MA Thesis Data Cleaning

Konrat Pekkip

2022-07-26

Data Cleaning and Preprocessing

The following code loads, cleans and preprocesses datasets relevant for the development of the natural language processing and machine learning models underlying my M.A. thesis. The majority of the data comes from the Open Discourse dataset, which was assembled and published by Richter et al. in 2020. It is available through the Harvard Dataverse; you can access it using this link. Further, I retrieved up-to-date geographic data provided by the Federal Statistical Office of Germany (Statistisches Bundesamt) through its *Gemeindeverzeichnis-Informationssystem* (GV-ISys). Use this link to access the GV-ISys. For the full analysis, make sure to run both this file to clean the data and the main file containing most of the relevant code. If you have trouble accessing any of the data or questions about my code, feel free to email me at kpekkip@uchicago.edu.

Loading Packages and Data

The following two code chunks load required packages and read in all relevant datasets.

```
#load packages
library(tidyverse)
library(readxl)
```

```
#load data
speeches <- read_csv("../data/speeches.csv")
politicians <- read_csv("../data/politicians.csv")
factions <- read_csv("../data/factions.csv")
cities <- read_excel("../data/german_cities2.xlsx",
                     sheet = "Onlineprodukt_Gemeinden",
                     skip = 3)
```

Data Preprocessing

In the following two code chunks, the `speeches` and `cities` data frames are amended to include variables indicating which *Bundestag* a speech was given in, assigning *Bundesländer* to each city, and indicating whether a given city is in East or West Germany.

```
#define variable indicating electoral term in more understandable terms
speeches <- speeches %>%
  mutate(bundestag = case_when(electoralTerm <= 10 ~ "pre-reunification",
                               electoralTerm == 11 ~ "11. Bundestag (1987-1990)",
                               electoralTerm == 12 ~ "12. Bundestag (1990-1994)",
                               electoralTerm == 13 ~ "13. Bundestag (1994-1998)",
                               electoralTerm == 14 ~ "14. Bundestag (1998-2002)",
                               electoralTerm == 15 ~ "15. Bundestag (2002-2005)",
```

```
                                electoralTerm == 16 ~ "16. Bundestag (2005-2009)",
                                electoralTerm == 17 ~ "17. Bundestag (2009-2013)",
                                electoralTerm == 18 ~ "18. Bundestag (2013-2017)",
                                electoralTerm == 19 ~ "19. Bundestag (2017-2021)"))

#subset speeches dataset to only include speeches given since reunification
speeches_subset <- speeches %>%
  filter(date >= "1990-10-03") %>%
  select(-firstName, -lastName) #remove names -- will be merged from politicians df later
```

```
#tidy up cities df, add variables indicating bundesland and east/west
cities_amended <- cities %>%
  select(Land, ...8) %>%
  rename(bundesland_code = Land,
         city = ...8) %>%
  slice(-1, -2, -3) %>%
  mutate(bundesland_code = as.numeric(bundesland_code),
         bundesland = case_when(bundesland_code == 01 ~ "Schleswig-Holstein",
                                bundesland_code == 02 ~ "Hamburg",
                                bundesland_code == 03 ~ "Niedersachsen",
                                bundesland_code == 04 ~ "Bremen",
                                bundesland_code == 05 ~ "Nordrhein-Westfalen",
                                bundesland_code == 06 ~ "Hessen",
                                bundesland_code == 07 ~ "Rheinland-Pfalz",
                                bundesland_code == 08 ~ "Baden-Württemberg",
                                bundesland_code == 09 ~ "Bayern",
                                bundesland_code == 10 ~ "Saarland",
                                bundesland_code == 11 ~ "Berlin",
                                bundesland_code == 12 ~ "Brandenburg",
                                bundesland_code == 13 ~ "Mecklenburg-Vorpommern",
                                bundesland_code == 14 ~ "Sachsen",
                                bundesland_code == 15 ~ "Sachsen-Anhalt",
                                bundesland_code == 16 ~ "Thüringen"),
         east_west = case_when(bundesland_code <= 10 ~ "West",
                               bundesland_code == 11 ~ "Berlin",
                               bundesland_code >= 12 ~ "East"))

#remove certain characters, standardize spelling of select cities in cities df
cities_amended$clean_city <- gsub('\\,.*', '', cities_amended$city)

cities_amended <- cities_amended %>%
  distinct(clean_city, .keep_all = TRUE) %>%
  mutate(clean_city = replace(clean_city, clean_city == "Frankfurt am Main", "Frankfurt_Main"),
         clean_city = replace(clean_city, clean_city == "Frankfurt (Oder)", "Frankfurt_Oder"),
         clean_city = replace(clean_city, clean_city == "Rüsselsheim am Main", "Rüsselsheim"))
```

Data Cleaning

The Open Discourse datasets do not contain all the contemporary names of cities and countries MdB were born in; instead providing only information on the city name and country at the time of an MdB's birth. Especially for older generations of MdB, this poses an issue, as many were born prior to the consolidation of modern European borders. Moreover, many German city names changed over time, especially during sweeping local government reorganization processes in the 1960s and 1970s. The code in the chunk below first

subsets the `politicians` data frame to only include those politicians who appear in the relevant subset of the `speeches` data frame, before manually updating city names in order to match the entries in the `cities` data frame. Finally, all cities that are falsely listed as German in the Open Discourse dataset are manually identified and coded as foreign cities.

```
#initial cleaning of city names in politicians df
politicians$clean_city <- gsub('\\ /.*', '', politicians$birthPlace)
politicians$clean_city <- gsub('\\,.*', '', politicians$clean_city)
politicians$clean_city <- gsub('\\-.*', '', politicians$clean_city)

#subset only politicians who appear in speeches_subset; manually update German city names, accurately c
politicians_subset <- politicians %>%
  rename(politicianId = id) %>%
  subset(., politicianId %in% speeches_subset$politicianId) %>%
  mutate(born_germany = ifelse(birthCountry != "Deutschland", 0, 1))  %>%
  mutate(clean_city = replace(clean_city, clean_city == "Frankfurt am Main" | clean_city == "Frankfurt/
         clean_city = replace(clean_city, clean_city == "Frankfurt (Oder)", "Frankfurt_Oder"),
         clean_city = replace(clean_city, clean_city == "Goslar am Harz", "Goslar"),
         clean_city = replace(clean_city, clean_city == "Drangstedt", "Geestland"),
         clean_city = replace(clean_city, clean_city == "Altenhaßlau", "Linsengericht"),
         clean_city = replace(clean_city, clean_city == "Ennigloh", "Bünde"),
         clean_city = replace(clean_city, clean_city == "Nahrstedt", "Stendal"),
         clean_city = replace(clean_city, clean_city == "Holz", "Heusweiler"),
         clean_city = replace(clean_city, clean_city == "Altenhaßlau", "Linsengericht"),
         clean_city = replace(clean_city, clean_city == "Kolenfeld", "Wunstorf"),
         clean_city = replace(clean_city, clean_city == "Neheim", "Arnsberg"),
         clean_city = replace(clean_city, clean_city == "Krofdorf", "Wettenberg"),
         clean_city = replace(clean_city, clean_city == "Schwarzenberg", "Schwarzenberg/Erzgeb."),
         clean_city = replace(clean_city, clean_city == "Röckersbühl", "Berngau"),
         clean_city = replace(clean_city, clean_city == "Tiengen", "Waldshut-Tiengen"),
         clean_city = replace(clean_city, clean_city == "Boizenburg", "Boizenburg/Elbe"),
         clean_city = replace(clean_city, clean_city == "Piesenkofen", "Obertraubling"),
         clean_city = replace(clean_city, clean_city == "Solscheid", "Buchholz (Westerwald)"),
         clean_city = replace(clean_city, clean_city == "Landau", "Landau in der Pfalz"),
         clean_city = replace(clean_city, clean_city == "Neuwürschnitz", "Oelsnitz/Erzgeb."),
         clean_city = replace(clean_city, clean_city == "Fussingen", "Waldbrunn (Westerwald)"),
         clean_city = replace(clean_city, clean_city == "Untermolbitz", "Rositz"),
         clean_city = replace(clean_city, clean_city == "Bevensen", "Bad Bevensen"),
         clean_city = replace(clean_city, clean_city == "Idar", "Idar-Oberstein"),
         clean_city = replace(clean_city, clean_city == "Rheinhausen (jetzt Duisburg)", "Duisburg"),
         clean_city = replace(clean_city, clean_city == "Groß Schwansee", "Kalkhorst"),
         clean_city = replace(clean_city, clean_city == "Kirchhellen", "Bottrop"),
         clean_city = replace(clean_city, clean_city == "Efferen", "Hürth"),
         clean_city = replace(clean_city, clean_city == "Walsum (jetzt Duisburg)", "Duisburg"),
         clean_city = replace(clean_city, clean_city == "Rheydt (jetzt Mönchengladbach)", "Mönchengladb
         clean_city = replace(clean_city, clean_city == "Dingen (jetzt Hamminken)", "Hamminkeln"),
         clean_city = replace(clean_city, clean_city == "Altenrheine", "Rheine"),
         clean_city = replace(clean_city, clean_city == "Schippach (jetzt Elsenfeld)", "Elsenfeld"),
         clean_city = replace(clean_city, clean_city == "Harmonie (jetzt Eitorf)", "Eitorf"),
         clean_city = replace(clean_city, clean_city == "Hüls (jetzt Krefeld)", "Krefeld"),
         clean_city = replace(clean_city, clean_city == "Alstätte (jetzt Ahaus)", "Ahaus"),
         clean_city = replace(clean_city, clean_city == "Erpen (jetzt Dissen)", "Dissen am Teutoburger
         clean_city = replace(clean_city, clean_city == "Elten (jetzt Emmerich)", "Emmerich am Rhein"),
         clean_city = replace(clean_city, clean_city == "Lank", "Meerbusch"),
```

```
        clean_city = replace(clean_city, clean_city == "Bad Godesberg (jetzt Bonn)", "Bonn"),
        clean_city = replace(clean_city, clean_city == "Bensberg (jetzt Bergisch Gladbach)", "Bergisch
        clean_city = replace(clean_city, clean_city == "Hehlrath (jetzt Eschweiler)", "Eschweiler"),
        clean_city = replace(clean_city, clean_city == "Greppin", "Bitterfeld-Wolfen"),
        clean_city = replace(clean_city, clean_city == "Baden", "Baden-Baden"),
        clean_city = replace(clean_city, clean_city == "Kirrlach", "Waghäusel"),
        clean_city = replace(clean_city, clean_city == "Ballenberg", "Ravenstein"),
        clean_city = replace(clean_city, clean_city == "Brünnau", "Prichsenstadt"),
        clean_city = replace(clean_city, clean_city == "Syrau", "Rosenbach/Vogtl."),
        clean_city = replace(clean_city, clean_city == "Elversberg", "Spiesen-Elversberg"),
        clean_city = replace(clean_city, clean_city == "Haltern", "Haltern am See"),
        clean_city = replace(clean_city, clean_city == "Altenbunnen", "Löningen"),
        clean_city = replace(clean_city, clean_city == "Eiweiler", "Heusweiler"),
        clean_city = replace(clean_city, clean_city == "Dorum", "Wurster Nordseeküste"),
        clean_city = replace(clean_city, clean_city == "Reideburg", "Halle (Saale)"),
        clean_city = replace(clean_city, clean_city == "Mahlow", "Blankenfelde-Mahlow"),
        clean_city = replace(clean_city, clean_city == "Bockel", "Wietzendorf"),
        clean_city = replace(clean_city, clean_city == "Steinehaig", "Frankenhardt"),
        clean_city = replace(clean_city, clean_city == "Großhennersdorf", "Herrnhut"),
        clean_city = replace(clean_city, clean_city == "Plöckendorf", "Rednitzhembach"),
        clean_city = replace(clean_city, clean_city == "Haren", "Haren (Ems)"),
        clean_city = replace(clean_city, clean_city == "Singen", "Singen (Hohentwiel)"),
        clean_city = replace(clean_city, clean_city == "Annerod bei Gießen", "Fernwald"),
        clean_city = replace(clean_city, clean_city == "Zerbst", "Zerbst/Anhalt"),
        clean_city = replace(clean_city, clean_city == "Karken", "Heinsberg"),
        clean_city = replace(clean_city, clean_city == "Heilberskofen b.Mamming", "Mamming"),
        clean_city = replace(clean_city, clean_city == "Hoffenheim", "Sinsheim"),
        clean_city = replace(clean_city, clean_city == "Reinharts", "Kempten (Allgäu)"),
        clean_city = replace(clean_city, clean_city == "Aistaig", "Oberndorf am Neckar"),
        clean_city = replace(clean_city, clean_city == "Ludwigshafen", "Ludwigshafen am Rhein"),
        clean_city = replace(clean_city, clean_city == "Wittichenau", "Wittichenau / Kulow"),
        clean_city = replace(clean_city, clean_city == "Krs. Havelberg", "Havelberg"),
        clean_city = replace(clean_city, clean_city == "Breitenhagen", "Barby"),
        clean_city = replace(clean_city, clean_city == "Mühlbach", "Pirna"),
        clean_city = replace(clean_city, clean_city == "Oschersleben", "Oschersleben (Bode)"),
        clean_city = replace(clean_city, clean_city == "Oldenburg (Oldb.)", "Oldenburg"),
        clean_city = replace(clean_city, clean_city == "Pippensen", "Buxtehude"),
        clean_city = replace(clean_city, clean_city == "Oberdielbach (Waldbrunn)", "Waldbrunn"),
        clean_city = replace(clean_city, clean_city == "Oberstetten", "Münsingen"),
        clean_city = replace(clean_city, clean_city == "Lutherstadt" | clean_city == "Lutherstadt Eisl
        clean_city = replace(clean_city, clean_city == "Lutherstadt Wittenberg", "Wittenberg"),
        clean_city = replace(clean_city, clean_city == "Neunkrichen (Saar)", "Neunkirchen"),
        clean_city = replace(clean_city, clean_city == "Homburg (Saar)", "Homburg"),
        clean_city = replace(clean_city, clean_city == "Kaulhausen", "Erkelenz"),
        clean_city = replace(clean_city, clean_city == "Kirchengel", "Sondershausen"),
        clean_city = replace(clean_city, clean_city == "Mauersberg", "Großrückerswalde"),
        clean_city = replace(clean_city, clean_city == "Riesenbeck", "Hörstel"),
        clean_city = replace(clean_city, clean_city == "Castrop", "Castrop-Rauxel"),
        clean_city = replace(clean_city, clean_city == "St. Hubert", "Kempen"),
        clean_city = replace(clean_city, clean_city == "Siegertsbrunn", "München"),
        clean_city = replace(clean_city, clean_city == "Leibolz", "Eiterfeld"),
        clean_city = replace(clean_city, clean_city == "Dessau", "Dessau-Roßlau"),
        clean_city = replace(clean_city, clean_city == "Merka", "Bautzen"),
```

```
        clean_city = replace(clean_city, clean_city == "Großeneder", "Höxter"),
        clean_city = replace(clean_city, clean_city == "Lingen", "Lingen (Ems)"),
        clean_city = replace(clean_city, clean_city == "Berzdorf", "Wesseling"),
        clean_city = replace(clean_city, clean_city == "Nunkirchen", "Wadern"),
        clean_city = replace(clean_city, clean_city == "Bernburg", "Bernburg (Saale)"),
        clean_city = replace(clean_city, clean_city == "Stolberg", "Stolberg (Rhld.)"),
        clean_city = replace(clean_city, clean_city == "Setterich", "Baesweiler"),
        clean_city = replace(clean_city, clean_city == "Gölriehenfeld", "Bockhorn"),
        clean_city = replace(clean_city, clean_city == "Wentorf", "Wentorf bei Hamburg"),
        clean_city = replace(clean_city, clean_city == "Wahlscheid", "Engelskirchen"),
        clean_city = replace(clean_city, clean_city == "Penzendorf", "Schwabach"),
        clean_city = replace(clean_city, clean_city == "Rheydt", "Mönchengladbach"),
        clean_city = replace(clean_city, clean_city == "Pippensen (heute Buxtehude)", "Buxtehude"),
        clean_city = replace(clean_city, clean_city == "Villingen", "Villingen-Schwenningen"),
        clean_city = replace(clean_city, clean_city == "Sohland", "Sohland a. d. Spree"),
        clean_city = replace(clean_city, clean_city == "Garching bei München", "Garching b.München"),
        clean_city = replace(clean_city, clean_city == "Zweenfurth bei Leipzig", "Borsdorf"),
        clean_city = replace(clean_city, clean_city == "Salmrohr", "Salmtal"),
        clean_city = replace(clean_city, clean_city == "Heiligenstadt", "Heilbad Heiligenstadt"),
        clean_city = replace(clean_city, clean_city == "Höhr", "Höhr-Grenzhausen"),
        clean_city = replace(clean_city, clean_city == "Brake", "Bielefeld"),
        clean_city = replace(clean_city, clean_city == "Heldenbergen", "Nidderau"),
        clean_city = replace(clean_city, clean_city == "Kölln", "Kölln-Reisiek"),
        clean_city = replace(clean_city, clean_city == "Gräfenroda", "Geratal"),
        clean_city = replace(clean_city, clean_city == "Süssen", "Süßen"),
        clean_city = replace(clean_city, clean_city == "Nienburg", "Nienburg (Weser)"),
        clean_city = replace(clean_city, clean_city == "Untersulmetingen", "Laupheim"),
        clean_city = replace(clean_city, clean_city == "Freiburg", "Freiburg im Breisgau"),
        clean_city = replace(clean_city, clean_city == "Mülheim", "Mülheim an der Ruhr"),
        clean_city = replace(clean_city, clean_city == "Laubusch", "Lauta"),
        clean_city = replace(clean_city, clean_city == "Barbecke", "Lengede"),
        clean_city = replace(clean_city, clean_city == "Jahnishausen", "Riesa"),
        clean_city = replace(clean_city, clean_city == "Altenheideck", "Heideck"),
        clean_city = replace(clean_city, clean_city == "Packebusch", "Kalbe (Milde)"),
        clean_city = replace(clean_city, clean_city == "St. Tönis", "Tönisvorst"),
        clean_city = replace(clean_city, clean_city == "Mossenberg", "Blomberg"),
        clean_city = replace(clean_city, clean_city == "Garbisdorf", "Göpfersdorf"),
        clean_city = replace(clean_city, clean_city == "Zella", "Zella-Mehlis"),
        clean_city = replace(clean_city, clean_city == "Warbeyen", "Kleve"),
        clean_city = replace(clean_city, clean_city == "Lohburg", "Bakum"),
        clean_city = replace(clean_city, clean_city == "Kehl am Rhein", "Kehl"),
        clean_city = replace(clean_city, clean_city == "Hornbostel", "Wietze"),
        clean_city = replace(clean_city, clean_city == "Wattenscheid", "Bochum"),
        clean_city = replace(clean_city, clean_city == "Niederwartha", "Dresden"),
        clean_city = replace(clean_city, clean_city == "Parberg", "Parsberg"),
        clean_city = replace(clean_city, clean_city == "Röthenbach bei St. Wolfgang", "Wendelstein"),
        clean_city = replace(clean_city, clean_city == "Wimmental", "Weinsberg"),
        clean_city = replace(clean_city, clean_city == "Dörnholthausen", "Sundern (Sauerland)"),
        clean_city = replace(clean_city, clean_city == "Kossa", "Laußig"),
        clean_city = replace(clean_city, clean_city == "Wildenhain", "Großenhain"),
        clean_city = replace(clean_city, clean_city == "Bad Gottleuba", "Bad Gottleuba-Berggießhübel")
        clean_city = replace(clean_city, clean_city == "Oberspay", "Spay"),
        clean_city = replace(clean_city, clean_city == "Hahnenknoop", "Loxstedt"),
```

```
        clean_city = replace(clean_city, clean_city == "Schirgiswalde", "Schirgiswalde-Kirschau"),
        clean_city = replace(clean_city, clean_city == "Niederlinxweiler", "St. Wendel"),
        clean_city = replace(clean_city, clean_city == "Oberrohr", "Ursberg"),
        clean_city = replace(clean_city, clean_city == "Schloß Zeil", "Leutkirch im Allgäu"),
        clean_city = replace(clean_city, clean_city == "Straußberg", "Sondershausen"),
        clean_city = replace(clean_city, clean_city == "Neustadt", "Neustadt an der Orla"),
        clean_city = replace(clean_city, clean_city == "Bernkastel", "Bernkastel-Kues"),
        clean_city = replace(clean_city, clean_city == "Heringen", "Heringen (Werra)"),
        clean_city = replace(clean_city, clean_city == "Prien", "Prien a.Chiemsee"),
        clean_city = replace(clean_city, clean_city == "Lintach", "Freudenberg"),
        clean_city = replace(clean_city, clean_city == "Obernburg", "Eisenbach (Hochschwarzwald)"),
        clean_city = replace(clean_city, clean_city == "Immenstadt", "Immenstadt i.Allgäu"),
        clean_city = replace(clean_city, clean_city == "Neudorf", "Diemelstadt"),
        clean_city = replace(clean_city, clean_city == "Imsum", "Geestland"),
        clean_city = replace(clean_city, clean_city == "Letmathe", "Iserlohn"),
        clean_city = replace(clean_city, clean_city == "Wölkau", "Leuna"),
        clean_city = replace(clean_city, clean_city == "Niedergrenzebach", "Schwalmstadt"),
        clean_city = replace(clean_city, clean_city == "Aumühle bei Hamburg", "Aumühle"),
        clean_city = replace(clean_city, clean_city == "Eickelborn", "Lippstadt"),
        clean_city = replace(clean_city, clean_city == "Pleinting", "Vilshofen an der Donau"),
        clean_city = replace(clean_city, clean_city == "Neumarkt", "Neumarkt i.d.OPf."),
        clean_city = replace(clean_city, clean_city == "Klosterneuendorf", "Gardelegen"),
        clean_city = replace(clean_city, clean_city == "Neuenburg", "Zetel"),
        clean_city = replace(clean_city, clean_city == "Neustadt a. d. Aisch", "Neustadt a.d.Aisch"),
        clean_city = replace(clean_city, clean_city == "Neu", "Lübstorf"),
        clean_city = replace(clean_city, clean_city == "Neuburg a. d. Donau", "Neuburg a.d.Donau"),
        clean_city = replace(clean_city, clean_city == "Altmugl", "Bad Neualbenreuth"),
        clean_city = replace(clean_city, clean_city == "Neustadt a.d. Waldnaab", "Neustadt a.d.Waldnaa
        clean_city = replace(clean_city, clean_city == "Krakow", "Krakow am See"),
        clean_city = replace(clean_city, clean_city == "Bischofferode", "Ellrich"),
        clean_city = replace(clean_city, clean_city == "Adorf", "Neukirchen/Erzgeb."),
        clean_city = replace(clean_city, clean_city == "Emmerich", "Emmerich am Rhein"),
        clean_city = replace(clean_city, clean_city == "Kiehl", "Kiel"),
        clean_city = replace(clean_city, clean_city == "Holthausen", "Meppen"),
        clean_city = replace(clean_city, clean_city == "Kausche", "Drebkau/Drjowk"),
        clean_city = replace(clean_city, clean_city == "Storkow", "Storkow (Mark)"),
        clean_city = replace(clean_city, clean_city == "Prieros", "Storkow (Mark)"),
        clean_city = replace(clean_city, clean_city == "Gemünden am Main", "Gemünden a.Main"),
        clean_city = replace(clean_city, clean_city == "Krumbach", "Krumbach (Schwaben)"),
        clean_city = replace(clean_city, clean_city == "Wulfen", "Dorsten"),
        clean_city = replace(clean_city, clean_city == "Spremberg", "Spremberg/Grodk"),
        clean_city = replace(clean_city, clean_city == "Denkwitz", "Bautzen"),
        clean_city = replace(clean_city, clean_city == "Heckershausen", "Ahnatal"),
        clean_city = replace(clean_city, clean_city == "Boeken", "Schwerin"),
        clean_city = replace(clean_city, clean_city == "Roßla", "Südharz"),
        clean_city = replace(clean_city, clean_city == "Augsdorf", "Gerbstedt"),
        clean_city = replace(clean_city, clean_city == "Lobenstein", "Bad Lobenstein"),
        clean_city = replace(clean_city, clean_city == "Urberach", "Rödermark"),
        clean_city = replace(clean_city, clean_city == "Koerbecke", "Möhnesee"),
        clean_city = replace(clean_city, clean_city == "Oberlucken", "Simbach"),
        clean_city = replace(clean_city, clean_city == "Wahlen", "Losheim am See"),
        clean_city = replace(clean_city, clean_city == "Lobsdorf", "St. Egidien"),
        clean_city = replace(clean_city, clean_city == "Stein bei Nürnberg", "Mittelfranken"),
```

```
        clean_city = replace(clean_city, clean_city == "Gablenz", "Gablenz / Jabłońc"),
        clean_city = replace(clean_city, clean_city == "Oberbubach", "Dingolfing"),
        clean_city = replace(clean_city, clean_city == "Lautenhausen", "Friedewald"),
        clean_city = replace(clean_city, clean_city == "Brackwede", "Bielefeld"),
        clean_city = replace(clean_city, clean_city == "Wilkau", "Wilkau-Haßlau"),
        clean_city = replace(clean_city, clean_city == "Großräschen", "Großräschen/Rań"),
        clean_city = replace(clean_city, clean_city == "Annaberg", "Annaberg-Buchholz"),
        clean_city = replace(clean_city, clean_city == "Sandershausen", "Niestetal"),
        clean_city = replace(clean_city, clean_city == "Quetzen", "Petershagen"),
        clean_city = replace(clean_city, clean_city == "Dinklar", "Schellerten"),
        clean_city = replace(clean_city, clean_city == "Gillrath", "Geilenkirchen"),
        clean_city = replace(clean_city, clean_city == "Hofheim", "Hofheim am Taunus"),
        clean_city = replace(clean_city, clean_city == "Wanne", "Herne"),
        clean_city = replace(clean_city, clean_city == "Collmen", "Colditz"),
        clean_city = replace(clean_city, clean_city == "Badersleben", "Huy"),
        clean_city = replace(clean_city, clean_city == "Reichau", "Boos (VGem)"),
        clean_city = replace(clean_city, clean_city == "Speinshard", "Speinshart"),
        clean_city = replace(clean_city, clean_city == "Geversmühlen", "Grevesmühlen"),
        clean_city = replace(clean_city, clean_city == "Elsterwerder", "Elsterwerda"),
        clean_city = replace(clean_city, clean_city == "Hirsau", "Calw"),
        clean_city = replace(clean_city, clean_city == "Rottenburg", "Rottenburg am Neckar"),
        clean_city = replace(clean_city, clean_city == "Krothmaißling", "Cham"),
        clean_city = replace(clean_city, clean_city == "Mengeringhausen", "Bad Arolsen"),
        clean_city = replace(clean_city, clean_city == "Lehrte bei Hannover", "Lehrte"),
        clean_city = replace(clean_city, clean_city == "Saalfeld", "Saalfeld/Saale"),
        clean_city = replace(clean_city, clean_city == "Graal", "Graal-Müritz"),
        clean_city = replace(clean_city, clean_city == "Garmisch", "Garmisch-Partenkirchen"),
        clean_city = replace(clean_city, clean_city == "Kreuzebra", "Dingelstädt"),
        clean_city = replace(clean_city, clean_city == "Bad Frankenhausen", "Bad Frankenhausen/Kyffhäus
        clean_city = replace(clean_city, clean_city == "Mehrstedt", "Nottertal-Heilinger Höhen"),
        clean_city = replace(clean_city, clean_city == "Töging", "Töging a.Inn"),
        clean_city = replace(clean_city, clean_city == "Heyrothsberge", "Biederitz"),
        clean_city = replace(clean_city, clean_city == "Kettwig", "Essen"),
        clean_city = replace(clean_city, clean_city == "Feiburg", "Freiburg im Breisgau"),
        clean_city = replace(clean_city, clean_city == "Gaschwitz Kreis Leipzig", "Markkleeberg"),
        clean_city = replace(clean_city, clean_city == "Weißenburg", "Weißenburg i.Bay."),
        clean_city = replace(clean_city, clean_city == "Weißandt", "Südliches Anhalt"),
        clean_city = replace(clean_city, clean_city == "Stedten", "Seegebiet Mansfelder Land"),
        clean_city = replace(clean_city, clean_city == "Bovenden bei Göttingen", "Bovenden"),
        clean_city = replace(clean_city, clean_city == "Frohnstetten", "Stetten am kalten Markt"),
        clean_city = replace(clean_city, clean_city == "Bork bei Münster", "Selm"),
        clean_city = replace(clean_city, clean_city == "Bonndorf", "Bonndorf im Schwarzwald"),
        clean_city = replace(clean_city, clean_city == "Verden an der Aller", "Verden"),
        clean_city = replace(clean_city, clean_city == "Morlautern", "Kaiserslautern"),
        clean_city = replace(clean_city, clean_city == "Rauenzell", "Herrieden"),
        clean_city = replace(clean_city, clean_city == "Großengottern", "Unstrut-Hainich"),
        clean_city = replace(clean_city, clean_city == "Tanndorf", "Colditz"),
        clean_city = replace(clean_city, clean_city == "Münden", "Lichtenfels"),
        clean_city = replace(clean_city, clean_city == "Lauf a. d. Pegnitz", "Lauf a.d.Pegnitz"),
        clean_city = replace(clean_city, clean_city == "Dillingen a. d. Donau", "Dillingen a.d.Donau")
        clean_city = replace(clean_city, clean_city == "Kamenz", "Kamenz / Kamjenc"),
        clean_city = replace(clean_city, clean_city == "Rössing", "Nordstemmen"),
        clean_city = replace(clean_city, clean_city == "Seelscheid", "Neunkirchen-Seelscheid"),
```

```
        clean_city = replace(clean_city, clean_city == "Geinsheim", "Neustadt an der Weinstraße"),
        clean_city = replace(clean_city, clean_city == "Erkenrechtsweiler", "Erkenbrechtsweiler"),
        clean_city = replace(clean_city, clean_city == "Ohrbeck", "Hasbergen"),
        clean_city = replace(clean_city, clean_city == "Möhnsee", "Möhnesee"),
        clean_city = replace(clean_city, clean_city == "Stirpe", "Erwitte"),
        clean_city = replace(clean_city, clean_city == "Rielasingen", "Rielasingen-Worblingen"),
        clean_city = replace(clean_city, clean_city == "Dannenberg", "Dannenberg (Elbe)"),
        clean_city = replace(clean_city, clean_city == "Helmern", "Bad Wünnenberg"),
        clean_city = replace(clean_city, clean_city == "Oberhochstatt", "Weißenburg i.Bay."),
        clean_city = replace(clean_city, clean_city == "Bardenberg", "Würselen"),
        clean_city = replace(clean_city, clean_city == "Gadderbaum", "Bielefeld"),
        clean_city = replace(clean_city, clean_city == "Zollbrück", "Kloster Veßra"),
        clean_city = replace(clean_city, clean_city == "Schwedt", "Schwedt/Oder"),
        clean_city = replace(clean_city, clean_city == "Altenoythe", "Friesoythe"),
        clean_city = replace(clean_city, clean_city == "Nahne", "Osnabrück"),
        clean_city = replace(clean_city, clean_city == "Schlema", "Aue-Bad Schlema"),
        clean_city = replace(clean_city, clean_city == "Freekenhorst", "Warendorf"),
        clean_city = replace(clean_city, clean_city == "Prisser", "Dannenberg (Elbe)"),
        clean_city = replace(clean_city, clean_city == "Ribnitz", "Ribnitz-Damgarten"),
        clean_city = replace(clean_city, clean_city == "Großkorbetha", "Weißenfels"),
        clean_city = replace(clean_city, clean_city == "Schwarzberg", "Wernberg-Köblitz"),
        clean_city = replace(clean_city, clean_city == "Krusel", "Kusel"),
        clean_city = replace(clean_city, clean_city == "Langholt", "Rhauderfehn"),
        clean_city = replace(clean_city, clean_city == "Hangard", "Neunkirchen"),
        clean_city = replace(clean_city, clean_city == "Hilten", "Neuenhaus"),
        clean_city = replace(clean_city, clean_city == "Belzig", "Bad Belzig"),
        clean_city = replace(clean_city, clean_city == "Wunstorf OT Idensen", "Wunstorf"),
        clean_city = replace(clean_city, clean_city == "Burgsteinfurt", "Steinfurt"),
        clean_city = replace(clean_city, clean_city == "Großburgwedel", "Burgwedel"),
        clean_city = replace(clean_city, clean_city == "Lauenburg/Elbe", "Lauenburg/ Elbe"),
        clean_city = replace(clean_city, clean_city == "Wiedenbrück", "Rheda-Wiedenbrück"),
        clean_city = replace(clean_city, clean_city == "Kamenz", "Kamenz / Kamjenc"),
        clean_city = replace(clean_city, clean_city == "Meerane (Sachsen)", "Meerane"),
        clean_city = replace(clean_city, clean_city == "Mittelberg", "Oy-Mittelberg"),
        clean_city = replace(clean_city, clean_city == "Johannisberg", "Geisenheim"),
        clean_city = replace(clean_city, clean_city == "Vorsfelde", "Wolfsburg"),
        clean_city = replace(clean_city, clean_city == "Herzberg/Elster", "Herzberg (Elster)"),
        clean_city = replace(clean_city, clean_city == "Vetschau", "Vetschau/Spreewald / Wětošow/Błota
        clean_city = replace(clean_city, clean_city == "Jugenheim", "Seeheim-Jugenheim"),
        clean_city = replace(clean_city, clean_city == "Schwäblishausen", "Pfullendorf"),
        clean_city = replace(clean_city, clean_city == "Ellenbach", "Floß"),
        clean_city = replace(clean_city, clean_city == "Andernach/Rhein", "Andernach"),
        clean_city = replace(clean_city, clean_city == "Nienburg  (Weser)", "Nienburg (Weser)"),
        clean_city = replace(clean_city, clean_city == "Arnsgereuth", "Saalfeld/Saale"),
        clean_city = replace(clean_city, clean_city == "Dreiborn bei Schleiden", "Schleiden"),
        clean_city = replace(clean_city, clean_city == "Schwandorf/Bayern", "Schwandorf"),
        clean_city = replace(clean_city, clean_city == "Bad Godesberg", "Bonn"),
        clean_city = replace(clean_city, clean_city == "Büdingen/Oberhessen", "Büdingen"),
        clean_city = replace(clean_city, clean_city == "Marburg/Lahn", "Marburg"),
        clean_city = replace(clean_city, clean_city == "Rulle", "Wallenhorst"),
        clean_city = replace(clean_city, clean_city == "Haren an der Ems", "Haren (Ems)"),
        clean_city = replace(clean_city, clean_city == "Ulm/Donau", "Ulm"),
        clean_city = replace(clean_city, clean_city == "Hagen (Westfalen)", "Hagen am Teutoburger Wald
```

```
        clean_city = replace(clean_city, clean_city == "Luhdorf", "Winsen (Luhe)"),
        clean_city = replace(clean_city, clean_city == "Drevenack", "Hünxe"),
        clean_city = replace(clean_city, clean_city == "Säckingen", "Bad Säckingen"),
        clean_city = replace(clean_city, clean_city == "Bad Griesbach i. Rottal", "Bad Griesbach i.Rot
        clean_city = replace(clean_city, clean_city == "Porz am Rhein", "Köln"),
        clean_city = replace(clean_city, clean_city == "Leinefelde", "Leinefelde-Worbis"),
        clean_city = replace(clean_city, clean_city == "Aken/Elbe", "Aken (Elbe)"),
        clean_city = replace(clean_city, clean_city == "Bauhaus", "Nentershausen"),
        clean_city = replace(clean_city, clean_city == "Saulgau", "Bad Saulgau"),
        clean_city = replace(clean_city, clean_city == "Meißen/Sachsen", "Meißen"),
        clean_city = replace(clean_city, clean_city == "Döbern", "Döbern-Land"),
        clean_city = replace(clean_city, clean_city == "Berleburg", "Bad Berleburg"),
        clean_city = replace(clean_city, clean_city == "Ober", "Wöllstadt"),
        clean_city = replace(clean_city, clean_city == "Bendorf am Rhein", "Bendorf"),
        clean_city = replace(clean_city, clean_city == "Welschen", "Kirchhundem"),
        clean_city = replace(clean_city, clean_city == "Ehlerstorf", "Wangels"),
        clean_city = replace(clean_city, clean_city == "Münster (Westf.)", "Münster"),
        clean_city = replace(clean_city, clean_city == "Weißwasser", "Weißwasser/O.L."),
        clean_city = replace(clean_city, clean_city == "Mallersdorf", "Mallersdorf-Pfaffenberg"),
        clean_city = replace(clean_city, clean_city == "Räckelwitz", "Räckelwitz / Worklecy"),
        clean_city = replace(clean_city, clean_city == "Hoyerswerda", "Hoyerswerda / Wojerecy"),
        clean_city = replace(clean_city, clean_city == "Heimboldshausen", "Philippsthal (Werra)"),
        clean_city = replace(clean_city, clean_city == "Limburg a.d. Lahn", "Limburg a. d. Lahn"),
        clean_city = replace(clean_city, clean_city == "Stolberg (Harz)", "Südharz"),
        clean_city = replace(clean_city, clean_city == "Dollerupholz", "Westerholz"),
        clean_city = replace(clean_city, clean_city == "Verden (Niedersachsen)", "Verden"),
        clean_city = replace(clean_city, clean_city == "Rhede (Westf.)", "Rhede"),
        clean_city = replace(clean_city, clean_city == "Herborn (Dillkreis)", "Herborn"),
        clean_city = replace(clean_city, clean_city == "Mühldorf a. Inn", "Mühldorf a.Inn"),
        clean_city = replace(clean_city, birthPlace == "Groß-Ottersleben", "Magdeburg"),
        clean_city = replace(clean_city, birthPlace == "Groß-Umstadt", "Groß-Umstadt"),
        clean_city = replace(clean_city, birthPlace == "Groß-Gerau", "Groß-Gerau"),
        clean_city = replace(clean_city, politicianId == 11000864, "Dresden"),
        clean_city = replace(clean_city, politicianId == 11002028, "Schwarzheide"),
        clean_city = replace(clean_city, politicianId == 11002105, "Gersdorf"),
        clean_city = replace(clean_city, politicianId == 11002736, "Grabs")) %>%
mutate(not_germany = case_when(clean_city == "Breslau" ~ 1,
                               clean_city == "Teheran/Iran" ~ 1,
                               clean_city == "Preßburg" ~ 1,
                               clean_city == "Gdynia" ~ 1,
                               clean_city == "Danzig" ~ 1,
                               clean_city == "Teschen" ~ 1,
                               clean_city == "Lauenburg" ~ 1,
                               clean_city == "Trautenau" ~ 1,
                               clean_city == "Elbing" ~ 1,
                               clean_city == "Graz" ~ 1,
                               clean_city == "Königshütte" ~ 1,
                               clean_city == "Hirschberg am See" ~ 1,
                               clean_city == "Gurschdorf" ~ 1,
                               clean_city == "Marktlangendorf" ~ 1,
                               clean_city == "Thomigsdorf" ~ 1,
                               clean_city == "Raase" ~ 1,
                               clean_city == "Kattowitz" ~ 1,
```

```
                        clean_city == "Gleiwitz" ~ 1,
                        clean_city == "Beuthen" ~ 1,
                        clean_city == "Kirchwalde" ~ 1,
                        clean_city == "Gräfenort" ~ 1,
                        clean_city == "Urbanstreben" ~ 1,
                        clean_city == "Neusalz" ~ 1,
                        clean_city == "Bankau" ~ 1,
                        clean_city == "Gut Quickendorf" ~ 1,
                        clean_city == "Sagan" ~ 1,
                        clean_city == "Bunzlau" ~ 1,
                        clean_city == "Lauban" ~ 1,
                        clean_city == "Liegnitz" ~ 1,
                        clean_city == "Schweidnitz" ~ 1,
                        clean_city == "Schönlanke" ~ 1,
                        clean_city == "Stettin" ~ 1,
                        clean_city == "Kolberg" ~ 1,
                        clean_city == "Jasenitz" ~ 1,
                        clean_city == "Freienwalde" ~ 1,
                        clean_city == "Stolp" ~ 1,
                        clean_city == "Altbeelitz" ~ 1,
                        clean_city == "Jagertow" ~ 1,
                        clean_city == "Schillen" ~ 1,
                        clean_city == "Rückers" ~ 1,
                        clean_city == "Eger" ~ 1,
                        clean_city == "Goldap" ~ 1,
                        clean_city == "Königsberg" ~ 1,
                        clean_city == "Schillwen" ~ 1,
                        clean_city == "Gerdauen" ~ 1,
                        clean_city == "Engelshöhe" ~ 1,
                        clean_city == "Insterburg" ~ 1,
                        clean_city == "Treuburg" ~ 1,
                        clean_city == "Weepers" ~ 1,
                        clean_city == "Saleschen" ~ 1,
                        clean_city == "Woopen" ~ 1,
                        clean_city == "Graudenz" ~ 1,
                        clean_city == "Mährisch" ~ 1,
                        clean_city == "Klum" ~ 1,
                        clean_city == "Eigenheim" ~ 1,
                        clean_city == "Varto" ~ 1,
                        clean_city == "Rütli" ~ 1,
                        clean_city == "Bielitz" ~ 1,
                        clean_city == "Mewe" ~ 1,
                        clean_city == "Neudamm" ~ 1,
                        clean_city == "Oppeln" ~ 1,
                        clean_city == "Schwaz" ~ 1,
                        clean_city == "Tevel" ~ 1,
                        clean_city == "Lechnitz" ~ 1,
                        clean_city == "Kopenhagen" ~ 1,
                        clean_city == "Prag" ~ 1,
                        clean_city == "Swinemünde" ~ 1,
                        clean_city == "Teplitz" ~ 1,
                        clean_city == "Plan bei Marienbad" ~ 1,
                        clean_city == "Moskau" ~ 1,
```

```
                                clean_city == "Cleveland" ~ 1,
                                clean_city == "Marsassoum" ~ 1,
                                clean_city == "Kelkit" ~ 1,
                                clean_city == "Yukaribalakur" ~ 1,
                                clean_city == "Nibbixwoud" ~ 1,
                                clean_city == "Adana" ~ 1,
                                clean_city == "Kastek" ~ 1,
                                clean_city == "Pitesti" ~ 1,
                                clean_city == "Uspenska" ~ 1,
                                clean_city == "Sabelowka" ~ 1,
                                clean_city == "Gobabis" ~ 1,
                                clean_city == "Rahmel" ~ 1,
                                clean_city == "Deutsch" ~ 1,
                                clean_city == "Korntal" ~ 1,
                                clean_city == "Damm" ~ 1,
                                clean_city == "Brüssel" ~ 1,
                                clean_city == "Most" ~ 1,
                                clean_city == "Agnetheln" ~ 1,
                                clean_city == "Craiova" ~ 1,
                                clean_city == "Skalica" ~ 1,
                                clean_city == "Arnswalde" ~ 1,
                                clean_city == "Gablonz an der Neiße" ~ 1,
                                clean_city == "Grabs" ~ 1,
                                TRUE ~ 0))
```

Merging Data Frames

In the next step, the preprocessed `politicians`, `cities`, `factions`, and `speeches` data frames are merged together to form `full_data`.

```
#merge politiciians, cities, and speeches dfs
politicians_cities_merged <- left_join(politicians_subset, cities_amended, by = "clean_city") %>%
  mutate(east_west = replace(east_west, born_germany == 0 | not_germany == 1, "Born Abroad"))

speeches_politicians_merged <- left_join(speeches_subset, politicians_cities_merged, by = "politicianId

factions_premerge <- factions %>%
  rename(factionId = id,
         party_abb = abbreviation,
         party_name = fullName)

full_data <- left_join(speeches_politicians_merged, factions_premerge, by = "factionId")
```

Further Preprocessing

The next code chunk adds several variables to the data frame and further subsets the data to only include relevant cases. For instance, speeches given by MdB with no party affiliation are dropped, as these are often given by external individuals addressing the *Bundestag*. Moreover, in addition to the `party` variable, `partygroup` indicates whether a speaker represents one of the more established, mainstream parties in German politics, the right-wing extremist AfD, or the far-left PDS/*Die Linke*. The reasons for this grouping are explained in the thesis itself. Finally, `born_system` indicates which German state a speaker was born into,

while `born_gdr` simplifies this a bit and indicates whether an MdB was born in the GDR, in East Germany but before or after the founding of the GDR, or elsewhere.

```
#add layered party, electoral term, and birthplace variables
full_data <- full_data %>%
  mutate(party = case_when(party_abb == "PDS" | party_abb == "DIE LINKE." ~ "PDS/Die Linke",
                           party_abb == "CDU/CSU" ~ "CDU/CSU",
                           party_abb == "SPD" ~ "SPD",
                           party_abb == "FDP" ~ "FDP",
                           party_abb == "Grüne" ~ "Bündnis 90/Die Grünen",
                           party_abb == "AfD" ~ "AfD",
                           party_abb == "not found" | party_abb == "Fraktionslos" | party_abb == "Gast"
  filter(party != "Other") %>%
  mutate(partygroup = case_when(party == "CDU/CSU" | party == "FDP" | party == "SPD" | party == "Bündnis
                                party == "AfD" ~ "AfD",
                                party == "PDS/Die Linke" ~ "PDS/Die Linke"),
         admin = case_when(electoralTerm == 11 | electoralTerm == 12 | electoralTerm == 13 ~ "Kohl",
                           electoralTerm == 14 | electoralTerm == 15 ~ "Schröder",
                           electoralTerm == 16 | electoralTerm == 17 | electoralTerm == 18 | electoralTe
         born_system = case_when((birthDate >= as.Date('1990-10-03')) & (east_west == "East" | east_west
                                 ((birthDate >= as.Date('1945-05-08')) & (east_west == "Born Abroad") |
                                 (birthDate < as.Date('1990-10-03') & birthDate >= as.Date('1945-05-08'
                                 (birthDate < as.Date('1990-10-03') & birthDate >= as.Date('1945-05-08'
                                 (birthDate < as.Date('1990-10-03') & birthDate >= as.Date('1945-05-08'
                                 (birthDate < as.Date('1945-05-08') & birthDate >= as.Date('1933-01-30'
                                 (birthDate < as.Date('1933-01-30') & birthDate >= as.Date('1918-11-09'
                                 (birthDate < as.Date('1918-11-09')) & birthCountry == "Deutschland" ~
         born_gdr = case_when(birthDate < as.Date('1990-10-03') & birthDate >= as.Date('1945-05-08') &
                              (birthDate > as.Date('1990-10-03') | birthDate <= as.Date('1945-05-08')) &
                              TRUE ~ "Elsewhere"))

#code year variable
full_data$year <- as.numeric(gsub('\\-.*', '', full_data$date))

#remove 11th Bundestag observations due to low number of observations
full_data <- full_data %>%
  filter(electoralTerm != 11)
```

Save to CSV

Finally, this last code chunk allows users to save the resulting cleaned and preprocessed data frame, `full_data`, to be saved locally.

```
#uncomment the following line to save df as a csv file
#write.csv(full_data, "data/full_data.csv")
```