Tidy Tuesday

# First install packages and load libraries

library('readr')  
library('readr')  
library('ggplot2')  
library("dplyr")  
library("stringr")  
library("scales")

# Read all the papers manually from the source

papers <- readr::read\_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2021/2021-09-28/papers.csv')

## Rows: 29434 Columns: 4

## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (2): paper, title  
## dbl (2): year, month

##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

glimpse(papers)

## Rows: 29,434  
## Columns: 4  
## $ paper <chr> "w0001", "w0002", "w0003", "w0004", "w0005", "w0006", "w0007", "…  
## $ year <dbl> 1973, 1973, 1973, 1973, 1973, 1973, 1973, 1973, 1973, 1973, 1973…  
## $ month <dbl> 6, 6, 6, 7, 7, 7, 8, 9, 9, 9, 9, 10, 10, 10, 10, 11, 11, 11, 12,…  
## $ title <chr> "Education, Information, and Efficiency", "Hospital Utilization:…

paper\_authors <- readr::read\_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2021/2021-09-28/paper\_authors.csv')

## Rows: 67090 Columns: 2

## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (2): paper, author

##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

glimpse(paper\_authors)

## Rows: 67,090  
## Columns: 2  
## $ paper <chr> "w0001", "w0002", "w0003", "w0004", "w0005", "w0006", "w0007", …  
## $ author <chr> "w0001.1", "w0002.1", "w0003.1", "w0004.1", "w0005.1", "w0006.1…

authors <- readr::read\_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2021/2021-09-28/authors.csv')

## Rows: 15437 Columns: 4

## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (4): author, name, user\_nber, user\_repec

##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

glimpse(authors)

## Rows: 15,437  
## Columns: 4  
## $ author <chr> "w0001.1", "w0002.1", "w0003.1", "w0004.1", "w0005.1", "w00…  
## $ name <chr> "Finis Welch", "Barry R Chiswick", "Swarnjit S Arora", "Lee…  
## $ user\_nber <chr> "finis\_welch", "barry\_chiswick", "swarnjit\_arora", NA, "jam…  
## $ user\_repec <chr> NA, "pch425", NA, "pli669", "psm28", NA, NA, NA, "pli259", …

paper\_programs <- readr::read\_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2021/2021-09-28/paper\_programs.csv')

## Rows: 53996 Columns: 2

## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (2): paper, program

##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

glimpse(paper\_programs)

## Rows: 53,996  
## Columns: 2  
## $ paper <chr> "w0074", "w0087", "w0087", "w0107", "w0116", "w0117", "w0129",…  
## $ program <chr> "EFG", "IFM", "ITI", "PE", "PE", "LS", "HE", "IFM", "ITI", "HE…

programs <- readr::read\_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2021/2021-09-28/programs.csv')

## Rows: 21 Columns: 3

## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (3): program, program\_desc, program\_category

##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

glimpse(programs)

## Rows: 21  
## Columns: 3  
## $ program <chr> "AG", "AP", "CF", "CH", "DAE", "DEV", "ED", "EEE", "E…  
## $ program\_desc <chr> "Economics of Aging", "Asset Pricing", "Corporate Fin…  
## $ program\_category <chr> "Micro", "Finance", "Finance", "Micro", "Micro", "Mic…

```

#View each of them and then write each of them as a csv file

write\_csv(x = papers, "Data/papers.csv")  
write\_csv(x = authors, "data/authors.csv")  
write\_csv(x = programs, "data/programs.csv")  
write\_csv(x = paper\_authors, "data/paper\_authors.csv")  
write\_csv(x = paper\_programs, "data/paper\_programs.csv")

#Clean the data and joined them [left joining]

joined <- left\_join(papers, paper\_authors) %>%   
 left\_join(authors) %>%   
 left\_join(paper\_programs) %>%   
 left\_join(programs)%>%   
 mutate(  
 catalogue\_group = str\_sub(paper, 1, 1),  
 catalogue\_group = case\_when(  
 catalogue\_group == "h" ~ "Historical",  
 catalogue\_group == "t" ~ "Technical",  
 catalogue\_group == "w" ~ "General"  
 ),  
 .after = paper  
 )

## Joining, by = "paper"

## Joining, by = "author"

## Joining, by = "paper"

## Joining, by = "program"

View(joined)

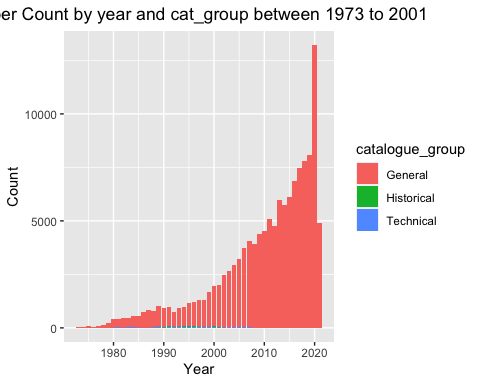
# See how many distinct values are there for each category

sapply(joined, n\_distinct)

## paper catalogue\_group year month   
## 29434 3 49 12   
## title author name user\_nber   
## 29419 15437 15398 14247   
## user\_repec program program\_desc program\_category   
## 5456 22 22 4

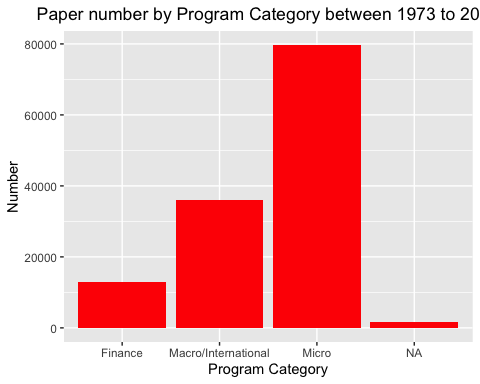
#Let’s see how many papers publishes each year throughout the 1973-2001 time

ggplot(joined, aes(x=year, fill=catalogue\_group)) + geom\_bar() + labs(title= "Paper Count by year and cat\_group between 1973 to 2001") + xlab("Year") + ylab("Count")



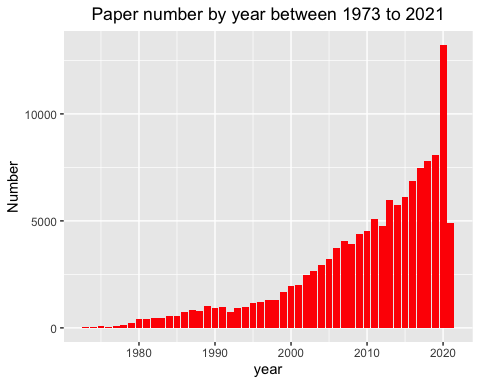
#It’s quite clear that the majority is “general” type. So, let’s do something else. What about finding number of papers according to program category? Lets do it

ggplot(joined, aes(x=program\_category)) +  
 geom\_bar(fill = "red") + labs(title= "Paper number by Program Category between 1973 to 2021") + xlab("Program Category") + ylab("Number")

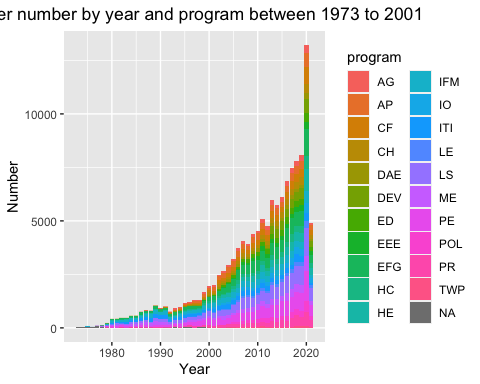


#What about numbe rof papper for each year?

ggplot(joined, aes(x=year)) +  
 geom\_bar(fill = "red") + labs(title= "Paper number by year between 1973 to 2021") + xlab("year") + ylab("Number")

 #What about number of paper for each year categorized by program name also?

ggplot(joined, aes(x=year, fill=program)) + geom\_bar() + labs(title= "Paper number by year and program between 1973 to 2001") + xlab("Year") + ylab("Number")



#Most number of papers were in 2000. Let’s look deeper in that

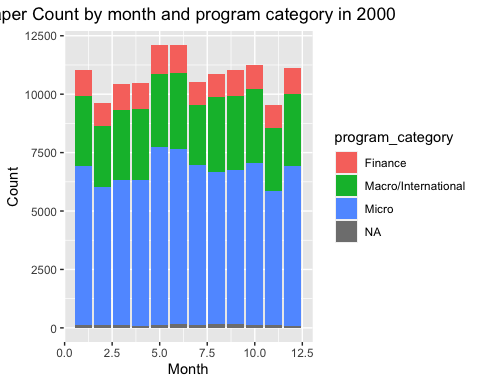
data\_2000 <- joined %>% filter(year ==2000)

#View the 2000 data

glimpse(data\_2000)

## Rows: 1,962  
## Columns: 12  
## $ paper <chr> "w6130", "w6130", "w6509", "w6509", "w6518", "w6518",…  
## $ catalogue\_group <chr> "General", "General", "General", "General", "General"…  
## $ year <dbl> 2000, 2000, 2000, 2000, 2000, 2000, 2000, 2000, 2000,…  
## $ month <dbl> 3, 3, 5, 5, 9, 9, 3, 3, 6, 6, 5, 5, 5, 5, 6, 6, 6, 6,…  
## $ title <chr> "Nonparametric Risk Management and Implied Risk Avers…  
## $ author <chr> "w5345.1", "t0059.1", "w2161.1", "t0097.1", "w0505.1"…  
## $ name <chr> "Yacine Ait-Sahalia", "Andrew W Lo", "Lucian Arye Beb…  
## $ user\_nber <chr> "yacine\_ait-sahalia", "andrew\_lo", "lucian\_bebchuk", …  
## $ user\_repec <chr> "pai23", "plo171", "pbe72", NA, "pmi37", "pes29", "pb…  
## $ program <chr> "AP", "AP", "LE", "LE", "ME", "ME", "CF", "CF", "IFM"…  
## $ program\_desc <chr> "Asset Pricing", "Asset Pricing", "Law and Economics"…  
## $ program\_category <chr> "Finance", "Finance", "Micro", "Micro", "Macro/Intern…

#Let's see how many papers from each program got published in each month of 2000?  
ggplot(joined, aes(x=month, fill=program\_category)) + geom\_bar() + labs(title= "Paper Count by month and program category in 2000") + xlab("Month") + ylab("Count")



#We got a problem here. The x axis (month number is not integer. Let's fix it.  
  
ggp <- ggplot(joined, aes(x=month, fill=program\_category)) + geom\_bar() + labs(title= "Paper Count by month and program category in 2000") + xlab("Month") + ylab("Count")  
  
#fixed  
ggp + scale\_x\_continuous(breaks = pretty\_breaks())

