Data Cleaning

As described in the pages of the book _____, Bill Adair and his research team sought to understand how politicians lie through data. Throughout our research, which is based on a dataset of existing fact-checks, we asked who lies more, who lies worse, what do they lie about, and where do they lie?

Much of this data analysis was initially done in spreadsheets, but has been replicated here in R so that it can be reproduced and shared with the public. We seek transparency in all our work, so feel free to explore the data yourself. Sometimes the changes may seem small, irrelevant and tedious, but we wanted to be fully transparent with each and every change we made.

This is the first of a series of Rmd files that walk a reader through our work. Here we describe our data management and cleaning process in detail. We used a version system to keep ourselves organized and archive previous versions of the data.

Each version represents a new dataset that has been altered from the previous one. This allows us to look back at previous steps and clearly communicate every change we made.

Version 0 - All Unsorted Data

Our data was drawn from Claim Review. ClaimReview is a tagging system developed by the Duke Reporters Lab and Google in 2015. When a fact-checker publishes a fact-check, they submit some baseline information to the ClaimReview database about the fact-check. Some of the categories are listed below.

url - Link to the fact-check. title - The title of the fact-check, often a summary. publisher.name - Name of the site that published the fact-check (PolitiFact). reviewDate - Date the fact-check was published. text - The actual claim being fact-checked. claimDate - The date the claim was made. claimant - The name of the speaker/person making the claim. textualRating - The rating assigned to the claim.

ClaimReview was initially developed to promote fact-checks in search results, incentivizing fact-checkers to participate. ClaimReview also had something of an unintended benefit: It created a giant, growing database of fact-checks uploaded by fact-checkers around the world. From US-based fact-checkers alone, there are 18,770 claims from Jan 2016 to June 2021.

This is the database from which we pulled in order to conduct our reserach.

The version 0 dataset was pulled by Duke Reporters Lab Joel Luther. The dataset includes all entries from Claim Review that were published by fact checkers based in the US (and in English). 14 unique fact-checking publishers were included. Included entries were those in which the reviewDate was between Jan 1st, 2016 and June 30th, 2021.

Table 1: All US Claims from ClaimReview 01/01/2016 - 06/30/2021

n
18770

Table 2: All US Claims from ClaimReview 01/01/2016 - 06/30/2021 Sorted by Publisher

publisher.site	n
cbsnews.com	240
checkyourfact.com	2108
factcheck.org	2310
factcheck.thedispatch.com	176
hoax-alert.leadstories.com	5
leadstories.com	3287
newsweek.com	189
nytimes.com	485
politifact.com	7770
polygraph.info	3
poynter.org	50
thegazette.com	9
usatoday.com	756
vox.com	2
washingtonpost.com	1380

To access V0, and any other code or datasets used in this project, feel free to visit our public github repository.

Version 1 - Only Political Figures

ClaimReview consists of fact-checks for all sorts of misinformation, not just the lies of politicians, which was the focus of our research. So we filtered the claimants so that the data was narrowed down to just claims made by political figures. That process is described below.

Asa Royal, the software engineer assisting on the project, built a small data pipeline that tried to identify whether the speaker of each claim was a person and whether they met our definition of a politician. For our purposes, a politician was anyone who'd held or run for partisan office, or who'd been hired or appointed by a such a person to serve on a campaign or in a government agency. For the remainder of our analysis we will be using the terms politician and political figure interchangably.

We identified human claimants by feeding claimant names in our dataset through an entity recognition API. Given text like ""The Sierra Club"" or ""Bill Murray"", the API tried to detect what it referred to – *e.g.* a person or an organization – letting us label claimants as human.

Next, we tried to categorize human claimants as politicians or non-politicians. We began by running claimant names through a Wikipedia API. If a given claimant had a Wikipedia article about them, our code checked whether the article contained any "politician indicators". For example, if the infobox in an article about a claimant said they'd worked as a politician or political operative, we accepted that as fact and marked them as a politician. Likewise, if in the first few paragraphs of the Wikipedia article, text matched certain keywords (*e.g.* in a single sentence, the article mentioned its subject "ran" in an "election" or "work"ed in the "White House"), the code inferred that they, too, were a politician.

Additionally, once the code deduced that a claimant was a politician, it scanned their Wikipedia infobox for reference to their party affiliation, recording it in our dataset. If there was no Wikipedia article about a human claimant, as was occasionally the case for unsuccessful candidates for local office, the data pipeline tried to find a corresponding page for the claimant on Ballotpedia, an online encyclopedia of American politics. If it found one, it searched the page for the same markers looked for on Wikipedia pages, and similarly tried to identify the claimant's party affiliation.

When the aforementioned process was complete for each claimant and we filtered out all claims in the dataset whose claimants weren't labeled as human politicians, we considered data collection to be finished. To begin

making this data workable, we undertook some light data cleaning. Namely, if we found that a claimant had been affiliated with several political parties, we gave them a party label corresponding to the party they'd identified with when they'd made the claim. In a small number of cases, our code categorized a claimant as a politician but failed to infer their party; those claimants' politician status and party ID were manually reviewed and edited at a later stage but were labelled as unknown_affiliation for the time being. The resulting dataset only contained claims by political figures with a corresponding party label.

Table 3: All Politician Claims

n
7635

Table 4: All Politician Claims Sorted by Publisher

publisher.site	n
cbsnews.com	167
checkyourfact.com	10
factcheck.org	1291
factcheck.thedispatch.com	59
leadstories.com	1
newsweek.com	66
nytimes.com	465
politifact.com	4282
polygraph.info	3
poynter.org	10
thegazette.com	7
usatoday.com	19
vox.com	2
washingtonpost.com	1253

Table 5: All Politician Claims Sorted by Party

claimant_party	n
[[Chinese Communist Party]]	1
[[Conservative Party (UK)|Conservative]]	1
[[Labour Party (UK)|Labour]]	2
[[Liberal Party of Canada|Liberal]]	1
[[Likud]]	2
[[Moderate Party|Moderate]]	1
[[UK Independence Party]] (2021–present)	1
[[Workers' Party of Korea]]	1
{{ubl|[[Peace and Freedom Party|Peace and Freedom]] (2012–2013)|[[Green Party of the United States|Green]] (2008–2012)}}	1
British Freedom Party (2018–present)	1
Democratic	2325
Independent	52
Libertarian	9
Republican	5119
unknown_affiliation	118

Version 2 - 5 Parties

As seen in Table 5, several of the claimants were assigned non-US political parties, suggesting they were not US political figures. At this point, all claims whose party affiliation in claimant_party was not Democratic, Republican, Libertarian, Independent, or unknown_affilition were removed. Version 2 only consists of claims made by political figures belonging to U.S. political parties or ones that we assigned an unknown affiliation. At this time, those with unknown_ affiliation were left unsorted. With further cleaning needing to be done, we held off on manually sorting these claims until we had further narrowed our dataset.

Table 6: All Claims in 5 Parties

n
7623

Table 7: 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

Table 8: All Claims in 5 Parties Sorted by Party

claimant_party	n
Democratic	2325
Independent	52
Libertarian	9
Republican	5119
unknown_affiliation	118

Version 3 - Deduped

At this step we filtered out duplicate claims where both the url and and text of the claim were identical. The first appearance of each claim remained in the dataset, others were removed. This resulted in a database of 7,360 unique claims.

Note: Later on in our subject tagging we uncovered more duplicates that had identical text, but had been

republished by a website under a different url. For transparency about our data cleaning process we are following our original steps and will remove those duplicates later.

Table 9: All Claims in 5 Parties - Deduped

n
7360

Table 10: All Claims in 5 Parties - Deduped and 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

Table 11: All Claims in 5 Parties - Deduped and Sorted by Party

claimant_party	n
Democratic	2265
Independent	50
Libertarian	9
Republican	4921
unknown_affiliation	115

Version 4 - Dejunked

For a final clean up before sorting by publisher and removing Trump, a member of our research team **manually identified 23 false positive rows** that either listed multiple claimants or had a false positive non-human claimant (e.g. "Donald Trump's campaign"). Those 23 claims were manually removed.

Table 12: All Claims in 5 Parties - Deduped and Dejunked

n
7337

Table 13: All Claims in 5 Parties - Deduped and Dejunked Sorted by Publisher

publisher.site	n
cbsnews.com	157
checkyourfact.com	9
factcheck.org	1237
factcheck.thedispatch.com	53
newsweek.com	62
nytimes.com	444
politifact.com	4167
polygraph.info	2
poynter.org	9
thegazette.com	7
usatoday.com	17
vox.com	1
washingtonpost.com	1172

Table 14: All Claims in 5 Parties - Deduped and Dejunked Sorted by Party

claimant_party	n
Democratic	2259
Independent	50
Libertarian	9
Republican	4907
unknown_affiliation	112

At this stage, the dataset of 7,337 claims had been filtered to fit the parameters of our query. The dataset that remains consists of claims that were made by U.S. political figures, and fact-checked by one of 13 U.S based publishers.

Version 5 - Split into Publisher data

From here on out, we began working with each publisher dataset separately. Only four publishers had enough remaining claims to move forward: FactCheck.Org, The New York Times, PolitiFact, and The Washington Post. This resulted in four individual datasets, one per publisher.

Table 15: FactCheck.Org claims

n
1237

Table 16: FactCheck.Org claims, Sorted by Party

claimant_party	n
Democratic	288
Independent	12

claimant_party	n
Libertarian	1
Republican	922
unknown_affiliation	14

Table 17: New York Times claims

n
444

Table 18: New York Times claims, Sorted by Party

claimant_party	n
Democratic	101
Republican	341
unknown_affiliation	2

Table 19: Politifact claims

n
4167

Table 20: Politifact claims, Sorted by Party

claimant_party	n
Democratic	1529
Independent	23
Libertarian	8
Republican	2535
unknown_affiliation	72

Table 21: Washington Post claims

n
1172

Table 22: Washington Post claims, Sorted by Party

claimant_party	n
Democratic	293
Independent	12
Republican	855

claimant_party	n
unknown_affiliation	12

Version 6 - Removing Trump

We made the decision to remove Donald Trump entirely from this dataset. He was overrepresented in the dataset and his large number of Republican falsehoods threatened to skew the data. To do so, we first identified all the different iterations of "Trump" within the `claimant` column. We could not simply remove all claimants whose names included Trump, because by our definitions, his daughter Ivanka Trump was a political figure.

Table 23: Names including 'Trump'

claimant	n
Donald J. Trump	205
Donald trump	1
Donald Trump	2587
Ivanka Trump	10
President Donald J. Trump	2

With these four versions of Trump's name, we manually removed them from the data. This created a dataset without Trump for each publisher, denoted V6.

Table 24: FactCheck.Org claims, no Trump

n
526

Table 25: FactCheck.Org claims, no Trump - Sorted by Party

claimant_party	n
Democratic	288
Independent	12
Libertarian	1
Republican	211
unknown_affiliation	14

Table 26: New York Times claims, no Trump

n
180

Table 27: New York Times claims, no Trump -Sorted by Party

claimant_party	n
Democratic	101
Republican	77
unknown_affiliation	2

Table 28: Politifact claims, no Trump

n
2951

Table 29: Politifact claims, no Trump - Sorted by Party

claimant_party	n
Democratic	1529
Independent	23
Libertarian	8
Republican	1319
unknown_affiliation	72

Table 30: Washington Post claims, no Trump

n
568

Table 31: Washington Post claims, no Trump - Sorted by Party

claimant_party	n
Democratic	293
Independent	12
Republican	251
unknown_affiliation	12

Version 7 - Sorting Unknown Affiliation

After we had divided the data into smaller publisher sub-sets and removed Trump from the data, our next step was to address missing data in the claimant_party section. As a reminder, the code that was used to assign party affiliations scraped Wikipedia data and Ballotpedia for politicians and party affiliations. 100 remaining claims were made by a claimant who was identified as a political figure but for whom a party could not be found. They were labeled unknown_affiliation. Having reduced our data to the selection of claims we planned to analyze, our next step was to manually assign party affiliations to the unknown_affiliation claims.

This was done by pulling the claims labeled unknown_affiliation for each publisher. Based on the name of the claimant and the content of the fact-check, claimants were either

1. Assigned a party (Democrat, Republican, Independent, Libertarian)
2. Marked to be removed from the dataset because they were not a political figure.

NOTE: If a claimant served in an administration and spoke on behalf of that administration or politician, they were considered a political figure and assigned the affiliation of the administration or politician they served/spoke on behalf of. This may not always align with the their personal political beliefs but represented the party they spoke for within their role. This applied to appointed cabinet members, political officials, and lawyers representing politicians. Political commentators, activists, and celebrities were removed and not considered political figures

If you are reproducing this, these steps may not be relevant to you. Our filtering algorithm was unable to assign these political figures accurately despite many of them having political affiliations based on our definition. We are providing this information for the sake of transparency and to exemplify how one might manually sort the unknown_affiliations.

FactCheck.Org unknown affiliation review First we reviewed the unknown_affiliation claims from the FactCheck.Org database.

Table 32: Unknown Affiliation Claimants to be Sorted - FactCheck.Org

claimant	n
Kirstjen Nielsen	4
Brett Giroir	2
Dr. Deborah Birx	1
Hal Turner	1
John Kelly	1
Kevin K. McAleenan	1
Loretta Lynch	1
Sean P. Conley	1
Sidney Powell	1
Steve Cortes	1

There were 14 claims and 10 unique claimants.

After a manual review, the following claimants were assigned Republican affiliation:

Brett Giroir who Served as Assistant Secretary of Health in the Trump Admin (2 claims)

Dr. Deborah Birx who served in the Trump admin(1 claim)

John Kelly, Former Chief of Staff in Trump admin (1 claim)

Kevin K. McAleenan, Sec. of Homeland Sec. Trump admin (1 claim)

Kirstjen Nielsen, Homeland Security Sec. Trump admin (4 claims)

Sidney Powell, Lawyer representing President Donald Trump (1 claim)

Steve Cortes, Former advisor to Donald Trump (1 claim)

One claimant was assigned a Democratic affiliation:

Loretta Lynch, Attorney General for Obama (1 claim)

2 claimants were removed for not being political figures:

Hal Turner, commentator, not a political figure (1 claim)

Sean P. Conley, Presidential physician, not a political figure (1 claim)

In the end, 11 total claims were assigned a Republican affiliation, 1 claim was assigned a Democratic affiliation, and 2 claims were removed from the FactCheck.Org dataset. Table 33 shows the count by party after this adjustment.

Table 33: Factcheck.org by Claimant Party, without Unknown Affiliation

claimant_party	n
Democratic	289
Independent	12
Libertarian	1
Republican	222

New York Times unknown affiliation review Next we reviewed the unknown_affiliation claims for the New York Times database.

Table 34: Unknown Affiliation Claimants to be Sorted - New York Times

claimant	n
Jesse Binnall	1
Sidney Powell	1

There were 2 unique claimants and 2 claims, both of which were assigned Republican affiliation:

Sidney Powell, Lawyer representing President Donald Trump (1 claim)

Jesse Binnall, Lawyer representing President Donald Trump (1 claim)

2 total claims were assigned a Republican affiliation Table 35 shows the count by party after this adjustment.

Table 35: New York Times by Claimant Party, without Unknown Affiliation

claimant_party	n
Democratic	101
Republican	79

PolitiFact unknown affiliation review Next we reviewed unknown_ affiliation claims for the PolitiFact database. There were 59 unique claimants and 72 claims.

Table 36: Unknown Affiliation Claimants to be Sorted - PolitiFact

claimant	n
Jake Tapper	3
John Kelly	3
Scott Jones	3
Brett Giroir	2

claimant	n
Daniel Kelly	2
Lowell Holtz	2
Maria Bartiromo	2
Michael Long	2
Nicholas Burns	2
Sidney Powell	2
Alejandro Mayorkas	1
Alexander Strenger	1
Ashley Smith	1
Beth Parlato	1
Bill Maher	1
Bob Donovan	1
Bob Spindell	1
Bono	1
Brett McGurk	1
Brit Hume	1
Charles Francis	1
Charles Ramsey	1
Chris Meagher	1
Clay Aiken	1
David Martin	1
Edward Flynn	1
Erin Burnett	1
George Papadopoulos	1
Giffords	1
H.R. McMaster	1
Hal Turner	1
J. Christian Adams	1
James Hauser	1
Jeffrey Zients	1
Jesse Kremer	1
John Kirby	1
John Patrick	1
Kevin Downing	1
Kirstjen Nielsen	1
Lawrence O'Donnell	1
Lisa Moore	1
Loretta Lynch	1
Louis Marinelli	1
Marc Butler	1
Mark Levin	1
Michael Screnock	1
Mike Collier	1
Mike Crute	1
Nate McMurray	1
Paul Maner	1
Ray Cross	1
Robert Sanborn	1
Ryan Frazier	1
Scott Dawson	1
Shelley Grogan	1
State representatives	1

claimant	n
Steve Cortes	1
Todd Wilcox	1
William Taylor	1

The following claimants were assigned a Republican affiliation:

Beth Parlato, NY Republican political candidate (1 claim)

Bob Spindell, Republican member of WI elections commission (1 claim)

Brett Giroir, Served as Assistant Secretary of Health in the Trump admin (2 claims)

Brett McGurk, National Security positions with Bush, Trump, and Obama (1 claim)

George Papadopoulos, Served on Trump campaign and congressional candidate(1 claim)

H.R. McMaster, National Security Advisor in Trump admin (1 claim)

Jesse Kremer, Republican WI legislator (1 claim)

John Kelly, Former Chief of Staff in Trump Admin (3 claims)

Kirstjen Nielsen, Homeland Security Sec. Trump Admin (1 claim)

Louis Marinelli, California Political Candidate (1 claim)

Marc Butler, New York Republican (1 claim)

Paul Maner, Georgia Republican candidate (1 claim)

Ryan Frazier, Colorado Republican candidate (1 claim)

Scott Dawson, Alabama Republican candidate (1 claim)

Scott Jones, Republican congressional candidate (3 claims)

Sidney Powell, Lawyer representing President Donald Trump (2 claims)

Steve Cortes, Former advisor to Donald Trump (1 claim)

State Representatives, link lists Wisconsin Assembly Republicans) (1 claim)

Todd Wilcox, Florida Republican candidate (1 claim)

The following claimants were assigned a Democratic affiliation:

Alejandro Mayorkas, Biden Sec. of Homeland Security (1 claim)

Chris Meagher, Dep. Press Sec. for Biden Admin (1 claim)

Jeffery Zients, Coronavirus Response Coordinator for Biden (1 claim)

John Kirby, Dept of Defense Obama/Biden, (1 claim)

Loretta Lynch, Attorney General for Obama (1 claim)

Mike Collier, Dem. Texas Lieutenant Gov. candidate (1 claim)

Nate McMurray, NY Democratic politician (1 claim)

Clay Aiken, NC Congressional candidate, (1 claim)

The following claimants were assigned an Independent affiliation:

Alexander Strenger, ran for city council in Austin (1 claim)

Bob Donovan, non-partisan Milwaukee Common Council member (1 claim)

Charles Francis, non-partisan Mayoral candidate (1 claim)

Lowell Holtz, non-partisan WI candidate (2 claims)

Michael Long, New York State Conservative Party Chairman (2 claims)

William Taylor, Ambassador to Ukraine, worked in both party administrations (1 claim)

The following claimants were removed from the dataset:

Ashley Smith, civilian, not political figure based on PF (1 claim)

Bill Maher, political comedian, not a political figure (1 claim)

Bono, singer, not political figure (1 claim)

Brit Hume, journalist, not a political figure (1 claim)

Charles Ramsey, police chief, not a political figure (1 claim)

Daniel Kelly, WI Supreme Court Justice, not a political figure (2 claims)

David Martin, CEO of M CAM Inc., not a political figure (1 claim)

Edward Flynn, Chief of Milwaukee Police dept., not a political figure (1 claim)

Erin Burnett, Political commentator, mot a political figure (1 claim)

Giffords, Gun violence activist group, not political figure (1 claim)

Hal Turner, commentator, Not a political figure (1 claim)

J. Christian Adams, conservative activist, not a political figure (1 claim)

Jake Tapper, journalist, Not a political figure (3 claims)

James Hauser, civilian, not a political figure, only post (1 claim)

John Patrick, President of the Texas AFL-CIO, not a political figure (1 claim)

Kevin Downing, Attorney to Paul Manafort, not a political figure (1 claim)

Lawrence O'Donnell, MSNBC Host, not a political figure (1 claim)

Lisa Moore, English Professor, not a political figure (1 claim)

Maria Bartiromo, FOX news host, not a political figure (2 claims)

Mark Levin, Talk show host, not a political figure (1 claim)

Michael Screnock, WI Supreme Court Justice, not a political figure (1 claim)

Mike Crute, radio show host, not a political figure (1 claim)

Nicholas Burns, Former ambassador, not a political figure (2 claims)

Ray Cross, University President, not a political figure (1 claim)

Robert Sanborn, nonprofit president, not a political figure (1 claim)

Shelley Grogan, WI judge non-partisan, not a political figure (1 claim)

In the end, 25 claims were assigned a Republican affiliation, 8 claims were assigned a Democratic affiliation, 8 claims were assigned an Independent affiliation, and 31 cliams were removed from the PolitiFact dataset. Table 37 shows the couns by party after the sort.

Table 37: PolitiFact by Claimant Party, without Unknown Affiliation

claimant_party	n
Democratic	1537
Independent	31
Libertarian	8
Republican	1344

The Washington Post unknown affiliation review Next we reviewed unknown_ affiliation claims for the Washington Post database. There were 9 unique claimants and 12 claims.

Table 38: Unknown Affiliation Claimants to be Sorted - The Washington Post

claimant	n
Kirstjen Nielsen	4
Anthony Fauci	1
Charmaine Yoest	1
H.R. McMaster	1
Jim Acosta	1
Jim Mattis	1
Loretta Lynch	1
Steve Cortes	1
Vanita Gupta	1

The following we assigned a Republican affiliation:

Charmaine Yoest, Trump admin appointee, (1 claim)

H.R. McMaster, National Security Advisor in Trump admin, (1 claim)

Jim Mattis, Trump Sec. of Defense, (1 claim)

Kirstjen Nielsen, Homeland Security Sec. Trump admin, (4 claims)

Steve Cortes, Former advisor to Donald Trump, (1 claim)

The following were assigned a Democratic affiliation:

Vanita Gupta, worked in the Attorney Gen. office for Obama and Biden, (1 claim)

Loretta Lynch, Attorney General for Obama, (1 claim)

The following were removed for not being political figures:

Anthony Fauci, not a political figure (1 claim)

Jim Acosta, Political commentator, not a political figure (1 claim)

In the end, 8 claims were assigned a Republican affiliation, 2 claims were assigned a Democratic affiliation and 2 claims were removed from the Washington Post dataset. Table 39 shows these updated counts.

Table 39: Washington Post by Claimant Party, without Unknown Affiliation

claimant_party	n
Democratic	295
Independent	12
Republican	259

Version 8 - Removing Libertarian and Independents

At this stage, it became clear that the Independent and Libertarian claims made up a very small part of each publisher dataset.

These categories were too small to draw statistically relevant conclusions from the data, and so to not distract from potential findings about the two major US parties, claims with an Independent or Libertarian party affiliation were removed from our analysis. Only Democratic and Republican claims remain in the datasets.

Table 40: Factcheck.org by Claimant Party

claimant_party	n
Democratic	289
Republican	222

Table 41: New York Times by Claimant Party

claimant_party	n
Democratic	101
Republican	79

Table 42: PolitiFact by Claimant Party

claimant_party	n
Democratic	1537
Republican	1344

Table 43: Washington Post by Claimant Party

claimant_party	n
Democratic	295
Republican	259

At this stage we moved forward with data set specific cleaning. Each publisher had it own rating scale, rating quirks, and duplications that needed to be addressed. Also we decided not to move forward with analysis of FactCheck.Org and the Washington Post. To see futher data cleaning or to understand why we did not move forward with our analysis see the following documents in GitHub:

PolitiFact_CleanUp

Washington-Post-Data-Analysis

FactCheck_CleanUp

NewYorkTimes_CleanUp