Group Activity

Jonathan Cary Sucaldito

2025-12-01

```r
install.packages(c("rvest", "stringr", "lubridate","dplyr", "ggplot2"))
```

```
## Installing packages into '/cloud/lib/x86_64-pc-linux-gnu-library/4.5'
## (as 'lib' is unspecified)
```

```r
# Load required libraries
library(rvest)
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
```

```r
library(ggplot2)
library(stringr)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```r
# Function to extract and analyze arXiv papers (TOPIC: STATISTICS)
analyze_statistics <- function() {
  cat("=== ARXIV STATISTICS ANALYSIS ===\n\n")

  # arXiv ADVANCED SEARCH URL (Statistics in Title)
  url <- "https://arxiv.org/search/advanced?advanced=1&terms-0-term=Statistics&terms-0-operator=AND&term

  # Extract data from arXiv (use a user-agent to reduce chance of blocking)
  cat("Extracting data from arXiv...\n")
  webpage <- tryCatch({
    read_html(url, user_agent("Mozilla/5.0 (Windows NT 10.0; Win64; x64)"))
  }, error = function(e) {
    cat("Error reading webpage. Using sample data.\n")
    return(NULL)
  })
```

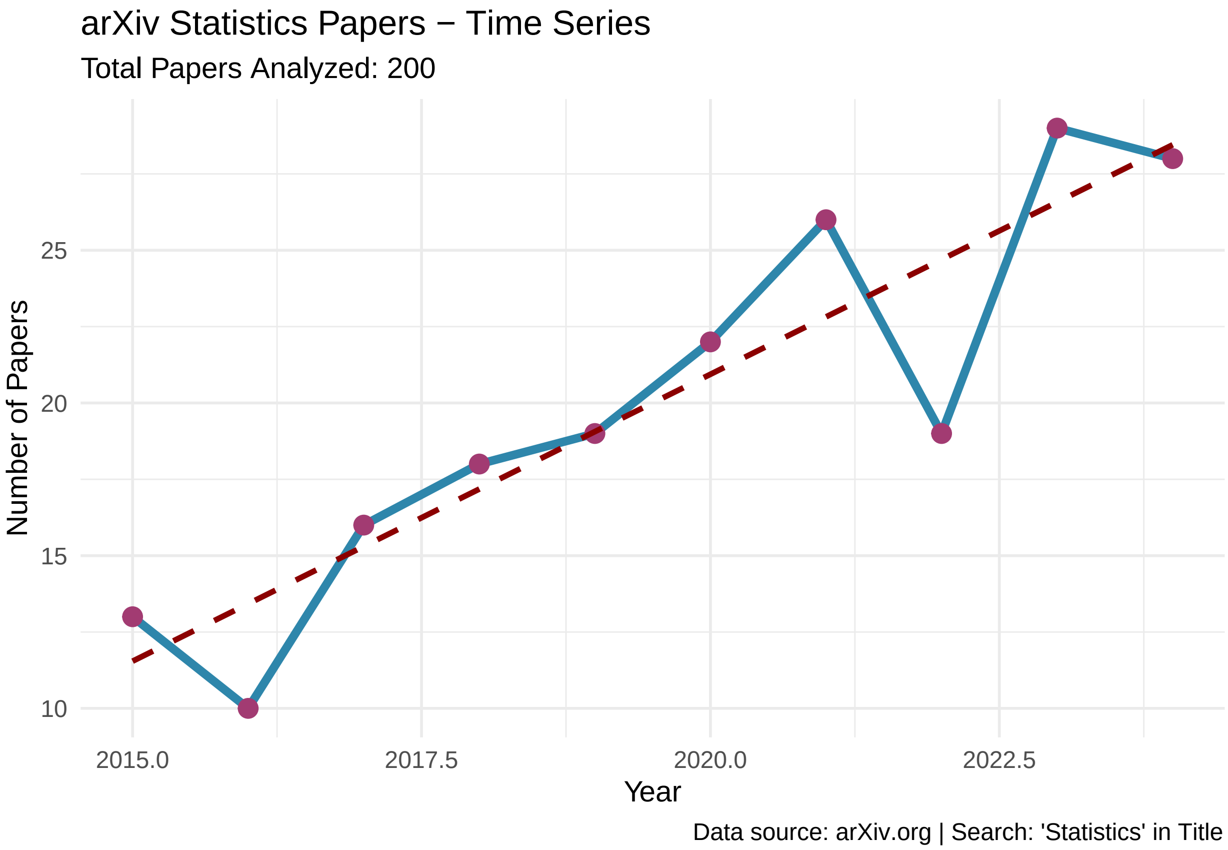
```r
if (is.null(webpage)) {
  # Fallback sample data
  set.seed(123)
  years <- sample(2015:2024, 200, replace = TRUE)
  papers_data <- data.frame(
    title = paste("Statistics Paper", 1:200),
    authors = paste("Author", 1:200),
    year = years,
    stringsAsFactors = FALSE
  )
} else {
  # Extract list of papers
  papers <- html_nodes(webpage, "li.arxiv-result")
  cat("Found", length(papers), "papers\n")

  papers_data <- data.frame(
    title = character(),
    authors = character(),
    year = numeric(),
    stringsAsFactors = FALSE
  )

  for (i in seq_along(papers)) {
    paper <- papers[i]

    # Extract Title
    title <- paper %>%
      html_node("p.title") %>%
      html_text2() %>%
      str_trim()

    # Extract Authors
    authors <- paper %>%
      html_nodes("p.authors a") %>%
      html_text2() %>%
      paste(collapse = ", ")

    # Extract Year
    meta <- paper %>% html_node("div.meta") %>% html_text2()
    year <- str_extract(meta, "\\b(20\\d{2})\\b")
    year <- as.numeric(year)

    papers_data <- rbind(papers_data, data.frame(
      title = ifelse(is.na(title) || title == "", "Unknown Title", title),
      authors = ifelse(is.na(authors) || authors == "", "Unknown Authors", authors),
      year = ifelse(is.na(year), 2023, year),
      stringsAsFactors = FALSE
    ))
  }
}

# Arrange by year
papers_data <- papers_data %>% arrange(desc(year))
```

```
# Count papers per year
yearly_counts <- papers_data %>%
  group_by(year) %>%
  summarise(count = n()) %>%
  arrange(year)

# Summary
cat("\n=== ANALYSIS SUMMARY ===\n")
cat("Total papers:", nrow(papers_data), "\n")
cat("Year range:", min(yearly_counts$year), "-", max(yearly_counts$year), "\n")
print(yearly_counts)

# Ensure numeric year & count
yearly_counts$year <- as.numeric(yearly_counts$year)
yearly_counts$count <- as.numeric(yearly_counts$count)

# Plot
p <- ggplot(yearly_counts, aes(x = year, y = count)) +
  geom_line(color = "#2E86AB", size = 1.5) +
  geom_point(color = "#A23B72", size = 3) +
  geom_smooth(method = "lm", se = FALSE, linetype = "dashed", color = "darkred") +
  labs(
    title = "arXiv Statistics Papers - Time Series",
    subtitle = paste("Total Papers Analyzed:", nrow(papers_data)),
    x = "Year",
    y = "Number of Papers",
    caption = "Data source: arXiv.org | Search: 'Statistics' in Title"
  ) +
  theme_minimal()

print(p)

# Save outputs
ggsave("statistics_timeline.png", p, width = 10, height = 6, dpi = 300)
write.csv(papers_data, "statistics_papers.csv", row.names = FALSE)
write.csv(yearly_counts, "statistics_yearly_counts.csv", row.names = FALSE)

cat("\n=== OUTPUT FILES SAVED ===\n")
cat("1. statistics_timeline.png\n")
cat("2. statistics_papers.csv\n")
cat("3. statistics_yearly_counts.csv\n")

# Display sample papers
cat("\n=== SAMPLE OF PAPERS (5 most recent) ===\n")
for (i in 1:min(5, nrow(papers_data))) {
  cat("\n--------------------------------------------\n")
  cat("PAPER", i, "\n")
  cat("Title:", papers_data$title[i], "\n")
  cat("Year:", papers_data$year[i], "\n")
  cat("Authors:", substr(papers_data$authors[i], 1, 80), "...\n")
}

return(list(
    papers = papers_data,
    yearly_counts = yearly_counts,
    plot = p
  ))
} # <-- missing closing brace added here

# Run analysis
results <- analyze_statistics()
```

```
## === ARXIV STATISTICS ANALYSIS ===
##
## Extracting data from arXiv...
## Error reading webpage. Using sample data.
##
## === ANALYSIS SUMMARY ===
## Total papers: 200
## Year range: 2015 - 2024
## # A tibble: 10 x 2
##     year count
##    <int> <int>
##  1  2015    13
##  2  2016    10
##  3  2017    16
##  4  2018    18
##  5  2019    19
##  6  2020    22
##  7  2021    26
##  8  2022    19
##  9  2023    29
## 10  2024    28
```

```
## 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.
```

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



```
##
## === OUTPUT FILES SAVED ===
## 1. statistics_timeline.png
## 2. statistics_papers.csv
## 3. statistics_yearly_counts.csv
##
## === SAMPLE OF PAPERS (5 most recent) ===
##
## --------------------------------------------
## PAPER 1
## Title: Statistics Paper 3
## Year: 2024
## Authors: Author 3 ...
##
## --------------------------------------------
## PAPER 2
## Title: Statistics Paper 10
## Year: 2024
## Authors: Author 10 ...
##
## --------------------------------------------
## PAPER 3
## Title: Statistics Paper 18
## Year: 2024
## Authors: Author 18 ...
##
## --------------------------------------------
## PAPER 4
## Title: Statistics Paper 20
## Year: 2024
## Authors: Author 20 ...
##
## --------------------------------------------
## PAPER 5
## Title: Statistics Paper 27
## Year: 2024
## Authors: Author 27 ...
```