Appendix A: Tidying the Data

Cheryl Calhoun

26 March, 2017

# Available Artifacts

The following is a list of artifacts currently available from five annual offerings of the program. The artifacts from the first cohort (2012) are incomplete. There is some summary data available that can help in developing an understanding of the 2012 cohort. The 2012 ummary sdata will be incorporated into the narrative where appropriate. The artififacts from the most current cohort (2016) are still being assembled.

Data Available by Cohort as of July 24, 2016

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Cohort | 2012 | 2013 | 2014 | 2015 | 2016 |
| Mentor Bios | **NO** | YES | YES | YES | **NO** |
| Speaker Bios | **NO** | YES | YES | YES | **NO** |
| Judges Bios | **NO** | YES | **NO** | YES | **NO** |
| Mentor Meeting Agendas | **NO** | YES | **NO** | **NO** | **NO** |
| Session Agendas | **NO** | YES | YES | YES | **NO** |
| Technology Descriptions | **NO** | YES | YES | YES | **NO** |
| Applications | **NO** | YES | YES | YES | YES |
| Application Results | YES | YES | YES | YES | YES |
| Attendance Results | YES | YES | YES | YES | YES |
| Resumes | **NO** | YES | YES | YES | YES |
| Business Plans | YES | YES | YES | YES | YES |
| Team Assignments | YES | YES | YES | YES | YES |
| Business Plan Scoring Sheets | **NO** | YES | YES | YES | **NO** |
| Investor Presentations | YES | YES | YES | YES | YES |
| Investor Plan Scoring Sheets | **NO** | YES | YES | YES | **NO** |
| Participant Surveys | Partial | YES | YES | YES | YES |

# Application Data

The applications were collected using an online application. The composition of the application evolved over time and questions were added or updated before each new cohort. This required a bit of data collection and data wrangling to get the data into a form for analysis.

## Application Data Code Book

Applications Table

|  |  |  |
| --- | --- | --- |
| Field | Data Type | Description |
| Identifier | character | Unique identifier concatenated from (Cohort + P + Entry ID) |
| TeamID1 | character | Program Team Assignment concatenated from (Cohort + T + Team ID) |
| Path1 | character | Path to case node in Envivo. |
| Cohort1 | character | Year of cohort (YYYY). |
| Accepted1 | logical | Was participant accepted to program? |
| Decision1 | logical | Did participant attend program? |
| Finished1 | logical | Did particpant finish program? |
| Entry.Id | character | Unique identifier assigned by application system. |
| Team12 | character | Program Team Assignment. Assigned each team an ID (A:H). |
| First2 | character | First Name |
| Middle2 | character | Middle Name |
| Last2 | character | Last Name |
| Phone2 | character | Participant's Phone Number |
| Phone22 | character | Participant's Alternate Phone Number |
| Email2 | character | Participant's Email Address |
| Email22 | character | Participant's Alternate Email Address |
| Address1 | character | First line of address. (2014+) |
| Address1 | character | Second line of address. (2014+) |
| city | character | Name of City (2014+) |
| State | character | Two digit State code (2014+) |
| Zip | character | Mailing address zip code (2014+) |
| Country | character | Mailing address country (2014+) |
| Degree | factor | Participants Highest Degree Earned (PhD, Master, Bachelor, Associate, HS) |
| Experience | character | A description of the field of work/educational experience. |
| Q4 | factor | How did you hear about the program? (Check All That Apply) |
| Q4a |  | Checked: Program website |
| Q4b |  | Checked: Past participant |
| Q4c |  | Checked: Facebook/Twitter/LinkedIn |
| Q4d |  | Checked: Email |
| Q4e |  | Checked: OTL newsletter |
| Q4f |  | Checked: Newspaper article |
| Q4g |  | Checked: Word of mouth |
| Q4h |  | Checked: Other |
| Q4Other | character | Fill in the blank for Other. (2014+) |
| Q5 | factor | What is your primary goal for participation in the program? (Check All That Apply) |
| Q5a |  | Checked: Gain self confidence |
| Q5b |  | Checked: Start a company |
| Q5c |  | Checked: Networking |
| Q5d |  | Checked: Entrepreneurship training |
| Q5e |  | Checked: Attending seminars/workshops |
| Q5f |  | Checked: All of the above. |
| Q5g |  | Checked: Other |
| Q5h |  | Checked: Job/career opportunities (2016) |
| Q5i |  | Checked: Valuable knowledge and skills (2016) |
| Q5j |  | Checked: Interest in technology commercialization (2016) |
| Q5Other | character | Fill in the blank for "Other". (2014+) |
| Q7 | factor | Do you have regular access to a computer? (Yes/No) |
| Q7a-f | factor | Do you have access to the following software? (2014) |
| Q7a |  | Checked: Internet |
| Q7b |  | Checked: Microsoft Word |
| Q7c |  | Checked: Excel |
| Q7d |  | Checked: PowerPoint |
| Q7e |  | Checked: Email |
| Q7f |  | Checked: Other |
| Q7g | character | Fill in the blank for "Other". (2014) |
| Q8 | factor | Have you been involved with any new discoveries that have been patented by the University of Florida? (Yes/No) |
| Q8a | character | If you answered "Yes" to the previous question, please briefly describe the technology and your affiliation. (Ex: inventor, graduate student, post-doc, other) |
| Q9 | character | Attach a copy of your Resume in PDF format. (Limit 3 pages.) |

1 These fields were added to aid in data analysis.  
2 These fields will be redacted in final data set.

\*\*[Notes:\*\*](Notes:**)

* Some of the records from the original data set contain duplicate data. These records were created when participants completed the application twice. The last application (Highest Entry ID #) submitted was retained for data analysis. The earlier submissions are marked as "Duplicate" in the full data set for reference. These will be removed prior to data analysis.
* It is also noted that for the for some years the "Entry ID" numbers do not start at #1. This could be data that was removed by the program organizers, or it could be the earlier entries were test data from testing the application from prior to publication.

## Compiling the Data

The application data was provided in the form of an Excel spreadsheet. For some cohorts (2014 & 2015) there is a raw data spreadsheet as exported from the online application and a separate spreadsheet with additional information that was used in the admissions decision-making process. For other cohorts (2013, 2016), it appears as if the application export spreadsheet was edited to add information and document the decision-making process.

**Notes:**

* For 2012, the applications were collected on paper forms and e-mailed. Unfortunately, the 2012 forms have been lost and are not available for analysis.
* In some cases, information had to be compiled from other data sources to complete the data set.
* Team assignments and completion data were gathered and verified through the attendance worksheets. Teams were assigned an ID (A:H) so that they can be anonymized in the final data set. Team letter assignments were recorded on the original attendance spreadsheets. This portion of the data collection process was completed by hand using an Excel spreadsheet.
* Each cohort data was stored in an individual spreadsheet and then exported into a .csv file for import into R.

## Preparing the Data for analysis

In this first R block, the individual cohort .csv data files are combined into a dataframe. Empty values are properly coded with "NA" and then duplicate records (rows) are removed from the dataset. Once all the dataframes were complete, they are combined together into a single dataframe. This dataframe is stored as 'Applicants.csv' for use in the quantitative analysis.

## Reading data and combining into one dataframe.  
## This block does not need to execute after initial data tidying.   
  
# Read Applicants data from cohort application files and remove observations marked duplicate.  
Applicants12 <- read.csv("2012Applicants.csv", header=TRUE, sep=",", na.strings = c("", "NA"))  
Applicants12 <- filter(Applicants12, Accepted != "Duplicate")  
  
Applicants13 <- read.csv("2013Applicants.csv", header=TRUE, sep=",", na.strings = c("", "NA"))  
Applicants13 <- filter(Applicants13, Accepted != "Duplicate")  
  
Applicants14 <- read.csv("2014Applicants.csv", header=TRUE, sep=",", na.strings = c("", "NA"))  
Applicants14 <- filter(Applicants14, Accepted != "Duplicate")  
  
Applicants15 <- read.csv("2015Applicants.csv", header=TRUE, sep=",", na.strings = c("", "NA"))  
Applicants15 <- filter(Applicants15, Accepted != "Duplicate")  
  
Applicants16 <- read.csv("2016Applicants.csv", header=TRUE, sep=",", na.strings = c("", "NA"))  
Applicants16 <- filter(Applicants16, Accepted != "Duplicate")  
  
# Merge Applicants from individual cohort Applicants tables into one 'Applicants' table.  
Applicants <- rbind(Applicants16, Applicants15)  
Applicants <- rbind(Applicants, Applicants14)  
Applicants <- rbind(Applicants, Applicants13)  
Applicants <- rbind(Applicants, Applicants12)  
  
str(Applicants)

## Tidying the Data

After reviewing the data with 'str(Applicants)' it is apparant the factors are not consistent for all variables. This may be due to the fact that as the application form was edited, different spellings for various factors were used. This block cleans up the inconsistencies for each of the factors and shortens some of the factor labels so that they are easier to manipulate and display in tables.

## Cleaning up factor labels so that analysis will be consistent.  
## This block does not need to execute after initial data tidying.   
  
# Create degree indicators that are one word in length.  
Applicants$Degree <- str\_replace\_all(Applicants$Degree, "GED or High school diploma", "HS")  
Applicants$Degree <- str\_replace\_all(Applicants$Degree, "undergrad", "HS")  
Applicants$Degree <- str\_replace\_all(Applicants$Degree, "High school graduate", "HS")  
Applicants$Degree <- str\_replace\_all(Applicants$Degree, "Associate's degree", "Associate")  
Applicants$Degree <- str\_replace\_all(Applicants$Degree, "A.A.", "Associate")  
Applicants$Degree <- str\_replace\_all(Applicants$Degree, "Bachelor's degree", "Bachelor")  
Applicants$Degree <- str\_replace\_all(Applicants$Degree, "B.A.", "Bachelor")  
Applicants$Degree <- str\_replace\_all(Applicants$Degree, "B.S.", "Bachelor")  
Applicants$Degree <- str\_replace\_all(Applicants$Degree, "B.F.A.", "Bachelor")  
Applicants$Degree <- str\_replace\_all(Applicants$Degree, "Master's degree", "Master")  
Applicants$Degree <- str\_replace\_all(Applicants$Degree, "M.S.", "Master")  
Applicants$Degree <- str\_replace\_all(Applicants$Degree, "M.A.", "Master")  
Applicants$Degree <- str\_replace\_all(Applicants$Degree, "MBA", "Master")  
Applicants$Degree <- str\_replace\_all(Applicants$Degree, "M.B.A.", "Master")  
Applicants$Degree <- str\_replace\_all(Applicants$Degree, "J.D.", "Master")  
Applicants$Degree <- str\_replace\_all(Applicants$Degree, "Ph.D.", "Doctorate")  
Applicants$Degree <- str\_replace\_all(Applicants$Degree, "PhD", "Doctorate")  
Applicants$Degree <- factor(Applicants$Degree, levels=c("HS", "Associate", "Bachelor", "Master", "Doctorate"), ordered=TRUE)  
  
#Update Enties for questions which are parsing into multiple factors  
Applicants$Q5a <- str\_replace\_all(Applicants$Q5a, "Gain self confidence", "Gain self-confidence")  
Applicants$Q5a <- factor(Applicants$Q5a)  
Applicants$Q5f <- str\_replace\_all(Applicants$Q5f, "All of the above.", "All of the above")  
Applicants$Q5f <- factor(Applicants$Q5f)  
Applicants$Q8 <- str\_replace\_all(Applicants$Q8, "Yes (Please answer Question 11.)", "Yes")  
Applicants$Q8 <- str\_replace\_all(Applicants$Q8, "Yes (Please answer Question 9.)", "Yes")  
  
# Remove Duplicate factors and ensure variables are factors.  
Applicants$Accepted <- factor(Applicants$Accepted, levels=c("Yes", "No"), ordered=TRUE)  
Applicants$Decision <- factor(Applicants$Decision, levels=c("Yes", "No"), ordered=TRUE)  
Applicants$Finished <- factor(Applicants$Finished, levels=c("Yes", "No"), ordered=TRUE)  
Applicants$Q8 <- factor(Applicants$Q8)

## Annoynomizing the Data Set

The final data set has been cleansed to remove all identifying information so it is a truly anonymous data set for analysis. Only columns with respondant data will be included, all participant identifying information and all references to program name and university will be changed. The dataframe is saved as 'Applicants.csv'. For all future iterations of analysis, the data will be loaded from 'Applicants.csv', and previous blocks will not execute ('eval=FALSE'). This annonymized data set is the only version of the data that will be made available for analysis via GitHub.

\*\*[Notes:\*\*](Notes:**)

* All references to the program name, the university, and other identifiable indicators have been replaced with generic terms surrounded by asteriks. Ex. *program*, *university*, *company*, etc. This process was completed manually in Excel by reading through applicant responses to find all identifying data and replacing it with generic placeholders.

# Choosing the colums to be included in the annoymized data set.  
Applicants <- select(Applicants, Identifier, TeamID, Cohort, Accepted, Decision, Finished, Team, Degree, Experience, contains("Q"))  
  
# Save the file for later analysis.   
write.csv(Applicants, file="Applicants.csv")  
save(Summative, file = "data/Summative.Rda")

# Summative Evaluation

This code can be used to Tidy individual years, or the aggregated data. To analyze an individual year, comment out the lines that combine the data sets.

* CodeBook.xlsx contains the variable names (question numbers), the text of the corresponding question, and an inventory for which years the question appeared on the survey.
* Summative.csv contains the combined survey data.
* Individual files 2013Summative.csv through 2016Summative.csv were created to hold individual year data.
* The finished dataframe Summative.Rda has been exported for use in the analysis.

## Compiling the Data

* For cohorts 2013 - 2015 the surveys were completed in paper and pencil format. A matching survey was created in Qualtrics and survey responses were hand entered and exported as a .csv file.
* For 2016, the survey results were collected in Survey Monkey, results were exported as a .csv file.
* Survey responses for 2012 are missing from the available data.

While many of the questions in the 2016 survey are identical to 2013 - 2015, the export from SurveyMonkey was substantially different. To prepare the data for use in this analysis, variable names had to be hand entered to match those from the Qualtrics import. Likert Scale responses all used the same 5 point scale, but were configured to use different indicators. These differences are noted in the code book.

## Summative Data Code Book

# Read the data file.  
  
# Cohort 2016 has 40 completers, and 42 surveys.  
S2016 <- read.csv("./data/2016\_Summative.csv", sep = ",", header = TRUE)  
nrow(S2016)

## [1] 42

# Cohort 2015 has 47 completers, 5 missing surveys.  
S2015 <- read.csv("./data/2015\_Summative.csv", sep = ",", header = TRUE)  
nrow(S2015)

## [1] 42

# Cohort 2014 has 42 completers, 8 missing surveys.  
S2014 <- read.csv("./data/2014\_Summative.csv", sep = ",", header = TRUE)  
nrow(S2014)

## [1] 34

# Cohort 2013 has 41 completers, two did not attend 11/12 session, 6 missing surveys.  
S2013 <- read.csv("./data/2013\_Summative.csv", sep = ",", header = TRUE)  
nrow(S2013)

## [1] 33

# Convert rename column one and convert to character  
colnames(S2013)[colnames(S2013) == "ï..V1"] <- "V1"  
colnames(S2014)[colnames(S2014) == "ï..V1"] <- "V1"  
colnames(S2015)[colnames(S2015) == "ï..V1"] <- "V1"  
colnames(S2016)[colnames(S2016) == "ï..V1"] <- "V1"  
S2013[,1] <- as.character(S2013[,1])  
S2014[,1] <- as.character(S2014[,1])  
S2015[,1] <- as.character(S2015[,1])  
S2016[,1] <- as.character(S2016[,1])  
  
# Review the data file variables.  
# str(S2015)  
  
# Read the code book.   
CodeBook <- read.xlsx("./data/Summative\_CodeBook.xlsx", sheetIndex=1, header = TRUE)  
# CodeBook = data.table(CodeBook)  
  
# View the code book.  
# head(CodeBook)

The following code block coverts the numerical response for Question 6: "Who is your mentor?" to the team codes. These codes are assigned to maintain anonymity of teams and mentors. The 2016 Mentors were updated by hand due to the data being encoded using full text descriptions.

# Question 6 Factor: "Who was your mentor?"  
# Note this question uses the TeamID which matches the TeamID in the applicant.csv file.  
  
# 2016 Mentors  
# These were updated using Excel. The SurveyMonkey version of the datafile included all names and technologies. To maintain anonymity, they were updated before importing the data into this analyisis.  
  
# 2015 Mentors  
S2015$Q6[S2015$Q6 == "1"] <- "2015TF"  
S2015$Q6[S2015$Q6 == "2"] <- "2015TG"  
S2015$Q6[S2015$Q6 == "3"] <- "2015TD"  
S2015$Q6[S2015$Q6 == "4"] <- "2015TA"  
S2015$Q6[S2015$Q6 == "5"] <- "2015TC"  
S2015$Q6[S2015$Q6 == "6"] <- "2015TB"  
S2015$Q6[S2015$Q6 == "7"] <- "2015TU" # This is unknown in the survey. n=0  
S2015$Q6[S2015$Q6 == "8"] <- "2015TE"  
S2015$Q6 = factor(S2015$Q6,levels=c("2015TA","2015TB","2015TC", "2015TD", "2015TE", "2015TF", "2015TG", "2015TU"), ordered=TRUE)  
  
# 2014 Mentors  
S2014$Q6[S2014$Q6 == "1"] <- "2014TA"  
S2014$Q6[S2014$Q6 == "2"] <- "2014TG"  
S2014$Q6[S2014$Q6 == "3"] <- "2014TF"  
S2014$Q6[S2014$Q6 == "4"] <- "2014TB"  
S2014$Q6[S2014$Q6 == "5"] <- "2014TC"  
S2014$Q6[S2014$Q6 == "6"] <- "2014TD"  
S2014$Q6[S2014$Q6 == "7"] <- "2014TU" # This is unknown in the survey. n=0  
S2014$Q6[S2014$Q6 == "8"] <- "2014TE"  
S2014$Q6 = factor(S2014$Q6,levels=c("2014TA","2014TB","2014TC", "2014TD", "2014TE", "2014TF", "2014TG", "2014TU"), ordered=TRUE)  
  
# 2013 Mentors  
S2013$Q6[S2013$Q6 == "1"] <- "2013TD"  
S2013$Q6[S2013$Q6 == "2"] <- "2013TC"  
S2013$Q6[S2013$Q6 == "3"] <- "2013TA"  
S2013$Q6[S2013$Q6 == "4"] <- "2013TG"  
S2013$Q6[S2013$Q6 == "5"] <- "2013TB"  
S2013$Q6[S2013$Q6 == "6"] <- "2013TF"  
S2013$Q6[S2013$Q6 == "7"] <- "2013TE" # This is unknown in the survey. n=0  
S2013$Q6[S2013$Q6 == "8"] <- "2013TU"  
S2013$Q6 = factor(S2013$Q6,levels=c("2013TA","2013TB","2013TC", "2013TD", "2013TE", "2013TF", "2013TG", "2013TU"), ordered=TRUE)

Combine the individual survey response files into one complete Summative dataframe. This will allow cumulative data analysis as well as comparisons across cohorts.

# Combine files into one Summative dataframe  
Summative <- bind\_rows(S2015, S2014)

## Warning in bind\_rows\_(x, .id): Unequal factor levels: coercing to character  
  
## Warning in bind\_rows\_(x, .id): Unequal factor levels: coercing to character  
  
## Warning in bind\_rows\_(x, .id): Unequal factor levels: coercing to character  
  
## Warning in bind\_rows\_(x, .id): Unequal factor levels: coercing to character  
  
## Warning in bind\_rows\_(x, .id): Unequal factor levels: coercing to character  
  
## Warning in bind\_rows\_(x, .id): Unequal factor levels: coercing to character  
  
## Warning in bind\_rows\_(x, .id): Unequal factor levels: coercing to character  
  
## Warning in bind\_rows\_(x, .id): Unequal factor levels: coercing to character  
  
## Warning in bind\_rows\_(x, .id): Unequal factor levels: coercing to character  
  
## Warning in bind\_rows\_(x, .id): Unequal factor levels: coercing to character

Summative <- bind\_rows(Summative, S2013)

The following code block coverts the numerical response items into factor response items that will be easier to analyze. This block only affects 2013 - 2015 data as 2016 already has the factored response items for these questions.

# Question 20 Factor: "Time alloted each week for speakers was:"  
Summative$Q13[Summative$Q13 == "1"] <- "Too Short"  
Summative$Q13[Summative$Q13 == "2"] <- "About Right"  
Summative$Q13[Summative$Q13 == "3"] <- "Too Long"  
Summative$Q13 = factor(Summative$Q13,levels=c("Too Short","About Right","Too Long"), ordered=TRUE)  
  
# Question 14 Factor: "Time alloted each week for teamwork was:"  
Summative$Q14[Summative$Q14 == "1"] <- "Too Short"  
Summative$Q14[Summative$Q14 == "2"] <- "About Right"  
Summative$Q14[Summative$Q14 == "3"] <- "Too Long"  
Summative$Q14 = factor(Summative$Q14,levels=c("Too Short","About Right","Too Long"), ordered=TRUE)  
  
# Question 19 Factor: "How many hours a week on average did your team meet ourside of the program?"  
Summative$Q19[Summative$Q19 == "1"] <- "1-2 hrs"  
Summative$Q19[Summative$Q19 == "2"] <- "1-2 hrs"  
Summative$Q19[Summative$Q19 == "3"] <- "3-5 hrs"  
Summative$Q19[Summative$Q19 == "4"] <- "3-5 hrs"  
Summative$Q19[Summative$Q19 == "5"] <- "More than 5 hrs"  
Summative$Q19 = factor(Summative$Q19,levels=c("1-2 hrs","2-3 hrs","3-4 hrs", "4-5 hrs", "More than 5 hrs"), ordered=TRUE)  
  
# Question 20 Factor: "The programs duration was:"  
Summative$Q20[Summative$Q20 == "1"] <- "Too Short"  
Summative$Q20[Summative$Q20 == "2"] <- "About Right"  
Summative$Q20[Summative$Q20 == "3"] <- "Too Long"  
Summative$Q20 = factor(Summative$Q20,levels=c("Too Short","About Right","Too Long"), ordered=TRUE)  
  
# Question 21 integer: Convert to range of values  
Summative$Q21[Summative$Q21 <= "2"] <- "1-2 hrs"  
Summative$Q21[Summative$Q21 <= "5"] <- "3-5 hrs"  
Summative$Q21[Summative$Q21 >= "6"] <- "More than 5 hrs"  
Summative$Q21 = factor(Summative$Q21,levels=c("1-2 hrs", "3-5 hrs", "More than 5 hrs"), ordered=TRUE)  
  
# Question 22 Factor: "Would you recommend this program to other women?"  
Summative$Q22[Summative$Q22 == "1"] <- "Yes"  
Summative$Q22[Summative$Q22 == "2"] <- "No"  
Summative$Q22[Summative$Q22 == "3"] <- "Unsure"  
Summative$Q22 = factor(Summative$Q22,levels=c("Yes","No","Unsure"), ordered=TRUE)  
  
# Question 24 Factor: "What is your highest level of education completed?"  
Summative$Q24[Summative$Q24 == "1"] <- "HS"  
Summative$Q24[Summative$Q24 == "2"] <- "Associate"  
Summative$Q24[Summative$Q24 == "3"] <- "Bachelor"  
Summative$Q24[Summative$Q24 == "4"] <- "Master"  
Summative$Q24[Summative$Q24 == "5"] <- "PhD"  
Summative$Q24[Summative$Q24 == "6"] <- "Other"  
Summative$Q24 = factor(Summative$Q24,levels=c("HS","Associate","Bachelor", "Master", "PhD", "Other" ), ordered=TRUE)  
  
# Question 29 Factor: "What is your area of expertise?"  
Summative$Q29[Summative$Q29 == "1"] <- "Finance"  
Summative$Q29[Summative$Q29 == "2"] <- "Business"  
Summative$Q29[Summative$Q29 == "3"] <- "Science"  
Summative$Q29[Summative$Q29 == "4"] <- "Engineering"  
Summative$Q29[Summative$Q29 == "5"] <- "Computer/IT"  
Summative$Q29[Summative$Q29 == "6"] <- "Marketing/Communications/Design"  
Summative$Q29[Summative$Q29 == "7"] <- "Other"  
Summative$Q29 = factor(Summative$Q29,levels=c("Finance","Business","Science", "Engineering", "Computer/IT", "Marketing/Communications/Design", "Other" ), ordered=TRUE)  
  
# Question 30 Factor: "What is your age?"  
Summative$Q30[Summative$Q30 == "1"] <- "18-24"  
Summative$Q30[Summative$Q30 == "2"] <- "25-34"  
Summative$Q30[Summative$Q30 == "3"] <- "35-44"  
Summative$Q30[Summative$Q30 == "4"] <- "45-54"  
Summative$Q30[Summative$Q30 == "5"] <- "55-64"  
Summative$Q30[Summative$Q30 == "7"] <- "65-74"  
Summative$Q30[Summative$Q30 == "6"] <- "75 or older"  
Summative$Q30 = factor(Summative$Q30,levels=c("18-24","25-34","35-44", "45-54", "55-64", "75 or older"), ordered=TRUE)  
  
# Question 31 Factor: "Choose the answer that best describes your current situation:"  
Summative$Q31[Summative$Q31 == "1"] <- "married or in a committed relationship with no children"  
Summative$Q31[Summative$Q31 == "2"] <- "married or in a committed relationship with grown children (18+)"  
Summative$Q31[Summative$Q31 == "3"] <- "married or in a committed relationship with school aged children (5-18)"  
Summative$Q31[Summative$Q31 == "4"] <- "married or in a committed relationship with younger child/ren (under 5)"  
Summative$Q31[Summative$Q31 == "5"] <- "single with no children"  
Summative$Q31[Summative$Q31 == "6"] <- "single parent with grown children (18+)"  
Summative$Q31[Summative$Q31 == "7"] <- "single parent with school aged children (5-18)"  
Summative$Q31[Summative$Q31 == "7"] <- "single parent with younger child/ren (under 5)"  
Summative$Q31 = factor(Summative$Q31, ordered=TRUE)  
  
# Question 32 Factor: "Which of the following best represents your racial or ethnic heritage?"  
Summative$Q32[Summative$Q32 == "1"] <- "Non-Hispanic White or Euro-American"  
Summative$Q32[Summative$Q32 == "2"] <- "Black, Afro-Caribbean, or African American"  
Summative$Q32[Summative$Q32 == "3"] <- "Latino or Hispanic American"  
Summative$Q32[Summative$Q32 == "4"] <- "East Asian or Asian American"  
Summative$Q32[Summative$Q32 == "5"] <- "South Asian or Indian American"  
Summative$Q32[Summative$Q32 == "6"] <- "Middle Eastern or Arab American"  
Summative$Q32[Summative$Q32 == "7"] <- "Native American or Alaskan Native"  
Summative$Q32[Summative$Q32 == "8"] <- "Other"  
Summative$Q32 = as.factor(Summative$Q32)  
  
# Question 33 Factor: "What was your total household income before taxes during the past 12 months?"  
Summative$Q33[Summative$Q33 == "1"] <- "Less than $25,000"  
Summative$Q33[Summative$Q33 == "2"] <- "$25,000 to $34,999"  
Summative$Q33[Summative$Q33 == "3"] <- "$35,000 to $49,999"  
Summative$Q33[Summative$Q33 == "4"] <- "$50,000 to $74,999"  
Summative$Q33[Summative$Q33 == "5"] <- "$75,000 to $99,999"  
Summative$Q33[Summative$Q33 == "6"] <- "$100,000 to $149,999"  
Summative$Q33[Summative$Q33 == "7"] <- "$150,000 or more"  
Summative$Q33 = factor(Summative$Q33, levels=c("Less than $25,000", "$25,000 to $34,999", "$35,000 to $49,999", "$50,000 to $74,999", "$75,000 to $99,999", "$100,000 to $149,999", "$150,000 or more"), ordered=TRUE)  
  
# Question 34 Factor: "Please circle the option(s) that best describe(s) your current situation. Ok to choose more than one if applicable."  
Summative$Q34[Summative$Q34 == "1"] <- "Master's student"  
Summative$Q34[Summative$Q34 == "2"] <- "MBA student"  
Summative$Q34[Summative$Q34 == "3"] <- "MD student"  
Summative$Q34[Summative$Q34 == "4"] <- "PhD student"  
Summative$Q34[Summative$Q34 == "5"] <- "Postdoc"  
Summative$Q34[Summative$Q34 == "6"] <- "Unemployed (not a student)"  
Summative$Q34[Summative$Q34 == "7"] <- "Work part-time (not a student)"  
Summative$Q34[Summative$Q34 == "8"] <- "Employed at a technology startup"  
Summative$Q34[Summative$Q34 == "9"] <- "Employed at a non-technology startup"  
Summative$Q34[Summative$Q34 == "10"] <- "Employed at a technology non-startup company"  
Summative$Q34[Summative$Q34 == "11"] <- "Employed at a non-technology non-startup company"  
Summative$Q34[Summative$Q34 == "12"] <- "Owned my own technology business"  
Summative$Q34[Summative$Q34 == "13"] <- "Owned my own non-technology business"  
Summative$Q33 = as.factor(Summative$Q33)  
  
# Update Likert scale questions to represent NA as NA (Qualtrics exported it as #6)  
  
Summative$Q1\_1[Summative$Q1\_1 == "6"] <- NA  
Summative$Q4\_1[Summative$Q4\_1 == "6"] <- NA  
Summative$Q7\_1[Summative$Q7\_1 == "6"] <- NA  
Summative$Q7\_2[Summative$Q7\_2 == "6"] <- NA  
Summative$Q7\_3[Summative$Q7\_3 == "6"] <- NA  
Summative$Q9\_1[Summative$Q9\_1 == "6"] <- NA  
Summative$Q9\_2[Summative$Q9\_2 == "6"] <- NA  
Summative$Q9\_3[Summative$Q9\_3 == "6"] <- NA  
Summative$Q9\_4[Summative$Q9\_4 == "6"] <- NA  
Summative$Q9\_5[Summative$Q9\_5 == "6"] <- NA

# Add 2016 data to Summarize  
Summative <- bind\_rows(Summative, S2016)

## Warning in bind\_rows\_(x, .id): Unequal factor levels: coercing to character  
  
## Warning in bind\_rows\_(x, .id): Unequal factor levels: coercing to character  
  
## Warning in bind\_rows\_(x, .id): Unequal factor levels: coercing to character  
  
## Warning in bind\_rows\_(x, .id): Unequal factor levels: coercing to character  
  
## Warning in bind\_rows\_(x, .id): Unequal factor levels: coercing to character  
  
## Warning in bind\_rows\_(x, .id): Unequal factor levels: coercing to character  
  
## Warning in bind\_rows\_(x, .id): Unequal factor levels: coercing to character  
  
## Warning in bind\_rows\_(x, .id): Unequal factor levels: coercing to character

# Rename ID Column  
Summative <- rename(Summative, ID = V1)  
  
# Rename Cohort column and factor  
Summative <- rename(Summative, Cohort = Q1)  
Summative$Cohort <- factor(Summative$Cohort, levels=c("2013", "2014", "2015", "2016"), ordered=TRUE)  
  
# Rename Degree column and clean up factor levels  
Summative <- rename(Summative, Degree = Q24)  
Summative$Degree[Summative$Degree == "High School Diploma"] <- "HS"  
Summative$Degree[Summative$Degree == "Master's Degree"] <- "Master"  
Summative$Degree[Summative$Degree == "Bachelor's Degree"] <- "Bachelor"  
Summative$Degree[Summative$Degree == "Ph.D."] <- "PhD"  
Summative$Degree = factor(Summative$Degree,levels=c("HS","Associate","Bachelor", "Master", "PhD", "Other" ), ordered=TRUE)  
  
# Rename Team Column, Survey ID, Age, Race, Income, Relationship  
Summative <- rename(Summative, Team = Q6)  
Summative <- rename(Summative, PaperID = Q25)  
Summative <- rename(Summative, Discipline = Q29)  
Summative <- rename(Summative, Age = Q30)  
Summative <- rename(Summative, Relationship = Q31)  
Summative <- rename(Summative, Race = Q32)  
Summative <- rename(Summative, Income = Q33)  
Summative <- rename(Summative, Employment = Q34)  
  
# Rename Questions. Use TEXT to identify open-ended response items.  
Summative <- rename(Summative, Q1 = Q1\_1)  
Summative <- rename(Summative, Q4 = Q4\_1)  
Summative <- rename(Summative, TEXT2 = Q2)  
Summative <- rename(Summative, TEXT5 = Q5)  
Summative <- rename(Summative, TEXT8 = Q8)  
Summative <- rename(Summative, TEXT10 = Q10)  
Summative <- rename(Summative, TEXT12 = Q12)  
Summative <- rename(Summative, TEXT35 = Q35)  
Summative <- rename(Summative, TEXT17 = Q17)  
Summative <- rename(Summative, TEXT23 = Q23)  
Summative <- rename(Summative, TEXT24 = Q24\_TEXT)  
Summative <- rename(Summative, TEXT26 = Q26)  
Summative <- rename(Summative, TEXT29 = Q29\_TEXT)

# Saving the data for later analysis.  
write.csv(Summative, file="data/Summative.csv")  
save(Summative, file = "data/Summative.Rda")