Cleaning Darkorchid4 Data

Eric Leinweber

```
#setwd("/Users/ericleinweber/Desktop/PLCY 715 GitHub Repository/final-team-projects-darkorchid/DarkOrchid/")
```

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✔ ggplot2 3.4.0     ✔ purrr   1.0.1
## ✔ tibble  3.1.8     ✔ dplyr   1.1.0
## ✔ tidyr   1.3.0     ✔ stringr 1.5.0
## ✔ readr   2.1.4     ✔ forcats 0.5.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(moderndive)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(janitor)
```

```
##
## Attaching package: 'janitor'
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(dplyr)
library(stringr)
library(tidyr)
```

```
#reading in the data
asthma_outcomes <- read_csv("/Users/ericleinweber/Desktop/PLCY 715 GitHub Repository/final-team-projects-darkorchid/DarkOrchid/DATA/Raw_Unprocessed/Health Outcomes/Asthma_Mortality_UNCLEAN.csv")
```

```
## Rows: 3196 Columns: 11
## ── Column specification ────────────────────────────────────────────────────────
────────────────────────────────
## Delimiter: ","
## chr (10): Location, Mortality Rate, 1980*, Mortality Rate, 1985*, Mortality ...
## dbl  (1): FIPS
##
## 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.
```

```
resp_outcomes <- read_csv("/Users/ericleinweber/Desktop/PLCY 715 GitHub Repository/final-team-projects-darkorchid/DarkOrchid/DATA/Raw_Unprocessed/Health Outcomes/ChronicRespiratoryDiseases_Mortality_UNCLEAN.csv")
```

```
## Rows: 3196 Columns: 11
## ── Column specification ────────────────────────────────────────────────────────
────────────────────────────────
## Delimiter: ","
## chr (10): Location, Mortality Rate, 1980*, Mortality Rate, 1985*, Mortality ...
## dbl  (1): FIPS
##
## 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.
```

```
cwp_outcomes <- read_csv("/Users/ericleinweber/Desktop/PLCY 715 GitHub Repository/final-team-projects-darkorchid/DarkOrchid/DATA/Raw_Unprocessed/Health Outcomes/CoalWorkersPneumoconiosis_Mortality_UNCLEAN.csv")
```

```
## Rows: 3196 Columns: 11
## ── Column specification ────────────────────────────────────────────────────────
────────────────────────────────
## Delimiter: ","
## chr (10): Location, Mortality Rate, 1980*, Mortality Rate, 1985*, Mortality ...
## dbl  (1): FIPS
##
## 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.
```

```
lung_outcomes <- read_csv("/Users/ericleinweber/Desktop/PLCY 715 GitHub Repository/final-team-projects-darkorchid/DarkOrchid/DATA/Raw_Unprocessed/Health Outcomes/LungDisease_Mortality_UNCLEAN.csv")
```

```
## Rows: 3196 Columns: 11
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (10): Location, Mortality Rate, 1980*, Mortality Rate, 1985*, Mortality ...
## dbl  (1): FIPS
##
## 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.
```

```
#selecting and renaming mortality rate rows for clarity
asthma_outcomes <- asthma_outcomes  %>% select(c("Location", "FIPS", `Mortality Rate, 2014*`, `% Change in Mortality Rate, 1980-2014`)) %>% mut
ate("mortality_ast_2014" = `Mortality Rate, 2014*`, "perc_mortality_change_ast" = `% Change in Mortality Rate, 1980-2014`) %>% select(-c(`% Chang
e in Mortality Rate, 1980-2014`, `Mortality Rate, 2014*`))

cwp_outcomes <- cwp_outcomes  %>% select(c("Location", "FIPS", `Mortality Rate, 2014*`, `% Change in Mortality Rate, 1980-2014`)) %>% mutate("m
ortality_cwp_2014" = `Mortality Rate, 2014*`, "perc_mortality_change_cwp" = `% Change in Mortality Rate, 1980-2014`) %>% select(-c(`% Change in
Mortality Rate, 1980-2014`, `Mortality Rate, 2014*`))

lung_outcomes <- lung_outcomes  %>% select(c("Location", "FIPS", `Mortality Rate, 2014*`, `% Change in Mortality Rate, 1980-2014`)) %>% mutate
("mortality_lung_2014" = `Mortality Rate, 2014*`, "perc_mortality_change_lung" = `% Change in Mortality Rate, 1980-2014`) %>% select(-c(`% Chan
ge in Mortality Rate, 1980-2014`, `Mortality Rate, 2014*`))

resp_outcomes <- resp_outcomes  %>% select(c("Location", "FIPS", `Mortality Rate, 2014*`, `% Change in Mortality Rate, 1980-2014`)) %>% mutate
("mortality_resp_2014" = `Mortality Rate, 2014*`, "perc_mortality_change_resp" = `% Change in Mortality Rate, 1980-2014`) %>% select(-c(`% Chang
e in Mortality Rate, 1980-2014`, `Mortality Rate, 2014*`))

#merging all health outcomes data into one, complete dataset
merge1 <- merge(asthma_outcomes, cwp_outcomes, by = c("FIPS", "Location"))
merge2 <- merge(merge1, lung_outcomes, by = c("FIPS", "Location"))
health_outcomes <- merge(merge2, resp_outcomes, by = c("FIPS", "Location"))

#using stringr to remove ranges in mortality rate variables
health_outcomes$mortality_ast_2014 <- stringr::str_extract(health_outcomes$mortality_ast, "^.{5}")
health_outcomes$mortality_resp_2014 <- stringr::str_extract(health_outcomes$mortality_resp, "^.{5}")
health_outcomes$mortality_lung_2014 <- stringr::str_extract(health_outcomes$mortality_lung, "^.{5}")
health_outcomes$mortality_cwp_2014 <- stringr::str_extract(health_outcomes$mortality_cwp, "^.{5}")

#using stringr to remove ranges in percent mortality rate variables
health_outcomes$perc_mortality_change_ast <- stringr::str_extract(health_outcomes$perc_mortality_change_ast, "^.{6}")
health_outcomes$perc_mortality_change_resp <- stringr::str_extract(health_outcomes$perc_mortality_change_resp, "^.{6}")
health_outcomes$perc_mortality_change_cwp <- stringr::str_extract(health_outcomes$perc_mortality_change_cwp, "^.{6}")
health_outcomes$perc_mortality_change_lung <- stringr::str_extract(health_outcomes$perc_mortality_change_lung, "^.{6}")

#using stringr to remove remaining parentheses in percent mortality rate variables
health_outcomes$perc_mortality_change_ast<- gsub("[()]", "", health_outcomes$perc_mortality_change_ast)
health_outcomes$perc_mortality_change_resp<- gsub("[()]", "", health_outcomes$perc_mortality_change_resp)
health_outcomes$perc_mortality_change_cwp<- gsub("[()]", "", health_outcomes$perc_mortality_change_cwp)
health_outcomes$perc_mortality_change_lung<- gsub("[()]", "", health_outcomes$perc_mortality_change_lung)

#changing all health outcome variables to numeric, now that there are no parentheses and range data
health_outcomes$mortality_ast_2014 <- as.numeric(health_outcomes$mortality_ast_2014)
health_outcomes$mortality_resp_2014 <- as.numeric(health_outcomes$mortality_resp_2014)
health_outcomes$mortality_lung_2014 <- as.numeric(health_outcomes$mortality_lung_2014)
health_outcomes$mortality_cwp_2014 <- as.numeric(health_outcomes$mortality_cwp_2014)
health_outcomes$perc_mortality_change_ast <- as.numeric(health_outcomes$perc_mortality_change_ast)
health_outcomes$perc_mortality_change_resp <- as.numeric(health_outcomes$perc_mortality_change_resp)
health_outcomes$perc_mortality_change_cwp <- as.numeric(health_outcomes$perc_mortality_change_cwp)
health_outcomes$perc_mortality_change_lung <- as.numeric(health_outcomes$perc_mortality_change_lung)

#removing "Parish," "County," and replacing "Saint" with "St." for all county names
```

```
health_outcomes <- health_outcomes %>%
  separate(Location, into = c("County", "State"), sep = ", ", remove = FALSE) %>% select(-c("Location"))
```

```
## Warning: Expected 2 pieces. Missing pieces filled with `NA` in 53 rows [1, 2, 56, 75,
## 143, 303, 309, 354, 457, 550, 650, 651, 765, 900, 972, 989, 1014, 1029, 1113,
## 1201, ...].
```

```
health_outcomes$County <- str_replace(health_outcomes$County, " County", "")

health_outcomes <- health_outcomes %>% mutate(County = str_replace(County, " Parish", ""))

health_outcomes <- health_outcomes %>% mutate(County = str_replace(health_outcomes$County, "Saint ", "St. "))

#removing NA's created from separating into state/county... these NA's represent cumulative state data for each state/the US, which is not needed for our a
nalysis
health_outcomes <- health_outcomes %>%
  filter(!is.na(State))

#creating cleaned CSV into shared folder
write_csv(health_outcomes, "/Users/ericleinweber/Desktop/PLCY 715 GitHub Repository/final-team-projects-darkorchid/DarkOrchid/DATA/Clean/health_
outcomes_CLEAN.csv")
```

```
#reading in the data
aqi_2014 <- read_csv("/Users/ericleinweber/Desktop/PLCY 715 GitHub Repository/final-team-projects-darkorchid/DarkOrchid/DATA/Raw_Unprocessed/a
nnual_aqi_by_county_2014.csv")
```

```
## Rows: 1036 Columns: 18
## ── Column specification ────────────────────────────────────────────────────────
──────────────────────────
## Delimiter: ","
## chr  (2): State, County
## dbl (16): Year, Days with AQI, Good Days, Moderate Days, Unhealthy for Sensi...
##
## 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.
```

```
#this data is already cleaned, so creating a new CSV into the shared folder
write_csv(aqi_2014, "/Users/ericleinweber/Desktop/PLCY 715 GitHub Repository/final-team-projects-darkorchid/DarkOrchid/DATA/Clean/aqi_2014_CLE
AN.csv")
```

```
coal_plants <- read_csv("/Users/ericleinweber/Desktop/PLCY 715 GitHub Repository/final-team-projects-darkorchid/DarkOrchid/DATA/Raw_Unprocesse
d/CoalPlants.csv")
```

```
## Rows: 13491 Columns: 37
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (27): Tracker ID, TrackerLOC, ParentID, Wiki page, Country, Subnational ...
## dbl  (8): Capacity (MW), RETIRED, Planned Retire, Latitude, Longitude, Annua...
## num  (2): Heat rate (Btu per kWh), Emission factor (kg of CO2 per TJ)
##
## 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.
```

```
#filtering for just the United States, as this dataset contains infornmation on the entire world
coal_plants <- coal_plants %>% filter(Country == "United States")

#this data is already cleaned, as we only need the latitude/longitude of each coal plant for the purposes of this data
write_csv(coal_plants,  "/Users/ericleinweber/Desktop/PLCY 715 GitHub Repository/final-team-projects-darkorchid/DarkOrchid/DATA/Clean/coal_plants_CLEAN")
```

```
smoking <- read.csv("/Users/ericleinweber/Desktop/PLCY 715 GitHub Repository/final-team-projects-darkorchid/DarkOrchid/DATA/Raw_Unprocessed/smoking.csv")

#Deleting the first row
smoking <- smoking[-c(1), ]

#Clean names
smoking <- clean_names(smoking)
names(smoking)
```

```
##  [1] "geography_type_description" "geography_name"
##  [3] "sits_in_state"           "geo_id"
##  [5] "formatted_geo_id"        "current_smokers"
##  [7] "data_time_period"        "geographic_vintage"
##  [9] "data_source"             "selected_location"
```

```
#Removing unnecessary columns
smoking <- smoking %>% select(-one_of('geography_type_description', 'data_time_period', 'geographic_vintage', 'data_source', 'selected_location', 'formatted_geo_id'))

smoking <- smoking %>% rename_at('sits_in_state', ~'state')
smoking <- smoking %>% rename_at('geography_name', ~'county')


#Chaning current_smokers to be numeric
smoking <- smoking %>% mutate(current_smokers = as.numeric(current_smokers))
```

```
## Warning: There was 1 warning in `mutate()`.
## ℹ In argument: `current_smokers = as.numeric(current_smokers)`.
## Caused by warning:
## ! NAs introduced by coercion
```

```
#creating cleaned CSV into shared folder
write.csv(smoking, "/Users/ericleinweber/Desktop/PLCY 715 GitHub Repository/final-team-projects-darkorchid/DarkOrchid/DATA/Clean/smoking_CLEAN.csv", row.names=FALSE)
```