# Clean and merge the Job Prep Datasets

Halle Prine

## Document Purpose

This document outlines the data cleaning process for the Job Preparation Study, including validation and formatting of IDs, cleaning of demographic, GPA, and survey data, and creation of composite variables needed for analysis.

## Data Import

First we will import the data and clean along the way whatever is necessary to clean first and foremost.

```
# Import raw demographics with unknown structure (extra column included)
demographics <- read_csv(here("Data", "Job Prep Demographics.csv"),
                         col_names = FALSE,
                         col_types = cols(.default = "c"))

# Import GPA Data
gpa <- read_csv(here("Data", "Job Prep GPA Data.csv"),
                skip = 5,
                col_types = "cnnfnnnn",
                col_names = c("ID", "school_code", "age", "gender",
                              "s1_credits", "s1_gpa", "s2_credits", "s2_gpa"))

# Import Survey Data
survey <- read_csv(here("Data", "Job Prep Survey Results.csv"),
                   skip = 4,
                   col_types = "cnnnnnnnnnnnnnn",
                   col_names = c("ID", paste0("item_", 1:14)))
```

### Data Cleaning

## Demographics Cleaning

Now we will clean each data frame starting with demographics

```
# Assign column names
colnames(demographics) <- c("ID", "school_code", "age", "gender", "discipline_raw")

# Flag and fill discipline rows
demographics <- demographics %>%
  mutate(discipline_marker = if_else(!is.na(ID) & is.na(school_code) & is.na(age) & is.na(ge
  fill(discipline_marker)

# Remove header junk and non-participant rows
demo_clean <- demographics %>%
  filter(!ID %in% c("ID", "Job Preparation Study", "Demographic Data"),
         !is.na(school_code), !is.na(age)) %>%
  transmute(ID = as.character(ID),
            school_code = as.character(school_code),
            age = as.numeric(age),
            gender = gender,
            discipline = discipline_marker)

# Remove identified junk rows and column
demo_tidy <- demographics %>%
  slice(-c(1:7, 146:150, 304:308, 464:468, 618:622)) %>%
  select(-discipline_raw)
```

## GPA Data Check and Cleanup

```
# Summary check of GPA values
gpa %>%
  summarise(
    min_s1 = min(s1_gpa, na.rm = TRUE),
    max_s1 = max(s1_gpa, na.rm = TRUE),
    min_s2 = min(s2_gpa, na.rm = TRUE),
    max_s2 = max(s2_gpa, na.rm = TRUE)
  )
```

```
# A tibble: 1 x 4
  min_s1 max_s1 min_s2 max_s2
   <dbl>  <dbl>  <dbl>  <dbl>
1   0.54   3.97   0.51   3.99
```

```
# Identify GPA values outside 1.0-4.3
gpa %>%
  filter(s1_gpa < 1 | s2_gpa < 1 | s1_gpa > 4.3 | s2_gpa > 4.3)
```

```
# A tibble: 37 x 8
   ID    school_code   age gender s1_credits s1_gpa s2_credits s2_gpa
   <chr>       <dbl> <dbl> <fct>       <dbl>  <dbl>      <dbl>  <dbl>
 1 308            22    22 f              16   1.38         12   0.53
 2 788             8    22 f              16   0.55         15   1.14
 3 326            15    23 f              15   0.96         16   1.19
 4 439            21    22 m              16   0.7          16   1.57
 5 854             9    22 f              16   1.3          17   0.87
 6 618            23    22 f              17   0.92         17   0.63
 7 516            13    22 f              18   1.99         15   0.95
 8 504            16    22 f              16   1.09         12   0.96
 9 123             7    22 f              15   0.85         15   1.03
10 435            19    23 m              15   1.14         16   0.62
# i 27 more rows
```

```
# Compute weighted GPA
gpa <- gpa %>%
  mutate(GPA_weighted = (s1_gpa * s1_credits + s2_gpa * s2_credits) / (s1_credits + s2_credi
```

## Survey Data Cleaning

```
# Our function that reverse codes and calculates composite scores
survey_tidy <- score_likert_scale(
  survey,
  scale_max = 6,
  reverse_items = c("item_4", "item_9", "item_12", "item_13")
) %>%
  rename(optimism_score = total_score)
```

### ID Checks

```
# Check uniqueness and formatting of IDs in each dataset
demographics %>% count(ID) %>% filter(n > 1)
```

```
# A tibble: 2 x 2
  ID        n
  <chr> <int>
1 ID        5
2 <NA>     15
```

```
gpa %>% count(ID) %>% filter(n > 1)
```

```
# A tibble: 1 x 2
  ID        n
  <chr> <int>
1 659       2
```

```
survey %>% count(ID) %>% filter(n > 1)
```

```
# A tibble: 0 x 2
# i 2 variables: ID <chr>, n <int>
```

```
# Check format (e.g., all numeric or 3-digit strings)
all_ids_valid <- function(id_vector) {
  all(str_detect(id_vector, "^\\d{3}$"))
}

all_ids_valid(demo_tidy$ID)
```

```
[1] TRUE
```

```
all_ids_valid(gpa$ID)
```

```
[1] TRUE
```

```
all_ids_valid(survey$ID)
```

```
[1] TRUE
```

## Data Merging

Now we will merge the data into one analytical dataset. Inorder to retain all participants from the demographics dataset (demo_tidy) we decided to use a left join and just identify missing survey or GPA data

```
# Make sure all IDs are character
demo_tidy <- demo_tidy %>% mutate(ID = as.character(ID))
gpa <- gpa %>% mutate(ID = as.character(ID))
survey_tidy <- survey_tidy %>% mutate(ID = as.character(ID))

# Merge datasets using left_join to retain all demographics
analytic_job_data <- demo_tidy %>%
  left_join(gpa, by = "ID") %>%
  left_join(survey_tidy, by = "ID")

# Check how many are missing GPA or survey data
summary(analytic_job_data)
```

```
      ID            school_code.x         age.x            gender.x
 Length:740         Length:740         Length:740         Length:740
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

```
 discipline_marker  school_code.y       age.y       gender.y     s1_credits
 Length:740         Min.   : 1.00   Min.   :20.00   f   :379   Min.   : 9.00
 Class :character   1st Qu.: 7.00   1st Qu.:22.00   m   :360   1st Qu.:15.00
 Mode  :character   Median :12.00   Median :22.00   NA's:  1   Median :16.00
                    Mean   :12.25   Mean   :22.01              Mean   :15.75
                    3rd Qu.:18.00   3rd Qu.:22.00              3rd Qu.:17.00
                    Max.   :23.00   Max.   :25.00              Max.   :18.00
                    NA's   :1       NA's   :1                  NA's   :1
     s1_gpa        s2_credits        s2_gpa       GPA_weighted
```

```
 Min.   :0.540   Min.   : 9.00   Min.   :0.510   Min.   :0.605
 1st Qu.:2.040   1st Qu.:15.00   1st Qu.:2.015   1st Qu.:2.076
 Median :2.610   Median :16.00   Median :2.640   Median :2.630
 Mean   :2.558   Mean   :15.69   Mean   :2.554   Mean   :2.556
 3rd Qu.:3.115   3rd Qu.:17.00   3rd Qu.:3.130   3rd Qu.:3.115
 Max.   :3.970   Max.   :18.00   Max.   :3.990   Max.   :3.975
 NA's   :1       NA's   :1       NA's   :1       NA's   :1
     item_1          item_2          item_3          item_4
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
 1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
 Median :3.000   Median :3.000   Median :3.000   Median :4.000
 Mean   :3.385   Mean   :3.469   Mean   :3.568   Mean   :3.582
 3rd Qu.:5.000   3rd Qu.:5.000   3rd Qu.:5.000   3rd Qu.:5.000
 Max.   :6.000   Max.   :6.000   Max.   :6.000   Max.   :6.000
 NA's   :207     NA's   :207     NA's   :207     NA's   :207
     item_5          item_6          item_7          item_8
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
 1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
 Median :4.000   Median :3.000   Median :3.000   Median :3.000
 Mean   :3.475   Mean   :3.497   Mean   :3.538   Mean   :3.484
 3rd Qu.:5.000   3rd Qu.:5.000   3rd Qu.:5.000   3rd Qu.:5.000
 Max.   :6.000   Max.   :6.000   Max.   :6.000   Max.   :6.000
 NA's   :207     NA's   :207     NA's   :207     NA's   :207
     item_9         item_10         item_11         item_12
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
 1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
 Median :3.000   Median :3.000   Median :3.000   Median :4.000
 Mean   :3.473   Mean   :3.484   Mean   :3.445   Mean   :3.572
 3rd Qu.:5.000   3rd Qu.:5.000   3rd Qu.:5.000   3rd Qu.:5.000
 Max.   :6.000   Max.   :6.000   Max.   :6.000   Max.   :6.000
 NA's   :207     NA's   :207     NA's   :207     NA's   :207
    item_13         item_14      optimism_score
 Min.   :1.000   Min.   :1.000   Min.   :30.00
 1st Qu.:2.000   1st Qu.:2.000   1st Qu.:42.00
 Median :3.000   Median :3.000   Median :48.00
 Mean   :3.411   Mean   :3.458   Mean   :48.84
 3rd Qu.:5.000   3rd Qu.:5.000   3rd Qu.:55.00
 Max.   :6.000   Max.   :6.000   Max.   :71.00
 NA's   :207     NA's   :207     NA's   :207
```

```r
# Count participants missing key fields
analytic_job_data %>%
```

```r
  summarise(
    missing_gpa = sum(is.na(s1_gpa) | is.na(s2_gpa)),
    missing_survey = sum(is.na(optimism_score))
  )
```

```
# A tibble: 1 x 2
  missing_gpa missing_survey
        <int>          <int>
1           1            207
```