Final Stats 141XP Cleaning

Suraj Rajan (UID: 206066871)

2024-06-03

```
chem_bio_df = read.csv("chemistry_and_biochemistry_department.csv")

comm_df <- read.csv("communication_department.csv")

econ_df <- read.csv("economics_department.csv")

math_df <- read.csv("mathematics_department2.csv")

physics_df <- read.csv("physics_department.csv")

poli_sci_df <- read.csv("political_science_department.csv")

psych_df <- read.csv("psychology_department.csv")



stats_df <- read.csv("statistics_department2.csv")
```

```
head(stats_df, 30)

dim(stats_df)
```

```
head(comm_df,45)

dim(comm_df)
```

```
head(chem_bio_df, 500 )
```

```
head(econ_df, 50)
```

Before we combine our dataframes for cleaning, we must make sure the Easiness, Clarity, Workload, and Helpfulness Ratings have the same datatype.

Right now the columns in a few of these dfs are of type character, while others are type numeric, so lets convert them all to to numeric.

```
column_names = c("Easiness", "Clarity", "Workload", "Helpfulness")
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.2.3

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
Edit_data <- function(data) {
  clean_data <- data %>%
    mutate(Easiness = as.numeric(Easiness),
           Clarity = as.numeric(Clarity),
           Workload = as.numeric(Workload),
           Helpfulness = as.numeric(Helpfulness))

  return(clean_data)
  }
```

```
data_frames <- list(chem_bio_df, comm_df, econ_df, math_df, physics_df, poli_sci_df, psych_df, stats_df

clean_list <- lapply(data_frames, Edit_data)
```

```
## Warning: There were 4 warnings in `mutate()`.
## The first warning was:
## i In argument: `Easiness = as.numeric(Easiness)`.
## Caused by warning:
## ! NAs introduced by coercion
## i Run `dplyr::last_dplyr_warnings()` to see the 3 remaining warnings.
## There were 4 warnings in `mutate()`.
## The first warning was:
## i In argument: `Easiness = as.numeric(Easiness)`.
## Caused by warning:
## ! NAs introduced by coercion
## i Run `dplyr::last_dplyr_warnings()` to see the 3 remaining warnings.
## There were 4 warnings in `mutate()`.
## The first warning was:
## i In argument: `Easiness = as.numeric(Easiness)`.
## Caused by warning:
## ! NAs introduced by coercion
## i Run `dplyr::last_dplyr_warnings()` to see the 3 remaining warnings.
## There were 4 warnings in `mutate()`.
## The first warning was:
## i In argument: `Easiness = as.numeric(Easiness)`.
## Caused by warning:
## ! NAs introduced by coercion
## i Run `dplyr::last_dplyr_warnings()` to see the 3 remaining warnings.
```

```
library(dplyr)

combined_df <- bind_rows(clean_list)

head(combined_df)
```

```
##                  Professor Class_Code
## 1            John S. Adams  CHEM 196A
## 2 Anastassia N Alexandrova  CHEM 196A
## 3           Anne M Andrews  CHEM 196A
## 4        Soumitra Athavale  CHEM 196A
## 5               Agape Awad  CHEM 196A
## 6          Alfred D Bacher  CHEM 196A
##                                          Class_Name Overall_Rating Easiness
## 1 Research Apprenticeship in Chemistry and Biochemistry           N/A        NA
## 2 Research Apprenticeship in Chemistry and Biochemistry           N/A        NA
## 3 Research Apprenticeship in Chemistry and Biochemistry           N/A        NA
## 4 Research Apprenticeship in Chemistry and Biochemistry           N/A        NA
## 5 Research Apprenticeship in Chemistry and Biochemistry           N/A        NA
## 6 Research Apprenticeship in Chemistry and Biochemistry           N/A        NA
##   Clarity Workload Helpfulness  A.   A A..1  B.   B B..1  C.   C C..1  D.   D
## 1      NA       NA          NA N/A N/A  N/A N/A N/A  N/A N/A N/A  N/A N/A N/A
## 2      NA       NA          NA N/A N/A  N/A N/A N/A  N/A N/A N/A  N/A N/A N/A
## 3      NA       NA          NA N/A N/A  N/A N/A N/A  N/A N/A N/A  N/A N/A N/A
## 4      NA       NA          NA N/A N/A  N/A N/A N/A  N/A N/A N/A  N/A N/A N/A
## 5      NA       NA          NA N/A N/A  N/A N/A N/A  N/A N/A N/A  N/A N/A N/A
## 6      NA       NA          NA N/A N/A  N/A N/A N/A  N/A N/A N/A  N/A N/A N/A
##   D..1   F Grade.Quarter Review Review.Quarter Reviewer.Grade
## 1  N/A N/A            N/A    N/A            N/A            N/A
## 2  N/A N/A            N/A    N/A            N/A            N/A
## 3  N/A N/A            N/A    N/A            N/A            N/A
## 4  N/A N/A            N/A    N/A            N/A            N/A
## 5  N/A N/A            N/A    N/A            N/A            N/A
## 6  N/A N/A            N/A    N/A            N/A            N/A
```

```
dim(combined_df)
```

```
## [1] 157013     25
```

```
head(combined_df)

dim(combined_df)
```

```
combined_df$Overall_Rating <- as.numeric(combined_df$Overall_Rating)
```

```
## Warning: NAs introduced by coercion
```

```
tail(combined_df)
```

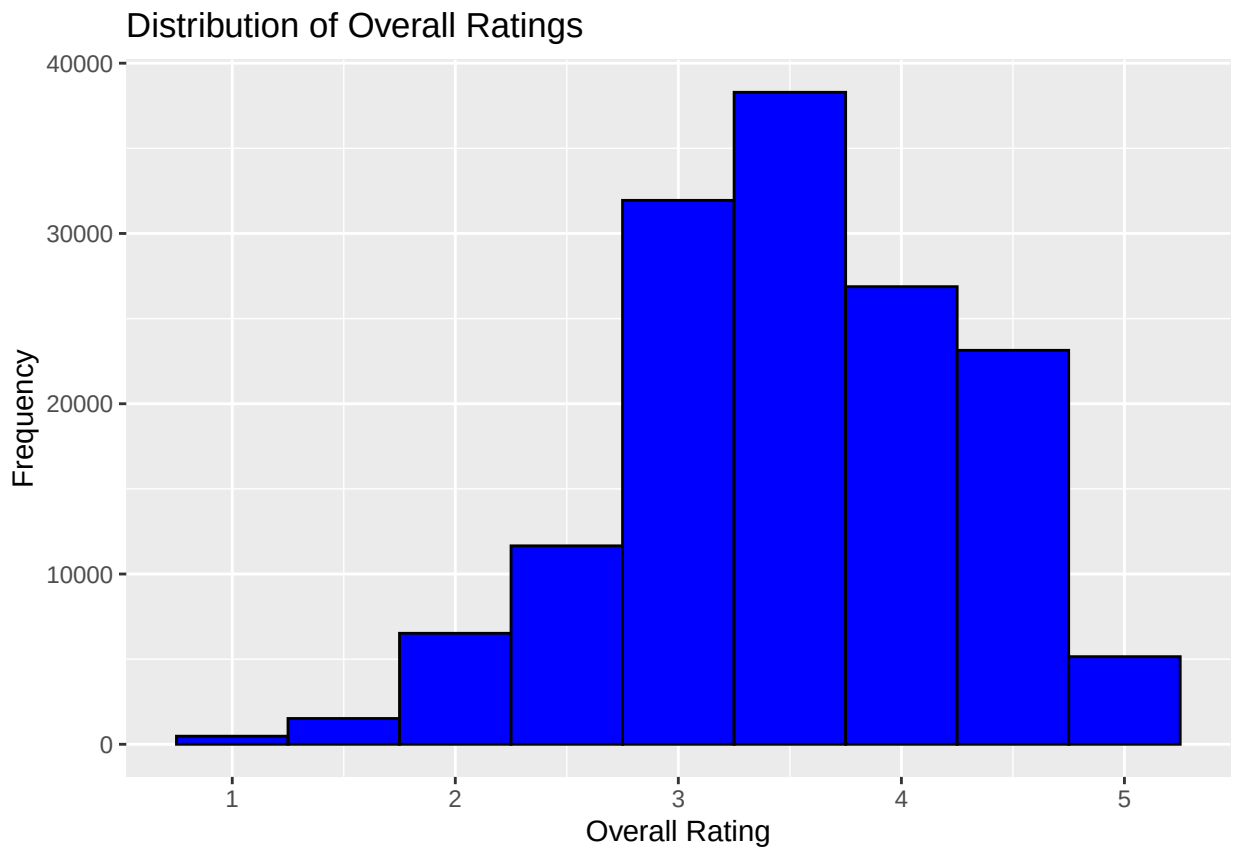
```r
grade_columns <- c("A.", "A", "A..1", "B.", "B", "B..1", "C.", "C", "C..1", "D.", "D", "D..1", "F")

combined_df[grade_columns] <- lapply(combined_df[grade_columns], function(x) as.numeric(sub("%", "", x)))
```

```
## Warning in FUN(X[[i]], ...): NAs introduced by coercion

## Warning in FUN(X[[i]], ...): NAs introduced by coercion

## Warning in FUN(X[[i]], ...): NAs introduced by coercion

## Warning in FUN(X[[i]], ...): NAs introduced by coercion

## Warning in FUN(X[[i]], ...): NAs introduced by coercion

## Warning in FUN(X[[i]], ...): NAs introduced by coercion

## Warning in FUN(X[[i]], ...): NAs introduced by coercion

## Warning in FUN(X[[i]], ...): NAs introduced by coercion

## Warning in FUN(X[[i]], ...): NAs introduced by coercion

## Warning in FUN(X[[i]], ...): NAs introduced by coercion

## Warning in FUN(X[[i]], ...): NAs introduced by coercion

## Warning in FUN(X[[i]], ...): NAs introduced by coercion
```

```r
head(combined_df)
```

```r
library(ggplot2)
ggplot(combined_df, aes(x = Overall_Rating)) +
  geom_histogram(binwidth = 0.5, fill = "blue", color = "black") +
  labs(title = "Distribution of Overall Ratings", x = "Overall Rating", y = "Frequency")
```

```
## Warning: Removed 11471 rows containing non-finite values (`stat_bin()`).
```




```
tail(combined_df)
```

```
combined_df[combined_df == "N/A"] <- NA
```

```
head(combined_df)
```

```
tail(combined_df)
```

```
classify_department <- function(code) {

  STEM <- c('^CHEM', '^MATH', '^PHYSICS', '^STATS')


  if (any(sapply(STEM, function(p) grepl(p, code, ignore.case = TRUE)))) {
    return('stem')
  } else {
    return('non-stem')
  }
}

combined_df <- combined_df %>%
  mutate(Dept.Type = sapply(Class_Code, classify_department))
```

```r
head(combined_df, 500)
```

```r
explore <- combined_df %>%
  filter(Dept.Type == "non-stem")

tail(explore, 50)
```

```r
final_df <- combined_df %>%
  select(Professor, Class_Code, Dept.Type, Class_Name, Overall_Rating, Easiness, Clarity, Workload, Hel

head(final_df)
tail(final_df)
```

```r
colnames(combined_df)
```

```
##  [1] "Professor"      "Class_Code"     "Class_Name"     "Overall_Rating"
##  [5] "Easiness"       "Clarity"        "Workload"       "Helpfulness"
##  [9] "A."             "A"              "A..1"           "B."
## [13] "B"              "B..1"           "C."             "C"
## [17] "C..1"           "D."             "D"              "D..1"
## [21] "F"              "Grade.Quarter"  "Review"         "Review.Quarter"
## [25] "Reviewer.Grade" "Dept.Type"
```

```r
colnames(final_df)
```

```
##  [1] "Professor"      "Class_Code"     "Dept.Type"      "Class_Name"
##  [5] "Overall_Rating" "Easiness"       "Clarity"        "Workload"
##  [9] "Helpfulness"    "A."             "A"              "A..1"
## [13] "B."             "B"              "B..1"           "C."
## [17] "C"              "C..1"           "D."             "D"
## [21] "D..1"           "F"              "Grade.Quarter"  "Review"
## [25] "Review.Quarter" "Reviewer.Grade"
```

```r
cols_to_check <- c("Overall_Rating", "Easiness", "Clarity", "Workload",
                   "Helpfulness", "A.", "A", "A..1", "B.", "B", "B..1",
                   "C.", "C", "C..1", "D.", "D", "D..1", "F", "Grade.Quarter", "Review",
                   "Review.Quarter", "Reviewer.Grade")


cleaned_df <- final_df %>%
  filter(rowSums(is.na(select(., all_of(cols_to_check)))) < length(cols_to_check))

head(cleaned_df)

dim(cleaned_df)
```

By eliminating observations where all of the above columns have NA values, we reduced the dataset's total observations by 6919 observations.

```r
table_stem <- table(cleaned_df$Dept.Type)

table_stem
```

```
##
## non-stem     stem
##    77461    72633
```

```r
cleaned_df <- cleaned_df %>%
  rename(A.Plus = A.,
         A.Minus = A..1,
         B.Plus = B.,
         B.Minus = B..1,
         C.Plus = C.,
         C.Minus= C..1,
         D.Plus = D.,
         D.Minus = D..1
         )
```

```r
head(cleaned_df)
```

```r
sentiment_df <- cleaned_df %>%
  filter(!is.na(Review)) %>%
  distinct(Review, .keep_all = TRUE) %>%
  select(-Grade.Quarter)
```

```r
table(sentiment_df$Dept.Type)
```

```
##
## non-stem     stem
##     9856    14655
```

```r
write.csv(sentiment_df, "sentiment_cleaned.csv", row.names = FALSE)
```

```r
grade_columns <- c("A.Plus", "A", "A.Minus",
                   "B.Plus", "B", "B.Minus",
                   "C.Plus", "C", "C.Minus",
                   "D.Plus", "D", "D.Minus", "F")


grade_cleaned <- cleaned_df %>%
  filter(rowSums(is.na(select(., all_of(grade_columns)))) < length(grade_columns))
```

```r
dim(grade_cleaned)
```

```
## [1] 146479     26
```

```r
write.csv(grade_cleaned, "grade_cleaned2.csv", row.names = FALSE)
```

```r
#head(grade_cleaned)
```

```r
explore_2 = grade_cleaned %>%
  filter(Dept.Type == 'non-stem')

#head(explore_2)
```

```r
cols_to_check <- c("Professor", "Class_Code", "Dept.Type", "Class_Name", "Overall_Rating",
                   "Easiness", "Clarity", "Workload", "Helpfulness", "A.Plus", "A", "A.Minus",
                   "B.Plus", "B", "B.Minus", "C.Plus", "C", "C.Minus", "D.Plus", "D", "D.Minus", "F")

grade_unique <- grade_cleaned %>%
  distinct(across(all_of(cols_to_check)), .keep_all = TRUE)

dim(grade_cleaned)
```

```
## [1] 146479     26
```

```r
dim(grade_unique)
```

```
## [1] 12306    26
```

```r
#head(grade_unique)

table(grade_unique$Dept.Type)
```

```
##
## non-stem     stem
##     6007     6299
```

```r
library(stringr)
depts <- c('CHEM', 'MATH', 'PHYSICS', 'STATS', 'ECON', 'COMM', 'PSYCH', 'POL SCI')
grade_unique <- grade_unique %>%
  mutate(Dept = str_extract(Class_Code, paste(depts, collapse = "|")))
```

```r
#head(grade_unique)

table(grade_unique$Dept)
```

```
##
##    CHEM    COMM    ECON    MATH PHYSICS POL SCI   PSYCH   STATS
##    1490     965    1356    2647    1373    1370    2316     789
```

```r
grade_unique <- grade_unique %>%
  select(Professor, Class_Code, Dept, everything())

head(grade_unique)
```

```
##                  Professor Class_Code Dept Dept.Type
## 1          Richard B Kaner   CHEM 189 CHEM      stem
## 2          Richard B Kaner   CHEM 189 CHEM      stem
## 3 Anastassia N Alexandrova CHEM C115A CHEM      stem
## 4          Richard B Kaner   CHEM 189 CHEM      stem
## 5          Richard B Kaner   CHEM 189 CHEM      stem
## 6          Richard B Kaner   CHEM 189 CHEM      stem
##                                                                                  Class_Name
## 1 Advanced Honors Seminars: Advanced Honors Seminar for Chemistry & Biochemistry 171, Lecture 1
## 2 Advanced Honors Seminars: Advanced Honors Seminar for Chemistry & Biochemistry 171, Lecture 1
## 3                                                                          Quantum Chemistry
## 4                                                                Advanced Honors Seminars
## 5                                                                Advanced Honors Seminars
## 6                                                                Advanced Honors Seminars
##   Overall_Rating Easiness Clarity Workload Helpfulness A.Plus    A A.Minus
## 1            5.0      5.0     5.0      5.0         5.0    0.0 89.7    10.3
## 2            5.0      5.0     5.0      5.0         5.0    0.0 74.0    22.0
## 3             NA       NA      NA       NA          NA   11.1 22.2    11.1
## 4            4.7      2.7     4.4      2.2         4.6    1.6 68.3    25.4
## 5            4.7      2.7     4.4      2.2         4.6    0.0 72.2    24.1
## 6            4.7      2.7     4.4      2.2         4.6    0.0 55.6    19.0
##   B.Plus    B B.Minus C.Plus    C C.Minus D.Plus D D.Minus   F Grade.Quarter
## 1    0.0  0.0       0      0  0.0       0      0 0       0 0.0     Fall 2020
## 2    4.0  0.0       0      0  0.0       0      0 0       0 0.0     Fall 2019
## 3   22.2 22.2       0      0 11.1       0      0 0       0 0.0     Fall 2015
## 4    4.8  0.0       0      0  0.0       0      0 0       0 0.0     Fall 2018
## 5    3.7  0.0       0      0  0.0       0      0 0       0 0.0     Fall 2017
## 6   17.5  6.3       0      0  0.0       0      0 0       0 1.6     Fall 2016
##
## 1 I really enjoyed this class. The focus is on "Materials World", so there are many engaging 50 minu
## 2 I really enjoyed this class. The focus is on "Materials World", so there are many engaging 50 minu
## 3
## 4                                                                 Very cute and essential class to tak
## 5                                                                 Very cute and essential class to tak
## 6                                                                 Very cute and essential class to tak
##   Review.Quarter Reviewer.Grade
## 1      Fall 2022              A
## 2      Fall 2022              A
## 3           <NA>           <NA>
## 4      Fall 2020              A
## 5      Fall 2020              A
## 6      Fall 2020              A
```

```
depts <- c('CHEM', 'MATH', 'PHYSICS', 'STATS', 'ECON', 'COMM', 'PSYCH', 'POL SCI')
sentiment_df <- sentiment_df %>%
  mutate(Dept = str_extract(Class_Code, paste(depts, collapse = "|"))) %>%
  select(Professor, Class_Code, Dept, everything())

head(sentiment_df)
```

```
##             Professor Class_Code Dept Dept.Type
## 1  Alexander Spokoyny  CHEM 196A CHEM      stem
## 2     Richard B Kaner   CHEM 189 CHEM      stem
## 3    Catherine Clarke   CHEM 147 CHEM      stem
```

```
## 4      Michael E Jung   CHEM 147 CHEM      stem
## 5  Alexander Spokoyny  CHEM 196B CHEM      stem
## 6 Benjamin J Schwartz CHEM C115B CHEM      stem
##                                                                                     Class_Name
## 1                                         Research Apprenticeship in Chemistry and Biochemistry
## 2 Advanced Honors Seminars: Advanced Honors Seminar for Chemistry & Biochemistry 171, Lecture 1
## 3                                                         Careers in Chemistry and Biochemistry
## 4                                                         Careers in Chemistry and Biochemistry
## 5                                         Research Apprenticeship in Chemistry and Biochemistry
## 6                                                                             Quantum Chemistry
##   Overall_Rating Easiness Clarity Workload Helpfulness A.Plus    A A.Minus
## 1              5        5       5        5           5     NA   NA      NA
## 2              5        5       5        5           5      0 89.7    10.3
## 3              5        5       5        5           5     NA   NA      NA
## 4              5        5       5        5           5     NA   NA      NA
## 5              5        5       5        5           5     NA   NA      NA
## 6              5        3       5        2           5     NA   NA      NA
##   B.Plus  B B.Minus C.Plus  C C.Minus D.Plus  D D.Minus  F
## 1     NA NA      NA     NA NA      NA     NA NA      NA NA
## 2      0  0       0      0  0       0      0  0       0  0
## 3     NA NA      NA     NA NA      NA     NA NA      NA NA
## 4     NA NA      NA     NA NA      NA     NA NA      NA NA
## 5     NA NA      NA     NA NA      NA     NA NA      NA NA
## 6     NA NA      NA     NA NA      NA     NA NA      NA NA
##
## 1
## 2                                                                               I really enjoyed thi
## 3
## 4 Very chill seminar with some interesting speakers. All you needed to do for the class was submit a
## 5
## 6
##   Review.Quarter Reviewer.Grade
## 1      Fall 2021              P
## 2      Fall 2022              A
## 3    Winter 2018              P
## 4    Winter 2023              P
## 5    Spring 2022              A
## 6    Winter 2019              A
```

```r
write.csv(grade_unique, "grade_cleaned3.csv", row.names = FALSE)
write.csv(sentiment_df, "sentiment_df_updated.csv", row.names = FALSE)
```