University of Auckland Citation Analysis: 2015-2024

Tom Saunders

Table of contents

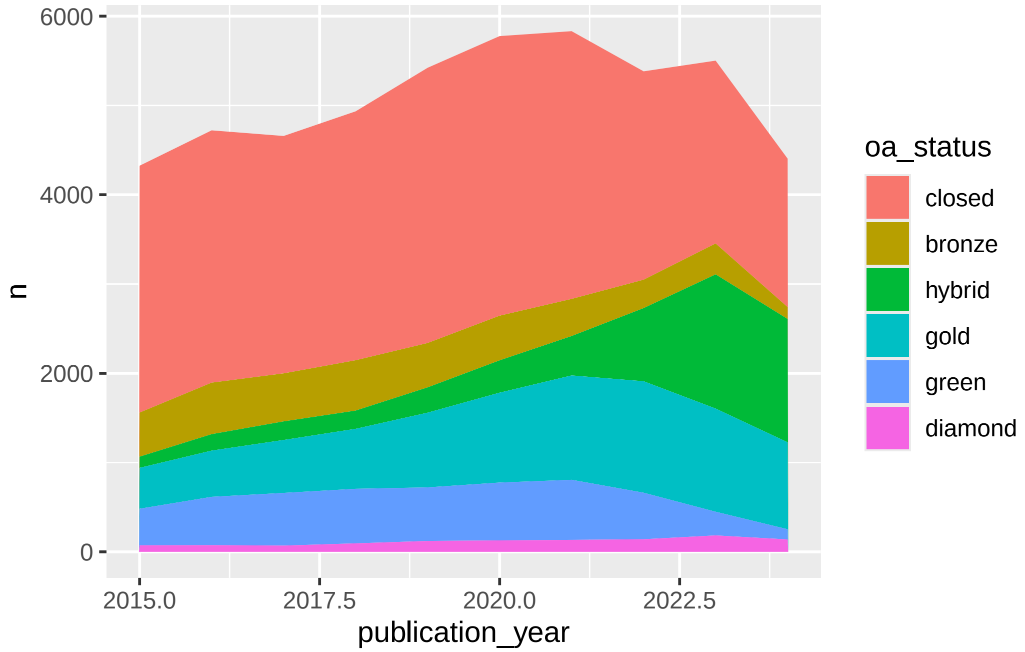
```
library(dplyr)
```

```


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
```

```
library(ggplot2)
library(scales)
library(tidyr)
```

```
citations <- read.csv("uoa-citations.csv")
```

```
# Publication output over time by oa status

citations$oa_status <- factor(citations$oa_status, levels = c("closed",
                                                              "bronze",
                                                              "hybrid",
                                                              "gold",
                                                              "green",
                                                              "diamond"))

citations |>
  group_by(publication_year, oa_status) |>
  tally() |>
  ggplot(aes(x = publication_year, y = n, fill = oa_status)) +
  geom_area()
```



```
# Number of different items in each type

citations |>
  count(type) |>
  mutate(freq = n / sum(n)) |>
  arrange(desc(n))
```

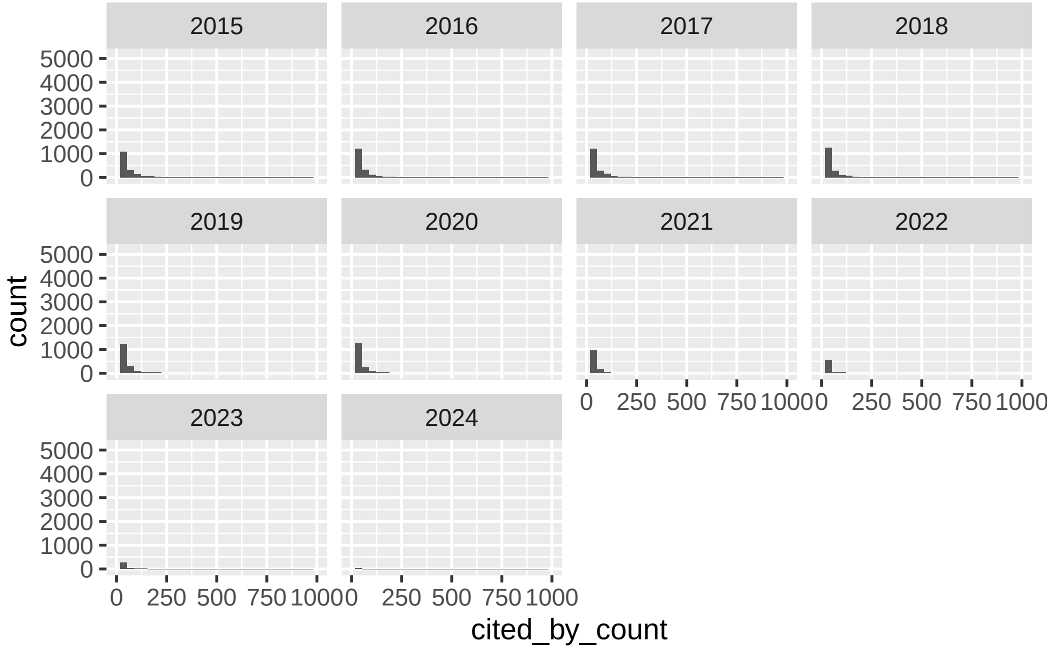
```
           type     n        freq
1       article 46659 0.915654375
2 book-chapter   3915 0.076829484
3          book   383 0.007516141
```

```
# Spread of citations for each year

citations |>
  ggplot(aes(x = cited_by_count)) +
  geom_histogram() +
  xlim(0,1000) +
  facet_wrap(vars(publication_year))
```

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

```
Warning: Removed 28 rows containing non-finite outside the scale range
(`stat_bin()`).
```

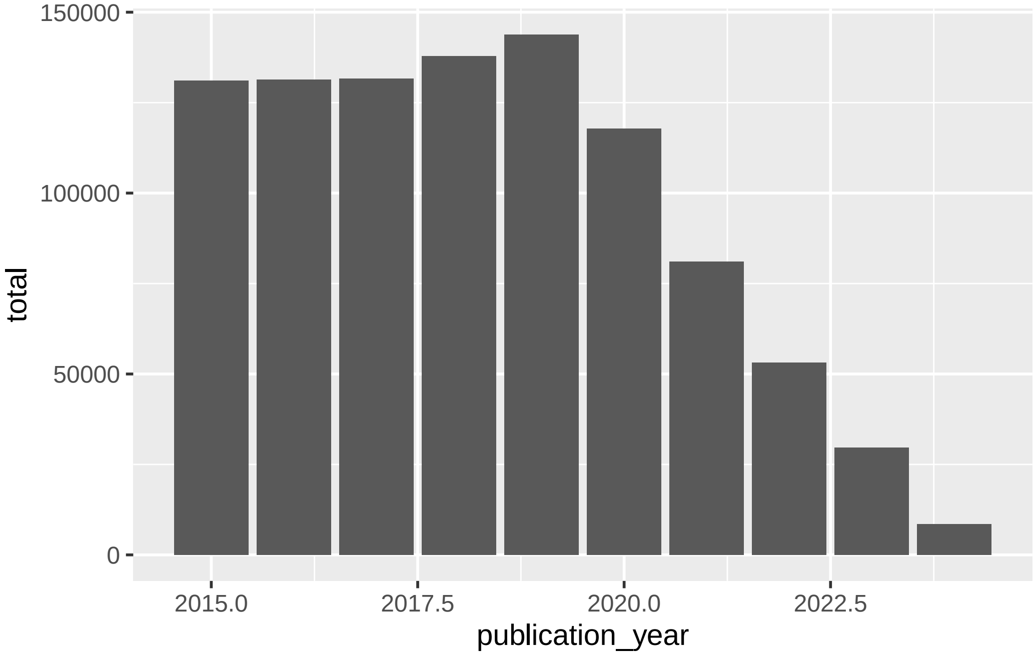
```
Warning: Removed 20 rows containing missing values or values outside the scale range
(`geom_bar()`).
```



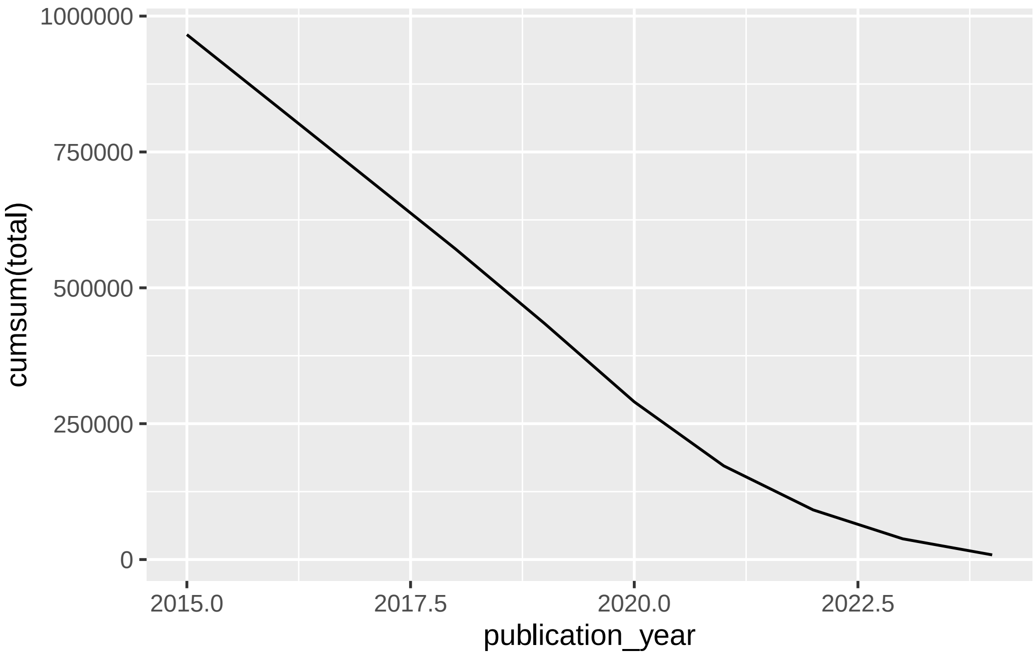
```
# Number of citations each year (table)

citations |>
  group_by(publication_year) |>
  summarise(
    total = sum(cited_by_count),
    median = median(cited_by_count),
    average = mean(cited_by_count)
  ) |>
  arrange(desc(publication_year))
```

```
# A tibble: 10 x 4
   publication_year  total median average
              <int>  <int>  <dbl>   <dbl>
 1             2024   8550      0    1.94
 2             2023  29592      2    5.38
 3             2022  53170      4    9.88
 4             2021  81081      6   13.9
 5             2020 117818      8   20.4
 6             2019 143803      9   26.5
 7             2018 137796     11   27.9
 8             2017 131633     11   28.3
 9             2016 131427     11   27.8
10             2015 131048     11   30.3
```

```
# Number of citations each year (plot)

citations |>
  group_by(publication_year) |>
  summarise(
    total = sum(cited_by_count),
    median = median(cited_by_count),
    average = mean(cited_by_count)
  ) |>
  arrange(desc(publication_year)) |>
  ggplot(aes(x = publication_year, y = total)) +
  geom_col()
```



```
# Cumulative number of citations over time

citations |>
  group_by(publication_year) |>
  summarise(
    total = sum(cited_by_count),
    median = median(cited_by_count),
    average = mean(cited_by_count)
  ) |>
  arrange(desc(publication_year)) |>
  ggplot(aes(x = publication_year, y = cumsum(total))) +
  geom_line()
```

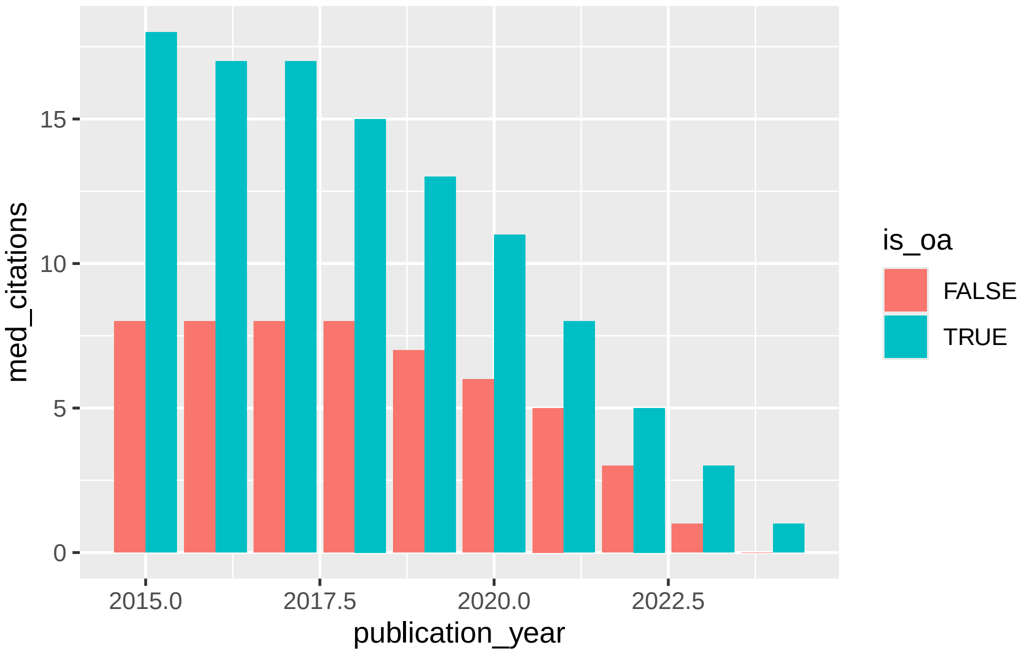


```
# Overall, mean & median citations, OA vs closed

citations |>
  group_by(is_oa) |>
  summarise(
    avg_citations = mean(cited_by_count),
    med_citations = median(cited_by_count)
  )
```

```
# A tibble: 2 x 3
  is_oa avg_citations med_citations
  <lgl>         <dbl>         <dbl>
1 FALSE          16.1             5
2 TRUE           22.0             6
```

```
# Each year, median citation OA vs closed

citations |>
  group_by(publication_year, is_oa) |>
  summarise(
    med_citations = median(cited_by_count)
  ) |>
  ggplot(aes(x = publication_year, y = med_citations, fill = is_oa)) +
  geom_bar(position="dodge", stat="identity")
```

```
`summarise()` has grouped output by 'publication_year'. You can override using
the `.groups` argument.
```



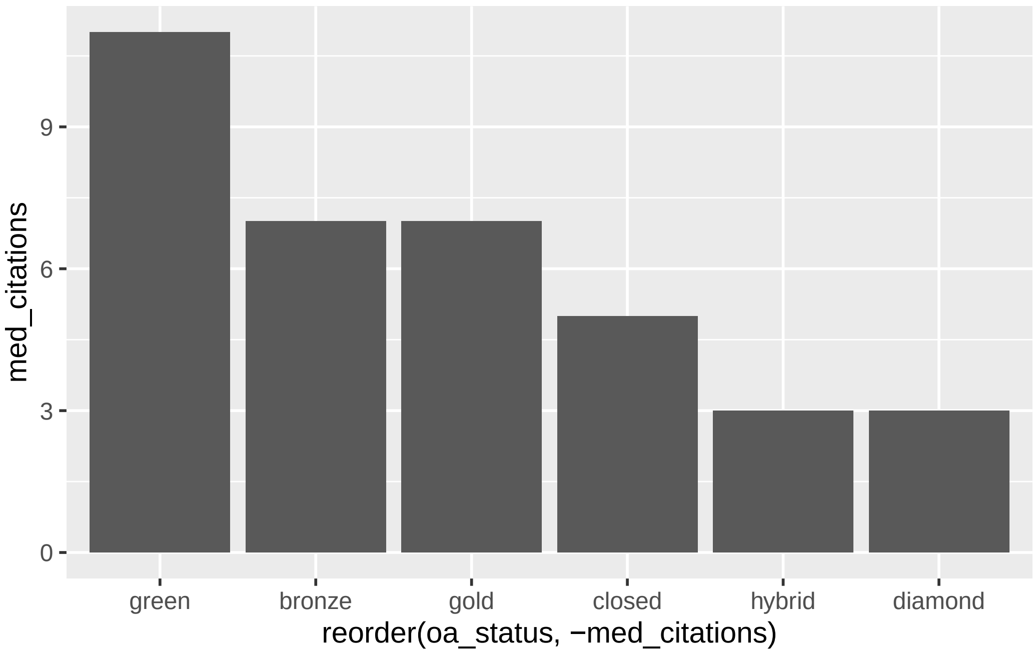
```
# Difference in citations each year between open vs closed

cit_diff_oa <-
citations |>
  group_by(publication_year, is_oa) |>
  summarise(
    med_citations = median(cited_by_count),
  ) |>
  pivot_wider(names_from = is_oa, values_from = med_citations) |>
  mutate(
    difference = `FALSE` / `TRUE`,
  )
```

```
`summarise()` has grouped output by 'publication_year'. You can override using
the `.groups` argument.
```

```
cat("Each year, open access items enjoy median citation rates between", scales::percent(min(
```

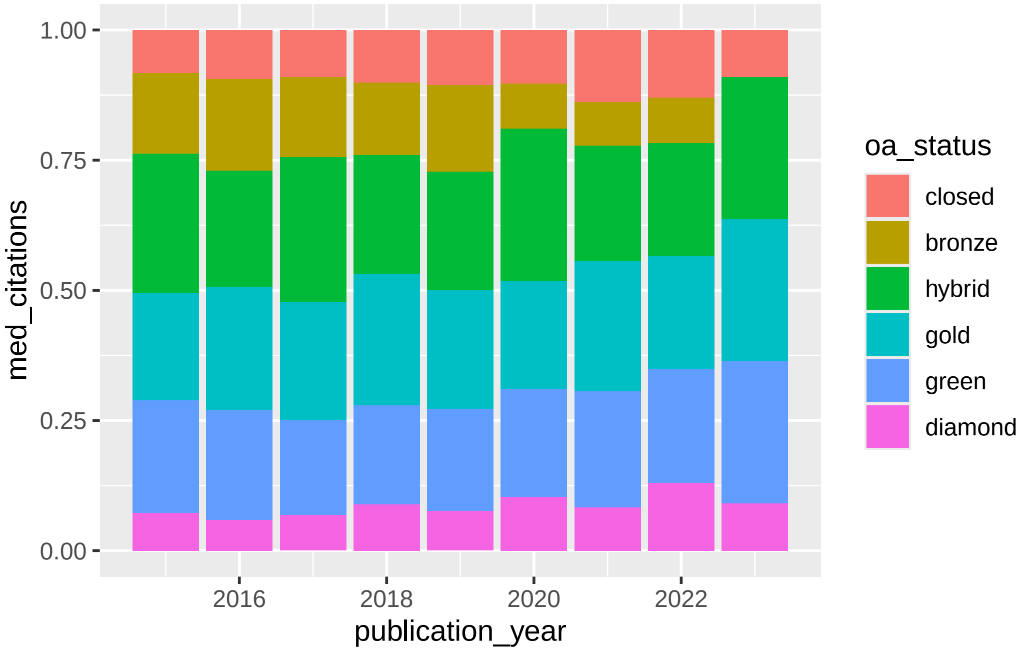
```
Each year, open access items enjoy median citation rates between 0% and 62% higher than clos
```

```
# Overall, median citations for each type of OA vs closed

citations |>
  group_by(oa_status) |>
  summarise(
    med_citations = median(cited_by_count)
  ) |>
  ggplot(aes(x = reorder(oa_status, -med_citations), y = med_citations)) +
  geom_col()
```



```
# Each year, median citations for each type of OA vs closed

citations |>
  filter(publication_year != 2024) |>
  group_by(publication_year, oa_status) |>
  summarise(
    med_citations = median(cited_by_count),
  ) |>
  ggplot(aes(x = publication_year, y = med_citations, fill = oa_status)) +
  geom_bar(position="fill", stat="identity")
```

```
`summarise()` has grouped output by 'publication_year'. You can override using
the `.groups` argument.
```


med_citations
publication_year
oa_status
closed
bronze
hybrid
gold
green
diamond