// Lab Exercise 4 Cleaning

```
library(readr)
```

```
## Warning: package ’readr’ was built under R version 4.3.3
```

```
library(stringr)
```

```
## Warning: package ’stringr’ was built under R version 4.3.2
```

```
library(dplyr)
```

```
## Warning: package ’dplyr’ was built under R version 4.3.3

##
## 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
```

```
# Load Arxiv Scraped Dataset
arxiv <- read_csv("ArxivPapers_Philosophy.csv")
```

```
## New names:
## * ‘‘ -> ‘...1‘

## Rows: 150 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): title, author, subject, abstract, meta
## dbl (1): ...1
##
## 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.
```

```
# Extracting the date from the meta column
arxiv_date_only <- str_extract(arxiv$meta, "\\d+\\s[A-Za-z]+\\s\\d+")

# Changing to date type
arxiv_date_type <- as.Date(arxiv_date_only, format = "%d %b %Y")
head(arxiv_date_type)
```

```
## [1] "2024-03-14" "2024-03-14" "2024-03-13" "2024-03-12" "2024-03-12"
## [6] "2024-03-12"
```

```
# Removing meta and number column and appending the new date column
# Mutating all while converting other columns to lowercase, removing parenthesis text in the subject co
cleaned_arxivpapers <- arxiv %>%
  mutate(date = arxiv_date_type,
         subject = gsub("\\s\\(.*\\)", "", subject),
         across(where(is.character), tolower)) %>%
  select(-meta, -...1)

# Writing to CSV
write.csv(cleaned_arxivpapers, file = "cleaned_ArxivPapers_Philosophy.csv", row.names = FALSE)
```

// Lab Exercise 5 Cleaning

```
library(readr)
library(stringr)
library(dplyr)

# Load movie Scraped Dataset
MovieReviews <- read_csv("AllMovies (1).csv")
```

```
## New names:
## Rows: 2500 Columns: 7
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (6): Movie Title, Review Title, Reviewer, Review, Date, Ratings dbl (1): ...1
## 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.
## * ‘‘ -> ‘...1‘
```

```
# Extracting the date from the meta column and changing to date type
reviews_date_type <- as.Date(str_extract(MovieReviews$Date, "\\d+\\s[A-Za-z]+\\s\\d+"), format = "%d %b

# Extracting the rating from the rating column and changing to integer
reviews_ratings_integer <- as.integer(str_extract(MovieReviews$Ratings, "\\d+\\.\\d+"))

# Removing all emoticons from the columns
MovieReviews$MovieTitle <- gsub("\\p{So}", "", MovieReviews$MovieTitle, perl = FALSE)
```

```
## Warning: Unknown or uninitialised column: ‘MovieTitle‘.
```

```
## Warning in gsub("\\p{So}", "", MovieReviews$MovieTitle, perl = FALSE): TRE
## pattern compilation error ’Invalid contents of {}’
```

```
## Error in gsub("\\p{So}", "", MovieReviews$MovieTitle, perl = FALSE): invalid regular expression ’\p{S
```

```
MovieReviews$Reviewer <- gsub("\\p{So}", "", MovieReviews$Reviewer, perl = TRUE)
MovieReviews$Review <- gsub("\\p{So}", "", MovieReviews$Review, perl = TRUE)

# Removing non-alphabetical languages from the columns
MovieReviews$MovieTitle <- gsub("[ˆa-zA-Z ]", "", MovieReviews$MovieTitle)
```

```
## Warning: Unknown or uninitialised column: ‘MovieTitle‘.

## Error in ‘$<-‘:
## ! Assigned data ‘gsub("[^a-zA-Z ]", "", MovieReviews$MovieTitle)‘ must
##   be compatible with existing data.
## x Existing data has 2500 rows.
## x Assigned data has 0 rows.
## i Only vectors of size 1 are recycled.
## Caused by error in ‘vectbl_recycle_rhs_rows()‘:
## ! Can’t recycle input of size 0 to size 2500.
```

```
MovieReviews$Reviewer <- gsub("[^a-zA-Z ]", "", MovieReviews$Reviewer)
MovieReviews$Review <- gsub("[^a-zA-Z ]", "", MovieReviews$Review)

# Replace all blank strings with NA
MovieReviews$MovieTitle <- na_if(MovieReviews$MovieTitle, "")
```

```
## Warning: Unknown or uninitialised column: ‘MovieTitle‘.

## Error in ‘vec_init()‘:
## ! ‘x‘ must be a vector, not ‘NULL‘.
```

```
MovieReviews$Reviewer <- na_if(MovieReviews$Reviewer, "")
MovieReviews$Review <- na_if(MovieReviews$Review, "")

# Converting all columns to lowercase
MovieReviews <- MovieReviews %>%
  mutate(across(where(is.character), tolower))

# Combine all together
cleaned_reviews <- MovieReviews %>%
  mutate(date = reviews_date_type, ratings = reviews_ratings_integer)

# Writing to CSV
write.csv(cleaned_reviews, file = "cleaned_MovieReviews.csv", row.names = FALSE)

View(MovieReviews)
```