**ANALYSIS FILES/STEPS:**

1. Data cleaning code (data-cleaning-code)
   1. PACKAGES:
      1. Tidyverse for ggplot, magriterr, etc
      2. Readxl for reading in data
      3. Writexl for write\_xlxs function
   2. Loaded individual datasets from each coder (measurement data only)
   3. Deleted one scan that was coded by two coders (Jared and Kayna, kept Kayna’s)
   4. Added “coder” column with name of each coder to each individual dataset
   5. Compiled all individual datasets into one dataset
   6. Renamed columns with abbreviations of measurements (Appendix A? of dissertation)
   7. Imported demographic dataset
   8. Recoded race/ethnicity with abbreviations in Table 3.3? of dissertation
   9. Added age group column
   10. Set and double checked variable classifications (numeric vs char vs factor)
   11. ***Created headscan\_full excel***
       1. ***All participants, all measurement values***
       2. ***Includes visual and statistic outliers, and NA values***
       3. ***All demographic categories are retained***
       4. ***Headscan measure data had 2021, demographic data had 2017***
          1. ***When joined, 2016 obs of 33 variables***
2. Demographic data exploration (demographic-data-exploration)
   1. PACKAGES:
      1. Tidyverse for ggplot, magriterr, etc
      2. Readxl for reading in data
      3. Extrafont to make plots in times new roman
      4. Flextable for flextable() with custom function (allows times new roman)
   2. Loaded headscan\_full
   3. Numeric age info:
      1. Table of sumstats
      2. Histogram of age frequency
   4. Race/ethnicity (all categories retained) info:
      1. Table of sumstats
      2. Boxplot of age by race/eth category
      3. Histogram of race/eth category frequencies
   5. Gender (all categories retained) info:
      1. Table of sumstats
      2. Boxplot of age by gender category
      3. Histogram of race/eth category frequencies
   6. Age group info:
      1. Table of sumstats
      2. Boxplot of age (num) by age group
      3. Histogram of age group frequencies
   7. Headscan sample demographics plot of age, sorted by race/eth, color by gender
   8. Couple of other plots (not that great)
3. Summary stats SA1 (summary-stats-SA1)
   1. PACKAGES:
      1. Tidyverse for ggplot, magriterr, etc
      2. Readxl for reading in data
      3. Extrafont to make plots in times new roman
      4. Flextable for flextable() with custom function (allows times new roman)
      5. Writexl for write\_xlxs function
      6. Forcats for fct\_reorder
   2. Loaded headscan\_full
   3. Transformed all measurement data from centimeters to millimeters
   4. ***Created measureNAs excel***
      1. ***Sum of NA values per measurement***
      2. ***Proportion of NA values per measurement***
   5. Measurement NA info:
      1. Table of proportion and sum of na values per each measurement location
      2. (CREATE NEW TABLE IN ORDER OF NA)- couldn’t figure out how to do this, flextable wants to do alphabetical
      3. Histogram of missing value ordered by measurement location (one for count, one for prop)
   6. Overall, race/eth, gender, age sumstats for all measurements!
      1. Histogram of measurement frequency
      2. Table of overall measurement sumstats
      3. Table of all race/eth categories sumstats
      4. Boxplots (ordered and not) of measurement by all race/eth categories
      5. Table of all gender categories sumstats
      6. Boxplots (ordered and not) of measurement by all gender categories
      7. Table of all age group categories sumstats
      8. Boxplots (ordered and not) of measurement by age group categories
   7. Removed visual outliers and redeveloped all of above for following measurements:
      1. GoSub\_C
      2. NRB\_L
      3. SelDH\_C
      4. SnasM\_C
      5. TrGo\_C
      6. TrSman\_C
      7. TrSnas\_C
4. Removing visual outliers (removing-visual-outliers)
   1. PACKAGES:
      1. Tidyverse for ggplot, magriterr, etc
      2. Readxl for reading in data
      3. Extrafont to make plots in times new roman
      4. Flextable for flextable() with custom function (allows times new roman)
      5. Writexl for write\_xlxs function
   2. Loaded headscan\_full
   3. Changed measurement data in cm to mm
   4. For all 7 above measures:
      1. Looking at histogram with all data
      2. Checking max value
      3. Change values close to max (visual outliers) to NA
      4. Checking how many values were filtered out
      5. Histogram of filtered data
   5. ***Created headscan\_full1 excel***
      1. ***2016 obs of 33 vars***
      2. ***26 individual measurement values changed to NA based on visual outlier status***
      3. ***Numeric measurement data changed from cm to mm***
   6. ***Created all\_vis\_out1 excel***
      1. ***2 columns:***
         1. ***ID: id number***
         2. ***Vis\_out: what was the measurement that was changed to NA in headscan\_full1 dataset***
5. Bivariate panel
   1. PACKAGES:
      1. Tidyverse for ggplot, magriterr, etc
      2. Readxl for reading in data
      3. Extrafont to make plots in times new roman
      4. rstatix for identify\_outliers
   2. Loaded headscan\_full1
   3. Made panel numbers dataframe to be able to use geom\_text inside geom\_rect(angle)
      1. From Zhuang 2007 article
   4. Generated bivariate panel with face width (Sellion menton length) and face width (bizygomatic length)
      1. Original zhaung 2007 panel overlaid
   5. Investigated and changed one outlier for face length to NA (using rstatix identify\_outliers)
   6. Generated same bivariate plot as above (d)
   7. Generated bivariate plot colored by gender
   8. Generated bivariate plot colored by age group
   9. Generated bivariate plot colored by race/eth
6. Exploring measurement variables
   1. PACKAGES:
      1. Tidyverse for ggplot, magriterr, etc
      2. Readxl for reading in data
      3. Extrafont to make plots in times new roman
      4. Flextable for flextable() with custom function (allows times new roman)
      5. Writexl for write\_xlxs function
      6. Corrr for correlate() function
      7. Ggcorrplot for correlation plot in times new roman
      8. Ggrepel for geom\_text\_repel
   2. Loaded headscan\_full1
   3. Created headscan\_num with only measurement variables (did not save excel)
   4. ***Used correlate from corrr package to create correlation\_data\_full dataset***
      1. ***Top and bottom correlation spreadsheet (duplicates)***
   5. ***Created correlation\_data and deleted code! Because it was edited in excel***
      1. ***Bottom only correlation spreadsheet (no duplicates)***
   6. Generated correlation plot using ggcorrplot (after some dataframe manipulation to wide dataframe)
   7. Loaded correlation\_data
   8. Pivoted longer and added column that lists “first & second” measure
   9. Removed missing values (same measure correlation)
   10. Generated table that lists first&second and correlation rounded to 4 digits
   11. Generated table as above sorted high to low correlation- couldn’t figure out how to do this, flextable wants to do alphabetical
   12. Created high\_cor dataframe with only correlation values over 0.7
       1. Considered highly correlated for this research
   13. Generated table of highly correlated measures
   14. Plot of correlations over 0.7
   15. From headscan\_full1 dataset, created dataset called coders (not to excel)
   16. Generated plot of how many participants each data collector digitized
   17. ***Loaded intraRR dataset (created in excel from Isabel thesis)***
   18. Generated plot of intraRR values wrapped by coder
   19. ***Created 4 excel spreadsheets for intraRR:***
       1. ***Kayna\_intra, Isabel\_intra, chandler\_intra, jared\_intra***
       2. ***Same as intraRR dataset but separated by coder***
   20. Generated plot of intraRR icc values ordered from low to high for all coders
   21. Created low\_intraRR dataframe by filtering to only icc values lower than 0.9
       1. Added column ‘percent coded’ indicating how much each data collector did
   22. Generated plot of low icc values wrapped by coder
   23. Generated dodge histogram of low icc values for all coders (color by coder)
   24. Created high\_cor\_full for values with high correlation
   25. Full joined low\_intraRR and high\_cor\_full by (first) measure
       1. Shows measurements of ‘concern’ or questionable measures
   26. Generated plot of correlation by intraRR icc values
   27. Generated plot of correlation by intraRR icc values wrapped by coder
   28. Generated plot of InterRR values ordered low to high icc
   29. Generated plot of just low interRR (after filtering to new dataframe)
   30. Generated plot of correlation by interRR icc values
   31. Generated plot of intraRR by interRR icc values wrapped by coder
   32. Generated plot of questionable stats by per measure (number)
   33. ***Created questionable measures excel, measures in this include:***
       1. ***All interRR iccs below 0.9***
       2. ***All intraRR iccs below 0.9***
       3. ***All correlation values 0.7***
       4. ***Indication for what measurements are involved***
7. Cors-interintra-nas
   1. PACKAGES:
      1. Tidyverse for ggplot, magriterr, etc
      2. Readxl for reading in data
      3. Extrafont to make plots in times new roman
      4. Flextable for flextable() with custom function (allows times new roman)
      5. Writexl for write\_xlxs function
   2. Loaded several datasets
   3. Generated table of all correlations, na proportion, all coder intraRR, interRR
      1. This was used by me to develop choosing-measurement-vars doc
   4. Generated tables for highly correlated variables, including all of above info
      1. AA\_C, ProA\_L, ProA\_C
      2. BiW\_C and BiW\_L
      3. SnasM\_L, SnasM\_C, SelM\_L
      4. TrSel\_C, TrSnas\_C, TrTr\_C, TrTr\_L (twice, oops)
      5. SelP\_L and SelP\_C
      6. ProS\_C and ProS\_L
      7. SmanM\_C and SmanM\_L
   5. All of this info was used in choosing measurement variables
8. PCA1
   1. PACKAGES:
      1. Tidyverse for ggplot, magriterr, etc
      2. Readxl for reading in data
      3. Extrafont to make plots in times new roman
      4. Flextable for flextable() with custom function (allows times new roman)
      5. Writexl for write\_xlxs function
      6. GGfortify for autoplot function
      7. Scales for percent function
   2. Loaded headscan\_full1 with visual outliers as NA values
   3. Created PCAdata\_num with only chosen numeric variables
      1. Chosen from choosing measure vars docx
   4. Created PCAdata\_full will all variables
      1. Chosen from choosing measure vars docx
   5. Loaded measureNAs dataset
   6. Generated table of NA values in chosen vars
   7. NEED plot of NA values
   8. Dropped NA values from measurement vars, added demographic variables back in
      1. PCAdata\_full dataframe
   9. Removed ID column from PCAdata\_num and PCAdata\_full
   10. Following link in R code, ran PCA:
       1. PCA data was scaled using prcomp(data, scale=TRUE)
       2. PCA data were rotated by -1
       3. Variance explained was gathered using equation:
          1. SD^2/sum(SD^2)
   11. Scree plot generated using qplot
   12. Variance explained manipulated into dataframe
   13. Scree plot generated using ggplot
   14. Manipulated x variable of PCA output to match PCA2.Rmd
       1. Plots were same, but mirrored and rotated
       2. Went with PCA2 as better due to
          1. more prevalence of that process online
          2. the ability to use that process with ggplot (and make ellipses)
   15. Used autoplot to generate plots of pc1 and pc2
       1. No color coding
       2. Race/eth
       3. Race/eth Loadings displayed
       4. Gender
       5. Gender loadings displayed
       6. Age group
       7. Age group loadings displayed
   16. ***Created PCAdata\_full excel***
       1. ***12 chosen measurement variables in mm***
       2. ***Gender, race\_eth, age all categories and NA values retained***
       3. ***NA values dropped (entire row)***
       4. ***1707 rows (used for plots above)***
9. PCA demographics
   1. PACKAGES:
      1. Tidyverse for ggplot, magriterr, etc
      2. Readxl for reading in data
      3. Extrafont to make plots in times new roman
      4. Flextable for flextable() with custom function (allows times new roman)
      5. Writexl for write\_xlxs function
      6. Forcats for fct\_reorder
   2. Loaded PCAdata\_full
   3. Generated histogram of race\_eth frequency
   4. Generated histogram of gender frequency
   5. Generated histogram of age group frequency
10. PCA2
    1. PACKAGES:
       1. Tidyverse for ggplot, magriterr, etc
       2. Readxl for reading in data
       3. Extrafont to make plots in times new roman
       4. Flextable for flextable() with custom function (allows times new roman)
       5. Fauxnaif for na\_if\_in
       6. Reshape2 for melt
       7. FactoMineR for PCA
    2. Centered and scaled data (function found online, link in code)
    3. Created PCA columns in dataframe
    4. Manipulated a couple of columns for data plotting
    5. Used ggplot to generate plots of pc1 and pc2
       1. No color coding
       2. Race/eth
       3. Race/eth with ellipses all categories
       4. Race/eth with ellipses four most common race/eth categories with other shown
       5. Race/eth with ellipses four most common race/eth categories with other not shown
       6. Black, Asian, other with ellipses
       7. Black and Asian with ellipses
       8. Gender
       9. Gender with ellipses all categories
       10. Gender with ellipses two most common gender categories with other shown
       11. Gender with ellipses two most common gender categories with other not shown
       12. Age
       13. Age with ellipses all categories
11. Imputing for PCA
12. PCA-imputed
13. PCA2-imputed
14. Removing statistical outliers (for MANOVA)
    1. PACKAGES
       1. Tidyverse for ggplot, magriterr, etc
       2. Readxl for reading in data
       3. Extrafont to make plots in times new roman
       4. Flextable for flextable() with custom function (allows times new roman)
       5. Writexl for write\_xlxs
       6. rstatix for identify\_outliers
    2. Loaded headscan\_full, with NAs and visual outliers
    3. Created SA3 dataframe with only 12 chosen measurement variables
    4. Generated table of proportion and count of NA values
    5. Rows with NA values were dropped (892 NA values, 306 rows dropped)
    6. Identified univariate outliers using identify\_outliers
       1. Univariate outliers identified overall, and for race/eth, gender,and age
       2. Generated tables that indicate outlier status of overall, and for each factor
       3. Generated bar charts of measurement distribution for all 12 measurements
    7. Identified multivariate outliers using mahalanobis and p values
    8. Added column indicating offending univariate outliers
    9. Created combin\_out dataframe combining indication of uni and multi outliers
    10. ***Created combin\_out1 excel***
        1. ***multivariate outlier status (True, False, not outlier)***
        2. ***univariate extreme overall outlier (True, False, not outlier)***
        3. ***race\_eth extreme outlier (True, False, not outlier)***
        4. ***gender extreme outlier (True, False, not outlier)***
        5. ***age group extreme outlier (True, False, not outlier)***
        6. ***offending univariate measurement (as applicable)***
        7. Generated table of all outlier IDs indicating all of above
    11. Removed 19 multivariate outliers from SA3\_data
        1. RETAINED 4 UNIVARIATE OUTLIERS, not worth deleting the full row of data
    12. ***Created SA3\_noout excel***
        1. ***Removed 19 outliers from full dataset AFTER na rows removed (1710)***
        2. ***12 chosen variables***
        3. ***All factor variables***
        4. ***ID variable***
15. compare these outliers and visual outliers
    1. PACKAGES:
       1. Tidyverse for ggplot, magriterr, etc
       2. Readxl for reading in data
       3. Extrafont to make plots in times new roman
       4. Flextable for flextable() with custom function (allows times new roman)
    2. Loaded all\_vis\_out1 (vis\_outliers) and combin\_out1 (stat\_outliers)
    3. Filtered visual outliers to only 12 chosen measurements
    4. Joined visual and stat outlier dataframes
    5. Manipulated dataframe to have ID’s not repeat
    6. Generated table that indicates visual, multivariate, and univariate outlier status (and offending measures)
16. MANOVA demographics
    1. PACKAGES:
       1. Tidyverse for ggplot, magriterr, etc
       2. Readxl for reading in data
       3. Extrafont to make plots in times new roman
       4. Flextable for flextable() with custom function (allows times new roman)
       5. Writexl for write\_xlxs function
       6. Forcats for fct\_reorder
    2. Loaded SA3\_noout
    3. Generated histogram of race\_eth frequency
    4. Generated histogram of gender frequency
    5. Generated histogram of age group frequency
17. MANOVA assumptions
18. MANOVA equalcovars
19. Running MANOVA
20. Imputing for MANOVA
    1. Started with imputed\_data