National Research University Higher School of Economics

Master's Programme 'Data Analytics and Social Statistics (DASS)'

Exploratory Data Analysis (prof. Batagelj)

Project 2: OpenAlex

Contents

The project was prepared by ***Timofei Korovin****, DASS student*

Short formulation of the task:

We should select an institution with at least 25000 published works. In our case it was MSU (Lomonosov Moscow State University). We should extract data from the OpenAlex database, then analyze the resulted. data set

Creation date: 24/04/2025

The last change date: 28/04/2025

Preface

In my project, I used the following strategy to offload data from OpenAlex. Instead of using the API right away, I simply uploaded the data in csv format. In the zip to the project, I will attach this dataset. The reason behind this decision is purely technical. I tried to use the API, but due to the large amount of papers, my session in R was hanging. So, first I used a csv file. Then, while analyzing the dataset, I realized that the maximum number of authors per one paper is 100. This was described in the OpenAlex documentation for the API use. For this reason, I unloaded the IDs of the papers that had exactly 100 authors and re-parsed the data for them using the API. Then I updated the data for analysis and performed the tasks specified in the project description.

CSV file import and preprocess

```r
df <- read.csv("works.csv")
```

```r
library(dplyr)
```

```
## Warning:     'dplyr'          R     4.4.3
```

```
## 
##           : 'dplyr'
```

```
##                    'package:stats':
## 
##     filter, lag
```

```
##                    'package:base':
## 
##     intersect, setdiff, setequal, union
```

```r
library(tidyr)
```

```
## Warning:     'tidyr'          R     4.4.3
```

```r
library(stringr)
library(purrr)


authors_clean2 <- df %>%
  select(id, authorships.author.display_name, authorships.author.id, authorships.countries, publication_
  mutate(
    author_names = str_split(authorships.author.display_name, pattern = "\\|"),
    author_ids = str_split(authorships.author.id, pattern = "\\|"),
    author_countries = str_split(authorships.countries, pattern = "\\|")
  ) %>%
  unnest_longer(author_names, indices_include = TRUE) %>%
  rename(author_position = author_names_id) %>%
  mutate(
    author_id = map2_chr(author_ids, author_position, ~ .x[.y]),
    author_country = map2_chr(author_countries, author_position, ~ .x[.y])
  ) %>%
  mutate(
    author_names = str_trim(author_names),
    author_id = str_trim(author_id),
    author_country = str_trim(author_country)
  ) %>%
  select(id, author_id, author_names, author_country, publication_year)
```

100 authors works detection

```
authors_per_work_test <- df %>%
  mutate(num_authors = str_count(authorships.author.display_name, "\\|") + 1) %>%
  select(id, num_authors)

authors_per_work_test %>%
  filter(num_authors == 100) %>%
  summarise(number_of_works = n())
```

```
##   number_of_works
## 1            1077
```

```
works_with_100_authors <- df %>%
  mutate(num_authors = str_count(authorships.author.display_name, "\\|") + 1) %>%
  filter(num_authors == 100) %>%
  select(id)

works_with_100_authors$cleaned_id <- gsub("https://openalex.org/", "", works_with_100_authors$id)
```

Parsing data for 100 authors papers

```
library(httr)
```

```
## Warning:     'httr'          R     4.4.3
```

```
library(jsonlite)
```

```
## Warning:     'jsonlite'          R     4.4.3
```

```
##
##           : 'jsonlite'
```

```
##                 'package:purrr':
##
##     flatten
```

```r
all_work_details <- list()


for (work_id in works_with_100_authors$cleaned_id) {
  url <- paste0("https://api.openalex.org/works/", work_id)
  res <- GET(url)

  if (status_code(res) == 200) {
    work_data <- fromJSON(rawToChar(res$content))
    all_work_details <- append(all_work_details, list(work_data))
  } else {
    message(paste("Error:", work_id))
  }
}

df_new_works2 <- map_dfr(all_work_details, ~ tibble(
  id = .x$id,
  publication_year = .x$publication_year,
  authorships = list(.x$authorships)
))

df_new_works2 <- df_new_works2 %>%
  unnest_longer(authorships) %>%
  unnest_wider(authorships)

df_new_works2 <- df_new_works2 %>%
  unnest_wider(author, names_sep = "_")

df_new_works2 <- df_new_works2 %>%
  unnest_longer(countries) %>%
  rename(country_code = countries)

df_new_works2 <- df_new_works2 %>%
  select(id, author_id, author_display_name, country_code, publication_year) %>%
  rename(author_names = author_display_name) %>%
  rename(author_country = country_code)

df_new_works2 <- df_new_works2 %>%
  mutate(publication_year = as.character(publication_year)) %>%
  mutate(author_country = as.character(author_country))

authors_clean2_updated <- authors_clean2 %>%
  filter(!id %in% df_new_works2$id)

authors_final <- bind_rows(authors_clean2_updated, df_new_works2)
```

Task 1

```r
author_total_works <- authors_final %>%
  group_by(author_id, author_names) %>%
  summarise(total_works = n(), .groups = 'drop')
```

```
authors_per_work <- authors_final %>%
  group_by(id) %>%
  summarise(n_authors = n(), .groups = 'drop')

authors_with_fraction <- authors_final %>%
  left_join(authors_per_work, by = "id") %>%
  mutate(fractional_contribution = 1 / n_authors)

author_total_fraction <- authors_with_fraction %>%
  group_by(author_id, author_names) %>%
  summarise(total_fractional_contribution = sum(fractional_contribution), .groups = 'drop')

work_internationality <- authors_final %>%
  group_by(id) %>%
  summarise(
    n_countries = n_distinct(author_country, na.rm = TRUE),
    .groups = 'drop'
  ) %>%
  mutate(international = ifelse(n_countries > 1, TRUE, FALSE))

authors_with_international_flag <- authors_final %>%
  left_join(work_internationality, by = "id")

author_international_works <- authors_with_international_flag %>%
  filter(international == TRUE) %>%
  group_by(author_id, author_names) %>%
  summarise(total_international_works = n(), .groups = 'drop')

final_summary <- author_total_works %>%
  left_join(author_total_fraction, by = c("author_id", "author_names")) %>%
  left_join(author_international_works, by = c("author_id", "author_names")) %>%
  mutate(total_international_works = ifelse(is.na(total_international_works), 0, total_international_wo
```

```
top10_authors <- final_summary %>%
  arrange(desc(total_works)) %>%
  slice_head(n = 10) %>%
  pull(author_id)

top10_authors
```

```
##  [1] "https://openalex.org/A5108065580" "https://openalex.org/A5000154066"
##  [3] "https://openalex.org/A5023708186" "https://openalex.org/A5106455834"
##  [5] "https://openalex.org/A5114375500" "https://openalex.org/A5021849393"
##  [7] "https://openalex.org/A5057736043" "https://openalex.org/A5106498549"
##  [9] "https://openalex.org/A5046346772" "https://openalex.org/A5107749023"
```

```
top_authors_data <- authors_final %>%
  filter(author_id %in% top10_authors)

top_authors_data <- top_authors_data %>%
  left_join(authors_per_work, by = "id") %>%
  mutate(fractional_contribution = 1 / n_authors) %>%
  left_join(work_internationality, by = "id")

top_authors_yearly_summary <- top_authors_data %>%
  group_by(author_id, author_names, publication_year) %>%
  summarise(
    works_per_year = n(),
    fractional_contribution_per_year = sum(fractional_contribution),
    international_works_per_year = sum(international, na.rm = TRUE),
    .groups = 'drop'
  )
```

Visual interpretation

```
library(tidyverse)
```

```
## Warning:     'tidyverse'         R     4.4.3
```

```
## Warning:     'ggplot2'         R     4.4.3
```

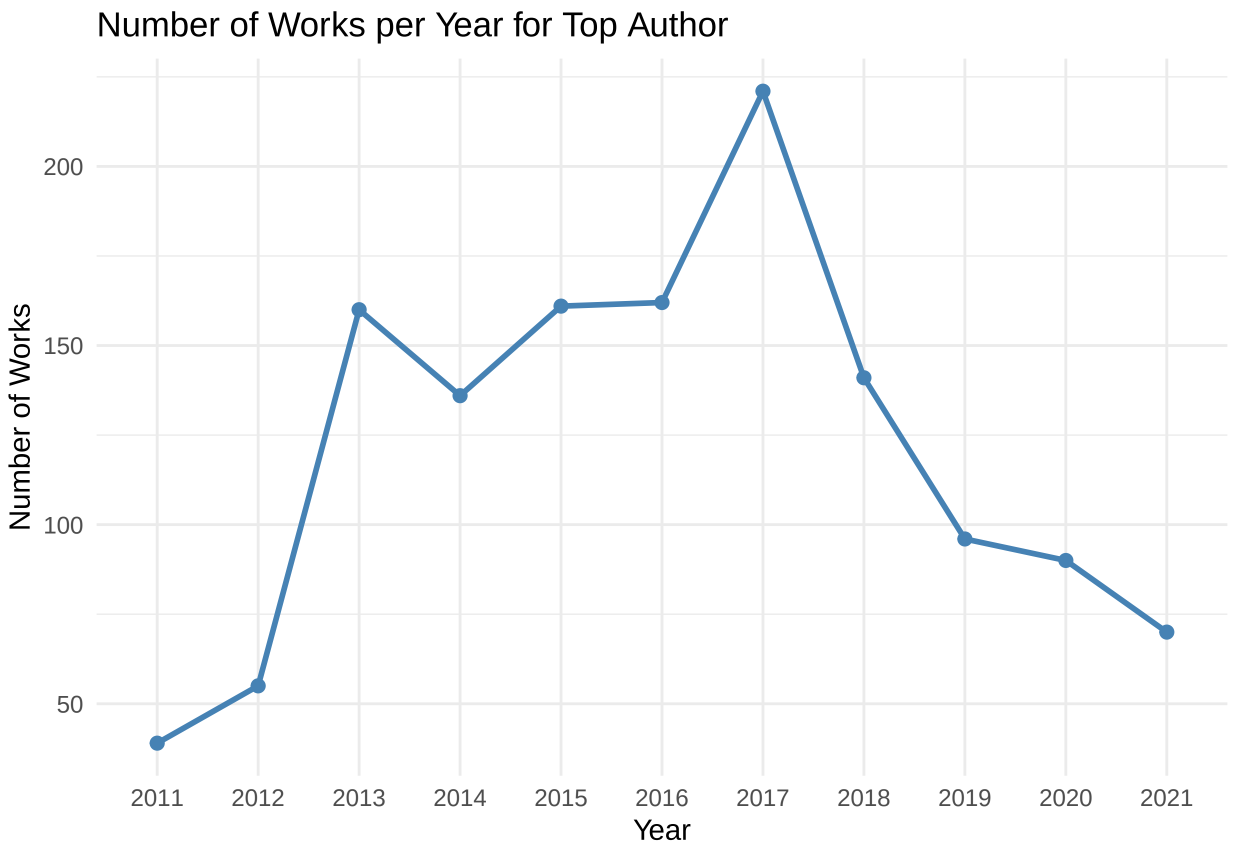
```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v readr     2.1.5
## v ggplot2   3.5.2     v tibble    3.2.1
## v lubridate 1.9.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter()     masks stats::filter()
## x jsonlite::flatten() masks purrr::flatten()
## x dplyr::lag()        masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
top_author_id_name <- final_summary %>%
  filter(total_works == max(total_works)) %>%
  pull(author_id, author_names)

top_author_data <- top_authors_yearly_summary %>%
  filter(author_id == top_author_id_name)

top_author_data
```

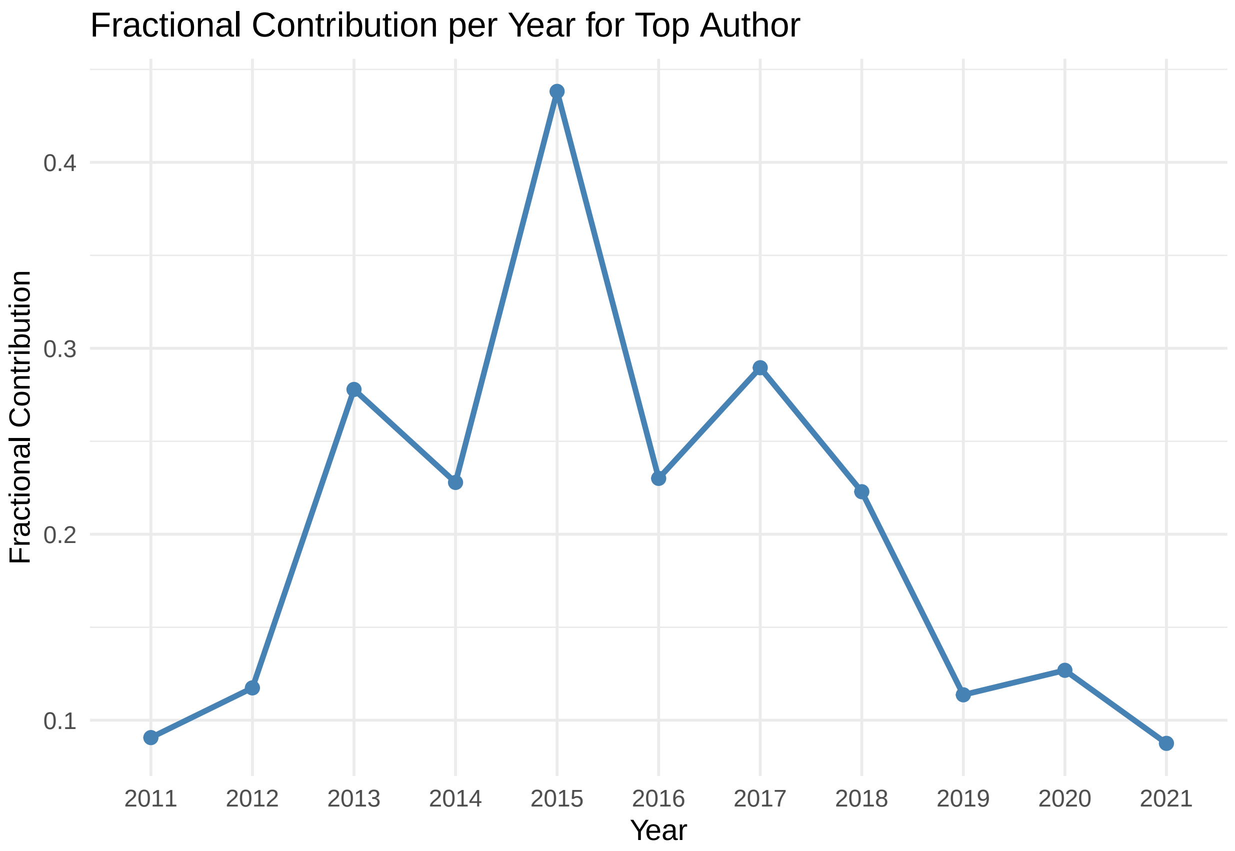
```
## # A tibble: 11 x 6
##    author_id author_names publication_year works_per_year fractional_contribut~1
##    <chr>     <chr>        <chr>                     <int>                  <dbl>
##  1 https://~ M. Williams  2011                         39                 0.0907
##  2 https://~ M. Williams  2012                         55                 0.117
##  3 https://~ M. Williams  2013                        160                 0.278
##  4 https://~ M. Williams  2014                        136                 0.228
##  5 https://~ M. Williams  2015                        161                 0.438
##  6 https://~ M. Williams  2016                        162                 0.230
##  7 https://~ M. Williams  2017                        221                 0.290
##  8 https://~ M. Williams  2018                        141                 0.223
##  9 https://~ M. Williams  2019                         96                 0.114
## 10 https://~ M. Williams  2020                         90                 0.127
## 11 https://~ M. Williams  2021                         70                 0.0875
## # i abbreviated name: 1: fractional_contribution_per_year
## # i 1 more variable: international_works_per_year <int>
```

```r
ggplot(top_author_data, aes(x = publication_year, y = works_per_year, group = 1)) +
  geom_line(color = "steelblue", size = 1) +
  geom_point(color = "steelblue", size = 2) +
  labs(title = "Number of Works per Year for Top Author",
       x = "Year",
       y = "Number of Works") +
  theme_minimal()
```

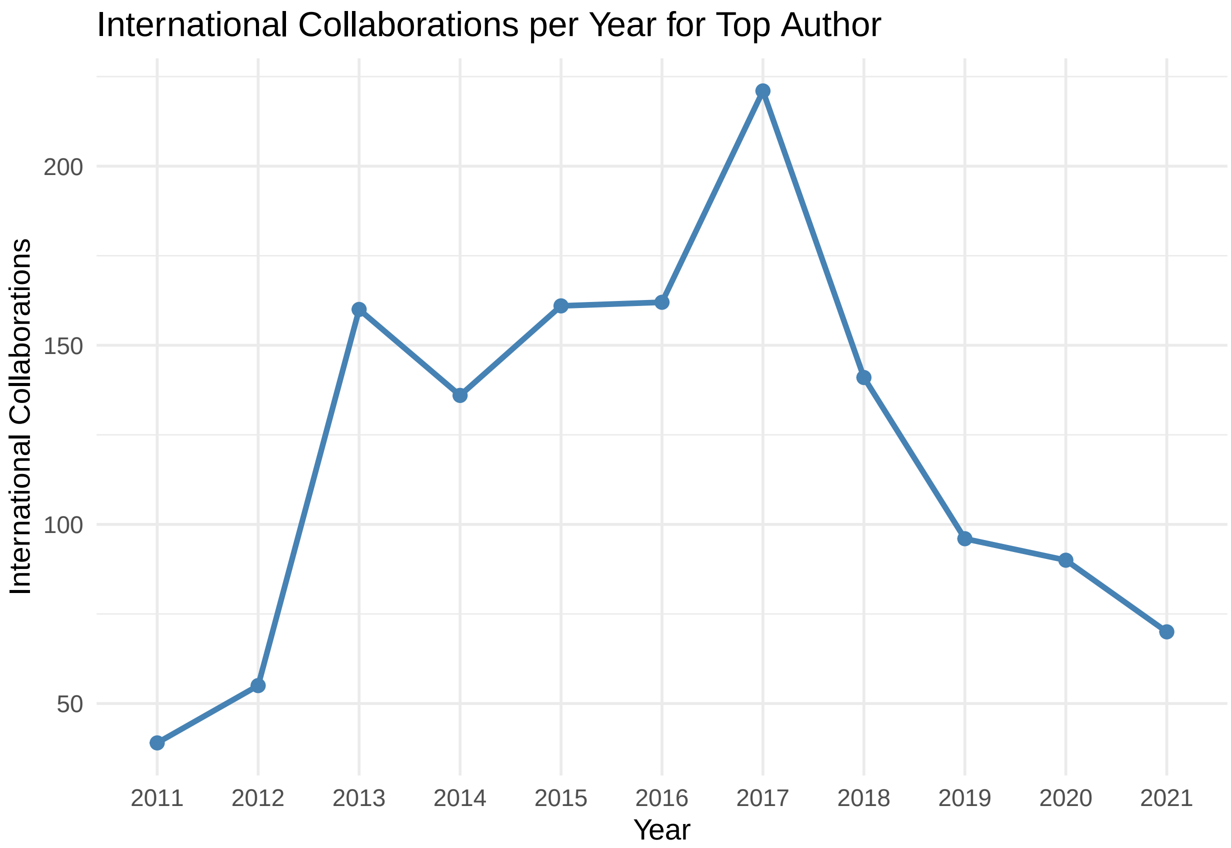
```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



```r
ggplot(top_author_data, aes(x = publication_year, y = fractional_contribution_per_year, group = 1)) +
  geom_line(color = "steelblue", size = 1) +
  geom_point(color = "steelblue", size = 2) +
  labs(title = "Fractional Contribution per Year for Top Author",
       x = "Year",
       y = "Fractional Contribution") +
  theme_minimal()
```

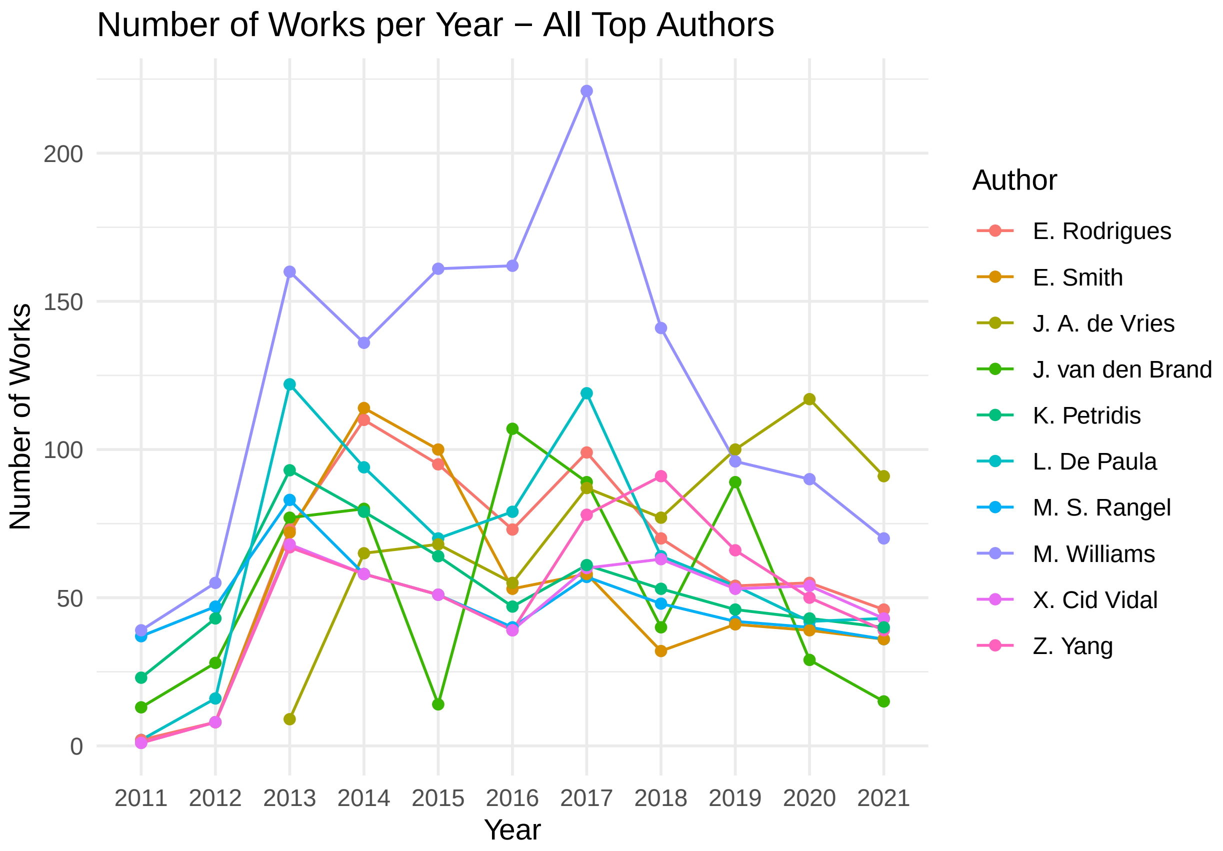


```
ggplot(top_author_data, aes(x = publication_year, y = international_works_per_year, group = 1)) +
  geom_line(color = "steelblue", size = 1) +
  geom_point(color = "steelblue", size = 2) +
  labs(title = "International Collaborations per Year for Top Author",
       x = "Year",
       y = "International Collaborations") +
  theme_minimal()
```



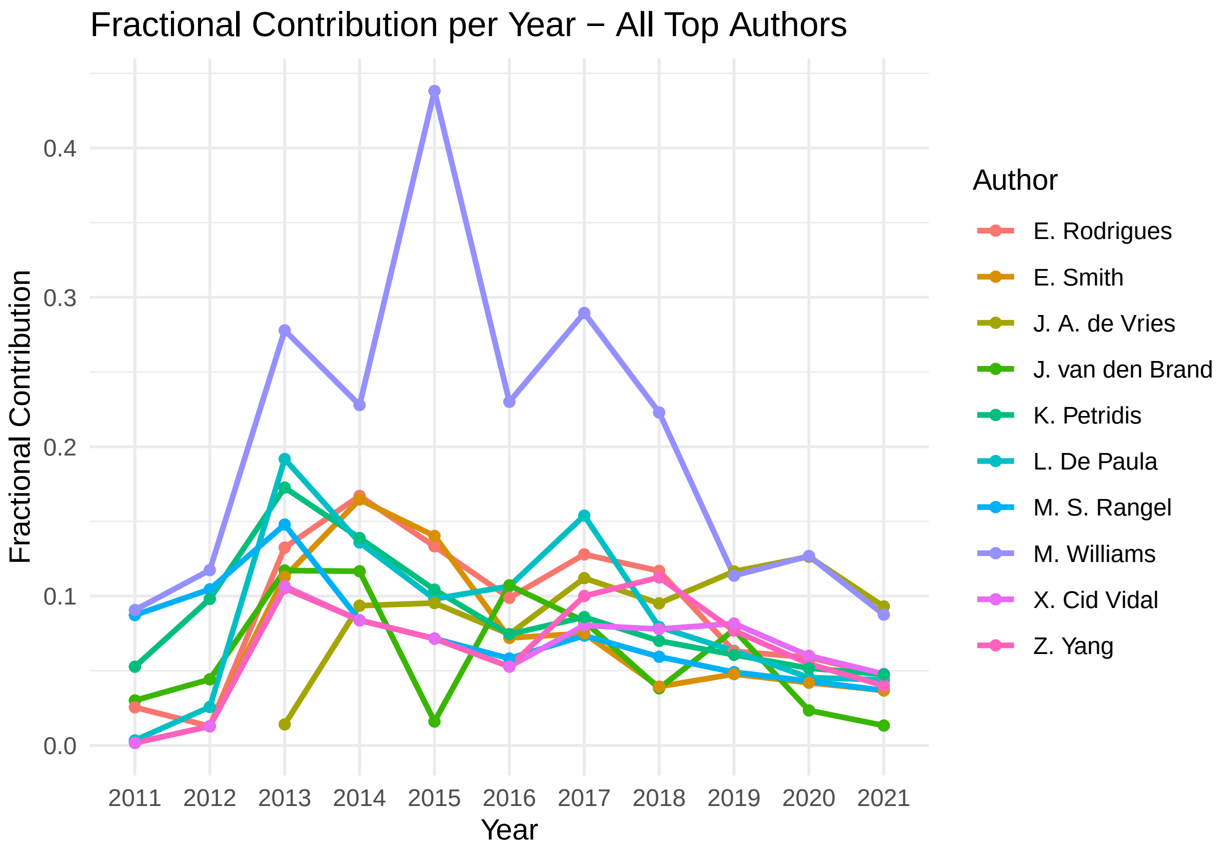
We will not interpret these charts right away since these trends will be covered below.

```
ggplot(top_authors_yearly_summary, aes(x = publication_year, y = works_per_year, color = author_names, g
  geom_line(size = 0.5) +
  geom_point() +
  labs(title = "Number of Works per Year - All Top Authors",
       x = "Year",
       y = "Number of Works",
       color = "Author") +
  theme_minimal()
```



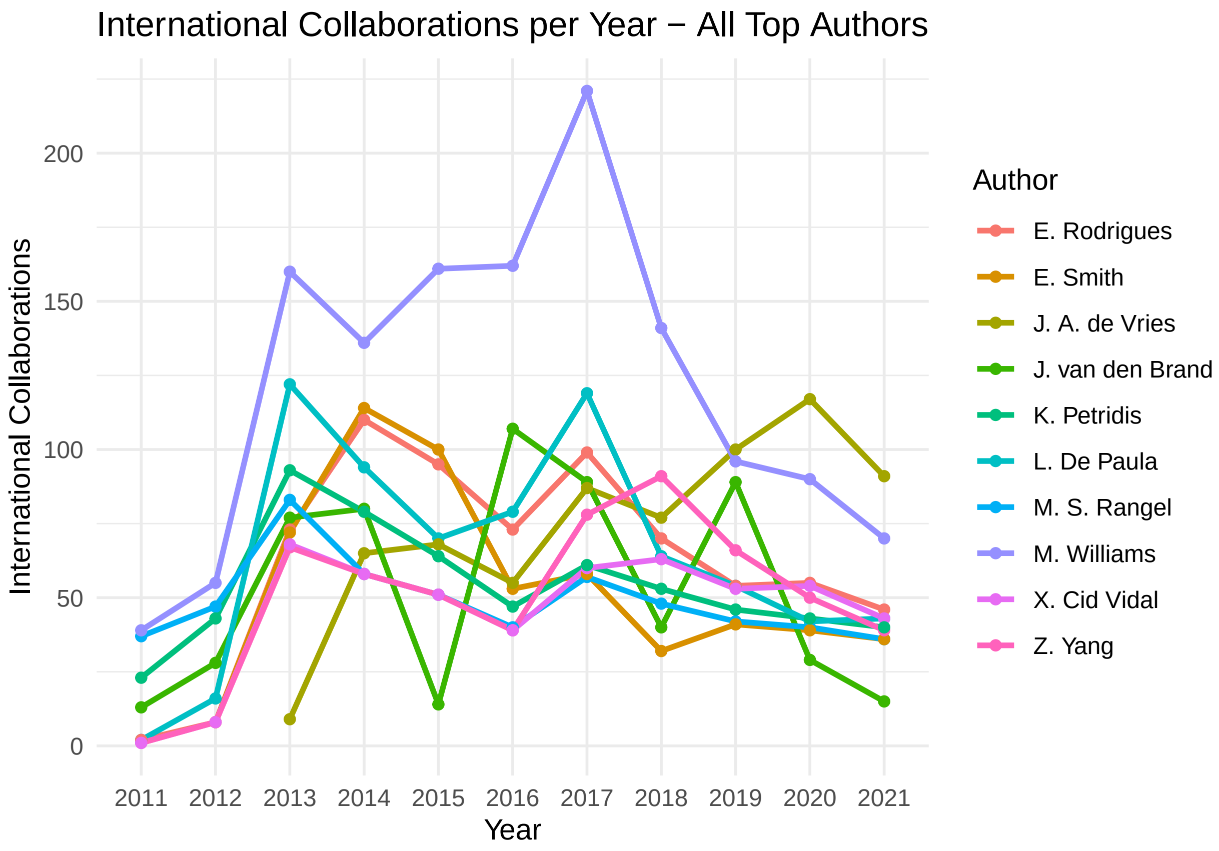
Most authors show an increase in productivity from 2011 to 2017. After that, there is a stabilization in the number of publications. Author M. Williams stands out from the rest (with picking number of 250 publications per year in 2017). Others show more moderate growth.

```
ggplot(top_authors_yearly_summary, aes(x = publication_year, y = fractional_contribution_per_year, colo
  geom_line(size = 1) +
  geom_point() +
  labs(title = "Fractional Contribution per Year - All Top Authors",
       x = "Year",
       y = "Fractional Contribution",
       color = "Author") +
  theme_minimal()
```



The fractional contribution of most authors repeats the dynamics of the total number of publications, but with smoother fluctuations. Again M.Williams demonstrate higher rates than other authors, indicating his significant role in the papers, perhaps because lower number of ollaborations with other researchers.

```
ggplot(top_authors_yearly_summary, aes(x = publication_year, y = international_works_per_year, color = a
  geom_line(size = 1) +
  geom_point() +
  labs(title = "International Collaborations per Year - All Top Authors",
       x = "Year",
       y = "International Collaborations",
       color = "Author") +
  theme_minimal()
```

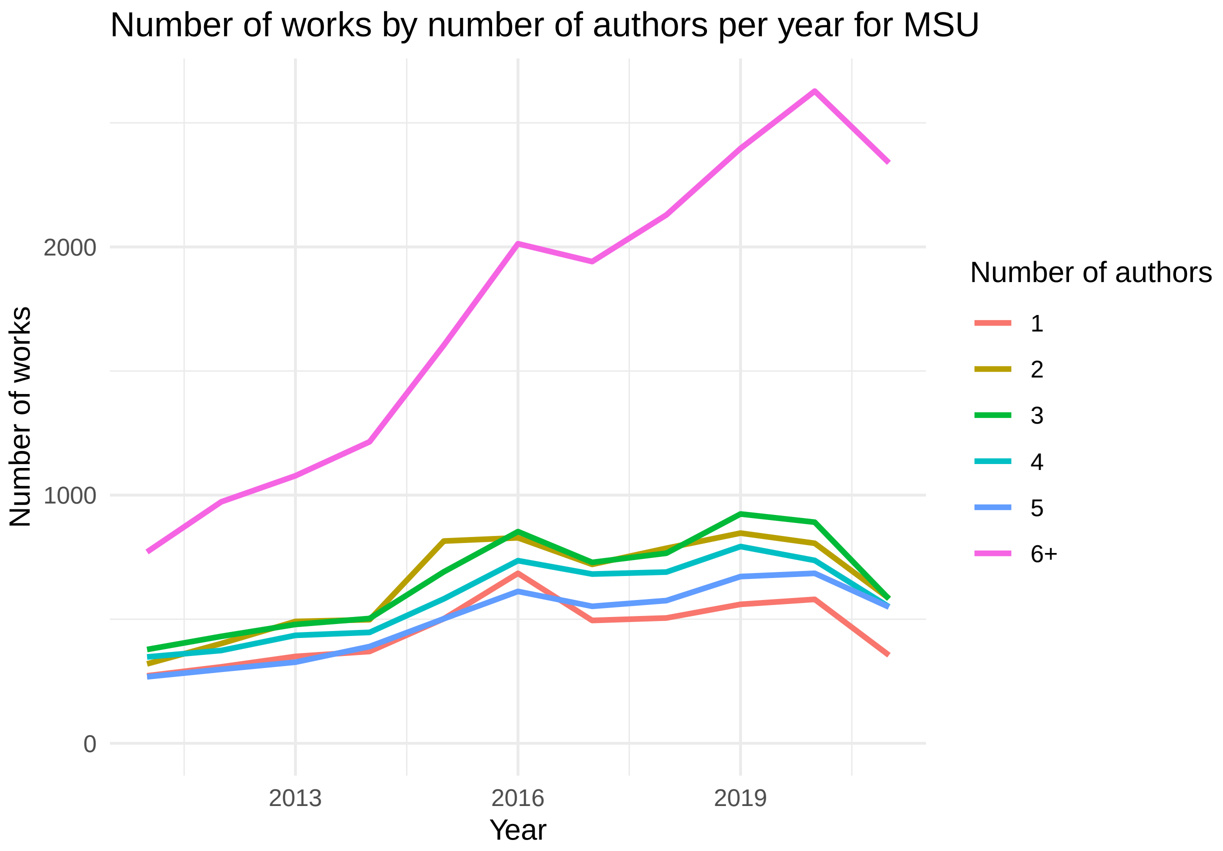


Again we see the similar patterns on the graph. M. Williams leads in the number of international collaborations. So we can conclude that international collaboration is an important component of top authors' scientific activity, especially during periods of high productivity rates.

Task 2

```
library(tidyr)

authors_per_work2 <- authors_final %>%
  group_by(id, publication_year) %>%
  summarise(n_authors = n(), .groups = 'drop')

works_by_authors_per_year <- authors_per_work2 %>%
  mutate(n_authors_group = case_when(
    n_authors == 1 ~ "1",
    n_authors == 2 ~ "2",
    n_authors == 3 ~ "3",
    n_authors == 4 ~ "4",
    n_authors == 5 ~ "5",
    n_authors >= 6 ~ "6+"
  )) %>%
  group_by(publication_year, n_authors_group) %>%
  summarise(num_works = n(), .groups = 'drop')
```

```r
works_by_authors_per_year$publication_year <- as.integer(works_by_authors_per_year$publication_year)


ggplot(works_by_authors_per_year, aes(x = publication_year, y = num_works, color = n_authors_group)) +
  geom_line(size = 1) +
  labs(
    title = "Number of works by number of authors per year for MSU",
    x = "Year",
    y = "Number of works",
    color = "Number of authors"
  ) +
  theme_minimal()
```



Overall, it might be seen on the graph that the 6+ line (purple) is noticeably higher than the others and has grown significantly over the years. In general, this is an expected result, as it is a fairly well-known fact that modern science tends to be highly collaborative and interdisciplinary, which explains the large number of papers with more than 6 authors. The rest of the lines show pretty much the same patterns.

Conclusion

To summarize, we can repeat the conclusion from the last section - that there is clearly a trend in contemporary science towards international collaboration . It was a small surprise that we see almost no Russian names among the top authors for Russian university. There can be several explanations for this trend, or hypotheses if you like. First, we uploaded the papers where MSU was mentioned. In fact, it is not the only university for these papers, since this is the way OpenAlex is organized. That is why in most cases we notice foreign surnames of authors. Thus, most of our papers are written in collaboration with other universities/authors from other countries. Whether this is a mistake or not - we cannot say now, since we interpreted the requirements to the project in such a way. The second reason may be that most Russian scientists are not inclined to collaborate with foreign authors, but this is more of a hypothesis than a fact.