

Faculty Engagement EDA

William Morgan

February 26, 2018

1. Import and Check Structure

Goals:

- Import data
- Find important variables
- Check if cleaning needs to be done

Preliminary thoughts:

- `course_year`, `cem_unique`, `cour_st_dt`, `cour_end_dt` appear redundant or unnecessary
- rename `sln`, `term_session` to `class_nbr`, `session_code` for consistency
- insert '|' between subject and catalog number in `course` for consistency
- `acad_career` is a student-level variable and may not make sense to use at this granularity
- This includes graduate courses, which for now I'll just get rid of

Things to inspect:

- what is the difference between `acad_group` and `acad_org`?
 - `acad_group` is kind of like college but not exactly? full data has weird cross tabs
 - `acad_org` appears to be the department
- what is the distribution of enrollment totals?
- any courses from campuses outside the normal group? (Poly, West, Tempe, DT, ASUO)
- how complete is the feedback data?

Update before moving on:

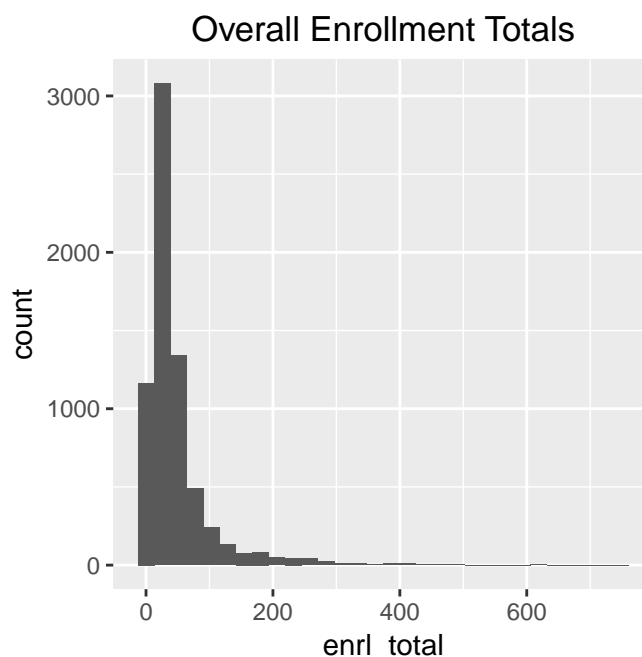
- Drops:
 - Graduate courses and missing courses
 - `course_year`, `cem_unique`, `cour_st_dt`, and `cour_end_dt` variables
 - Rename `sln` and `term_session`
 - Recreate `course` based on previous work

```
forum <- forum %>%
  filter(course != '-' & as.numeric(str_sub(course, 4, 6)) < 500) %>%
  select(-c(cour_st_dt, cour_end_dt, cem_unique)) %>%
  rename(session_code = term_session) %>%
  mutate(subject = str_sub(course, 1, 3),
         catalog_nbr = str_sub(course, 4, 6),
         course = paste(subject, catalog_nbr, sep = '|'),
         enr1_total = pass_ct + wdnw_ct + fail_ct)
```

2. Further inspection

What's the distribution of enrollment totals?

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Okay, so a handful of courses with super high enrollment are throwing off this graph, so let's figure out what these courses are and redo the graph without them to get a better understanding of the majority. We're not really interested in how many students are in these courses, just what the courses are

```
## # A tibble: 22 x 2
## # Groups:   course [22]
##   course  enrl_total
##   <chr>      <dbl>
## 1 ENG|102      748
## 2 SOC|352      553
## 3 ASB|300      484
## 4 BIO|112      476
## 5 CDE|312      459
## 6 SOC|101      456
## 7 ASB|100      427
## 8 SOC|424      412
## 9 MCO|194      403
## 10 FIN|380      402
## # ... with 12 more rows
```

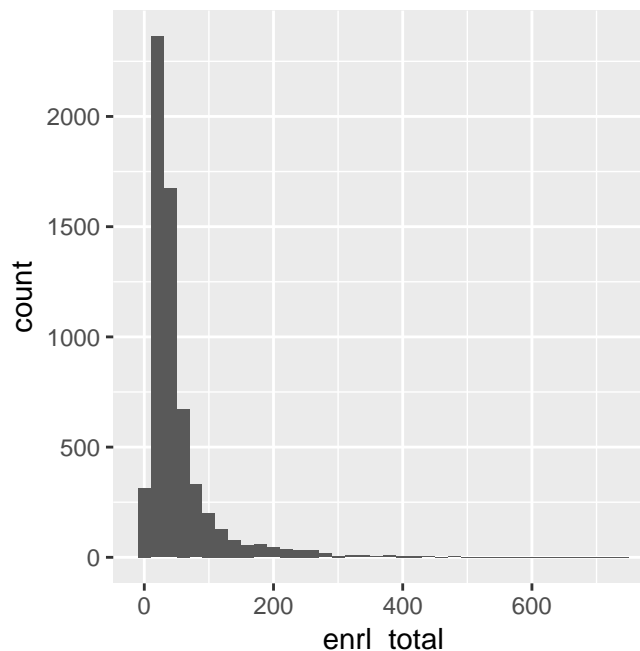
These are all intro freshman courses, so everything looks to be as expected there. Now let's check out low enrollment courses:

```
## # A tibble: 152 x 1
## # Groups:   course [152]
##   course
##   <chr>
## 1 ACC|320
```

```
## 2 ACC|310
## 3 ACC|241
## 4 AGB|250
## 5 AGB|294
## 6 ARS|250
## 7 ARS|480
## 8 ART|206
## 9 ASM|201
## 10 ASU|101
## # ... with 142 more rows
```

These courses are going to be a pain in the butt to filter through because a lot of them appear to be topic courses or internships (courses with catalog_nbr of 384, 394, or 494 are generally specially designed courses). It would probably be a good idea to merge `descr` and `descr2` from the peoplesoft tables to give us more info.

For now, let's just get rid of any courses with less than 5 students because **a.)** it isn't that many courses and **b.)** we know that some of those are irregularly taught courses that are probably irrelevant to the analysis



Before we move on, I'm going to assume that this super-low enrollment courses will actually turn out to be irrelevant for the study, so I will drop them entirely from the data set. To recap, the cumulative list of drops I've made is:

- Graduate courses and missing courses
- Courses with less than 5 students enrolled

```
forum <- forum %>% filter(enrl_total >= 5)
```

Campus/Location Variable

___ THIS SECTION IS NOW UNNECESSARY DUE TO CHANGES IN THE DATA ___

In the full data ASUO is designated as a unique location and observations having "ASUONLINE" `location` values should (for the most part) only have "ONLNE" as their `campus`. In other words, online students have their own campus that is different from Tempe, Poly, etc.

For some reason, this data seems to have a variety of values for campus despite the fact that there should only be one campus in the data. I don't know what the cause of this is and it could definitely be the case that I'm interpreting something wrong on the `full_data` end. To illustrate, here's a cross-tab of location and campus in the forum data:

I won't do anything with this for now because it really isn't that big of a deal, but whatever is causing this needs to be understood

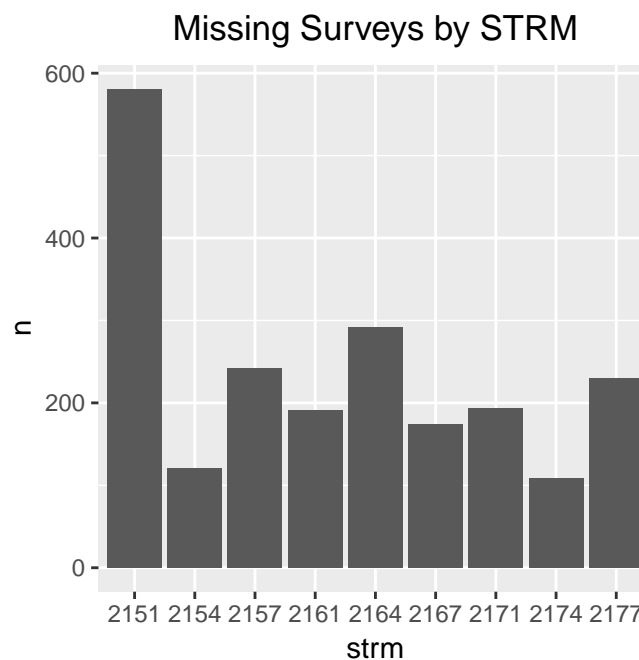
Feedback Data Completeness

When I initially glanced at the data, it seemed like there were a ton of missing values for the variables around student-course reviews. Let's check first if there were any missing values for `num_evals_expected`. We shouldn't have any, hopefully.

```
## [1] 2133
```

Okay, so about 2100 courses have missing values for `num_evals_expected`. It is not included here but these courses also have missing `responserate`, so I will just assume that these guys are also missing the average response for each of the questions.

Let's investigate a little more into these courses - is it limited to a specific time frame? course? department? (for the moment we assume that `acad_org` is the department; still working on confirming this)

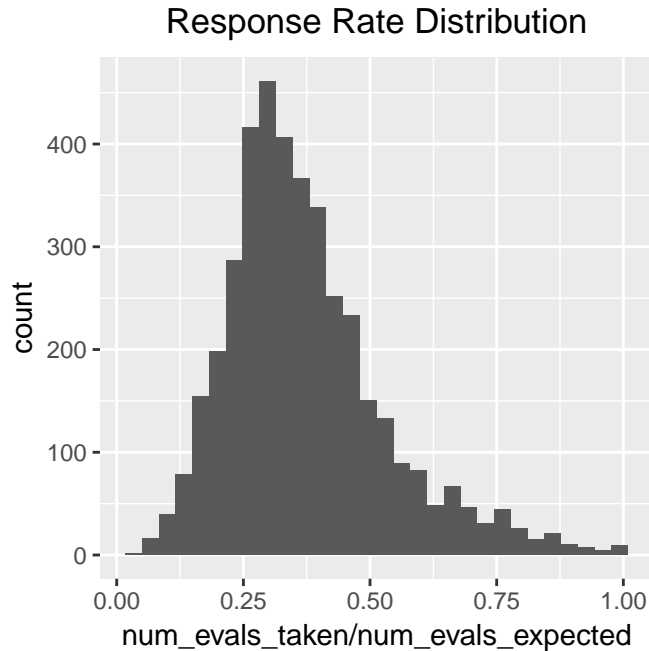


It is probably reasonable to assume that the missing values from earlier terms are just flat out missing from the warehouse, but I'm not sure what is going on with the more recent terms. For now, we'll just drop the course-sections with missing `expectedsurveys` and move on to checking out `responserate`. Just as a quick recap, the current list of drops is:

- graduate courses or missing courses (~2100)
- courses with less than 5 students (~700)
- courses with missing surveys (~2100)

Onto the graphing:

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



This looks pretty good, no reason for concern. Since most of this looks fine, let's just move on to some feature engineering before we get to modeling

3. Faculty Posts

We want to investigate the distribution of the faculty postings. In particular, we look at `num_fac_ta_posts`, `total_len_fac_ta_posts`, and `num_fac_ta`

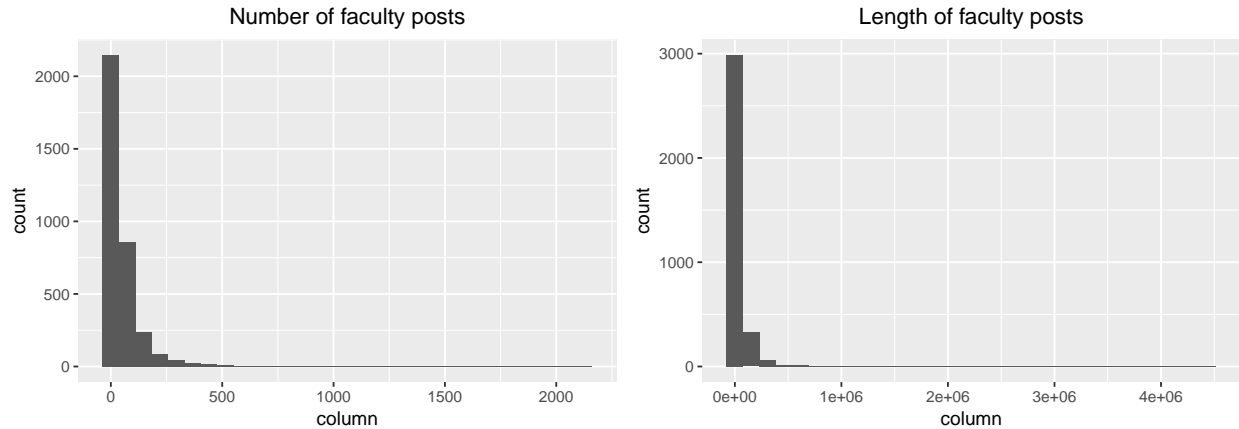
```
## # A tibble: 8 x 2
##   num_fac_ta     n
##   <dbl> <int>
## 1      0     590
## 2     1.00  2859
## 3     2.00   377
## 4     3.00    99
## 5     4.00    33
## 6     5.00    19
## 7     6.00    13
## 8     7.00    12
```

So it turns out that there are a lot of courses that don't have any faculty. We need to drop those courses for now until we have an answer as to why this is the case

```
forum <- filter(forum, num_fac_ta > 0)
```

Now that that is out of the way, let's look at the distributions for the other two variables of interest:

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



There are some huge outliers that are throwing off the graphs. To condense this slightly, let's drop anything above the 95th quantile for the faculty post length. This is done assuming that the reason for these crazy high totals is because an instructor might post fat blocks of text describing an assignment. These observations are less valuable because they are less of a representation of faculty engagement and more of how the instructor uses the blackboard shell.

We leave in observations with really high faculty post counts because that is likely indicative of instructors who reply to a lot and post often, which is exactly what we want to identify.

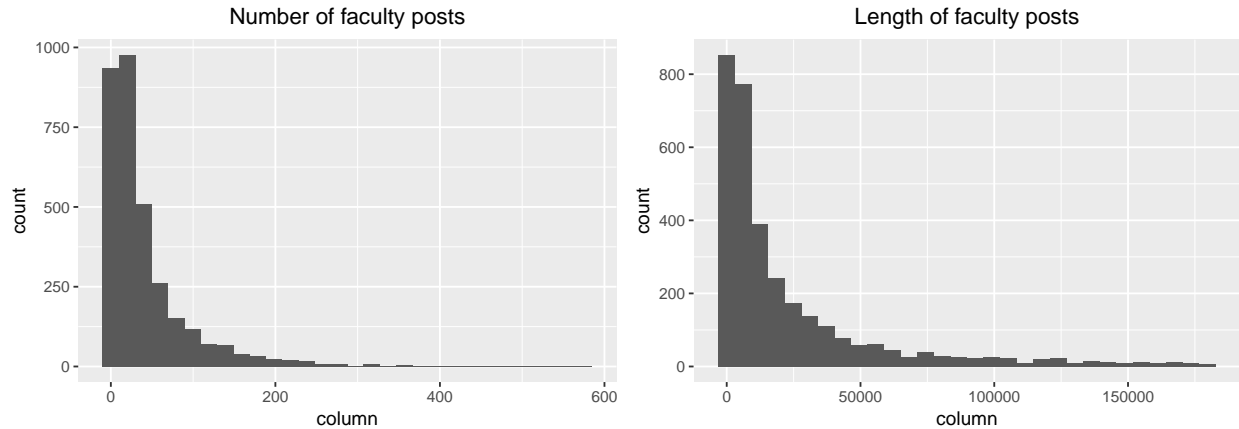
```
forum <- forum %>%
  filter(total_len_fac_ta_posts <= quantile(total_len_fac_ta_posts, .95))

plotHist <- function(data, column, title){
  data %>%
    ggplot(aes(column)) +
    geom_histogram() +
    ggtitle(title) +
    theme(plot.title = element_text(hjust = .5))
}

num <- plotHist(forum, forum$num_fac_ta_posts, "Number of faculty posts")
len <- plotHist(forum, forum$total_len_fac_ta_posts, "Length of faculty posts")

grid.arrange(num, len, nrow = 1)

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
rm(num, len)
```

Distributions are still ultra skewed - let's get a numeric summary real quick and keep it for later

```
summary(forum$num_fac_ta_posts)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      1.00   8.00   22.00  41.93  51.00  576.00
```

```
summary(forum$total_len_fac_ta_posts)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##         0   2912   9406  23665  28937 179661
```

4. Missing Values

This will be a relatively short section; we just want to check out which variables have missing values and guess why that might be. More than likely we will have to go back to Mike for an explanation. Any variables with missing values will be dropped

```
sapply(forum, function(x) sum(is.na(x)))
```

```
##      myasu_course      course      strm
##           0           0           0
##      sln_ct      non_ol_ct      sync_st_dt
##           0           0           0
##      sync_end_dt      session_code      acad_group
##           0           0           0
##      acad_org      campus      sum_enrl_total
##           0           0           0
##      ocourse_ct      icourse_ct      other_course_ct
##           0           0           0
##      pass_ct      fail_ct      wdown_ct
##           0           0           0
##      drop_ct      sum_gpa      num_fac_ta
##           0           0           0
##      has_hallway      fac_posted_wk0      fac_posted_wk1
##           0           0           0
##      fac_posted_wk2      fac_posted_wk3      fac_posted_wk4
##           0           0           0
```

```
##      fac_posted_wk5      fac_posted_wk6      fac_posted_wk7
##      0                0                0
##      fac_posted_wk8      fac_posted_wk9      fac_posted_wk10
##      0                0                0
##      fac_posted_wk11     fac_posted_wk12     fac_posted_wk13
##      0                0                0
##      fac_posted_wk14     fac_posted_wk15     num_stu_posts
##      0                0                0
##      num_stu_with_posts   num_fac_ta_posts   total_len_fac_ta_posts
##      0                0                0
##      num_evals_expected   num_evals_taken    q_success
##      0                0                0
##      q_prepared          q_presentations     q_navigate
##      0                0                1
##      q_feedback          q_responded         q_present
##      0                1                0
##      subject            catalog_nbr         enrl_total
##      0                0                0
```

```
forum <- forum %>%
  filter_all(all_vars(!is.na(.)))
```

Summary of drops so far:

- graduate courses or missing courses (~2100)
- courses with less than 5 students (~700)
- courses with missing surveys (~2100)
- 0 faculty/ta (~600)
- faculty length outliers (~200)
- missing values (~5?)

5. Feature Engineering

This section is probably just going to be condensing a couple of variables (like “faculty posted in week X”) and maybe coming up with some extra dependent variables

Possible ideas:

- change num_stu_with_posts to proportion
- condense survey questions into faculty/design split
- indicator for having posted in first or last week?

What is the gpa of students who withdraw??? It should be 0

New find: the question responses are not numeric and have to be changed

I’m actually just going to drop all the C session courses cause they represent like none of the sample

```
# Variables I don't want to keep for modeling
drop_vars <- c("myasu_course", "sync_st_dt", "sync_end_dt",
              "acad_group", "acad_career", "acad_org", "campus",
              "location", "subject", "catalog_nbr")
```



```

# Grab question names
questions <- forum %>%
  select(starts_with('q_')) %>%
  names()

# Force questions to numeric
forum <- forum %>%
  mutate_at(questions, as.numeric)

# Add new variables
forum_cleaned <- forum %>%
  mutate(prop_stu_posted = num_stu_with_posts / enr1_total,
         avg_gpa = sum_gpa / enr1_total,
         prop_withdrw = wdwn_ct / enr1_total,
         instr_score = (q_responded + q_present + q_feedback) / 3,
         design_score = (q_success + q_prepared + q_presentations + q_navigate) / 4,
         posts_per_student = num_stu_posts / enr1_total,
         posts_per_fac = num_fac_ta_posts / num_fac_ta,
         pass_rate = pass_ct / enr1_total,
         avg_fac_post_len = total_len_fac_ta_posts / num_fac_ta_posts,
         upper_division = if_else(
           as.numeric(str_sub(course, 5, 7)) >= 300, 1, 0),
         fac_posted_boc = if_else(
           (fac_posted_wk1 == 1 | fac_posted_wk2 == 1), 1, 0),
         fac_posted_eoc = if_else(
           (fac_posted_wk7 == 1 | fac_posted_wk8 == 1), 1, 0),
         pass_rate = pass_ct / enr1_total,
         response_rate = num_evals_taken / num_evals_expected) %>%
  rename(course_id = myasu_course) %>%
  select(course_id, strm, session_code, enr1_total, response_rate, pass_ct, fail_ct,
         wdwn_ct, has_hallway, fac_posted_boc, fac_posted_eoc,
         total_len_fac_ta_posts, prop_stu_posted, avg_gpa,
         prop_withdrw, instr_score, design_score, posts_per_fac,
         posts_per_student, avg_fac_post_len, upper_division) %>%
  filter(session_code %in% c("A", "B"))

# Do a final check on missing values
supply(forum_cleaned, function(x) sum(is.na(x)))

```

```

##          course_id          strm          session_code
##              0              0              0
##      enr1_total      response_rate          pass_ct
##              0              0              0
##          fail_ct          wdwn_ct          has_hallway
##              0              0              0
##      fac_posted_boc      fac_posted_eoc total_len_fac_ta_posts
##              0              0              0
##      prop_stu_posted          avg_gpa          prop_withdrw
##              0              0              0
##          instr_score          design_score          posts_per_fac
##              0              0              0
##      posts_per_student      avg_fac_post_len          upper_division
##              0              0              0

```

```
write_csv(forum_cleaned, "Data/forum_cleaned.csv")
```

Final summary of drops (we started with $n = 8916$)

- graduate courses or missing courses (~2100)
- courses with less than 5 students (~700)
- courses with missing surveys (~2100)
- 0 faculty/ta (~600)
- faculty length outliers (~200)
- missing values (~5?)
- session C courses (~25)

Final tally: $N = 3237$