Units, based on terms and courses

Dominik Glandorf

2022-08-19

```
students = get_student_sub()
bg = get_student_background_data(ids = students$mellon_id)
# get terms
terms = get_term_data(ids = students$mellon_id)
terms = merge(terms, students[,c('mellon_id','admitdate','dropout')], by="mellon_id")
# enumerate terms order by their date
terms_ordered = terms[order(terms$term_code),c('mellon_id','term_code')]
terms_ordered$term_num = ave(terms_ordered$mellon_id, terms_ordered$mellon_id, FUN=seq_along)
terms = merge(terms, terms_ordered, by=c("mellon_id","term_code"))
# get courses
courses = get_course_data(ids = students$mellon_id)
```

Availability of units information

```
units_summary_terms = is.na(terms[,names(terms)[86:121]])
colSums(units_summary_terms)[order(colSums(units_summary_terms))]
```

```
## current_units_completed_transfer    current_units_completed_total
##                                0                                0
## cumulative_units_completed_trans cumulative_units_completed_total
##                                0                                0
##   cumulative_units_attempted_pnp   cumulative_units_completed_pnp
##                           366765                           366765
##    current_units_attempted_grade      current_units_attempted_pnp
##                           474356                           474356
## current_units_attempted_upperdiv current_units_attempted_graduate
##                           474356                           474356
##   current_units_attempted_online current_units_attempted_oncampus
##                           474356                           474356
##   current_units_completed_graded      current_units_completed_pnp
##                           474356                           474356
## current_units_completed_lowerdiv current_units_completed_upperdiv
##                           474356                           474356
##   current_units_completed_online current_units_completed_oncampus
##                           474356                           474356
##    current_units_attempted_total cumulative_units_attempted_grade
##                           474384                           474657
## cumulative_units_completed_grade cumulative_units_completed_onlin
```

```
##                            474657                            474657
## cumulative_units_completed_oncam cumulative_units_attempted_total
##                            474657                            474685
## current_units_attempted_lowerdiv current_units_attempted_transfer
##                            489477                            516372
## current_units_completed_graduate cumulative_units_attempted_lower
##                            516372                            516372
## cumulative_units_attempted_upper cumulative_units_attempted_gradu
##                            516372                            516372
## cumulative_units_attempted_onlin cumulative_units_attempted_oncam
##                            516372                            516372
## cumulative_units_attempted_trans cumulative_units_completed_lower
##                            516372                            516372
## cumulative_units_completed_upper cumulative_units_completed_gradu
##                            516372                            516372
```

```
units_summary_courses = is.na(courses[,names(courses)[44:47]])
colSums(units_summary_courses)[order(colSums(units_summary_courses))]
```

```
##         units_attempted         units_completed include_units_attempted
##                  214891                 1484594                 3251125
## include_units_completed
##                 3251125
```

Units completed total

In term data

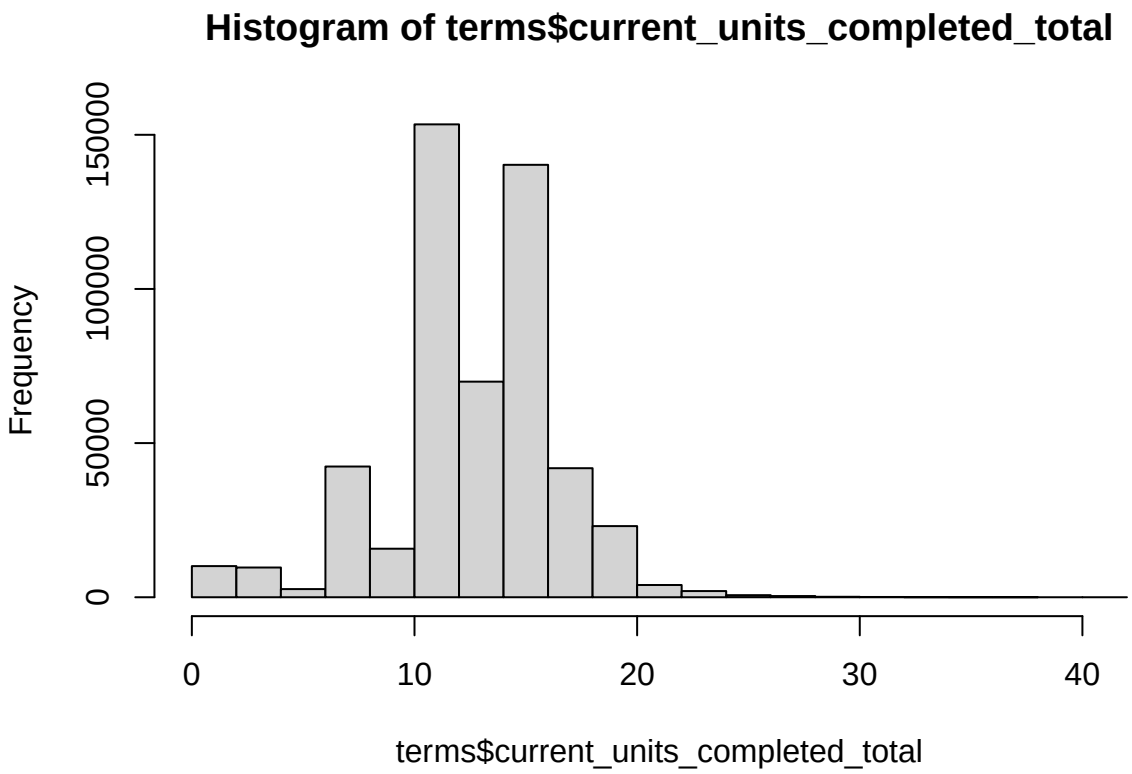
About how much percent of terms per user do we have units completed?

```
mean(!is.na(terms$current_units_completed_total))
```

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

This information exists for every term of all students in term data.

```
hist(terms$current_units_completed_total)
```

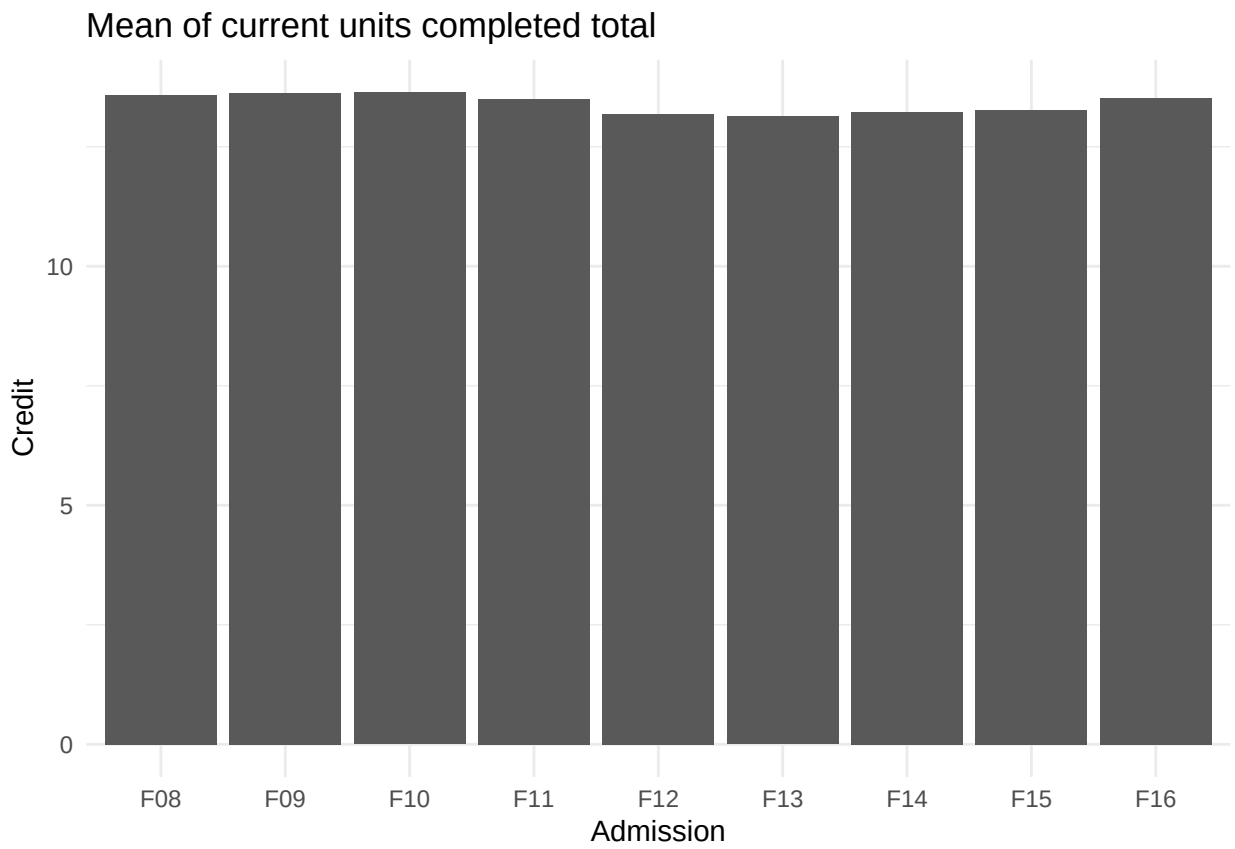


Hypothesis: users that have one term with very low units completed have a higher dropout chance

Some plots about current units completed:

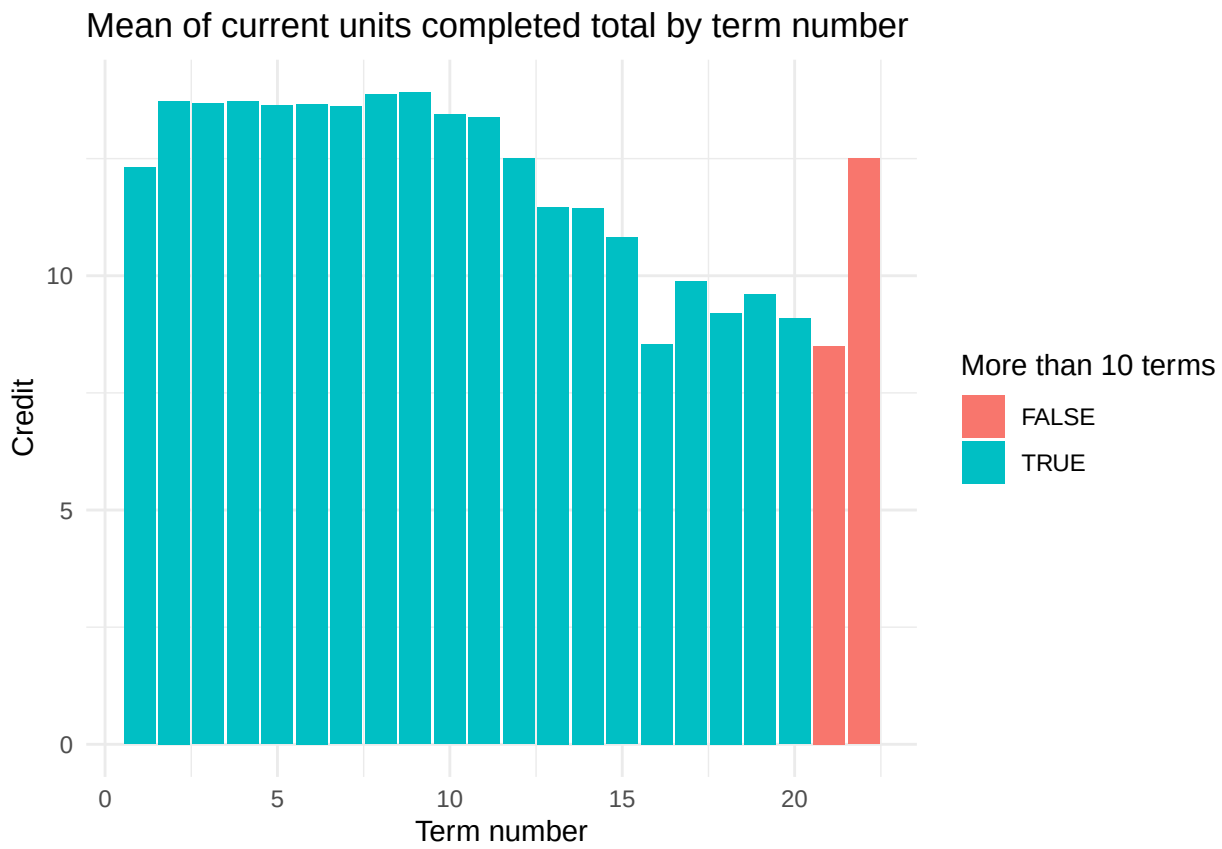
```
avg_credits_admission = aggregate(terms$current_units_completed_total, by=list("admitdate"=terms$admitda
avg_credits_major = do.call(data.frame, aggregate(current_units_completed_total ~ major_name_1, terms, 
avg_credits_term = do.call(data.frame, aggregate(current_units_completed_total ~ term_num, terms, FUN =
avg_credits_major_term = do.call(data.frame, aggregate(current_units_completed_total ~ major_name_1 + te
```

```
ggplot(avg_credits_admission, aes(x=admitdate, y=x)) +
  geom_bar(stat="identity") +
  theme_minimal() +
  labs(title = "Mean of current units completed total",
       x = "Admission",
       y = "Credit")
```



Current units does not seem to considerable vary between cohorts.

```
ggplot(avg_credits_term, aes(x=term_num, y=current_units_completed_total.mean, fill=current_units_compl
  geom_bar(stat="identity") +
  theme_minimal() +
  labs(title = "Mean of current units completed total by term number",
       x = "Term number",
       y = "Credit",
       fill = "More than 10 terms")
```

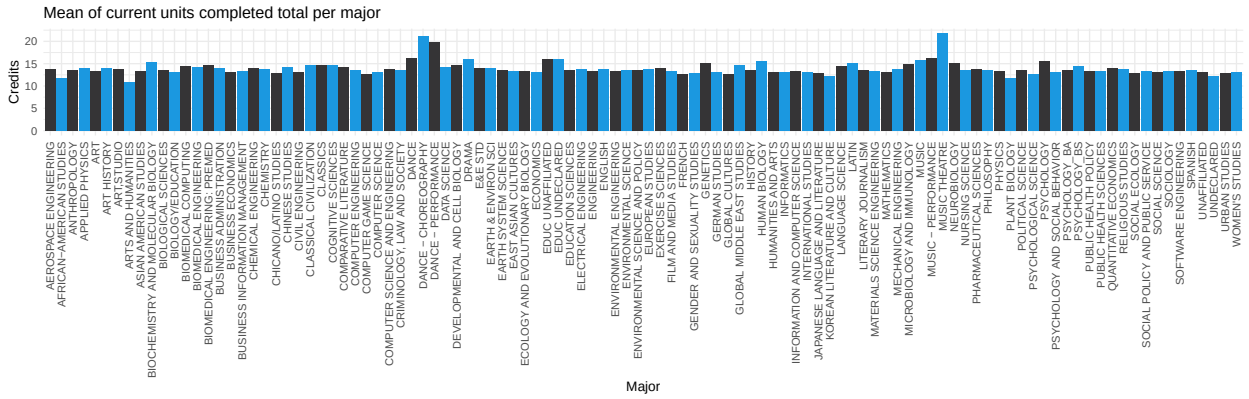


Units in the first and 12th term are lower. If people study longer, units keep decreasing.

```
fill=c(rep(c("1","2"), 53))
print(avg_credits_major$major_name_1[avg_credits_major$current_units_completed_total.n<10])
```

```
## [1] "EDUC UNAFFILIATED"   "EDUC UNDECLARED"     "HUMANITIES AND ARTS"
## [4] "PLANT BIOLOGY"
```

```
ggplot(avg_credits_major, aes(x=major_name_1, y=current_units_completed_total.mean, fill=fill)) +
  geom_bar(stat="identity") +
  theme_minimal() +
  labs(title = "Mean of current units completed total per major",
       x = "Major",
       y = "Credits", fill="")+
  scale_x_discrete(guide = guide_axis(angle = 90)) +
  scale_fill_manual(values = c("1"="#353436", "2"="#1b98e0"))+
  theme(legend.position = "none")
```




```
ggplot(avg_credits_major_term, aes(x=term_num, y=current_units_completed_total.mean, fill=(current_units
  geom_bar(stat="identity") +
  theme_minimal() +
  labs(title = "Mean of current units completed total per term",
       x = "Term",
       y = "Credits", fill="")+
  scale_x_discrete(guide = guide_axis(angle = 90)) +
  theme(legend.position = "none") +
  facet_wrap(vars(major_name_1), ncol = 8)
```


Mean of current units completed total per term
ANTHROPOLOGY
APPLIED PHYSICS
ART
ART HISTORY
ART,STUDIO
ANTHROPOLOGY
APPLIED PHYSICS
ART
ART HISTORY
ART,STUDIO
BIOLOGICAL SCIENCES
BIOLOGY/EDUCATION
CHEMISTRY
CHINESE STUDIES
CIVIL ENGINEERING
CLASSICS
COGNITIVE SCIENCES
COMPUTER SCIENCE
DANCE
DATA SCIENCE
DRAMA
E&E STD
EARTH & ENVIRON SCI
ECONOMICS
EDUC UNAFFILIATED
EDUC UNDECLARED
EDUCATION SCIENCES
ENGINEERING
ENGLISH
EUROPEAN STUDIES
EXERCISE SCIENCES
FRENCH
GENETICS
GERMAN STUDIES
GLOBAL CULTURES
HISTORY
HUMAN BIOLOGY
INFORMATICS
LANGUAGE SCIENCE
LATIN
MATHEMATICS
MUSIC
MUSIC THEATRE
NEUROBIOLOGY
NURSING SCIENCE
PHILOSOPHY
PHYSICS
PLANT BIOLOGY
POLITICAL SCIENCE
PSYCHOLOGY
PSYCHOLOGY_BA
PSYCHOLOGY_BS
RELIGIOUS STUDIES
SOCIAL ECOLOGY
SOCIAL SCIENCE
SOCIOLOGY
SPANISH
UNAFFILIATED
UNDECLARED
URBAN STUDIES
WOMEN'S STUDIES
0
10
20
Credits
Term

In courses data

```
names(courses)
```

```
##  [1] "mellon_id"                "group_a"
##  [3] "mellon_enr_dt_a"          "group_b"
##  [5] "mellon_enr_dt_b"          "term_code"
##  [7] "term_desc"                "course_code"
##  [9] "course_code_concat"       "course_title"
## [11] "course_dept_code_and_num" "course_section_num"
## [13] "course_section_title"     "school_code"
## [15] "school_name_abbrev"       "dept_code"
## [17] "dept_name_abbrev"         "course_level"
## [19] "course_type"              "meeting_schedule"
## [21] "meeting_location"         "enroll_restrictions"
## [23] "enroll_restrictions_desc" "class_weeks"
## [25] "instructor_id"            "instructor_title"
## [27] "honors_course"            "transferring_inst"
## [29] "online_course"            "min_units"
## [31] "max_units"                "max_seats"
## [33] "students_waitlisted"      "seats_requested"
## [35] "in_progress"              "enrollment_status"
## [37] "course_syllabus"          "section_start_at"
## [39] "section_end_at"           "credit_code"
## [41] "repeat_code"              "workload_credit_only"
## [43] "include_gpa"              "include_units_attempted"
## [45] "include_units_completed"  "units_attempted"
## [47] "units_completed"          "midterm_grade"
## [49] "final_score_canvas"       "final_grade"
## [51] "final_grade_date"         "final_grade_official"
## [53] "year"                     "acadyr"
## [55] "mellon_yr"
```

Let's see how much missing data we have here.

```
mean(is.na(courses$units_completed))
```

```
## [1] 0.4566401
```

This can be related to listing a lot of courses twice:

```
courses[courses$mellon_id==162784,c('course_code','course_dept_code_and_num','meeting_schedule','units_
```

```
## # A tibble: 77 x 4
##    course_code course_dept_code_and_num meeting_schedule         units_completed
##          <dbl> <chr>                    <chr>                              <dbl>
##  1       62000 ECON 1                   Tu  Th 02:00 PM - 03:20~              4
##  2       62003 ECON 1                   We 04:00 PM - 04:50 PM               NA
##  3        4000 MUSIC 8                  Tu 06:30 PM - 09:20 PM               NA
##  4       63010 SOCSCI 5D                Mo  We  Fr 09:00 AM - 0~              4
##  5       62040 ECON 20A                 <NA>                                 4
```

```
##  6       62043 ECON 20A                Mo 05:00 PM - 05:50 PM                NA
##  7       44240 MATH 2A                 Mo  We  Fr 04:00 PM - 0~               4
##  8       44242 MATH 2A                 Tu  Th 02:00 PM - 02:50~              NA
##  9       30300 PHILOS 1                Tu  Th 09:30 AM - 10:50~               4
## 10       30306 PHILOS 1                Tu 00:00 PM - 00:50 PM                NA
## # ... with 67 more rows
```

Why are some courses listed twice? It looks like there are different time slots for the same course.

Let's count units from courses.

```
units_from_courses=aggregate(cbind(units_attempted, units_completed)~term_code+mellon_id, courses,FUN=su
```

And compare to term data.

```
terms_merged=left_join(x=terms, y=units_from_courses, by=c("mellon_id","term_code"))
```

Do we have the information still for all the terms?

```
mean(!is.na(terms_merged$units_completed))
```

```
## [1] 0.9071387
```

Which terms are missing?

```
aggregate(!is.na(units_completed)~term_code,terms_merged,FUN=mean)
```

```
##    term_code !is.na(units_completed)
## 1     200892               0.9928292
## 2     200903               0.9945031
## 3     200914               0.9955713
## 4     200992               0.9925899
## 5     201003               0.9936021
## 6     201014               0.9927034
## 7     201092               0.9855473
## 8     201103               0.9822640
## 9     201114               0.9766332
## 10    201192               0.9777061
## 11    201203               0.9726144
## 12    201214               0.9717022
## 13    201292               0.9741494
## 14    201303               0.9709668
## 15    201314               0.9690867
## 16    201392               0.9754517
## 17    201403               0.9752616
## 18    201414               0.9671822
## 19    201492               0.9763290
## 20    201503               0.9728796
## 21    201514               0.9702727
## 22    201592               0.9746302
## 23    201603               0.9738353
## 24    201614               0.9677653
```

```
## 25    201692                0.9738256
## 26    201703                0.9756993
## 27    201714                0.9730991
## 28    201792                0.9707521
## 29    201803                0.9683804
## 30    201814                0.9623719
## 31    201892                0.9574394
## 32    201903                0.0000000
## 33    201914                0.0000000
## 34    201992                0.9803428
## 35    202003                0.0000000
## 36    202014                0.0000000
## 37    202051                0.0000000
## 38    202092                0.0000000
## 39    202103                0.0000000
## 40    202114                0.0000000
## 41    202192                0.0000000
## 42    202203                0.0000000
## 43    202214                0.0000000
```

It looks like units_completed is mostly missing in courses data from 2019 on. How well do both data sources match apart from availability?

```
mean(terms_merged$units_completed==terms_merged$current_units_completed_total, na.rm=T)
```

```
## [1] 0.4890003
```

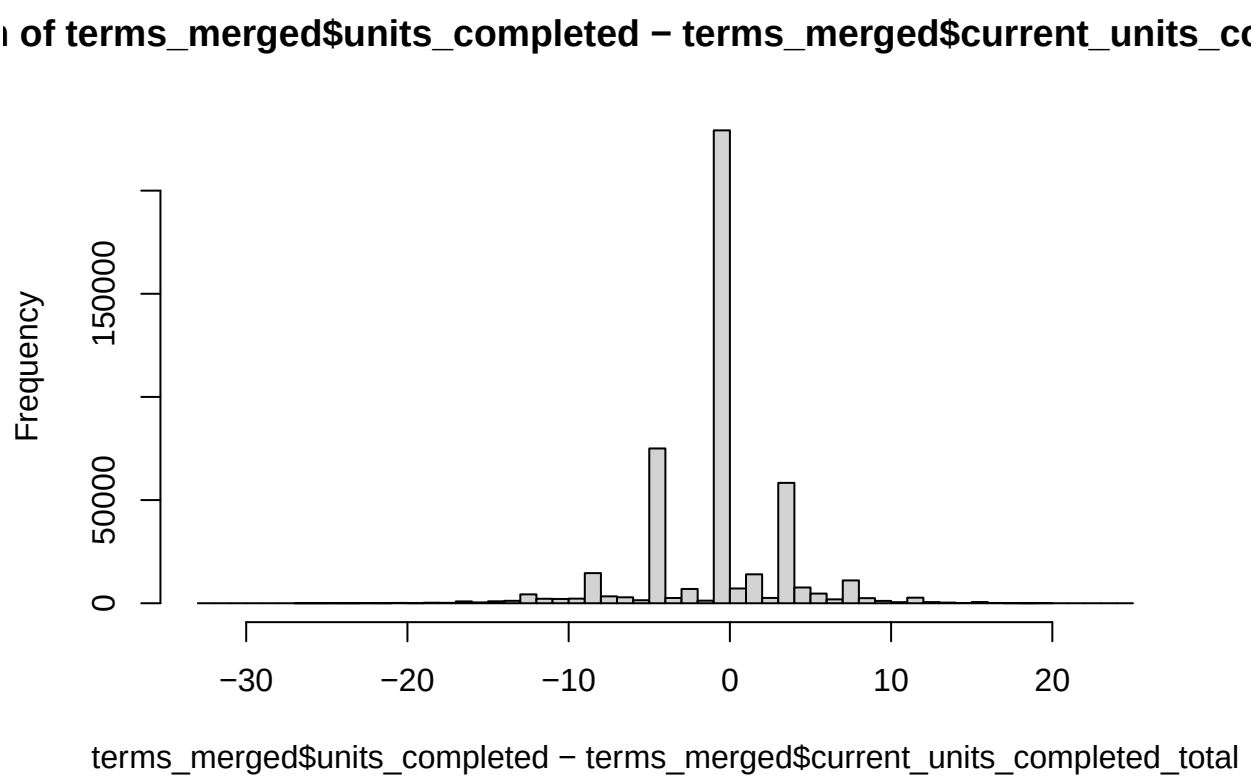
In 48.9% of the cases where we have both sources, they match exactly.

```
mean(terms_merged$units_completed-terms_merged$current_units_completed_total, na.rm=T)
```

```
## [1] -0.2838496
```

Mean difference is not very big in terms of the absolute values. Let's check distribution.

```
hist(terms_merged$units_completed-terms_merged$current_units_completed_total, breaks=50)
```



It looks like often 0, 4 or 8 units difference is there.

```
aggregate(units_completed-current_units_completed_total~admitdate, terms_merged, FUN=mean, na.action = r
```

```
##   admitdate units_completed - current_units_completed_total
## 1       F08                                     -0.80716534
## 2       F09                                     -0.90331400
## 3       F10                                     -0.74344217
## 4       F11                                     -0.52717988
## 5       F12                                     -0.18460799
## 6       F13                                      0.02015271
## 7       F14                                     -0.01281511
## 8       F15                                      0.14165686
## 9       F16                                      0.31936070
```

Slight differences between cohorts.

Some example students:

```
courses[courses$mellon_id==166410,c('term_desc','course_dept_code_and_num','units_attempted','units_comp
```

```
## # A tibble: 85 x 4
##    term_desc   course_dept_code_and_num units_attempted units_completed
##    <chr>       <chr>                              <dbl>           <dbl>
##  1 Fall 2015   ACENG 20A                              5               5
##  2 Fall 2015   ACENG 20A                              0              NA
```

```
##  3 Fall 2015   MATH 2A                                4              4
##  4 Fall 2015   MATH 2A                                0             NA
##  5 Fall 2015   PHYSICS 2                              4              4
##  6 Fall 2015   PHYSICS 2                              0             NA
##  7 Fall 2015   PHYSICS 2                              0             NA
##  8 Fall 2015   UNISTU 1                               2              2
##  9 Fall 2015   UNISTU 1                               0             NA
## 10 Winter 2016 ACENG 20B                              5              5
## # ... with 75 more rows
```

Another student:

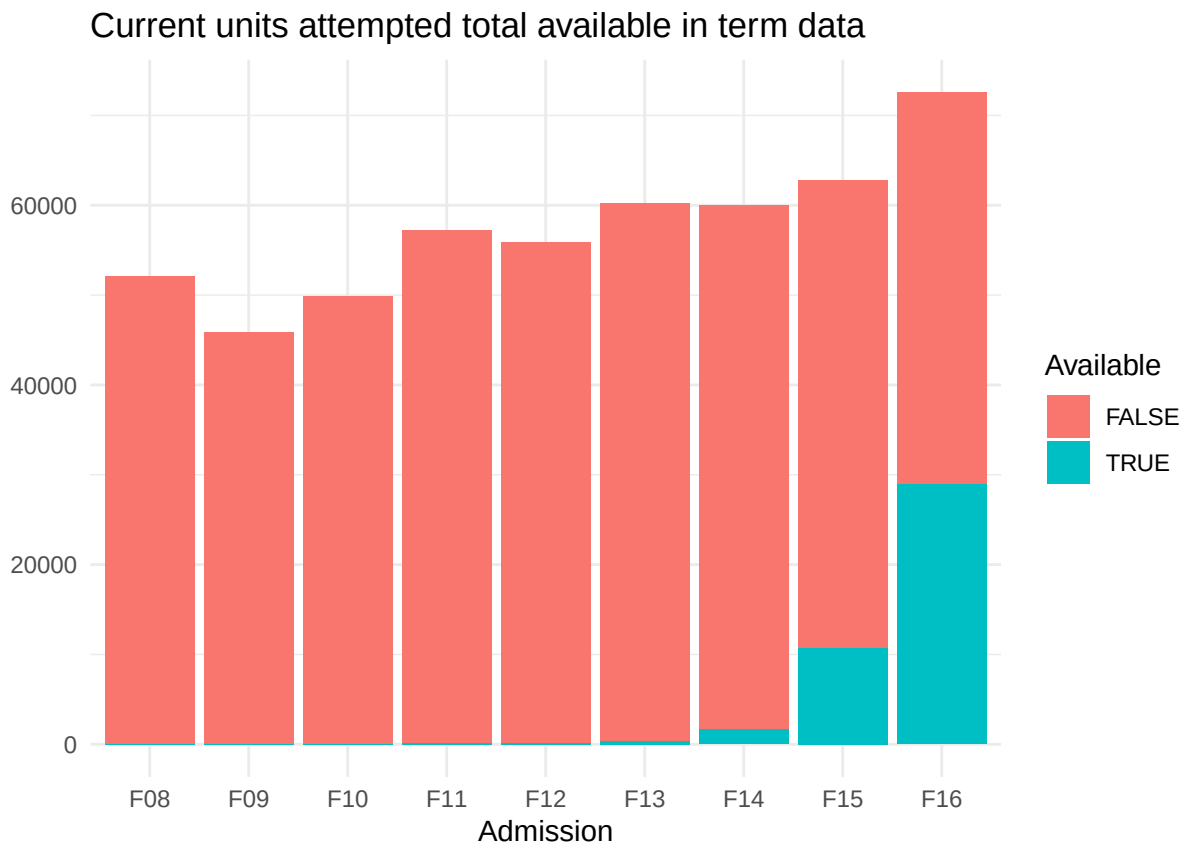
```
courses[courses$mellon_id==166128,c('term_desc','course_dept_code_and_num','units_attempted','units_comp
```

```
## # A tibble: 89 x 4
##    term_desc   course_dept_code_and_num units_attempted units_completed
##    <chr>       <chr>                              <dbl>           <dbl>
##  1 Fall 2016   BME 1                                  3               3
##  2 Fall 2016   CHEM 1A                                4               4
##  3 Fall 2016   CHEM 1A                                0              NA
##  4 Fall 2016   ENGR 93                                1               1
##  5 Fall 2016   WRITING 39B                            4               4
##  6 Winter 2017 CHEM 1B                                4               4
##  7 Winter 2017 CHEM 1B                                0              NA
##  8 Winter 2017 MATH 2D                                4               4
##  9 Winter 2017 MATH 2D                                0              NA
## 10 Winter 2017 PHILOS 4                               4               4
## # ... with 79 more rows
```

Units attempted total

In term data

```
ggplot(terms, aes(x=admitdate, fill=!is.na(current_units_attempted_total))) +
  geom_bar() +
  theme_minimal() +
  labs(title = "Current units attempted total available in term data",
       x = "Admission",
       y = "", fill="Available")
```




```
length(unique(terms[!is.na(terms$current_units_attempted_total),'mellon_id']))
```

```
## [1] 11454
```

For 11612 distinct students we have information about current_units_attempted_total in term data

In courses data

Let's see availability from courses data.

```
mean(!is.na(terms_merged$units_attempted))
```

```
## [1] 0.9071387
```

Which terms are missing?

```
aggregate(!is.na(units_attempted)~term_code,terms_merged,FUN=mean)
```

```
##    term_code !is.na(units_attempted)
## 1     200892               0.9928292
## 2     200903               0.9945031
## 3     200914               0.9955713
## 4     200992               0.9925899
```

```
## 5     201003               0.9936021
## 6     201014               0.9927034
## 7     201092               0.9855473
## 8     201103               0.9822640
## 9     201114               0.9766332
## 10    201192               0.9777061
## 11    201203               0.9726144
## 12    201214               0.9717022
## 13    201292               0.9741494
## 14    201303               0.9709668
## 15    201314               0.9690867
## 16    201392               0.9754517
## 17    201403               0.9752616
## 18    201414               0.9671822
## 19    201492               0.9763290
## 20    201503               0.9728796
## 21    201514               0.9702727
## 22    201592               0.9746302
## 23    201603               0.9738353
## 24    201614               0.9677653
## 25    201692               0.9738256
## 26    201703               0.9756993
## 27    201714               0.9730991
## 28    201792               0.9707521
## 29    201803               0.9683804
## 30    201814               0.9623719
## 31    201892               0.9574394
## 32    201903               0.0000000
## 33    201914               0.0000000
## 34    201992               0.9803428
## 35    202003               0.0000000
## 36    202014               0.0000000
## 37    202051               0.0000000
## 38    202092               0.0000000
## 39    202103               0.0000000
## 40    202114               0.0000000
## 41    202192               0.0000000
## 42    202203               0.0000000
## 43    202214               0.0000000
```

It looks like units_attempted is mostly missing in courses data from 2019 on. How well do both data sources match apart from availability?

```
mean(terms_merged$units_attempted==terms_merged$current_units_attempted_total, na.rm=T)
```

```
## [1] 0.9943856
```

In 99.4% of the cases where we have both sources, they match exactly.

```
mean(terms_merged$units_attempted-terms_merged$current_units_attempted_total, na.rm=T)
```

```
## [1] -0.02887392
```

Taking units attempted from courses data could be a good idea, given that we find this information from 2019 on.