generalized_data

Madison Thompson

2025-12-09

Cleaning the datasets for incidence and mortality. - needs to be pivoted - needs to be joined and verified

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.1     v stringr   1.5.2
## v ggplot2   4.0.0     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(dplyr)
library(readr)

# load datasets
incidence_data <- read_csv("nc_cancer_incidence.csv")
```

```
## Rows: 101 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): County
## dbl (3): ColorectalRate_I, LungRate_I, BreastRate_I
##
## 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.
```

```
mortality_data <- read_csv("nc_cancer_mortality.csv")
```

```
## Rows: 101 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): County
## dbl (3): ColorectalRate_M, LungRate_M, BreastRate_M
##
## 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.
```

```r
# pivot data to match county-column layout of health data (cleaned)
long_i <- incidence_data %>%
  pivot_longer(
    cols = ends_with("_I"),
    names_to = "Metric",
    values_to = "Rate"
  )

# repeat for m
long_m <- mortality_data %>%
  pivot_longer(
    cols = ends_with("_M"),
    names_to = "Metric",
    values_to = "Rate"
  )

# peek data
#head(long_i)
#head(long_m)

# join data
combined_data <- bind_rows(long_i, long_m)

#head(combined_data)

# need to pivot combined data
combined_data <- combined_data %>%
  pivot_wider(
    names_from = Metric,
    values_from = Rate
  )

head(combined_data)
```

```
## # A tibble: 6 x 7
##   County    ColorectalRate_I LungRate_I BreastRate_I ColorectalRate_M LungRate_M
##   <chr>                <dbl>      <dbl>        <dbl>             <dbl>      <dbl>
## 1 NC                    34.7       58.6         175.              12.8       36.8
## 2 Alamance              40.1       64.7         183.              13.1       43.5
## 3 Alexander             38.8       65.8         177.              14.3       43.7
## 4 Alleghany             48         74.1         192.              NA         37.4
## 5 Anson                 47.8       68.4         204.              20.2       43.9
## 6 Ashe                  34.2       73           209.               8.8       36.5
## # i 1 more variable: BreastRate_M <dbl>
```

```r
# save data in case of future FUBAR
write_csv(combined_data, "nc_cancer_incidence_mortality_long.csv")


# join to cleaned health data
health_data <- read_csv("nc_county_health_data_2021_zscores.csv")
```

```
## Rows: 1000 Columns: 5
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): Metric, County, Value
## dbl (2): Value_Clean, Z_Score
##
## 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.
```

```
head(health_data)
```

```
## # A tibble: 6 x 5
##   Metric     County    Value_Clean Z_Score Value
##   <chr>      <chr>           <dbl>   <dbl> <chr>
## 1 Population Alamance       169509   0.380 <NA>
## 2 Population Alexander       37497  -0.396 <NA>
## 3 Population Alleghany       11137  -0.551 <NA>
## 4 Population Anson           24446  -0.473 <NA>
## 5 Population Ashe            27203  -0.457 <NA>
## 6 Population Avery           17557  -0.513 <NA>
```

```
final_data <- combined_data %>%
  left_join(health_data, by = "County")

# show final data
head(final_data)
```

```
## # A tibble: 6 x 11
##   County   ColorectalRate_I LungRate_I BreastRate_I ColorectalRate_M LungRate_M
##   <chr>               <dbl>      <dbl>        <dbl>            <dbl>      <dbl>
## 1 NC                   34.7       58.6         175.             12.8       36.8
## 2 Alamance             40.1       64.7         183.             13.1       43.5
## 3 Alamance             40.1       64.7         183.             13.1       43.5
## 4 Alamance             40.1       64.7         183.             13.1       43.5
## 5 Alamance             40.1       64.7         183.             13.1       43.5
## 6 Alamance             40.1       64.7         183.             13.1       43.5
## # i 5 more variables: BreastRate_M <dbl>, Metric <chr>, Value_Clean <dbl>,
## #   Z_Score <dbl>, Value <chr>
```

```
# save final data
#write_csv(final_data, "nc_county_health_data_final.csv")


### okay that actually doesn't work for LASSO, need one row per column
df <- final_data %>%
  filter(County != "NC") %>%
  filter(!is.na(Metric))

# reformatting to have 100 rows
df_wide <- df %>%
  select(County, Metric, Z_Score, ColorectalRate_I, LungRate_I, BreastRate_I,
         ColorectalRate_M, LungRate_M, BreastRate_M) %>%
  pivot_wider(
```

```
    names_from = Metric,
    values_from = Z_Score
  )

# cleaning up column names
colnames(df_wide) <- gsub(" ", "_", colnames(df_wide))

head(df_wide)
```

```
## # A tibble: 6 x 17
##   County    ColorectalRate_I LungRate_I BreastRate_I ColorectalRate_M LungRate_M
##   <chr>                <dbl>      <dbl>        <dbl>             <dbl>      <dbl>
## 1 Alamance              40.1       64.7         183.              13.1       43.5
## 2 Alexander             38.8       65.8         177.              14.3       43.7
## 3 Alleghany             48         74.1         192.              NA         37.4
## 4 Anson                 47.8       68.4         204.              20.2       43.9
## 5 Ashe                  34.2       73           209.               8.8       36.5
## 6 Avery                 27.3       65           171.              NA         27.6
## # i 11 more variables: BreastRate_M <dbl>, Population <dbl>,
## #   College_Graduation <dbl>, Uninsured_Adults <dbl>,
## #   `Health_Care_Workforce_-_Primary_Care_Physicians` <dbl>,
## #   Adult_Smoking <dbl>, Poverty <dbl>, Food_Insecurity <dbl>,
## #   Transportation <dbl>, Air_Pollution <dbl>,
## #   Metropolitan_or_Nonmetropolitan <dbl>
```

```
# overwrite old file
write_csv(df_wide, "nc_county_health_data_final.csv")
```

Verifying data using metrics outlined in Section G