

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("allUS_polclaimsV1_sheets.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("gracesoriginaldata.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

<u>publisher.site</u>	<u>n</u>
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!

```
v2_5parties_2 <- v2_5parties %>%
  filter(publisher.site == "factcheck.org"|
         publisher.site == "politifact.com"|
         publisher.site == "nytimes.com"|
         publisher.site == "washingtonpost.com")

v2_5parties_2 %>%
  group_by(publisher.site) %>%
  count()
```

```
## # A tibble: 4 x 2
## # Groups:   publisher.site [4]
##   publisher.site      n
```

```
##   <chr>           <int>
## 1 factcheck.org    1290
## 2 nytimes.com      464
## 3 politifact.com   4275
## 4 washingtonpost.com 1252
```

```
gracev2_2 <- grace_v2 %>%
  filter(publisher.site == "factcheck.org"|
         publisher.site == "politifact.com"|
         publisher.site == "nytimes.com"|
         publisher.site == "washingtonpost.com")

gracev2_2 %>%
  group_by(publisher.site) %>%
  count()
```

```
## # A tibble: 4 x 2
## # Groups:   publisher.site [4]
##   publisher.site      n
##   <chr>           <int>
## 1 factcheck.org    1290
## 2 nytimes.com      464
## 3 politifact.com   4275
## 4 washingtonpost.com 1252
```

```
anti_join(gracev2_2, v2_5parties_2, by = "claimant") %>%
  print()
```

```
## # A tibble: 14 x 13
##   ...1 url      title textualRating languageCode publisher.name publisher.site
##   <dbl> <chr>   <chr>   <chr>      <chr>      <chr>      <chr>
## 1 16719 https:~ FactC~ FALSE      en        FactCheck.org factcheck.org
## 2 17036 https:~ Polit~ We Explain t~ en        FactCheck.org factcheck.org
## 3 17984 https:~ Not '~ Trump Admin.~ en        FactCheck.org factcheck.org
## 4 5007 http:/~ Does ~ Mostly True  en        PolitiFact    politifact.com
## 5 1506 http:/~ Juliá~ TRUE        en        PolitiFact    politifact.com
## 6 6319 http:/~ Mary ~ Mostly True  en        PolitiFact    politifact.com
## 7 6479 http:/~ Do Ca~ Mostly True  en        PolitiFact    politifact.com
## 8 5818 http:/~ Juliá~ Mostly True  en        PolitiFact    politifact.com
## 9 7176 https:~ Yes; ~ Half True    en        PolitiFact    politifact.com
## 10 6917 https:~ When ~ Mostly True  en        PolitiFact    politifact.com
## 11 1584 https:~ Fact~ Metering aff~ en        PolitiFact    politifact.com
## 12 2527 https:~ O'Rou~ Mostly True  en        PolitiFact    politifact.com
## 13 893 https:~ Compa~ Mostly False  en        PolitiFact    politifact.com
## 14 14705 https:~ Analy~ Not the whol~ en        The Washingto~ washingtonpos~
## # ... with 6 more variables: reviewDate <dtm>, text <chr>, claimant <chr>,
## #   claimDate <dtm>, claimant_party <chr>, reason <chr>
```

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

```
v3_deduped <- v2_5parties %>%
  select(url:claimant_party) %>%
```

```
distinct()

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

<u>publisher.site</u>	<u>n</u>
cbsnews.com	157
checkyourfact.com	10
factcheck.org	1238
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	1180

```
dim(v3_deduped)
```

```
## [1] 7362  11
```

Loading Grace version

```
grace_v3 <- read_csv("gracededupedata.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

<u>publisher.site</u>	<u>n</u>
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

<hr/>	
publisher.site	n
thegazette.com	7
usatoday.com	17
vox.com	1
washingtonpost.com	1179
<hr/>	

```
dim(grace_v3)
```

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

Data Counts are not the same: why?