181_Final Group Project

2023-10-20

Preparation

Please note that data cleaning and wrangling tasks were also executed using Excel prior to conducting the R analysis below. For more information on the Excel wrangling section, please refer back to our GitHub Repository: https://github.com/fmalzy/Insomnia/tree/main

```
# Set working directory
setwd("/Users/freyam/Desktop/Dartmouth/Courses/002 Fall Term/001_Core Courses/QBS 181 - Data Wrangling /

# Read in the data
Insomnia <- read.csv("insomnia_data_cleaned.csv")

# Dataset inspection
head(Insomnia)
```

```
##        ID Group SubGroup Remote Sex  Age American_Indian Asian Native_Hawaiian
## 1 sub_001     0        0      0   0 19.3               0     0               0
## 2 sub_002     0        0      0   0 19.3               0     0               0
## 3 sub_003     0        0      0   1 18.8               0     0               0
## 4 sub_004     0        0      0   0 18.8               0     0               0
## 5 sub_005     1        2      0   1 19.6               0     0               0
## 6 sub_006     0        0      0   0 19.1               0     0               0
##   Black White unknown_Race Hispanic NotHispanic unknown_Etnicity PDS_FEMALE
## 1     0     1            0        0           1                0         16
## 2     0     1            0        0           1                0         13
## 3     0     1            0        0           1                0         NA
## 4     0     1            0        0           1                0         17
## 5     0     1            0        0           1                0         NA
## 6     0     1            0        0           1                0         16
##   PDS_MALE ISI_total PSQI_total BDI_total ASHS_total ASHS_physiological
## 1       NA         0          0         0       5.43                5.8
## 2       NA         1          2         3       5.00                5.6
## 3       16         2          4         0       4.86                5.0
## 4       NA         1          5         2       4.39                4.4
## 5       14        10          7         5       4.14                4.4
## 6       NA         0          4         0       4.79                5.6
##   ASHS_cognitive ASHS_emotional ASHS_SleepEnvirnmont ASHS_DaytimeSleep
## 1           5.83           6.00                 5.75                 6
## 2           4.50           6.00                 5.75                 6
## 3           5.33           5.33                 5.75                 6
## 4           4.00           5.33                 5.00                 6
## 5           2.33           4.00                 5.75                 6
## 6           4.33           4.67                 6.00                 6
##   ASHS_substances ASHS_bedtimeRoutine ASHS_sleepStability ASHS_BedroomSharing
## 1             6.0                   1                5.25                3.5
## 2             5.5                   2                4.50                3.5
```

```
## 3            6.0                  1              3.50               3.5
## 4            5.5                  3              3.50               3.5
## 5            6.0                  3              4.25               3.5
## 6            6.0                  1              4.00               3.5
##   DBAS_total FIRST_total GCTI_total GCTI_anxiety GCTI_reflection GCTI_worries
## 1      28.75          16         30            9               7            4
## 2      11.81          15         31            9               6            5
## 3         NA          14         30            8               7            5
## 4      19.63          12         34            9               6            6
## 5      55.31          28         70           23              13           11
## 6      31.88          15         28            9               4            5
##   GCTI_thoughts GCTI_negativeAffect STAI_Y_total NEO_neuroticism
## 1             4                   3          43              12
## 2             4                   3          47               9
## 3             3                   4          50              19
## 4             5                   4          43              15
## 5             5                   8          52              15
## 6             3                   4          46              17
##   NEO_extraversion NEO_openness NEO_agreeableness NEO_Conscientiousness
## 1               30           25                18                    33
## 2               23           24                18                    30
## 3               27           29                21                    33
## 4               33           22                18                    27
## 5               28           28                33                    34
## 6               29           26                23                    27
##   MEQr_total PSRS_PrR PSRS_RWO PSRS_RSC PSRS_FRa PSRS_RSE PSRS_total PSS_total
## 1         58        7        1        6        5        4         23        23
## 2         57        6        1        4        5        4         20        18
## 3         35        6        4        4        5        4         23        19
## 4         54        6        2        4        5        4         21        16
## 5         52        4        6        5        3        7         25        25
## 6         60        7        3        6        6        6         28        18
##   TCQI_R_Total TCQIR_Aggressive_supression TCQIR_cognitive_distraction
## 1           68                          18                         15
## 2           68                          10                         14
## 3           63                           8                         12
## 4           60                          12                         11
## 5           71                          11                          9
## 6           63                          14                          9
##   TCQIR_reappraisal TCQIR_behavtioral_distraction TCQIR_social_avoidance
## 1                13                             9                      6
## 2                19                            11                      5
## 3                15                            15                      5
## 4                13                             9                      5
## 5                18                            11                      6
## 6                15                             9                      5
##   TCQIR_worry ACE_tot asq_home asq_school asq_attendance asq_romantic asq_peer
## 1           7       0       13         11              3            5        7
## 2           9       0       15          9              3            5        7
## 3           8       0       19         20              6            8        8
## 4          10       1       12          9              3            9        9
## 5          16       0       26         18              6           15       10
## 6          11       0       12         12              4            5        8
##   asq_teacher asq_future asq_leisure asq_finance asq_responsibility casq_total
```

```
## 1           7          6           6           4                  3         35
## 2           7          5           5           4                  5         35
## 3          10          6          10           7                  4         36
## 4           7          8           5           4                  3         37
## 5           8          7          17          10                  4         36
## 6           7          6          10           4                  3         40
##   casq_sleepy casq_alert cope_disengage_su cope_growth cope_disengage_mental
## 1          15         10                 4          16                    7
## 2          14          9                 4          15                   10
## 3          16         10                 4          14                    9
## 4          15          8                 4          11                    7
## 5          23         17                 8          10                    8
## 6          19          9                 4          16                   10
##   cope_emotions cope_socialsupp_instr cope_active cope_denial cope_religion
## 1             5                    12          13           4             6
## 2             5                    12          12           5             5
## 3            10                    14           7           4             4
## 4            13                     9          11           4             4
## 5             5                     5           8           4             4
## 6             9                    11          12           4             7
##   cope_humor cope_disengage_emo cope_restraint cope_socialsupp_emo cope_acccept
## 1          9                  4             14                  14           9
## 2          5                  4              8                   8          15
## 3          9                  6              6                  16          10
## 4          8                  4              7                  13           6
## 5         11                  4              6                   9           5
## 6          6                  5             10                  13           8
##   cope_suppression cope_planning ders_nonaccpetance ders_total ders_goals
## 1               13            13                  6         87         20
## 2                9            14                  8         81         12
## 3               12             9                  9         70          8
## 4                5             9                  8         82         14
## 5                7             8                 12         91         18
## 6                6            12                 14         85          9
##   ders_impulse ders_awareness ders_strategies ders_clarity ZISI_total
## 1           10             26              12           13  -1.413051
## 2           10             26              12           13  -1.209039
## 3            9             21              12           11  -1.005027
## 4           10             23              14           13  -1.209039
## 5           11             20              15           15   0.627068
## 6           10             26              14           12  -1.413051
##   ZPSQI_total  ZBDI_total ZASHS_total ZASHS_physiological ZASHS_cognitive
## 1  -1.9136582 -1.01514988   2.2939259           1.0353556       2.2335101
## 2  -1.2378309 -0.38619832   1.2566092           0.7162624       0.8512074
## 3  -0.5620037 -1.01514988   0.9108369          -0.2410172       1.7151466
## 4  -0.2240901 -0.59584884  -0.2129229          -1.1982968       0.3328439
## 5   0.4517372  0.03310271  -0.8180244          -1.1982968      -1.3950344
## 6  -0.5620037 -1.01514988   0.7379508           0.7162624       0.6784196
##   ZASHS_emotional ZASHS_SleepEnvirnmont ZASHS_DaytimeSleep ZASHS_substances
## 1       1.4312285             0.6791653          0.9153677        0.3620167
## 2       1.4312285             0.6791653          0.9153677       -1.3575628
## 3       0.5679478             0.6791653          0.9153677        0.3620167
## 4       0.5679478            -0.6199092          0.9153677       -1.3575628
## 5      -1.1586135             0.6791653          0.9153677        0.3620167
```

```
## 6      -0.2953329             1.1121901          0.9153677        0.3620167
##   ZASHS_bedtimeRoutine ZASHS_sleepStability ZASHS_BedroomSharing ZDBAS_total
## 1           -0.9804062            1.8739264          -0.01211277  -0.6476796
## 2           -0.2527610            0.8662490          -0.01211277  -1.9098422
## 3           -0.9804062           -0.4773209          -0.01211277          NA
## 4            0.4748843           -0.4773209          -0.01211277  -1.3276638
## 5            0.4748843            0.5303565          -0.01211277   1.3317267
## 6           -0.9804062            0.1944641          -0.01211277  -0.4148082
##   ZFIRST_total ZGCTI_total ZGCTI_anxiety ZGCTI_reflection ZGCTI_worries
## 1   -0.5882030  -1.1482108    -0.9133689       -0.7160901    -1.3265786
## 2   -0.7661625  -1.0727229    -0.9133689       -1.0722606    -0.9654756
## 3   -0.9441220  -1.1482108    -1.1170544       -0.7160901    -0.9654756
## 4   -1.3000410  -0.8462591    -0.9133689       -1.0722606    -0.6043725
## 5    1.5473111   1.8713054     1.9382287        1.4209327     1.2011428
## 6   -0.7661625  -1.2991866    -0.9133689       -1.7846015    -0.9654756
##   ZGCTI_thoughts ZGCTI_negativeAffect ZSTAI_Y_total ZNEO_neuroticism
## 1     -0.4465522           -1.2091717    0.07112872       -1.2043743
## 2     -0.4465522           -1.2091717    0.68542222       -1.7652349
## 3     -0.9286255           -0.6348152    1.14614234        0.1043004
## 4      0.0355212           -0.6348152    0.07112872       -0.6435137
## 5      0.0355212            1.6626112    1.45328908       -0.6435137
## 6     -0.9286255           -0.6348152    0.53184884       -0.2696067
##   ZNEO_extraversion ZNEO_openness ZNEO_agreeableness ZNEO_Conscientiousness
## 1        0.26828373   -0.21630530         -1.3729993              0.55020625
## 2       -1.25657677   -0.43491172         -1.3729993              0.00383419
## 3       -0.38522791    0.65812037         -0.9093680              0.55020625
## 4        0.92179537   -0.87212455         -1.3729993             -0.54253787
## 5       -0.16739070    0.43951396          0.9451571              0.73233027
## 6        0.05044651    0.00230112         -0.6002805             -0.54253787
##   ZMEQr_total  ZPSRS_PrR    ZPSRS_RWO    ZPSRS_RSC  ZPSRS_FRa  ZPSRS_RSE
## 1   1.0699790  1.5076986 -1.299595985  0.728257383  0.3576728 -0.2663142
## 2   0.9403259  0.9084042 -1.299595985 -0.713085354  0.3576728 -0.2663142
## 3  -1.9120416  0.9084042 -0.009056418 -0.713085354  0.3576728 -0.2663142
## 4   0.5513667  0.9084042 -0.869416129 -0.713085354  0.3576728 -0.2663142
## 5   0.2920606 -0.2901847  0.851303293  0.007586014 -1.1868233  1.4987918
## 6   1.3292851  1.5076986 -0.439236274  0.728257383  1.1299208  0.9104231
##   ZPSRS_total  ZPSS_total ZTCQI_R_Total ZTCQIR_Aggressive_supression
## 1   0.1095509  1.12300069    -0.0285142                    2.3876285
## 2  -0.5276330 -0.06238893    -0.0285142                   -1.0819655
## 3   0.1095509  0.17468900    -0.5702839                   -1.9493640
## 4  -0.3152384 -0.53654477    -0.8953457                   -0.2145670
## 5   0.5343402  1.59715653     0.2965476                   -0.6482662
## 6   1.1715242 -0.06238893    -0.5702839                    0.6528315
##   ZTCQIR_cognitive_distraction ZTCQIR_reappraisal
## 1                   1.46037913        -0.63390345
## 2                   1.08028045         1.00848276
## 3                   0.32008310        -0.08644138
## 4                  -0.06001558        -0.63390345
## 5                  -0.82021293         0.73475172
## 6                  -0.82021293        -0.08644138
##   ZTCQIR_behavtioral_distraction ZTCQIR_social_avoidance ZTCQIR_worry
## 1                     -0.4937992              -0.5867699   -1.5028207
## 2                      0.2063639              -1.2130974   -0.8952974
## 3                      1.6066901              -1.2130974   -1.1990591
```

```
## 4                    -0.4937992             -1.2130974   -0.5915358
## 5                     0.2063639             -0.5867699    1.2310340
## 6                    -0.4937992             -1.2130974   -0.2877742
##     ZACE_tot  Zasq_home Zasq_school Zasq_attendance Zasq_romantic  Zasq_peer
## 1 -0.5951622 -1.0682846  -0.9420565      -1.2664608    -0.7703171 -1.0880463
## 2 -0.5951622 -0.8183165  -1.2957932      -1.2664608    -0.7703171 -1.0880463
## 3 -0.5951622 -0.3183804   0.6497584       0.3448844     0.1043457 -0.8131410
## 4  0.5827630 -1.1932686  -1.2957932      -1.2664608     0.3959000 -0.5382357
## 5 -0.5951622  0.5565079   0.2960217       0.3448844     2.1452258 -0.2633304
## 6 -0.5951622 -1.1932686  -0.7651882      -0.7293457    -0.7703171 -0.8131410
##   Zasq_teacher Zasq_future Zasq_leisure Zasq_finance Zasq_responsibility
## 1  -0.77442798  -0.8704381   -1.3175245   -1.0024147          -0.7828529
## 2  -0.77442798  -1.2079549   -1.6026380   -1.0024147           0.2154159
## 3   0.08772817  -0.8704381   -0.1770705    0.0289506          -0.2837185
## 4  -0.77442798  -0.1954045   -1.6026380   -1.0024147          -0.7828529
## 5  -0.48704260  -0.5329213    1.8187241    1.0603159          -0.2837185
## 6  -0.77442798  -0.8704381   -0.1770705   -1.0024147          -0.7828529
##   Zcasq_total Zcasq_sleepy Zcasq_alert Zcope_growth Zcope_disengage_mental
## 1 -0.23806237   -0.9618201  -0.8241381    1.6161790             -0.8314592
## 2 -0.23806237   -1.1490596  -1.0504188    1.2284593              0.6588922
## 3 -0.07879529   -0.7745805  -0.8241381    0.8407396              0.1621084
## 4  0.08047179   -0.9618201  -1.2766995   -0.3224196             -0.8314592
## 5 -0.07879529    0.5360964   0.7598268   -0.7101393             -0.3346754
## 6  0.55827303   -0.2128618  -1.0504188    1.6161790              0.6588922
##   Zcope_emotions Zcope_socialsupp_instr Zcope_active Zcope_denial
## 1     -0.9880625              0.5479410    1.1102859   -0.5410135
## 2     -0.9880625              0.5479410    0.6797669    0.3155912
## 3      0.6415545              1.1751024   -1.4728282   -0.5410135
## 4      1.6193247             -0.3928011    0.2492478   -0.5410135
## 5     -0.9880625             -1.6471239   -1.0423092   -0.5410135
## 6      0.3156311              0.2343603    0.6797669   -0.5410135
##   Zcope_religion  Zcope_humor Zcope_disengage_emo Zcope_restraint
## 1      0.1579507  0.002954759          -0.7835810       2.4245103
## 2     -0.2268011 -1.119853536          -0.7835810      -0.3505316
## 3     -0.6115529  0.002954759           0.3192367      -1.2755456
## 4     -0.6115529 -0.277747315          -0.7835810      -0.8130386
## 5     -0.6115529  0.564358906          -0.7835810      -1.2755456
## 6      0.5427025 -0.839151462          -0.2321722       0.5744824
##   Zcope_socialsupp_emo Zcope_disengage_su Zcope_acccept Zcope_suppression
## 1            1.0128177         -0.2393677    -0.7015620         2.2153540
## 2           -0.5731880         -0.2393677     1.5450129         0.3278327
## 3            1.5414863         -0.2393677    -0.3271328         1.7434737
## 4            0.7484834         -0.2393677    -1.8248494        -1.5596887
## 5           -0.3088538          3.2590839    -2.1992786        -0.6159280
## 6            0.7484834         -0.2393677    -1.0759911        -1.0878084
##   Zcope_planning Zders_nonaccpetance Zders_goals Zders_impulse Zders_awareness
## 1      0.7401369          -1.1228046   1.9057268    -0.1145649       1.0830870
## 2      1.0781803          -0.5523959  -0.3021274    -0.1145649       1.0830870
## 3     -0.6120363          -0.2671915  -1.4060546    -0.4255269       0.2716259
## 4     -0.6120363          -0.5523959   0.2498361    -0.1145649       0.5962104
## 5     -0.9500796           0.5884217   1.3537633     0.1963970       0.1093337
## 6      0.4020936           1.1588305  -1.1300728    -0.1145649       1.0830870
##   Zders_strategies Zders_clarity ZDERS_total
## 1       -0.6560514     0.5380155   0.5758063
```

```
## 2       -0.6560514     0.5380155   0.1539431
## 3       -0.6560514    -0.2606013  -0.6194729
## 4       -0.1164420     0.5380155   0.2242536
## 5        0.1533627     1.3366323   0.8570485
## 6       -0.1164420     0.1387071   0.4351853
```

```
colnames(Insomnia)
```

```
##   [1] "ID"                                "Group"
##   [3] "SubGroup"                          "Remote"
##   [5] "Sex"                               "Age"
##   [7] "American_Indian"                   "Asian"
##   [9] "Native_Hawaiian"                   "Black"
##  [11] "White"                             "unknown_Race"
##  [13] "Hispanic"                          "NotHispanic"
##  [15] "unknown_Etnicity"                  "PDS_FEMALE"
##  [17] "PDS_MALE"                          "ISI_total"
##  [19] "PSQI_total"                        "BDI_total"
##  [21] "ASHS_total"                        "ASHS_physiological"
##  [23] "ASHS_cognitive"                    "ASHS_emotional"
##  [25] "ASHS_SleepEnvirnmont"              "ASHS_DaytimeSleep"
##  [27] "ASHS_substances"                   "ASHS_bedtimeRoutine"
##  [29] "ASHS_sleepStability"               "ASHS_BedroomSharing"
##  [31] "DBAS_total"                        "FIRST_total"
##  [33] "GCTI_total"                        "GCTI_anxiety"
##  [35] "GCTI_reflection"                   "GCTI_worries"
##  [37] "GCTI_thoughts"                     "GCTI_negativeAffect"
##  [39] "STAI_Y_total"                      "NEO_neuroticism"
##  [41] "NEO_extraversion"                  "NEO_openness"
##  [43] "NEO_agreeableness"                 "NEO_Conscientiousness"
##  [45] "MEQr_total"                        "PSRS_PrR"
##  [47] "PSRS_RWO"                          "PSRS_RSC"
##  [49] "PSRS_FRa"                          "PSRS_RSE"
##  [51] "PSRS_total"                        "PSS_total"
##  [53] "TCQI_R_Total"                      "TCQIR_Aggressive_supression"
##  [55] "TCQIR_cognitive_distraction"       "TCQIR_reappraisal"
##  [57] "TCQIR_behavtioral_distraction"     "TCQIR_social_avoidance"
##  [59] "TCQIR_worry"                       "ACE_tot"
##  [61] "asq_home"                          "asq_school"
##  [63] "asq_attendance"                    "asq_romantic"
##  [65] "asq_peer"                          "asq_teacher"
##  [67] "asq_future"                        "asq_leisure"
##  [69] "asq_finance"                       "asq_responsibility"
##  [71] "casq_total"                        "casq_sleepy"
##  [73] "casq_alert"                        "cope_disengage_su"
##  [75] "cope_growth"                       "cope_disengage_mental"
##  [77] "cope_emotions"                     "cope_socialsupp_instr"
##  [79] "cope_active"                       "cope_denial"
##  [81] "cope_religion"                     "cope_humor"
##  [83] "cope_disengage_emo"                "cope_restraint"
##  [85] "cope_socialsupp_emo"               "cope_acccept"
##  [87] "cope_suppression"                  "cope_planning"
##  [89] "ders_nonaccpetance"                "ders_total"
##  [91] "ders_goals"                        "ders_impulse"
##  [93] "ders_awareness"                    "ders_strategies"
```

```
##  [95] "ders_clarity"                    "ZISI_total"
##  [97] "ZPSQI_total"                     "ZBDI_total"
##  [99] "ZASHS_total"                     "ZASHS_physiological"
## [101] "ZASHS_cognitive"                 "ZASHS_emotional"
## [103] "ZASHS_SleepEnvirnmont"           "ZASHS_DaytimeSleep"
## [105] "ZASHS_substances"                "ZASHS_bedtimeRoutine"
## [107] "ZASHS_sleepStability"            "ZASHS_BedroomSharing"
## [109] "ZDBAS_total"                     "ZFIRST_total"
## [111] "ZGCTI_total"                     "ZGCTI_anxiety"
## [113] "ZGCTI_reflection"                "ZGCTI_worries"
## [115] "ZGCTI_thoughts"                  "ZGCTI_negativeAffect"
## [117] "ZSTAI_Y_total"                   "ZNEO_neuroticism"
## [119] "ZNEO_extraversion"               "ZNEO_openness"
## [121] "ZNEO_agreeableness"              "ZNEO_Conscientiousness"
## [123] "ZMEQr_total"                     "ZPSRS_PrR"
## [125] "ZPSRS_RWO"                       "ZPSRS_RSC"
## [127] "ZPSRS_FRa"                       "ZPSRS_RSE"
## [129] "ZPSRS_total"                     "ZPSS_total"
## [131] "ZTCQI_R_Total"                   "ZTCQIR_Aggressive_supression"
## [133] "ZTCQIR_cognitive_distraction"    "ZTCQIR_reappraisal"
## [135] "ZTCQIR_behavtioral_distraction" "ZTCQIR_social_avoidance"
## [137] "ZTCQIR_worry"                    "ZACE_tot"
## [139] "Zasq_home"                       "Zasq_school"
## [141] "Zasq_attendance"                 "Zasq_romantic"
## [143] "Zasq_peer"                       "Zasq_teacher"
## [145] "Zasq_future"                     "Zasq_leisure"
## [147] "Zasq_finance"                    "Zasq_responsibility"
## [149] "Zcasq_total"                     "Zcasq_sleepy"
## [151] "Zcasq_alert"                     "Zcope_growth"
## [153] "Zcope_disengage_mental"          "Zcope_emotions"
## [155] "Zcope_socialsupp_instr"          "Zcope_active"
## [157] "Zcope_denial"                    "Zcope_religion"
## [159] "Zcope_humor"                     "Zcope_disengage_emo"
## [161] "Zcope_restraint"                 "Zcope_socialsupp_emo"
## [163] "Zcope_disengage_su"              "Zcope_acccept"
## [165] "Zcope_suppression"               "Zcope_planning"
## [167] "Zders_nonaccpetance"             "Zders_goals"
## [169] "Zders_impulse"                   "Zders_awareness"
## [171] "Zders_strategies"                "Zders_clarity"
## [173] "ZDERS_total"
```

Data Cleaning

```
library(mice)
```

```
##
## Attaching package: 'mice'
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(tidyr)

# Impute missing data using the mice package
mice_mod <- mice(Insomnia, m = 1, method = 'pmm', maxit = 5)
```

```
##
##  iter imp variable
##   1   1  Group  SubGroup*  Remote*  Sex*  Age*  American_Indian  Asian*  Native_Hawaiian*  Black*  Wh
##   2   1  Group  SubGroup  Remote  Sex  Age  American_Indian*  Asian  Native_Hawaiian  Black  White  I
##   3   1  Group*  SubGroup*  Remote  Sex  Age  American_Indian  Asian*  Native_Hawaiian  Black*  Whit
##   4   1  Group  SubGroup  Remote  Sex  Age  American_Indian  Asian  Native_Hawaiian  Black  White*  I
##   5   1  Group  SubGroup*  Remote  Sex  Age  American_Indian  Asian  Native_Hawaiian  Black*  White*
```

```
## Warning: Number of logged events: 634
```

```r
Insomnia_clean <- complete(mice_mod)

# Remove rows with any missing value if still present
Insomnia_clean <- Insomnia_clean %>%
  drop_na()

# Save the cleaned dataset
write.csv(Insomnia_clean, "Insomnia_final.csv", row.names = FALSE)
```

```r
# Read in the final dataset again
Insomnia_Final <- read.csv("Insomnia_final.csv")

# Dataset inspection
head(Insomnia_Final)
```

```
##        ID Group SubGroup Remote Sex  Age American_Indian Asian Native_Hawaiian
## 1 sub_001     0        0      0   0 19.3               0     0               0
## 2 sub_002     0        0      0   0 19.3               0     0               0
## 3 sub_004     0        0      0   0 18.8               0     0               0
## 4 sub_005     1        2      0   1 19.6               0     0               0
## 5 sub_006     0        0      0   0 19.1               0     0               0
## 6 sub_007     0        0      0   0 19.0               0     0               0
##   Black White unknown_Race Hispanic NotHispanic unknown_Etnicity PDS_FEMALE
## 1     0     1            0        0           1                0         16
## 2     0     1            0        0           1                0         13
## 3     0     1            0        0           1                0         17
## 4     0     1            0        0           1                0         13
## 5     0     1            0        0           1                0         16
## 6     0     1            0        1           0                0         16
```

```
##   PDS_MALE ISI_total PSQI_total BDI_total ASHS_total ASHS_physiological
## 1       19         0          0         0       5.43                5.8
## 2       20         1          2         3       5.00                5.6
## 3       12         1          5         2       4.39                4.4
## 4       14        10          7         5       4.14                4.4
## 5       13         0          4         0       4.79                5.6
## 6       16         0          1         3       4.21                3.6
##   ASHS_cognitive ASHS_emotional ASHS_SleepEnvirnmont ASHS_DaytimeSleep
## 1           5.83           6.00                 5.75                 6
## 2           4.50           6.00                 5.75                 6
## 3           4.00           5.33                 5.00                 6
## 4           2.33           4.00                 5.75                 6
## 5           4.33           4.67                 6.00                 6
## 6           3.33           4.33                 5.75                 5
##   ASHS_substances ASHS_bedtimeRoutine ASHS_sleepStability ASHS_BedroomSharing
## 1             6.0                   1                5.25                 3.5
## 2             5.5                   2                4.50                 3.5
## 3             5.5                   3                3.50                 3.5
## 4             6.0                   3                4.25                 3.5
## 5             6.0                   1                4.00                 3.5
## 6             5.5                   3                4.50                 3.5
##   DBAS_total FIRST_total GCTI_total GCTI_anxiety GCTI_reflection GCTI_worries
## 1      28.75          16         30            9                7            4
## 2      11.81          15         31            9                6            5
## 3      19.63          12         34            9                6            6
## 4      55.31          28         70           23               13           11
## 5      31.88          15         28            9                4            5
## 6      65.44          25         45           14               10            8
##   GCTI_thoughts GCTI_negativeAffect STAI_Y_total NEO_neuroticism
## 1             4                   3          43              12
## 2             4                   3          47               9
## 3             5                   4          43              15
## 4             5                   8          52              15
## 5             3                   4          46              17
## 6             3                   5          46              18
##   NEO_extraversion NEO_openness NEO_agreeableness NEO_Conscientiousness
## 1               30           25                18                    33
## 2               23           24                18                    30
## 3               33           22                18                    27
## 4               28           28                33                    34
## 5               29           26                23                    27
## 6               33           23                21                    28
##   MEQr_total PSRS_PrR PSRS_RWO PSRS_RSC PSRS_FRa PSRS_RSE PSRS_total PSS_total
## 1         58        7        1        6        5        4         23        23
## 2         57        6        1        4        5        4         20        18
## 3         54        6        2        4        5        4         21        16
## 4         52        4        6        5        3        7         25        25
## 5         60        7        3        6        6        6         28        18
## 6         61        3        4        5        5        3         20        17
##   TCQI_R_Total TCQIR_Aggressive_supression TCQIR_cognitive_distraction
## 1           68                         18                         15
## 2           68                         10                         14
## 3           60                         12                         11
## 4           71                         11                          9
```

```
## 5           63                         14                          9
## 6           71                         15                          7
##   TCQIR_reappraisal TCQIR_behavtioral_distraction TCQIR_social_avoidance
## 1                13                            9                      6
## 2                19                           11                      5
## 3                13                            9                      5
## 4                18                           11                      6
## 5                15                            9                      5
## 6                17                           10                      7
##   TCQIR_worry ACE_tot asq_home asq_school asq_attendance asq_romantic asq_peer
## 1           7       0       13         11              3            5        7
## 2           9       0       15          9              3            5        7
## 3          10       1       12          9              3            9        9
## 4          16       0       26         18              6           15       10
## 5          11       0       12         12              4            5        8
## 6          15       1       29         15              4           16       17
##   asq_teacher asq_future asq_leisure asq_finance asq_responsibility casq_total
## 1           7          6           6           4                  3         35
## 2           7          5           5           4                  5         35
## 3           7          8           5           4                  3         37
## 4           8          7          17          10                  4         36
## 5           7          6          10           4                  3         40
## 6          16          9          16          11                 10         38
##   casq_sleepy casq_alert cope_disengage_su cope_growth cope_disengage_mental
## 1          15         10                 4          16                     7
## 2          14          9                 4          15                    10
## 3          15          8                 4          11                     7
## 4          23         17                 8          10                     8
## 5          19          9                 4          16                    10
## 6          22         14                 4          14                    12
##   cope_emotions cope_socialsupp_instr cope_active cope_denial cope_religion
## 1             5                    12          13           4             6
## 2             5                    12          12           5             5
## 3            13                     9          11           4             4
## 4             5                     5           8           4             4
## 5             9                    11          12           4             7
## 6             9                     8           9           4             4
##   cope_humor cope_disengage_emo cope_restraint cope_socialsupp_emo cope_acccept
## 1          9                  4            14                  14           9
## 2          5                  4             8                   8          15
## 3          8                  4             7                  13           6
## 4         11                  4             6                   9           5
## 5          6                  5            10                  13           8
## 6         12                  8             9                   9          14
##   cope_suppression cope_planning ders_nonaccpetance ders_total ders_goals
## 1               13            13                  6         87         20
## 2                9            14                  8         81         12
## 3                5             9                  8         82         14
## 4                7             8                 12         91         18
## 5                6            12                 14         85          9
## 6               10            10                 18        100         17
##   ders_impulse ders_awareness ders_strategies ders_clarity ZISI_total
## 1           10             26              12           13  -1.413051
## 2           10             26              12           13  -1.209039
```

```
## 3           10             23             14           13  -1.209039
## 4           11             20             15           15   0.627068
## 5           10             26             14           12  -1.413051
## 6            9             23             19           14  -1.413051
##   ZPSQI_total  ZBDI_total ZASHS_total ZASHS_physiological ZASHS_cognitive
## 1  -1.9136582 -1.01514988   2.2939259           1.0353556       2.2335101
## 2  -1.2378309 -0.38619832   1.2566092           0.7162624       0.8512074
## 3  -0.2240901 -0.59584884  -0.2129229          -1.1982968       0.3328439
## 4   0.4517372  0.03310271  -0.8180244          -1.1982968      -1.3950344
## 5  -0.5620037 -1.01514988   0.7379508           0.7162624       0.6784196
## 6  -1.5757445 -0.38619832  -0.6451382          -2.4746696      -0.3583074
##   ZASHS_emotional ZASHS_SleepEnvirnmont ZASHS_DaytimeSleep ZASHS_substances
## 1       1.4312285             0.6791653         0.91536769        0.3620167
## 2       1.4312285             0.6791653         0.91536769       -1.3575628
## 3       0.5679478            -0.6199092         0.91536769       -1.3575628
## 4      -1.1586135             0.6791653         0.91536769        0.3620167
## 5      -0.2953329             1.1121901         0.91536769        0.3620167
## 6      -0.7269732             0.6791653         0.03698455       -1.3575628
##   ZASHS_bedtimeRoutine ZASHS_sleepStability ZASHS_BedroomSharing ZDBAS_total
## 1           -0.9804062            1.8739264          -0.01211277  -0.6476796
## 2           -0.2527610            0.8662490          -0.01211277  -1.9098422
## 3            0.4748843           -0.4773209          -0.01211277  -1.3276638
## 4            0.4748843            0.5303565          -0.01211277   1.3317267
## 5           -0.9804062            0.1944641          -0.01211277  -0.4148082
## 6            0.4748843            0.8662490          -0.01211277   2.0862298
##   ZFIRST_total ZGCTI_total ZGCTI_anxiety ZGCTI_reflection ZGCTI_worries
## 1   -0.5882030 -1.14821077    -0.9133689       -0.7160901    -1.3265786
## 2   -0.7661625 -1.07272286    -0.9133689       -1.0722606    -0.9654756
## 3   -1.3000410 -0.84625915    -0.9133689       -1.0722606    -0.6043725
## 4    1.5473111  1.87130544     1.9382287        1.4209327     1.2011428
## 5   -0.7661625 -1.29918658    -0.9133689       -1.7846015    -0.9654756
## 6    1.0134326 -0.01589219     0.1050589        0.3524213     0.1178336
##   ZGCTI_thoughts ZGCTI_negativeAffect ZSTAI_Y_total ZNEO_neuroticism
## 1     -0.4465522          -1.20917175    0.07112872      -1.20437434
## 2     -0.4465522          -1.20917175    0.68542222      -1.76523494
## 3      0.0355212          -0.63481517    0.07112872      -0.64351374
## 4      0.0355212           1.66261116    1.45328908      -0.64351374
## 5     -0.9286255          -0.63481517    0.53184884      -0.26960667
## 6     -0.9286255          -0.06045859    0.53184884      -0.08265314
##   ZNEO_extraversion ZNEO_openness ZNEO_agreeableness ZNEO_Conscientiousness
## 1        0.26828373   -0.21630530         -1.3729993             0.55020625
## 2       -1.25657677   -0.43491172         -1.3729993             0.00383419
## 3        0.92179537   -0.87212455         -1.3729993            -0.54253787
## 4       -0.16739070    0.43951396          0.9451571             0.73233027
## 5        0.05044651    0.00230112         -0.6002805            -0.54253787
## 6        0.92179537   -0.65351814         -0.9093680            -0.36041385
##   ZMEQr_total  ZPSRS_PrR    ZPSRS_RWO    ZPSRS_RSC  ZPSRS_FRa  ZPSRS_RSE
## 1   1.0699790  1.5076986 -1.299595985  0.728257383  0.3576728 -0.2663142
## 2   0.9403259  0.9084042 -1.299595985 -0.713085354  0.3576728 -0.2663142
## 3   0.5513667  0.9084042 -0.869416129 -0.713085354  0.3576728 -0.2663142
## 4   0.2920606 -0.2901847  0.851303293  0.007586014 -1.1868233  1.4987918
## 5   1.3292851  1.5076986 -0.439236274  0.728257383  1.1299208  0.9104231
## 6   1.4589382 -0.8894791 -0.009056418  0.007586014  0.3576728 -0.8546829
##   ZPSRS_total  ZPSS_total ZTCQI_R_Total ZTCQIR_Aggressive_supression
```

```
## 1   0.1095509  1.12300069    -0.0285142                   2.3876285
## 2  -0.5276330 -0.06238893    -0.0285142                  -1.0819655
## 3  -0.3152384 -0.53654477    -0.8953457                  -0.2145670
## 4   0.5343402  1.59715653     0.2965476                  -0.6482662
## 5   1.1715242 -0.06238893    -0.5702839                   0.6528315
## 6  -0.5276330 -0.29946685     0.2965476                   1.0865307
##   ZTCQIR_cognitive_distraction ZTCQIR_reappraisal
## 1                   1.46037913        -0.63390345
## 2                   1.08028045         1.00848276
## 3                  -0.06001558        -0.63390345
## 4                  -0.82021293         0.73475172
## 5                  -0.82021293        -0.08644138
## 6                  -1.58041029         0.46102069
##   ZTCQIR_behavtioral_distraction ZTCQIR_social_avoidance ZTCQIR_worry
## 1                     -0.4937992             -0.58676992   -1.5028207
## 2                      0.2063639             -1.21309736   -0.8952974
## 3                     -0.4937992             -1.21309736   -0.5915358
## 4                      0.2063639             -0.58676992    1.2310340
## 5                     -0.4937992             -1.21309736   -0.2877742
## 6                     -0.1437177              0.03955752    0.9272724
##     ZACE_tot  Zasq_home Zasq_school Zasq_attendance Zasq_romantic  Zasq_peer
## 1 -0.5951622 -1.0682846  -0.9420565      -1.2664608    -0.7703171 -1.0880463
## 2 -0.5951622 -0.8183165  -1.2957932      -1.2664608    -0.7703171 -1.0880463
## 3  0.5827630 -1.1932686  -1.2957932      -1.2664608     0.3959000 -0.5382357
## 4 -0.5951622  0.5565079   0.2960217       0.3448844     2.1452258 -0.2633304
## 5 -0.5951622 -1.1932686  -0.7651882      -0.7293457    -0.7703171 -0.8131410
## 6  0.5827630  0.9314600  -0.2345832      -0.7293457     2.4367800  1.6610069
##   Zasq_teacher Zasq_future Zasq_leisure Zasq_finance Zasq_responsibility
## 1   -0.7744280  -0.8704381   -1.3175245    -1.002415          -0.7828529
## 2   -0.7744280  -1.2079549   -1.6026380    -1.002415           0.2154159
## 3   -0.7744280  -0.1954045   -1.6026380    -1.002415          -0.7828529
## 4   -0.4870426  -0.5329213    1.8187241     1.060316          -0.2837185
## 5   -0.7744280  -0.8704381   -0.1770705    -1.002415          -0.7828529
## 6    1.8120405   0.1421123    1.5336105     1.404104           2.7110879
##   Zcasq_total Zcasq_sleepy Zcasq_alert Zcope_growth Zcope_disengage_mental
## 1 -0.23806237   -0.9618201 -0.82413812    1.6161790             -0.8314592
## 2 -0.23806237   -1.1490596 -1.05041882    1.2284593              0.6588922
## 3  0.08047179   -0.9618201 -1.27669951   -0.3224196             -0.8314592
## 4 -0.07879529    0.5360964  0.75982676   -0.7101393             -0.3346754
## 5  0.55827303   -0.2128618 -1.05041882    1.6161790              0.6588922
## 6  0.23973887    0.3488569  0.08098467    0.8407396              1.6524599
##   Zcope_emotions Zcope_socialsupp_instr Zcope_active Zcope_denial
## 1     -0.9880625              0.5479410    1.1102859   -0.5410135
## 2     -0.9880625              0.5479410    0.6797669    0.3155912
## 3      1.6193247             -0.3928011    0.2492478   -0.5410135
## 4     -0.9880625             -1.6471239   -1.0423092   -0.5410135
## 5      0.3156311              0.2343603    0.6797669   -0.5410135
## 6      0.3156311             -0.7063818   -0.6117902   -0.5410135
##   Zcope_religion  Zcope_humor Zcope_disengage_emo Zcope_restraint
## 1      0.1579507  0.002954759          -0.7835810       2.4245103
## 2     -0.2268011 -1.119853536          -0.7835810      -0.3505316
## 3     -0.6115529 -0.277747315          -0.7835810      -0.8130386
## 4     -0.6115529  0.564358906          -0.7835810      -1.2755456
## 5      0.5427025 -0.839151462          -0.2321722       0.5744824
```

```
## 6     -0.6115529  0.845060980           1.4220545       0.1119754
##   Zcope_socialsupp_emo Zcope_disengage_su Zcope_acccept Zcope_suppression
## 1            1.0128177         -0.2393677     -0.701562         2.2153540
## 2           -0.5731880         -0.2393677      1.545013         0.3278327
## 3            0.7484834         -0.2393677     -1.824849        -1.5596887
## 4           -0.3088538          3.2590839     -2.199279        -0.6159280
## 5            0.7484834         -0.2393677     -1.075991        -1.0878084
## 6           -0.3088538         -0.2393677      1.170584         0.7997130
##   Zcope_planning Zders_nonaccpetance Zders_goals Zders_impulse Zders_awareness
## 1      0.7401369          -1.1228046   1.9057268    -0.1145649       1.0830870
## 2      1.0781803          -0.5523959  -0.3021274    -0.1145649       1.0830870
## 3     -0.6120363          -0.5523959   0.2498361    -0.1145649       0.5962104
## 4     -0.9500796           0.5884217   1.3537633     0.1963970       0.1093337
## 5      0.4020936           1.1588305  -1.1300728    -0.1145649       1.0830870
## 6     -0.2739930           2.2996480   1.0777815    -0.4255269       0.5962104
##   Zders_strategies Zders_clarity ZDERS_total
## 1       -0.6560514     0.5380155   0.5758063
## 2       -0.6560514     0.5380155   0.1539431
## 3       -0.1164420     0.5380155   0.2242536
## 4        0.1533627     1.3366323   0.8570485
## 5       -0.1164420     0.1387071   0.4351853
## 6        1.2325813     0.9373239   1.4898434
```

Preliminary Step No.1: Compare Age and Race by subgroups

We first delved into our dataset by examining its structure, focusing on comparing Age and Race across various subgroups.

There are 3 different sub-groups in this dataset: *0 = Control*: Individuals who do not have insomnia.

1 = Clean Insomnia: Individuals who meet all the criteria for a diagnosis of insomnia without having other complciating factors or conditions (e.g., individuals who have sleep issues that are not caused by another mental health condition, medication, or a medical problem)

2 = Sub-clinical Insomnia: Individuals who have symptoms of insomnia, but these symptoms do not meet the full diagnostic criteria for clinical insomnia. (e.g., these symptoms are less severe, do not occur as frequently, or have not been occuring for long enough to warrant a full diagnosis of insomnia. However, people with sub-clinical insomnia may still experience significant distress or impairment in daytime functioning, but not to the extent where it would be considered a clinical disorder)

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v readr     2.1.4
## v ggplot2   3.4.4     v stringr   1.5.0
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks mice::filter(), stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggplot2)
library(dplyr)
library(viridis)
```

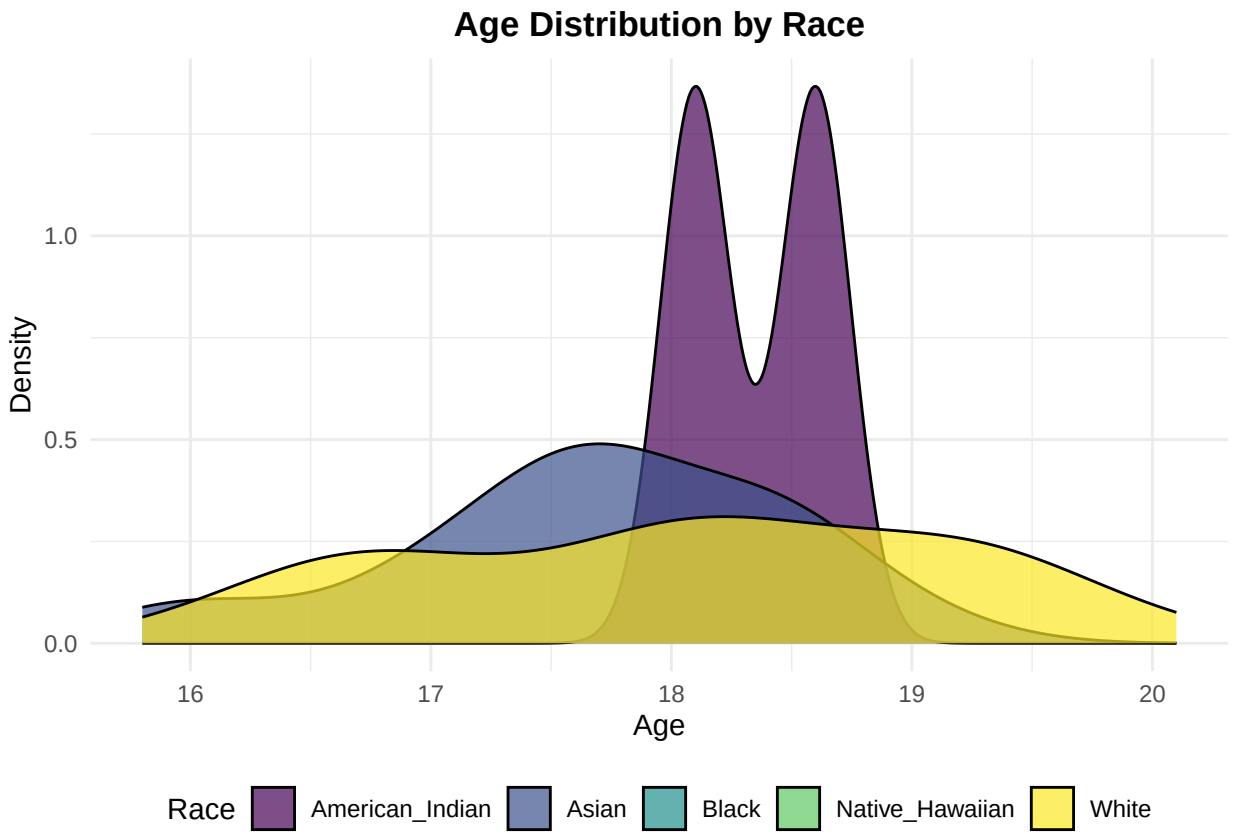
```
## Loading required package: viridisLite
```

```r
# Combine individual race columns into one for the Insomnia_Final dataframe
Insomnia_Final <- Insomnia_Final %>%
  mutate(
    Race = case_when(
      American_Indian == 1 ~ "American_Indian",
      Asian == 1 ~ "Asian",
      Native_Hawaiian == 1 ~ "Native_Hawaiian",
      Black == 1 ~ "Black",
      White == 1 ~ "White",
      TRUE ~ "Other"
    )
  )

# Density plot for age distribution within each race
densityplot_age_by_race <- ggplot(Insomnia_Final, aes(x = Age, fill = Race)) +
  geom_density(alpha = 0.7) +
  scale_fill_viridis(discrete = TRUE) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    legend.position = "bottom"
  ) +
  labs(
    title = "Age Distribution by Race",
    x = "Age",
    y = "Density"
  )

# Bar plot for subgroup distribution by race with subgroup names and smaller labels
barplot_group_by_race <- ggplot(Insomnia_Final, aes(x = as.factor(SubGroup), fill = Race)) +
  geom_bar(position = position_dodge()) +
  scale_fill_viridis(discrete = TRUE) +
  scale_x_discrete(labels = c("0" = "0 = Control", "1" = "1 = Clean Insomnia", "2" = "2 = Sub-clinical
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    legend.position = "bottom",
    axis.text.x = element_text(angle = 45, hjust = 1, face = "plain", size = 8)
  ) +
  labs(
    title = "SubGroup Distribution by Race",
    x = "SubGroup",
    y = "Count"
  )

# Render the plots
print(densityplot_age_by_race)
```

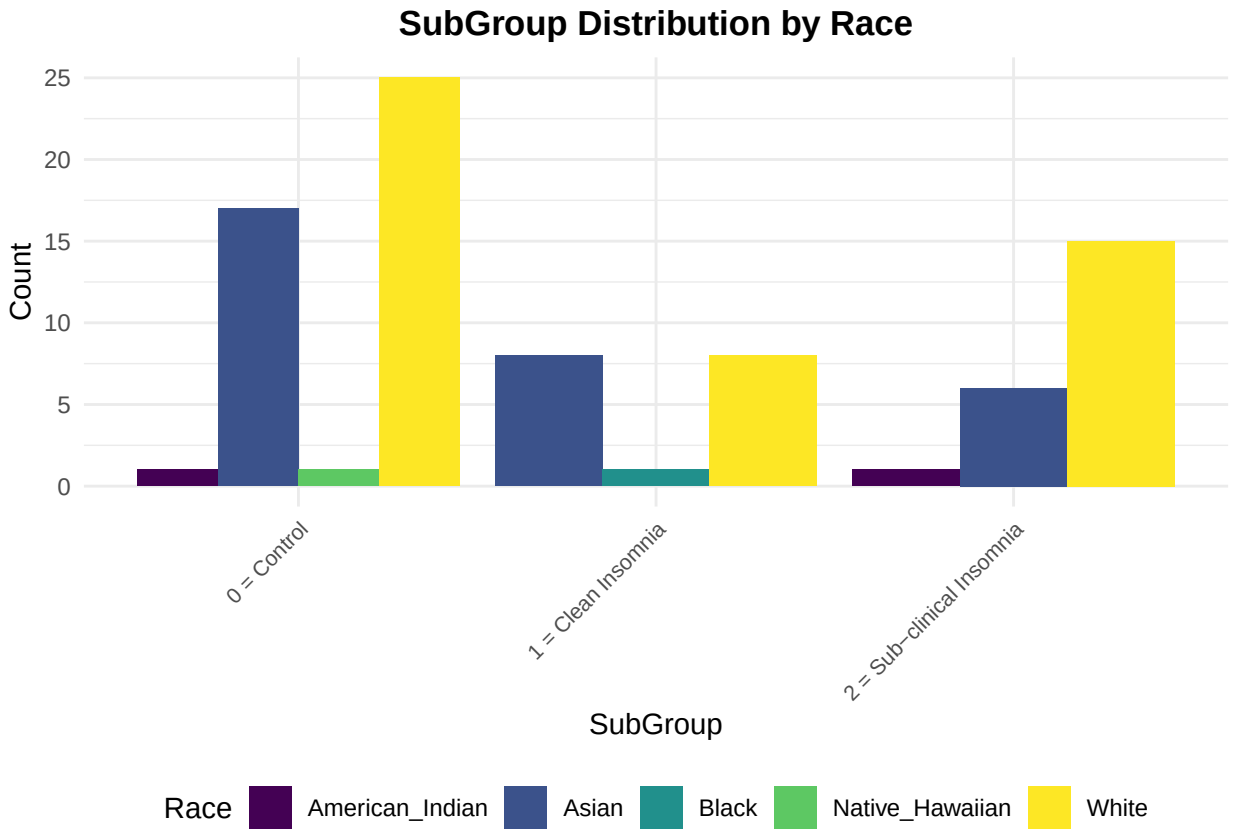
```
## Warning: Groups with fewer than two data points have been dropped.
## Groups with fewer than two data points have been dropped.
```

```
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning
## -Inf
```

```
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning
## -Inf
```




```
print(barplot_group_by_race)
```



Key Findings: Based on the 'Age Distribution by Race' plot above, we can see that Both 'White' and 'Asian' age distribution are pretty evenly spread whereas 'American Indian' almost only has age 18-19 dataset. However, since this entire analysis is based on adolescent, 'Age' won't be our main focus here and we would only use it as a reference for future analysis.

As shown in the 'SubGroup Distribution by Race' plot above, we do not have all races in each group (only 'White' and 'Asian' are presented in all three sub groups), which brings us the question of whether to include every single race in our future analysis.

Variable Selection

Now, we will move on to compare and test correlations between different variables to see which variables are the most suitable for our analysis. Since it's not efficient to generate all plots combination all at once, we can now conduct a pairwise correlation test among all variables to perform variable selection.

```
# List of column names to calculate pairwise correlations
columns_to_analyze <- c("PDS_FEMALE", "PDS_MALE", "ISI_total", "PSQI_total", "BDI_total",
                        "ASHS_total", "ASHS_physiological", "ASHS_cognitive", "ASHS_emotional",
                        "ASHS_SleepEnvirnmont", "ASHS_DaytimeSleep", "ASHS_substances",
                        "ASHS_bedtimeRoutine", "ASHS_sleepStability", "ASHS_BedroomSharing",
                        "DBAS_total", "FIRST_total", "GCTI_total", "GCTI_anxiety", "GCTI_reflection",
                        "GCTI_worries", "GCTI_thoughts", "GCTI_negativeAffect", "STAI_Y_total",
                        "NEO_neuroticism", "NEO_extraversion", "NEO_openness", "NEO_agreeableness",
                        "NEO_Conscientiousness", "MEQr_total", "PSRS_PrR", "PSRS_RWO", "PSRS_RSC",
                        "PSRS_FRa", "PSRS_RSE", "PSRS_total", "PSS_total", "TCQI_R_Total",
                        "TCQIR_Aggressive_supression", "TCQIR_cognitive_distraction", "TCQIR_reappraisal
                        "TCQIR_behavtioral_distraction", "TCQIR_social_avoidance", "TCQIR_worry",
                        "ACE_tot", "asq_home", "asq_school", "asq_attendance", "asq_romantic", "asq_peer
```

```
                        "asq_teacher", "asq_future", "asq_leisure", "asq_finance", "asq_responsibility"
                        "casq_total", "casq_sleepy", "casq_alert", "cope_disengage_su", "cope_growth",
                        "cope_disengage_mental", "cope_emotions", "cope_socialsupp_instr", "cope_active
                        "cope_denial", "cope_religion", "cope_humor", "cope_disengage_emo", "cope_restra
                        "cope_socialsupp_emo", "cope_acccept", "cope_suppression", "cope_planning",
                        "ders_nonaccpetance", "ders_total", "ders_goals", "ders_impulse", "ders_awarenes
                        "ders_strategies", "ders_clarity")

# Create an empty list to store the results
correlation_results <- list()

# Generate all combinations of the column names
column_combinations <- combn(columns_to_analyze, 2)

# Function to calculate Pearson correlation for a pair
calculate_correlation <- function(column_pair) {
  col1 <- Insomnia[[column_pair[1]]]
  col2 <- Insomnia[[column_pair[2]]]
  complete_cases <- complete.cases(col1, col2)

  # Check if there are enough complete cases
  if (sum(complete_cases) > 2) {  # More than two observations are needed to calculate correlation
    result <- cor.test(col1[complete_cases], col2[complete_cases], method = "pearson")
    return(list(column_pair = column_pair, correlation_coefficient = result$estimate,
                p_value = result$p.value, method = "pearson"))
  } else {
    return(list(column_pair = column_pair, correlation_coefficient = NA,
                p_value = NA, method = "pearson",
                warning = "Not enough finite observations"))
  }
}

# Apply the function to each pair
correlation_results <- apply(column_combinations, 2, calculate_correlation)

# Convert the results to a data frame, including a warning column for pairs with insufficient data
correlation_df <- do.call(rbind, lapply(correlation_results, function(x) {
  data.frame(column1 = x$column_pair[1], column2 = x$column_pair[2],
             correlation_coefficient = x$correlation_coefficient,
             p_value = x$p_value, warning = ifelse(is.na(x$correlation_coefficient), x$warning, ""))
}))

# View the dataframe
head(correlation_df,10)
```

```
##         column1              column2 correlation_coefficient    p_value
## 1    PDS_FEMALE            PDS_MALE                      NA         NA
## cor  PDS_FEMALE            ISI_total            -0.18812968 0.15360598
## cor1 PDS_FEMALE           PSQI_total            -0.17943143 0.17388840
## cor2 PDS_FEMALE            BDI_total            -0.29147820 0.02509976
## cor3 PDS_FEMALE           ASHS_total             0.14057581 0.28825183
## cor4 PDS_FEMALE   ASHS_physiological            -0.04335602 0.74657877
## cor5 PDS_FEMALE       ASHS_cognitive             0.26348574 0.04376304
## cor6 PDS_FEMALE       ASHS_emotional            -0.03191713 0.81034907
```

```
## cor7 PDS_FEMALE ASHS_SleepEnvirnmont              0.09716665 0.46409706
## cor8 PDS_FEMALE    ASHS_DaytimeSleep             -0.18393988 0.16314540
##                               warning
## 1    Not enough finite observations
## cor
## cor1
## cor2
## cor3
## cor4
## cor5
## cor6
## cor7
## cor8
```

```
#view(correlation_df)
```

Based on our the above battery tests, variable selections, and our defined data dictionaries, we have chosen the following variables for further investigation:

ISI_total: Insomnia severity Index *BDI_total*: Becks Depression Inventory *GCTI_total*: The Glasgow Content of Thoughts Inventory (e.g., anxiety, reflection, worries, thoughts, negativeAffect) *ASHS_total*: Adolescent Sleep Hygiene Scale (e.g., physiological, emotional, SleepEnvironment, Substances, bedtimeRoutine)

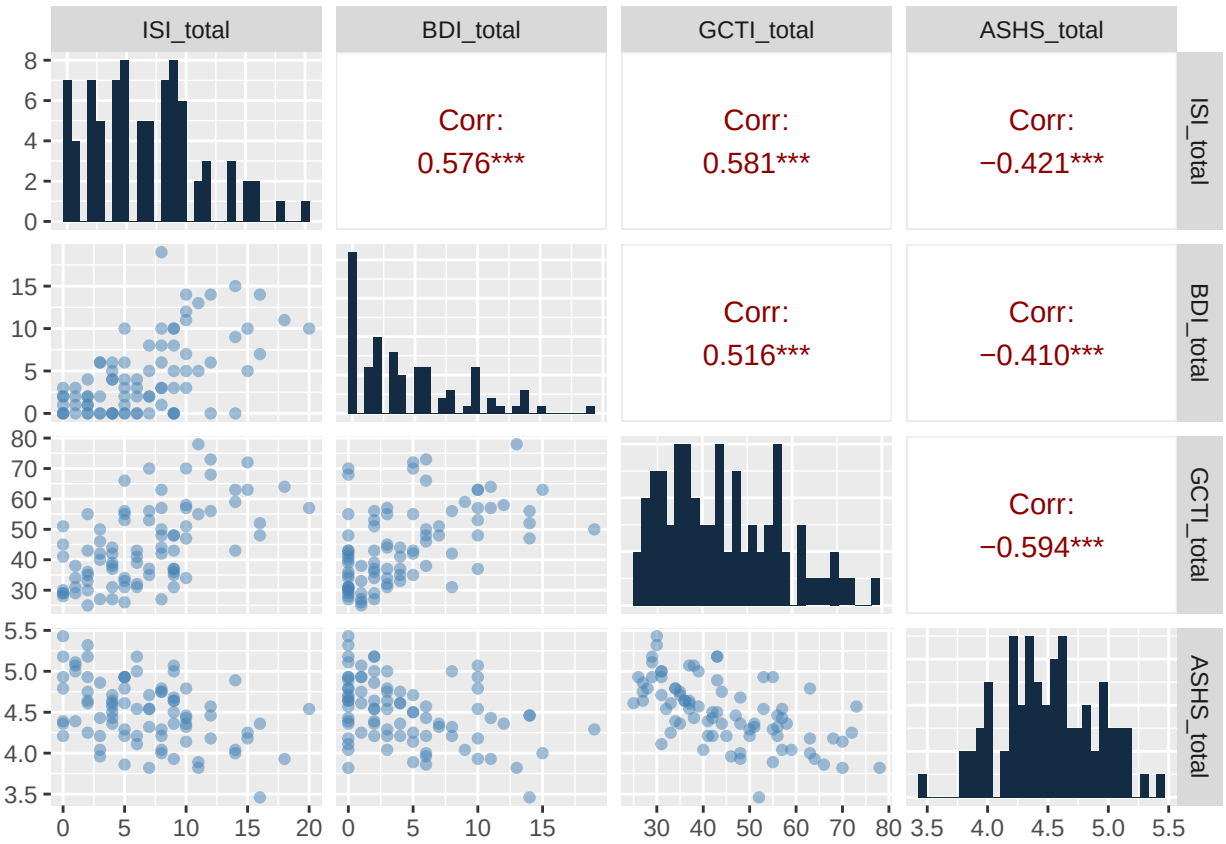
```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(ggplot2)

# Define the columns of interest
cols_of_interest <- c("ISI_total", "BDI_total", "GCTI_total", "ASHS_total")

# Create the scatter plot matrix with a specified color
scatter_plot_matrix <- ggpairs(
  Insomnia_Final[, cols_of_interest],
  lower = list(
    continuous = wrap("points", color = "steelblue", alpha = 0.5)
  ),
  upper = list(
    continuous = wrap("cor", color = "darkred")
  ),
  diag = list(
    continuous = wrap("barDiag", fill = "#132B43")
  )
)

# Print the plot
print(scatter_plot_matrix)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Key Findings: Positive Correlations: (For example) As the level of depression increases, the severity of Insomnia will also increase.Negative Correlations: (For example) As the level of depression increases, the adolescent sleep hygiene scale (sleep environment, the use of substances, bedtime routines) will decrease.In summary, insomnia, depression, as well as the thoughts inventory all have positive correlations, whereas the ASHS_total score (physiological factors) has negative correlations with all three other variables.

After investigating the total scores, we can now dig deeper into individual psychological factors and how they affect the severity of insomnia and depression. For example, since we are particularly interested in anxiety, worries, sleep stability, and the use of substances, we will examine the following four variables in relationship with insomnia and depression severity.

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(ggplot2)

# Define the variables and titles
variables <- c("ASHS_sleepStability", "ASHS_substances", "GCTI_anxiety", "GCTI_worries")
titles <- c("Insomnia Severity", "Insomnia Severity", "Depression", "Depression")

# Define a function to plot different plots into one grid
plot_variable <- function(df, x_var, y_var, title_prefix) {
  p <- ggplot(df, aes_string(x = x_var, y = y_var)) +
    geom_point(aes_string(color = x_var), alpha = 0.6) +
```

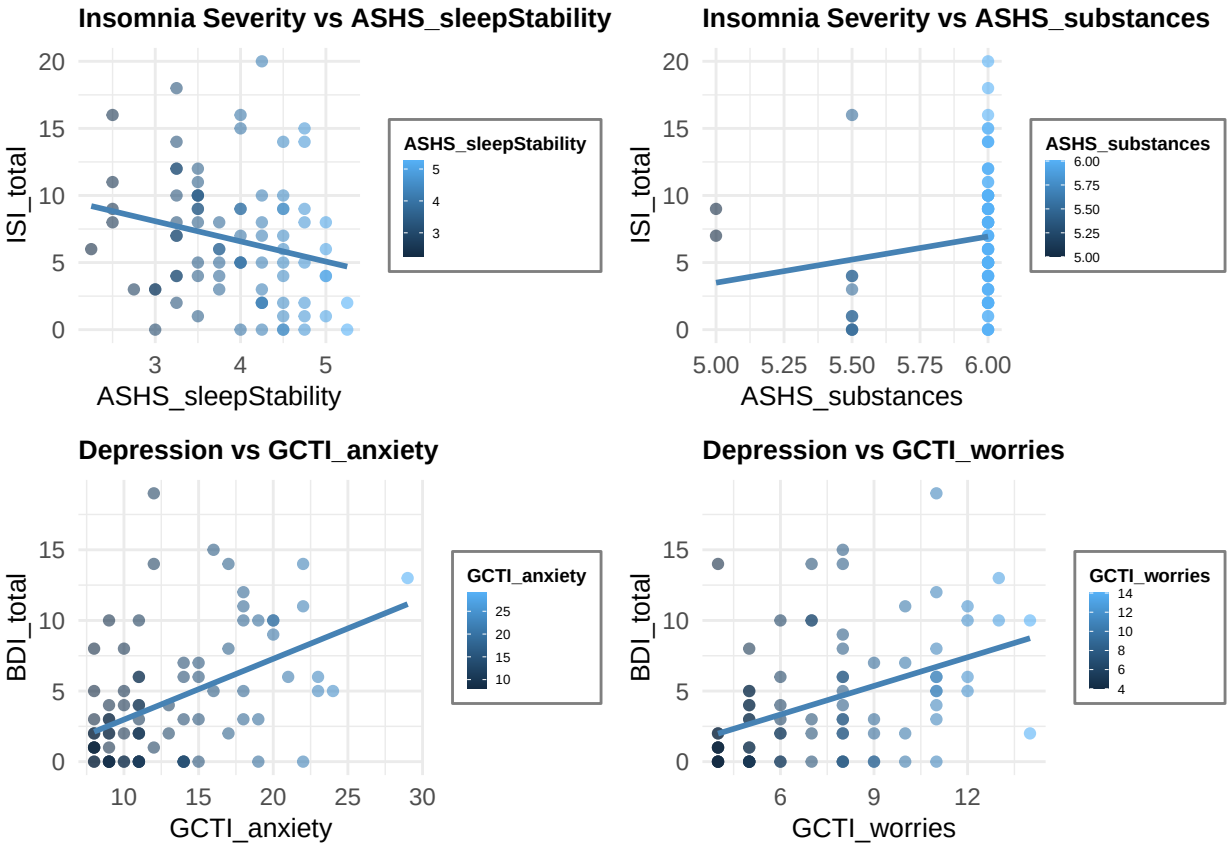
```
    geom_smooth(method = "lm", se = FALSE, color = "steelblue") +
    labs(title = paste(title_prefix, "vs", x_var),
         x = x_var,
         y = y_var) +
    theme_minimal() +
    theme(
      plot.title = element_text(face = "bold", size = 10),
      axis.title = element_text(size = 10),
      legend.position = "right",
      legend.background = element_rect(fill = "white", colour = "grey50"),
      legend.text = element_text(size = 5),
      legend.title = element_text(face = "bold", size = 7),
      legend.key.size = unit(0.5, "lines")
    )
  return(p)
}

# Create a list to hold plots
plot_list <- list()

# Generate plots
for (i in 1:length(variables)) {
  if (titles[i] == "Insomnia Severity") {
    plot_list[[i]] <- plot_variable(Insomnia_Final, variables[i], "ISI_total", titles[i])
  } else {
    plot_list[[i]] <- plot_variable(Insomnia_Final, variables[i], "BDI_total", titles[i])
  }
}
```

```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with `aes()`.
## i See also `vignette("ggplot2-in-packages")` for more information.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
# Arrange the plots in a grid
do.call(gridExtra::grid.arrange, c(plot_list, ncol = 2))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



Key Findings: There are 3 positive correlations and 1 negative correlation in the grid on the left-hand side, indicating that only Insomnia Severity and sleep stability is negatively correlated (As the score of sleep stability increases, the level of insomnia severity decreases, which makes a lot of sense because both teenagers and adults need to sleep well in order to avoid insomnia and even depression)

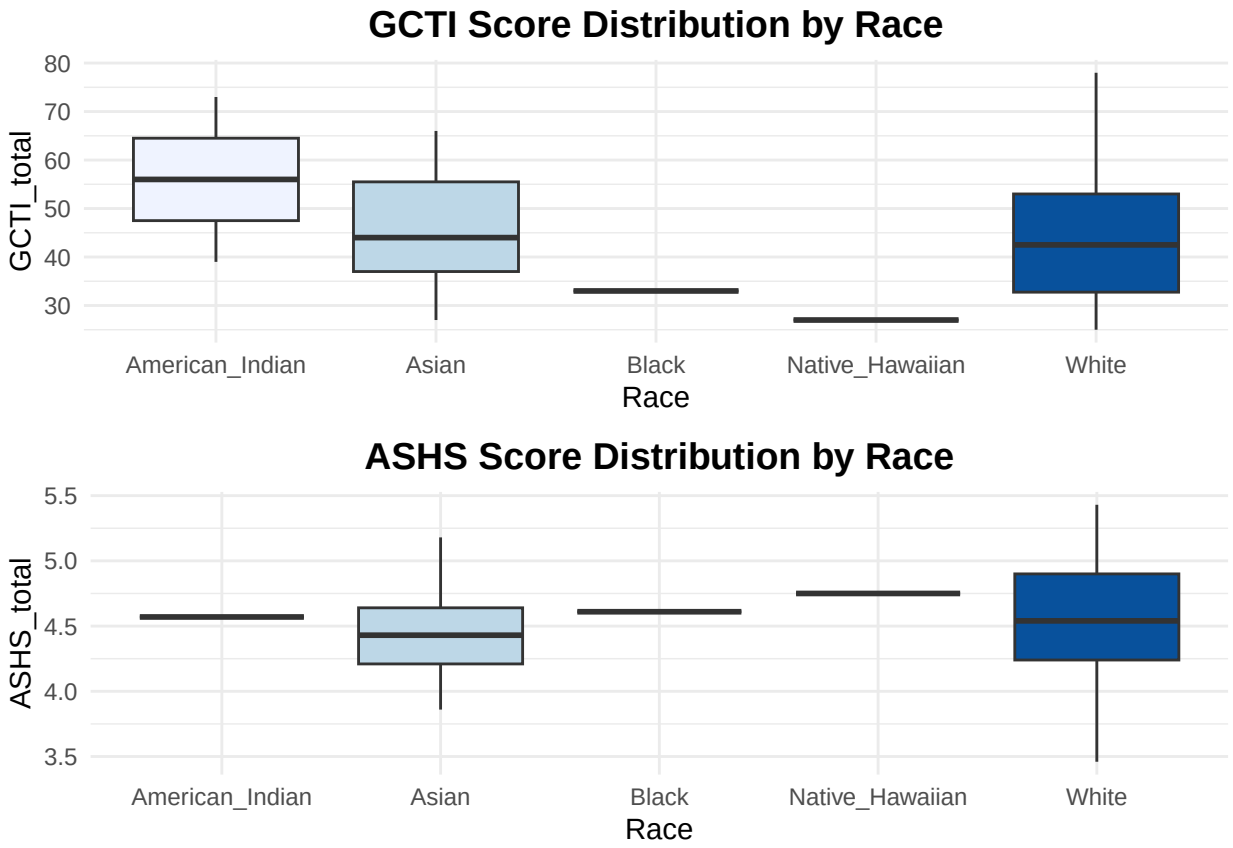
We are now ready to conduct our actual aims:

Aim #1 (1.1) Part A

We can first take a look at GCTI and ASHS score distributions for different ages and races.

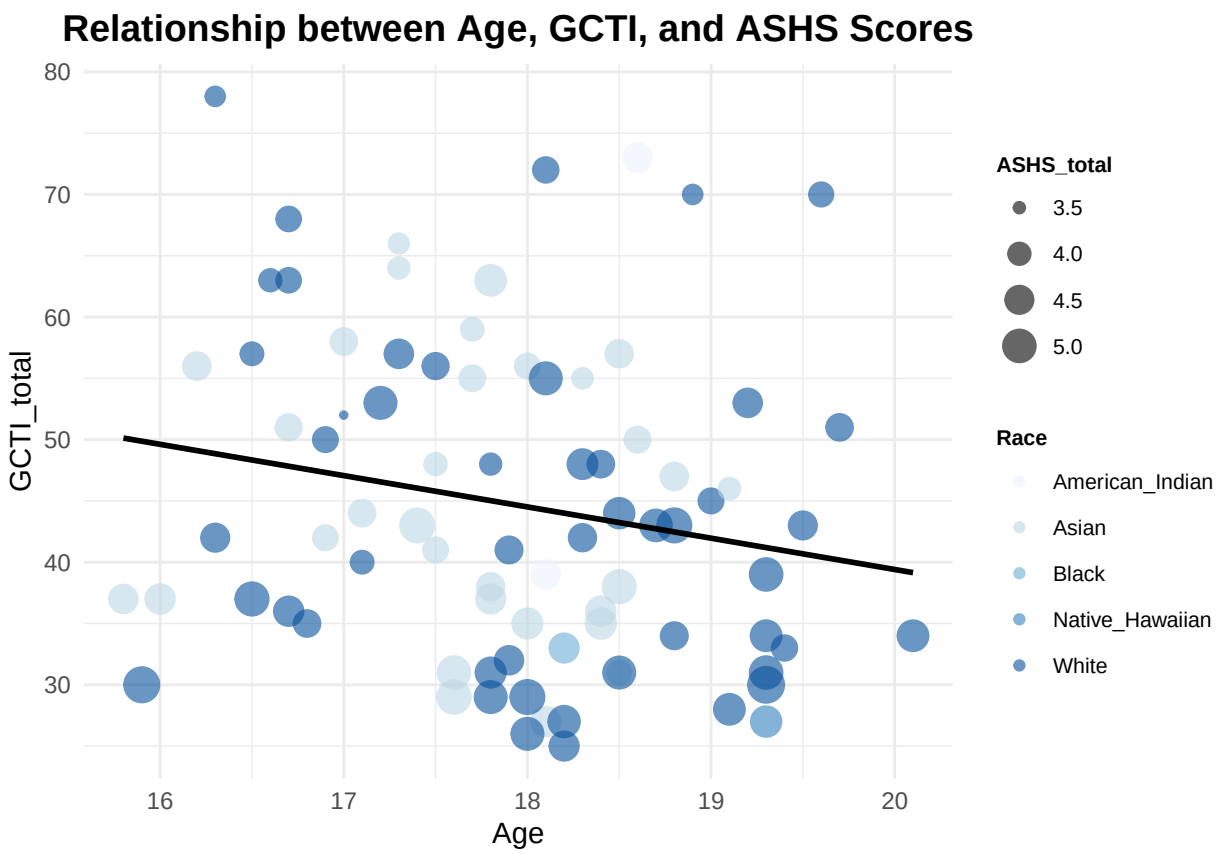
```
library(ggplot2)
library(dplyr)
library(gridExtra)
library(RColorBrewer)

# Set the base theme for our plots
base_theme <- theme_minimal() +
              theme(plot.title = element_text(face = "bold", hjust = 0.5, size = 14),
                    plot.subtitle = element_text(hjust = 0.5))
color_palette <- scale_fill_brewer(palette = "Blues")

# GCTI Score Distribution by Race
gcti_race_boxplot <- ggplot(Insomnia_Final, aes(x = Race, y = GCTI_total, fill = Race)) +
  geom_boxplot() +
  labs(title = "GCTI Score Distribution by Race") +
  color_palette +
  base_theme +
```

```
  theme(legend.position = "none")

# ASHS Score Distribution by Race
ashs_race_boxplot <- ggplot(Insomnia_Final, aes(x = Race, y = ASHS_total, fill = Race)) +
  geom_boxplot() +
  labs(title = "ASHS Score Distribution by Race") +
  color_palette +
  base_theme +
  theme(legend.position = "none")

# Relationship between Age, GCTI, and ASHS Scores with Race color distinction
age_gcti_ashs_scatter <- ggplot(Insomnia_Final, aes(x = Age, y = GCTI_total, color = Race)) +
  geom_point(aes(size = ASHS_total), alpha = 0.6) +
  geom_smooth(aes(y = GCTI_total), method = "lm", se = FALSE, color = "black") +
  labs(title = "Relationship between Age, GCTI, and ASHS Scores") +
  scale_color_brewer(palette = "Blues") +
  base_theme +
  theme(legend.title = element_text(size = 8, face = "bold"),
        legend.text = element_text(size = 8))

# Arrange the boxplots in a grid for better comparison
grid.arrange(gcti_race_boxplot, ashs_race_boxplot, nrow = 2)
```




```
# Print the scatter plot
print(age_gcti_ashs_scatter)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

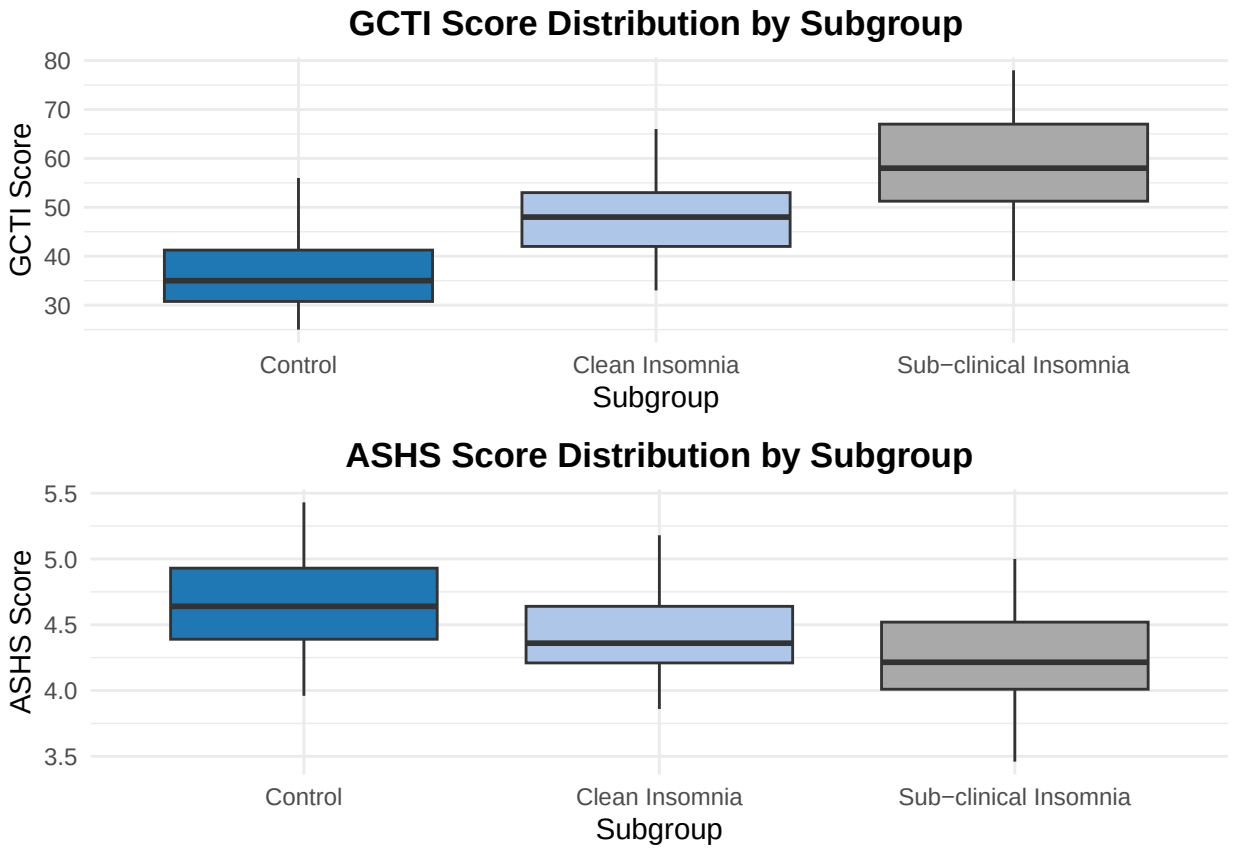


Aim #1 (1.1) Part B

We can now take a look at GCTI and ASHS score distributions for different subgroups.

```
# Load necessary library for grid arranging
library(gridExtra)

# Update the ggplot calls to add labels for subgroups
gcti_boxplot <- ggplot(Insomnia_Final, aes(x = factor(SubGroup), y = GCTI_total, fill = factor(SubGroup
  geom_boxplot() +
  scale_fill_manual(values = c("#1f77b4", "#aec7e8", "darkgrey")) +
  scale_x_discrete(labels = c("0" = "Control", "1" = "Clean Insomnia", "2" = "Sub-clinical Insomnia")) +
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5),
    legend.position = "none"
  ) +
  labs(
    title = "GCTI Score Distribution by Subgroup",
    x = "Subgroup",
    y = "GCTI Score"
  )

ashs_boxplot <- ggplot(Insomnia_Final, aes(x = factor(SubGroup), y = ASHS_total, fill = factor(SubGroup
  geom_boxplot() +
  scale_fill_manual(values = c("#1f77b4", "#aec7e8", "darkgrey")) +
```

```
  scale_x_discrete(labels = c("0" = "Control", "1" = "Clean Insomnia", "2" = "Sub-clinical Insomnia")) +
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5),
    legend.position = "none"
  ) +
  labs(
    title = "ASHS Score Distribution by Subgroup",
    x = "Subgroup",
    y = "ASHS Score"
  )

# Combine the plots into a grid
combined_plots <- grid.arrange(gcti_boxplot, ashs_boxplot, nrow = 2)
```




```
# Print the combined plot grid
print(combined_plots)
```

```
## TableGrob (2 x 1) "arrange": 2 grobs
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (2-2,1-1) arrange gtable[layout]
```

Key Findings: For the distribution by subgroup plot, 'Sub-clinical Insomnia (Sub group 2)' has the highest median score for thoughts inventory factors, which might imply more severe cognitive impacts or a greater need for cognitive therapy. For the sleep hygiene factors however, it is the other way around. We will go over what they each mean more in depth in our discussion section.

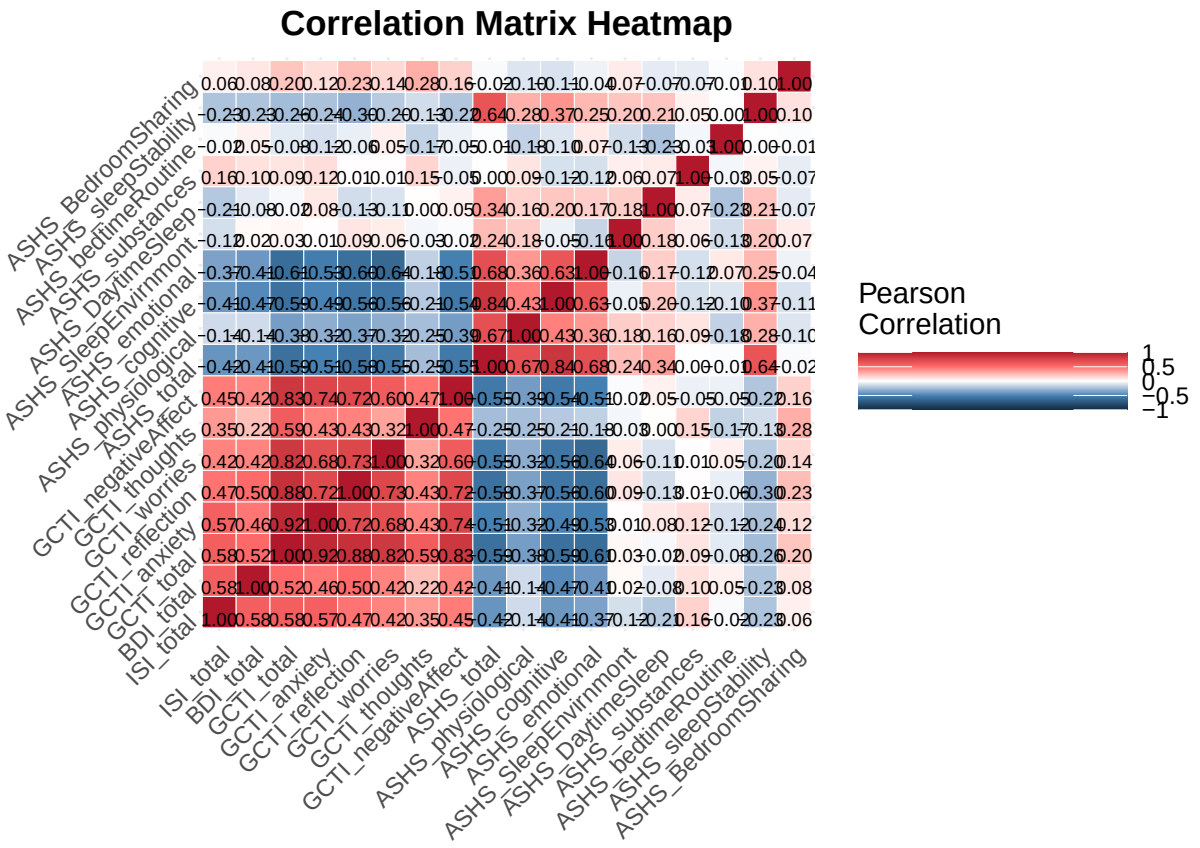
Aim #1 (1.2) Part A

```
library(ggplot2)
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
```

```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
library(dplyr)

# Select variables starting with GCTI and ASHS along with ISI_total and BDI_total for the correlation m
selected_vars <- Insomnia_Final %>%
  select(ISI_total, BDI_total, starts_with("GCTI"), starts_with("ASHS"))

# Compute correlation matrix
cor_matrix <- cor(selected_vars, use = "complete.obs")

# Melt the correlation matrix for ggplot2
cor_data <- melt(cor_matrix)

# Define specific colors for blue-white-red gradient
blue_white_red_colors <- c("#132B43", "steelblue", "white", "indianred1", "#B2182B")

# Visualize the correlation matrix
correlation_matrix_heatmap <- ggplot(data = cor_data, aes(Var1, Var2, fill = value)) +
  geom_tile(color = "white") +
  geom_text(aes(label = sprintf("%.2f", value)), vjust = 1, color = "black", size = 2.5) +
  scale_fill_gradientn(colors = blue_white_red_colors,
                       limits = c(-1, 1),
                       breaks = c(-1, -0.5, 0, 0.5, 1),
                       labels = c("-1", "-0.5", "0", "0.5", "1"),
                       name="Pearson\nCorrelation") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust=1),
        axis.text.y = element_text(angle = 45, vjust = 1, hjust=1),
        axis.title = element_blank(),
        plot.title = element_text(hjust = 0.5, face = "bold")) +
  labs(title = "Correlation Matrix Heatmap") +
  guides(fill = guide_colorbar(barwidth = 7, barheight = 1.5))

# Print the plot
print(correlation_matrix_heatmap)
```



We have perfect positive correlations when it equals to 1 or red, perfect negative correlations when it equals to -1 or blue, and the values and different colors schemes in between.

Key findings: There are relatively strong positive correlations between the insomnia severity index and the depression index, as well as the thoughts inventory factors such as anxiety, reflection, and negative affect. This finding suggests a relationship between cognitive and hypnosis-related factors and the severity of insomnia symptoms.

Sleep hygiene factors (except for the use of substances) however, are mostly negatively correlated with the severity of insomnia, which suggests that as the level of sleep stability, bedtime routine increases, the severity of insomnia decreases.

Aim #2 (2.1) Part A

Now moving to our Aim #2, which is to predict depression based on insomnia and sleep hygiene, we approached this aim with running the ANOVA test (i.e., two-ways, three-ways, and four-ways) to find their significance. We will first start with a two-way ANOVA test with the Insomnia Severity Index (ISI), GCTI (Thought Inventories), BDI (Depression Index), and the sleep hygiene scale.

```
# Create a new race factor variable where 1 represents 'White' and 2 represents 'Asian'
Insomnia_Final$Race <- NA  # Create an empty column for race
Insomnia_Final$Race[Insomnia_Final$White == 1] <- 'White'
Insomnia_Final$Race[Insomnia_Final$Asian == 1] <- 'Asian'

# Convert the new Race column to a factor
Insomnia_Final$Race <- factor(Insomnia_Final$Race)

# Now, select only the rows where Race is 'Asian' or 'White'
```

```r
Insomnia_Final_subset <- Insomnia_Final[Insomnia_Final$Race %in% c('Asian', 'White'), ]

# Run a two-way ANOVA with Race and ISI_total
result_gcti_isi <- aov(ISI_total ~ Race + GCTI_total + Race:GCTI_total, data = Insomnia_Final_subset)
summary(result_gcti_isi)
```

```
##                 Df Sum Sq Mean Sq F value   Pr(>F)
## Race             1   13.3    13.3   0.901    0.346
## GCTI_total       1  605.7   605.7  40.918 1.17e-08 ***
## Race:GCTI_total  1    4.3     4.3   0.291    0.591
## Residuals       76 1125.0    14.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Aim #2 (2.1) Part B

```r
# Run a two-way ANOVA with Race and ISI_total
result_isi_bdi <- aov(BDI_total ~ Race + ISI_total + Race:ISI_total, data = Insomnia_Final_subset)
summary(result_isi_bdi)
```

```
##                Df Sum Sq Mean Sq F value   Pr(>F)
## Race            1  102.5   102.5   7.869  0.00638 **
## ISI_total       1  504.7   504.7  38.755 2.43e-08 ***
## Race:ISI_total  1   20.9    20.9   1.608  0.20869
## Residuals      76  989.8    13.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Aim #2 (2.1) Part C

```r
# Run a two-way ANOVA with Race and ISI_total
result_ashs_bdi <- aov(BDI_total ~ Race + ASHS_total + Race:ASHS_total, data = Insomnia_Final_subset)
summary(result_ashs_bdi)
```

```
##                 Df Sum Sq Mean Sq F value   Pr(>F)
## Race             1  102.5  102.48   6.061 0.016090 *
## ASHS_total       1  229.7  229.69  13.583 0.000426 ***
## Race:ASHS_total  1    0.6    0.64   0.038 0.845993
## Residuals       76 1285.2   16.91
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

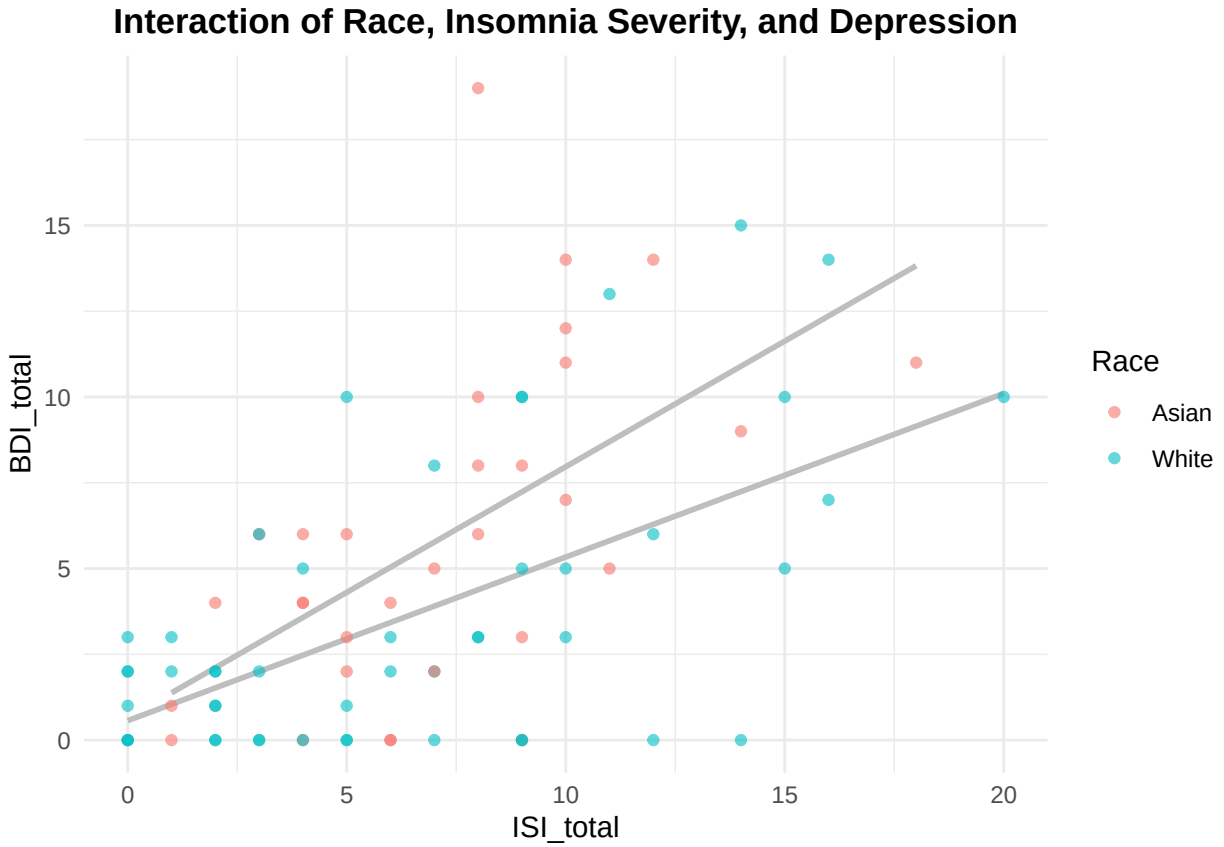
Based on these results, we can see that race effect was not significant, while the insomnia severity level and sleep hygiene were significantly correlated with depression level.

Aim #2 (2.1) Part D

```r
# Interaction plot for Race and ISI_total
ggplot(Insomnia_Final_subset, aes(x = ISI_total, y = BDI_total, color = Race, group = Race)) +
  geom_smooth(aes(y = BDI_total), method = "lm", se = FALSE, color ="grey") +
  geom_point(aes(color = Race), alpha = 0.6) +
  theme_minimal() +
```

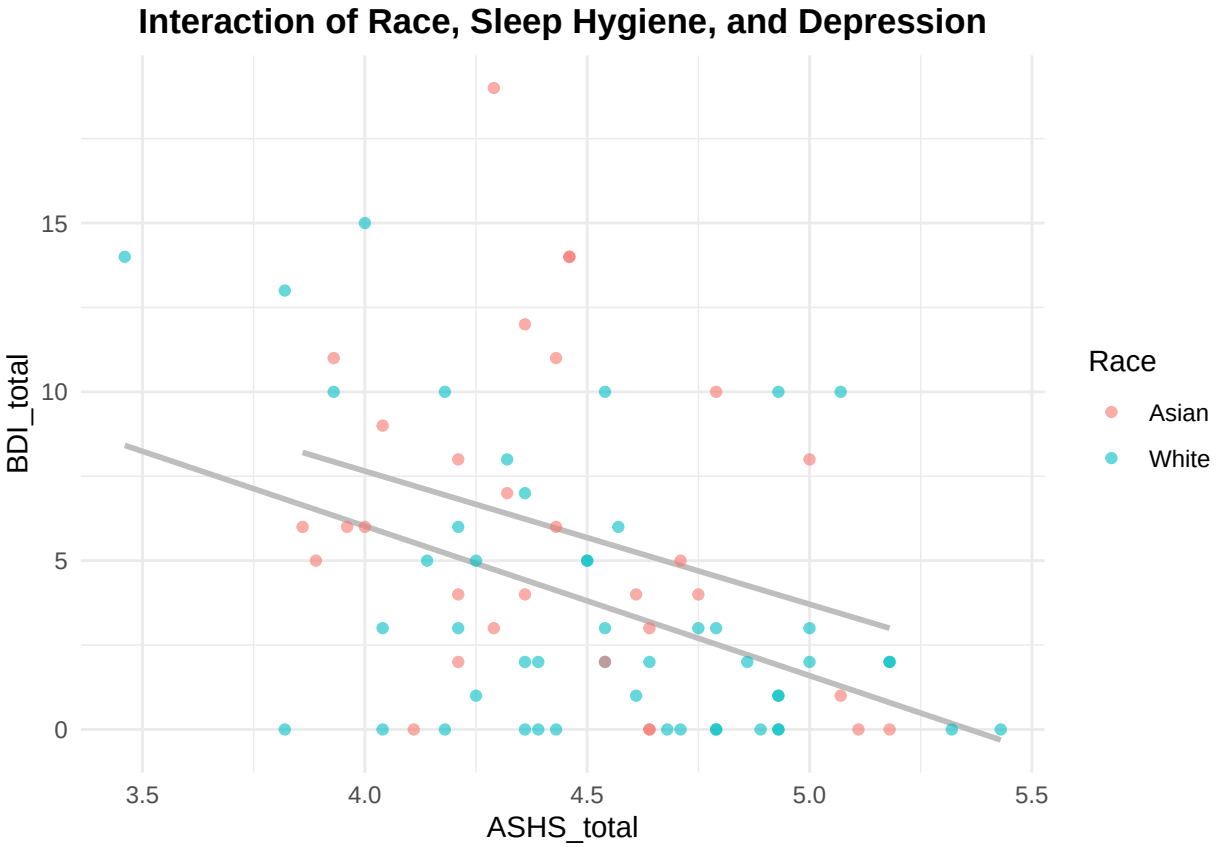
```
  theme(plot.title = element_text(hjust = 0.5, face = "bold")) +
  labs(title = "Interaction of Race, Insomnia Severity, and Depression", x = "ISI_total", y = "BDI_total
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
# Interaction plot for Race and ASHS_total
ggplot(Insomnia_Final_subset, aes(x = ASHS_total, y = BDI_total, color = Race, group = Race)) +
  geom_smooth(aes(y = BDI_total), method = "lm", se = FALSE, color ="grey") +
  geom_point(aes(color = Race), alpha = 0.6) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold")) +
  labs(title = "Interaction of Race, Sleep Hygiene, and Depression", x = "ASHS_total", y = "BDI_total")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Again, based on the visualization shown above, it re-emphazie on the idea that race effect is not significant enough to determine a relationship between Race and Insomnia, and therefore cannot be used as a predictor to make prediction for Insomnia.

Please also note that only 'White' and 'Asian' are chosen here for the two-way ANOVA test based on previous data explorations and preliminary steps (e.g., since other races might not have data presented in all subgroups).

Aim #2 (2.2)

To further investigate on their relationships, we will perform interaction of depression, insomnia, sleep hygiene and race on a three-way ANOVA test.

```
# Three way ANOVA
Insomnia_Final$Race <- factor(Insomnia_Final$Race)

# Perform the three-way ANOVA
three_way_anova_result <- aov(BDI_total ~ Race * ISI_total * ASHS_total, data = Insomnia_Final)

# Get the summary of the ANOVA
summary(three_way_anova_result)
```

```
##                          Df Sum Sq Mean Sq F value   Pr(>F)
## Race                      1  102.5   102.5   8.464  0.00482 **
## ISI_total                 1  504.7   504.7  41.686 1.09e-08 ***
## ASHS_total                1   40.6    40.6   3.351  0.07132 .
## Race:ISI_total            1   19.9    19.9   1.644  0.20383
## Race:ASHS_total           1    3.2     3.2   0.262  0.61044
```

```
## ISI_total:ASHS_total       1   20.7    20.7   1.709  0.19525
## Race:ISI_total:ASHS_total  1   54.6    54.6   4.512  0.03710 *
## Residuals                 72  871.8    12.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 3 observations deleted due to missingness
```

Based on this three-way ANOVA result, we can see that the race did not correlate with other sleep factors again. It is interesting to note that three way interaction effect were marginally more significant than the two-way ANOVA test.

For better comparison, we generated a R Shiny Dashboard to see the difference and changes from an one-way ANOVA analysis to a three-way ANOVA analysis among our chosen variables.

```r
library(shiny)
library(ggplot2)

# Define UI for the dashboard
ui <- fluidPage(
    titlePanel("ANOVA Comparisons on Insomnia Data"),

    sidebarLayout(
        sidebarPanel(
            selectInput("anovaType", "Select ANOVA Type:",
                        choices = c("One-Way", "Two-Way", "Three-Way")),
            selectInput("dependentVar", "Dependent Variable",
                        choices = c("BDI_total", "ISI_total", "ASHS_total", "GCTI_total")),
            selectInput("independentVar1", "Independent Variable 1",
                        choices = c("Race", "BDI_total", "ISI_total", "ASHS_total", "GCTI_total")),
            selectInput("independentVar2", "Independent Variable 2",
                        choices = c("Race", "BDI_total", "ISI_total", "ASHS_total", "GCTI_total")),
            selectInput("independentVar3", "Independent Variable 3",
                        choices = c("Race", "BDI_total", "ISI_total", "ASHS_total", "GCTI_total"))
        ),

        mainPanel(
           textOutput("anovaResult"),
           plotOutput("plot")
        )
    )
)

# Define server logic
server <- function(input, output) {

    output$anovaResult <- renderText({
        anovaType <- input$anovaType
        dependentVar <- input$dependentVar
        independentVar1 <- input$independentVar1
        independentVar2 <- input$independentVar2
        independentVar3 <- input$independentVar3

        if(anovaType == "One-Way") {
            result <- summary(aov(as.formula(paste(dependentVar, "~", independentVar1)), data = Insomnia
        } else if(anovaType == "Two-Way") {
```

```
            result <- summary(aov(as.formula(paste(dependentVar, "~", independentVar1, "*", independentV
        } else {
            result <- summary(aov(as.formula(paste(dependentVar, "~", independentVar1, "*", independentV
        }

        return(capture.output(print(result)))
    })

    output$plot <- renderPlot({
        anovaType <- input$anovaType
        dependentVar <- input$dependentVar
        independentVar1 <- input$independentVar1
        independentVar2 <- input$independentVar2
        independentVar3 <- input$independentVar3

        if(anovaType == "One-Way") {
            ggplot(Insomnia_Final, aes_string(x = independentVar1, y = dependentVar)) +
                geom_boxplot() +
                labs(title = "One-Way ANOVA", x = independentVar1, y = dependentVar)
        } else if(anovaType == "Two-Way") {
            ggplot(Insomnia_Final, aes_string(x = independentVar1, y = dependentVar, fill = independentV
                geom_boxplot() +
                labs(title = "Two-Way ANOVA", x = independentVar1, y = dependentVar)
        } else {
            ggplot(Insomnia_Final, aes_string(x = independentVar1, y = dependentVar, color = independen
                geom_point() +
                labs(title = "Three-Way ANOVA", x = independentVar1, y = dependentVar)
        }
    })
}

# Run the application
shinyApp(ui = ui, server = server)
```

```
## PhantomJS not found. You can install it with webshot::install_phantomjs(). If it is installed, pleas
```

Last but not least, we can conduct a machine learning (Random Forest) model with tidyverse and rsample to predict Depression based on the Insomnia Severity index total score.

```
library(randomForest)
```

```
## randomForest 4.7-1.1
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:gridExtra':
##
##     combine
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
## The following object is masked from 'package:dplyr':
```

```
##
##     combine
```

```
library(tidyverse)
library(rsample)

# Using tidyverse and rsample for data manipulation
set.seed(181)

# Splitting the data into training and testing sets
split <- initial_split(Insomnia_Final, prop = 0.8)
train_data <- training(split)
test_data <- testing(split)

# Training a Random Forest model
model <- randomForest(BDI_total ~ ISI_total, data = train_data, ntree = 500)
print(model)
```

```
##
## Call:
##  randomForest(formula = BDI_total ~ ISI_total, data = train_data,      ntree = 500)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 1
##
##           Mean of squared residuals: 15.69761
##                     % Var explained: 18.06
```

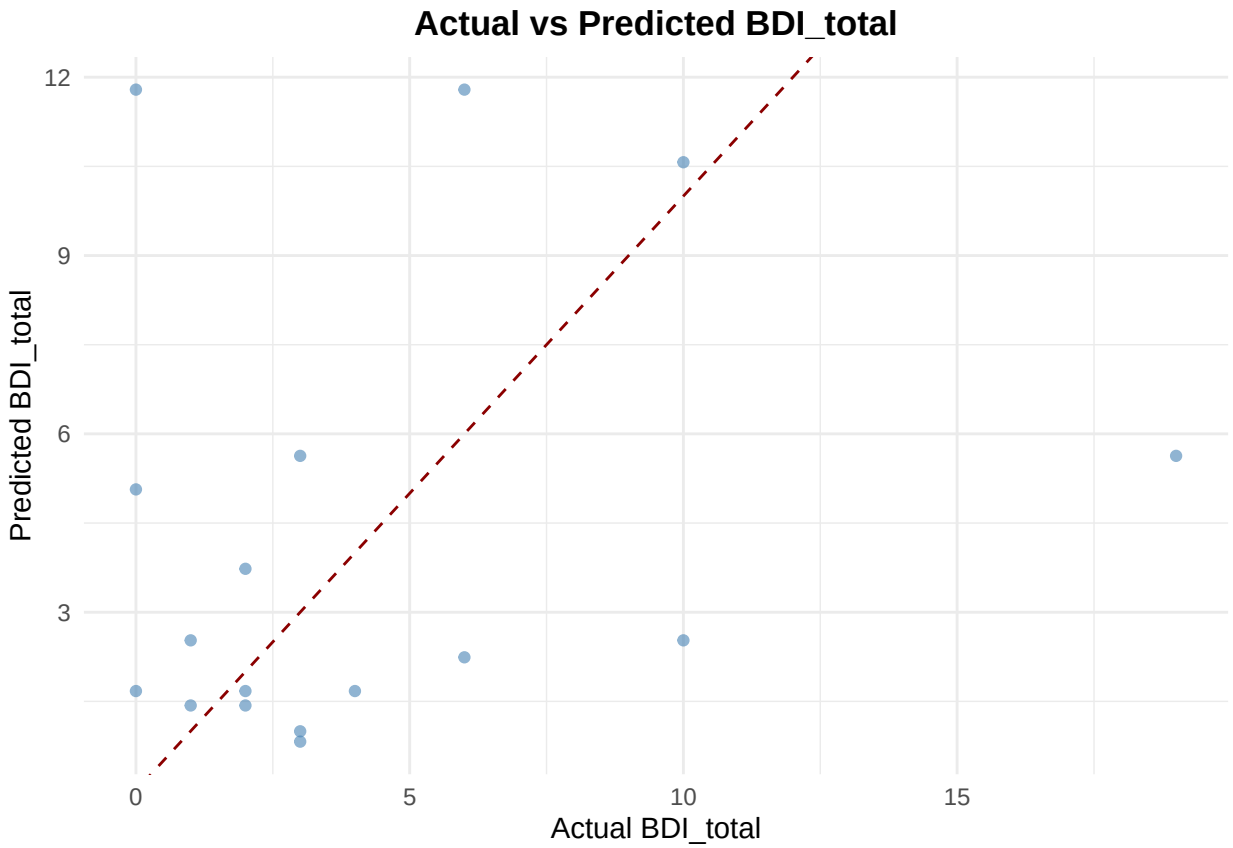
```
# Predicting on the test data
predictions <- predict(model, test_data)

# Evaluating model performance
# For a regression task, we can use RMSE (Root Mean Squared Error)
test_data$predicted_BDI <- predictions
RMSE <- sqrt(mean((test_data$BDI_total - test_data$predicted_BDI)^2))
RMSE
```

```
## [1] 5.297605
```

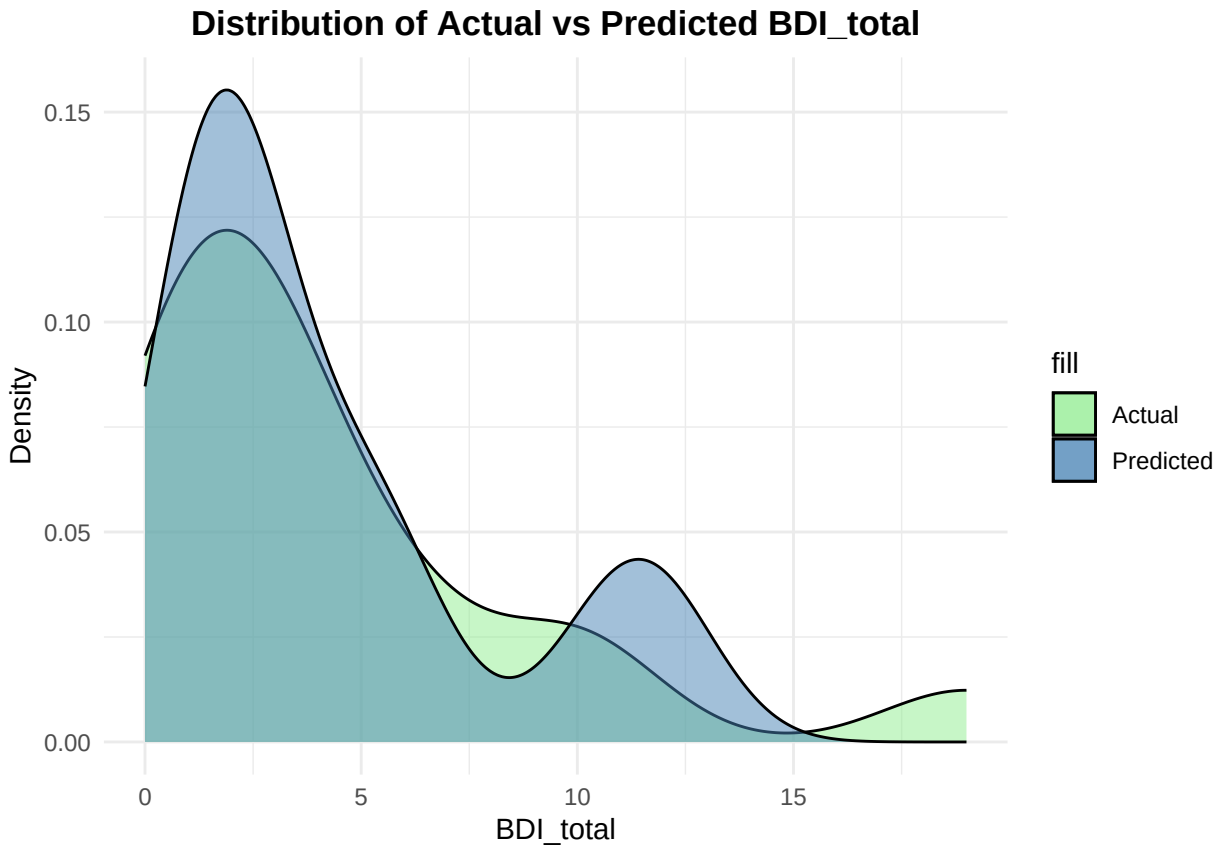
Key Findings: This Random Forest model predicting depression (measured by BDI_total) from insomnia severity (ISI_total) exhibits modest accuracy. With a mean of squared residuals at 15.697 and an explanation of 18.06% of the variance in BDI_total, the model indicates a moderate relationship between insomnia and depression. However, the Root Mean Squared Error (RMSE) of 5.297605 suggests that while there is some predictive capability, precision is limited. This outcome reveals that insomnia severity alone is not a highly precise predictor of depression levels, suggesting the need for us to consider additional factors or variables to enhance the model's predictive power and accuracy.

```
library(ggplot2)

# Adding predictions to the test dataset
test_data$predicted_BDI <- predictions

# Scatter plot of actual vs predicted values
ggplot(test_data, aes(x = BDI_total, y = predicted_BDI)) +
  geom_point(color = 'steelblue', alpha = 0.6) +
  geom_abline(intercept = 0, slope = 1, linetype = "dashed", color = "darkred") +
  labs(title = "Actual vs Predicted BDI_total",
```

```
       x = "Actual BDI_total",
       y = "Predicted BDI_total") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```




```
# Distribution of predictions and actual values
ggplot(test_data) +
  geom_density(aes(x = BDI_total, fill = "Actual"), alpha = 0.5) +
  geom_density(aes(x = predicted_BDI, fill = "Predicted"), alpha = 0.5) +
  labs(title = "Distribution of Actual vs Predicted BDI_total", x = "BDI_total", y = "Density") +
  scale_fill_manual(values = c("lightgreen", "steelblue")) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```



Key Findings: *Distribution of Actual vs Predicted BDI_total Plot*: Since there is an overlap between the two distributions, this suggests that while the model can capture the general trend in depression scores, there are differences, particularly in the tails, where the model seems less accurate.

Actual vs Predicted BDI_total Plot: The spread of points suggests the model has moderate predictive power, with some variance not captured by ISI_total alone, as evidenced by points falling away from the line.

Results summary, limitations, and future steps The observed correlation between increasing insomnia severity and higher depression scores underscores the complex interplay between sleep disturbances and mood disorders. It is important to consider, however, that the generalizability of these findings from our analysis is constrained by the demographic composition of the dataset, which predominantly includes White and Asian adolescents. This demographic limitation highlights the necessity for further research that encompasses a broader, more diverse population to validate and extend these findings.

Additional Wrangling and a taste of further investigation Given the above-mentioned limitations of our dataset, we also decided to play around with 'webscraping' to compare results out of curiosity.

By using a 'webscraping' method, we can use APIs to collect and 'scrap' data from websites and compare them with our insomnia data in the downloaded dataset. We will first need to generate an API key or a token from here: https://dev.elsevier.com/sc_apis.html

Webscraping Analysis and additional data collection for comparison

Let's first start with scraping 'Insomnia' related articles in PMC.

```
library(tidyverse)
library(httr)
library(jsonlite)
```

```
##
## Attaching package: 'jsonlite'

## The following object is masked from 'package:shiny':
##
##     validate

## The following object is masked from 'package:purrr':
##
##     flatten
```

```
library(xml2)

# Set my own api token
rscopus::set_api_key("074bdd40f4bdeb52fc8892db76f8a0e9")

# Base URL for E-utilities API
base_url <- "https://eutils.ncbi.nlm.nih.gov/entrez/eutils/"

# Update the search term to include both "Insomnia" and "adolescents"
search_term <- "Insomnia[Title/Abstract] AND adolescents[Title/Abstract]"

# Construct the search query URL with the updated search term
search_url <- paste0(base_url, "esearch.fcgi?db=pmc&term=",
                     gsub(" ", "+", search_term), "&retmode=xml")

# Send the search request to the NCBI API
search_response <- GET(search_url)

# Parse the XML response
search_content <- content(search_response, "text")
search_xml <- read_xml(search_content)

# Extract the PMC article IDs from the search result
pmc_ids <- xml_find_all(search_xml, ".//Id") %>% xml_text()

# Print the PMC IDs
print(pmc_ids)
```

```
##  [1] "10605192" "10566088" "10519740" "10478937" "10339205" "10339185"
##  [7] "10228684" "10104418" "9923883"  "9632537"  "9838014"  "9739758"
## [13] "9441791"  "9212947"  "9384123"  "9006600"  "9058217"  "8807915"
## [19] "8559387"  "8533758"
```

Now, with all the IDs, let's scrape content from three randomly chosen PMC_IDs (articles) for specific key words related to insomnia (e.g., Insomina, Depression, Worries, Sleep Environment, Sleep Routine, Substance).

```
library(httr)
library(XML)

# Function to fetch the abstract of a PMC article using E-utilities
fetch_abstract <- function(pmc_id) {
  base_url <- "https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi"
  query_list <- list(db="pmc", id=pmc_id, retmode="xml")

  # Fetch the abstract using E-utilities
```

```
  response <- GET(url=base_url, query=query_list)

  if (status_code(response) == 200) {
    # Parse the XML content
    content <- content(response, "text", encoding = "UTF-8")
    xml_content <- xmlParse(content)

    # Extract the abstract
    abstract_node <- getNodeSet(xml_content, "//abstract//p")
    abstract <- xmlValue(abstract_node[[1]])
    return(abstract)
  } else {
    return(NULL)
  }
}

# Keywords to search within the abstract
keywords <- c("Depression", "Anxiety", "Sleep Environment", "Insomnia", "Worries", "Substance", "Adoles

# PMC IDs to fetch abstracts for
pmc_ids <- c("8533758", "9632537", "9006600")

# Fetch abstracts and check for keywords
results <- lapply(pmc_ids, function(id) {
  abstract <- fetch_abstract(id)
  sapply(keywords, function(keyword) grepl(keyword, abstract, ignore.case = TRUE))
})
print(results)
```

```
## [[1]]
##       Depression          Anxiety Sleep Environment          Insomnia
##             TRUE             TRUE             FALSE             TRUE
##          Worries        Substance        Adolescent
##            FALSE            FALSE             TRUE
##
## [[2]]
##       Depression          Anxiety Sleep Environment          Insomnia
##            FALSE            FALSE             FALSE             TRUE
##          Worries        Substance        Adolescent
##            FALSE            FALSE             TRUE
##
## [[3]]
## [[3]]$Depression
## logical(0)
##
## [[3]]$Anxiety
## logical(0)
##
## [[3]]$`Sleep Environment`
## logical(0)
##
## [[3]]$Insomnia
## logical(0)
##
```

```
## [[3]]$Worries
## logical(0)
##
## [[3]]$Substance
## logical(0)
##
## [[3]]$Adolescent
## logical(0)
```

Based on these results, we can see all PMC articles contain 'Insomnia' and 'Adolescent', but not all of them mentioned other psychological factors. However, based on this small random trial, we can assume that 'anxiety' and 'Depression' are typically more correlated with 'Insomnia' and especially among 'Adolescent' from published papers.

Let's now fetch texts containing the above-mentioned keywords, and see the number of time they got mentioned in the three journal articles.

```
library(httr)
library(xml2)
library(stringr)

pmc_ids <- c("8533758", "9632537", "9006600")

# Create a dataframe for keywords
keywords <- c("Depression", "Anxiety", "Sleep Environment", "Insomnia", "Worries", "Substance", "Adoles

# Function to fetch abstracts and count keywords
fetch_and_count_keywords <- function(pmc_id, keywords) {
  base_url <- "https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi"

  # Fetch the content of the article
  response <- GET(url = paste0(base_url, "?db=pmc&id=", pmc_id, "&retmode=xml"))
  content <- content(response, "text", encoding = "UTF-8")

  # Parse the content
  parsed_content <- read_xml(content)

  # Extract the full text or abstract
  text_nodes <- xml_find_all(parsed_content, "//body")
  full_text <- paste(xml_text(text_nodes), collapse = " ")

  # Initialize a list to store counts
  keyword_counts <- list()

  # Loop through keywords and count their occurrence
  for (keyword in keywords) {
    sentences <- str_extract_all(full_text, str_glue("(?i)[^.]*\\b{keyword}\\b[^.]*\\."))
    count <- length(sentences[[1]])
    keyword_counts[[keyword]] <- count
  }

  return(keyword_counts)
}

# Loop over PMC IDs and apply the function
```

```r
results <- lapply(pmc_ids, fetch_and_count_keywords, keywords)
print(results)
```

```
## [[1]]
## [[1]]$Depression
## [1] 12
##
## [[1]]$Anxiety
## [1] 10
##
## [[1]]$`Sleep Environment`
## [1] 0
##
## [[1]]$Insomnia
## [1] 16
##
## [[1]]$Worries
## [1] 0
##
## [[1]]$Substance
## [1] 0
##
## [[1]]$Adolescent
## [1] 0
##
##
## [[2]]
## [[2]]$Depression
## [1] 3
##
## [[2]]$Anxiety
## [1] 3
##
## [[2]]$`Sleep Environment`
## [1] 0
##
## [[2]]$Insomnia
## [1] 26
##
## [[2]]$Worries
## [1] 0
##
## [[2]]$Substance
## [1] 1
##
## [[2]]$Adolescent
## [1] 0
##
##
## [[3]]
## [[3]]$Depression
## [1] 2
##
## [[3]]$Anxiety
```

```
## [1] 49
##
## [[3]]$`Sleep Environment`
## [1] 0
##
## [[3]]$Insomnia
## [1] 38
##
## [[3]]$Worries
## [1] 0
##
## [[3]]$Substance
## [1] 0
##
## [[3]]$Adolescent
## [1] 2
```

NLP Analysis

We can now conduct a NLP (sentiment analysis) now and calculate the average sentiment score for the randomly chosen articles.

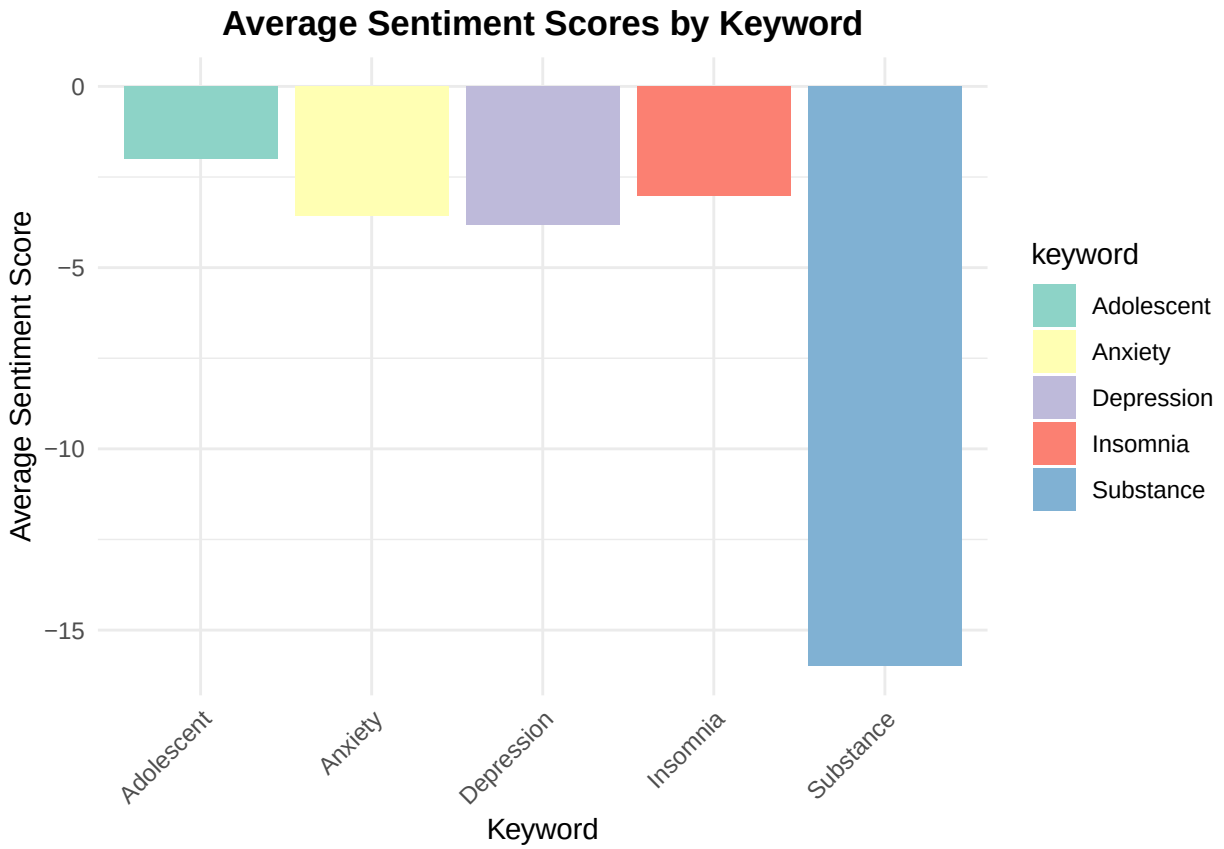
```
library(httr)
library(xml2)
library(stringr)
library(syuzhet)
library(dplyr)

# Function to fetch content, extract sentences with keywords, and analyze sentiment
analyze_sentiment <- function(pmc_id, keywords) {
  base_url <- "https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi"

  # Fetch the content of the article
  response <- GET(url = paste0(base_url, "?db=pmc&id=", pmc_id, "&retmode=xml"))
  content <- content(response, "text", encoding = "UTF-8")

  # Parse the content
  parsed_content <- read_xml(content)

  # Extract the full text or abstract
  text_nodes <- xml_find_all(parsed_content, "//body")
  full_text <- paste(xml_text(text_nodes), collapse = " ")

  # Initialize a data frame to store results
  sentiment_df <- data.frame(keyword = character(), sentence = character(), score = numeric())

  # Loop through keywords and perform sentiment analysis on sentences containing each keyword
  for (keyword in keywords) {
    keyword_sentences <- str_extract_all(full_text, str_glue("(?i)[^.?!]*\\b{keyword}\\b[^.?!]*[.?!]"))
    keyword_sentences <- unlist(keyword_sentences)

    # Get sentiment scores for sentences containing the keyword
    if (length(keyword_sentences) > 0) {
      scores <- get_sentiment(keyword_sentences, method = "afinn")
```

```
      # Combine sentences and scores into a data frame
      keyword_sentiment_df <- data.frame(keyword = keyword, sentence = keyword_sentences, score = scores

      # Bind to the overall sentiment data frame
      sentiment_df <- rbind(sentiment_df, keyword_sentiment_df)
    }
  }

  return(sentiment_df)
}

# Loop over PMC IDs and apply the sentiment analysis function
sentiment_analysis_results <- lapply(pmc_ids, analyze_sentiment, keywords)

# Summarize the sentiment scores for each keyword across all articles
sentiment_summary <- bind_rows(sentiment_analysis_results) %>%
  group_by(keyword) %>%
  summarise(average_score = mean(score, na.rm = TRUE), .groups = 'drop')

# Print the sentiment summary
print(sentiment_summary)
```

```
## # A tibble: 5 x 2
##   keyword    average_score
##   <chr>              <dbl>
## 1 Adolescent         -2
## 2 Anxiety            -3.56
## 3 Depression         -3.82
## 4 Insomnia           -3.01
## 5 Substance         -16
```

Visualization for sentiment analysis above

```
library(ggplot2)

# Use sentiment_summary dataframe to create a combined bar plot
ggplot(sentiment_summary, aes(x = keyword, y = average_score, fill = keyword)) +
  geom_bar(stat = "identity", position = position_dodge()) +
  scale_fill_brewer(palette = "Set3") +
  labs(x = "Keyword", y = "Average Sentiment Score", title = "Average Sentiment Scores by Keyword") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```



Key Findings It's without a doubt that both Insomnia and psychological factors (e.g., depression, anxiety) are negative, and 'substance' has the highest negative sentiment score (i.e., -16) among all variables due to its definition and impression. However, 'Adolescent' is also surprisingly a negative word based on the sentiment analysis.

Final Notes Insomnia remains a multifaceted challenge with many dimensions yet to be fully understood. The intricacies of its causes, effects, and treatments invite a continuous and evolving inquiry. Current research has laid a substantial foundation, but there is a vast expanse of knowledge that beckons for deeper investigation. Future studies should strive to unravel the complexities of insomnia, exploring the interplay of genetic, environmental, and psychological factors. It is essential that we persist in our pursuit of more sophisticated and nuanced analysis to develop targeted interventions. As our understanding grows, so too will our ability to offer more effective, personalized solutions for those afflicted by this pervasive sleep disorder. The journey of discovery is far from complete, and the path ahead promises to yield invaluable insights that will enrich our scientific and clinical approaches to managing insomnia. Hope you all enjoy this dataset and each step of our Insomnia analysis as much as we do!

The End of the Tutorial