Data Science Salaries - DATA 557 WI2020

06 February, 2020

```
# Load data and libraries

library(tidyverse)
```

```
## -- Attaching packages ----------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts -------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
dataraw <- read_csv('multipleChoiceResponses.csv')
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   Age = col_double(),
##   LearningCategorySelftTaught = col_double(),
##   LearningCategoryOnlineCourses = col_double(),
##   LearningCategoryWork = col_double(),
##   LearningCategoryUniversity = col_double(),
##   LearningCategoryKaggle = col_double(),
##   LearningCategoryOther = col_double(),
##   WorkToolsFrequencyKNIMECommercial = col_logical(),
##   TimeGatheringData = col_double(),
##   TimeModelBuilding = col_double(),
##   TimeProduction = col_double(),
##   TimeVisualizing = col_double(),
##   TimeFindingInsights = col_double(),
##   TimeOtherSelect = col_double(),
##   CompensationAmount = col_number()
## )
```

```
## See spec(...) for full column specifications.
## Warning: 43 parsing failures.
##  row                               col           expected           actual
file
## 1209 CompensationAmount                a number           -                'multipleChoiceRes
ponses.csv'
## 1316 WorkToolsFrequencyKNIMECommercial 1/0/T/F/TRUE/FALSE Rarely           'multipleChoiceRes
ponses.csv'
## 1594 WorkToolsFrequencyKNIMECommercial 1/0/T/F/TRUE/FALSE Often            'multipleChoiceRes
ponses.csv'
## 2443 WorkToolsFrequencyKNIMECommercial 1/0/T/F/TRUE/FALSE Most of the time 'multipleChoiceRes
ponses.csv'
## 2466 WorkToolsFrequencyKNIMECommercial 1/0/T/F/TRUE/FALSE Most of the time 'multipleChoiceRes
ponses.csv'
## .... ................................. .................. ................
.............................
## See problems(...) for more details.
```

```
# check NAs
check_df_nulls <- function(df) {
  ## Function to check the number of NAs in a df
  Column.Name <- rep(NA, dim(df)[2])
  Col.Type <- sapply(df, typeof)
  Column.Type <- rep(NA, dim(df)[2])
  Number.NAs<-rep(NA, dim(df)[2])
  per.of.NAs<-rep(NA, dim(df)[2])


  for(i in 1:dim(df)[2])
  {
    #cat(sprintf('Column %.0f: %30s \t Number of NAs: %.0f \t Percent NA data: %.0f%% \n'
    #             , i,names(df)[i], length(which(is.na(df[,i]))),100*length(which(is.na(df[,
i])))/dim(df)[1]))
    Column.Name[i] <- names(df)[i]
    Number.NAs[i]  <- length(which(is.na(df[,i])))
    Column.Type[i] <- Col.Type[[i]]

  }
  df_NAs <- data.frame(Column.Name, Column.Type, Number.NAs, per.of.NAs)
  df_NAs$per.of.NAs <- round(100*(Number.NAs/dim(df)[1]),0)
  print(df_NAs)

}
```

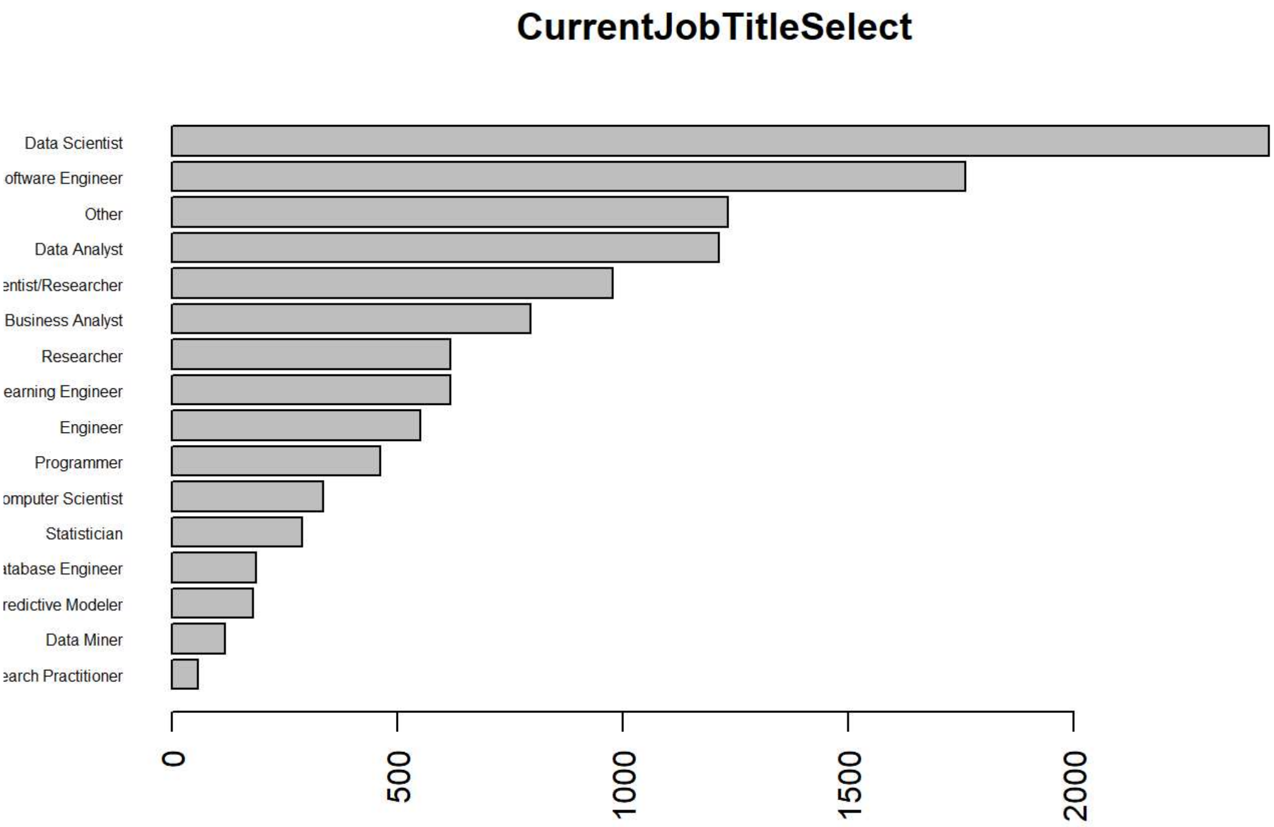
```
# Dimensions of data
dim(dataraw)
```

```
## [1] 16716   228
```

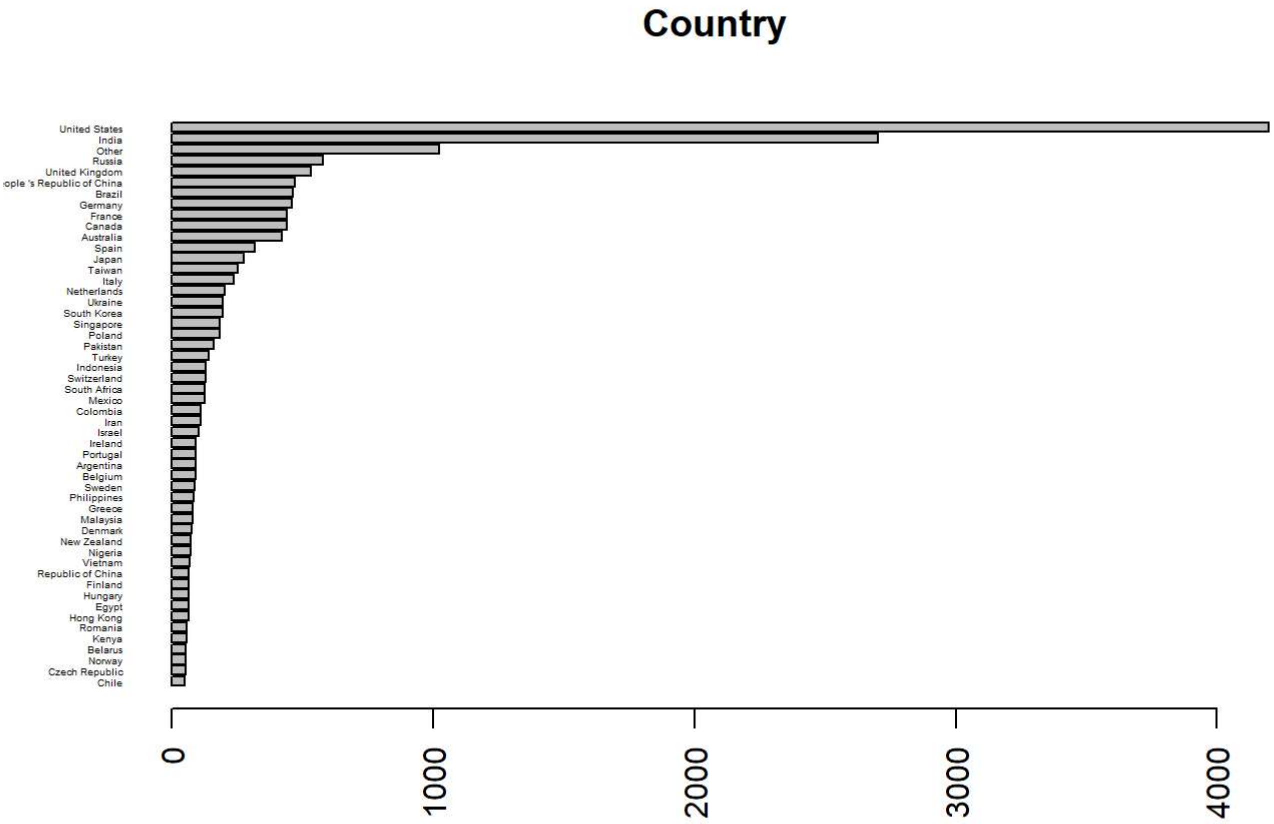
```
#  checking the structure and NAs of data through function
check_df_nulls(dataraw)
```

```
##                                 Column.Name Column.Type Number.NAs
## 1                              GenderSelect   character         95
## 2                                   Country   character        121
## 3                                       Age      double        331
## 4                          EmploymentStatus   character          0
## 5                             StudentStatus   character      15436
## 6                       LearningDataScience   character      15432
## 7                                CodeWriter   character       3530
## 8                            CareerSwitcher   character      13704
## 9                     CurrentJobTitleSelect   character       4886
## 10                                 TitleFit   character       5212
## 11                      CurrentEmployerType   character       5115
## 12                     MLToolNextYearSelect   character       5718
## 13                   MLMethodNextYearSelect   character       5883
## 14             LanguageRecommendationSelect   character       5718
## 15                     PublicDatasetsSelect   character       5920
## 16                   LearningPlatformSelect   character       5445
## 17          LearningPlatformUsefulnessArxiv   character      14325
## 18          LearningPlatformUsefulnessBlogs   character      11951
## 19        LearningPlatformUsefulnessCollege   character      13357
## 20        LearningPlatformUsefulnessCompany   character      15735
## 21    LearningPlatformUsefulnessConferences   character      14534
## 22        LearningPlatformUsefulnessFriends   character      15135
## 23         LearningPlatformUsefulnessKaggle   character      10133
## 24    LearningPlatformUsefulnessNewsletters   character      15627
## 25    LearningPlatformUsefulnessCommunities   character      15574
## 26  LearningPlatformUsefulnessDocumentation   character      14395
## 27        LearningPlatformUsefulnessCourses   character      10724
## 28       LearningPlatformUsefulnessProjects   character      11922
## 29       LearningPlatformUsefulnessPodcasts   character      15502
## 30             LearningPlatformUsefulnessSO   character      11076
## 31       LearningPlatformUsefulnessTextbook   character      12535
## 32      LearningPlatformUsefulnessTradeBook   character      16383
## 33       LearningPlatformUsefulnessTutoring   character      15290
## 34        LearningPlatformUsefulnessYouTube   character      11487
## 35           BlogsPodcastsNewslettersSelect   character       8576
## 36                  LearningDataScienceTime   character      12367
## 37                  JobSkillImportanceBigData   character      12760
## 38                   JobSkillImportanceDegree   character      12807
## 39                    JobSkillImportanceStats   character      12756
## 40          JobSkillImportanceEnterpriseTools   character      13022
## 41                   JobSkillImportancePython   character      12685
## 42                        JobSkillImportanceR   character      12772
## 43                      JobSkillImportanceSQL   character      12824
## 44            JobSkillImportanceKaggleRanking   character      12846
## 45                     JobSkillImportanceMOOC   character      12884
## 46           JobSkillImportanceVisualizations   character      12849
## 47             JobSkillImportanceOtherSelect1   character      16358
## 48             JobSkillImportanceOtherSelect2   character      16544
## 49             JobSkillImportanceOtherSelect3   character      16603
## 50                     CoursePlatformSelect   character      14420
## 51           HardwarePersonalProjectsSelect   character      12510
## 52                        TimeSpentStudying   character      12688
## 53                      ProveKnowledgeSelect   character      12555
## 54                 DataScienceIdentitySelect   character       4045
## 55                           FormalEducation   character       1701
## 56                               MajorSelect   character       3435
## 57                                    Tenure   character       3184
## 58                       PastJobTitlesSelect   character       2524
## 59                       FirstTrainingSelect   character       2004
## 60               LearningCategorySelftTaught      double       3607
## 61             LearningCategoryOnlineCourses      double       3590
## 62                      LearningCategoryWork      double       3605
## 63                LearningCategoryUniversity      double       3594
## 64                    LearningCategoryKaggle      double       3590
## 65                     LearningCategoryOther      double       3622
## 66                            MLSkillsSelect   character       3963
## 67                        MLTechniquesSelect   character       4132
## 68                          ParentsEducation   character       4048
## 69                          EmployerIndustry   character       5970
## 70                              EmployerSize   character       8945
## 71                        EmployerSizeChange   character       9159
## 72                            EmployerMLTime   character       9065
## 73                      EmployerSearchMethod   character       8988
## 74                      UniversityImportance   character       8618
## 75                         JobFunctionSelect   character       8735
## 76                        WorkHardwareSelect   character       8698
## 77                        WorkDataTypeSelect   character       8692
## 78                   WorkProductionFrequency   character       9594
## 79                           WorkDatasetSize   character       9628
## 80                      WorkAlgorithmsSelect   character       9415
## 81                           WorkToolsSelect   character       8761
## 82                WorkToolsFrequencyAmazonML   character      16317
## 83                     WorkToolsFrequencyAWS   character      14896
## 84                  WorkToolsFrequencyAngoss   character      16694
## 85                       WorkToolsFrequencyC   character      15215
## 86                WorkToolsFrequencyCloudera   character      16277
## 87               WorkToolsFrequencyDataRobot   character      16653
## 88                   WorkToolsFrequencyFlume   character      16575
## 89                     WorkToolsFrequencyGCP   character      16192
## 90                  WorkToolsFrequencyHadoop   character      15379
## 91               WorkToolsFrequencyIBMCognos   character      16561
## 92          WorkToolsFrequencyIBMSPSSModeler   character      16450
## 93       WorkToolsFrequencyIBMSPSSStatistics   character      16258
## 94               WorkToolsFrequencyIBMWatson   character      16473
## 95                  WorkToolsFrequencyImpala   character      16464
## 96                    WorkToolsFrequencyJava   character      15311
## 97                   WorkToolsFrequencyJulia   character      16536
## 98                 WorkToolsFrequencyJupyter   character      13547
## 99         WorkToolsFrequencyKNIMECommercial     logical      16716
## 100              WorkToolsFrequencyKNIMEFree   character      16449
## 101            WorkToolsFrequencyMathematica   character      16424
## 102                 WorkToolsFrequencyMATLAB   character      15292
## 103                  WorkToolsFrequencyAzure   character      16141
## 104                  WorkToolsFrequencyExcel   character      15668
## 105       WorkToolsFrequencyMicrosoftRServer   character      16344
## 106           WorkToolsFrequencyMicrosoftSQL   character      16277
## 107                  WorkToolsFrequencyMinitab   character      16572
## 108                    WorkToolsFrequencyNoSQL   character      15238
## 109                   WorkToolsFrequencyOracle   character      16508
## 110                   WorkToolsFrequencyOrange   character      16594
## 111                     WorkToolsFrequencyPerl   character      16414
## 112                   WorkToolsFrequencyPython   character      10726
## 113                     WorkToolsFrequencyQlik   character      16352
## 114                        WorkToolsFrequencyR   character      12078
## 115     WorkToolsFrequencyRapidMinerCommercial   character      16642
## 116           WorkToolsFrequencyRapidMinerFree   character      16389
## 117                  WorkToolsFrequencySalfrod   character      16684
## 118       WorkToolsFrequencySAPBusinessObjects   character      16625
## 119                  WorkToolsFrequencySASBase   character      16001
## 120            WorkToolsFrequencySASEnterprise   character      16329
## 121                   WorkToolsFrequencySASJMP   character      16601
## 122                    WorkToolsFrequencySpark   character      15391
## 123                      WorkToolsFrequencySQL   character      12528
## 124                     WorkToolsFrequencyStan   character      16564
## 125               WorkToolsFrequencyStatistica   character      16674
## 126                  WorkToolsFrequencyTableau   character      15132
## 127               WorkToolsFrequencyTensorFlow   character      14494
## 128                    WorkToolsFrequencyTIBCO   character      16577
## 129                     WorkToolsFrequencyUnix   character      14888
## 130                  WorkToolsFrequencySelect1   character      16030
## 131                  WorkToolsFrequencySelect2   character      16581
## 132                      WorkFrequencySelect3   character      16635
## 133                         WorkMethodsSelect   character       8943
## 134                    WorkMethodsFrequencyA/B   character      14846
## 135       WorkMethodsFrequencyAssociationRules   character      15620
## 136               WorkMethodsFrequencyBayesian   character      14871
## 137                   WorkMethodsFrequencyCNNs   character      15363
## 138 WorkMethodsFrequencyCollaborativeFiltering   character      15955
## 139       WorkMethodsFrequencyCross-Validation   character      12956
## 140      WorkMethodsFrequencyDataVisualization   character      11810
## 141          WorkMethodsFrequencyDecisionTrees   character      13134
## 142        WorkMethodsFrequencyEnsembleMethods   character      14733
## 143 WorkMethodsFrequencyEvolutionaryApproaches   character      16302
## 144                   WorkMethodsFrequencyGANs   character      16486
## 145                    WorkMethodsFrequencyGBM   character      15211
## 146                   WorkMethodsFrequencyHMMs   character      16317
## 147                    WorkMethodsFrequencyKNN   character      14171
## 148           WorkMethodsFrequencyLiftAnalysis   character      16093
## 149     WorkMethodsFrequencyLogisticRegression   character      12544
## 150                    WorkMethodsFrequencyMLN   character      16474
## 151             WorkMethodsFrequencyNaiveBayes   character      14910
## 152                    WorkMethodsFrequencyNLP   character      14840
## 153         WorkMethodsFrequencyNeuralNetworks   character      14006
## 154                    WorkMethodsFrequencyPCA   character      14014
## 155   WorkMethodsFrequencyPrescriptiveModeling   character      15899
## 156          WorkMethodsFrequencyRandomForests   character      13360
## 157     WorkMethodsFrequencyRecommenderSystems   character      15604
## 158                   WorkMethodsFrequencyRNNs   character      15868
## 159           WorkMethodsFrequencySegmentation   character      14739
## 160             WorkMethodsFrequencySimulation   character      15365
## 161                    WorkMethodsFrequencySVMs   character      14813
## 162            WorkMethodsFrequencyTextAnalysis   character      14385
## 163      WorkMethodsFrequencyTimeSeriesAnalysis   character      13644
## 164                 WorkMethodsFrequencySelect1   character      16483
## 165                 WorkMethodsFrequencySelect2   character      16677
## 166                 WorkMethodsFrequencySelect3   character      16623
## 167                           TimeGatheringData      double       9186
## 168                           TimeModelBuilding      double       9188
## 169                              TimeProduction      double       9199
## 170                             TimeVisualizing      double       9187
## 171                         TimeFindingInsights      double       9193
## 172                             TimeOtherSelect      double       9203
## 173                 AlgorithmUnderstandingLevel   character       9306
## 174                        WorkChallengesSelect   character       9340
## 175              WorkChallengeFrequencyPolitics   character      14036
## 176         WorkChallengeFrequencyUnusedResults   character      14972
## 177 WorkChallengeFrequencyUnusefulInstrumenting   character      16077
## 178            WorkChallengeFrequencyDeployment   character      15869
## 179             WorkChallengeFrequencyDirtyData   character      13165
## 180            WorkChallengeFrequencyExplaining   character      15131
## 181                  WorkChallengeFrequencyPass   character      16292
## 182           WorkChallengeFrequencyIntegration   character      15744
## 183                WorkChallengeFrequencyTalent   character      13720
## 184             WorkChallengeFrequencyDataFunds   character      15764
## 185       WorkChallengeFrequencyDomainExpertise   character      15308
## 186                    WorkChallengeFrequencyML   character      15951
## 187                 WorkChallengeFrequencyTools   character      15537
## 188          WorkChallengeFrequencyExpectations   character      15582
## 189        WorkChallengeFrequencyITCoordination   character      15547
## 190           WorkChallengeFrequencyHiringFunds   character      15429
## 191               WorkChallengeFrequencyPrivacy   character      15294
## 192               WorkChallengeFrequencyScaling   character      15883
## 193          WorkChallengeFrequencyEnvironments   character      15463
## 194               WorkChallengeFrequencyClarity   character      14537
## 195            WorkChallengeFrequencyDataAccess   character      14526
## 196           WorkChallengeFrequencyOtherSelect   character      16439
## 197                      WorkDataVisualizations   character       9837
## 198                 WorkInternalVsExternalTools   character       9959
## 199                        WorkMLTeamSeatSelect   character      10028
## 200                                WorkDatasets   character      14508
## 201                       WorkDatasetsChallenge   character      14148
## 202                             WorkDataStorage   character      10201
## 203                             WorkDataSharing   character      10214
## 204                            WorkDataSourcing   character      16335
## 205                             WorkCodeSharing   character      10513
## 206                                  RemoteWork   character      10619
## 207                          CompensationAmount      double      11499
## 208                        CompensationCurrency   character      12186
## 209                                SalaryChange   character      10327
## 210                             JobSatisfaction   character      10039
## 211                           JobSearchResource   character      12977
## 212                                 JobHuntTime   character      12985
## 213                           JobFactorLearning   character      13165
## 214                             JobFactorSalary   character      13231
## 215                             JobFactorOffice   character      13248
## 216                          JobFactorLanguages   character      13241
## 217                            JobFactorCommute   character      13269
## 218                         JobFactorManagement   character      13282
## 219                    JobFactorExperienceLevel   character      13279
## 220                         JobFactorDepartment   character      13300
## 221                              JobFactorTitle   character      13302
## 222                     JobFactorCompanyFunding   character      13305
## 223                             JobFactorImpact   character      13322
## 224                             JobFactorRemote   character      13292
## 225                           JobFactorIndustry   character      13307
## 226                   JobFactorLeaderReputation   character      13315
## 227                          JobFactorDiversity   character      13306
## 228              JobFactorPublishingOpportunity   character      13292
##     per.of.NAs
## 1            1
## 2            1
## 3            2
## 4            0
## 5           92
## 6           92
## 7           21
## 8           82
## 9           29
## 10          31
## 11          31
## 12          34
## 13          35
## 14          34
## 15          35
## 16          33
## 17          86
## 18          71
## 19          80
## 20          94
## 21          87
## 22          91
## 23          61
## 24          93
## 25          93
## 26          86
## 27          64
## 28          71
## 29          93
## 30          66
## 31          75
## 32          98
## 33          91
## 34          69
## 35          51
## 36          74
## 37          76
## 38          77
## 39          76
## 40           78
## 41           76
## 42           76
## 43           77
## 44           77
## 45           77
## 46           77
## 47           98
## 48           99
## 49           99
## 50           86
## 51           75
## 52           76
## 53           75
## 54           24
## 55           10
## 56           21
## 57           19
## 58           15
## 59           12
## 60           22
## 61           21
## 62           22
## 63           22
## 64           21
## 65           22
## 66           24
## 67           25
## 68           24
## 69           36
## 70           54
## 71           55
## 72           54
## 73           54
## 74           52
## 75           52
## 76           52
## 77           52
## 78           57
## 79           58
## 80           56
## 81           52
## 82           98
## 83           89
## 84          100
## 85           91
## 86           97
## 87          100
## 88           99
## 89           97
## 90           92
## 91           99
## 92           98
## 93           97
## 94          99
## 95          98
## 96          92
## 97          99
## 98          81
## 99         100
## 100         98
## 101         98
## 102         91
## 103         97
## 104         94
## 105         98
## 106         97
## 107         99
## 108         91
## 109         99
## 110         99
## 111         98
## 112         64
## 113         98
## 114         72
## 115        100
## 116         98
## 117        100
## 118         99
## 119         96
## 120         98
## 121         99
## 122         92
## 123         75
## 124         99
## 125        100
## 126         91
## 127         87
## 128         99
## 129         89
## 130         96
## 131         99
## 132        100
## 133         53
## 134         89
## 135         93
## 136         89
## 137         92
## 138         95
## 139         78
## 140         71
## 141         79
## 142         88
## 143         98
## 144         99
## 145         91
## 146         98
## 147         85
## 148    96
## 149    75
## 150    99
## 151    89
## 152    89
## 153    84
## 154    84
## 155    95
## 156    80
## 157    93
## 158    95
## 159    88
## 160    92
## 161    89
## 162    86
## 163    82
## 164    99
## 165   100
## 166    99
## 167    55
## 168    55
## 169    55
## 170    55
## 171    55
## 172    55
## 173    56
## 174    56
## 175    84
## 176    90
## 177    96
## 178    95
## 179    79
## 180    91
## 181    97
## 182    94
## 183    82
## 184    94
## 185    92
## 186    95
## 187    93
## 188    93
## 189    93
## 190    92
## 191    91
## 192    95
## 193    93
## 194    87
## 195    87
## 196    98
## 197    59
## 198    60
## 199    60
## 200    87
## 201    85
## 202         61
## 203         61
## 204         98
## 205         63
## 206         64
## 207         69
## 208         73
## 209         62
## 210         60
## 211         78
## 212         78
## 213         79
## 214         79
## 215         79
## 216         79
## 217         79
## 218         79
## 219         79
## 220         80
## 221         80
## 222         80
## 223         80
## 224         80
## 225         80
## 226         80
## 227         80
## 228         80
```

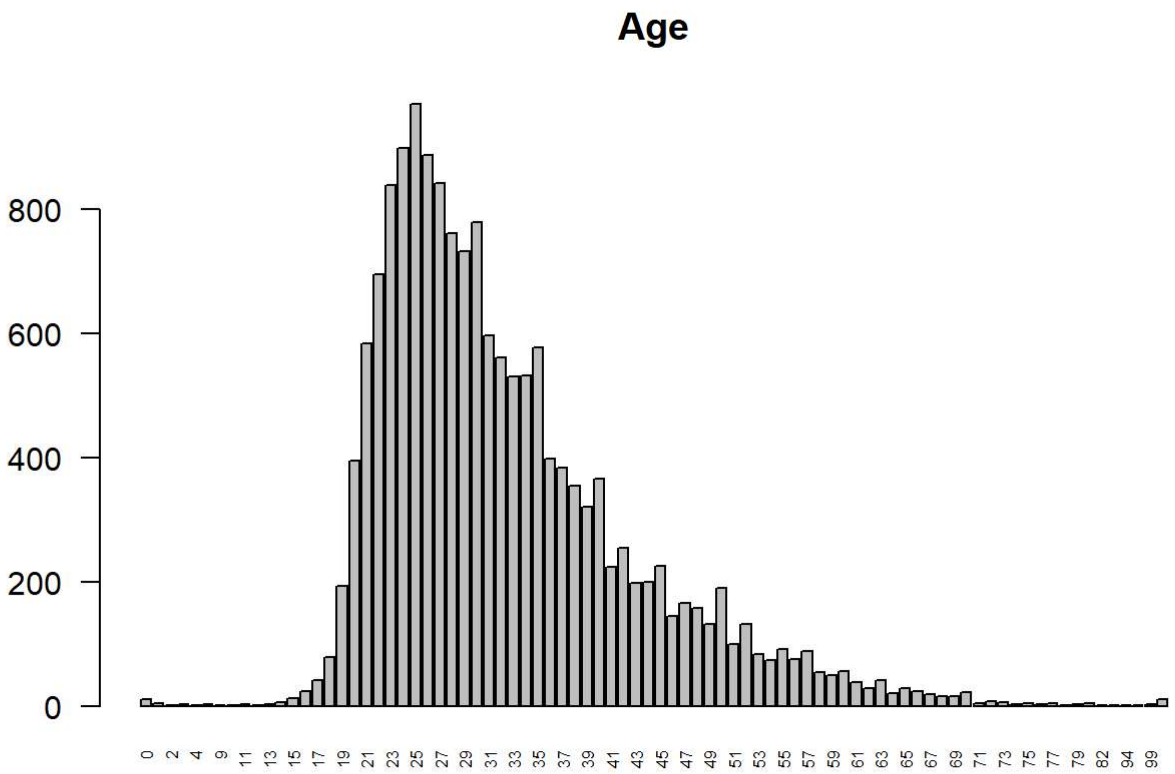
```
barplot(sort(table(dataraw[,'CurrentJobTitleSelect'])), main = 'CurrentJobTitleSelect',horiz = T
, cex.names=0.5, las = 2)
```



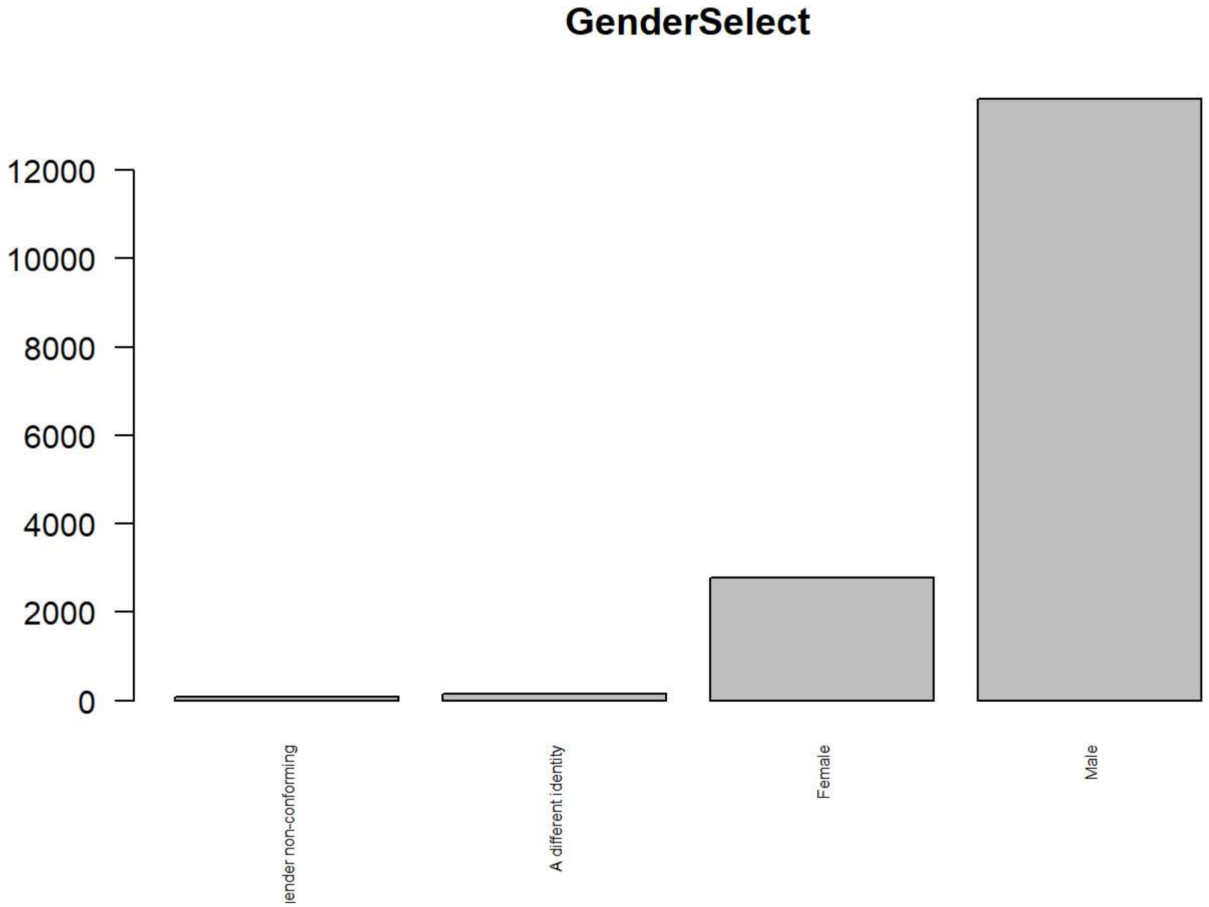
```
barplot(sort(table(dataraw[,'Country'])), main = 'Country',horiz = T, cex.names=0.3, las = 2)
```



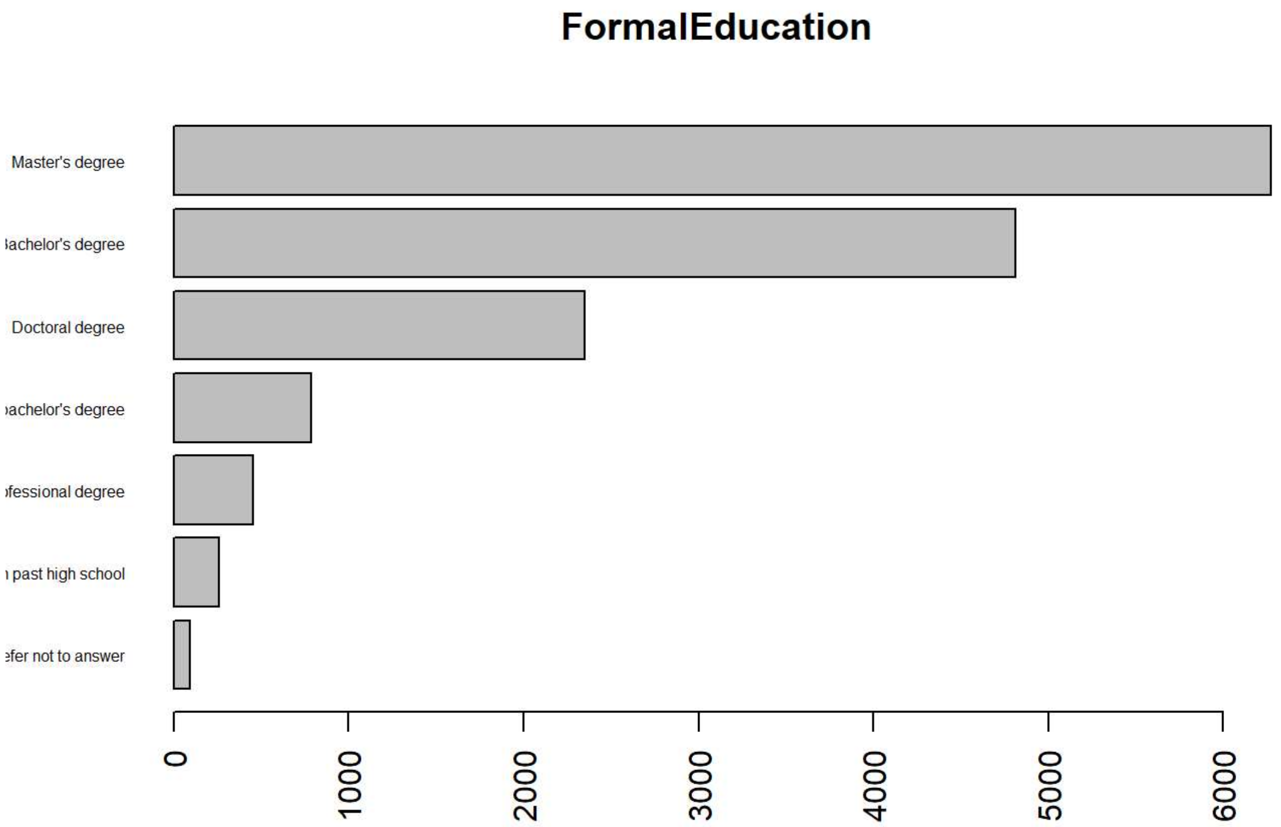
```
barplot((table(dataraw[,'Age'])), main = 'Age',horiz = F, cex.names=0.5, las = 2)
```



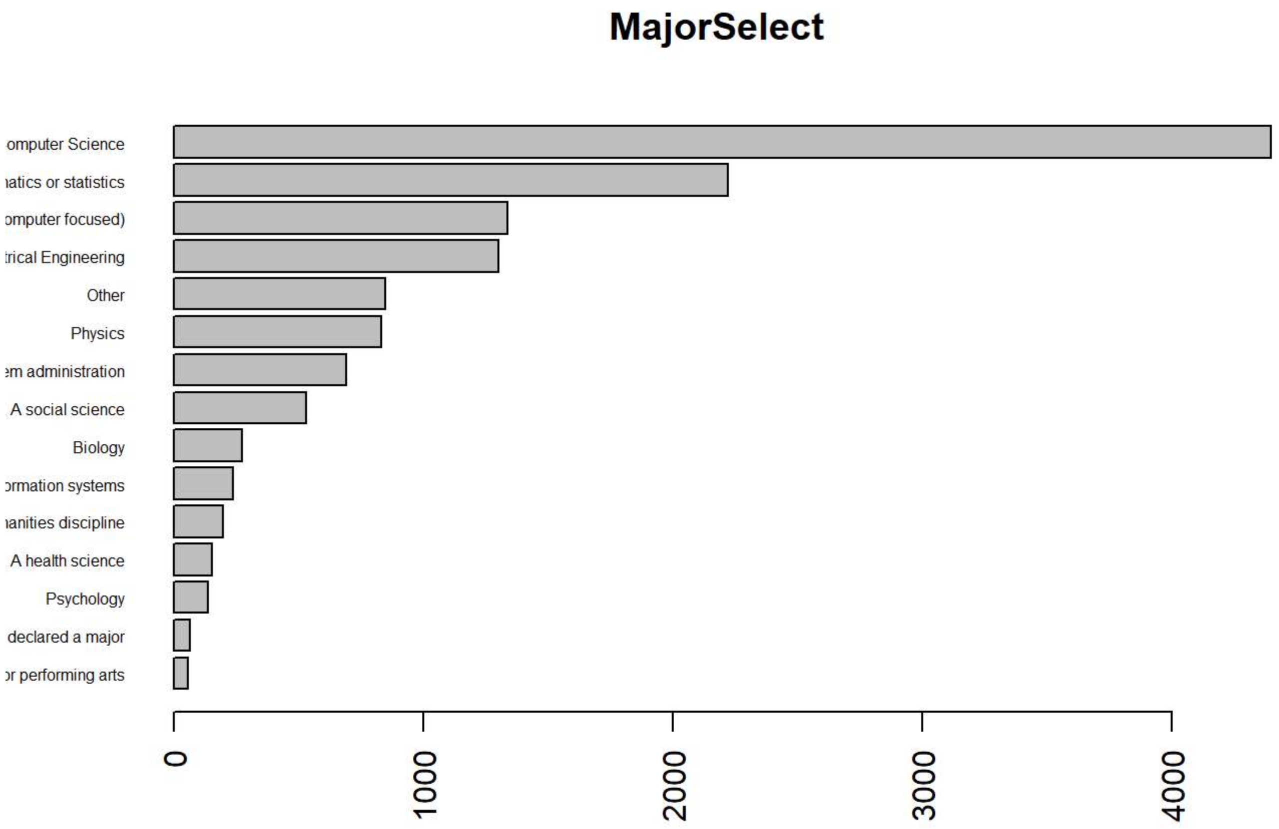
```
barplot(sort(table(dataraw[,'GenderSelect'])), main = 'GenderSelect',horiz = F, cex.names=0.5, las = 2)
```



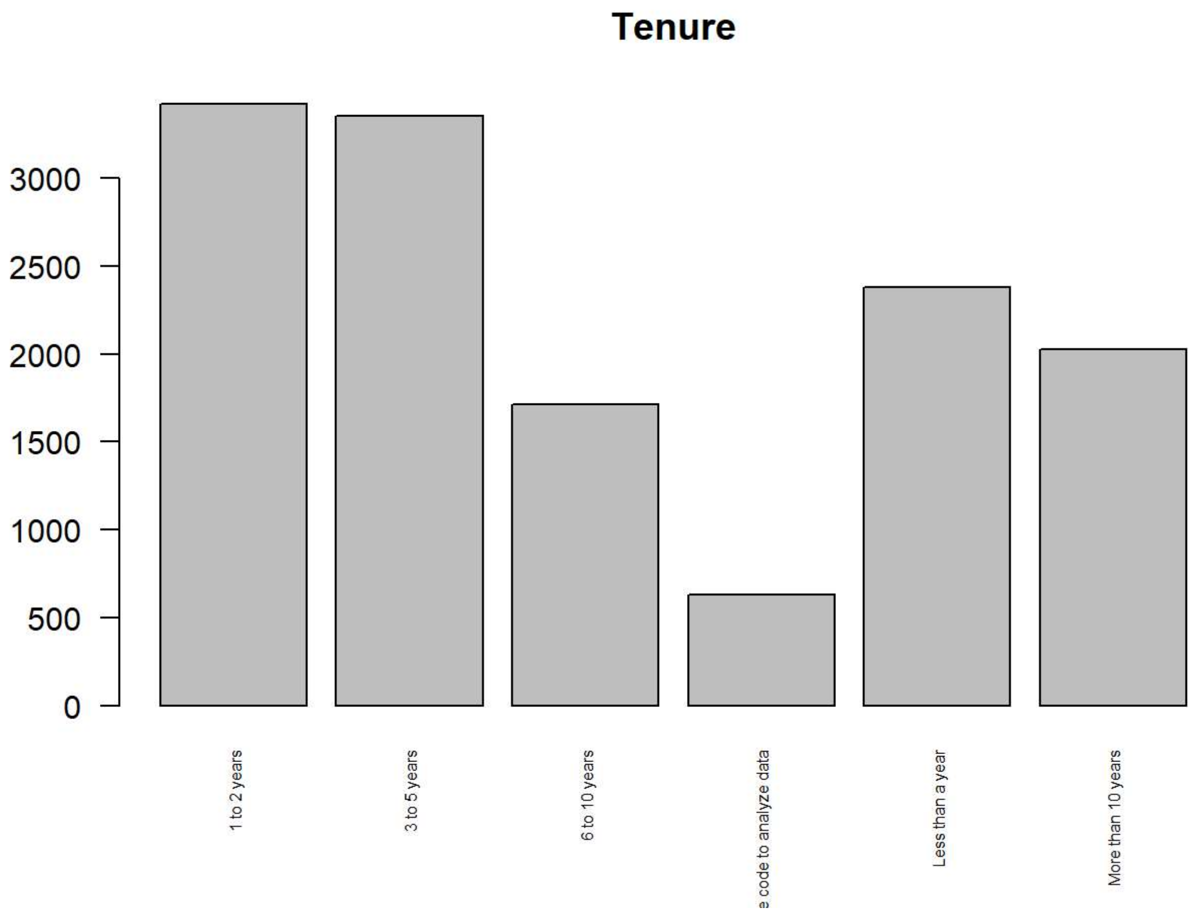
```
barplot(sort(table(dataraw[,'FormalEducation'])), main = 'FormalEducation',horiz = T, cex.names=
0.5, las = 2)
```



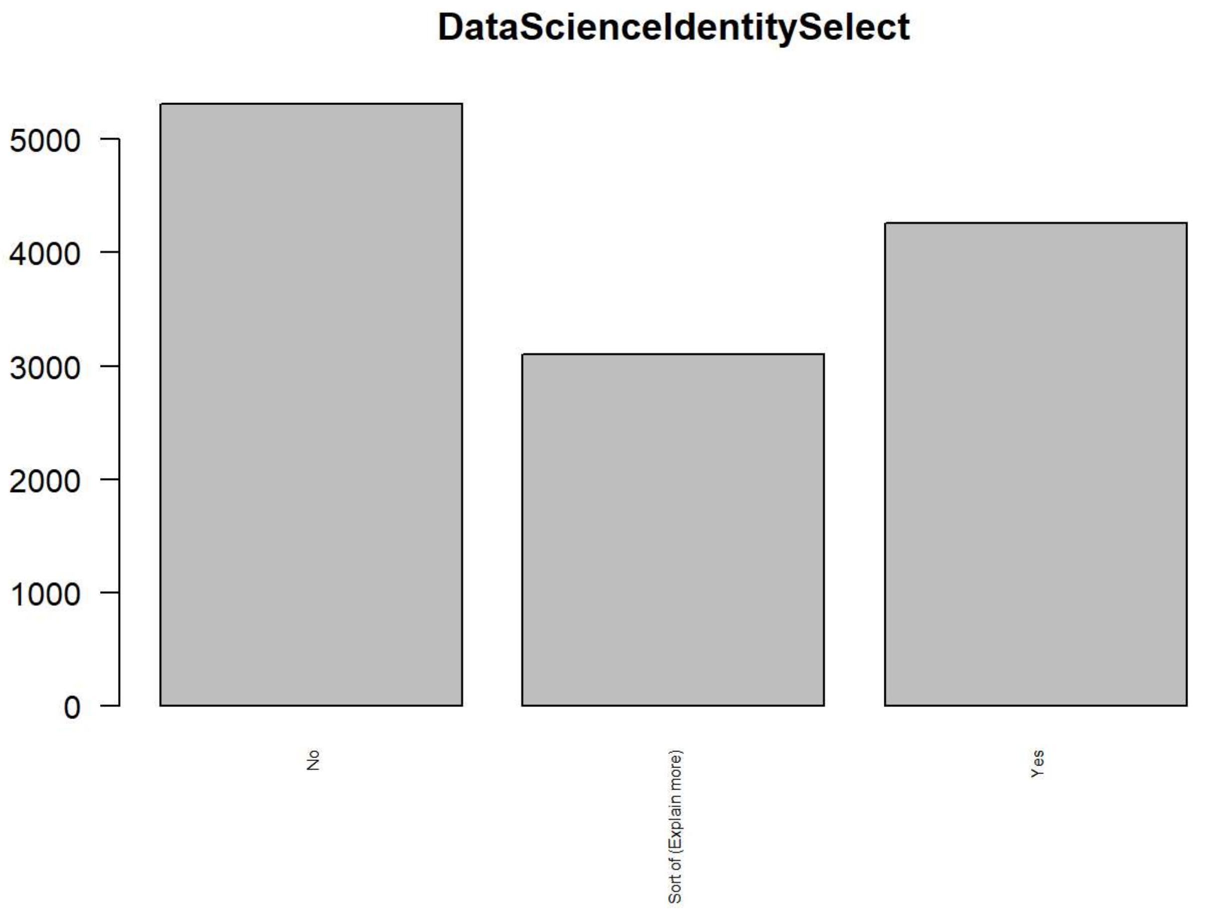
```
barplot(sort(table(dataraw[,'MajorSelect'])), main = 'MajorSelect',horiz = T, cex.names=0.5, las = 2)
```



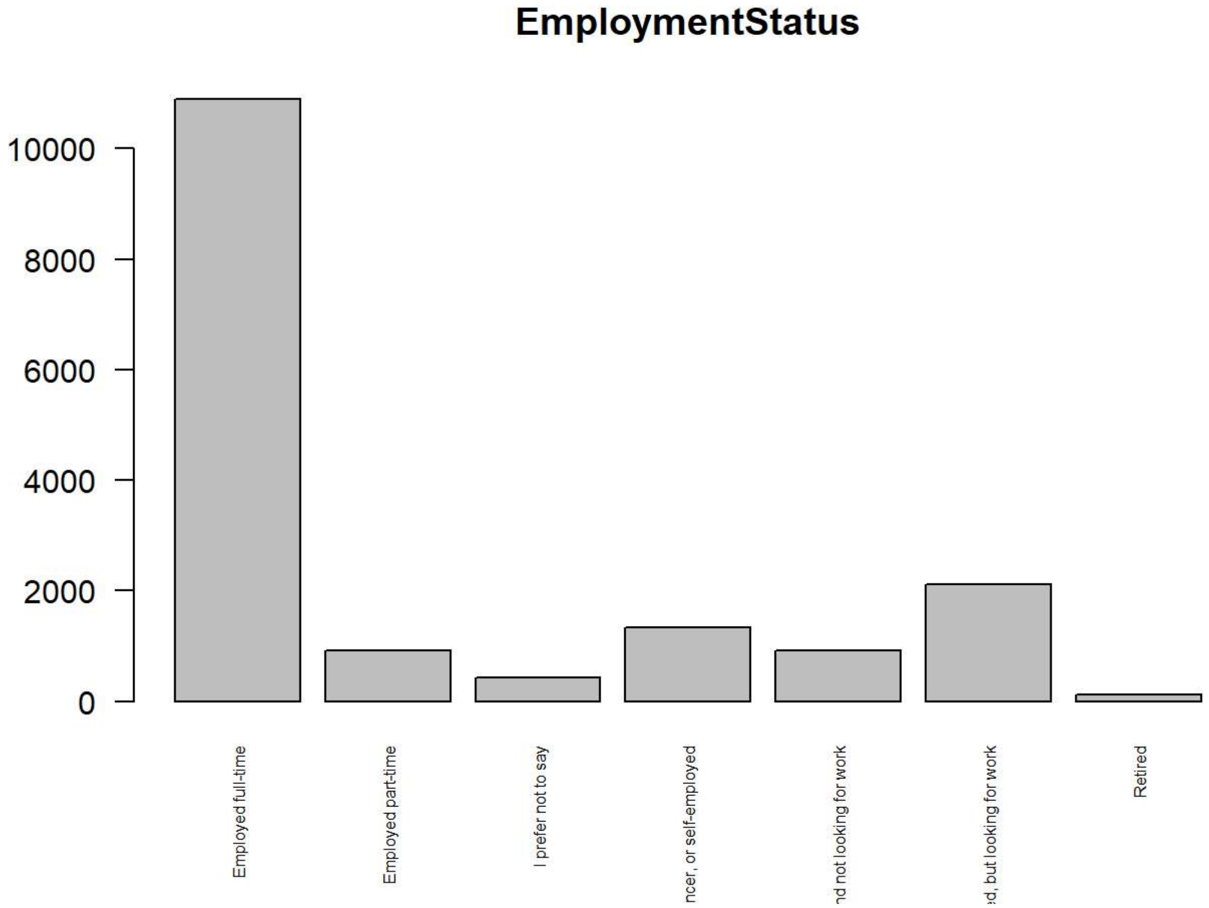
```
barplot((table(dataraw[,'Tenure'])), main = 'Tenure' ,horiz = F, cex.names=0.5, las = 2)
```



```
barplot((table(dataraw[,'DataScienceIdentitySelect'])), main = 'DataScienceIdentitySelect' ,hori
z = F, cex.names=0.5, las = 2)
```



```
barplot((table(dataraw[,'EmploymentStatus'])), main = 'EmploymentStatus',horiz = F, cex.names=0.
5, las = 2)
```



EmploymentStatus