.96404
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               0.87951    0.08765  10.034 1.43e-15 ***
## taskData visualization -0.44945    0.10556  -4.258 5.84e-05 ***
## taskData wrangling     -0.28545    0.12592  -2.267   0.0262 *
## firstJava              -0.27979    0.16836  -1.662   0.1007
## firstMatlab            -0.09811    0.17530  -0.560   0.5773
## firstOther              0.05421    0.13368   0.406   0.6862
## firstPython             0.11091    0.12582   0.882   0.3808
## firstR                 -0.39410    0.16286  -2.420   0.0179 *
## firstSAS               -0.38555    0.20386  -1.891   0.0624 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1546241)
##
##     Null deviance: 19.412  on 84  degrees of freedom
## Residual deviance: 11.751  on 76  degrees of freedom
## AIC: 93.032
##
## Number of Fisher Scoring iterations: 2
```

```
#Adjusting for p-values.(Not required anymore after chat with Tiffany)
#p.vals <- summary(model)$coef[,4]
#p.adjust(p.vals ,method = "BH") < 0.05
```

```
#https://stackoverflow.com/questions/11767602/backward-elimination-in-r?utm_medium=organic&utm_source=g
mod <- glm(binary ~ task + background + experience + attitude + first + active, data = data)
be_mod <- step(mod, direction = "both", trace=FALSE)
summary(be_mod)
```

```
##
## Call:
## glm(formula = binary ~ task + first, data = data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.93373  -0.19996   0.06627   0.12049   0.96404
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            0.43006    0.12617   3.409  0.00105 **
```

```
## taskData wrangling    0.16400    0.13703   1.197  0.23512
## taskMachine Learning  0.44945    0.10556   4.258 5.84e-05 ***
## firstJava            -0.27979    0.16836  -1.662  0.10065
## firstMatlab          -0.09811    0.17530  -0.560  0.57733
## firstOther            0.05421    0.13368   0.406  0.68623
## firstPython           0.11091    0.12582   0.882  0.38081
## firstR               -0.39410    0.16286  -2.420  0.01791 *
## firstSAS             -0.38555    0.20386  -1.891  0.06240 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1546241)
##
##     Null deviance: 19.412  on 84  degrees of freedom
## Residual deviance: 11.751  on 76  degrees of freedom
## AIC: 93.032
##
## Number of Fisher Scoring iterations: 2
```

```
#Not Required to Relevel the confounders.
#Releveling the reference task from Data Viz -> Machine Learning
#Releveling the reference first language from C -> R
data_relevel <- data
data_relevel$task <-relevel(data$task,ref="Machine Learning")
data_relevel$first <-relevel(data$first,ref="R")

model <- glm(binary ~ task + first, data = data_relevel)
summary(model)
```