Final Project - USPTO Examination Analysis

2024-02-09

Introduction

This analysis aims to explore organizational and social factors affecting the length of patent application prosecution and examiner attrition at the U.S. Patent and Trademark Office (USPTO), with a particular focus on the role of gender, race, and ethnicity.

Data Loading

First, we need to load the dataset.

```
# Read the dataset
applications <-
read.csv("/Users/sheidamajidi/Desktop/Winter2024/COURSES/ORGB671/Final
Project/app_data_starter.csv")
applications
```

```
##      application_number filing_date examiner_name_last examiner_name_first
## 1               8284457  2000-01-26             HOWARD          JACQUELINE
## 2               8413193  2000-10-11           YILDIRIM               BEKIR
## 3               8531853  2000-05-17           HAMILTON             CYNTHIA
## 4               8637752  2001-07-20             MOSHER                MARY
## 5               8682726  2000-04-10               BARR             MICHAEL
## 6               8687412  2000-04-28               GRAY               LINDA
## 7               8716371  2004-01-26          MCMILLIAN                KARA
## 8               8765941  2000-06-23               FORD             VANESSA
## 9               8776818  2000-02-04         STRZELECKA              TERESA
## 10              8809677  2002-02-20                KIM                 SUN
## 11              8836939  2000-06-13               WOOD           ELIZABETH
## 12              8901519  2000-09-26               DENT               ALANA
## 13              8913518  2004-04-06           AFTERGUT              JEFFRY
## 14              8930379  2002-04-08              KUMAR          SHAILENDRA
## 15              8945309  2000-06-15           STARSIAK                JOHN
## 16              8952426  2000-08-21               TRAN               SUSAN
## 17              8973360  2000-02-09                 LI                QIAN
## 18              8974843  2000-01-11              PEESO              THOMAS
## 19              8981219  2000-07-27              DAVIS              ROBERT
## 20              8994479  2001-01-30              BOYER             CHARLES
## 21              9000004  2001-05-02           SAUNDERS               DAVID
## 22              9011027  2000-05-01           LANDSMAN              ROBERT
## 23              9011075  2000-05-03             FORMAN               BETTY
## 24              9029401  2000-02-08            ANTHONY              JOSEPH
## 25              9036107  2000-08-16                COE              PHILIP
## 26              9043351  2000-03-13           NAKARANI           DHIRAJLAL
## 27              9043825  2000-02-22           ROBINSON               ALLEN
## 28              9043931  2000-04-21             MACKEY               JAMES
## 29              9043944  2000-10-06                LIU              SAMUEL
## 30              9051571  2001-12-14             SEAMAN          D MARGARET
## 31              9063553  2000-06-14            SZEKELY               PETER
## 32              9068704  2001-02-23           ROBINSON               BINTA
## 33              9077252  2000-05-20        PADMANABHAN              KARTIC
## 34              9077619  2000-03-31         STRZELECKA              TERESA
## 35              9077671  2000-06-19   NAZARIO GONZALEZ            PORFIRIO
## 36              9077740  2000-01-12              NOLAN             PATRICK
## 37              9091473  2000-01-20              JUSKA              CHERYL
## 38              9091481  2000-06-27           ROBINSON               BINTA
## 39              9091683  2000-11-24               YOON                 TAE
## 40              9091815  2000-04-13             NICKOL                GARY
## 41              9101427  2000-09-16               BARR             MICHAEL
## 42              9101566  2000-10-18            TENTONI                 LEO
## 43              9102914  2000-02-16             BERMAN               SUSAN
## 44              9106157  2000-02-07             KRAMER                DEAN
## 45              9117087  2000-07-28             SALIMI                 ALI
## 46              9117089  2000-09-05              KUMAR          SHAILENDRA
## 47              9117222  2000-12-08             NASHED             NASHAAT
## 48              9117365  2000-01-05     MCGUTHRY BANKS                TIMA
## 49              9117588  2000-03-08            SAUCIER              SANDRA
## 50              9119563  2000-02-03              EGWIM             KELECHI
## 51              9125199  2000-03-23             ROTMAN                ALAN
## 52              9125738  2000-01-04             SHOSHO              CALLIE
## 53              9129758  2010-09-15                PAK             MICHAEL
## 54              9142080  2000-05-11            FREDMAN             JEFFREY
## 55              9142120  2000-03-20              WELLS              LAUREN
## 56              9142313  2000-03-15               None                None
## 57              9142314  2000-06-01             SISSON             BRADLEY
## 58              9144927  2007-03-12              WALKE              AMANDA
## 59              9147226  2000-05-31                YEE             DEBORAH
## 60              9147568  2000-10-30              HUANG              EVELYN
## 61              9147572  2000-07-31        PADMANABHAN              KARTIC
## 62              9147994  2000-01-12                LUK            EMMANUEL
## 63              9155146  2000-02-03           WOODWARD                 ANA
## 64              9155842  2001-04-06           FERGUSON            LAWRENCE
## 65              9171311  2000-09-12               NAFF               DAVID
## 66              9171344  2000-05-19                FAY              ZOHREH
## 67              9171573  2000-01-04                ULM                JOHN
## 68              9171671  2000-05-01         WESSENDORF              TERESA
## 69              9171909  2001-05-09          SCHWADRON              RONALD
## 70              9179478  2000-11-17             KNABLE            GEOFFREY
## 71              9180120  2000-04-05               WONG              LESLIE
## 72              9180805  2002-06-07              SMITH               DUANE
## 73              9194043  2001-06-01     MCGUTHRY BANKS                TIMA
## 74              9194075  2000-10-31          HEITBRINK             TIMOTHY
## 75              9194374  2000-07-25             GORDON               BRIAN
## 76              9194619  2003-08-25             KOLKER              DANIEL
## 77              9194664  2002-02-14            SHEEHAN                JOHN
## 4728            9485269  2000-05-12          METZMAIER              DANIEL
## 4729            9485270  2000-02-07              GROUP                KARL
## 4730            9485273  2000-05-31          GALLAGHER                JOHN
## 4731            9485274  2001-01-10            OLTMANS              ANDREW
## 4732            9485275  2000-04-05            CAMERON                ERMA
## 4733            9485277  2000-05-17              JUSKA              CHERYL
## 4734            9485279  2000-04-03              BROWN         CHRISTOPHER
## 4735            9485280  2000-02-04              CHANG               CELIA
## 4736            9485281  2000-06-15              UPTON         CHRISTOPHER
## 4737            9485283  2000-04-24           TARAZANO              DONALD
## 4738            9485284  2000-02-07             TOOMER              CEPHIA
## 4739            9485286  2000-02-07            DEBERRY              REGINA
## 4740            9485288  2000-02-07            SZEKELY               PETER
## 4741            9485292  2000-05-03              CHANG               CELIA
## 4742            9485293  2000-02-08               YUAN             DAH WEI
## 4743            9485297  2000-02-08                KIM            JENNIFER
## 4744            9485298  2000-02-08                KIM               YOUNG
## 4745            9485300  2000-04-27              GHALI                ISIS
## 4746            9485303  2000-02-08             SAYALA              CHHAYA
## 4747            9485307  2000-05-15           BUCHANAN         CHRISTOPHER
## 4748            9485309  2000-05-18              CHANG               CELIA
## 4749            9485313  2000-05-09               BASI              NIRMAL
## 4750            9485314  2000-04-24              SINES               BRIAN
## 4751            9485316  2000-02-04             TAYLOR              JANELL
## 4752            9485320  2000-02-08               MARX               IRENE
## 4753            9485321  2000-07-20            PESELEV                ELLI
## 4754            9485322  2000-06-21    BALASUBRAMANIAN        VENKATARAMAN
## 4755            9485323  2000-02-07                WAX              ROBERT
## 4756            9485326  2000-02-07             PARTON               KEVIN
## 4757            9485329  2002-07-29          BLACKWELL           GWENDOLYN
## 4758            9485335  2000-05-10               DOVE               TRACY
## 4759            9485336  2000-03-20             SIEGEL                ALAN
## 4760            9485337  2000-05-30            NICOLAS              WESLEY
## 4761            9485339  2000-05-15        WILKINS III               HARRY
##       examiner_name_middle examiner_id examiner_art_unit uspc_class
## 1                       V       96082              1764        508
## 2                       L       87678              1764        208
## 3                               63213              1752        430
## 4                               73788              1648        530
## 5                       E       77294              1762        427
## 6                   LAMEY       68606              1734        156
## 7                  RENITA       89557              1627        424
## 8                       L       97543              1645        424
## 9                       E       98714              1637        435
## 10                      U       65530              1723        210
## 11                      D       77112              1755        106
## 12                 HARRIS       92931              1642        435
## 13                      H       75406              1733        156
## 14                              95054              1672        514
## 15                      S       99360              1753        204
## 16                      T       73198              1615        424
## 17                 JANICE       76132              1632        514
## 18                      R       77284              2132        713
## 19                      B       63176              1722        425
## 20                      I       59816              1751        510
## 21                      A       64507              1644        435
## 22                      S       98520              1647        424
## 23                      J       59480              1634        435
## 24                  DAVID       82563              1714        252
## 25                      R       75700              1746        134
## 26                      S       68153              1773        428
## 27                    JAY       64054              1616        514
## 28                      P       77298              1722        425
## 29                      W       66510              1653        800
## 30                      M       84317              1625        546
## 31                      A       76370              1714        523
## 32                      M       97637              1625        514
## 33                             64900              1641        435
## 34                      E       98714              1637        435
## 35                             90096              1671        556
## 36                      J       97461              1644        435
## 37                    ANN       99249              1771        428
## 38                      M       97637              1625        549
## 39                      H       95839              1714        428
## 40                      B       65024              1642        435
## 41                      E       77294              1762        427
## 42                      B       96556              1732        264
## 43                      W       88204              1711        429
## 44                      J       67890              2167        294
## 45                   REZA       69410              1648        435
## 46                             95054              1672        548
## 47                      T       60331              1652        435
## 48                MICHELE       68970              1742        075
## 49                      E       59730              1651        424
## 50                  CHIDI       91482              1713        525
## 51                      L       60246              1625        546
## 52                      E       63829              1714        523
## 53                      D       95205              1646        435
## 54                 NORMAN       99343              1637        435
## 55                      Q       75736              1617        424
## 56                   None          NA              1615        424
## 57                      L       62984              1634        435
## 58                      C       75563              1722        430
## 59                             71101              1742        420
## 60                    MEI       70017              1625        514
## 61                             64900              1641        436
## 62                      S       97657              1722        164
## 63               LUCRECIA       92836              1711        525
## 64                      D       99518              1774        428
## 65                      M       79886              1651        435
## 66                      A       69138              1614        514
## 67                      D       73081              1646        514
## 68                      D       73880              1639        435
## 69                      B       71437              1644        514
## 70                      L       79330              1733        156
## 71                      A       60400              1761        426
## 72                              66387              1724        095
## 73                MICHELE       68970              1742        075
## 74                      W       64074              1722        264
## 75                      R       91423              1743        422
## 76                      E       93839              1649        435
## 77                      P       97553              1742        148
## 78                      E       63829              1714        524
## 79                  COMBS       62010              1742        148
## 80                      L       61876              1624        514
## 81                      C       67636              1648        435
## 82                  DAVID       82563              1714        252
## 83                      D       67998              1621        554
## 84                              70824              1631        800
## 85                      A       71777              1712        525
## 86                      D       67998              1671        554
## 87                     XU       67140              1639        435
## 88                      M       64839              1616        424
## 89                      B       59768              2155        709
## 90                      F       98254              2121        700
## 91                              64187              2151        709
## 92                      J       65111              1657        435
## 93              DOROSHENK       90241              1764        422
## 94                      E       84127              1744        015
## 95                      M       71268              2154        709
## 96                  DAVID       82563              1714        252
## 97                      A       88202              1762        427
## 98                              60130              2171        707
## 99               SHARIDAN       85332              1746        134
## 100               WEI MIN       70264              1655        435
## 101                     A       89539              1734        156
## 102                     C       98937              1724        210
## 103                     K       94939              1625        568
## 104                     T       72253              1632        514
## 105                             91867              1773        428
## 106                     T       61541              1774        428
## 107                             91867              1773        156
## 108                     F       65737              1621        568
## 109                     H       95415              1614        514
## 110                     A       71595              1741        204
## 111                             96458              1653        514
## 112                             59645              1625        514
## 113                     L       66941              1652        435
## 114                     N       94911              1652        435
## 115                     U       65530              1723        422
## 4716                    M       81537              1751        510
## 4717                    I       59816              1751        510
## 4718                            60520              1617        514
## 4719                            65829              1614        514
## 4720                    H       95415              1614        514
## 4721                    J       98205              1614        514
## 4722                    F       61790              1671        568
## 4723                    S       63649              1745        429
## 4724                            77979              1617        514
## 4725                            71101              1742        420
## 4726                            87124              1711        528
## 4727                            92485              1617        514
## 4728                    S       98582              1712        516
## 4729                    E       72666              1755        501
## 4730                    J       68386              1733        156
## 4731                    L       86424              1742        148
## 4732                    C       94341              1762        427
## 4733                  ANN       99249              1771        428
## 4734                    J       61925              2134        713
## 4735                    C       65536              1625        536
## 4736                            74727              1724        210
## 4737             LAWRENCE       71035              1773        428
## 4738                    D       75336              1714        524
## 4739                    M       72069              1647        435
## 4740                    A       76370              1714        524
## 4741                    C       65536              1625        514
## 4742                    D       69378              1745        429
## 4743                    M       61529              1617        514
## 4744                    J       78019              1637        435
## 4745                  A D       92187              1615        424
## 4746                    D       69350              1761        426
## 4747                    R       64013              2167        180
## 4748                    C       65536              1625        514
## 4749                SINGH       75730              1646        435
## 4750                    J       72122              1743        422
## 4751                    E       90546              1656        514
## 4752                            65264              1651        424
## 4753                            62253              1623        514
## 4754                            84504              1624        544
## 4755                    A       67669              1653        514
## 4756                    S       69109              2153        709
## 4757                            98245              1775        428
## 4758                  MAE       63557              1745        429
## 4759                    M       66593              1621        570
## 4760                    A       71595              1741        205
## 4761                    D       98852              1742        075
##      uspc_subclass patent_number patent_issue_date abandon_date
disposal_type
## 1           273000       6521570        2003-02-18
ISS
```

```
## 2           179000       6440298        2002-08-27
ISS
## 3           271100       5607816        1997-03-04
ISS
## 4           388300       6927281        2005-08-09
ISS
## 5           430100                                   2000-12-27
ABN
## 6           204000       6267836        2001-07-31
ISS
## 7           401000
PEND
## 8           001210                                   2001-08-22
ABN
## 9           006000                                   2002-07-15
ABN
## 10          645000       6858146        2005-02-22
ISS
## 11          479000       6358307        2002-03-19
ISS
## 12          007230       6261766        2001-07-17
ISS
## 13          148000       7005024        2006-02-28
ISS
## 14          617000       6670400        2003-12-30
ISS
## 15          604000                                   2003-11-26
ABN
## 16          451000       6428808        2002-08-06
ISS
## 17          044000                                   2002-11-14
ABN
## 18          168000                                   2003-01-13
ABN
## 19          133100                                   2003-06-10
ABN
## 20          511000       7256169        2007-08-14
ISS
## 21          007210       6780603        2004-08-24
ISS
## 22          085200       6387364        2002-05-14
ISS
## 23          006000       6562566        2003-05-13
ISS
## 24          186100                                   2005-06-07
ABN
## 25          002000       6240934        2001-06-05
ISS
## 26          336000       6248440        2001-06-19
ISS
```

```
## 27          386000       6417216        2002-07-09
ISS
## 28          595000                                   2002-05-08
ABN
## 29          013000       6787641        2004-09-07
ISS
## 30          153000       6479660        2002-11-12
ISS
## 31          335000       6300392        2001-10-09
ISS
## 32          354000                                   2001-08-09
ABN
## 33          007920       6514716        2003-02-04
ISS
## 34          006000       6500614        2002-12-31
ISS
## 35          028000       6294682        2001-09-25
ISS
## 36          069100                                   2001-11-05
ABN
## 37          090000       6479416        2002-11-12
ISS
## 38          305000       6482959        2002-11-19
ISS
## 39          35500R       6632522        2003-10-14
ISS
## 40          069100                                   2001-10-17
ABN
## 41          008000       6455097        2002-09-24
ISS
## 42          103000       6413631        2002-07-02
ISS
## 43          185000       6146789        2000-11-14
ISS
## 44          102200       6227587        2001-05-08
ISS
## 45          005000       6365343        2002-04-02
ISS
## 46          566000       6429317        2002-08-06
ISS
## 47          183000                                   2002-10-17
ABN
## 48          010670       6355085        2002-03-12
ISS
## 49          531000       6312733        2001-11-06
ISS
## 50          061000       6096826        2000-08-01
ISS
## 51          126000       6268498        2001-07-31
ISS
```

```
## 4752        093440       6716424        2004-04-06
ISS
## 4753        100000       6340746        2002-01-22
ISS
## 4754        182000                                   2000-12-27
ABN
## 4755        012000       6818617        2004-11-16
ISS
## 4756        223000       6766366        2004-07-20
ISS
## 4757        432000       6942925        2005-09-13
ISS
## 4758        217000       6555268        2003-04-29
ISS
## 4759        171000       6441256        2002-08-27
ISS
## 4760        618000                                   2002-05-29
ABN
## 4761        241000       6375707        2002-04-23
ISS
##      appl_status_code   appl_status_date   tc gender     race
earliest_date
## 1                 150 30jan2003 00:00:00 1700 female    white    2000-01-
10
## 2                 250 27sep2010 00:00:00 1700           white    2000-01-
04
## 3                 250 30mar2009 00:00:00 1700 female    white    2000-01-
06
## 4                 250 07sep2009 00:00:00 1600 female    white    2000-01-
04
## 5                 161 19apr2001 00:00:00 1700   male    white    2000-01-
03
## 6                 150 16jul2001 00:00:00 1700 female    white    2000-01-
04
## 7                 135 15may2017 00:00:00 1600 female    black    2001-12-
19
## 8                 161 03apr2002 00:00:00 1600 female    white    2000-02-
08
## 9                 161 27nov2002 00:00:00 1600 female    white    2000-01-
21
## 10                250 23mar2009 00:00:00 1700 female    Asian    2000-01-
03
## 11                250 19apr2006 00:00:00 1700 female    white    2000-01-
05
## 12                250 17aug2009 00:00:00 1600 female    white    2000-01-
03
## 13                250 28mar2014 00:00:00 1700   male    white    2000-01-
10
## 14                250 30jan2012 00:00:00 1600           Asian    2000-01-
07
```

```
## 15                168 08dec2003 00:00:00 1700   male    white    2000-01-
04
## 16                250 06sep2010 00:00:00 1600 female    Asian    2000-01-
14
## 17                161 19may2003 00:00:00 1600   male    Asian    2000-01-
12
## 18                168 23jan2003 00:00:00 2100   male    white    2000-01-
03
## 19                161 25sep2003 00:00:00 1700   male    white    2000-01-
18
## 20                250 11sep2015 00:00:00 1700   male    white    2000-01-
06
## 21                250 22sep2008 00:00:00 1600   male    white    2000-01-
14
## 22                150 26apr2002 00:00:00 1600   male    white    2000-01-
11
## 23                250 18jun2007 00:00:00 1600 female    white    2000-01-
25
## 24                161 06dec2005 00:00:00 1700   male    white    2000-01-
05
## 25                250 06jul2005 00:00:00 1700   male    white    2000-01-
06
## 26                250 20jul2009 00:00:00 1700           white    2000-01-
05
## 27                250 09aug2006 00:00:00 1600   male    white    2000-01-
05
## 28                164 08may2002 00:00:00 1700   male    white    2000-01-
13
## 29                250 08oct2012 00:00:00 1600   male    Asian    2000-01-
20
## 30                250 05dec2014 00:00:00 1600           white    2000-01-
13
## 31                150 21sep2001 00:00:00 1700   male    white    2000-01-
03
## 32                161 16nov2001 00:00:00 1600 female    white    2000-01-
03
## 33                150 16jan2003 00:00:00 1600           Asian    2000-01-
19
## 34                250 31jan2011 00:00:00 1600 female    white    2000-01-
21
## 35                250 18oct2013 00:00:00 1600   male Hispanic    2000-01-
05
## 36                161 25feb2002 00:00:00 1600   male    white    2000-01-
11
## 37                150 24oct2002 00:00:00 1700 female    white    2000-01-
06
## 38                150 31oct2002 00:00:00 1600 female    white    2000-01-
03
## 39                250 06nov2015 00:00:00 1700   male    Asian    2000-01-
11
```

```
## 4740              150 03jun2004 00:00:00 1700   male    white    2000-01-
03
## 4741              250 03sep2012 00:00:00 1600 female    Asian    2000-01-
07
## 4742              250 27sep2010 00:00:00 1700           Asian    2000-01-
05
## 4743              150 29mar2001 00:00:00 1600 female    Asian    2000-01-
05
## 4744              161 20feb2007 00:00:00 1600   male    Asian    2000-01-
18
## 4745              250 27jun2014 00:00:00 1600 female    white    2000-01-
03
## 4746              250 12jan2005 00:00:00 1700           white    2000-01-
06
## 4747              150 12sep2002 00:00:00 2100   male    white    2000-03-
16
## 4748              250 11jan2010 00:00:00 1600 female    Asian    2000-01-
07
## 4749              161 06feb2003 00:00:00 1600   male    Asian    2000-01-
10
## 4750              250 13mar2015 00:00:00 1700   male    white    2000-01-
05
## 4751              250 28feb2014 00:00:00 1600 female    white    2000-01-
13
## 4752              250 07may2012 00:00:00 1600 female    white    2000-01-
13
## 4753              250 14feb2014 00:00:00 1600 female    white    2000-01-
04
## 4754              161 09may2001 00:00:00 1600           Asian    2000-01-
05
## 4755              250 17dec2012 00:00:00 1600   male    white    2000-02-
04
## 4756              150 01jul2004 00:00:00 2100   male    white    2000-01-
18
## 4757              150 24aug2005 00:00:00 1700 female    white    2000-01-
07
## 4758              250 22may2015 00:00:00 1700 female    white    2000-01-
03
## 4759              250 27sep2010 00:00:00 1600   male    white    2000-01-
04
## 4760              161 04sep2002 00:00:00 1700   male Hispanic    2000-01-
14
## 4761              250 24may2006 00:00:00 1700   male    white    2000-01-
03
##      latest_date tenure_days
## 1     2016-04-01        5926
## 2     2016-09-09        6093
## 3     2017-05-20        6344
## 4     2017-05-05        6331
## 5     2017-05-05        6332
## 6     2017-05-19        6345
## 7     2017-05-23        5634
## 8     2019-11-16        7221
## 9     2017-05-22        6331
## 10    2017-05-18        6345
## 11    2017-05-22        6347
## 12    2017-05-23        6350
## 13    2017-05-23        6343
## 14    2017-05-12        6335
## 15    2017-01-19        6225
## 16    2017-05-19        6335
## 17    2017-05-22        6340
## 18    2017-04-28        6325
## 19    2017-05-23        6335
## 20    2017-05-23        6347
## 21    2017-05-12        6328
## 22    2017-05-22        6341
## 23    2915-06-29      334352
## 24    2017-05-22        6347
## 25    2015-10-23        5769
## 26    2017-05-12        6337
## 27    2015-01-30        5504
## 28    2017-05-20        6337
## 29    2017-05-19        6329
## 30    2017-05-23        6340
## 31    2017-05-12        6339
## 32    2017-03-10        6276
## 33    2016-09-30        6099
## 34    2017-05-22        6331
## 35    2017-05-18        6343
## 36    2017-03-24        6282
## 37    2017-05-19        6343
## 38    2017-03-10        6276
## 39    2017-05-23        6342
## 40    2017-05-10        6322
## 41    2017-05-05        6332
## 42    2017-05-23        6349
## 43    2017-05-12        6332
## 44    2013-12-13        5082
## 45    2017-05-05        6328
## 46    2017-05-12        6335
## 47    2017-05-05        6330
## 48    2017-05-22        6349
## 49    2017-04-28        6309
## 50    2017-05-23        6349
## 51    2015-07-02        5657
## 52    2017-04-28        6324
## 53    2017-05-19        6315
## 54    2017-05-12        6324
## 55    2017-05-05        6323
## 56    2915-06-29      334373
## 57    2017-05-22        6347
## 58    2017-05-23        6347
## 59    2017-05-23        6348
## 60    2017-03-24        6283
## 61    2016-09-30        6099
## 62    2017-05-18        6336
## 63    2017-05-19        6330
## 64    2017-05-23        6342
## 65    2017-05-05        6328
## 66    2017-05-23        6348
## 67    2017-05-22        6348
## 68    2050-12-30       18610
## 69    2017-05-18        6327
## 70    2017-05-19        6342
## 71    2017-05-22        6338
## 72    2017-05-19        6338
## 73    2017-05-22        6349
## 74    2017-05-12        6337
## 75    2017-05-23        6342
## 76    2017-05-19        6245
## 77    2017-03-31        6296
## 78    2017-04-28        6324
## 79    2017-05-22        6349
## 80    2017-04-28        6324
## 81    2016-09-09        6084
## 82    2017-05-22        6347
## 83    2017-05-22        6341
## 84    2017-03-09        6265
## 85    2017-01-27        6224
## 86    2017-05-22        6341
## 87    2017-05-16        6049
## 88    2017-02-17        6254
## 89    2017-05-18        6323
## 90    2016-01-22        5860
## 91    2017-01-13        6209
## 92    2017-05-23        6350
## 93    2017-05-15        6321
## 94    2017-04-07        6288
## 95    2017-01-19        6226
## 96    2017-05-22        6347
## 97    2017-05-22        6342
## 98    2017-05-19        6339
## 99    2017-05-19        6338
## 100   2017-05-22        6347
## 101   2017-04-28        6318
## 102   2017-05-05        6321
## 103   2017-05-12        6336
## 104   2017-05-22        6349
## 105   2017-05-23        6347
## 4756  2016-11-18        6149
## 4757  2017-05-12        6335
## 4758  2017-05-19        6346
## 4759  2015-03-27        5561
## 4760  2017-03-31        6286
## 4761  2017-05-23        6350
##  [ reached 'max' / getOption("max.print") -- omitted 2013716 rows ]
```

Data Preparation

We need to prepare the data by ensuring correct data types and creating necessary features.

```
applications <- applications %>%
  mutate(filing_date = ymd(filing_date),
         patent_issue_date = ymd(patent_issue_date),
         abandon_date = ymd(abandon_date),
         prosecution_length = if_else(!is.na(patent_issue_date),
as.numeric(difftime(patent_issue_date, filing_date, units = "days")),
                                       as.numeric(difftime(abandon_date,
filing_date, units = "days"))))
```

Data Cleaning and Exploration We’ll examine the dataset for missing values and summarize its key statistics.

```
# Check for missing values
summary(applications)

##  application_number  filing_date         examiner_name_last
examiner_name_first
##  Min.   : 8284457   Min.   :2000-01-02   Length:2018477     Length:2018477
##  1st Qu.:10975476   1st Qu.:2005-03-30   Class :character   Class
:character
##  Median :12491809   Median :2009-07-23   Mode  :character   Mode
:character
##  Mean   :12477062   Mean   :2009-03-23
##  3rd Qu.:13892722   3rd Qu.:2013-05-22
##  Max.   :95002230   Max.   :2017-05-26
##
##  examiner_name_middle  examiner_id    examiner_art_unit  uspc_class
##  Length:2018477       Min.   :59012   Min.   :1600      Length:2018477
##  Class :character     1st Qu.:66476   1st Qu.:1671      Class :character
##  Mode  :character     Median :75243   Median :1773      Mode  :character
##                       Mean   :78712   Mean   :1928
##                       3rd Qu.:93754   3rd Qu.:2171
##                       Max.   :99990   Max.   :2498
##                       NA's   :9229
##  uspc_subclass      patent_number      patent_issue_date
##  Length:2018477     Length:2018477     Min.   :1997-03-04
##  Class :character   Class :character   1st Qu.:2008-04-29
##  Mode  :character   Mode  :character   Median :2012-05-22
##                                         Mean   :2011-06-20
##                                         3rd Qu.:2015-01-20
##                                         Max.   :2017-06-20
##                                         NA's   :931178
##   abandon_date         disposal_type      appl_status_code appl_status_date
##  Min.   :1965-07-20   Length:2018477     Min.   :  1.0    Length:2018477
##  1st Qu.:2008-06-23   Class :character   1st Qu.:150.0    Class :character
##  Median :2011-04-19   Mode  :character   Median :150.0    Mode  :character
##  Mean   :2011-01-28                      Mean   :145.9
##  3rd Qu.:2014-04-15                      3rd Qu.:161.0
##  Max.   :2050-06-30                      Max.   :865.0
##  NA's   :1417057                         NA's   :4609
##        tc          gender              race           earliest_date
##  Min.   :1600   Length:2018477     Length:2018477     Length:2018477
##  1st Qu.:1600   Class :character   Class :character   Class :character
##  Median :1700   Mode  :character   Mode  :character   Mode  :character
##  Mean   :1877
##  3rd Qu.:2100
##  Max.   :2400
##
##  latest_date         tenure_days      prosecution_length
##  Length:2018477     Min.   :     27   Min.   :-13636
##  Class :character   1st Qu.:   4963   1st Qu.:   765
##  Mode  :character   Median :   6094   Median :  1079
##                     Mean   :  10282   Mean   :  1190
##                     3rd Qu.:   6336   3rd Qu.:  1481
##                     Max.   :2727903   Max.   : 17898
##                                       NA's   :329761

# Quick summary of data columns
skim(applications)
```

Data summary

Name	applications
Number of rows	2018477
Number of columns	22
Column type frequency:	
character	12
Date	3
numeric	7
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
examiner_name_last	0	1	2	17	0	3806	0
examiner_name_first	0	1	1	12	0	2595	0
examiner_name_middle	0	1	0	12	471770	516	0
uspc_class	0	1	0	3	4	417	0
uspc_subclass	0	1	0	6	1677	6155	0
patent_number	0	1	0	7	931651	1086825	0
disposal_type	0	1	3	4	0	3	0
appl_status_date	0	1	0	18	4610	5706	0
gender	0	1	0	6	303859	3	0
race	0	1	5	8	0	5	0
earliest_date	0	1	10	10	0	2325	0
latest_date	0	1	10	10	0	888	0

Variable type: Date

skim_variable	n_missing	complete_rate	min	max	median	n_unique
filing_date	0	1.00	2000-01-02	2017-05-26	2009-07-23	6204
patent_issue_date	931178	0.54	1997-03-04	2017-06-20	2012-05-22	891
abandon_date	1417057	0.30	1965-07-20	2050-06-30	2011-04-19	5052

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
application_number	0	1.00	12477061.84	2198067.04	8284457	10975476	12491809	13892722	95002230	▇▁▁▁

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
examiner_id	9229	1.00	78712.39	13606.61	59012	66476	75243	93754	99990	▁▇▇▂▂
examiner_art_unit	0	1.00	1928.02	304.38	1600	1671	1773	2171	2498	▇▇▂▁▂
appl_status_code	4609	1.00	145.94	51.72	1	150	150	161	865	▂▇▁▁▁
tc	0	1.00	1876.91	298.82	1600	1600	1700	2100	2400	▁▇▁▁▂
tenure_days	0	1.00	10282.35	87390.08	27	4963	6094	6336	2727903	▂▇▁▁▁
prosecution_length	329761	0.84	1190.22	620.88	-13636	765	1079	1481	17898	▁▁▇▁▁

Handling those missing values:

Imputation for prosecution_length:

For prosecution_length, since it’s numerical, using the median might be more robust to outliers than the mean.

```
median_prosecution_length <- median(applications$prosecution_length, na.rm =
TRUE)
```

```
applications$prosecution_length[is.na(applications$prosecution_length)] <-
median_prosecution_length
```

Examiner Gender and Race Estimation

```
library(gender)
library(dplyr)

# Step 1: Generate gender predictions
examiner_names <- applications %>%
  distinct(examiner_name_first)

# Use gender() on the unique list of first names
gender_predictions <- gender(examiner_names$examiner_name_first, method =
"ssa", years = c(1940, 2020))

## Warning in gender(examiner_names$examiner_name_first, method = "ssa",
years =
## c(1940, : The year range provided has been trimmed to fit within 1880 to
2012.

# Convert gender_predictions to a dataframe and prepare for join
gender_df <- as.data.frame(gender_predictions) %>%
  rename(examiner_name_first = name) %>%
  select(examiner_name_first, gender)

# Step 2: Join gender predictions back to applications
applications <- applications %>%
  left_join(gender_df, by = "examiner_name_first")

# Print the result to check the first few rows, including the newly added
gender column
head(applications)

##   application_number filing_date examiner_name_last examiner_name_first
## 1            8284457  2000-01-26             HOWARD          JACQUELINE
## 2            8413193  2000-10-11           YILDIRIM               BEKIR
## 3            8531853  2000-05-17           HAMILTON             CYNTHIA
## 4            8637752  2001-07-20             MOSHER                MARY
## 5            8682726  2000-04-10               BARR             MICHAEL
## 6            8687412  2000-04-28               GRAY               LINDA
##   examiner_name_middle examiner_id examiner_art_unit uspc_class
uspc_subclass
## 1                    V       96082              1764        508
273000
## 2                    L       87678              1764        208
179000
## 3                            63213              1752        430
271100
## 4                            73788              1648        530
388300
```

```
## 5                    E       77294              1762        427
430100
## 6                LAMEY       68606              1734        156
204000
##   patent_number patent_issue_date abandon_date disposal_type
appl_status_code
## 1       6521570        2003-02-18         <NA>           ISS
150
## 2       6440298        2002-08-27         <NA>           ISS
250
## 3       5607816        1997-03-04         <NA>           ISS
250
## 4       6927281        2005-08-09         <NA>           ISS
250
## 5                            <NA>   2000-12-27           ABN
161
## 6       6267836        2001-07-31         <NA>           ISS
150
##     appl_status_date   tc gender.x  race earliest_date latest_date
tenure_days
## 1 30jan2003 00:00:00 1700   female white    2000-01-10  2016-04-01
5926
## 2 27sep2010 00:00:00 1700          white    2000-01-04  2016-09-09
6093
## 3 30mar2009 00:00:00 1700   female white    2000-01-06  2017-05-20
6344
## 4 07sep2009 00:00:00 1600   female white    2000-01-04  2017-05-05
6331
## 5 19apr2001 00:00:00 1700     male white    2000-01-03  2017-05-05
6332
## 6 16jul2001 00:00:00 1700   female white    2000-01-04  2017-05-19
6345
##   prosecution_length gender.y
## 1               1119   female
## 2                685     <NA>
## 3              -1170   female
## 4               1481   female
## 5                261     male
## 6                459   female
```

Examiner Race Estimation

First, we need to start by predicting race based on surname using the WRU package:

```
library(dplyr)
library(wru)

# Step 1: Get unique surnames
examiner_surnames <- applications %>%
  distinct(examiner_name_last) %>%
```

```r
  rename(surname = examiner_name_last)

# Preparing the voter_file with surnames
#voter_file <- applications %>%
#  distinct(examiner_name_last) %>%
#  mutate(surname = tolower(examiner_name_last)) %>%
#  select(surname)

# Call to predict_race() adjusted for simplicity
#race_predictions <- predict_race(voter_file, census = "2010", surname.only =
TRUE)

# Step 2: Use predict_race() to estimate race
race_predictions <- predict_race(examiner_surnames, surname.only = TRUE)
```

```
## Proceeding with last name predictions...

## ℹ All local files already up-to-date!

## 701 (18.4%) individuals' last names were not matched.
```

```r
print(colnames(race_predictions))
```

```
## [1] "surname"  "pred.whi" "pred.bla" "pred.his" "pred.asi" "pred.oth"
```

```r
# Process the race predictions to identify the most probable race
race_predictions <- race_predictions %>%
  rowwise() %>%
  mutate(most_probable_race = case_when(
    pred.whi == max(c(pred.whi, pred.bla, pred.his, pred.asi, pred.oth),
na.rm = TRUE) ~ "White",
    pred.bla == max(c(pred.whi, pred.bla, pred.his, pred.asi, pred.oth),
na.rm = TRUE) ~ "Black or African American",
    pred.his == max(c(pred.whi, pred.bla, pred.his, pred.asi, pred.oth),
na.rm = TRUE) ~ "Hispanic",
    pred.asi == max(c(pred.whi, pred.bla, pred.his, pred.asi, pred.oth),
na.rm = TRUE) ~ "Asian",
    pred.oth == max(c(pred.whi, pred.bla, pred.his, pred.asi, pred.oth),
na.rm = TRUE) ~ "Other",
    TRUE ~ "Unknown"
  )) %>%
  ungroup()

# Join the race predictions back to the applications dataframe
# 'examiner_name_last' in 'applications' matches 'surname' in
'race_predictions'
applications_with_race <- applications %>%
  left_join(race_predictions %>% select(surname, most_probable_race), by =
c("examiner_name_last" = "surname"))

# Note: The above operation selects only the relevant columns ('surname' and
'most_probable_race')


# View the first few rows of the updated dataframe to check the join results
head(applications_with_race)

##   application_number filing_date examiner_name_last examiner_name_first
## 1            8284457  2000-01-26             HOWARD          JACQUELINE
## 2            8413193  2000-10-11           YILDIRIM               BEKIR
## 3            8531853  2000-05-17           HAMILTON             CYNTHIA
## 4            8637752  2001-07-20             MOSHER                MARY
## 5            8682726  2000-04-10               BARR             MICHAEL
## 6            8687412  2000-04-28               GRAY               LINDA
##   examiner_name_middle examiner_id examiner_art_unit uspc_class
uspc_subclass
## 1                    V       96082              1764        508
273000
## 2                    L       87678              1764        208
179000
## 3                            63213              1752        430
271100
## 4                            73788              1648        530
388300
## 5                    E       77294              1762        427
430100
## 6                LAMEY       68606              1734        156
204000
##   patent_number patent_issue_date abandon_date disposal_type
appl_status_code
## 1       6521570        2003-02-18         <NA>           ISS
150
## 2       6440298        2002-08-27         <NA>           ISS
250
## 3       5607816        1997-03-04         <NA>           ISS
250
## 4       6927281        2005-08-09         <NA>           ISS
250
## 5                            <NA>   2000-12-27           ABN
161
## 6       6267836        2001-07-31         <NA>           ISS
150
##     appl_status_date   tc gender.x  race earliest_date latest_date
tenure_days
## 1 30jan2003 00:00:00 1700   female white    2000-01-10  2016-04-01
5926
## 2 27sep2010 00:00:00 1700          white    2000-01-04  2016-09-09
6093
## 3 30mar2009 00:00:00 1700   female white    2000-01-06  2017-05-20
6344
## 4 07sep2009 00:00:00 1600   female white    2000-01-04  2017-05-05
```

```
6331
## 5 19apr2001 00:00:00 1700     male white    2000-01-03  2017-05-05
6332
## 6 16jul2001 00:00:00 1700   female white    2000-01-04  2017-05-19
6345
##   prosecution_length gender.y most_probable_race
## 1               1119   female              White
## 2                685     <NA>              White
## 3              -1170   female              White
## 4               1481   female              White
## 5                261     male              White
## 6                459   female              White
```

Dropping outliers

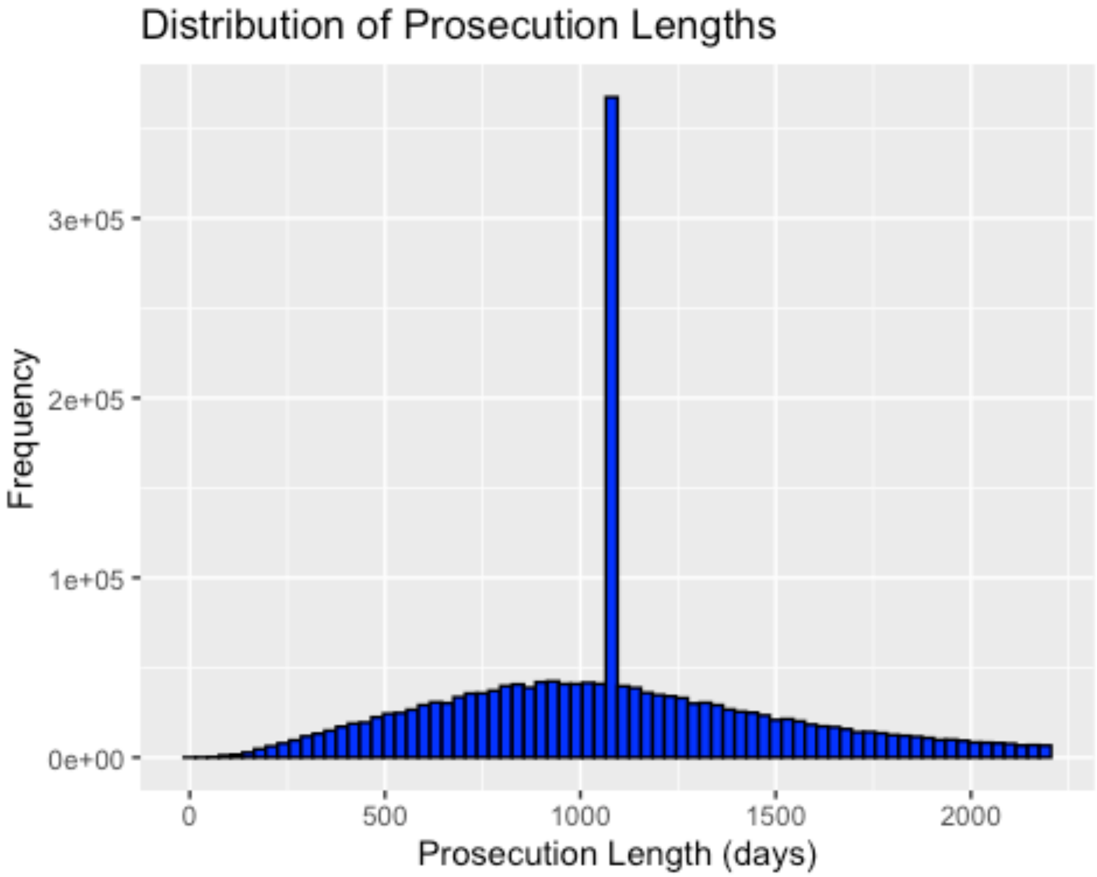
```
# Calculating the IQR for prosecution_length
Q1 <- quantile(applications_with_race$prosecution_length, 0.25, na.rm = TRUE)
Q3 <- quantile(applications_with_race$prosecution_length, 0.75, na.rm = TRUE)
IQR <- Q3 - Q1

# Defining the lower and upper bounds for what's considered an outlier
lower_bound <- Q1 - 1.5 * IQR
upper_bound <- Q3 + 1.5 * IQR

# Filtering out the outliers
applications_with_race <- applications_with_race %>%
  filter(prosecution_length >= lower_bound & prosecution_length <=
upper_bound)

# Checking the result after dropping outliers
summary(applications_with_race$prosecution_length)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       5     807    1079    1076    1290    2205
```

Part 1 - 1 ##############

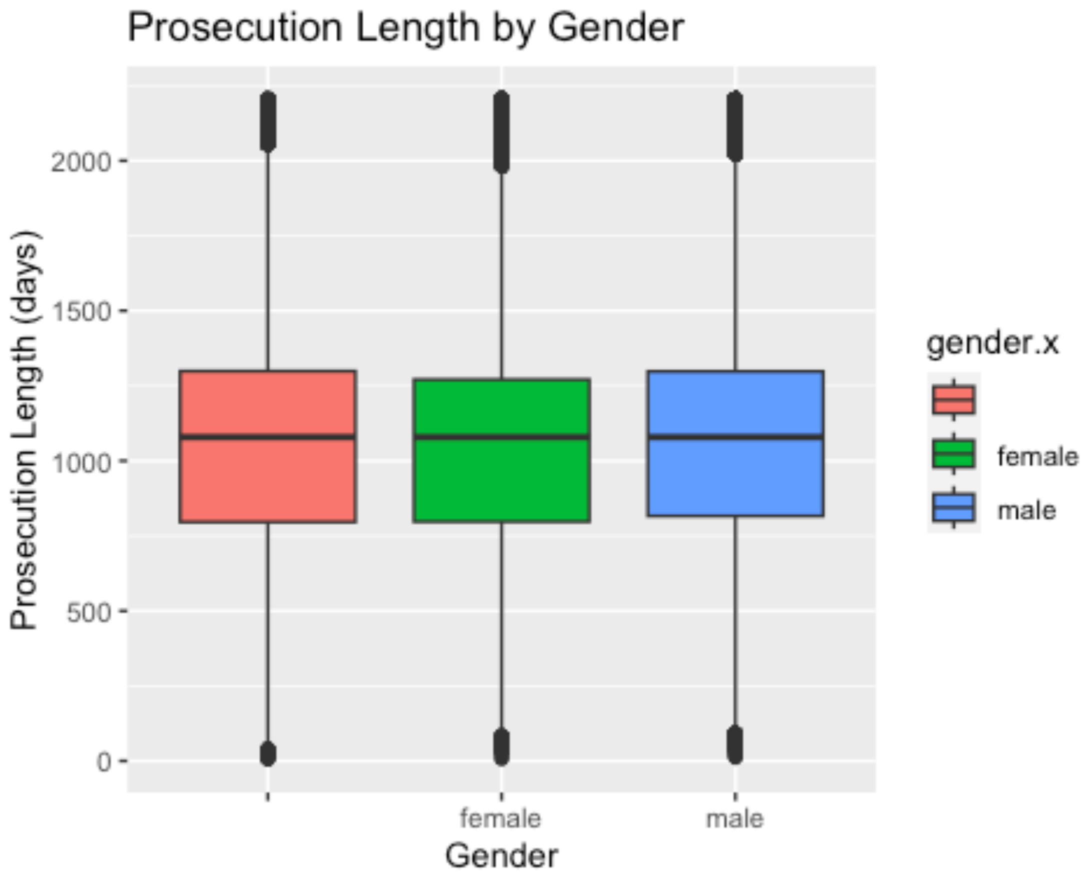
1. Organizational and Social Factors Associated with the Length of Patent Application Prosecution For this analysis, we consider factors such as examiner's art unit (examiner_art_unit), the USPC class/subclass, and demographic attributes (gender, most_probable_race).

Exploratory Analysis:

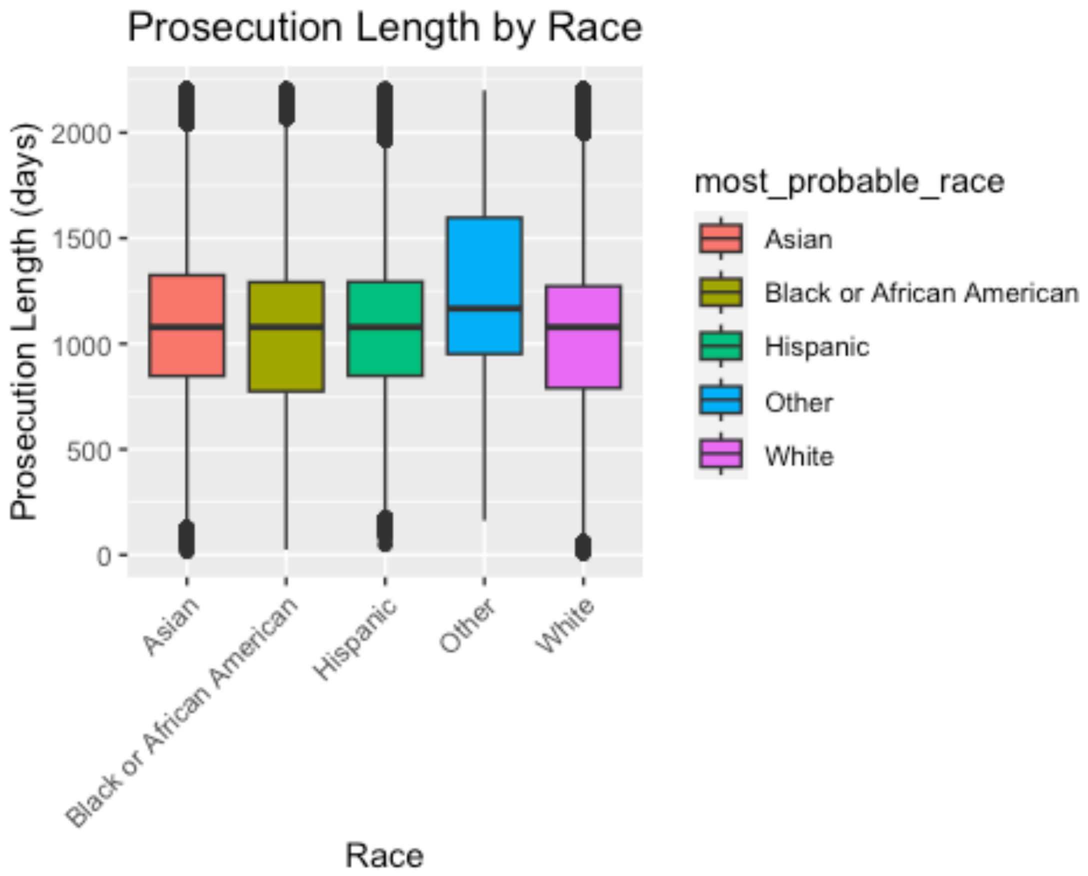
```
library(ggplot2)

# Distribution of prosecution lengths
ggplot(applications_with_race, aes(x = prosecution_length)) +
  geom_histogram(binwidth = 30, fill = "blue", color = "black") +
  labs(title = "Distribution of Prosecution Lengths", x = "Prosecution Length
(days)", y = "Frequency")
```



```
# Prosecution length by gender
ggplot(applications_with_race, aes(x = gender.x, y = prosecution_length, fill
= gender.x)) +
  geom_boxplot() +
  labs(title = "Prosecution Length by Gender", x = "Gender", y = "Prosecution
Length (days)")
```



```r
# Prosecution length by race
ggplot(applications_with_race, aes(x = most_probable_race, y =
prosecution_length, fill = most_probable_race)) +
  geom_boxplot() +
  labs(title = "Prosecution Length by Race", x = "Race", y = "Prosecution
Length (days)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Regression Analysis A linear regression can be used to quantify the impact of these factors on prosecution_length.

```
lm_result <- lm(prosecution_length ~ examiner_art_unit + uspc_class +
gender.x + most_probable_race, data = applications_with_race)
summary(lm_result)

##
## Call:
## lm(formula = prosecution_length ~ examiner_art_unit + uspc_class +
##     gender.x + most_probable_race, data = applications_with_race)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1327.83  -250.20   -25.74   212.57  1552.00
##
## Coefficients:
##                                                 Estimate Std. Error t value
## (Intercept)                                    6.513e+02  1.999e+02   3.257
## examiner_art_unit                             -2.200e-01  5.978e-03 -36.809
## uspc_class001                                  1.205e+03  4.465e+02   2.699
## uspc_class002                                  8.229e+02  2.078e+02   3.960
## uspc_class004                                  7.942e+02  2.283e+02   3.478
## uspc_class005                                7.874e+02  2.283e+02   3.449
## uspc_class007                                4.297e+02  3.458e+02   1.243
## uspc_class008                                6.964e+02  1.997e+02   3.487
## uspc_class012                                8.501e+02  2.824e+02   3.011
## uspc_class014                                6.043e+02  2.824e+02   2.140
## uspc_class015                                6.122e+02  1.998e+02   3.065
## uspc_class016                                8.461e+02  2.130e+02   3.972
## uspc_class019                                9.483e+02  2.824e+02   3.358
## uspc_class023                                7.056e+02  2.016e+02   3.499
## uspc_class024                                7.023e+02  2.503e+02   2.806
## uspc_class026                                9.367e+02  2.219e+02   4.221
## uspc_class027                                5.450e+02  3.050e+02   1.787
## uspc_class028                                7.784e+02  2.362e+02   3.295
## uspc_class029                                7.996e+02  2.000e+02   3.998
## uspc_class030                                8.307e+02  2.305e+02   3.603
## uspc_class033                                4.590e+02  3.458e+02   1.327
## uspc_class034                                7.602e+02  2.046e+02   3.716
## uspc_class036                                8.214e+02  2.445e+02   3.359
## uspc_class037                                1.267e+03  4.465e+02   2.837
## uspc_class038                                9.604e+02  4.465e+02   2.151
## uspc_class040                                7.826e+02  2.264e+02   3.457
## uspc_class042                                1.111e+03  2.400e+02   4.631
## uspc_class043                                5.311e+02  2.207e+02   2.406
## uspc_class044                                8.365e+02  1.998e+02   4.187
## uspc_class047                                7.242e+02  2.029e+02   3.568
## uspc_class048                                9.332e+02  1.998e+02   4.670
## uspc_class049                                9.764e+02  2.824e+02   3.458
## uspc_class051                                7.443e+02  1.999e+02   3.724
## uspc_class052                                7.648e+02  2.029e+02   3.769
## uspc_class053                                8.588e+02  2.122e+02   4.048
## uspc_class054                                6.218e+02  3.050e+02   2.039
## uspc_class055                                6.961e+02  1.997e+02   3.485
## uspc_class056                                6.091e+02  2.445e+02   2.491
## uspc_class057                                7.274e+02  2.163e+02   3.363
## uspc_class059                                7.940e+02  4.465e+02   1.779
## uspc_class060                                8.305e+02  2.049e+02   4.053
## uspc_class062                                7.227e+02  2.063e+02   3.503
## uspc_class063                                7.896e+02  3.050e+02   2.589
## uspc_class065                                8.383e+02  1.997e+02   4.197
## uspc_class066                                1.061e+03  3.458e+02   3.067
## uspc_class068                                9.143e+02  1.998e+02   4.576
## uspc_class069                                1.123e+03  3.458e+02   3.248
## uspc_class070                                7.022e+02  2.066e+02   3.400
## uspc_class071                                6.812e+02  2.000e+02   3.406
## uspc_class072                                8.449e+02  2.108e+02   4.009
## uspc_class073                                8.570e+02  2.031e+02   4.219
## uspc_class074                                1.025e+03  2.099e+02   4.883
## uspc_class075                                7.168e+02  1.997e+02   3.589
## uspc_class081                                9.655e+02  2.824e+02   3.419
## uspc_class082                                7.240e+02  2.080e+02   3.481
```

Part 1 -2 ##############

2. Organizational and Social Factors Associated with Examiner Attrition

As there's no straightforward attrition flag or direct indicator of examiner leaving, and considering attrition is meant to reflect whether an examiner has stopped working on applications (which from this dataset might not be directly inferable), we are considering alternative approaches to approximate this concept. Without a clear attrition indicator, we'll focus on what can be analyzed given the dataset.

Alternative Analysis Approach:

Since attrition directly cannot be assessed, let's pivot towards understanding the factors that influence application outcomes (e.g., issued patents vs. abandoned applications), which might provide insights into examiner behavior or organizational processes affecting application processing times or outcomes.

Analyzing Factors Influencing Application Outcomes:

We use the disposal_type as a proxy to differentiate between applications that were successfully issued a patent versus those abandoned or otherwise disposed of. This will allow us to analyze how different factors, including examiner demographics and organizational units, may influence these outcomes.

1- Preparing Data: First, we create a binary outcome variable based on disposal_type.

```
applications_with_race <- applications_with_race %>%
  mutate(outcome = ifelse(disposal_type == "ISS", 1, 0)) # "ISS" indicates
issued patents
```

2- Logistic Regression for Analyzing Influencing Factors: Since we don't have attrition_flag, we'll adjust the focus to the outcome variable just created.

```
# Logistic regression to explore factors influencing application outcomes
glm_outcomes <- glm(outcome ~ examiner_art_unit + tenure_days + gender.x +
race,
                    family = binomial(link = "logit"),
                    data = applications_with_race)

summary(glm_outcomes)

##
## Call:
## glm(formula = outcome ~ examiner_art_unit + tenure_days + gender.x +
##     race, family = binomial(link = "logit"), data =
applications_with_race)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -4.089e-01  1.110e-02 -36.841  < 2e-16 ***
## examiner_art_unit  3.068e-04  5.085e-06  60.328  < 2e-16 ***
## tenure_days       -3.982e-07  1.771e-08 -22.487  < 2e-16 ***
## gender.xfemale    -1.572e-01  4.849e-03 -32.416  < 2e-16 ***
## gender.xmale       5.963e-02  4.450e-03  13.400  < 2e-16 ***
## raceblack          2.526e-01  7.596e-03  33.260  < 2e-16 ***
## raceHispanic      -1.539e-01  8.968e-03 -17.163  < 2e-16 ***
## raceother          2.226e-01  4.981e-02   4.469 7.86e-06 ***
## racewhite         -6.838e-02  3.486e-03 -19.613  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2627056  on 1900660  degrees of freedom
## Residual deviance: 2613941  on 1900652  degrees of freedom
## AIC: 2613959
##
## Number of Fisher Scoring iterations: 4
```

3- Indirect Analysis for Examiner Attrition Given the limitations, we'll focus on exploring patterns that might suggest attrition-like behavior, such as examiners with shorter tenures possibly having different outcome patterns or being associated with specific organizational or social factors.

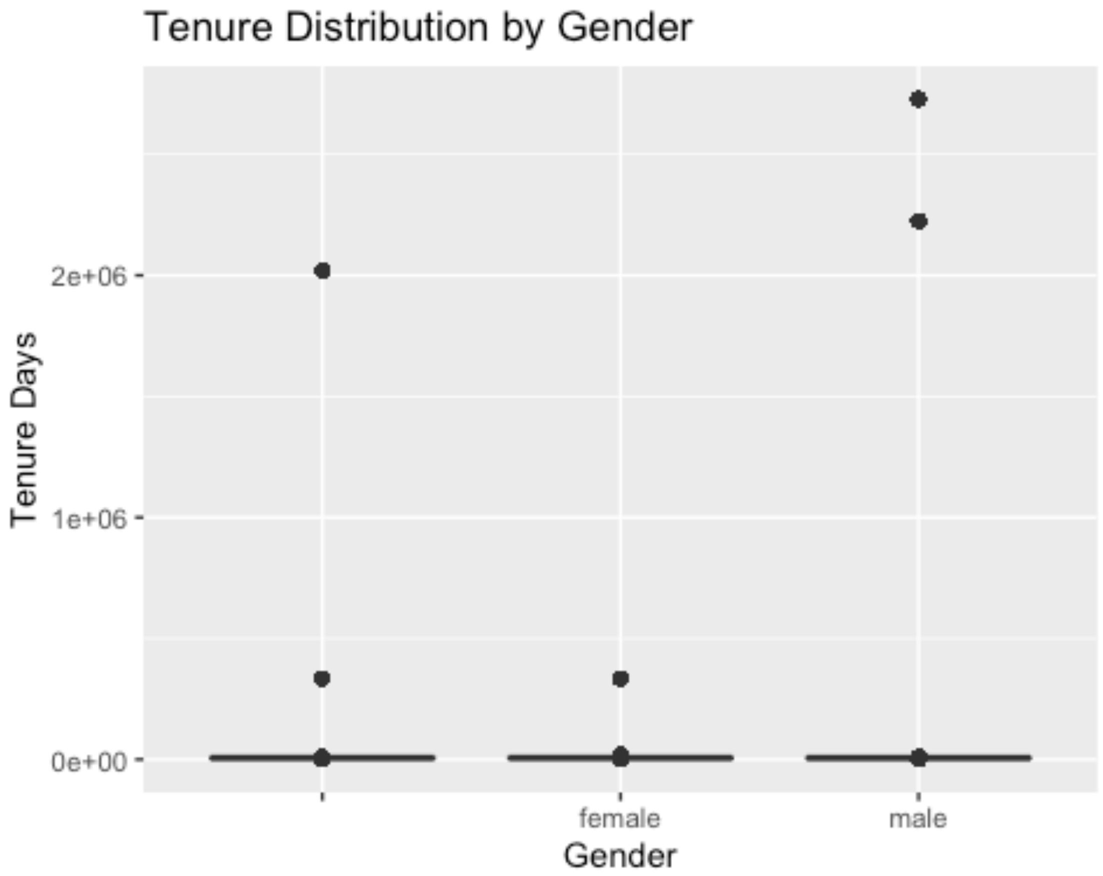
3 - 1 - Aggregate Analysis on Tenure and Outcomes First, we explore the relationship between tenure_days and application outcomes, assuming that changes in tenure_days distribution might reflect on engagement or attrition-like behavior.

```
# Aggregate analysis to explore tenure distribution across different
application outcomes
applications_with_race %>%
  group_by(outcome) %>%
  summarise(Average_Tenure = mean(tenure_days, na.rm = TRUE),
            Median_Tenure = median(tenure_days, na.rm = TRUE)) %>%
  print()
```

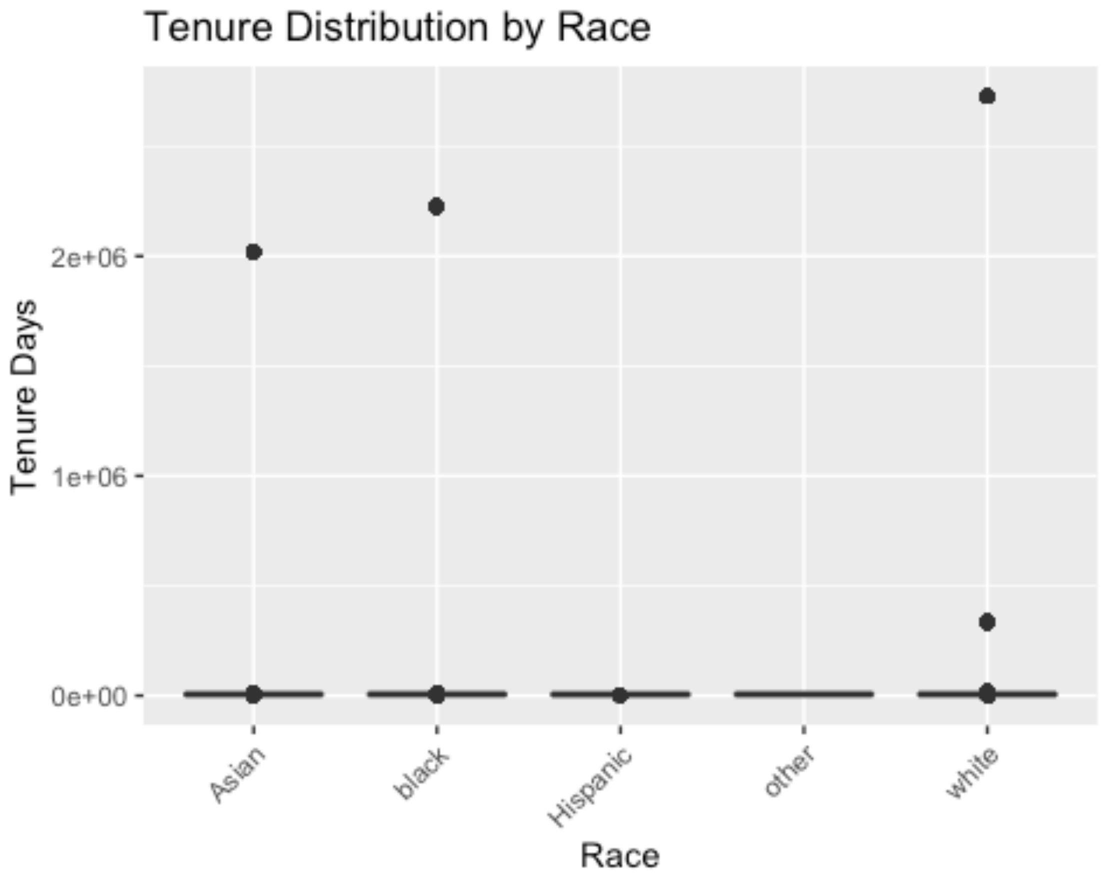
```
## # A tibble: 2 × 3
##   outcome Average_Tenure Median_Tenure
##     <dbl>          <dbl>         <dbl>
## 1       0         11608.          5849
## 2       1          9098.          6272
```

4. Exploratory Analysis on Tenure by Gender and Race Exploring tenure_days distribution across gender.x and race might also offer insights into differing patterns that could indirectly relate to attrition or engagement dynamics.

```
# Tenure distribution by gender
ggplot(applications_with_race, aes(x = gender.x, y = tenure_days)) +
  geom_boxplot() +
  labs(title = "Tenure Distribution by Gender", x = "Gender", y = "Tenure
Days")
```



```
# Tenure distribution by race
ggplot(applications_with_race, aes(x = race, y = tenure_days)) +
  geom_boxplot() +
  labs(title = "Tenure Distribution by Race", x = "Race", y = "Tenure Days")
+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Part 1 - 3 ##############

3- Role of gender, race and ethnicity here in the processes

Analyzing the Impact of Gender, Race, and Ethnicity Given the results from your logistic regression model for application outcomes and the EDA conducted for tenure distributions, you can interpret the roles of gender, race, and ethnicity as follows:

1. Interpretation from Logistic Regression Model The coefficients from the logistic regression model (glm_outcomes) give us quantitative insights into how gender and race are associated with the likelihood of an application being issued a patent (outcome).

```
# Recap the summary of the logistic regression model for reference
summary(glm_outcomes)

##
## Call:
## glm(formula = outcome ~ examiner_art_unit + tenure_days + gender.x +
##     race, family = binomial(link = "logit"), data =
applications_with_race)
##
```

```
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)    
## (Intercept)       -4.089e-01  1.110e-02 -36.841  < 2e-16 ***
## examiner_art_unit  3.068e-04  5.085e-06  60.328  < 2e-16 ***
## tenure_days       -3.982e-07  1.771e-08 -22.487  < 2e-16 ***
## gender.xfemale    -1.572e-01  4.849e-03 -32.416  < 2e-16 ***
## gender.xmale       5.963e-02  4.450e-03  13.400  < 2e-16 ***
## raceblack          2.526e-01  7.596e-03  33.260  < 2e-16 ***
## raceHispanic      -1.539e-01  8.968e-03 -17.163  < 2e-16 ***
## raceother          2.226e-01  4.981e-02   4.469 7.86e-06 ***
## racewhite         -6.838e-02  3.486e-03 -19.613  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 2627056  on 1900660  degrees of freedom
## Residual deviance: 2613941  on 1900652  degrees of freedom
## AIC: 2613959
## 
## Number of Fisher Scoring iterations: 4
```

Part 1 - 3 - Interpretation ###########################

From the output of summary(glm_outcomes), we can interpret the coefficients related to gender.x and race as follows:

Gender (gender.x): Coefficients for gender variables (e.g., gender.xfemale, gender.xmale) indicate the influence of an examiner's gender on the likelihood of a patent application being issued. A positive coefficient suggests that being of that gender increases the likelihood of an application resulting in an issuance compared to the baseline gender category, while a negative coefficient suggests a decrease.

Race (race): Similarly, coefficients for different race categories show how the race of an examiner might impact the outcome of patent applications. Positive values indicate an increased likelihood of issuance for examiners of that race, while negative values indicate a decreased likelihood, both relative to the baseline race category.

```r
# Exploring interaction effects (example)
glm_interactions <- glm(outcome ~ examiner_art_unit * gender.x * race + 
tenure_days,
                        family = binomial(link = "logit"),
                        data = applications_with_race)
summary(glm_interactions)
```

```
## 
## Call:
## glm(formula = outcome ~ examiner_art_unit * gender.x * race + 
##     tenure_days, family = binomial(link = "logit"), data = 
applications_with_race)
## 
## Coefficients: (2 not defined because of singularities)
##                                                  Estimate Std. Error z 
value
## (Intercept)                                   -9.154e-01  3.064e-02 -
29.880
## examiner_art_unit                              5.788e-04  1.517e-05  
38.164
## gender.xfemale                                -3.133e-01  4.858e-02  -
6.450
## gender.xmale                                   6.356e-01  4.034e-02  
15.754
## raceblack                                      2.250e+00  7.855e-02  
28.647
## raceHispanic                                  -9.019e-01  1.824e-01  -
4.943
## raceother                                     -1.662e+01  5.312e+00  -
3.129
## racewhite                                     -2.277e-01  5.167e-02  -
4.407
## tenure_days                                   -3.672e-07  1.838e-08 -
19.984
## examiner_art_unit:gender.xfemale               1.147e-04  2.448e-05   
4.684
## examiner_art_unit:gender.xmale                -3.391e-04  1.969e-05 -
17.220
## examiner_art_unit:raceblack                   -9.262e-04  3.872e-05 -
23.920
## examiner_art_unit:raceHispanic                 3.862e-04  9.563e-05   
4.038
## examiner_art_unit:raceother                    6.966e-03  2.159e-03   
3.227
## examiner_art_unit:racewhite                    4.725e-07  2.583e-05   
0.018
## gender.xfemale:raceblack                      -1.856e+00  1.193e-01 -
15.552
## gender.xmale:raceblack                        -1.655e+00  1.220e-01 -
13.561
## gender.xfemale:raceHispanic                    1.301e+00  2.145e-01   
6.068
## gender.xmale:raceHispanic                      1.742e+00  1.997e-01   
8.724
## gender.xfemale:raceother                              NA         NA      
NA
## gender.xmale:raceother                        -1.031e+00  5.397e+00  -
0.191
## gender.xfemale:racewhite                       9.019e-01  6.928e-02  
13.018
```

```
## gender.xmale:racewhite                         2.573e-01  6.005e-02
4.284
## examiner_art_unit:gender.xfemale:raceblack     7.358e-04  6.120e-05
12.022
## examiner_art_unit:gender.xmale:raceblack       7.771e-04  5.820e-05
13.352
## examiner_art_unit:gender.xfemale:raceHispanic -6.718e-04  1.122e-04  -
5.989
## examiner_art_unit:gender.xmale:raceHispanic   -8.613e-04  1.035e-04  -
8.326
## examiner_art_unit:gender.xfemale:raceother            NA         NA
NA
## examiner_art_unit:gender.xmale:raceother       1.693e-03  2.207e-03
0.767
## examiner_art_unit:gender.xfemale:racewhite    -4.461e-04  3.545e-05 -
12.586
## examiner_art_unit:gender.xmale:racewhite      -7.557e-06  2.987e-05  -
0.253
##                                               Pr(>|z|)
## (Intercept)                                    < 2e-16 ***
## examiner_art_unit                              < 2e-16 ***
## gender.xfemale                                1.12e-10 ***
## gender.xmale                                   < 2e-16 ***
## raceblack                                      < 2e-16 ***
## raceHispanic                                  7.67e-07 ***
## raceother                                      0.00176 **
## racewhite                                     1.05e-05 ***
## tenure_days                                    < 2e-16 ***
## examiner_art_unit:gender.xfemale              2.81e-06 ***
## examiner_art_unit:gender.xmale                 < 2e-16 ***
## examiner_art_unit:raceblack                    < 2e-16 ***
## examiner_art_unit:raceHispanic                5.39e-05 ***
## examiner_art_unit:raceother                    0.00125 **
## examiner_art_unit:racewhite                    0.98541
## gender.xfemale:raceblack                       < 2e-16 ***
## gender.xmale:raceblack                         < 2e-16 ***
## gender.xfemale:raceHispanic                   1.30e-09 ***
## gender.xmale:raceHispanic                      < 2e-16 ***
## gender.xfemale:raceother                            NA
## gender.xmale:raceother                         0.84847
## gender.xfemale:racewhite                       < 2e-16 ***
## gender.xmale:racewhite                        1.83e-05 ***
## examiner_art_unit:gender.xfemale:raceblack     < 2e-16 ***
## examiner_art_unit:gender.xmale:raceblack       < 2e-16 ***
## examiner_art_unit:gender.xfemale:raceHispanic 2.11e-09 ***
## examiner_art_unit:gender.xmale:raceHispanic    < 2e-16 ***
## examiner_art_unit:gender.xfemale:raceother          NA
## examiner_art_unit:gender.xmale:raceother       0.44304
## examiner_art_unit:gender.xfemale:racewhite     < 2e-16 ***
## examiner_art_unit:gender.xmale:racewhite       0.80027
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2627056  on 1900660  degrees of freedom
## Residual deviance: 2610065  on 1900632  degrees of freedom
## AIC: 2610123
##
## Number of Fisher Scoring iterations: 4
```