# Thesis Data Analysis: The Psychology of Scientific Fraud

Benjamin Zubaly

2024-03-04

## Table of contents



## Introduction

This document is intended to track the data analysis for my undergraduate thesis project on the psychology of scientific fraud. I have already planned out (and preregistered) my data analysis plan (see here), and I will note throughout the document if I deviate from this plan and why. This project is also being tracked in a private GitHub repository in case I need to revert back to a previous version of the project due to a fatal error.

## Overview of Sections

- **Data Cleaning:** The data cleaning section will take the finalized dataset that I have included in the project directory and use it to create the remaining variables that we need for our analysis. We will also save this dataset.
- **Data Exploration:** In the data exploration section, we will check for and deal with missing data, run descriptive statistics of our outcome variables, visualize our data, and run bivariate correlations.
- **Testing Hypotheses:** In the hypothesis testing section, we will test each of our hypotheses one-by-one, according to our analysis plan.

## Directory Set-Up

In order to ensure that this analysis is computationally reproducible, I have included everything that is needed to complete this analysis (just the finalized data file, "study_dataset.csv") in the current working directory. The code will be displayed before the output of each analysis.

## Variable Definitions

Although the following variables may not exist until after the data cleaning section, here is what each variable name refers to, so that you can refer to them while examining the code.

`DOI`: Unique identifier for each paper (i.e., the paper DOI).

`PaperType`: Categorical variable indicating SAFP, SAGP, MAFP, or MAGP.

`LingObf`: Continuous variable for linguistic obfuscation.

`CertSent`: Continuous variable for certainty sentiment.

`Refs`: Count variable for references.

`FraudCorrAuth`: Dichotomous variable indicating if the fraudulent author is the corresponding author (1) or is not (0). Unknown cases will be marked in a seperate variable with this variable left blank.

`NumAuth`: Count variable indicating the number of authors for each paper.

`abstraction`: Abstraction index composed of the sum of standardized scores for `article`, `prep`, and `quantity`.

`article`: Articles from LIWC.

`prep`: Prepositions from LIWC.

`quantity`: Quantities from LIWC.

`cause`: Causation terms from LIWC.

`jargon`: The percent of words not captured by LIWC (100-`Dic`).

`Dic`: The percentage of words captured by all LIWC dictionaries.

`emo_pos`: Positive emotion terms from LIWC.

`flesch_re`: Flesch Reading Ease from ARTE.

### Initial Package Installations

If you would like to reproduce this analysis, here are the package I will be using, so that they can be cued before starting.

```
  install.packages("readr")       # For reading data
```

```
Installing package into '/Users/benjaminzubaly/Library/R/x86_64/4.3/library'
(as 'lib' is unspecified)
```

```
The downloaded binary packages are in
    /var/folders/38/1ybnplc53zdb089bn6drqfn00000gn/T//Rtmp94Cdsa/downloaded_packages
```

```
  install.packages("dplyr")       # For data manipulation and handling of missing data
```

```
Installing package into '/Users/benjaminzubaly/Library/R/x86_64/4.3/library'
(as 'lib' is unspecified)
```

```
The downloaded binary packages are in
    /var/folders/38/1ybnplc53zdb089bn6drqfn00000gn/T//Rtmp94Cdsa/downloaded_packages
```

```
  install.packages("psych")       # For descriptive statistics, correlations
```

```
Installing package into '/Users/benjaminzubaly/Library/R/x86_64/4.3/library'
(as 'lib' is unspecified)
```

```
The downloaded binary packages are in
    /var/folders/38/1ybnplc53zdb089bn6drqfn00000gn/T//Rtmp94Cdsa/downloaded_packages
```

```r
install.packages("ggplot2")     # For data visualization
```

```
Installing package into '/Users/benjaminzubaly/Library/R/x86_64/4.3/library'
(as 'lib' is unspecified)
```

```
The downloaded binary packages are in
    /var/folders/38/1ybnplc53zdb089bn6drqfn00000gn/T//Rtmp94Cdsa/downloaded_packages
```

```r
install.packages("car")         # For diagnostic tests such as Levene's test
```

```
Installing package into '/Users/benjaminzubaly/Library/R/x86_64/4.3/library'
(as 'lib' is unspecified)
```

```
The downloaded binary packages are in
    /var/folders/38/1ybnplc53zdb089bn6drqfn00000gn/T//Rtmp94Cdsa/downloaded_packages
```

```r
install.packages("rcompanion")  # For Games-Howell post-hoc test (if applicable)
```

```
Installing package into '/Users/benjaminzubaly/Library/R/x86_64/4.3/library'
(as 'lib' is unspecified)
```

```
The downloaded binary packages are in
    /var/folders/38/1ybnplc53zdb089bn6drqfn00000gn/T//Rtmp94Cdsa/downloaded_packages
```

```r
install.packages("dunn.test")   # For Dunn post-hoc test (if applicable)
```

```
Installing package into '/Users/benjaminzubaly/Library/R/x86_64/4.3/library'
(as 'lib' is unspecified)
```

```
The downloaded binary packages are in
    /var/folders/38/1ybnplc53zdb089bn6drqfn00000gn/T//Rtmp94Cdsa/downloaded_packages
```

### Initial Import of Data

We will not load in the dataset "study_dataset.csv" from the working directory.

```r
library(readr) # Loading the readr package

data <- read_csv("study_dataset.csv") # Loading in study dataset as "data"
```

```
Rows: 88 Columns: 207
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (26): DOI, Title, Subject, Institution, Journal, Publisher, Country, Au...
dbl (181): Record ID, RetractionPubMedID, OriginalPaperDate, year, OriginalP...

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.
```

I have viewed the data frame, and the data seems to have loaded correctly.

## Data Cleaning

To conduct the analysis, we will first need to calculate the `LingObf` variable by calculating the `abstraction` index and `jargon` words; creating standardized scores for `abstraction`, `cause`, `jargon`, `emo_pos`, and `flesch_re`; and calculating the `LingObf` composite variable from these standardized scores.

1. First, we will calculate the `abstraction` index by creating standardized scores for `article`, `prep`, and `quantity` and summing them.

```r
# Calculate standardized scores for article, prep, and quantity and add them to the datase
data$articles_standardized <- scale(data$article)
data$prep_standardized <- scale(data$prep)
data$quantity_standardized <- scale(data$quantity)

# Create the new variable 'abstraction' as the sum of the three standardized variables
data$abstraction <- (data$articles_standardized + data$prep_standardized + data$quantity_s
```

- After viewing the data, the transformations and variable calculation seem to have occurred appropriately.

2. Next, we will calculate the `jargon` words by subtracting `Dic` from 100.

```r
# Calculate the new variable 'jargon' by subtracting 'Dic' from 100
data$jargon <- (100 - data$Dic)
```

- After viewing the data, the variable calculation seem to have occurred appropriately.

3. Next, we will create standardized scores for each subcomponent of the `LingObf`.

```
# Standardize the new set of variables and add them to the dataset
data$abstraction_standardized <- scale(data$abstraction)
data$cause_standardized <- scale(data$cause)
data$jargon_standardized <- scale(data$jargon)
data$emo_pos_standardized <- scale(data$emo_pos)
data$flesch_re_standardized <- scale(data$flesch_re)
```

- After viewing the data, the variable transformations seem to have occurred appropriately.

4. Now we will calculate the `LingObf` variable using the following formula: [cause_standardized + abstraction_standardized + jargon_standardized] – [emo_pos_standardized + flesch_re_standardized].

```
# Calculate 'LingObf'
data$LingObf <- (data$cause_standardized + data$abstraction_standardized + data$jargon_sta
```

- After viewing the data, the variable calculation seem to have occurred appropriately.

5. Lastly, our variable that indicates certainty sentiment is currently `certainty_avg`, but to make things easier I am going to copy this data into a new variables called `CertSent`.

```
data$CertSent <- data$certainty_avg
```

To ensure that our clean data is saved, we will write the dataset to the current working directory.

```
# Writing our data as a csv file in the current working directory
write.csv(data, "clean_study_data.csv")
```

- I have opened the saved data file outside of Rstudio, and it seems to have been written correctly.

## Data Exploration

### Dealing with Missing Data

1. Data will be first inspected for missing scores.

```r
library(dplyr) # Loading the dplyr package for data manipulation and handling missing valu
```

```
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
# To summarize the number of missing values in each column
missing_data_summary <- sapply(data, function(x) sum(is.na(x)))

print(missing_data_summary) # To see summary of missing values for all columns
```

```
                DOI                 Record ID
                  0                        44
              Title                   Subject
                  0                        44
        Institution                   Journal
                  0                         0
          Publisher                   Country
                  0                         0
             Author                      URLS
                  0                        68
        ArticleType            RetractionDate
                  0                        44
      RetractionDOI         RetractionPubMedID
                 48                        51
  OriginalPaperDate                      year
                 44                         0
   OriginalPaperDOI       OriginalPaperPubMedID
                  0                        51
   RetractionNature                    Reason
                 44                        44
         Paywalled                     Notes
                 44                        70
```

| | |
|---|---|
| simple_reason | NumAuth |
| 44 | 0 |
| inst_pres | PaperType |
| 0 | 0 |
| gender | matched_DOI_SAFP_MAFP |
| 0 | 44 |
| matched_DOI_SAFP_SAGP | matched_DOI_MAFP_MAGP |
| 44 | 44 |
| matching_concessions | FraudCorrAuth |
| 42 | 80 |
| FCA_unknown | FCA_notes |
| 66 | 66 |
| different_country | year_difference |
| 22 | 22 |
| different_gender | different_inst_pres |
| 22 | 22 |
| different_journal | Refs |
| 66 | 0 |
| Abstract | flesch_re |
| 0 | 0 |
| wordcount | valence_min |
| 0 | 0 |
| valence_max | valence_avg |
| 0 | 0 |
| valence_std | extremity_min |
| 0 | 0 |
| extremity_max | extremity_avg |
| 0 | 0 |
| extremity_std | extremity_min_pos |
| 0 | 0 |
| extremity_max_pos | extremity_avg_pos |
| 0 | 0 |
| extremity_std_pos | extremity_min_neg |
| 0 | 1 |
| extremity_max_neg | extremity_avg_neg |
| 1 | 1 |
| extremity_std_neg | extremity_PosMinNeg |
| 5 | 0 |
| emotionality_min | emotionality_max |
| 0 | 0 |
| emotionality_avg | emotionality_std |
| 0 | 0 |
| emotionality_min_pos | emotionality_max_pos |

```
                        0                         0
      emotionality_avg_pos       emotionality_std_pos
                        0                         0
      emotionality_min_neg       emotionality_max_neg
                        1                         1
      emotionality_avg_neg       emotionality_std_neg
                        1                         5
    emotionality_PosMinNeg count_unique_evaluative_pos
                        0                         0
count_total_evaluative_pos count_unique_evaluative_neg
                        0                         0
count_total_evaluative_neg     count_unique_evaluative
                        0                         0
    count_total_evaluative  count_evaluative_PosMinNeg
                        0                         0
                ambivalent             pos_dichotomous
                        0                        87
             certainty_min               certainty_max
                        0                         0
             certainty_avg               certainty_std
                        0                         0
    count_unique_certainty       count_total_certainty
                        0                         0
                   Segment                        WC
                        0                         0
                  Analytic                     Clout
                        0                         0
                 Authentic                     Tone
                        0                         0
                       WPS                   BigWords
                        0                         0
                       Dic                  Linguistic
                        0                         0
                  function                    pronoun
                        0                         0
                     ppron                         i
                        0                         0
                        we                       you
                        0                         0
                     shehe                      they
                        0                         0
                     ipron                       det
                        0                         0
```

| | |
|---|---|
| article | number |
| 0 | 0 |
| prep | auxverb |
| 0 | 0 |
| adverb | conj |
| 0 | 0 |
| negate | verb |
| 0 | 0 |
| adj | quantity |
| 0 | 0 |
| Drives | affiliation |
| 0 | 0 |
| achieve | power |
| 0 | 0 |
| Cognition | allnone |
| 0 | 0 |
| cogproc | insight |
| 0 | 0 |
| cause | discrep |
| 0 | 0 |
| tentat | certitude |
| 0 | 0 |
| differ | memory |
| 0 | 0 |
| Affect | tone_pos |
| 0 | 0 |
| tone_neg | emotion |
| 0 | 0 |
| emo_pos | emo_neg |
| 0 | 0 |
| emo_anx | emo_anger |
| 0 | 0 |
| emo_sad | swear |
| 0 | 0 |
| Social | socbehav |
| 0 | 0 |
| prosocial | polite |
| 0 | 0 |
| conflict | moral |
| 0 | 0 |
| comm | socrefs |
| 0 | 0 |
| family | friend |

| | |
|---|---|
| 0 | 0 |
| female | male |
| 0 | 0 |
| Culture | politic |
| 0 | 0 |
| ethnicity | tech |
| 0 | 0 |
| Lifestyle | leisure |
| 0 | 0 |
| home | work |
| 0 | 0 |
| money | relig |
| 0 | 0 |
| Physical | health |
| 0 | 0 |
| illness | wellness |
| 0 | 0 |
| mental | substances |
| 0 | 0 |
| sexual | food |
| 0 | 0 |
| death | need |
| 0 | 0 |
| want | acquire |
| 0 | 0 |
| lack | fulfill |
| 0 | 0 |
| fatigue | reward |
| 0 | 0 |
| risk | curiosity |
| 0 | 0 |
| allure | Perception |
| 0 | 0 |
| attention | motion |
| 0 | 0 |
| space | visual |
| 0 | 0 |
| auditory | feeling |
| 0 | 0 |
| time | focuspast |
| 0 | 0 |
| focuspresent | focusfuture |
| 0 | 0 |

```
          Conversation                netspeak
                     0                       0
                assent                  nonflu
                     0                       0
                filler                 AllPunc
                     0                       0
                Period                   Comma
                     0                       0
                 QMark                  Exclam
                     0                       0
               Apostro                  OtherP
                     0                       0
                 Emoji articles_standardized
                     0                       0
     prep_standardized quantity_standardized
                     0                       0
           abstraction                  jargon
                     0                       0
abstraction_standardized    cause_standardized
                     0                       0
    jargon_standardized     emo_pos_standardized
                     0                       0
  flesch_re_standardized                  LingObf
                     0                       0
              CertSent
                     0
```

- Taking a look at our outcome variables, there are no missing scores, so we do not need to impute any values.

## Descriptive Statistics

1. I will now generate descriptive statistics for each relevant variable in the dataset. I originally (in the preregistered plan) was simply going to deploy the `describe()` function on the entire dataset, but because I retained all of the columns from the Retraction Watch Database (*Retraction Watch Database*, 2023) and all of the output from the text analysis packages (Aggarwal, 2022; Boyd et al., 2022; Rocklage et al., 2023) there are currently 219 variables. Because this would be unmanageable, I am going to only calculate descriptive statistics for a selection of variables of interest. I will first create a dataframe with only the continuous variables that I am interested in generating descriptive statistics for, and I will use the `psych` package to produce the descriptive statistics for these variables.

```r
library(psych) # Loading the psych package

# Selecting the continuous variables I am interested in getting descriptive statistics for
continuous_data_for_descriptives <- data[c("year", "Refs", "flesch_re", "WC", "abstraction

# Generating descriptive statistics for variables of interest
continuous_descriptive_stats_all <- describe(continuous_data_for_descriptives)

# Displaying the results of the descriptive stats for the variables listed above
continuous_descriptive_stats_all
```

```
             vars  n    mean      sd  median trimmed     mad     min      max
year            1 88 2009.70   10.47 2013.00 2011.14    7.41 1980.00  2022.00
Refs            2 88   45.82   23.47   39.00   43.21   20.02   12.00   146.00
flesch_re       3 88   34.90   10.94   36.45   35.46   10.28  -10.89    56.59
WC              4 88 4436.25 2248.24 3951.50 4238.17 1956.29 1124.00 13901.00
abstraction     5 88    0.00    1.83    0.09    0.04    1.86   -4.46     3.55
jargon          6 88   33.33    7.47   34.76   33.50    7.84   17.91    46.67
CertSent        7 88    6.17    0.25    6.18    6.18    0.22    5.21     6.76
LingObf         8 88    0.00    1.67   -0.14   -0.01    1.54   -5.84     4.28
cause           9 88    2.80    1.22    2.62    2.69    1.02    0.89     6.01
emo_pos        10 88    0.09    0.22    0.04    0.05    0.06    0.00     1.67
article        11 88    8.38    2.20    8.31    8.33    2.13    3.53    14.47
prep           12 88   15.43    1.68   15.66   15.51    1.34    9.76    20.67
quantity       13 88    4.23    1.32    4.08    4.14    1.33    1.69     8.66
PaperType*     14 88    2.50    1.12    2.50    2.50    1.48    1.00     4.00
               range  skew kurtosis     se
year           42.00 -1.25     0.74   1.12
Refs          134.00  1.41     2.87   2.50
flesch_re      67.48 -0.85     2.16   1.17
WC          12777.00  1.21     2.18 239.66
abstraction     8.01 -0.20    -0.42   0.20
jargon         28.76 -0.27    -0.94   0.80
CertSent        1.55 -0.69     1.42   0.03
LingObf        10.12 -0.21     1.40   0.18
cause           5.12  0.80     0.14   0.13
emo_pos         1.67  5.66    34.22   0.02
article        10.94  0.33    -0.19   0.23
prep           10.91 -0.44     1.22   0.18
quantity        6.97  0.67     0.42   0.14
PaperType*      3.00  0.00    -1.40   0.12
```

- I won't make too many comments here, because for most of these measures there are no formal or informal norms against which to judge them. That being said:
  - **Year:** The mean year is mid-2009, with a standard deviation of 10 years (max 2022 and min 1980), indicating that the sample is recent enough to be relevant but also spans quite a number of years. This I think is good insofar as the recent history of academic publishing is represented more fully (some recent investigations limited their search to the three years prior to publication). There is some negative skew, which I suspect is due to the 1980 paper being quite a bit older than most papers.
  - **Refs:** The mean number of references is 45.82 with a standard deviation of 23.47. At least intuitively, this seems like a pretty standard distribution of references if we were to randomly select papers from the literature. However, the range is huge, with one paper showing 146 references—perhaps why the skew is positive. This may be something to consider when making group comparisons, as this point may have significant leverage.
  - **WC:** The average word count is 4436.25, and the standard deviation is 2248.24. It is also positively skewed, and although this should not affect our linguistic dependent variables it may be a confounding factor for references. Indeed, taking a look at the dataset I can see that the second highest value for word count (9670) also has the maximum value for references (the outlier noted above at 146 references). This will be something to watch out for.
- Now we will create descriptive statistics for each of the continuous variables above within the `PaperType` groups.

```
# Making PaperType a factor variable in the continuous variable dataframe to allow for gro
continuous_data_for_descriptives$PaperType <- factor(continuous_data_for_descriptives$Pape

# Generating descriptive statistics within PaperType groups
descriptive_stats_by_PaperType <- describeBy(continuous_data_for_descriptives, group = con

descriptive_stats_by_PaperType
```

```
 Descriptive statistics by group
group: Single-Authored Fraudulent Papers
            vars  n    mean      sd  median trimmed     mad     min     max
year           1 22 2008.95   11.74 2013.00 2010.39    6.67 1983.00 2022.00
Refs           2 22   44.18   19.93   36.00   42.89   17.79   18.00   87.00
flesch_re      3 22   36.09    9.17   37.06   36.29   10.53   18.49   54.36
WC             4 22 3998.91 1741.31 3796.50 3911.39 1829.53 1273.00 7330.00
abstraction    5 22    0.21    1.54    0.04    0.27    1.25   -3.27    2.95
jargon         6 22   34.23    8.33   36.94   34.90    8.49   17.91   44.35
CertSent       7 22    6.21    0.21    6.18    6.20    0.18    5.87    6.66
```

```
LingObf        8 22   -0.24    1.79   -0.04   -0.05    1.97   -4.23    2.25
cause          9 22    2.71    1.48    2.50    2.56    1.07    0.89    6.01
emo_pos       10 22    0.15    0.35    0.06    0.08    0.06    0.00    1.67
article       11 22    8.35    2.60    8.48    8.41    2.90    3.53   12.25
prep          12 22   15.75    1.60   15.75   15.79    1.43   11.79   18.42
quantity      13 22    4.27    0.97    4.11    4.21    1.13    2.59    6.59
PaperType     14 22    1.00    0.00    1.00    1.00    0.00    1.00    1.00
              range  skew kurtosis     se
year          39.00 -1.02    -0.29   2.50
Refs          69.00  0.56    -1.06   4.25
flesch_re     35.87 -0.01    -0.63   1.95
WC          6057.00  0.54    -0.67 371.25
abstraction    6.21 -0.23    -0.33   0.33
jargon        26.44 -0.56    -1.04   1.78
CertSent       0.78  0.44    -0.57   0.04
LingObf        6.48 -0.70    -0.32   0.38
cause          5.12  0.93    -0.05   0.32
emo_pos        1.67  3.81    13.82   0.07
article        8.72 -0.16    -1.18   0.55
prep           6.63 -0.34    -0.15   0.34
quantity       4.00  0.49    -0.41   0.21
PaperType      0.00   NaN      NaN   0.00
------------------------------------------------------------
group: Multi-Authored Fraudulent Papers
            vars  n    mean      sd  median trimmed     mad     min     max
year           1 22 2010.14    9.59 2012.50 2011.67    8.15 1982.00 2022.00
Refs           2 22   44.00   19.66   39.50   42.06   17.05   18.00   94.00
flesch_re      3 22   34.18    9.64   31.41   33.58    9.38   20.08   56.59
WC             4 22 4705.91 2047.63 4274.00 4559.39 1931.83 1823.00 9004.00
abstraction    5 22   -0.45    1.83   -0.53   -0.44    1.95   -4.10    3.55
jargon         6 22   35.97    6.97   37.36   36.23    6.28   23.85   46.48
CertSent       7 22    6.22    0.18    6.20    6.23    0.24    5.82    6.47
LingObf        8 22    0.47    1.43    0.18    0.36    1.57   -1.66    4.13
cause          9 22    2.84    1.08    2.74    2.75    1.03    1.14    5.66
emo_pos       10 22    0.03    0.04    0.01    0.02    0.02    0.00    0.15
article       11 22    8.19    1.94    8.51    8.12    1.53    4.90   12.66
prep          12 22   14.84    1.64   15.19   14.84    1.75   11.80   17.58
quantity      13 22    4.22    1.39    4.10    4.15    1.54    2.27    6.95
PaperType     14 22    2.00    0.00    2.00    2.00    0.00    2.00    2.00
              range  skew kurtosis     se
year          40.00 -1.37     1.52   2.05
Refs          76.00  0.86    -0.01   4.19
flesch_re     36.51  0.59    -0.62   2.05
```

```
WC          7181.00  0.56    -0.81 436.56
abstraction    7.65  0.02    -0.64   0.39
jargon        22.63 -0.38    -1.16   1.49
CertSent       0.64 -0.27    -1.00   0.04
LingObf        5.79  0.68    -0.15   0.30
cause          4.52  0.78     0.30   0.23
emo_pos        0.15  1.55     1.39   0.01
article        7.76  0.16    -0.34   0.41
prep           5.78 -0.11    -1.18   0.35
quantity       4.68  0.46    -1.11   0.30
PaperType      0.00   NaN      NaN   0.00
------------------------------------------------------------
group: Single-Authored Genuine Papers
            vars  n    mean      sd  median trimmed     mad     min     max
year           1 22 2008.95   12.24 2013.50 2010.61    5.93 1980.00 2022.00
Refs           2 22   51.05   32.53   40.50   46.28   28.91   12.00  146.00
flesch_re      3 22   34.46   15.72   38.40   36.15   10.34  -10.89   54.97
WC             4 22 4632.36 2424.81 4148.00 4491.00 2364.01 1141.00 9670.00
abstraction    5 22    0.80    1.89    1.00    0.91    2.13   -3.39    3.43
jargon         6 22   29.50    6.19   31.79   29.61    5.77   19.26   38.66
CertSent       7 22    6.07    0.28    6.13    6.07    0.31    5.55    6.49
LingObf        8 22   -0.25    1.77   -0.30   -0.15    1.56   -5.84    2.97
cause          9 22    2.73    1.04    2.55    2.67    1.05    1.18    4.91
emo_pos       10 22    0.12    0.26    0.06    0.06    0.09    0.00    1.23
article       11 22    9.13    2.45    8.61    8.95    2.64    5.91   14.47
prep          12 22   15.66    1.15   15.89   15.83    0.76   12.82   17.22
quantity      13 22    4.66    1.64    4.46    4.55    1.43    2.05    8.66
PaperType     14 22    3.00    0.00    3.00    3.00    0.00    3.00    3.00
              range  skew kurtosis     se
year          42.00 -1.11    -0.07   2.61
Refs         134.00  1.32     1.43   6.94
flesch_re     65.86 -1.16     0.90   3.35
WC          8529.00  0.42    -0.93 516.97
abstraction    6.82 -0.42    -0.84   0.40
jargon        19.40 -0.30    -1.34   1.32
CertSent       0.94 -0.27    -1.13   0.06
LingObf        8.81 -1.03     2.29   0.38
cause          3.73  0.42    -0.61   0.22
emo_pos        1.23  3.38    11.22   0.06
article        8.56  0.48    -0.91   0.52
prep           4.40 -1.22     0.76   0.25
quantity       6.61  0.62    -0.27   0.35
PaperType      0.00   NaN      NaN   0.00
```

```
------------------------------------------------------------
group: Multi-Authored Genuine Papers
            vars  n    mean      sd  median trimmed     mad     min      max
year           1 22 2010.77    8.50 2012.50 2011.89    8.15 1990.00  2022.00
Refs           2 22   44.05   20.05   36.50   41.94   15.57   19.00    91.00
flesch_re      3 22   34.86    8.34   36.17   35.12    6.35   15.95    49.45
WC             4 22 4407.82 2741.91 3624.00 4009.67 1307.65 1124.00 13901.00
abstraction    5 22   -0.56    1.83   -0.42   -0.48    1.65   -4.46     3.07
jargon         6 22   33.61    7.19   35.03   33.51    6.42   20.86    46.67
CertSent       7 22    6.19    0.31    6.22    6.20    0.19    5.21     6.76
LingObf        8 22    0.02    1.69   -0.45   -0.16    1.05   -2.50     4.28
cause          9 22    2.90    1.30    2.44    2.83    0.79    1.04     5.56
emo_pos       10 22    0.04    0.05    0.04    0.04    0.05    0.00     0.17
article       11 22    7.88    1.62    7.71    7.78    1.06    4.22    12.06
prep          12 22   15.48    2.15   15.66   15.58    1.41    9.76    20.67
quantity      13 22    3.75    1.09    3.74    3.76    1.21    1.69     5.53
PaperType     14 22    4.00    0.00    4.00    4.00    0.00    4.00     4.00
               range  skew kurtosis     se
year           32.00 -0.95     0.15   1.81
Refs           72.00  0.77    -0.61   4.27
flesch_re      33.50 -0.41    -0.37   1.78
WC          12777.00  1.90     3.92 584.58
abstraction     7.53 -0.26    -0.37   0.39
jargon         25.81 -0.16    -1.01   1.53
CertSent        1.55 -1.02     2.09   0.07
LingObf         6.78  1.08     0.44   0.36
cause           4.52  0.65    -0.77   0.28
emo_pos         0.17  0.88     0.00   0.01
article         7.84  0.52     0.94   0.35
prep           10.91 -0.37     1.33   0.46
quantity        3.84 -0.05    -1.19   0.23
PaperType       0.00   NaN      NaN   0.00
```

- Some quick notes:
  - **Year:** The year seems to be relatively similar between groups, with the single-author groups (SAGP and SAFP) both having mid-2008 as the mean and the multi-author groups both having mid-2010 as the mean. This may be partially due to there being single outliers (matched papers that are both outliers) at the lower end of the distribution.
  - **Refs:** As I suspected, the outlier of 146 seems to be pulling the mean for SAGP up to 51.05 (the means for all other groups are around 44). This is corroborated by the skew statistics which show that only the SAGP group has a skew value above

1. I may run the analyses where references is the outcome variable both with and without this case in the model, or I may use word count as a covariate in the model to try to account for it.

- **WC:** For some reason, there seems to be a notable difference in mean word count between the single-authored fraudulent group and the rest of the groups. I could speculate as to why this is the case, but I may do an unplanned follow-up analysis of this.

2. Frequencies will now be produced for categorical variables, both for the data in general and within `PaperType` groups.

- First, we will make the variables `inst_pres`, `gender`, `simple_reason`, `Country`, and `PaperType` factor variables.

```
# Changing inst_pres (institutional prestige), gender, simple_reason, Country, and PaperTy
data$inst_pres <- factor(data$inst_pres, levels = c(0, 1), labels = c("Not Major Research
data$gender <- factor(data$gender, levels = c("FEMALE", "MALE"), labels = c("Female", "Mal
data$simple_reason <- factor(data$simple_reason, levels = c("f_data", "f_image", "m_image"
data$Country <- factor(data$Country)
data$PaperType <- factor(data$PaperType)
```

- Next, we will produce frequency tables for each categorical variable of interest for the whole dataset.

```
# Creating the frequency tables for each categorical variable
freq_tab_inst_pres <- table(data$inst_pres)
freq_tab_gender <- table(data$gender)
freq_tab_simple_reason <- table(data$simple_reason)
freq_tab_Country <- table(data$Country)
freq_tab_PaperType <- table(data$PaperType)

# Displaying the frequency tables
freq_tab_inst_pres
```

```
Not Major Research Institution     Major Research Institution
                            70                             18
```

```
freq_tab_gender
```

```
Female   Male
    15     73
```

```
freq_tab_simple_reason
```

```
        Fabricated/Falsified Data          Fabricated/Falsified Image
                               27                                   3
              Manipulated Image Fabricated/Falsified Data and Image
                               12                                   2
```

```
freq_tab_Country
```

```
     Australia         Belgium           China           Egypt        Ethiopia
             1               1              10               3               2
         India            Iran         Israel           Italy           Japan
            10               1               1               1               3
        Latvia        Malaysia    Netherlands        Pakistan          Poland
             1               2               6               3               2
      Portugal    South Africa    South Korea          Taiwan          Turkey
             2               1               2               1               2
United Kingdom   United States
             4              29
```

```
freq_tab_PaperType
```

```
MAFP MAGP SAFP SAGP
  22   22   22   22
```

- Let's take a look at the way these are split:
  - First, there seems to ba a preponderance of lower status (not major) research institutions in the whole dataset (70:18).
  - Next, most of the papers were retracted due to fabricated/falsified data ($n = 27$), followed by those that manipulated images ($n = 12$), fabricated/falsified images ($n = 3$), or both fabricated/falsified data and images ($n = 2$).
  - For country, the most common category of papers seem to be from the United States ($n = 29$), followed by China and India (both $n = 10$) and the Netherlands ($n = 6$), with no other countries totaling more than $n = 4$.
  - Finally, we can see that the papers are evenly split between each of our groups ($n = 22$; $N = 88$).

- Next, we will produce the frequencies within `PaperType` groups. Because we were not able to perfectly match across our matching characteristics, we will take this opportunity to look at how far off we were. A proportion table will be produced to more easily compare frequencies across `PaperType` groups.
    - First, we need to make the matching dummy variables into factor variables with labels to make them more interpretable.

```
# Creating factor variables out of the dummy variables that track matching across groups
data$different_country <- factor(data$different_country, levels = c(0, 1), labels = c("Sam
data$different_gender <- factor(data$different_gender, levels = c(0, 1), labels = c("Same
data$different_inst_pres <- factor(data$different_inst_pres, levels = c(0, 1), labels = c(
data$different_journal <- factor(data$different_journal, levels = c(0, 1), labels = c("Sam
```

- I have checked the dataset, and the labels seem to have been assigned correctly.


Aggarwal, S. B., Chaitanya. (2022). *textstat: Calculate statistical features from text* [Python]. https://github.com/shivam5992/textstat

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin*, 1–47. https://www.researchgate.net/profile/Ryan-Boyd-8/publication/358725479_The_Development_and_Psychometric_Properties_of_LIWC-22/links/6210f62c4be28e145ca1e60b/The-Development-and-Psychometric-Properties-of-LIWC-22.pdf

*Retraction Watch Database*. (2023). http://retractiondatabase.org/RetractionSearch.aspx?

Rocklage, M. D., He, S., Rucker, D. D., & Nordgren, L. F. (2023). Beyond Sentiment: The Value and Measurement of Consumer Certainty in Language. *Journal of Marketing Research*, *60*(5), 870–888. https://doi.org/10.1177/00222437221134802