FinalProjectProposal

Chang Lu

2024-11-7

```
chronicdata <- read.csv("C:/Users/beiq1/Desktop/678Finalproject/chronicdata.csv")

head(chronicdata)
```

```
##   YearStart YearEnd LocationAbbr         LocationDesc DataSource  Topic
## 1      2014    2014           AR             Arkansas  SEDD; SID Asthma
## 2      2018    2018           CO             Colorado  SEDD; SID Asthma
## 3      2018    2018           DC District of Columbia  SEDD; SID Asthma
## 4      2017    2017           GA              Georgia  SEDD; SID Asthma
## 5      2010    2010           MI             Michigan  SEDD; SID Asthma
## 6      2015    2015           MT              Montana  SEDD; SID Asthma
##                      Question Response DataValueUnit DataValueType DataValue
## 1 Hospitalizations for asthma       NA                      Number       916
## 2 Hospitalizations for asthma       NA                      Number      2227
## 3 Hospitalizations for asthma       NA                      Number       708
## 4 Hospitalizations for asthma       NA                      Number      3520
## 5 Hospitalizations for asthma       NA                      Number       123
## 6 Hospitalizations for asthma       NA                      Number
##   DataValueAlt DataValueFootnoteSymbol DatavalueFootnote LowConfidenceLimit
## 1          916                                                          NA
## 2         2227                                                          NA
## 3          708                                                          NA
## 4         3520                                                          NA
## 5          123                                                          NA
## 6           NA                       - No data available                NA
##   HighConfidenceLimit StratificationCategory1 Stratification1
## 1                  NA                 Gender            Male
## 2                  NA                Overall         Overall
## 3                  NA                Overall         Overall
## 4                  NA                 Gender          Female
## 5                  NA          Race/Ethnicity        Hispanic
## 6                  NA          Race/Ethnicity        Hispanic
##   StratificationCategory2 Stratification2 StratificationCategory3
## 1                      NA              NA                      NA
## 2                      NA              NA                      NA
## 3                      NA              NA                      NA
## 4                      NA              NA                      NA
## 5                      NA              NA                      NA
## 6                      NA              NA                      NA
##   Stratification3                                  GeoLocation ResponseID
## 1              NA   POINT (-92.27449074299966 34.74865012400045)         NA
## 2              NA POINT (-106.13361092099967 38.843840757000464)         NA
```

```
## 3              NA                    POINT (-77.036871 38.907192)         NA
## 4              NA   POINT (-83.62758034599966 32.83968109300048)         NA
## 5              NA    POINT (-84.71439026999968 44.6613195430005)         NA
## 6              NA  POINT (-109.42442064499971 47.06652897200047)         NA
##   LocationID TopicID QuestionID DataValueTypeID StratificationCategoryID1
## 1          5     AST     AST3_1            NMBR                     GENDER
## 2          8     AST     AST3_1            NMBR                    OVERALL
## 3         11     AST     AST3_1            NMBR                    OVERALL
## 4         13     AST     AST3_1            NMBR                     GENDER
## 5         26     AST     AST3_1            NMBR                       RACE
## 6         30     AST     AST3_1            NMBR                       RACE
##   StratificationID1 StratificationCategoryID2 StratificationID2
## 1              GENM                        NA                NA
## 2               OVR                        NA                NA
## 3               OVR                        NA                NA
## 4              GENF                        NA                NA
## 5               HIS                        NA                NA
## 6               HIS                        NA                NA
##   StratificationCategoryID3 StratificationID3
## 1                        NA                NA
## 2                        NA                NA
## 3                        NA                NA
## 4                        NA                NA
## 5                        NA                NA
## 6                        NA                NA
```

```
dim(chronicdata)
```

```
## [1] 1185676      34
```

```
colnames(chronicdata)
```

```
##  [1] "YearStart"                 "YearEnd"
##  [3] "LocationAbbr"              "LocationDesc"
##  [5] "DataSource"                "Topic"
##  [7] "Question"                  "Response"
##  [9] "DataValueUnit"             "DataValueType"
## [11] "DataValue"                 "DataValueAlt"
## [13] "DataValueFootnoteSymbol"   "DatavalueFootnote"
## [15] "LowConfidenceLimit"        "HighConfidenceLimit"
## [17] "StratificationCategory1"   "Stratification1"
## [19] "StratificationCategory2"   "Stratification2"
## [21] "StratificationCategory3"   "Stratification3"
## [23] "GeoLocation"               "ResponseID"
## [25] "LocationID"                "TopicID"
## [27] "QuestionID"                "DataValueTypeID"
## [29] "StratificationCategoryID1" "StratificationID1"
## [31] "StratificationCategoryID2" "StratificationID2"
## [33] "StratificationCategoryID3" "StratificationID3"
```

```
# Omit rows where "DataValueAlt" is NA or 0
chronicdata_filtered <- subset(chronicdata, !is.na(DataValueAlt) & DataValueAlt != 0)

dim(chronicdata_filtered)
```

```
## [1] 803878     34
```

```
# Display the unique values in the "Topic" and "TopicID" column
unique_topics <- unique(chronicdata_filtered$Topic)
unique_topicsID <- unique(chronicdata_filtered$TopicID)

unique_topics
```

```
##  [1] "Asthma"
##  [2] "Cancer"
##  [3] "Chronic Kidney Disease"
##  [4] "Chronic Obstructive Pulmonary Disease"
##  [5] "Cardiovascular Disease"
##  [6] "Diabetes"
##  [7] "Disability"
##  [8] "Reproductive Health"
##  [9] "Alcohol"
## [10] "Arthritis"
## [11] "Tobacco"
## [12] "Nutrition, Physical Activity, and Weight Status"
## [13] "Mental Health"
## [14] "Older Adults"
## [15] "Oral Health"
## [16] "Overarching Conditions"
## [17] "Immunization"
```

```
unique_topicsID
```

```
##  [1] "AST"  "CAN"  "CKD"  "COPD" "CVD"  "DIA"  "DIS"  "RPH"  "ALC"  "ART"
## [11] "TOB"  "NPAW" "MTH"  "OLD"  "ORH"  "OVC"  "IMM"
```

```
# Display the count of rows for each unique type in the "Topic" and "TopicID" column
topic_counts <- table(chronicdata_filtered$Topic)

topicID_counts <- table(chronicdata_filtered$TopicID)

# Print the counts
topic_counts
```

```
##
##                                  Alcohol
##                                    42930
##                                Arthritis
##                                    54809
##                                   Asthma
##                                    39845
##                                   Cancer
##                                   130377
##                     Cardiovascular Disease
##                                   113167
##                     Chronic Kidney Disease
##                                    18467
```

```
##           Chronic Obstructive Pulmonary Disease
##                                           94562
##                                        Diabetes
##                                           84627
##                                      Disability
##                                            3239
##                                    Immunization
##                                            8949
##                                  Mental Health
##                                           10716
## Nutrition, Physical Activity, and Weight Status
##                                           63090
##                                    Older Adults
##                                           19251
##                                     Oral Health
##                                           16936
##                           Overarching Conditions
##                                           60946
##                              Reproductive Health
##                                            5510
##                                         Tobacco
##                                           36457
```

```
topicID_counts
```

```
##
##    ALC    ART    AST    CAN    CKD   COPD    CVD    DIA    DIS    IMM    MTH
##  42930  54809  39845 130377  18467  94562 113167  84627   3239   8949  10716
##   NPAW    OLD    ORH    OVC    RPH    TOB
##  63090  19251  16936  60946   5510  36457
```

```
# Print the count of rows where they are different
num_different_years
```

```
## [1] 161284
```

```
num_same_years
```

```
## [1] 1024392
```

```
unique_location <- unique(chronicdata_filtered$LocationDesc)

unique_location
```

```
##  [1] "Arkansas"             "Colorado"             "District of Columbia"
##  [4] "Georgia"              "Michigan"             "Oregon"
##  [7] "Wisconsin"            "Alabama"              "Idaho"
## [10] "Illinois"             "Kansas"               "Louisiana"
## [13] "Massachusetts"        "Maryland"             "Minnesota"
## [16] "Mississippi"          "North Carolina"       "New Mexico"
## [19] "Texas"                "New York"             "Indiana"
```

```
## [22] "Nevada"               "South Carolina"       "Virginia"
## [25] "Iowa"                 "Utah"                 "Wyoming"
## [28] "Alaska"               "California"           "Ohio"
## [31] "United States"        "Hawaii"               "Washington"
## [34] "South Dakota"         "Delaware"             "Kentucky"
## [37] "Rhode Island"         "Vermont"              "Arizona"
## [40] "Florida"              "Nebraska"             "New Jersey"
## [43] "Missouri"             "Maine"                "Connecticut"
## [46] "Tennessee"            "Montana"              "Pennsylvania"
## [49] "North Dakota"         "Puerto Rico"          "Oklahoma"
## [52] "New Hampshire"        "Guam"                 "West Virginia"
## [55] "Virgin Islands"
```

Analysis on alcohol issue

```
# Filter rows where the "Topic" or "Question" column mentions alcohol
alcohol_data <- subset(chronicdata_filtered, grepl("alcohol", Topic, ignore.case = TRUE) |
                                  grepl("alcohol", Question, ignore.case = TRUE))

head(alcohol_data)
```

```
##     YearStart YearEnd LocationAbbr  LocationDesc DataSource   Topic
## 195      2017    2017           MA Massachusetts       NVSS Alcohol
## 197      2015    2015           CO      Colorado      PRAMS Alcohol
## 200      2011    2011           NJ    New Jersey       NVSS Alcohol
## 201      2017    2017           ME         Maine      YRBSS Alcohol
## 203      2013    2013           MA Massachusetts       NVSS Alcohol
## 205      2013    2013           AK        Alaska       NVSS Alcohol
##                            Question Response     DataValueUnit
## 195 Chronic liver disease mortality       NA
## 197    Alcohol use before pregnancy       NA                 %
## 200 Chronic liver disease mortality       NA
## 201         Alcohol use among youth       NA                 %
## 203 Chronic liver disease mortality       NA cases per 100,000
## 205 Chronic liver disease mortality       NA cases per 100,000
##         DataValueType DataValue DataValueAlt DataValueFootnoteSymbol
## 195            Number       251        251.0
## 197  Crude Prevalence      18.8         18.8                      &
## 200            Number       784        784.0
## 201  Crude Prevalence      22.2         22.2
## 203 Age-adjusted Rate       6.9          6.9
## 205        Crude Rate      11.4         11.4
##                                  DatavalueFootnote LowConfidenceLimit
## 195                                                                NA
## 197 Less than 60 respondents, interpret with caution               9.0
## 200                                                                NA
## 201                                                              20.8
## 203                                                               4.5
## 205                                                               8.3
##     HighConfidenceLimit StratificationCategory1          Stratification1
## 195                  NA                  Gender                   Female
## 197                35.1          Race/Ethnicity Asian or Pacific Islander
```

```
## 200                  NA                Overall                    Overall
## 201                23.7          Race/Ethnicity       White, non-Hispanic
## 203                10.2          Race/Ethnicity                   Hispanic
## 205                15.3                  Gender                       Male
##     StratificationCategory2 Stratification2 StratificationCategory3
## 195                      NA              NA                       NA
## 197                      NA              NA                       NA
## 200                      NA              NA                       NA
## 201                      NA              NA                       NA
## 203                      NA              NA                       NA
## 205                      NA              NA                       NA
##     Stratification3                                    GeoLocation ResponseID
## 195              NA   POINT (-72.08269067499964 42.27687047000046)         NA
## 197              NA POINT (-106.13361092099967 38.843840757000464)         NA
## 200              NA   POINT (-74.27369128799967 40.13057004800049)         NA
## 201              NA  POINT (-68.98503133599962 45.254228894000505)         NA
## 203              NA   POINT (-72.08269067499964 42.27687047000046)         NA
## 205              NA  POINT (-147.72205903599973 64.84507995700051)         NA
##     LocationID TopicID QuestionID DataValueTypeID StratificationCategoryID1
## 195         25     ALC     ALC6_0            NMBR                    GENDER
## 197          8     ALC     ALC1_2         CRDPREV                      RACE
## 200         34     ALC     ALC6_0            NMBR                   OVERALL
## 201         23     ALC     ALC1_1         CRDPREV                      RACE
## 203         25     ALC     ALC6_0      AGEADJRATE                      RACE
## 205          2     ALC     ALC6_0         CRDRATE                    GENDER
##     StratificationID1 StratificationCategoryID2 StratificationID2
## 195              GENF                        NA               NA
## 197               API                        NA               NA
## 200               OVR                        NA               NA
## 201               WHT                        NA               NA
## 203               HIS                        NA               NA
## 205              GENM                        NA               NA
##     StratificationCategoryID3 StratificationID3
## 195                        NA               NA
## 197                        NA               NA
## 200                        NA               NA
## 201                        NA               NA
## 203                        NA               NA
## 205                        NA               NA
```

```
summary(alcohol_data$DataValueAlt)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##     0.02     5.70     8.70    51.89    17.70 51642.00
```

```
# Find how many types of the dataset
Types_alcohol_data <- table(alcohol_data$DataValueType)

Types_alcohol_data
```

```
##
##              Age-adjusted Mean        Age-adjusted Prevalence
##                           6097                           7209
```

```
##              Age-adjusted Rate                Crude Prevalence
##                           3030                           13404
##                     Crude Rate                            Mean
##                           3030                            6080
##                         Number Per capita alcohol consumption
##                           3030                             312
##                     US Dollars
##                            738
```

```r
# Split data by DataValueType
numbers_data <- subset(alcohol_data, DataValueType == "Number")
crude_prevalence_data <- subset(alcohol_data, DataValueType == "Crude Prevalence")
age_adjusted_rate_data <- subset(alcohol_data, DataValueType == "Age-adjusted Rate")
```