Data 630 9040

Machine Learning 2215

Professor Bati Firdu

Melissa Hunfalvay

Date: 8-3-2021

Assignment 5

Introduction

Objective

The dataset used for this project was an eye tracking dataset from RightEye, LLC (www.RightEye.com). Patients were asked to sit in front of a computer and watch stimuli move on the screen. One example is watching a dot move around in a circle. During this time their eyes were being tracked using a Tobii I15 device (Holmqvist & Nystrom, 2011; see Appendix A). The data from the eye tracker is recorded based on the x and y positions of the eyes (Figure 1). Data is then saved and calculated using various algorithms that comparing eye and stimuli positions on the screen. Demographic variables are also collected and appear in the dataset.

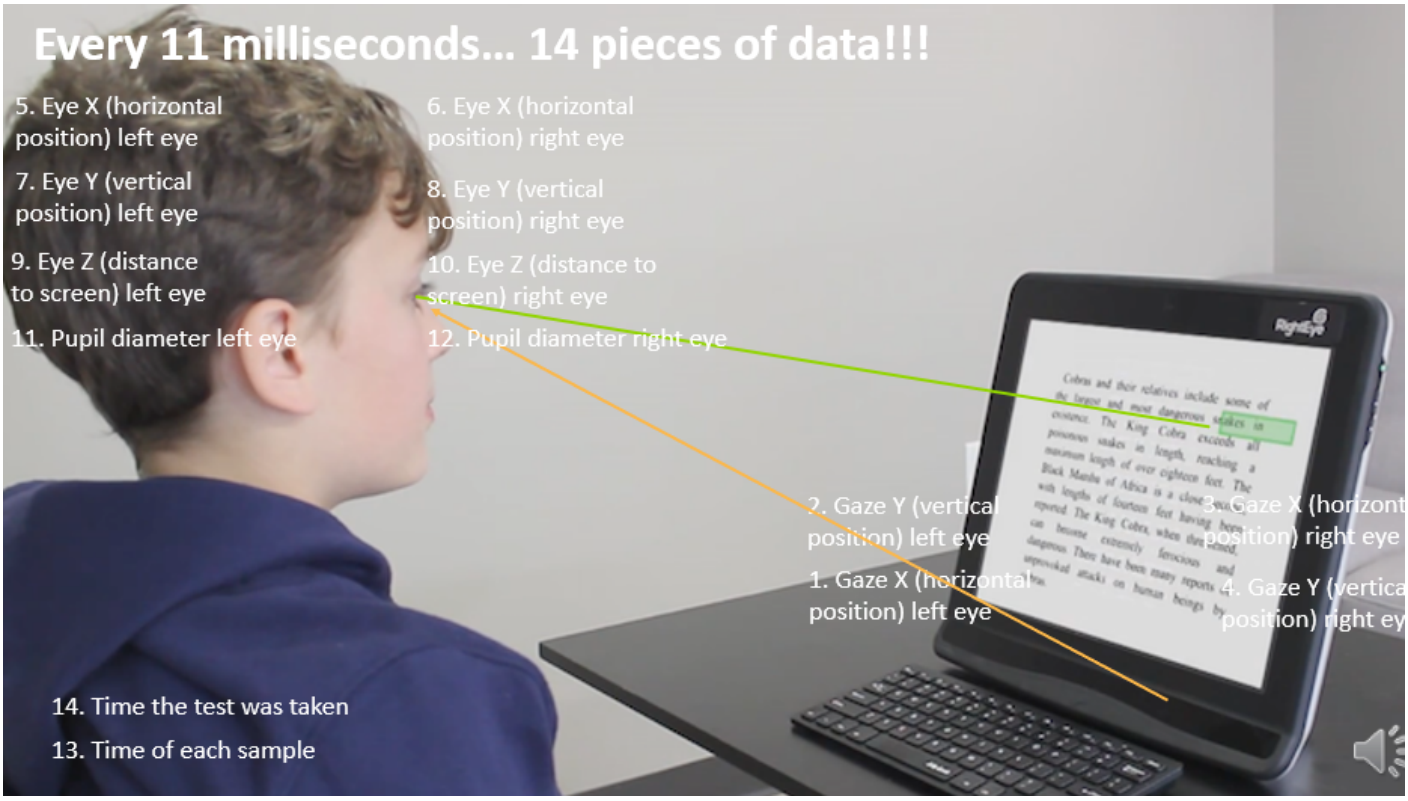


Figure 1: Shows the use and data output from the eye tracker

The objective of the analysis was to explore the variables that group together (cluster) to find similarities and relationships within the dataset. More specifically, K-Means cluster analysis

is a method of unsupervised learning, the objective is to sort the items into groups by similarity. For this dataset, the objective was to compare how the method clustered the data with age. Understanding how eye movements change over the lifespan is critical in determining thresholds of normality for baseline comparisons of peoples with suspected injury, clinical condition and for elite levels of performance (Murray, Hunfalvay, & Bolte, 2017; Lange, Hunfalvay, Murray, Roberts, & Bolte, 2018).

Unsupervised learning is used find insights into the data without using a target or ground truth. Unlike supervised learning, unsupervised learning is usually more exploratory in nature, with the objective of understanding information about the data such as trends, patterns, and groupings (Ng, 2021).

Problem Domain

Eyes are often said to be windows to the soul. The oculomotor system is an indicator of the neurological status of an individual, as each movement of our eyes can be mapped to locations within the brain (Figure 2). Therefore, in a non-invasive fashion, measuring eye movements is also measuring (and mapping) brain function.

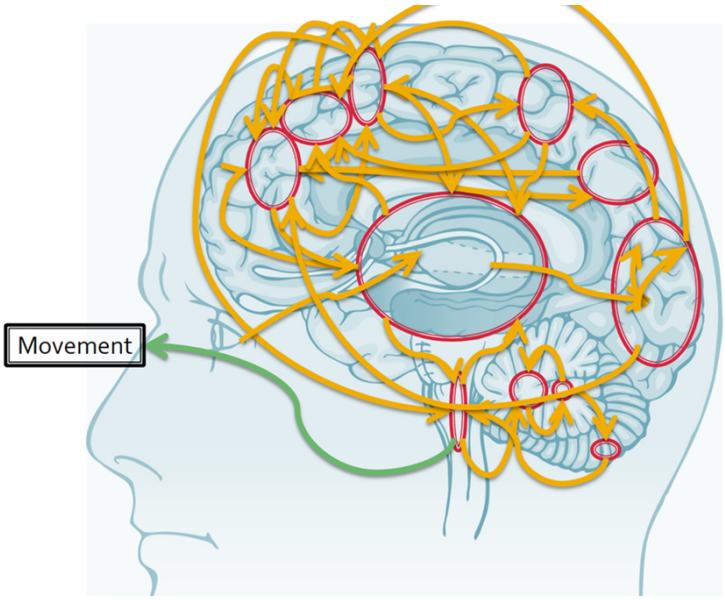


Figure 2: Eye movements mapped to brain locations

Our eyes can provide information on our state of health including early predictors of disease (e.g. Parkinsonism; Gitchel, Wetzel, & Baron, 2012). They can even assist in providing a more complete picture of elite performers such as athletes and military personnel (Hunfalvay, Orr, Murray, & Roberts, 2017).

Oculomotor behavior is measured using an eye tracker. The eye tracker used in this dataset was a Tobii I15 vision 15" monitor fitted with a Tobii 90Hz remote eye tracker (see appendix A). Eye tracking works by using a computer, fitted with an eye tracker that omits near infrared light. The light is reflected by the persons eyes and picked up by the eye tracking cameras. Through this method the system knows where and how the persons eyes are moving.

Limitations in eye tracking research include:

1. Highly specific research studies for certain clinical populations (e.g., concussion research; Ciuffreda, Kapoor, Rutner, Suchoff, Han, & Craig, 2007).

2. Very small sample sizes (e.g. sample size of ten; Naicker, Anoopkumar-Dukie, Grant, & Kavanagh, 2017).

3. Examining elite performance (e.g.: elite golfers; Hunfalvay, Orr, Murray, & Roberts, 2017; various sports, Tenebaum, 2003; various sports, Williams & Ward, 2003).

Furthermore, practical, real-world challenges whereby participants come to be tested who have never been tested previously, therefore they do not have an individual baseline for comparative purposes. They make have been referred for testing because of a suspected injury, such as a concussion. Or, they may be part of a very large organization where there is no practical way to test everyone ahead of time. Yet later testing and analysis of results are needed, perhaps as a way to select a person for a certain elite program, one such example is the U.S. Military recruitment program for Special Forrces. If there are no age-based comparisons, there is no way for a clinician to examine if this person's eye movements are in a state of normality; are clinically disrupted or are in fact, elite.

To date, we know relatively little about the movements of human eyes across the lifespan in non-clinical or elite populations. Without this knowledge, scientists and clinicians have no baselines for comparison and are at a loss as to some basic questions such as:

What should be considered "normal" and appropriate eyes movements?

What should eye movements look like in younger versus older persons? Are they the same? Do they differ?

Do all the eye movements differ or only some? Which ones? And are the differences significant?

Are other variables such as gender, handedness or ethnicity important considerations impacting eye movements across the lifespan?

Therefore, the purpose of this analysis is to examine a large volume of eye movement data across a "normal" population to determine if there are similarities in eye movement variables that cluster together as we age.

Method Rationale

The main methodology chosen was a K-means cluster analysis, which is an unsupervised classification technique.

The rationale for *unsupervised learning* methodology includes:

a) The exploratory nature of the analysis

b) Lack of a known outcome or "right answer"

The rationale for using clustering includes:

a) The research question is quantitative in nature requiring a tool for quantitative modeling

b) The need to explore similarities within the data set

c) The need to explore how differences drive the clustering

d) The research question is looking to identify the relationship among a set of variables or patterns in the data (Maimon & Rokach, 2010).

e) The nature of the problem is to inform RightEye, its customers, clinicians, patients, and researchers, as to a baseline for normal (non-clinical) eye movements over the lifespan. Therefore, a clustering model, identifying age was built to compare the cluster results with the actual age later in the process.

RightEye, LLC., customers, clinicians, patients, and researchers all want to ask questions of a dataset like this to include:

1. What do normal eye movement look like?
2. What should this patient be compared to? For example, can a 17-year-old be compared to other 17-year-olds? Is age a factor for similarity?
3. In contrast, who should that 17-year-old not be compared to? Can s/he be compared to 22-year-olds? Or 25-year-olds? When do the differences in age create different levels of similarity (i.e. different clusters)?
4. What are the primary factors contributing to normality? There are many variables. Are they all equally distant within a cluster? If not, what should the clinician pay most attention to when examining the eye movements of the patient in front of them?
5. By understanding these factors, decisions can be made as to what to do next. Next steps may include:
    a. No further action as the eye movements are normal; or
    b. The need for rest and not to "go back in the game" for 72 hours to let the brain heal from a head injury; or
    c. The need for a follow-up visit; or
    d. Medication, eyeglasses, or vision therapy; or
    e. A referral for surgery or to a specialist like a neurologist

Analysis

Data

This data set was collected by customers with RightEye systems, located throughout the world (Figure 3).

Figure 3: Locations where data was collected

The process for data collection is illustrated in Figure 4 and includes:

Step 1 Test: Participant sits in front of the eye tracking system; watches stimuli and completes testing.

Step 2 Insight platform: The results are set via API's to a cloud based repository where they are stored. Algorithms are then applied to the raw data.

Step 3 Results: Results are then available to the customer in a web-based portal. In the case of the dataset used for this analysis a data pull was conducted to obtain data across multiple participants within the one data sheet.

Figure 4: Process of data collection, storage, report generation and output

Parameters of the data pull included:

1. Only people without clinically identified conditions
2. Only people who had completed all tests
3. Only people who scored above 70 on all tests
4. Only one person per test (i.e. removal of all duplicate identifiers)
5. Only tests from the latest system version (Tobii I15 vision 15" monitor fitted with a Tobii 90Hz remote eye tracker)
6. Removal of known "junk" data such as invalid inputs for age
7. Only if the patient showed a testing distance within required parameters (55-60cm

from the device)

8. Data was pulled from April 2019 (as this was when the I15 system was first implemented in production) to March 2020 (as this was the last time, we undertook a large data pull like this which is costly and time consuming).

The steps conducted prior to the data for this analysis and the reduction in the data can be seen in Figure 5. Therefore, the final csv data file for this analysis included 17,776 rows. These were unique patient test results. There were 137 variables within the dataset.

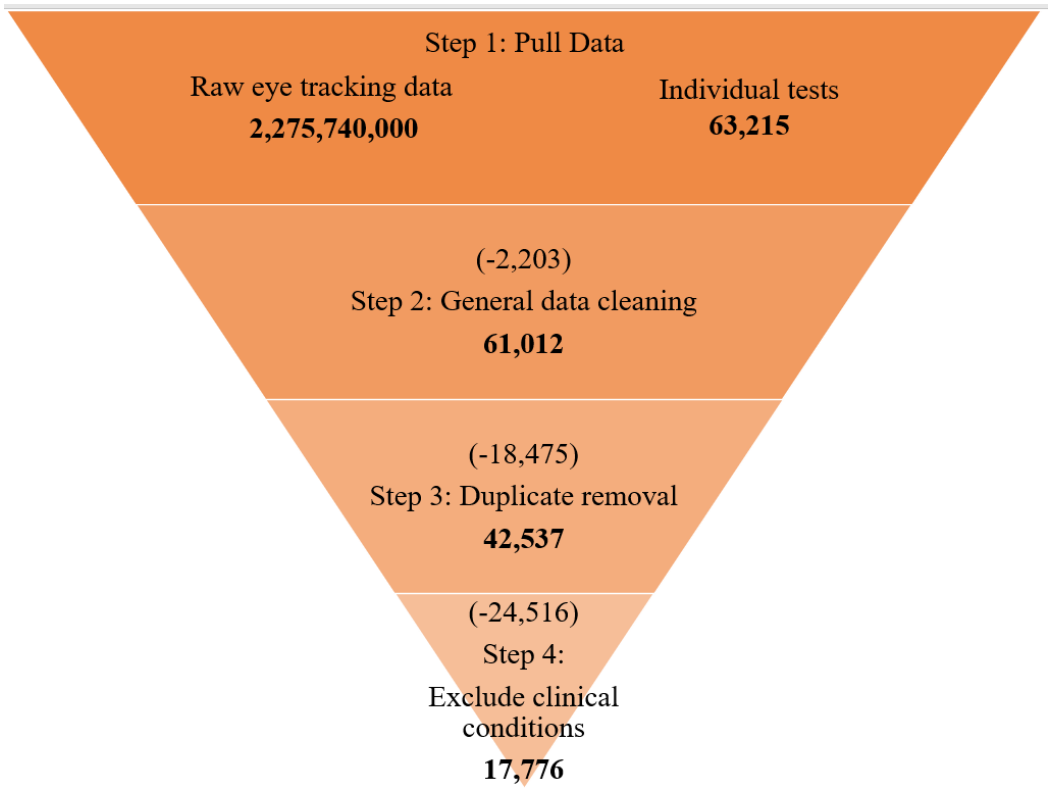


Figure 5: Data reduction and cleaning prior to this analysis

Demographic variables included: age, gender, and handedness.

Eye tracking variables were reported from six difference testing stimuli. These tests were: fixation stability (FS); circular smooth pursuit (CSP); horizontal smooth pursuit (HSP); vertical

smooth pursuit (VSP); horizontal saccades (HS) and vertical saccades (VS). Details of the tests can be found in Appendix C; Murray, Kubitz, Roberts, Hunfalvay, Bolte, Tyagi, 2019; Hunfalvay, Murray, Carrick, 2020).

These tests examined the major eye movements, of which there are three: fixation, pursuits, and saccades. Fixations are stopping points of the eyes and hold the image of a stationary target on the fovea (Komogortsev & Karpov, 2013). Pursuit eye movements occur when the eyes track a moving stimulus to stabilize the image on the fovea, a site of high visual acuity (Barnes, 2008; Duchowski, 2007; Poole & Ball, 2005). Saccades are rapid movements of the fovea between fixation points (Mollenbach, Hansen, Lillholm, 2013).

The three main eye movements are elicited from each of the tests. Then subsequent algorithms further divide the data to reveal additional information about the oculomotor behavior. For example, the speed of the eye and targeting location of the eye compared to the stimuli. Data is also split between left eye, right eye and both eyes (an average between left and right eye locations at a particular time).

Exploratory Analysis and Preprocessing

The exploratory analysis and preprocessing of this data set was highly coupled and iterative. Therefore, these sections were written together for ease of understanding these processes. Furthermore, as the data set was large, global commands were used as a way to examine, at a high level, various visualizations before diving deeper into individual variables.

After the initial data upload to R Studio there were 137 variables, a portion of these variables are shown by the str function (Figure 6). All variables were initially categorized by R as either numeric (num) or factors. There were 17,776 rows (or observations) of data. Each row represents one unique participant.

```
> str(Three_MH_excel_cleaning) # string command
'data.frame':   17776 obs. of  137 variables:
 $ X                                                  : int  2 7 13 14 17 20 22 23 26 27 ...
 $ AGE                                                : num  12 58 7 10 28 22 48 22 12 23 ...
 $ ASSESS_ID                                          : Factor w/ 17776 levels "0000579b-6dc7-d805-c5e6-2a3321c6e3b5",..: 1 2 3 4 5 6 7 8 9 1
0 ...
 $ BLINKAGG_blink_duration_pc                         : num  1.57 4.17 4.49 1.08 11.94 ...
 $ BLINKAGG_blink_num                                 : num  4 65 93 12 63 26 29 40 37 12 ...
 $ BLINKAGG_blink_rate                                : num  NA 1.6 1.64 NA 1.65 3.7 2.72 3.46 2.54 6.55 ...
 $ BLINKAGG_ext_blinks                                : num  0 0 0 0 1 1 1 2 1 1 ...
 $ CALIBRATION_mean_pupil_diameter.both_avg           : num  3.02 3.03 4.1 4.98 4.43 ...
 $ CALIBRATION_pupilary_distance                      : num  55.6 54.9 55.9 62.1 62 ...
 $ CHOICE_RT_REACTION_TIME_processing_speed           : num  851 NA NA NA NA 385 NA 396 NA 438 ...
 $ CHOICE_RT_REACTION_TIME_saccadic_latency.avg       : num  300 NA NA NA NA 309 NA 240 NA 270 ...
 $ CHOICE_RT_REACTION_TIME_visual_reaction_speed      : num  342 NA NA NA NA 62 NA 285 NA 318 ...
 $ DISCRIMINATE_RT_REACTION_TIME_processing_speed     : num  503 NA NA 279 348 241 379 238 354 376 ...
 $ DISCRIMINATE_RT_REACTION_TIME_saccadic_latency.avg : num  304 NA NA 320 255 257 356 254 273 299 ...
 $ DISCRIMINATE_RT_REACTION_TIME_visual_reaction_speed: num  350 NA NA 100 51 81 32 272 45 343 ...
 $ ETHNICITY                                          : Factor w/ 75 levels "","African American and White",..: 66 66 62 66 5 66 66 66 62 66
```

Figure 6: str function output without any pre-processing

Phase 1 cleaning to remove some variables and make the dataset more manageable was conducted after the initial upload of the data to R Studio. Variables were removed for different reasons. For example, identifiers, duplicate columns, partial values, type of eye tracker and other variables were removed as they would not contribute to the model. This reduced the variable size from 137 to 105, a reduction of 32 variables.

Missing data values were checked. Many variables had missing data (See a portion of the output in Figure 7).

```
# Missing Values

# Data Exploration:  Check for missing values
colSums(is.na(Three_MH_excel_cleaning))
```

AGE	0	BLINKAGG_blink_duration_pc	1245
BLINKAGG_blink_num	1279	BLINKAGG_blink_rate	3450
BLINKAGG_ext_blinks	1982	CALIBRATION_mean_pupil_diameter.both_avg	1148
CHOICE_RT_REACTION_TIME_processing_speed	4654	CHOICE_RT_REACTION_TIME_saccadic_latency.avg	4654
CHOICE_RT_REACTION_TIME_visual_reaction_speed	4654	DISCRIMINATE_RT_REACTION_TIME_processing_speed	3081
DISCRIMINATE_RT_REACTION_TIME_saccadic_latency.avg	3081	DISCRIMINATE_RT_REACTION_TIME_visual_reaction_speed	3081
EYEQ_SCORE	0	FIXATION_STABILITY_bcea	298
FIXATION_STABILITY_convergence_point.diff	576	FIXATION_STABILITY_depth	489
FIXATION_STABILITY_fixation_dispersion.both_avg	443	FIXATION_STABILITY_gaze_positions_band1.both_avg	5
FIXATION_STABILITY_gaze_positions_band2.both_avg	1033	FIXATION_STABILITY_gaze_positions_band3.both_avg	2539
FIXATION_STABILITY_gaze_positions_band4.both_avg	3883	FIXATION_STABILITY_gaze_positions_lessthan4deg.both_avg	3441

Figure 7: apply command to check for the missing values within each variable in the data set

Given the number of variables and variety of missing data several steps were taken:

Step 1: Determine the percentage of missing data for each variable (see a portion of this output in Figure 8).

Step 2: Remove variables where the missing data exceeded 4%

```
> # percentage of missing data in each column
> library(dplyr)
> Three_MH_excel_cleaning %>%
+   summarise_each(funs(round(100*mean(is.na(.)))))
  AGE BLINKAGG_blink_duration_pc BLINKAGG_blink_num BLINKAGG_blink_rate BLINKAGG_ext_blinks CALIBRATION_mean_pupil_diameter.both_avg
1   0                          7                 7                  19                 11                                         6
  CHOICE_RT_REACTION_TIME_processing_speed CHOICE_RT_REACTION_TIME_saccadic_latency.avg CHOICE_RT_REACTION_TIME_visual_reaction_speed
1                                       26                                          26                                          26
  DISCRIMINATE_RT_REACTION_TIME_processing_speed DISCRIMINATE_RT_REACTION_TIME_saccadic_latency.avg
1                                             17                                               17
  DISCRIMINATE_RT_REACTION_TIME_visual_reaction_speed EYEQ_SCORE FIXATION_STABILITY_bcea FIXATION_STABILITY_convergence_point.diff
1                                                  17          0                       2                                          3
  FIXATION_STABILITY_depth FIXATION_STABILITY_fixation_dispersion.both_avg FIXATION_STABILITY_gaze_positions_band1.both_avg
1                        3                                               2                                             0
```

Figure 8: Portion of the output of the percentage of missing data per variable

This resulted in the variables being further reduced to 38. Data was then checked again for missing values within the rows using the summary command (see a portion of this output in Figure 9).

```
> #Pre-processing: remove observations with missing values
> summary(Three_MH_excel_cleaning)
      AGE          EYEQ_SCORE     FIXATION_STABILITY_bcea FIXATION_STABILITY_convergence_point.diff FIXATION_STABILITY_depth
 Min.   : 1.00   Min.   : 70.00   Min.   :3.321           Min.   :-234.20                           Min.   :-116.062
 1st Qu.:16.00   1st Qu.: 79.00   1st Qu.:4.405           1st Qu.: -48.44                           1st Qu.: -34.408
 Median :25.00   Median : 86.00   Median :4.908           Median : -16.83                           Median :  -7.565
 Mean   :29.77   Mean   : 85.15   Mean   :5.023           Mean   : -14.26                           Mean   :  -7.482
 3rd Qu.:43.00   3rd Qu.: 92.00   3rd Qu.:5.533           3rd Qu.:  20.36                           3rd Qu.:  20.038
 Max.   :94.00   Max.   :100.00   Max.   :7.617           Max.   : 154.20                           Max.   :  91.306
                                  NA's   :298             NA's   :576                               NA's   :489
 FIXATION_STABILITY_fixation_dispersion.both_avg FIXATION_STABILITY_gaze_positions_band1.both_avg
 Min.   : 1.746                                  Min.   :  0.00
 1st Qu.: 4.720                                  1st Qu.: 43.43
 Median : 6.208                                  Median : 64.78
 Mean   : 6.593                                  Mean   : 61.57
 3rd Qu.: 8.062                                  3rd Qu.: 84.60
 Max.   :16.442                                  Max.   :100.00
 NA's   :443                                     NA's   :5
 FIXATION_STABILITY_targeting_displacement_vertical.both_avg GENDER_female    HORIZONTAL_SACCADES_qratio.both_avg
 Min.   :-2.4150                                             Min.   :0.0000   Min.   :1.565
 1st Qu.:-0.9250                                             1st Qu.:0.0000   1st Qu.:2.073
 Median :-0.5050                                             Median :0.0000   Median :2.249
 Mean   :-0.5093                                             Mean   :0.4984   Mean   :2.303
 3rd Qu.:-0.1000                                             3rd Qu.:1.0000   3rd Qu.:2.493
 Max.   : 1.7000                                             Max.   :1.0000   Max.   :3.322
 NA's   :449                                                                  NA's   :241
```

Figure 9: Shows a portion of the output from the summary command revealing missing data in variables as seen in NA's

The rows with missing data were eliminated (see a portion of this output in Figure 10). This resulted in an initial cleaned dataset of 38 variables and 10,539 observations.

```
> df<-na.omit(Three_MH_excel_cleaning)
> str(df)
'data.frame':   10539 obs. of  38 variables:
 $ AGE                                                  : num  7 10 48 12 23 60 56 56 27 26 ...
 $ EYEQ_SCORE                                           : num  87 73 98 91 98 89 81 90 80 82 ...
 $ FIXATION_STABILITY_bcea                              : num  6.76 4.95 5.07 6.67 5.07 ...
 $ FIXATION_STABILITY_convergence_point.diff            : num  -124.99 8.87 -18.45 -78.84 0.09 ...
 $ FIXATION_STABILITY_depth                             : num  -91.27 23.57 -13.75 -57.25 3.43 ...
 $ FIXATION_STABILITY_fixation_dispersion.both_avg      : num  7.92 7.54 4.22 6.94 4.01 ...
 $ FIXATION_STABILITY_gaze_positions_band1.both_avg     : num  51.9 44 87.7 73.8 88.7 ...
 $ FIXATION_STABILITY_targeting_displacement_vertical.both_avg: num  -0.405 -0.82 -0.535 -0.62 -0.215 -0.
 $ GENDER_female                                        : num  0 0 1 1 0 0 1 0 0 0 ...
 $ HORIZONTAL_SACCADES_qratio.both_avg                  : num  2.2 2.26 2.15 2.46 2.19 ...
 $ HORIZONTAL_SACCADES_saccadic_efficiency_left_side.both_avg : num  5.02 9.13 3.97 5.21 3.1 ...
 $ SPEM_fixation_pc.both_avg                            : num  0 0.0807 0.0391 0.0743 4.859 ...
 $ SPEM_H_eccentric_gaze_mean_left_side.both_avg        : num  -1.432 -1.891 -1.843 1.091 0.288 ...
 $ SPEM_H_eccentric_gaze_mean_middle.both_avg           : num  -1.903 -5.256 -2.657 -1.121 0.892 ..
 $ SPEM_H_eccentric_gaze_mean_right_side.both_avg       : num  -0.4166 -3.517 -1.0191 -2.2053 0.07
 $ SPEM_H_eccentric_gaze_variability_middle.both_avg    : num  3.39 2.32 1.91 2.4 1.27 ...
 $ SPEM_H_eccentric_gaze_variability_right_side.both_avg : num  2.498 2.184 1.572 1.551 0.794 ...
 $ SPEM_H_eye_target_vel_err.both_avg                   : num  19.8 18 18.4 18.7 18.4 ...
```

Figure 10: Shows a portion of the string function after initial cleaning of missing variables

To further explore the data, correlations between variables were examined (See Figure 11). The remaining variables were not found to be correlated and therefore no further-preprocessing was needed due to correlations.

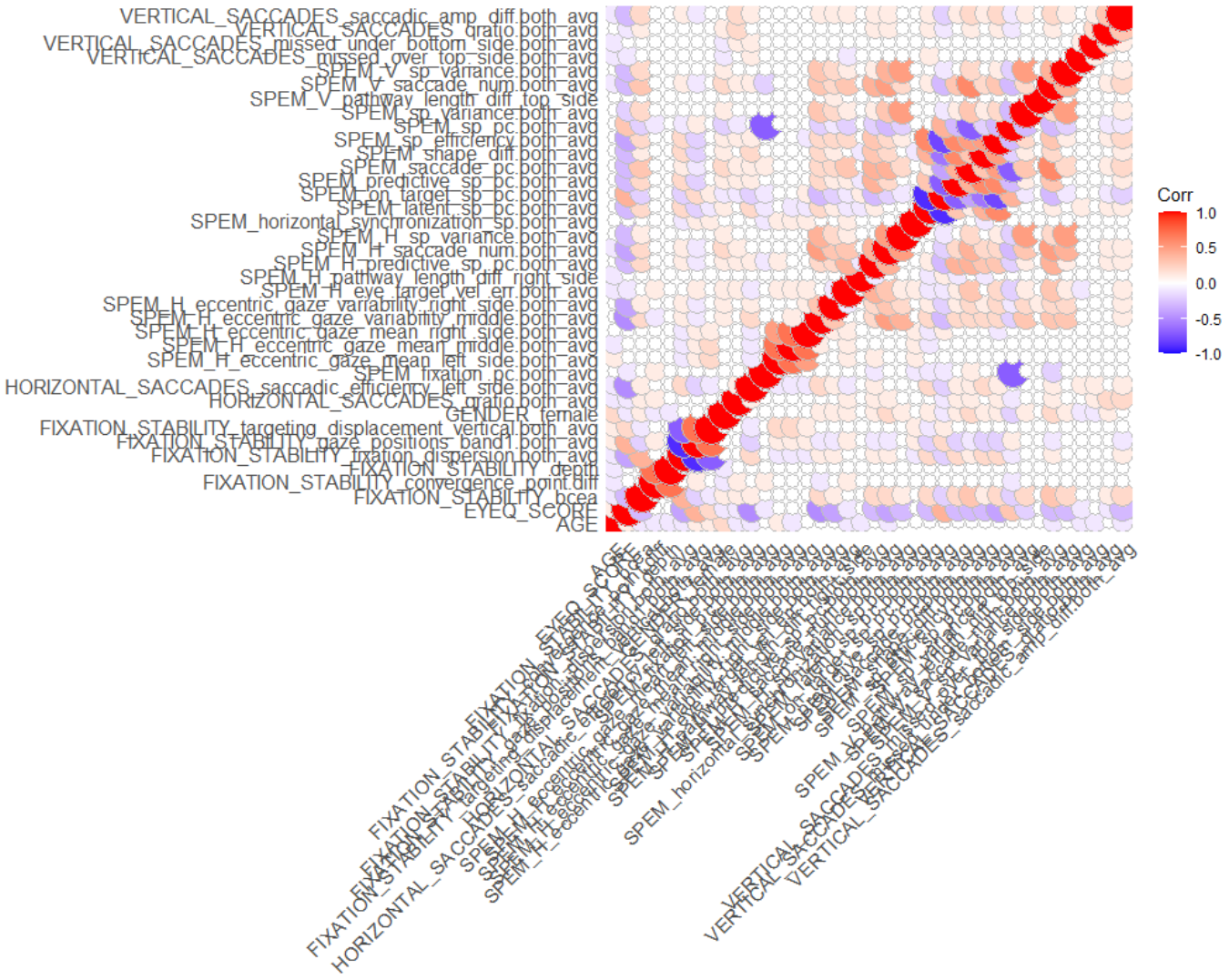


Figure 11: Correlation matrix for remaining 38 variables in the data set

Outliers were examined using a global boxplot command (see Figure 12). After initial review of all the variables for potential outliers, several variables were examined individually.

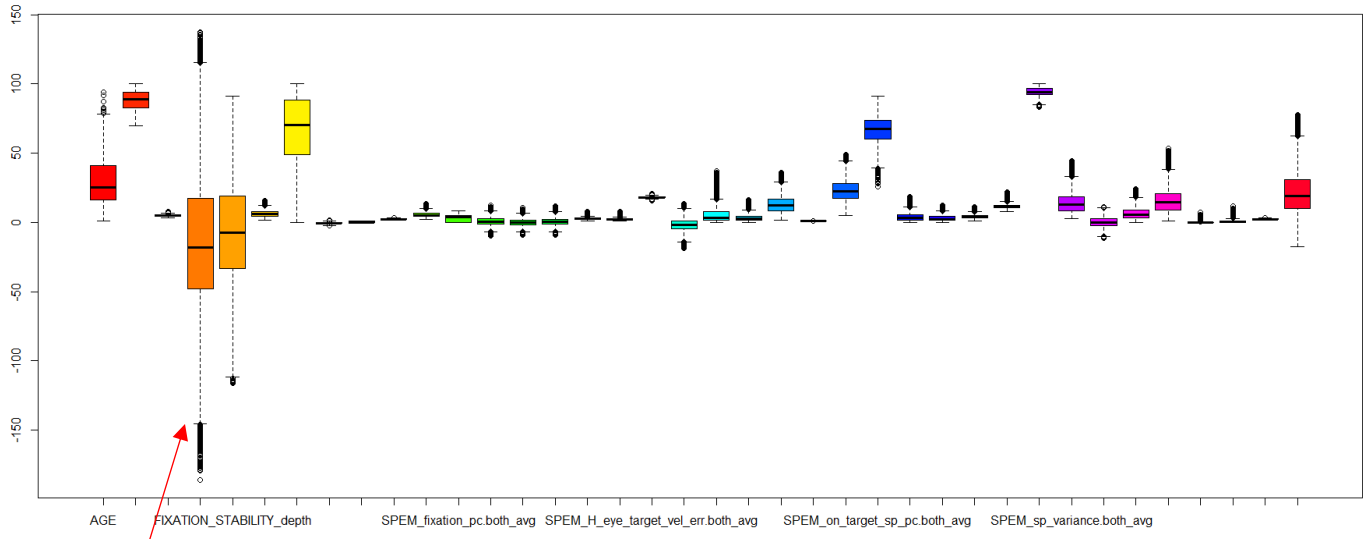


Figure 12: Box plot of 38 variables within the data set

FIXATION_STABILITY_convergence_point.diff was examined with an individual box plot (Figure 13). This variable refers to the location of where the eyes meet (or converge) in relation to the screen. Negative numbers reveal the eye converge before the screen, and positive numbers reveal the eyes converge after the screen. Therefore, after examining this variable in more detail, what looks like outliers are in fact legitimate data and no further processing was needed.

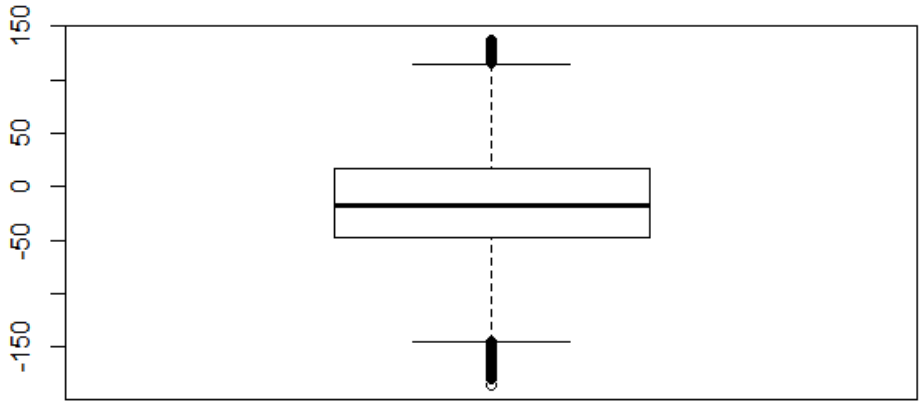


Figure 13: FIXATION_STABILITY_convergence_point.diff boxplot

Initial age range in the data revealed a minimum age of 1 and maximum age of 94 (Figure 14). As it is very difficult to get young children below the age of 5 to reliably attend to the testing stimuli a decision was made to remove those participants (n=3). Furthermore, participants 90 years of age and older (n = 2) were suspicious data and may have been added by Quality Assurance to check for system functioning. Therefore, over 90 years of age were also removed from the dataset. Therefore, the remaining age range was 5 to 87 years of age (see Figure 15).

```
> summary (df$AGE) # Descriptive Statistics
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00   16.00   25.00   29.63   41.00   94.00
```

```
> summary (df$AGE) # Descriptive Statistics
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   5.00   16.00   25.00   29.62   41.00   87.00
```

Figure 14: Summary command for age before pre-processing; Figure 15: After preprocessing

To further explore the data histograms were created for all numeric variables. It is common for eye tracking variables to be skewed (Komogortsev & Karpov, 2013). Some variables were found to be skewed such as VERTICAL_SACCADES_missed_under_bottom_side.both_avg (Figure 16). This variable represents the eyes stopping short of a target when looking down. Some people will have zero for this variable meaning that they precisely target each time, others will have more than zero meaning they stopped short. Each time they stop short a 1 is added to the variable. Therefore, even though this variable is skewed it is accurate and no further processing is needed.

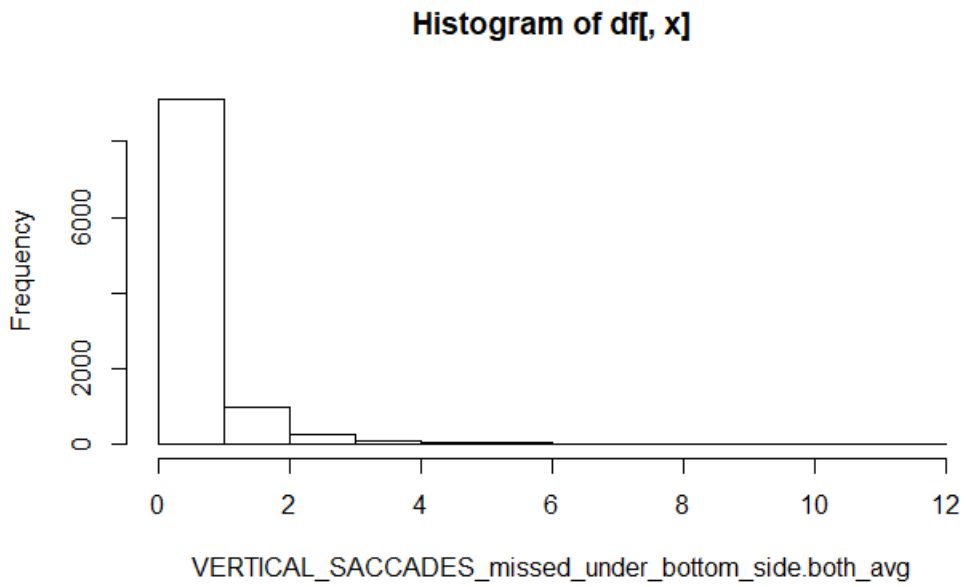


Figure 16: VERTICAL_SACCADES_missed_under_bottom_side.both_avg histogram showing skewness

Other variables were found to be relatively normally distributed. One such example was VERTICAL_SACCADES_saccadic_amp_diff.both_avg which indicates the speed (amplitude) of the eyes when doing a saccade (Figure 17). This distribution is also expected for non-clinical participants and therefore further validates the data, and no further processing is needed.

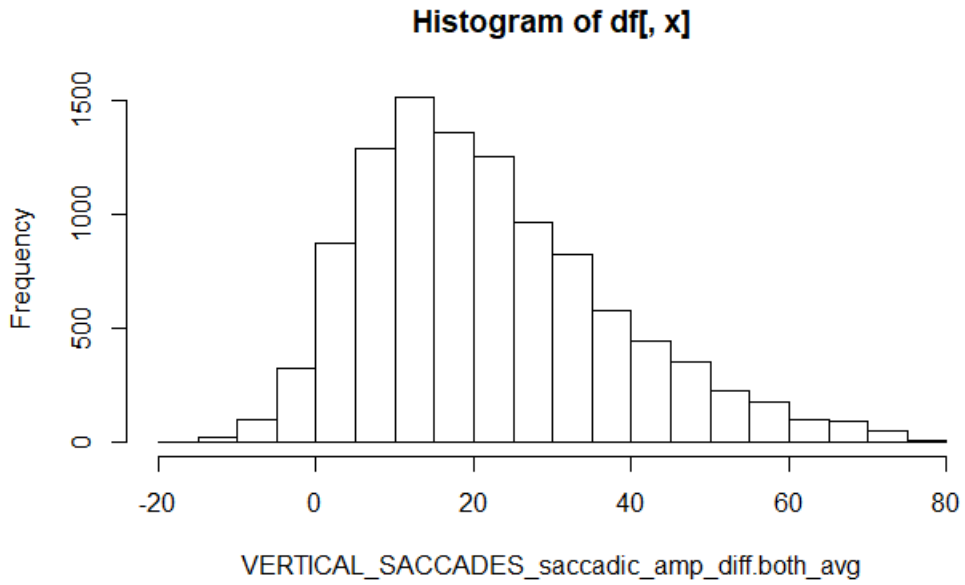


Figure 17: VERTICAL_SACCADES_saccadic_amp_diff.both_avg Histogram showing normal distribution

After the exploratory analysis and pre-processing a final phase of pre-processing was conducted to prepare the data for model generation. This included dropping the age variable, so it does not influence the results during the cluster iterations. The objective of the analysis was to identify age and therefore it is removed while the model is being built and added in later to determine where the ages fit within the clusters.

A second step for preparation of the model was to scale the data in the K-means because the results largely depend on the variability of the variables. If the variables have a large variance, they impact the clustering more than if the variance is low. Therefore, the data was scaled to avoid the issue of a few variables with large variances dominating the result (the clustering). Scaling normalizes the data prior to the analysis by using the sum of the sum of squares for each variable.

A final check of the data was conducted using the head, summary, view, and string commands. In summary, after pre-processing was complete there remained 10,534 observations, 7,242 observations were removed. Thirty-seven variables remained; 100 variables were removed. The gender was transformed from a numeric to a factor variable. A portion of the string command can be seen below in Figure 18.

```
> str(df)
'data.frame':   10534 obs. of  37 variables:
 $ EYEQ_SCORE                                          : num [1:10534, 1] -0.107 -2.045 1.416 0.447 1.416 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr  "3" "4" "7" "9" ...
  .. ..$ : chr "EYEQ_SCORE"
  ..- attr(*, "scaled:center")= Named num 87.8
  .. ..- attr(*, "names")= chr "EYEQ_SCORE"
  ..- attr(*, "scaled:scale")= Named num 7.22
  .. ..- attr(*, "names")= chr "EYEQ_SCORE"
 $ FIXATION_STABILITY_bcea                             : num [1:10534, 1] 2.3527 0.0727 0.2241 2.2406 0.2185 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr  "3" "4" "7" "9" ...
  .. ..$ : chr "FIXATION_STABILITY_bcea"
  ..- attr(*, "scaled:center")= Named num 4.9
  .. ..- attr(*, "names")= chr "FIXATION_STABILITY_bcea"
  ..- attr(*, "scaled:scale")= Named num 0.794
  .. ..- attr(*, "names")= chr "FIXATION_STABILITY_bcea"
 $ FIXATION_STABILITY_convergence_point.diff           : num [1:10534, 1] -2.1682 0.5046 -0.0409 -1.2467 0.3293 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr  "3" "4" "7" "9" ...
  .. ..$ : chr "FIXATION_STABILITY_convergence_point.diff"
  ..- attr(*, "scaled:center")= Named num -16.4
  .. ..- attr(*, "names")= chr "FIXATION_STABILITY_convergence_point.diff"
  ..- attr(*, "scaled:scale")= Named num 50.1
  .. ..- attr(*, "names")= chr "FIXATION_STABILITY_convergence_point.diff"
```

Figure 18: String command showing changes in variables, observations, and data categories after pre-processing

Algorithm Intuition

Clustering is a process of finding groups with similar attributes whilst also differentiating groups that differ from one another.

The overarching principle of clustering is to find a partition (or divide) in the data with preset number of groups. Then to minimize the variation within each group (cluster) and to maximize the inter cluster distances. The variation within each group was measured by the sum of the sum of squares once the data has been normalized.

Cluster analysis is generally considered exploratory and is useful in large data sets specially to help understand relationships within the data d across the variables that are not "visible" via other means, such as plotting histograms of one variable at a time.

There are various types of clustering algorithms such as hierarchical clustering, which is a nested cluster; and k-means clustering which is partitional. There are also various ways to measure distance-based similarities within clusters for example, single linkage, average, complete linkage, and centroid. This discussion will focus on K-mean clustering whereby similarity is defined as the distance a data point has from the center of a cluster.

The objective of K-means clustering is to find the separation (or partition) of the observations into a predefined number of groups in which the within group variation is minimal. Groups are identified as *k*. The formula for the minimization within each group is showing in Figure 19, whereby the difference between the value of the variable (x) and the mean value within a cluster (centroid; c) is then squared. This is repeated for every feature within a cluster. The process is then done for every cluster and total is the value that is trying to be minimized in this process.

$$minimize\left\{\sum_{i=1}^{k}\sum_{j=1}^{p}(x_{ij}-c_i)^2\right\}$$

Figure 19: Formula for the minimization within each group

To find the variation of a single variable within a single cluster (Figure 20) is calculated by taking the difference between that variable (x) and the mean value within that cluster. Then, dividing that value by the number of observations in that group.

$$s_{j(g)}^2=\sum_{i=1}^{n_g}\frac{\left(x_{ij}-\bar{x}_j\right)^2}{n_g-1}$$

Figure 20: Formula for the variation of a single variable within a single cluster

To measure the overall variation within each group a combination of the individual group variation is calculated (s^2). Therefore, all input variables for k-means clustering must be numeric values.

The k-means algorithm is iterative, taking the following steps:

Step 1: Define the number of clusters (k values) or centroids.

Step 2: Assign observations to the closest cluster

Step 3: Calculate the centroid based on the newly assigned observations.

Step 4: Repeat steps 2 and 3 until the observations and centroid location no longer change. At which point the output will reveal the final cluster centers and assignments of the training data.

As the number of clusters are pre-defined, a method of evaluating if the number of clusters are optimal for the dataset a method known as "*the elbow method*" may be used (Figure 21).

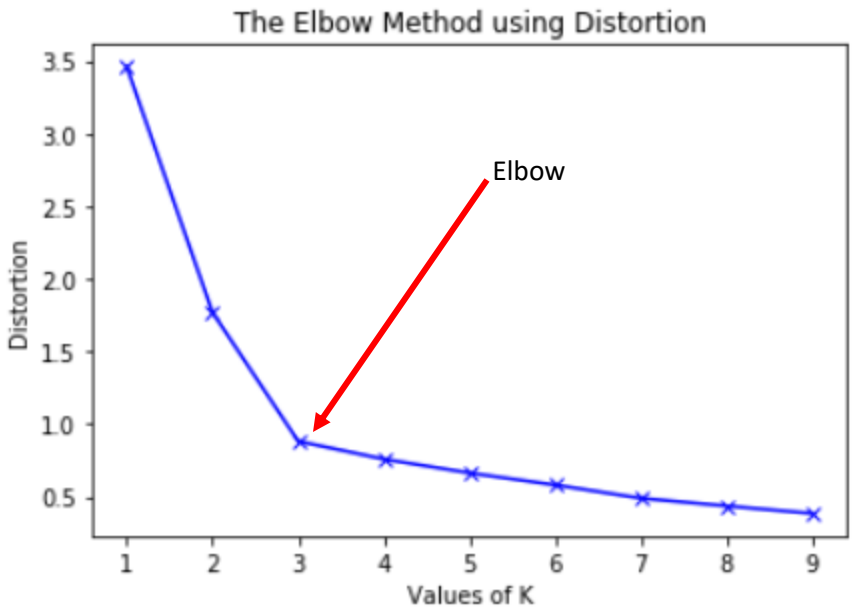


Figure 21: Plot of the sum of squares as k values increase. The elbow can be seen at the third cluster in this example.

The elbow method is a direct method that evaluates the percentage of variance explained as a function of the number of clusters (k). Typically, an increase in the number of clusters leads to a decrease in the sum of squares. The evaluation of the number of clusters can be made by looking for the rate at which the reduction of sum of squares in negligible, causing an "elbow" when graphed. At this point the increase in clusters gives no further reduction to the sum of squares and therefore indicating the value of k. Within sum of square formula is shown in Figure 22 (https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/).

$$\boldsymbol{wss} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{k}} \sum_{\boldsymbol{j=1}}^{\boldsymbol{n_i}} (\boldsymbol{x_{ij}} - \boldsymbol{c_i})^{\boldsymbol{2}}$$

Figure 22: Formula for the within sum of squares that can be plotted to find the optimal k value using the elbow method.

Model Fitting

The key steps used to fit the model were:

Step 1: To make sure the results were reproducible by using the set.seed command

Step 2: To identify an initial number of clusters (k values). An arbitrary starting point was 4 clusters.

Step 3: Build the model using the kmeans method (Figure 23)

```
> #Run the method
> kc<-kmeans(df, 4)
```

Figure 23: Shows code to run the kc method

Step 4: Print and inspect the model from the first iteration. Pay special attention the total within-ness and between-ness and the total sum of squares output.

Step 5: Build the cross-tabulation to compare how the method clustered age with the actual age class. Bin the ages into initial groups due to the number of unique age values. Print and inspect the tabulations especially looking for dominance within the clusters.

Step 6: Create a new data frame called "results" that shows each observation and which cluster it was assigned to. Examine the data for trends and inconsistencies.

Step 7: Due to the amount of data examine with different clustering plots to include colors and numbers of variables.

Step 8: Visualize the sum of squares and number of clusters to examine the elbow method to guide the next iteration of k values.

Step 9: Repeat the process using the following evaluation guidance:

a) Are the clusters separated in at least some of the clusters in the cluster plot?

b) Are there any clusters with only a few data points? If yes, decrease the value of k

c) Are the splits on the variables what is expected? If yes, increase the value of k.

d) Do any of the centroids seem to close together? If yes, decrease the value of k.

e) Check for outliers using anomaly detection, specifically:

    a. Inspect the outliers

    b. Use the method results to calculate the distance between each age and its corresponding cluster

    c. Identify the top ages with dissimilarity and consider removing them from the data set

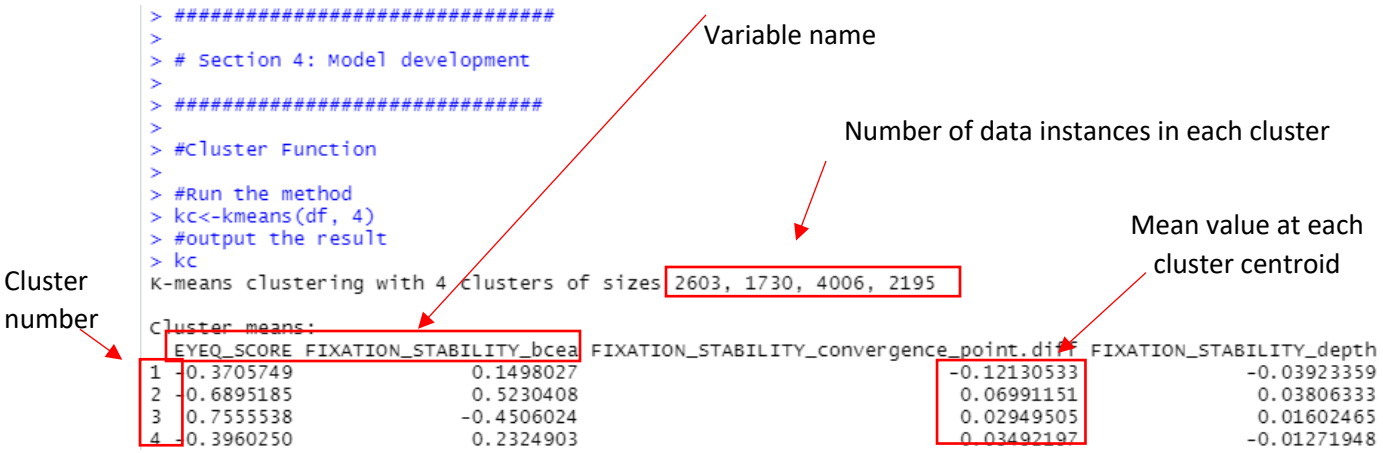
Step 10: Final evaluation of the model using decision making logic as it pertains to the objective or question being asked of the data.

Step 17: Do a final summary and inspection of the model in relation to the stated objectives

Results

Output

The first iteration of the model was pre-set with four clusters. Results of the textual output for the summary model can be seen in Figure 23. There were eight iterations to form this output.




```
> ################################
>
> # Section 4: Model development
>
> ################################
>
> #Cluster Function
>
> #Run the method
> kc<-kmeans(df, 4)
> #output the result
> kc
K-means clustering with 4 clusters of sizes 2603, 1730, 4006, 2195

Cluster means:
  EYEQ_SCORE FIXATION_STABILITY_bcea FIXATION_STABILITY_convergence_point.diff FIXATION_STABILITY_depth
1 -0.3705749              0.1498027                            -0.12130533            -0.03923359
2 -0.6895185              0.5230408                             0.06991151             0.03806333
3  0.7555538             -0.4506024                             0.02949505             0.01602465
4 -0.3960250              0.2324903                             0.03492197            -0.01271948
```

Figure 23: First iteration of the model output

Figure 24 shows the sum of square between clusters. The between sum of squares clusters is the sum of squares distance between the instance and the center of the clusters an instance does *not* belong to. Unlike the within sum of squares which shows the sum of squares distance between the instance and the center of the clusters an instance belongs to.

```
Within cluster sum of squares by cluster:
[1] 83906.10 71027.65 89397.19 69556.06
 (between_SS / total_SS =  17.8 %)
```

Figure 24: Within cluster sum of squares for each of the 4 clusters

Total sum of squares is the sum of squares within a cluster plus the sum of squares between clusters and remains the same irrespective of the number of clusters.

```
> kc$betweenss
[1] 67933.33
> kc$tot.withinss
[1] 313887
> kc$totss
[1] 381820.3
> kc$iter
[1] 8
```

Figure 25: Between cluster sum of squares, total within cluster sum of squares, total sum of squares and number of iterations taken to create the model.

Age of the participants was used to build the cross-tabulation to compare how the method clusters the ages with the actual age class (Figure 26). As there were too many unique age values to create the cross-tabulation ages were grouped (bined). Grouping of the age variable, for the purposes of determining actual age class, were determined based on past research. More specifically, a prior version of the age-based clusters was used as a precedent for evaluation (Hunfalvay & Murray, 2019). Age groups were 0-11, 11-17, 17-29, 29-42, 42-56, 56-65, 65-90 years of age (Figure 26).

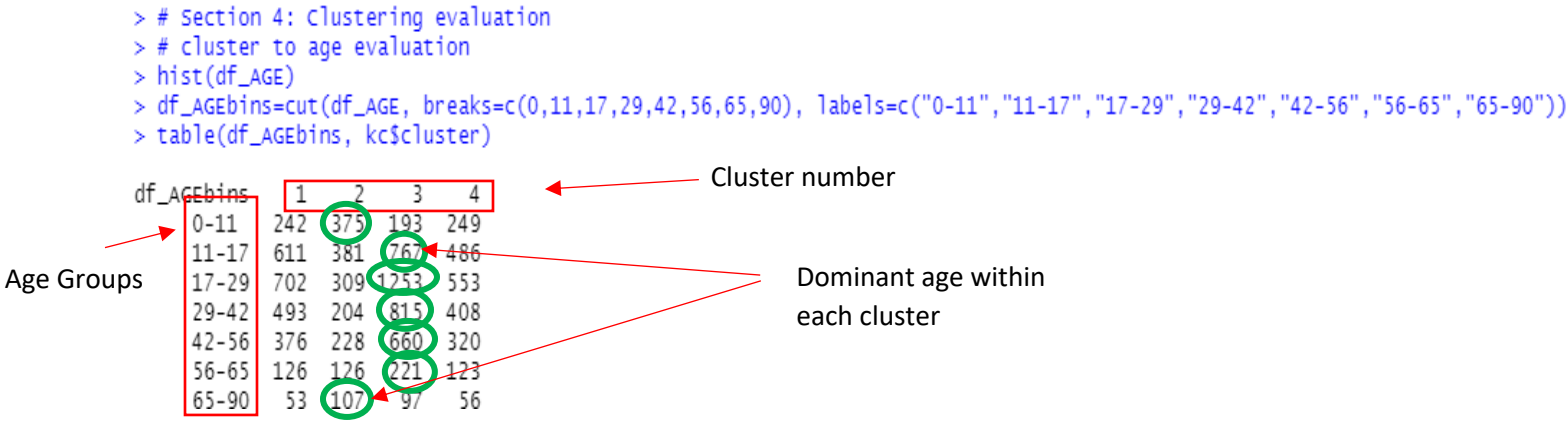


Figure 26: Cross-tabulation of age group with cluster number, showing dominant age within each cluster

A cluster plot was created for the data to visually examine the clusters, specifically, the withinness and betweenness (Figure 27). There are four clusters. Each cluster is numbered and has a different color code.

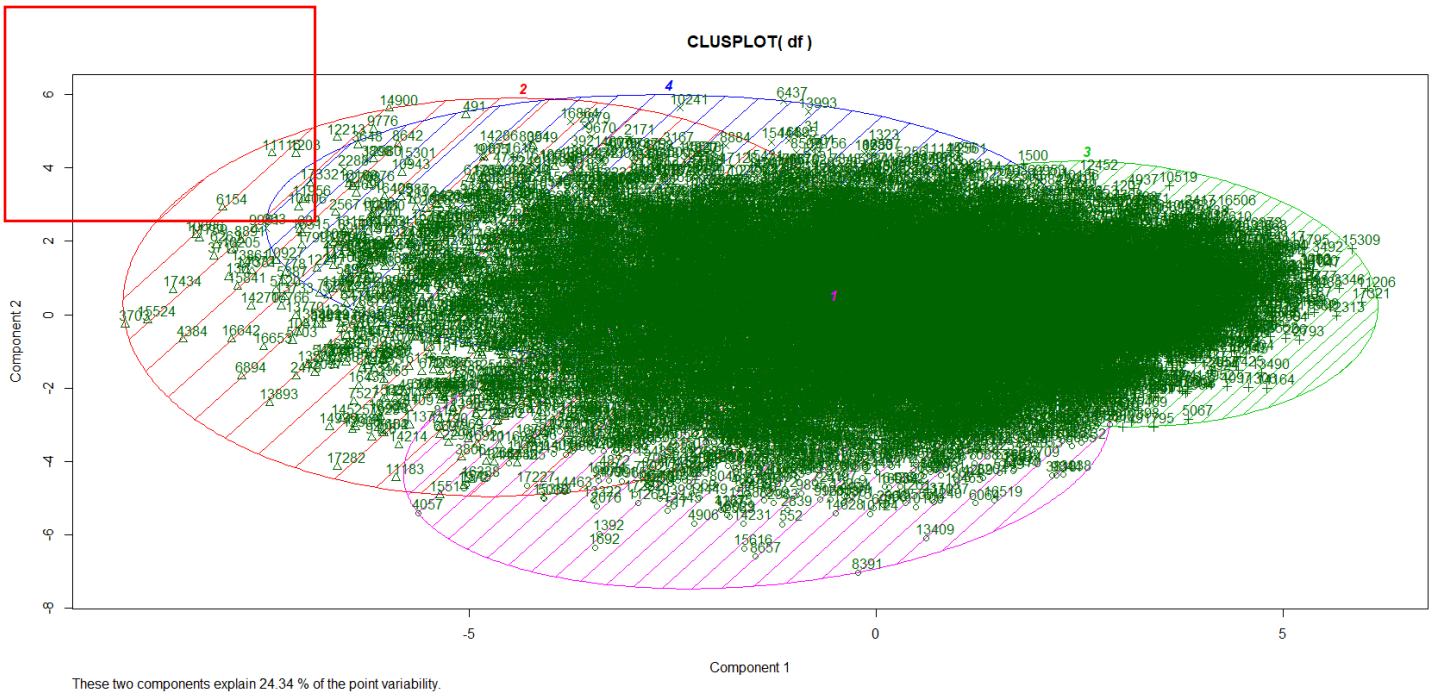


Figure 27: Cluster plot showing four clusters

Numbers within the clusters correspond to the instance numbers in the dataset. For example, 6154 is in the second cluster as is 1111. Figure 28 shows a specific portion of the cluster plot where these values fall.

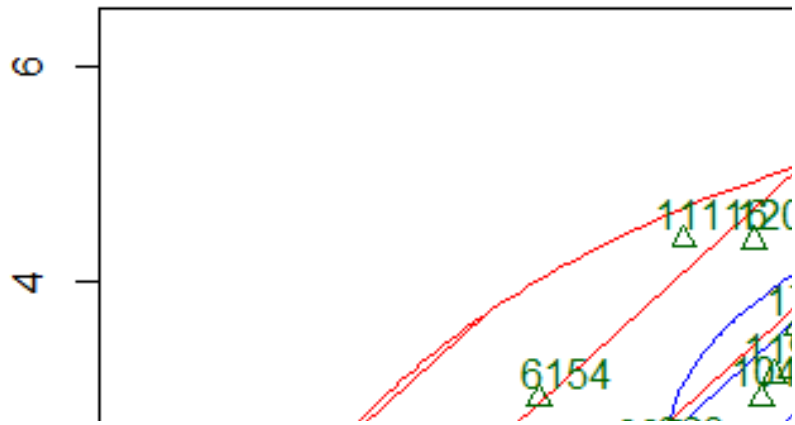


Figure 28: Portion of the cluster plot from Figure 27

The cluster plot shows that the cluster overlap. The goal is to minimize the sum of squared distances within the cluster (creating a tight circle) while maximizing the sum of squared

distances between clusters. That is; showing distinct circles that visually show distances from one another and are not overlapping.

Additional exploration of the dataset to determine the optimal number of clusters were examined by looking at betweenness and withinness values and are seen in table 1.

Table 1: Exploration of number of clusters

Number of Clusters	Betweenness	Withinness
11	102,875	278,945
10	99,688	282,131
9	94,980	286,839
12	103,621	278,199
20	121,571	260,248

After each iteration and exploration, the cross-tabulation was also examined. The optimal number of clusters use the code in Figure 29 was also run. However, due to the amount of data the code never completed processing in R Studio and therefore was unable to be used to provide additional insights or guidance.

```
#install.packages("fpc")
library(fpc)
best<-pamk(df) # shows optimal number of clusters
clusplot(df,best$pamobject[["clustering"]] , color=T)
plotcluster(df,best$pamobject[["clustering"]] )
```

Figure 29: Code to examine the optimal number of clusters

The elbow method was employed to evaluate the percentage of variance explained as a function of the number of clusters (k; Figure 30).

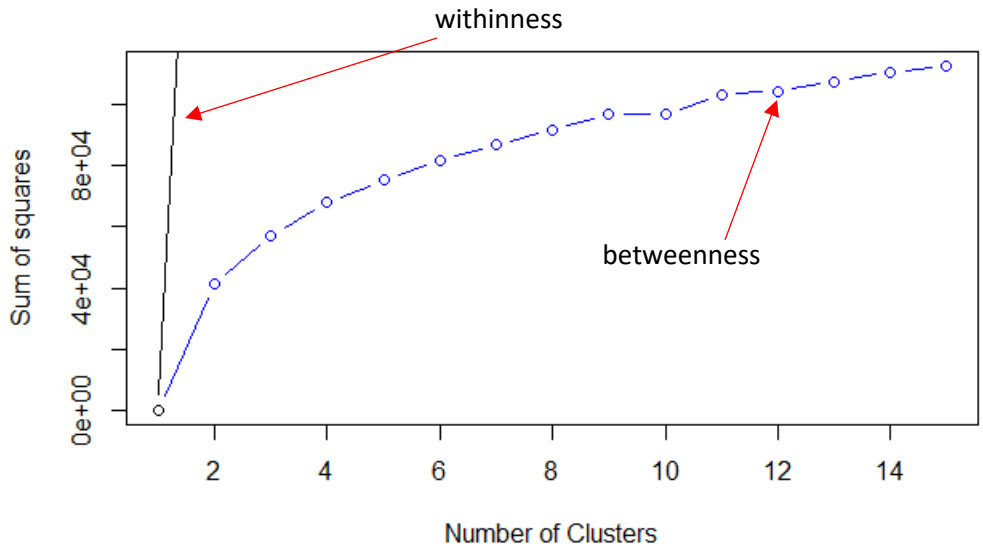


Figure 30: Graphical output of the elbow method for evaluation of clusters based on percentage of variance

After reviewing all these iterations, the total withinness values were much higher than the total betweenness values. This indicates that the data was not able to explain age well. The trend was consistent, irrespective of number of clusters. The cluster plots continued to explain less than 30% of the point variability. Furthermore, due to the amount of data it was very difficult to visually inspect the clusters via the cluster plot. Results from the elbow method did not show a reduction in variance and there was no "elbow" found. Therefore, a re-examination of the objective in relation to the data and method was considered. Options considered included:

a) A reduction in variables

b) A reduction in observations

A reduction in variables: Fixations are the base of all eye movements (Leigh & Zee, 2010). Fixations allow humans to stand, to balance, and to direct attention. Fixations are located within the brain stem, the oldest part of the brain, as they form the foundation of all other movements. Fixations occurred in human evolution before saccades and smooth pursuits (Leigh

& Zee, 2011). Therefore, when reviewing the variables within the context of the objective, it was deemed appropriate to remove all variables that were not related to fixations. This left seven variables for further examination (Figure 31).

```
> colnames(Three_MH_excel_cleaning)
[1] "AGE"                                                    "FIXATION_STABILITY_bcea"
[3] "FIXATION_STABILITY_convergence_point.diff"              "FIXATION_STABILITY_depth"
[5] "FIXATION_STABILITY_fixation_dispersion.both_avg"        "FIXATION_STABILITY_gaze_positions_band1.both_avg"
[7] "FIXATION_STABILITY_targeting_displacement_vertical.both_avg"
```

Figure 31: Variables remaining from which to build the clustering before age is dropped

A reduction in observations: Observations were reduced to the first 1000 in the dataset. This was done after consulting research in eye tracking suggesting appropriate numbers for normative analysis (Fleiss,). Numbers of 1000 were deemed appropriate and had precedence in the industry. Then the objective of the analysis was considered in relation to the results.

After further iterations and examination of the number of clusters an acceptable cluster output was achieved that also fit the objective of the analysis. The number of clusters was determined to be five and the iterations to generate the model was five. The total betweenness was 3011 and withinness was 2497. The withinness was now smaller than the betweenness values. Furthermore, the cluster plots continued to explain well over half the variability (71%). Numbers of observations within each cluster showed no clusters with too few observations (140, 251, 166, 142, 220).

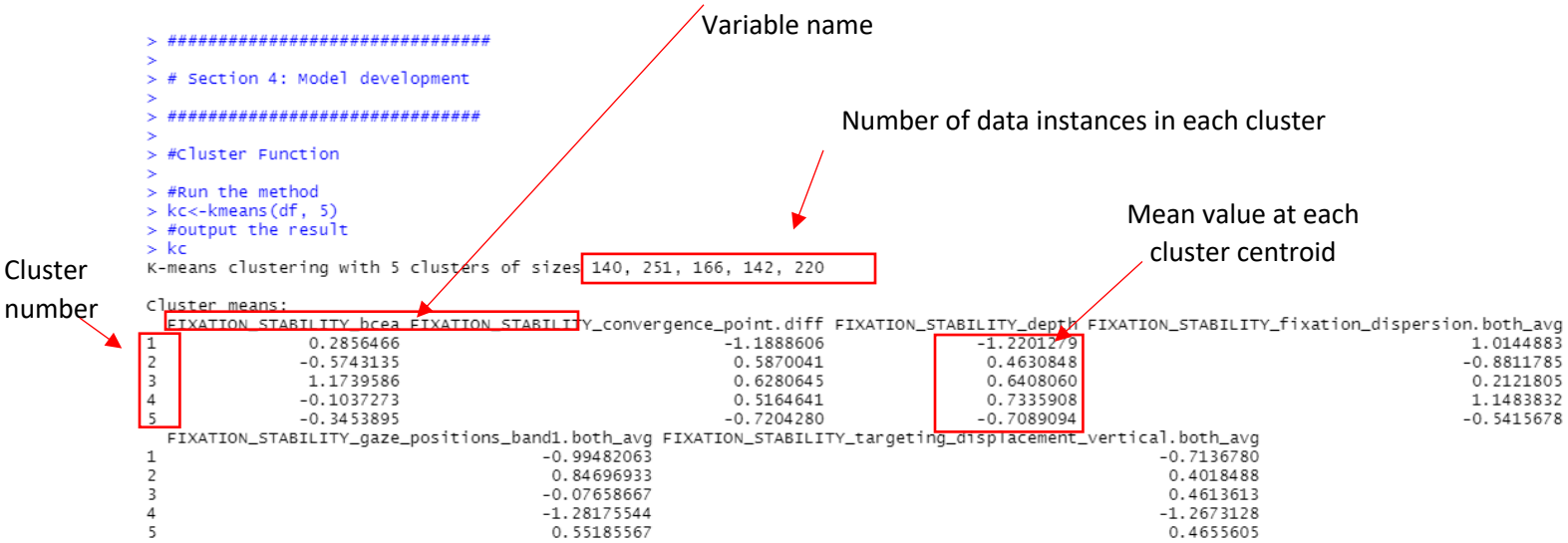
The objective of the analysis was to explore the variables that group together (cluster) to find similarities and relationships within the dataset. For this dataset, the objective was to compare how the method clustered the data with age. Understanding how eye movements change over the lifespan is critical in determining thresholds of normality for baseline comparisons of peoples with suspected injury, clinical condition and for elite levels of

performance (Murray, Hunfalvay, & Bolte, 2017; Lange, Hunfalvay, Murray, Roberts, & Bolte, 2018).

Given that fixations are the evolutionary base of all human eye movements and that the cluster analysis was able to determine higher values between clusters, lower values within clusters and over 70% of the variability, the stated objective was deemed to be successfully achieved by this analysis.

Model Properties

The final clustering characteristics included five clusters built upon six variables (each a fixation variable, Figure 32) and, after cleaning 919 observations. The total betweenness was 3011 and withinness was 2497 and total sum of squares was 5508. Numbers of observations within each cluster were 140, 251, 166, 142, 220.




```
> ################################
>
> # Section 4: Model development
>
> ################################
>
> #Cluster Function
>
> #Run the method
> kc<-kmeans(df, 5)
> #output the result
> kc
K-means clustering with 5 clusters of sizes 140, 251, 166, 142, 220

Cluster means:
  FIXATION_STABILITY_bcea FIXATION_STABILITY_convergence_point.diff FIXATION_STABILITY_depth FIXATION_STABILITY_fixation_dispersion.both_avg
1               0.2856466                              -1.1888606            -1.2201279                                      1.0144883
2              -0.5743135                               0.5870041             0.4630848                                     -0.8811785
3               1.1739586                               0.6280645             0.6408060                                      0.2121805
4              -0.1037273                               0.5164641             0.7335908                                      1.1483832
5              -0.3453895                              -0.7204280            -0.7089094                                     -0.5415678
  FIXATION_STABILITY_gaze_positions_band1.both_avg FIXATION_STABILITY_targeting_displacement_vertical.both_avg
1                                      -0.99482063                                            -0.7136780
2                                       0.84696933                                             0.4018488
3                                      -0.07658667                                             0.4613613
4                                      -1.28175544                                            -1.2673128
5                                       0.55185567                                             0.4655605
```

Figure 32: Output of the final model showing variables, mean values at each cluster centroid, number of observations within each cluster

The cluster plot (Figure 33) shows that the cluster overlap. The goal is to minimize the sum of squared distances within the cluster (creating a tight circle) while maximizing the sum of

squared distances between clusters. There is some overlap between clusters, especially on the left of the cluster plot. Therefore, it would be expected that when viewing the cross-tabulation, there would be observations in cluster 2 and 5 that hard to distinguish due to the overlap in these two clusters.

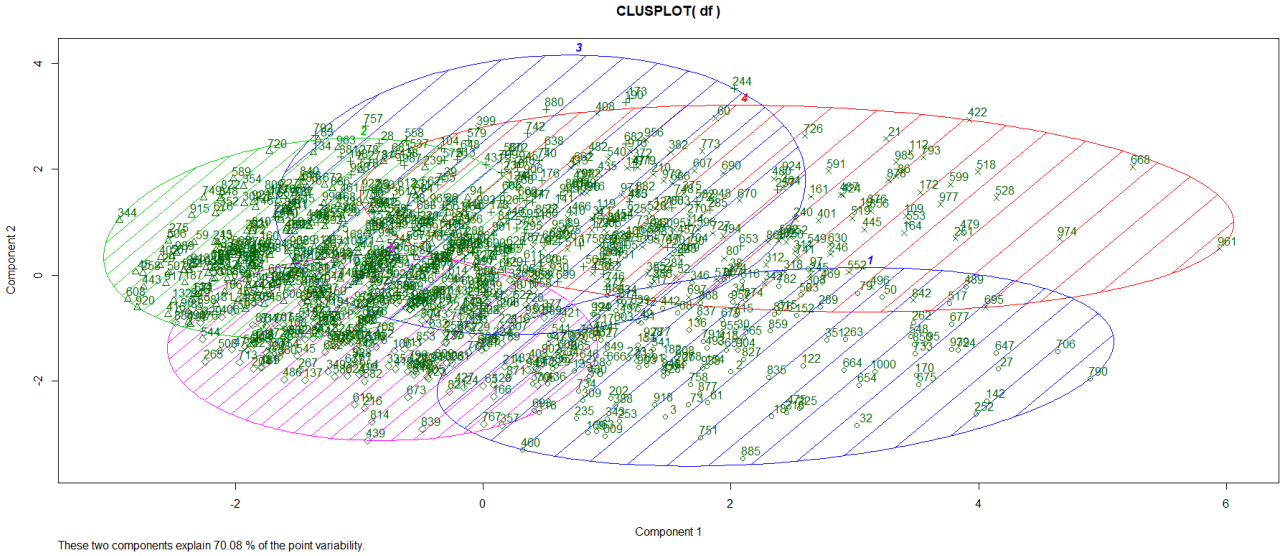


Figure 33: Cluster point with observations and final five clusters

Figure 34 was generated to try and provide a clearer picture of the clusters and the overlap.

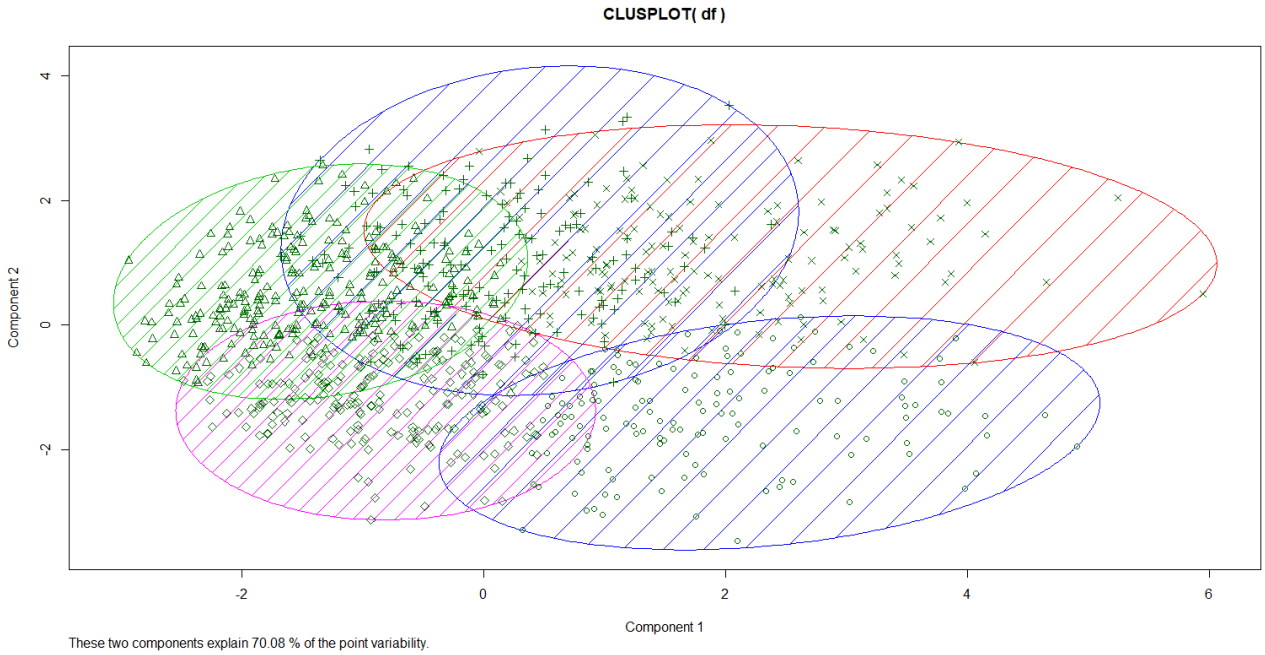


Figure 34: Cluster plot after removing observation numbers.

Evaluation

The model was evaluated from several different perspectives. It was deemed appropriate to look at different aspects of evaluation as the unsupervised nature of the exploratory analysis needed to be explored in relation to what made practical sense in explaining the output. Therefore, evaluation methods included:

1. The cross-tabulation matrix
2. The elbow method
3. The optimal number of clusters as recommended by the best$pamobject using the fpc library
4. Via anomaly detection
5. Via common sense

1. *The Cross-tabulation method:* shows the dominant class for each cluster (Figure 35). Each row is a age class, of which there are seven age classes. Each column is a cluster number, of which there are five clusters ($k = 5$).

```
> # cluster to age evaluation
> hist(df_AGE)
> df_AGEbins=cut(df_AGE, breaks=
> table(df_AGEbins, kc$cluster)

df_AGEbins  1  2  3  4  5
     0-11  14 35 33 16 19
     11-17 34 36 43 30 35
     17-29 29 73 29 51 46
     29-42 27 39 23 21 41
     42-56 24 43 17 13 54
     56-65  7 17 13  7 16
     65-90  5  8  8  4  9
```

Figure 35: Cross-tabulation of age group with cluster number, showing dominant age within each

The dominant cluster for 0-11 years of age is cluster 2 (n = 35; Figure 35). The dominant cluster for 11-17 years of age is cluster 3 (n = 46); for 17-29 is cluster 2 (n = 73); for 29-42 years of age is cluster 5 (n = 41); for 42-56 years of age is cluster 5 (n = 54); for 56-65 years of age is cluster 2 (n = 17); for 65-90 is cluster 5 (n = 9).

Each row adds up to the sum of that age. For 0-11 age group the total in the data set is 117 observations. For 11-17 the total observations are 178, for 17-29 are 228; for 29-42 are 151; for 42- 56 are 151; for 56-65 are 60; for 65-90 are 34.

In each row we can see the instances in each cluster. For instance, for 29-42 years of age, there are 27 instances in cluster 1; 39 instances in cluster 2; 23 in cluster 3; 21 in cluster 4 and 41 in cluster 5.

Moving vertically to the columns; the columns show the numbers of each age group in that cluster. In column 1 (cluster 1) there are 14 of age 0-11; 34 of age 11-17; 29 of age 17-29; 27 of age 29-42; 24 of age 42-56; 7 of age 56-65; 5 of age 65-90. Totaling these numbers reveals the number of instances in a cluster. For cluster 1 the total number of instances are 14 + 34 + 29 + 27 + 24 + 7 + 5 = 140. Column 2, 3, 4 and 5 can also be evaluated the same way. The total for each column represents the number of instances within that cluster.

To determine how many instances the age cluster assignment matches the actual age, the highest numbers in each row are added and then divided by the total number of instances of age. 35 + 43 + 73 + 41 + 54 + 17 + 9 = 272 out of 919. Therefore, the age was clustered in agreement with the age group 272/919 = 0.2960 or **30%.**

2. *Via the elbow method:* The elbow method was employed to evaluate the percentage of variance explained as a function of the number of clusters (k; Figure 36). The black line indicates the total withinness. The blue line represents the betweenness. Where they intersect and level out indicates the elbow as represented by the green circle and occurs at 4 clusters. When reviewing 4 versus 5 clusters however, the 5 clusters made more practical sense (described later in the practical sense section).

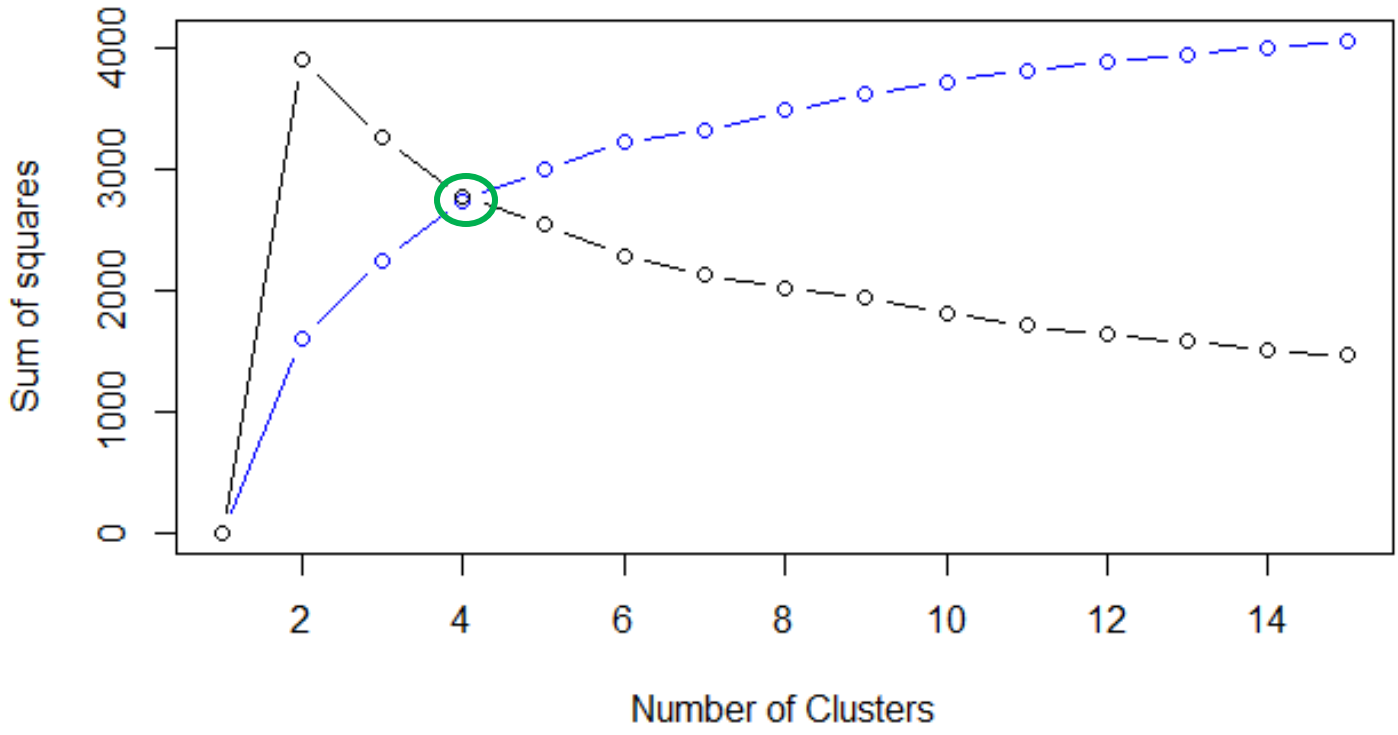


Figure 36: Graphical representation of the elbow method

3. *Via the optimal number of clusters as recommended by the best$pamobject using the fpc library.* This method of evaluation was employed to show the goodness of fit for each observation within a cluster. This reveals two clusters as seen in figure 37 and 38. When reviewing 2 versus 5 clusters however, the 5 clusters made more practical sense (described later in the practical sense section).

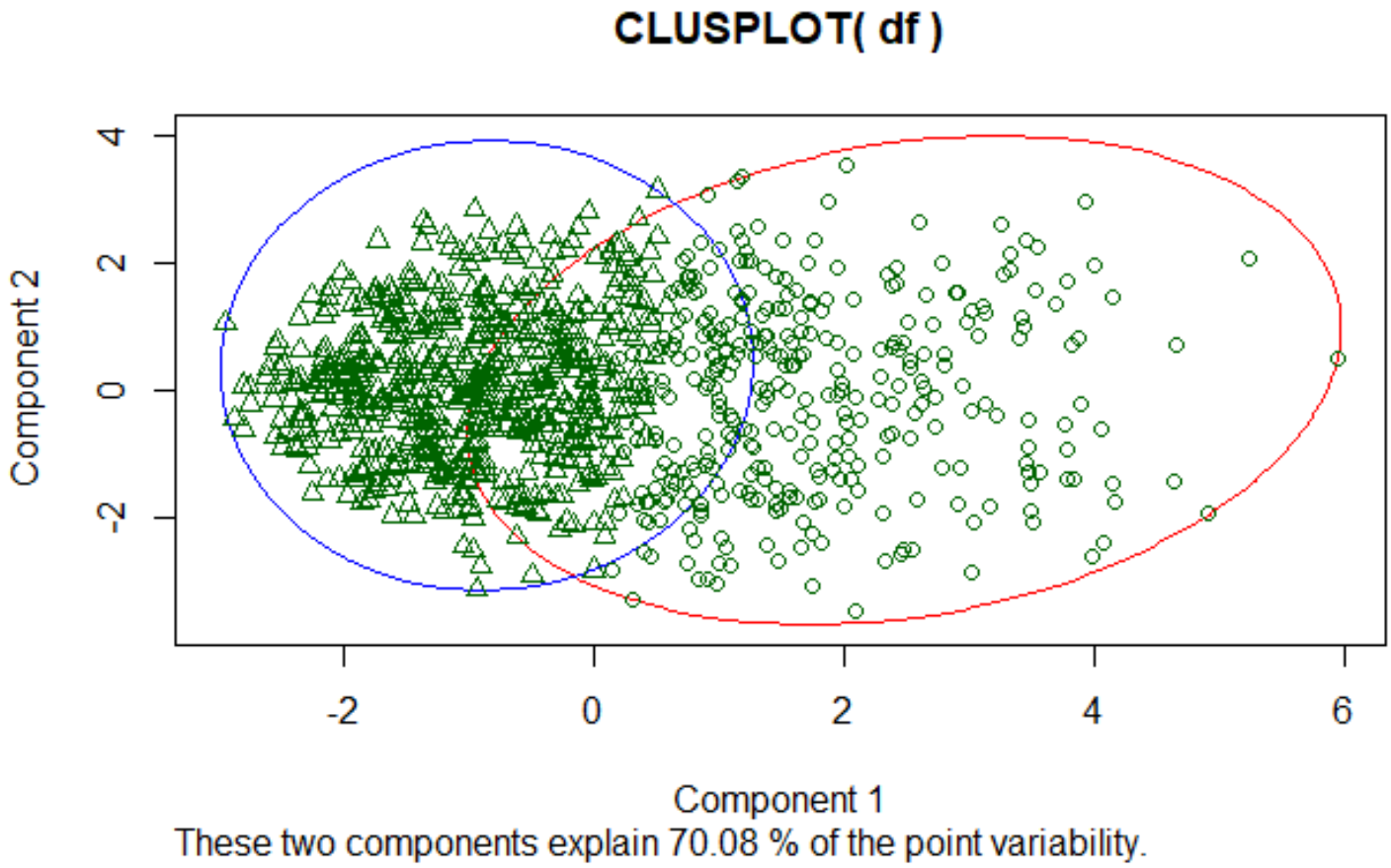


Figure 37: Cluster plot using best pamo object method via symbols

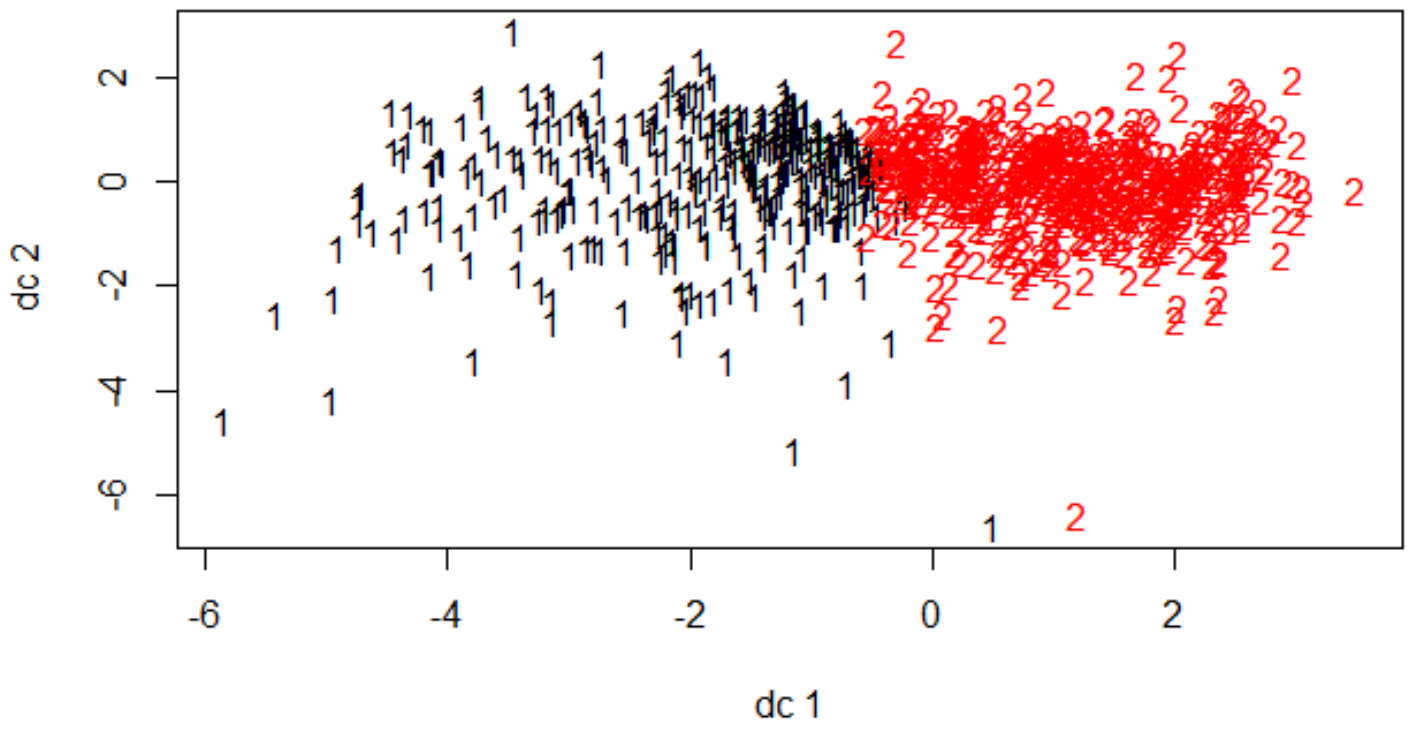


Figure 37: Cluster plot using best pamo object method via numbers

4. *Via anomaly detection:* Outliers are determined by taking the distance from the centroid within a cluster. The further the point the more it outlies.

Outliers are instances with the largest distance from the center of the cluster. Detection of outliers may represent bad data, or something interesting in the data that needs further investigation.

Figure 38 shows the cluster, the variable, and the distance from the centroid.

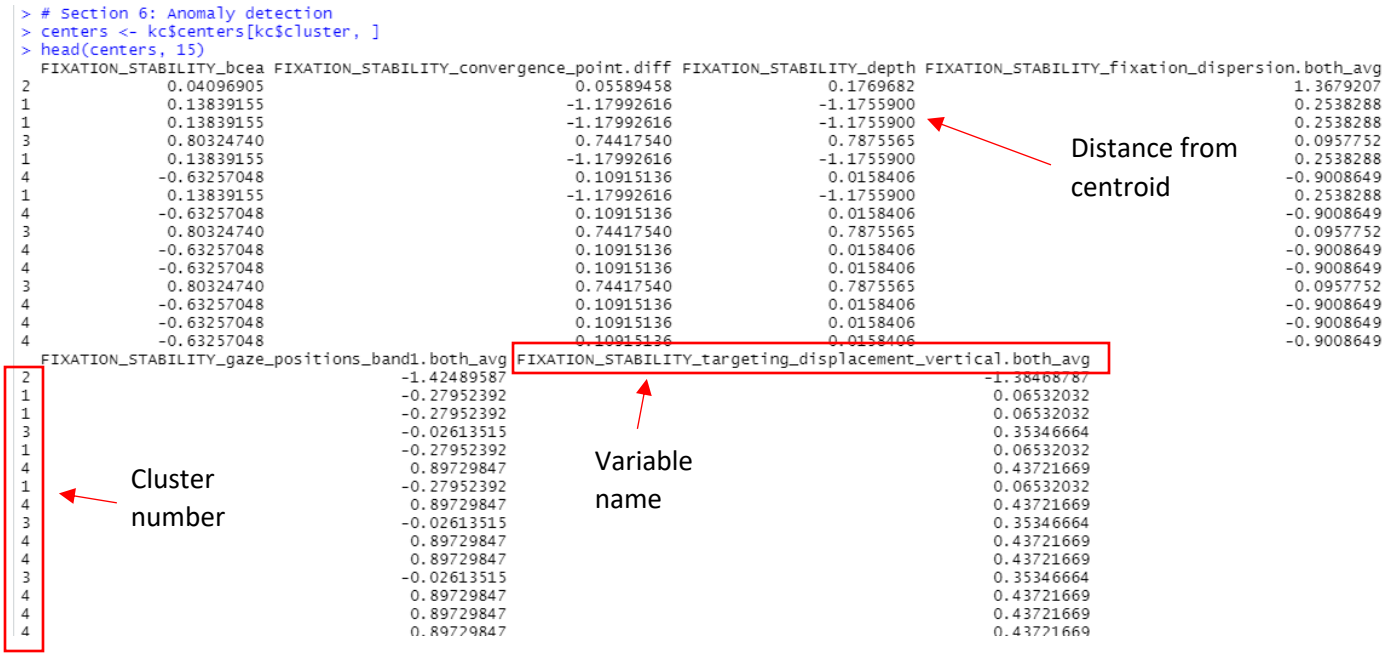


Figure 38: Variables, clusters, and distance from centroid

The first five outliers, as defined as those with the greatest distance from the centroid, are seen in figure 39. They are identified as observation 886, 832, 730, 356 and 577.

```
> outliers <- order(distances, decreasing=T)[1:5]
> outliers
[1] 886 832 730 356 577
```

Figure 39: Identifies the first 5 outliers

5. *Via common sense:* As cluster analysis is an unsupervised and exploratory method, it is important to view the output with a practical mindset and to include a subject matter expert (SME). After considering all the evaluation methods it was deemed appropriate to use 5 clusters even when some of the evaluation methods suggested other cluster numbers.

Two clusters as recommended by the pamo object were examined. When viewing the cross-tabulation method there was not enough clusters to provide practical understanding of age.

Specifically, young, and old age groups fell within the same clusters. More than two clusters were needed.

When reviewing the elbow method, 4 clusters were examined. Although 4 clusters may have been used, when viewing five clusters the dominant age group for each cluster made more practical sense. Specifically, oculomotor behavior is expected to decline over age. Therefore, if a younger age group and an older age group were in the same cluster (which occurred more often with 4 clusters), it did not make sense.

When reviewing the cross-tabulation for 5 clusters the dominant cluster for each age group differed enough between age groups *and* made practical and better scientific sense. Therefore, five clusters were determined to be optimal for the data set.

Conclusion

Summary

The objective of the analysis was to explore the variables that group together (cluster) to find similarities and relationships within the dataset. For this dataset, the objective was to compare how the method clustered the data with age. Understanding how eye movements change over the lifespan is critical in determining thresholds of normality for baseline comparisons of peoples with suspected injury, clinical condition and for elite levels of performance (Murray, Hunfalvay, & Bolte, 2017; Lange, Hunfalvay, Murray, Roberts, & Bolte, 2018).

Finding revealed the following outcomes:

1. The dominant age cluster for middle age was the same, that is for 29-42 and 42-56 years of age the cluster was number 5. This makes sense as eye movements are stable

and should not change in these age ranges unless there is a clinical incident, such as early onset Parkinsonism (Gitchel, Wetzel, & Baron, 2012).

2. Very young persons, up to 11 years of age differ from those in the teenage years (11-17). Developmentally this makes sense as the coordination of the muscles around the eyes develops during the early years and should therefore be different after the developmental stage (Lange, Hunfalvay, Murray, Roberts, & Bolte, 2018).
3. Age ranges show different cluster dominance indicating lifespan development changes influence oculomotor behavior in non-clinical persons.

Armed with this information the following guidance may be given to RightEye customers, clinicians, patients, and researchers. Understanding how eye movements change over the lifespan is critical in determining thresholds of normality for baseline comparisons of peoples with suspected injury, clinical condition and for elite levels of performance.

Table 2: Summary Outcome and Strategies for Oculomotor Change throughout the Lifespan

Outcome	**Strategy**
Eye movements should remain relatively stable and without much change from ages 29-42 and 42-56.	Clinician guidance: If a baseline is not obtained from a patient, compare the results to others within either of these age groups to obtain comparative information seems a reasonable strategy.

Young persons, up to 11 years of age differ from those in the teenage years (11-17 years)	Clinician guidance: clinicians should not compare those under 11 with teenage normative values as they differ.
Age ranges show different cluster dominance indicating lifespan development changes influence oculomotor behavior in non-clinical persons.	Research interpretation: eye movements differ across the lifespan. More research is needed to understand these differences.
Age ranges show different cluster dominance indicating lifespan development changes influence oculomotor behavior in non-clinical persons.	Clinician and Patients: should expect to see changes in oculomotor behavior as they age and prepare for mitigating circumstances such as eyeglasses, or vision training exercises.

In conclusion, through awareness, there are various strategies that can ensure patients and clinicians understand to expect changes in eye movements across the lifespan. These can inform strategies and ease anxieties of patients who may see changes as they age – as these changes are a normal part of the aging process. It is important for all members of the community and those who serve the patients to be aware, educated, and supportive of these changes in order to increase their well-being in the hopes of a happy, healthy, safe and long life.

Limitations

Many limitations were found while conducting the analysis. Data limitations were significant and included the following:

1. *The number of variables was large:* 137. This made the analysis difficult to truly understand in great depth.
2. *Subset variables:* Many of the variables were subsets of larger variables making it difficult to know what "level" of variable to consider for the analysis. Even though correlations revealed no highly correlated variables (which was surprising) the variable subsets made the practical understanding of the output difficult especially before the reduction to use of only fixation variables.
3. *The data was messy.* This is to be expected by "real world" data, nevertheless, the volume and amount of cleaning of this data should be explored in more depth with further analysis. One example can be to examine if production versions of the data differed over the span of the data pull. Sometime metric algorithms are updated or further sliced which can impact the output of the data.
4. *There was a lot of missing data:* Only 17 observations had every single cell complete with data. This was to be expected as some data cells only get populated if certain events occur. For instance, missing a target with the eyes. If the patient never missed the target the cell remains NA. This is a good outcome, but it makes the analysis difficult and further pre-processing should be conducted.
5. *The quality of the data:* as in many real-world datasets, the quality of the data is hard to validate. For instance, it is known that many persons taking the tests simply don't want to add in any clinical information. There are various reasons for this, often it relates to time to do so. Therefore, if no clinical condition is added to the data, then the data is assumed to be from a non-clinical, normal patient. This may or may not be the case. Verifying every single observation for clinical conditions is impractical.

6. *Age group stratification:* The age group sample sizes differed across groups. Especially in the elderly group, there were too few observations (n = 34).

There were also many limitations in the analysis, including:

1. *The initial visualizations:* the initial cluster plot was so heavy with data points and overlap (even when trying new types of cluster plots) that it was difficult to make sense of.
2. *Differing guidance:* the various evaluation methods were contradictory especially when looking at the practical and scientific explanations of the data. Elbow method suggested 4 clusters, best$pamobject suggested 2 and practical guidance would indicate 5 clusters.
3. *Some of the results do not make sense:* even though some guidance can be extracted from the analysis, there are several dominant clusters that do not make sense. Specially, age 0-11, 17-29 and 56-65 all fell in the same cluster (2). Scientifically this does not make sense. What would make sense is if the young and old clustered together and middle age clustered together. This did not happen and therefore, when viewing the results globally, I would not feel comfortable taking this forward to a production environment.
4. *Even with some good evaluation numbers the results are misleading:* The withinness was smaller than the betweenness values. Furthermore, the cluster plots continued to explain well over half the variability (71%). Numbers of observations within each cluster showed no clusters with too few observations (140, 251, 166, 142, 220). Even with these initial evaluation techniques the output of the model makes little global

sense. Yet, these were used to guide the number of clusters as other methods were found to be contradictory.

5. *The number of clusters:* Determining the number of clusters became a guessing game. Hours were spent modifying the data samples, variables, number of clusters and then re-evaluating the results in the context of the objective.

6. *The inability to run validation techniques.* On the larger dataset (prior to the 1000 observations) several parts of the code were unable to run including the best$pamobject. Although left to run overnight at one point the method still did not complete and had to be aborted. By the way, this also occurred to many of us in the group project. Therefore, this information was not known in the larger dataset and was a limitation of the code/data.

Improvement Areas

Improvement areas include:

1. *Verifying the version of production:* the data was obtained from in order to make sure all the algorithms (variables) are calculated in the same manner.

2. *Stratifying age groups.* Figure 40 shows the distribution of age in the final dataset after cleaning. An improvement would be to have the same number of observations in each age group. A further improvement would be to ensure the age group met the normative data standards for grouping as set by the National Institute of Health, whereby each group should have a minimum of 100 per group (Fleiss, 1986).

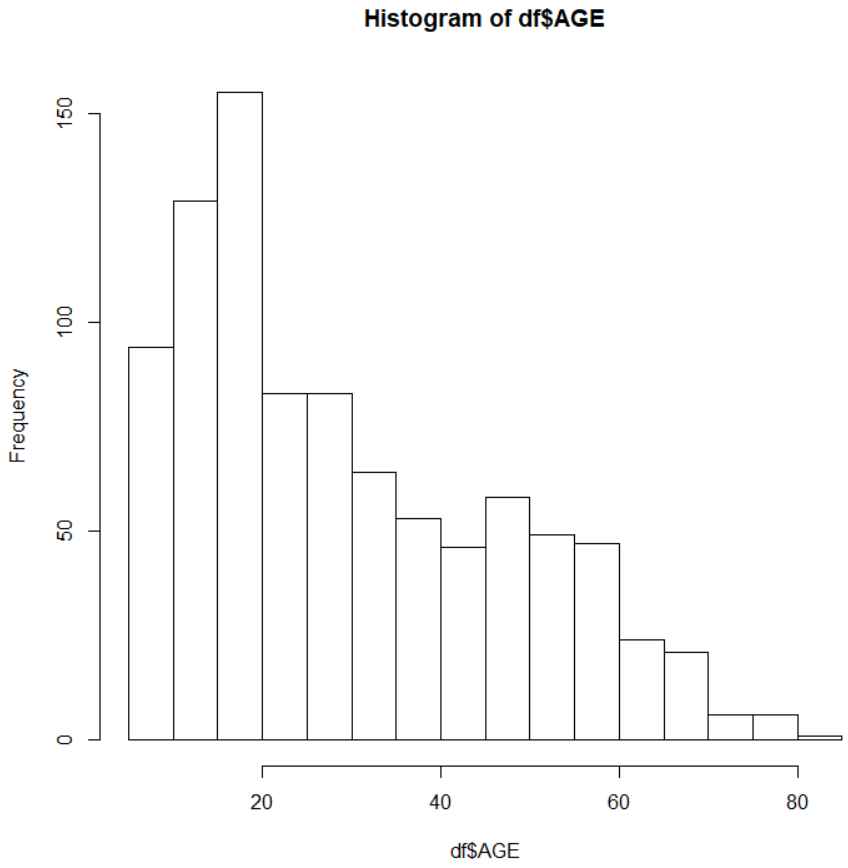


Figure 40: Histogram of age after final cleaning

3. *Cleaning of data:* many steps are needed to further clean and understand the data set. Some examples include:

    a. Potentially adding dichotomous variables to indicate positive or negative actions e.g., 0 = no misses of a target, and 1 = 1 or more misses. This was NA outcomes which show no data are not excluded. Another option can be to change the NA to a zero output.

    b. Removing outliers found in anomaly detection. This can help to further normalize the data.

    c. Prior to analysis including data that is age stratified that falls only within the

25/75th percentiles. This again may help to normalize the data by age.

4. *Where possible, validate the data.* There are some options to reach out to friendly customer to have them spot check validate the patients as falling within the "normal" category. This will not completely solve incorrect input of data but may guide us to remove data from certain companies if we know their practice is not to include clinical conditions.

5. *Consider other clustering techniques.* Given the issues with k-means, it may be a more appropriate path to consider hierarchical clustering. This may be useful given the number of variables. It could also be interesting to potentially go back and forth between the hierarchical clustering method, once the dissimilarities are known as such information may guide the variable selection for a K-means analysis. The disadvantage of hierarchical clustering would be that the hierarchy would be very large.

6. *Consider other agglomeratives:* the centroid method was used in this analysis which examines each data point in relation to the center of the matrix. Perhaps an average distance between all pairwise proximities may be more appropriate as we are looking for more comparative datapoints between clusters.

7. *Subject Matter Expertise.* Without a level of SME expertise on this data I think it would have been very difficult to know how to proceed with the analysis. This leads to a limitation in either a) resources who are analysts who know the data, b) the need for SME intervention and input/time allocation as high. As a business owner, both of these are potential limitations, especially to a small business.

8. *Consider other analytical options.* Finally, other analytical options should certainly be considered. These may include:

    a. ANOVA between variables that could then be used for further analysis in either clustering or regression

    b. Regression to determine weights of differences and significance that could be used to assist in variable reduction as well as a more informed algorithm for normative data comparisons. Specifically, it is anticipated that not all variables are equally important to determining levels of normality and a regression could help to understand which variables are important, especially when comparing clinical to non-clinical groups.

    c. Neural networks. Although NN can be difficult to understand due to the black box, it may be an appropriate technique given the amount of data and number of variables in this dataset.

On a side note, thank you Professor, for allowing me to use my own data for this project. It was very helpful to understand the data in the context of clustering. It was disappointing the results were not stronger, but it helps me to understand some potential next steps. Thank you.

References

Barnes, G.R. (2008). Cognitive processes involved in smooth pursuit eye movements. *Brain Cognition, 68*(3):309–26. doi:10.1016/j.bandc.2008.08.020.

Ciuffreda, K. J., Kapoor, N., Rutner, D., Suchoff, I. B., Han, M., & Craig, S. (2007). Occurrence of oculomotor dysfunctions in acquired brain injury: A retrospective analysis. *Optometry - Journal of the American Optometric Association, 78*, 155-161.

Duchowski, A. (2007). *Eye tracking methodology: theory and practice.* London: Springer.

Fleiss, J.L. (1986) Design and analysis of clinical experiments. Willey, New York.

Gitchel, G.T., Wetzel., P.A., & Baron, M.S. (2012). Pervasive ocular tremor in patients with Parkinson disease. *Archives of Neurology, 70,* E1-7. doi:10.1001/archneurol.2012.70.

Holmqvist K, & Nystrom M. (2011). *Eye Tracking: A Comprehensive Guide to Methods and Measures.* Oxford University Press.

Hunfalvay, M., Orr, R., Murray, N., & Roberts, C.M. (2017). Evaluation of Stereo Acuity in Professional Baseball and LPGA Athletes Compared to Non-Athletes. *Vision Development and Rehabilitation, 3 (10) 33-41.*

Hunfalvay, M., Murray, N.P., Carrick, F.R. (2020). Fixation stability as a biomarker for differentiating mild traumatic brain injury from age matched controls in pediatrics. Brain Injury. https://doi.org/10.1080/02699052.2020.1865566

Komogortsev, O.V. & Karpov, A. (2013). Automated classification and scoring of smooth pursuit eye movements in the presence of fixations and saccades. *Behavior Research Methods. 45*(1):203–15. doi:10.3758/s13428-012-0234-9.

Lange, B., Hunfalvay, M., Murray, N., Roberts, C.-M. & Bolte, T. (2018). Reliability of computerized eye-tracking reaction time tests in non-athletes, athletes, and individuals with traumatic brain injury. *Optometry & Visual Performance*, 6 (3), 165-180.

Maimon, O., Rokach, L. (2010). *Data Mining and Knowledge Discovery Handbook*, 2nd ed., DOI 10.1007/978-0-387-09823-4_21, Springer Science+Business Media, LLC.

Mollenbach, E., Hansen, J.P., Lillholm, M. (2013). Eye movements in gaze interaction. *Journal of Eye Movement Research. 6*(2):1–15. doi:10.16910/jemr.6.2.1.

Murray, N. P., Hunfalvay, M., & Bolte, T. (2017). The Reliability, Validity, and Normative Data of Interpupillary Distance and Pupil Diameter Using Eye-Tracking Technology. *Translational Vision Science & Technology, 6* (4), 1-12.

Murray, N.P., Kubitz, K., Roberts, C-M., Bolte, T., & Tyagi, A. (2019). An examination of the oculomotor metrics within a suite of digitized eye tracking tests. *Vision Development and Rehabilitation, 5*(4): 269-284.

Naicker, P., Anoopkumar-Dukie, S., Grant, G. D., & Kavanagh, J. J. (2017). Medications influencing central cholinergic neurotransmission affect saccadic and smooth pursuit eye movements in healthy young adults. *Psychopharmacology, 234*, 63-71. doi:10.1007/s00213-016-4436-1.

Poole, A. & Ball, L,J. (2005). Eye tracking in human-computer interaction and usability research: current status and future prospects. In: Ghaoui C, editor. *Encyclopedia of human-computer interaction.* Hershey: Pennsylvania Idea Group; 2005.

Tenenbaum, G. (2003). Expert athletes: An integrated approach to decision making. In J. L., Starkes, J & K. A., Ericsson (Eds.). *Expert performance in sports: Advances in research on Ssort expertise* by Human Kinetics, Champaign, IL.

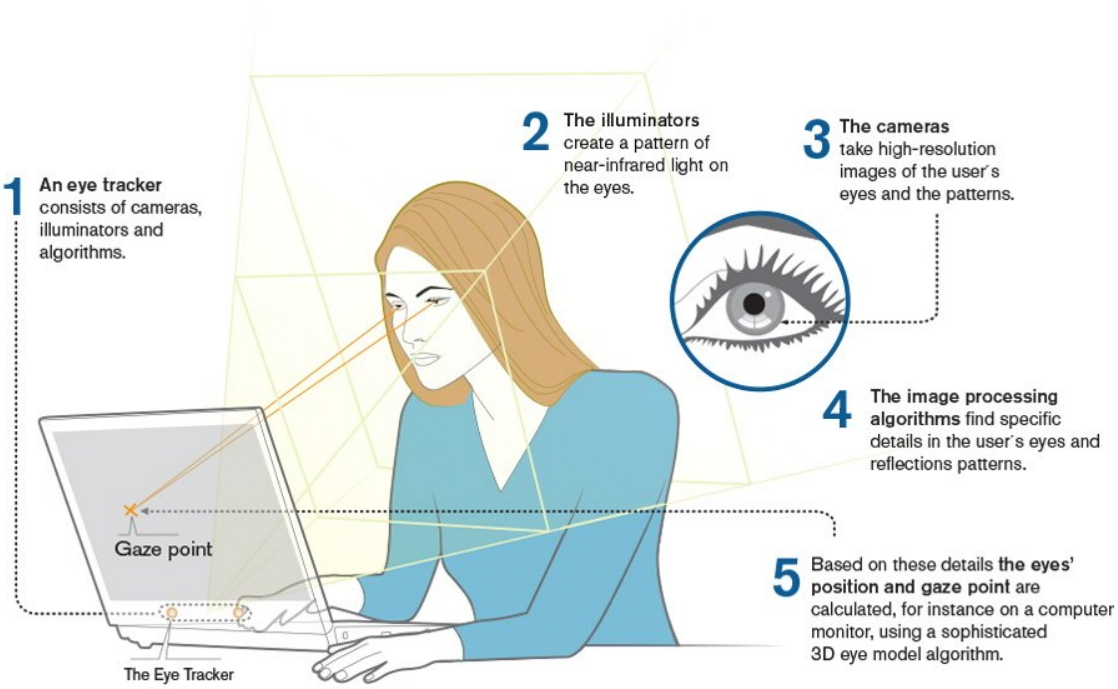
Williams, M.A. & Ward, P. (2003) Perceptual expertise: Development in sport. In J. L., Starkes, J & K. A., Ericsson (Eds.). *Expert performance in sports: Advances in research on Sport Expertise* by Human Kinetics, Champaign, IL.

Appendix A

Tobii I15 Eye Tracking Device

Appendix B

How eye tracking works



Appendix C

Direct language from:

Murray, N.P., Kubitz, K., Roberts, C-M., Bolte, T., & Tyagi, A. (2019). An examination of the oculomotor metrics within a suite of digitized eye tracking tests. *Vision Development and Rehabilitation, 5*(4): 269-284.

Oculomotor Tasks

Five RightEye oculomotor tests are described below. From these 5 tests, 54 different metrics of digitized oculomotor function was assessed (for full description see Appendix I).

Circular smooth pursuit test (CSP). In the CSP test, participants were instructed to track a target stimulus, a black dot of 0.2 degrees' diameter at a 10-degree radius at a rate of 0.4Hz, in a clockwise direction, for 15 seconds. The 0.4 Hz = 1 revolution/0.4 revolutions per sec = 2.5 sec. To find linear velocity, we multiply the angular velocity. The CSP test provides measures of fixation percentages, saccade percentages, latent smooth pursuit, and smooth pursuit target accuracy.

Horizontal smooth pursuit test (HSP). In the HSP test, participants were asked to focus on a dot (same size and speed as the CSP test) on the screen and follow the dot horizontally across the screen for 25 seconds, moving to the far right, then to the far left, and back to the center. The stimuli moved in a sinusoidal way from the left to right and right to left in a straight line. For a participant to be considered "on target," they were required to follow the stimuli within an error of 2.4 degrees. A participant could also be ahead or behind a stimulus and can still be labeled as 'following' if they are within an error of 4.8 degrees. The HSP test also provides measures of fixation percentages, saccade percentages, latent smooth pursuit, and smooth pursuit target accuracy.

Vertical smooth pursuit test (VSP). The protocol for the VSP test was the same as the protocol for the HSP test. However, the VSP test was in a vertical plane.

Horizontal saccades test (HS). In the HS test, participants were asked to look at a countdown of 3, 2, 1 in the center of the screen before moving their eyes back and forth between 2 dots. Their goal was to "target each dot" on the left and right of the screen as quickly and accurately as possible. The dots on the screen turned green when the participants' eyes hit the targets. The

test lasted 10 seconds. The HS test provides measures of fixation percentages, saccade percentages, and target accuracy.

Vertical saccades test (VS). The protocol for the VS test was the same as that for the HS test. However, the VS test was in a vertical plane.

Direct language from:

Hunfalvay, M., Murray, N.P., Carrick, F.R. (2020). Fixation stability as a biomarker for differentiating mild traumatic brain injury from age matched controls in pediatrics. Brain Injury. https://doi.org/10.1080/02699052.2020.1865566

Clinical diagnosis of mTBI was based on the American Congress of Rehabilitation Medicine (ACRM) definition (38). All participants were additionally examined using the GCS and scored between 13 and 15 on the scale. Although the GCS is widely used it is not necessarily the best measure of pediatric mTBI (21). Furthermore, clinicians do not usually use imagining for pediatric mTBI cases (39). Therefore, the Graded Symptoms Checklist (GSC) in the Standardized Assessment of Concussion (SAC) (40) was also used as a secondary clinical tool for measurement of mTBI as recommended by the Journal of the American Medical Association Pediatrics clinical guidelines (39,41). Using results from Grubenhoff, Kirkwood, Gao, Deakyne, and Wathen (19) and the American Academy of Neurology (AAN) (42) concussion grading scale, pediatric patients (6–18 years of age) were evaluated as having mTBI if their GCS score was between 7.7 and 19.3. According to Grubenhoff et al. (19), this yielded a 95% confidence interval for case-patients with an AAN grade 1 TBI (7.7–10.7) or grade 2 TBI (11.5–19.3). Therefore, participants in the mTBI group in this study scored between 13–15 on the GCS *and* 7.7–19.3 on the GCS.

Apparatus

Stimuli were presented using the RightEye tests on a Tobii I15 vision 15" monitor fitted with a Tobii 90 Hz remote eye tracker and a Logitech (model Y-R0017) wireless keyboard and mouse. The participants were seated in a stationary (nonwheeled) chair that could not be adjusted in height. They sat in front of a desk in a quiet, private room. Participants' heads were unconstrained. The accuracy of the Tobii eye tracker was 0.4° within the desired headbox of 32 cm × 21 cm at 56 cm from the screen. For standardization of testing, participants were asked to sit in front of the eye-tracking system at an exact measured distance of 56 cm (ideal positioning within the headbox range of the eye tracker).

Oculomotor task

The RightEye Fixation Stability oculomotor test included viewing six targets, presented one at a time, for 7 seconds each, with a break of 3 s between targets. Before each target was presented, identical verbal instructions were given to every participant: "Move your eyes to the center of the target. Keep your eyes as still as possible, until the target disappears." The tester then asked, "Are you looking at the center of the target?" Once the participant confirmed with a verbal "Yes" the tester pressed the spacebar and the 7-s time began.

The same order of targets was used for each participant and used in past fixation stability research from Bellmann and colleagues (33): Target 1 was a 1° cross, T2 was a 1° filled circle, T3 was a small 4-point diamond (3° point separation) using dimensions as in the Humphrey Field Analyzer (Carl Zeiss Meditec, Dublin, CA), T4 was a large 4-point diamond (7° point separation) using dimensions as in the Humphrey Field Analyzer, T5 was large-crossover whole-image diagonal with open 1° center, T6 was a 1° letter x (Figure 1). The following metrics were used to examine fixations; Bivariate Contour Ellipse Area (BCEA), Convergence Point, Depth, Disassociated Phoria, and Targeting Displacement (See Table 2 for further information).

Procedure

Participants were recruited through RightEye clinical providers. The study was conducted in accordance with the tenets of the Declaration of Helsinki. The study protocols were approved by the Institutional Review Board of East Carolina University. The nature of the study was explained to the participants and all participants provided written consent to participate. Participants were excluded from the study if they had more than one single discrete episode of mTBI (n = 21). Following informed consent, participants were asked to complete a prescreening questionnaire and an acuity vision screening where they were required to identify four shapes at 4 mm in diameter. If any of the prescreening questions were answered positively and any of the vision screening shapes were not correctly identified, then the participant was excluded from the study (n = 3). Additionally, participants were excluded from the study if they reported any of the following conditions, which may have prevented successful test calibration during the prescreening process: this included vision-related issues such as extreme tropias, phorias, static visual acuity of >20/400, nystagmus, cataracts, or eyelash impediments or if they had consumed drugs or alcohol within 24 hours of testing (n = 1) (43–47). Participants were also excluded if they were unable to pass a nine-point calibration sequence. As a result of the pre-screening, the total participants excluded from the study was 25.

Figure 1. Targets used for fixation stability testing. Adapted from Bellmann et al. (43).

Table 2. Fixation metrics.

Fixation Metrics	Definition
Bivariate Contour Ellipse Area (BCEA)	Microsaccades and drifts of the human eye cause corrections of the eye back to a central point. These slight eye movements form an area of dispersion in the shape of an ellipse that is measured by the BCEA.
Convergence Point	The average distance between the "point of convergence of eyes" from the stimuli location on