Comparing Datasets

Sofia Bliss-Carrascosa

9/12/2022

```
library(tidyverse)
library(knitr)
```

V2: TESTING THE R-VERSION COMPARED TO ORIGINAL VERSION Loading R-version

```
v0_all_us_unsorted <- read_csv("allUS_unsorted.csv")
```

```
v1_all_us_polclaims <- read_csv("all_us_pol_claims_updated_sep15.csv")
```

```
v2_5parties <- v1_all_us_polclaims %>%
  filter(claimant_party == "Republican"|
           claimant_party == "Democratic"|
         claimant_party == "Independent"|
         claimant_party == "Libertarian"|
         claimant_party == "unknown_affiliation")

v2_5parties %>%
  count(publisher.site) %>%
  kable(caption = "All Claims in 5 Parties
        Sorted by Publisher")
```

Table 1: All Claims in 5 Parties Sorted by Publisher

publisher.site	n
cbsnews.com	167
checkyourfact.com	10
factcheck.org	1290
factcheck.thedispatch.com	58
newsweek.com	66
nytimes.com	464
politifact.com	4275
polygraph.info	3
poynter.org	10
thegazette.com	7
usatoday.com	19
vox.com	2
washingtonpost.com	1252

```
dim(v2_5parties)
```

```
## [1] 7623   13
```

Loading Grace Original Version

```
grace_v2 <- read_csv("grace_v2_CSVversion.csv")

grace_v2 %>%
  count(publisher.site) %>%
  kable(caption = "All Claims in 5 Parties
        Sorted by Publisher")
```

Table 2: All Claims in 5 Parties Sorted by Publisher

publisher.site	n
cbsnews.com	167
checkyourfact.com	10
factcheck.org	1290
factcheck.thedispatch.com	58
newsweek.com	66
nytimes.com	464
politifact.com	4275
polygraph.info	3
poynter.org	10
thegazette.com	7
usatoday.com	19
vox.com	2
washingtonpost.com	1252

```
dim(grace_v2)
```

```
## [1] 7623   13
```

Data counts work out!

```
anti_join(v2_5parties, grace_v2)
```

```
## # A tibble: 0 x 13
## # ... with 13 variables: ...1 <dbl>, url <chr>, title <chr>,
## #   textualRating <chr>, languageCode <chr>, publisher.name <chr>,
## #   publisher.site <chr>, reviewDate <dttm>, text <chr>, claimant <chr>,
## #   claimDate <dttm>, claimant_party <chr>, reason <chr>
```

ALL MATCH!! WOOT!!

V3: TESTING THE R-VERSION COMPARED TO ORIGINAL VERSION Loading R-version

```
v3_deduped <- v2_5parties %>%
  distinct(url, text, .keep_all = TRUE)

v3_deduped %>%
  count(publisher.site) %>%
  kable(caption = "All Claims in 5 Parties
        Sorted by Publisher")
```

Table 3: All Claims in 5 Parties Sorted by Publisher

publisher.site	n
cbsnews.com	157
checkyourfact.com	10
factcheck.org	1237
factcheck.thedispatch.com	58
newsweek.com	63
nytimes.com	444
politifact.com	4176
polygraph.info	2
poynter.org	9
thegazette.com	7
usatoday.com	17
vox.com	1
washingtonpost.com	1179

```
dim(v3_deduped)
```

```
## [1] 7360   13
```

Loading Grace version

```
grace_v3 <- read_csv("grace_v3_deduped_CSVversion.csv")

grace_v3 %>%
  count(publisher.site) %>%
  kable(caption = "All Claims in 5 Parties
        Sorted by Publisher")
```

Table 4: All Claims in 5 Parties Sorted by Publisher

publisher.site	n
cbsnews.com	157
checkyourfact.com	10
factcheck.org	1237
factcheck.thedispatch.com	58
newsweek.com	63
nytimes.com	444
politifact.com	4176
polygraph.info	2

publisher.site	n
poynter.org	9
thegazette.com	7
usatoday.com	17
vox.com	1
washingtonpost.com	1179

```
dim(grace_v3)
```

```
## [1] 7360   13
```

Data Counts are not the same: why?

```
anti_join(v3_deduped, grace_v3)
```

```
## # A tibble: 0 x 13
## # ... with 13 variables: ...1 <dbl>, url <chr>, title <chr>,
## #   textualRating <chr>, languageCode <chr>, publisher.name <chr>,
## #   publisher.site <chr>, reviewDate <dttm>, text <chr>, claimant <chr>,
## #   claimDate <dttm>, claimant_party <chr>, reason <chr>
```

```
identified_dupes <- anti_join(v2_5parties, v3_deduped)

dim(identified_dupes)
```

```
## [1] 263  13
```

LOADING V4 – Manual removal by ASA

```
grace_v4 <- read_csv("grace_v4_dejunkedCSVversion.csv")
```

```
manual_removes <- anti_join(v3_deduped, grace_v4)
```