Engagement Data Quality Check

William Morgan

13 July, 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?

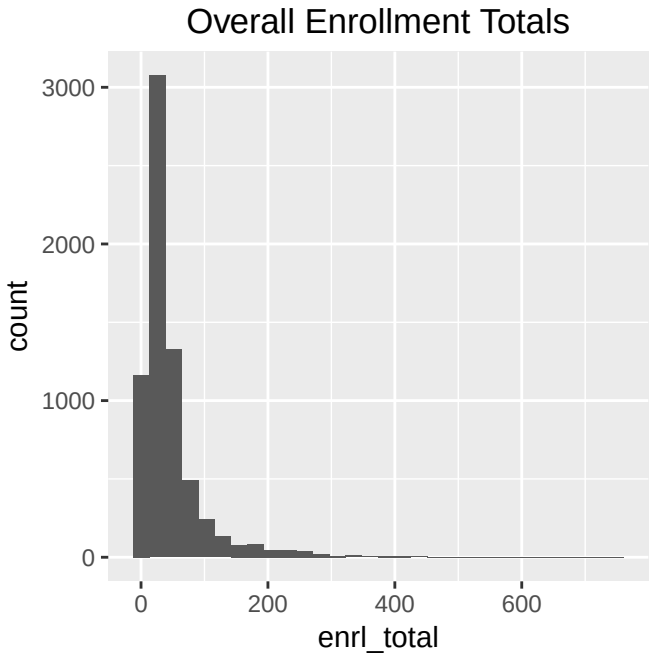
Update before moving on:

- 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

---

2. Further inspection

What's the distribution of enrollment totals?



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
```

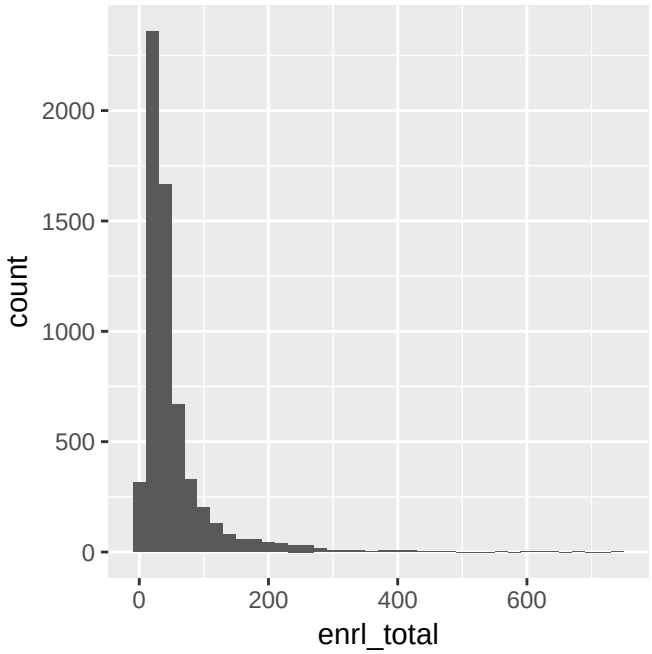
These are all intro freshman courses, so everything looks to be as expected there. Now let's check out low enrollment courses:

```
## # 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 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

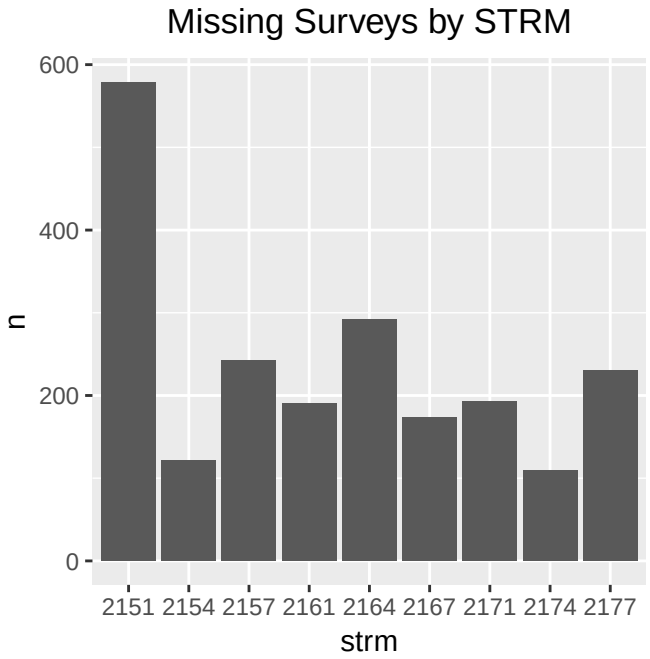
Feedback Data Completeness

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] 2131
```

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.

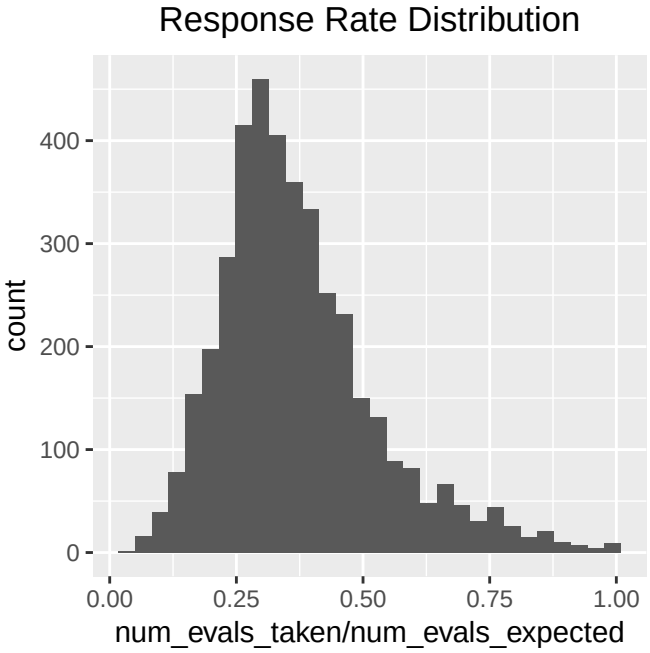
Let's investigate a little more into these courses - is it limited to a specific time frame? course? department? (for the moment we assume that `acad_org` is the department; still working on confirming this)



It is probably reasonable to assume that the missing values from earliers 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:



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

---

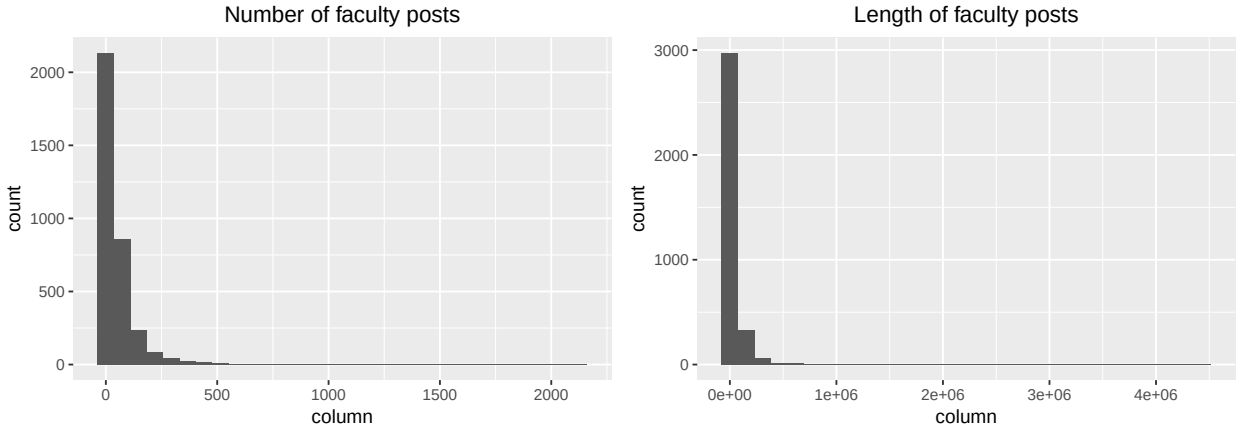
3. Faculty Posts

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  2844
## 3       2.00   377
## 4       3.00    99
## 5       4.00    33
## 6       5.00    19
## 7       6.00    13
## 8       7.00    12
```

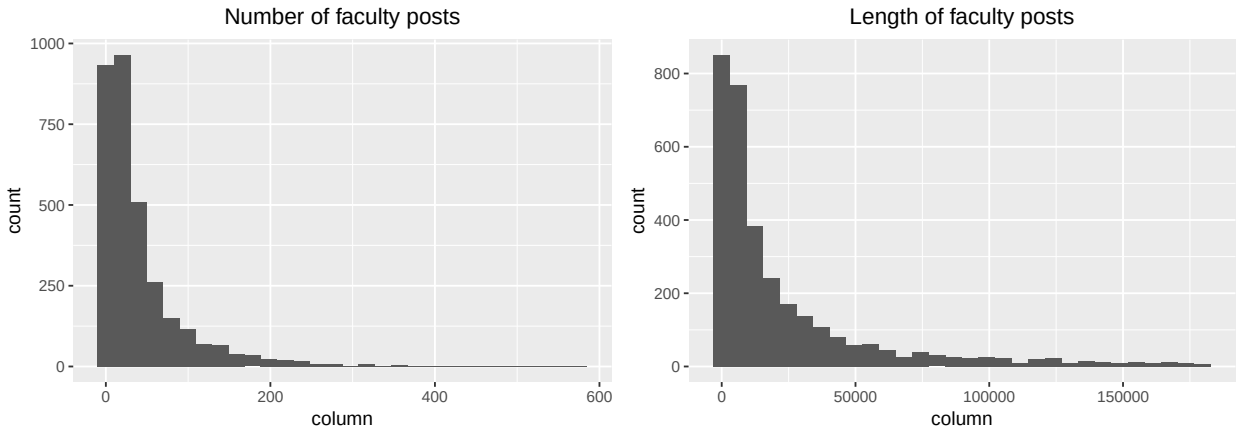
So it turns out that there are a lot of courses that don't have any faculty. We need to drop those courses for now until we have an answer as to why this is the case

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



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.



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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    8.00   22.00   42.06   51.00  576.00
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0    2892    9399   23720   29017  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

```
##  q_navigate q_responded
##          80          82
```

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

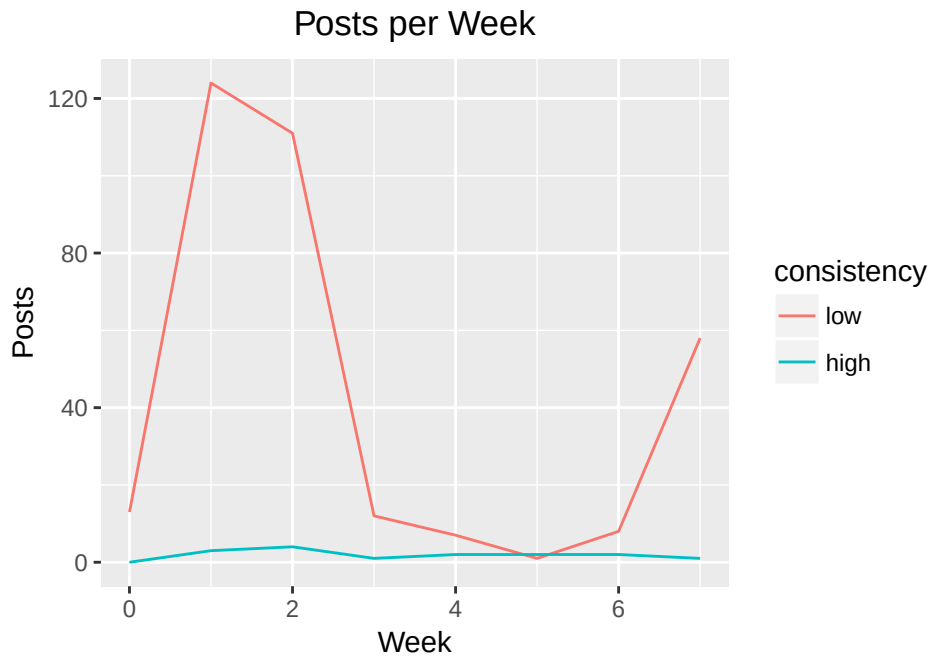
Possible ideas:

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

Something that came out of our discussion was the idea of faculty and student post consistency from week to week. This is a relatively easy measure to define, as we know how many posts were made in each week. For these two new variables, we take for each observation the sequence of numbers defining how many posts in each week and then take the std. dev. of that set. To make for a slightly easier interpretation, we take the reciprocal of that value so that higher values imply more consistency.

(we might get rid of the reciprocal transformation if its too much of a hassle to discuss)

To make it super clear, we plot the number of posts per week for two observations - one instructor with very high consistency and one instructor with very low consistency.




```
## Saving 5 x 3.5 in image
```

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)
- student data on these is still good in case we need/want it at some point
- faculty length outliers (~200)
- missing values (~5?)
- session C courses (~25)

Final tally: N = 3237