# EDUCE survey data cleaning

*Kim Dill-McFarland*

*version July 11, 2019*

## Setup

### Load packages

```
# Data manipulation
library(tidyverse)
```

### R session

```
sessionInfo()
```

```
## R version 3.5.3 (2019-03-11)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS Mojave 10.14.5
##
## Matrix products: default
## BLAS: /Library/Frameworks/R.framework/Versions/3.5/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.5/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_CA.UTF-8/en_CA.UTF-8/en_CA.UTF-8/C/en_CA.UTF-8/en_CA.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] forcats_0.4.0   stringr_1.4.0   dplyr_0.8.3     purrr_0.3.2
## [5] readr_1.3.1     tidyr_0.8.3     tibble_2.1.3    ggplot2_3.2.0
## [9] tidyverse_1.2.1
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.1       cellranger_1.1.0 pillar_1.4.2     compiler_3.5.3
##  [5] tools_3.5.3      zeallot_0.1.0    digest_0.6.20    lubridate_1.7.4
##  [9] jsonlite_1.6     evaluate_0.14    nlme_3.1-140     gtable_0.3.0
## [13] lattice_0.20-38  pkgconfig_2.0.2  rlang_0.4.0      cli_1.1.0
## [17] rstudioapi_0.10  yaml_2.2.0       haven_2.1.1      xfun_0.8
## [21] withr_2.1.2      xml2_1.2.0       httr_1.4.0       knitr_1.23
## [25] vctrs_0.2.0      generics_0.0.2   hms_0.5.0        grid_3.5.3
## [29] tidyselect_0.2.5 glue_1.3.1       R6_2.4.0         readxl_1.3.1
## [33] rmarkdown_1.13   modelr_0.1.4     magrittr_1.5     backports_1.1.4
## [37] scales_1.0.0     htmltools_0.3.6  rvest_0.3.4      assertthat_0.2.1
## [41] colorspace_1.4-1 stringi_1.4.3    lazyeval_0.2.2   munsell_0.5.0
## [45] broom_0.5.2      crayon_1.3.4
```

# Data cleaning: Fall 2017 surveys

Surveys were collected in FluidSurveys in Fall 2017 prior to UBC's switch to Qualtrics in Spring 2018. ## Load raw data Load pre- and post-survey data from Fall 2017.

```
pre2017 <- read_tsv("data_raw/2017Fall_EDUCE_pre-survey.tsv",
                    na=c("na", "NA","n/a","N/A"))
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   `Internal ID` = col_double(),
##   Username = col_logical(),
##   `GET Variables` = col_logical(),
##   `Number of Saves` = col_logical(),
##   `Weighted Score` = col_double(),
##   `Completion Time` = col_time(format = ""),
##   `Invite Code` = col_logical(),
##   `Invite Email` = col_logical(),
##   `Invite Name` = col_logical(),
##   Collector = col_logical(),
##   `For which course are you completing this survey? [MICB301: Microbial Ecophysiology]` = col_double
##   `For which course are you completing this survey? [MICB405: Bioinformatics]` = col_double(),
##   `For which course are you completing this survey? [MICB421: Experimental Microbiology]` = col_logi
##   `For which course are you completing this survey? [MICB425: Microbial Ecological Genomics]` = col_
##   `Please indicate which interfaces and/or programs you have used prior to this course. [Unix comman
##   `Please indicate which interfaces and/or programs you have used prior to this course. [R]` = col_d
##   `Please indicate which interfaces and/or programs you have used prior to this course. [mothur]` = 
##   `Please indicate which interfaces and/or programs you have used prior to this course. [QIIME/QIIME
##   `Please indicate which interfaces and/or programs you have used prior to this course. [Metagenomic
##   `Please indicate which interfaces and/or programs you have used prior to this course. [None of the
##   # ... with 4 more columns
## )
```

```
## See spec(...) for full column specifications.
```

```
post2017 <- read_tsv("data_raw/2017Fall_EDUCE_post-survey.tsv",
                     na=c("na", "NA","n/a","N/A"))
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   `Internal ID` = col_double(),
##   `For which course are you completing this survey? [MICB301: Microbial Ecophysiology]` = col_double
##   `For which course are you completing this survey? [MICB405: Bioinformatics]` = col_double(),
##   `Introductory [Introduction to R]` = col_double(),
##   `Introductory [Data carpentry: Basic data organization, manipulation, and visualization in R]` = co
##   `Introductory [Software carpentry: Programming in Python, the Unix shell, and version control with
##   `Intermediate [Working with tabular data in R]` = col_double(),
##   `Intermediate [Graphics with ggplot in R]` = col_double(),
##   `Intermediate [Statistical models in R]` = col_double(),
##   `Intermediate [Functions, workflows, and reproducible research in R and Git]` = col_double(),
##   `Microbiome [Amplicon sequence data analysis in mothur or QIIME]` = col_double(),
##   `Microbiome [Microbiome analysis in R]` = col_double(),
##   `Microbiome [Metagenomics analysis]` = col_double(),
##   `Microbiome [Metatranscriptomics analysis]` = col_double()
## )
## See spec(...) for full column specifications.
```

## Clean data

### Pre-surveys

```
pre2017_clean <- pre2017 %>%
  # Select vars of interest
  select("Updated At", "Weighted Score", "Internal ID",
         "For which course are you completing this survey? [MICB301: Microbial Ecophysiology]":
           "If so, please specify._1") %>%
  rename(RecordedDate = "Updated At",
         Progress = "Weighted Score",
         InternalID = "Internal ID",
         MICB301 =
        "For which course are you completing this survey? [MICB301: Microbial Ecophysiology]",
         MICB405 =
        "For which course are you completing this survey? [MICB405: Bioinformatics]",
         MICB421 =
        "For which course are you completing this survey? [MICB421: Experimental Microbiology]",
         MICB425 =
        "For which course are you completing this survey? [MICB425: Microbial Ecological Genomics]",
         ID =
        "Please create an anonymous identifier that is the first three letters of your mother’s first na
         Pre_DS = "Have you heard the term "data science"?",
         Pre_DS_define = "How would you define data science?",
         Pre_DS_where =
        "Where have you been exposed to data science? Courses, popular media, books, etc.",
         Pre_BI = "Have you heard the term "bioinformatics"?",
         Pre_BI_define = "How would you define bioinformatics?",
         Pre_BI_where =
        "Where have you been exposed to bioinformatics? Courses, popular media, books, etc.",
         Prev_unix =
        "Please indicate which interfaces and/or programs you have used prior to this course. [Unix com
         Prev_R =
        "Please indicate which interfaces and/or programs you have used prior to this course. [R]",
         Prev_mothur =
    "Please indicate which interfaces and/or programs you have used prior to this course. [mothur]",
         Prev_QIIME =
        "Please indicate which interfaces and/or programs you have used prior to this course. [QIIME/QI
         Prev_metaG =
  "Please indicate which interfaces and/or programs you have used prior to this course. [Metagenomic to
         Prev_none =
        "Please indicate which interfaces and/or programs you have used prior to this course. [None of
         Pre_Interest_MICB = "How would you rate your interest in  | microbiology?",
         Pre_Interest_BI = "How would you rate your interest in  | bioinformatics?",
         Pre_Interest_STAT = "How would you rate your interest in  | statistics?",
         Pre_Interest_CPSC = "How would you rate your interest in  | computer science?",
         Pre_Exp_MICB = "What level of experience do you have in | microbiology?",
         Pre_Exp_BI = "What level of experience do you have in | bioinformatics?",
         Pre_Exp_STAT = "What level of experience do you have in | statistics?",
         Pre_Exp_CPSC = "What level of experience do you have in | computer science?",
         Major = "What is your academic major?",
         Prev_EDUCE_YN = "Have you previously completed EDUCE module(s)?",
         Prev_EDUCE_MICB301 = "If so, in which course(s)? [MICB301: Microbial Ecophysiology]",
         Prev_EDUCE_MICB405 = "If so, in which course(s)? [MICB405: Bioinformatics]",
         Prev_EDUCE_MICB421 = "If so, in which course(s)? [MICB421: Experimental Microbiology]",
         Prev_EDUCE_MICB425 = "If so, in which course(s)? [MICB425: Microbial Ecological Genomics]",
         Pre_MICB301 = "Microbiology | MICB301: Microbial Ecophysiology",
         Pre_MICB405 = "Microbiology | MICB405: Bioinformatics",
         Pre_MICB421 = "Microbiology | MICB421: Experimental Microbiology",
         Pre_MICB425 = "Microbiology | MICB425: Microbial Ecological Genomics",
         Pre_STAT200 = "Statistics | STAT200: Elementary Statistics for Applications",
         Pre_STAT203 = "Statistics | STAT203: Statistical Methods",
         Pre_STAT241 = "Statistics | STAT241: Introductory Probability and Statistics",
         Pre_STAT251 = "Statistics | STAT251: Elementary Statistics",
         Pre_BIOL300 = "Statistics | BIOL300: Introduction to Biostatistics",
         Pre_CPSC100 = "Computer science | CPSC100: Computational Thinking",
         Pre_CPSC101 = "Computer science | CPSC101: Introduction to Systematic Program Design",
         Pre_CPSC110 = "Computer science | CPSC110: Computation, Programs, and Programming",
         Pre_CPSC121 = "Computer science | CPSC121: Models of Computation",
         Pre_CPSC301 = "Computer science | CPSC301: Computing in the Life Sciences",
         Pre_other_complete =
    "Please list any other relevant courses which you completed previously. This could be from the depa
         Pre_other_current =
    "Please list any other relevant courses in which you are currently enrolled. This could be from the
         Pre_other_future =
    "Please list any other relevant courses in which you plan to enroll. This could be from the departm
         Prev_cocurric_YN =
    "Have you participated in any data science-related activities outside of class? Workshops, hackatho
         Prev_cocurric = "If so, please specify.",
         Pre_Interest_cocurric_YN =
    "Are you interested in participating in workshops outside of class to develop or continue to develo
         Pre_Know_cocurric_YN = "Are you aware of any such workshops currently offered at UBC?",
         Pre_Know_cocurric = "If so, please specify._1") %>%
  # Reformt Previous experience values to remove description
  separate(Pre_Exp_BI, into="Pre_Exp_BI", sep=" ", extra="drop") %>%
  separate(Pre_Exp_CPSC, into="Pre_Exp_CPSC", sep=" ", extra="drop") %>%
  separate(Pre_Exp_STAT, into="Pre_Exp_STAT", sep=" ", extra="drop") %>%
  separate(Pre_Exp_MICB, into="Pre_Exp_MICB", sep=" ", extra="drop") %>%
  # Change Very high to just high to match 2018 data
  mutate_at(vars(starts_with("Pre_Exp_")), funs(ifelse(. == "Very", "High", .))) %>%
  # Gather course ID into 1 variable
  gather(key="Course", value="value", MICB301:MICB425) %>%
  drop_na(value) %>%
  select(-value) %>%
  # Separate date and time
  separate(RecordedDate, into=c("Date", "time"), sep=" ")  %>%
  mutate(Date = as.Date(Date)) %>%
  select(-time) %>%
  # Force all ids to be upper case
  mutate(ID = toupper(ID)) %>%
  # Remove incomplete surveys
  filter(Progress >= 70) %>%
  # Remove test surveys
  filter(!(ID %in% c("WOW1234", "NAN1828"))) %>%
  # Remove duplicate surveys, keeping newest response
  arrange(desc(Date)) %>%
  distinct(Course, ID, .keep_all=TRUE) %>%
  # Create categorical year
  mutate(year = ifelse(Date >= "2017-09-01" & Date <= "2018-06-01" ,
                       "2017/18",
                ifelse(Date >= "2018-09-01" & Date <= "2019-06-01",
                       "2018/19", NA))) %>%
  # Remove unneeded variables
  select(-c(Progress, Date))
```

```
## Warning: funs() is soft deprecated as of dplyr 0.8.0
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## This warning is displayed once per session.
```

```
## Warning: Expected 2 pieces. Additional pieces discarded in 152 rows [1, 2,
## 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].
```

*Warning confirms dropping of “PM” from end of times.*

Explore cleaned data

```
dim(pre2017_clean)
```

```
## [1] 142  52
```

```
colnames(pre2017_clean)
```

```
##  [1] "InternalID"               "ID"
##  [3] "Pre_DS"                   "Pre_DS_define"
##  [5] "Pre_DS_where"             "Pre_BI"
##  [7] "Pre_BI_define"            "Pre_BI_where"
##  [9] "Prev_unix"                "Prev_R"
## [11] "Prev_mothur"              "Prev_QIIME"
## [13] "Prev_metaG"               "Prev_none"
## [15] "Pre_Interest_MICB"        "Pre_Interest_BI"
## [17] "Pre_Interest_STAT"        "Pre_Interest_CPSC"
## [19] "Pre_Exp_MICB"             "Pre_Exp_BI"
## [21] "Pre_Exp_STAT"             "Pre_Exp_CPSC"
## [23] "Major"                    "Prev_EDUCE_YN"
## [25] "Prev_EDUCE_MICB301"       "Prev_EDUCE_MICB405"
## [27] "Prev_EDUCE_MICB421"       "Prev_EDUCE_MICB425"
## [29] "Pre_MICB301"              "Pre_MICB405"
## [31] "Pre_MICB421"              "Pre_MICB425"
## [33] "Pre_STAT200"              "Pre_STAT203"
## [35] "Pre_STAT241"              "Pre_STAT251"
## [37] "Pre_BIOL300"              "Pre_CPSC100"
## [39] "Pre_CPSC101"              "Pre_CPSC110"
## [41] "Pre_CPSC121"              "Pre_CPSC301"
## [43] "Pre_other_complete"       "Pre_other_current"
## [45] "Pre_other_future"         "Prev_cocurric_YN"
## [47] "Prev_cocurric"            "Pre_Interest_cocurric_YN"
## [49] "Pre_Know_cocurric_YN"     "Pre_Know_cocurric"
## [51] "Course"                   "year"
```

### Post-surveys

```
post2017_clean <- post2017 %>%
  # Rename variables
  rename(Progress = "Status",
         InternalID = "Internal ID",
         MICB301 =
           "For which course are you completing this survey? [MICB301: Microbial Ecophysiology]",
         MICB405 = "For which course are you completing this survey? [MICB405: Bioinformatics]",
         ID =
  "Please create an anonymous identifier that is the first three letters of your mother’s first name and
         Post_DS = "Have you heard the term "data science"?",
         Post_DS_define = "How would you define data science?",
         Post_DS_where =
    "Where have you been exposed to data science? Courses, popular media, books, etc.",
         Post_BI = "Have you heard the term "bioinformatics"?",
         Post_BI_define = "How would you define bioinformatics?",
         Post_BI_where =
    "Where have you been exposed to bioinformatics? Courses, popular media, books, etc.",
         Post_Interest_MICB = "How would you rate your interest in  | microbiology?",
         Post_Interest_BI = "How would you rate your interest in  | bioinformatics?",
         Post_Interest_STAT = "How would you rate your interest in  | statistics?",
         Post_Interest_CPSC = "How would you rate your interest in  | computer science?",
         Post_Exp_MICB = "What level of experience do you have in | microbiology?",
         Post_Exp_BI = "What level of experience do you have in | bioinformatics?",
         Post_Exp_STAT = "What level of experience do you have in | statistics?",
         Post_Exp_CPSC = "What level of experience do you have in | computer science?",
         Post_MICB301 = "Microbiology | MICB301: Microbial Ecophysiology",
         Post_MICB405 = "Microbiology | MICB405: Bioinformatics",
         Post_MICB421 = "Microbiology | MICB421: Experimental Microbiology",
         Post_MICB425 = "Microbiology | MICB425: Microbial Ecological Genomics",
         Post_STAT200 = "Statistics | STAT200: Elementary Statistics for Applications",
         Post_STAT203 = "Statistics | STAT203: Statistical Methods",
         Post_STAT241 = "Statistics | STAT241: Introductory Probability and Statistics",
         Post_STAT251 = "Statistics | STAT251: Elementary Statistics",
         Post_BIOL300 = "Statistics | BIOL300: Introduction to Biostatistics",
         Post_CPSC100 = "Computer science | CPSC100: Computational Thinking",
         Post_CPSC101 = "Computer science | CPSC101: Introduction to Systematic Program Design",
         Post_CPSC110 = "Computer science | CPSC110: Computation, Programs, and Programming",
         Post_CPSC121 = "Computer science | CPSC121: Models of Computation",
         Post_CPSC301 = "Computer science | CPSC301: Computing in the Life Sciences",
         Post_course_other =
    "Please list any other relevant courses in which you are more or less likely to enroll based on your
         Post_cocurric_YN =
    "Have you taken a data science workshop outside of class this semester?",
         Post_cocurric = "If so, please specify.",
         Post_Interest_cocurric_YN =
    "Are you interested in participating in workshops outside of class to develop or continue to develop
         Post_Interest_introR = "Introductory [Introduction to R]",
         Post_Interest_DC =
    "Introductory [Data carpentry: Basic data organization, manipulation, and visualization in R]",
         Post_Interest_SC =
    "Introductory [Software carpentry: Programming in Python, the Unix shell, and version control with G
         Post_Interest_data = "Intermediate [Working with tabular data in R]",
         Post_Interest_ggplot = "Intermediate [Graphics with ggplot in R]",
         Post_Interest_statmodel = "Intermediate [Statistical models in R]",
         Post_Interest_repro =
    "Intermediate [Functions, workflows, and reproducible research in R and Git]",
         Post_Interest_mothur_QIIME =
    "Microbiome [Amplicon sequence data analysis in mothur or QIIME]",
         Post_Interest_microbiomeR = "Microbiome [Microbiome analysis in R]",
         Post_Interest_metaG = "Microbiome [Metagenomics analysis]",
         Post_Interest_metaT = "Microbiome [Metatranscriptomics analysis]",
         Minority = "Do you identify as a racial or ethnic minority?",
         First_gen = "Do you identify as a first-generation university student?",
         Non_trad = "Do you identify as a non-traditional university student?",
         Addtl_comments = "Please use this space to provide any additional feedback.") %>%
  # Reformat post experience values to remove description
  separate(Post_Exp_BI, into="Post_Exp_BI", sep=" ", extra="drop") %>%
  separate(Post_Exp_CPSC, into="Post_Exp_CPSC", sep=" ", extra="drop") %>%
  separate(Post_Exp_STAT, into="Post_Exp_STAT", sep=" ", extra="drop") %>%
  separate(Post_Exp_MICB, into="Post_Exp_MICB", sep=" ", extra="drop") %>%
  # Change Very high to just high to match 2018 data
  mutate_at(vars(starts_with("Pre_Exp_")), funs(ifelse(. == "Very", "High", .))) %>%
# Gather course ID into 1 variable
  gather(key="Course", value="value", MICB301:MICB405) %>%
  drop_na(value) %>%
  select(-value) %>%
  # Create date group
  mutate(year = "2017/18") %>%
  # Force all ids to be upper case
  mutate(ID = toupper(ID)) %>%
  # Remove test surveys
  filter(!(ID %in% c("WOW1234", "NAN1828", ""))) %>%
  # Remove duplicate surveys, keeping newer response
  arrange(ID, desc(InternalID)) %>%
  distinct(Course, ID, .keep_all=TRUE) %>%
  # Remove unneeded variables
  select(-Progress)
```

Explore cleaned data

```
dim(post2017_clean)
```

```
## [1] 165  51
```

```
colnames(post2017_clean)
```

```
##  [1] "InternalID"                  "ID"
##  [3] "Post_DS"                     "Post_DS_define"
##  [5] "Post_DS_where"               "Post_BI"
##  [7] "Post_BI_define"              "Post_BI_where"
##  [9] "Post_Interest_MICB"        "Post_Interest_BI"
## [11] "Post_Interest_STAT"        "Post_Interest_CPSC"
## [13] "Post_Exp_MICB"             "Post_Exp_BI"
## [15] "Post_Exp_STAT"             "Post_Exp_CPSC"
## [17] "Post_MICB301"              "Post_MICB405"
## [19] "Post_MICB421"              "Post_MICB425"
## [21] "Post_STAT200"              "Post_STAT203"
## [23] "Post_STAT241"              "Post_STAT251"
## [25] "Post_BIOL300"              "Post_CPSC100"
## [27] "Post_CPSC101"              "Post_CPSC110"
## [29] "Post_CPSC121"              "Post_CPSC301"
## [31] "Post_course_other"         "Post_cocurric_YN"
## [33] "Post_cocurric"             "Post_Interest_cocurric_YN"
## [35] "Post_Interest_introR"      "Post_Interest_DC"
## [37] "Post_Interest_SC"          "Post_Interest_data"
## [39] "Post_Interest_ggplot"      "Post_Interest_statmodel"
## [41] "Post_Interest_repro"       "Post_Interest_mothur_QIIME"
## [43] "Post_Interest_microbiomeR" "Post_Interest_metaG"
## [45] "Post_Interest_metaT"       "Minority"
## [47] "First_gen"                 "Non_trad"
## [49] "Addtl_comments"            "Course"
## [51] "year"
```

# Data cleaning: 2018-2019 surveys

In 2018, UBC switched survey tools from FluidSurvey to Qualtrics. All surveys from 2018 forward were collected in a single Qualtrics survey form.

## Load pre- and post-survey data from 2018-2019.

Load remaining survey data, Spring 2018 - Spring 2019.

```
pre2018 = read_tsv("data_raw/20190325_EDUCE_pre-course_survey.tsv",
                na=c("na", "NA","n/a","N/A"))
```

```
## Warning: Duplicated column names deduplicated: 'Q17' => 'Q17_1' [35]
```

```
## Parsed with column specification:
## cols(
##   .default = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

```
post2018 = read_tsv("data_raw/20190325_EDUCE_post-course_survey.tsv",
                 na=c("na", "NA","n/a","N/A"))
```

```
## Parsed with column specification:
## cols(
##   .default = col_character()
## )
## See spec(...) for full column specifications.
```

## Clean data

### Pre-surveys

```r
pre2018_clean <- pre2018 %>%
  # Select vars of interest
  select(RecordedDate, Progress, ResponseId, Q4:Q30_1_TEXT,
         #Q26_ options cover this information
         -Q26) %>%
  # Rename variables
  rename(Course = Q4, ID = Q7, InternalID = ResponseId,
         Pre_DS = Q10, Pre_DS_define = Q11, Pre_DS_where = Q12,
         Pre_BI = Q13, Pre_BI_define = Q14, Pre_BI_where = Q15,
         Pre_programs = Q17,
         Pre_Interest_MICB = Q14_1, Pre_Interest_BI = Q14_2,
         Pre_Interest_STAT = Q14_3, Pre_Interest_CPSC = Q14_4,
         Pre_Exp_MICB = Q15_1, Pre_Exp_BI = Q15_2,
         Pre_Exp_STAT = Q15_3, Pre_Exp_CPSC = Q15_4,
         Major = Q17_1,
         Prev_EDUCE_YN = Q18, Prev_EDUCE_course = Q19,
         Pre_MICB301 = Q21_1, Pre_MICB405 = Q21_2,
         Pre_MICB421 = Q21_3, Pre_MICB425 = Q21_4,
         Pre_MICB447 = Q21_5,
         Pre_STAT200 = Q23_1,
         Pre_STAT203 = Q23_2, Pre_STAT241 = Q23_3,
         Pre_STAT251 = Q23_4, Pre_BIOL300 = Q23_5,
         Pre_CPSC100 = Q24_1, Pre_CPSC101 = Q24_2,
         Pre_CPSC110 = Q24_3, Pre_CPSC121 = Q24_4,
         Pre_CPSC301 = Q24_5, Pre_other_complete = Q26_1_TEXT,
         Pre_other_current = Q26_2_TEXT, Pre_other_future = Q26_3_TEXT,
         Prev_cocurric_YN = Q28, Prev_cocurric = Q28_1_TEXT,
         Pre_Interest_cocurric_YN = Q29, Pre_Know_cocurric_YN = Q30,
         Pre_Know_cocurric = Q30_1_TEXT) %>%
  # Separate list variables into columns
  mutate(Prev_unix = ifelse(grepl("Unix", Pre_programs), 1, NA),
         Prev_R = ifelse(grepl("R", Pre_programs), 1, NA),
         Prev_mothur = ifelse(grepl("mothur", Pre_programs), 1, NA),
         Prev_QIIME = ifelse(grepl("QIIME", Pre_programs), 1, NA),
         Prev_metaG = ifelse(grepl("Metagenomics", Pre_programs), 1, NA)) %>%
  mutate(Prev_EDUCE_MICB301 = ifelse(grepl("MICB301", Prev_EDUCE_course), 1, NA),
         Prev_EDUCE_MICB405 = ifelse(grepl("MICB405", Prev_EDUCE_course), 1, NA),
         Prev_EDUCE_MICB421 = ifelse(grepl("MICB421", Prev_EDUCE_course), 1, NA),
         Prev_EDUCE_MICB425 = ifelse(grepl("MICB425", Prev_EDUCE_course), 1, NA)) %>%

  # Remove full question text and {} notation
  slice(-1,-2) %>%
  # Separate date and time
  separate(RecordedDate, into=c("Date", "time"), sep=" ") %>%
  mutate(Date = as.Date(Date)) %>%
  select(-time) %>%
  # Force all ids to be upper case
  mutate(ID = toupper(ID)) %>%
  # Remove incomplete surveys
  filter(Progress == 100) %>%
  # Remove test surveys
  filter(!(ID %in% c("WOW1234", "NAN1828"))) %>%
  # Remove duplicate surveys, keeping newest response
  arrange(desc(Date)) %>%
  distinct(Course, ID, .keep_all=TRUE) %>%
  # Format course
  separate(Course, into=c("Course", "course_name"), sep=": ") %>%
  # Create categorical year
  mutate(year = ifelse(Date >= "2017-09-01" & Date <= "2018-06-01" ,
                       "2017/18",
                ifelse(Date >= "2018-09-01" & Date <= "2019-06-01",
                       "2018/19", NA))) %>%
  # Remove unneeded variables
  select(-c(Progress, course_name, Prev_EDUCE_course, Pre_programs, Date))
```

Explore cleaned data

```
dim(pre2018_clean)
```

```
## [1] 297  52
```

```
colnames(pre2018_clean)
```

```
##  [1] "InternalID"               "Course"
##  [3] "ID"                       "Pre_DS"
##  [5] "Pre_DS_define"            "Pre_DS_where"
##  [7] "Pre_BI"                   "Pre_BI_define"
##  [9] "Pre_BI_where"             "Pre_Interest_MICB"
## [11] "Pre_Interest_BI"          "Pre_Interest_STAT"
## [13] "Pre_Interest_CPSC"        "Pre_Exp_MICB"
## [15] "Pre_Exp_BI"               "Pre_Exp_STAT"
## [17] "Pre_Exp_CPSC"             "Major"
## [19] "Prev_EDUCE_YN"            "Pre_MICB301"
## [21] "Pre_MICB405"              "Pre_MICB421"
## [23] "Pre_MICB425"              "Pre_MICB447"
## [25] "Pre_STAT200"              "Pre_STAT203"
## [27] "Pre_STAT241"              "Pre_STAT251"
## [29] "Pre_BIOL300"              "Pre_CPSC100"
## [31] "Pre_CPSC101"              "Pre_CPSC110"
## [33] "Pre_CPSC121"              "Pre_CPSC301"
## [35] "Pre_other_complete"       "Pre_other_current"
## [37] "Pre_other_future"         "Prev_cocurric_YN"
## [39] "Prev_cocurric"            "Pre_Interest_cocurric_YN"
## [41] "Pre_Know_cocurric_YN"     "Pre_Know_cocurric"
## [43] "Prev_unix"                "Prev_R"
## [45] "Prev_mothur"              "Prev_QIIME"
## [47] "Prev_metaG"               "Prev_EDUCE_MICB301"
## [49] "Prev_EDUCE_MICB405"       "Prev_EDUCE_MICB421"
## [51] "Prev_EDUCE_MICB425"       "year"
```

### Post-surveys

```
post2018_clean <- post2018 %>%
  # Select vars of interest
  select(RecordedDate, Progress, ResponseId, Q3:Q35) %>%
```

```r
# Rename variables
rename(Course = Q3, ID = Q4, InternalID = ResponseId,
       Post_DS = Q6, Post_DS_define = Q7, Post_DS_where = Q8,
       Post_BI = Q9, Post_BI_define = Q10, Post_BI_where = Q11,
       Post_Interest_MICB = Q14_1, Post_Interest_BI = Q14_2,
       Post_Interest_STAT = Q14_3, Post_Interest_CPSC = Q14_4,
       Post_Exp_MICB = Q15_1, Post_Exp_BI = Q15_2,
       Post_Exp_STAT = Q15_3, Post_Exp_CPSC = Q15_4,

       Post_MICB301 = Q21_1, Post_MICB405 = Q21_2,
       Post_MICB421 = Q21_3, Post_MICB425 = Q21_4,
       Post_MICB447 = Q21_5,
       Post_STAT200 = Q22_1,
       Post_STAT203 = Q22_2, Post_STAT241 = Q22_3,
       Post_STAT251 = Q22_4, Post_BIOL300 = Q22_5,
       Post_CPSC100 = Q23_1, Post_CPSC101 = Q23_2,
       Post_CPSC110 = Q23_3, Post_CPSC121 = Q23_4,
       Post_CPSC301 = Q23_5, Post_course_other = Q24,
       Post_cocurric_YN = Q26, Post_cocurric = Q26_1_TEXT,
       Post_Interest_cocurric_YN = Q27, Post_cocurric_1 = Q28,
       Post_cocurric_2 = Q30, Post_cocurric_3 = Q31,
       Minority = Q33, First_gen = Q34, Non_trad = Q35) %>%
# Separate list variables into columns
mutate(Post_Interest_introR = ifelse(!is.na(Post_cocurric_1), 1, NA),

       Post_Interest_data = ifelse(grepl("Working with", Post_cocurric_2), 1, NA),
       Post_Interest_ggplot = ifelse(grepl("Graphics with", Post_cocurric_2), 1, NA),
       Post_Interest_statmodel = ifelse(grepl("Statistical model", Post_cocurric_2), 1, NA),
       Post_Interest_repro = ifelse(grepl("Reproducible research", Post_cocurric_2), 1,
                             ifelse(grepl("R programming", Post_cocurric_3), 1, NA)),

       Post_Interest_mothur= ifelse(grepl("mothur", Post_cocurric_3), 1, NA),
       Post_Interest_QIIME = ifelse(grepl("QIIME", Post_cocurric_3), 1, NA),
       Post_Interest_microbiomeR = ifelse(grepl("vegan and phyloseq", Post_cocurric_3), 1, NA)) %>%
  # Remove full question text and {} notation
slice(-1,-2) %>%
# Separate date and time
separate(RecordedDate, into=c("Date", "time"), sep=" ") %>%
mutate(Date = as.Date(Date)) %>%
select(-time) %>%
# Force all ids to be upper case
mutate(ID = toupper(ID)) %>%
# Remove incomplete surveys
filter(Progress == 100) %>%
# Remove test surveys
filter(!(ID %in% c("WOW1234", "NAN1828"))) %>%
# Remove duplicate surveys, keeping newest response
arrange(desc(Date)) %>%
distinct(Course, ID, .keep_all=TRUE) %>%
# Format course
separate(Course, into=c("Course", "course_name"), sep=": ") %>%
# Create categorical year
mutate(year = ifelse(Date >= "2017-09-01" & Date <= "2018-06-01" ,
                      "2017/18",
               ifelse(Date >= "2018-09-01" & Date <= "2019-06-01",
                      "2018/19", NA))) %>%
# Remove unneeded variables
select(-c(Progress, course_name, Post_cocurric_1, Post_cocurric_2, Post_cocurric_3, Date))
```

Explore cleaned data

```
dim(post2018_clean)
```

```
## [1] 235  48
```

```
colnames(post2018_clean)
```

```
##  [1] "InternalID"              "Course"
##  [3] "ID"                      "Post_DS"
##  [5] "Post_DS_define"          "Post_DS_where"
##  [7] "Post_BI"                 "Post_BI_define"
##  [9] "Post_BI_where"           "Post_Interest_MICB"
## [11] "Post_Interest_BI"        "Post_Interest_STAT"
## [13] "Post_Interest_CPSC"      "Post_Exp_MICB"
## [15] "Post_Exp_BI"             "Post_Exp_STAT"
## [17] "Post_Exp_CPSC"           "Post_MICB301"
## [19] "Post_MICB405"            "Post_MICB421"
## [21] "Post_MICB425"            "Post_MICB447"
## [23] "Post_STAT200"            "Post_STAT203"
## [25] "Post_STAT241"            "Post_STAT251"
## [27] "Post_BIOL300"            "Post_CPSC100"
## [29] "Post_CPSC101"            "Post_CPSC110"
## [31] "Post_CPSC121"            "Post_CPSC301"
## [33] "Post_course_other"       "Post_cocurric_YN"
## [35] "Post_cocurric"           "Post_Interest_cocurric_YN"
## [37] "Minority"                "First_gen"
## [39] "Non_trad"                "Post_Interest_introR"
## [41] "Post_Interest_data"      "Post_Interest_ggplot"
## [43] "Post_Interest_statmodel" "Post_Interest_repro"
## [45] "Post_Interest_mothur"    "Post_Interest_QIIME"
## [47] "Post_Interest_microbiomeR" "year"
```

## Combine data

### Pre-surveys

```
pre_all <- pre2017_clean %>%
  # Remove/modify columns that do not exactly match in 2017 vs 2018-19
  ## Remove "none" selected for previous programm/software experience
  select(-Prev_none) %>%
  ## Convert InternalID to character to match 2018 data format
  mutate(InternalID = as.character(InternalID)) %>%
  # Combine with 2018-19 data
  bind_rows(pre2018_clean)

dim(pre_all)
```

```
## [1] 439  52
```

### Post-surveys

```
post2017_clean2 <- post2017_clean %>%
  # Remove/modify columns that do not exactly match in 2017 vs 2018-19
  ## Remove workshops no longer offered
  select(-Post_Interest_DC, -Post_Interest_SC,
         -Post_Interest_metaG, -Post_Interest_metaT) %>%
  ## Remove addtl comments section
  select(-Addtl_comments) %>%
  ## Convert InternalID to character to match 2018 data format
  mutate(InternalID = as.character(InternalID))

post_all <- post2018_clean %>%
  # Remove/modify columns that do not exactly match in 2017 vs 2018-19
  ## Combine mothur and QIIME workshop data
  mutate(Post_Interest_mothur_QIIME = ifelse(Post_Interest_mothur == 1, 1,
                                             Post_Interest_QIIME)) %>%
  select(-Post_Interest_mothur, -Post_Interest_QIIME) %>%
  # Combine with 2017 data
  bind_rows(post2017_clean2)

dim(post_all)
```

```
## [1] 400  47
```

### Matched pre-post responses

Combine

```
survey <- full_join(pre_all, post_all, by=c("Course","ID", "year")) %>%
  #rename internal ID variables
  rename(InternalID_pre = InternalID.x,
         InternalID_post = InternalID.y)
```

## Student consent

Load consent form data.

```
consent = read_csv("data_raw/20190325_consent_NoNames.csv") %>%
  # Force IDs to uppercase
  mutate(ID= toupper(ID), survey=toupper(survey), written=toupper(written)) %>%
  rename(Course=course)
```

```
## Parsed with column specification:
## cols(
##   course = col_character(),
##   year = col_character(),
##   ID = col_character(),
##   survey = col_character(),
##   written = col_character()
## )
```

Select responses with survey consent.

```
survey_consent <- consent %>%
  select(year, Course, ID, survey) %>%
  right_join(survey, by=c("Course","ID","year")) %>%
  #Keep only consenting students
  filter(survey == "A") %>%
  #Also remove BIOL436 which is no longer considered part of EDUCE
  filter(Course != "BIOL436") %>%
  #Sort for viewing
  arrange(year, Course, ID) %>%
  #Remove potential duplicates
  distinct() # Removes 2 responses

dim(survey_consent)
```

```
## [1] 398  97
```

Save to disk.

```
write_csv(survey_consent, "data_clean/2017.18.19_survey_clean.csv")
```

---

END