EDA Report

Jessica Riedy

2025-04-06

```
library(readxl)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```
library(scales)

data <- read_excel("Data/googleTrendsMH.xlsx", sheet = "googleTrendsMH")
```

```
# Check for columns with  missing values
na_counts <- colSums(is.na(data))
na_columns <- na_counts[na_counts > 0]

if (any(na_counts > 0)){
  paste("List of columns with missing values for examination")
  print(names(na_columns))
} else {
  print("There are zero columns with null values.")
}
```

```
## [1] "There are zero columns with null values."
```

```
# Show summary statistics for numeric columns
numeric_cols <- select_if(data, is.numeric)
summary(numeric_cols)
```

```
##       year           fips       population_est       anxiety_ct
##  Min.   :2013   Min.   : 1.00   Min.   :  579054   Min.   :   217
##  1st Qu.:2015   1st Qu.:16.00   1st Qu.: 1945163   1st Qu.:  4604
##  Median :2018   Median :30.00   Median : 4636208   Median : 14139
##  Mean   :2018   Mean   :29.38   Mean   : 6690537   Mean   : 24279
##  3rd Qu.:2020   3rd Qu.:42.00   3rd Qu.: 7502082   3rd Qu.: 31821
##  Max.   :2022   Max.   :56.00   Max.   :39437610   Max.   :177155
```

```
##  trauma_stress_ct    adhd_ct        bipolar_ct     depression_ct
##  Min.   :   254   Min.   :    0   Min.   :   415   Min.   :   779
##  1st Qu.:  4473   1st Qu.: 2114   1st Qu.:  4234   1st Qu.:  7854
##  Median : 12913   Median : 6441   Median : 11342   Median : 23356
##  Mean   : 23569   Mean   :13244   Mean   : 16884   Mean   : 35634
##  3rd Qu.: 31016   3rd Qu.:18996   3rd Qu.: 21377   3rd Qu.: 45868
##  Max.   :142480   Max.   :76030   Max.   :113619   Max.   :201222
##  comm_psych_care  outpatient_util    state_psych_care  private_psych_care
##  Min.   :  9426   Min.   :0.002403   Min.   :  14158   Min.   :  13238
##  1st Qu.: 32606   1st Qu.:0.012931   1st Qu.:  63549   1st Qu.:  64165
##  Median : 88102   Median :0.019947   Median : 171969   Median : 173716
##  Mean   :137769   Mean   :0.025145   Mean   : 265977   Mean   : 263710
##  3rd Qu.:153101   3rd Qu.:0.031510   3rd Qu.: 293019   3rd Qu.: 290914
##  Max.   :784665   Max.   :0.105976   Max.   :1478138   Max.   :1465591
##  resid_psych_care  total_inpatient   inpatient_util    total_civilian
##  Min.   :  11622   Min.   :  39018   Min.   :0.01279   Min.   :  50826
##  1st Qu.:  64165   1st Qu.: 192051   1st Qu.:0.07423   1st Qu.: 224384
##  Median : 173264   Median : 519936   Median :0.11536   Median : 607942
##  Mean   : 267244   Mean   : 796930   Mean   :0.14564   Mean   : 934699
##  3rd Qu.: 294089   3rd Qu.: 876400   3rd Qu.:0.18728   3rd Qu.:1029324
##  Max.   :1479114   Max.   :4422843   Max.   :0.62662   Max.   :5207508
##    total_util       median_adhd     median_ptsd    median_anxiety
##  Min.   :0.01520   Min.   :11.50   Min.   : 0.00   Min.   :34.00
##  1st Qu.:0.08709   1st Qu.:21.00   1st Qu.:11.50   1st Qu.:62.00
##  Median :0.13466   Median :23.00   Median :13.00   Median :75.25
##  Mean   :0.17079   Mean   :26.59   Mean   :13.08   Mean   :72.22
##  3rd Qu.:0.21907   3rd Qu.:26.50   3rd Qu.:14.50   3rd Qu.:84.00
##  Max.   :0.73260   Max.   :64.00   Max.   :21.00   Max.   :92.50
##  median_bipolar  median_depression median_mental_hospital
##  Min.   :14.00   Min.   :36.5      Min.   : 0.00
##  1st Qu.:19.50   1st Qu.:62.0      1st Qu.:30.62
##  Median :21.00   Median :67.0      Median :38.50
##  Mean   :20.67   Mean   :66.9      Mean   :34.93
##  3rd Qu.:22.00   3rd Qu.:72.0      3rd Qu.:45.88
##  Max.   :26.00   Max.   :85.0      Max.   :78.00
##  median_psychiatrists_near_me median_psychologist_near_me
##  Min.   : 0.0000              Min.   : 0.000
##  1st Qu.: 0.0000              1st Qu.: 0.000
##  Median : 0.0000              Median : 0.000
##  Mean   : 0.6561              Mean   : 5.219
##  3rd Qu.: 0.0000              3rd Qu.:12.000
##  Max.   :17.0000              Max.   :25.500
##  median_therapist_near_me median_all_trends   mean_adhd       mean_ptsd
##  Min.   : 0.00            Min.   : 0.00     Min.   :12.67   Min.   : 3.083
##  1st Qu.: 0.00            1st Qu.:19.50     1st Qu.:21.08   1st Qu.:11.667
##  Median :16.00            Median :21.50     Median :23.00   Median :13.250
##  Mean   :27.98            Mean   :23.95     Mean   :26.52   Mean   :13.126
##  3rd Qu.:55.75            3rd Qu.:25.50     3rd Qu.:26.67   3rd Qu.:14.583
##  Max.   :95.50            Max.   :57.00     Max.   :60.58   Max.   :21.500
##   mean_anxiety    mean_bipolar   mean_depression mean_mental_hospital
##  Min.   :33.83   Min.   :14.42   Min.   :38.75   Min.   : 0.00
##  1st Qu.:62.19   1st Qu.:19.67   1st Qu.:61.83   1st Qu.:30.50
##  Median :75.79   Median :20.83   Median :66.58   Median :38.08
##  Mean   :72.31   Mean   :20.78   Mean   :66.21   Mean   :35.88
```

```
##  3rd Qu.:84.06   3rd Qu.:21.92   3rd Qu.:71.17   3rd Qu.:45.29
##  Max.   :91.92   Max.   :25.33   Max.   :78.83   Max.   :77.00
##  mean_psychiatrists_near_me mean_psychologist_near_me mean_therapist_near_me
##  Min.   : 0.0000            Min.   : 0.000            Min.   : 0.00
##  1st Qu.: 0.0000            1st Qu.: 0.000            1st Qu.: 0.00
##  Median : 0.0000            Median : 1.583            Median :16.71
##  Mean   : 0.8446            Mean   : 5.522            Mean   :28.01
##  3rd Qu.: 0.7500            3rd Qu.:11.479            3rd Qu.:53.58
##  Max.   :17.3333            Max.   :25.167            Max.   :91.75
##  mean_all_trends
##  Min.   :15.34
##  1st Qu.:24.02
##  Median :28.65
##  Mean   :29.91
##  3rd Qu.:35.96
##  Max.   :45.21
```

```r
table1 <- data %>%
  group_by(region) %>%
  summarise(
    avg_outpatient_util = mean(outpatient_util, na.rm = TRUE),
    avg_inpatient_util = mean(inpatient_util, na.rm = TRUE),
    avg_total_util = mean(total_util, na.rm = TRUE),
    avg_median_trend = mean(median_all_trends, na.rm = TRUE)
  )
print(table1)
```

```
## # A tibble: 4 x 5
##   region  avg_outpatient_util avg_inpatient_util avg_total_util avg_median_trend
##   <chr>                 <dbl>              <dbl>          <dbl>            <dbl>
## 1 Atlant~              0.0302              0.174          0.204             22.9
## 2 Central              0.0258              0.150          0.175             24.1
## 3 South                0.0181              0.105          0.123             25.6
## 4 West P~              0.0273              0.159          0.186             23.1
```

```r
table2 <- data %>%
  arrange(desc(total_util)) %>%
  select(state, year, total_util, outpatient_util, inpatient_util, median_all_trends) %>%
  head(10)
print(table2)
```

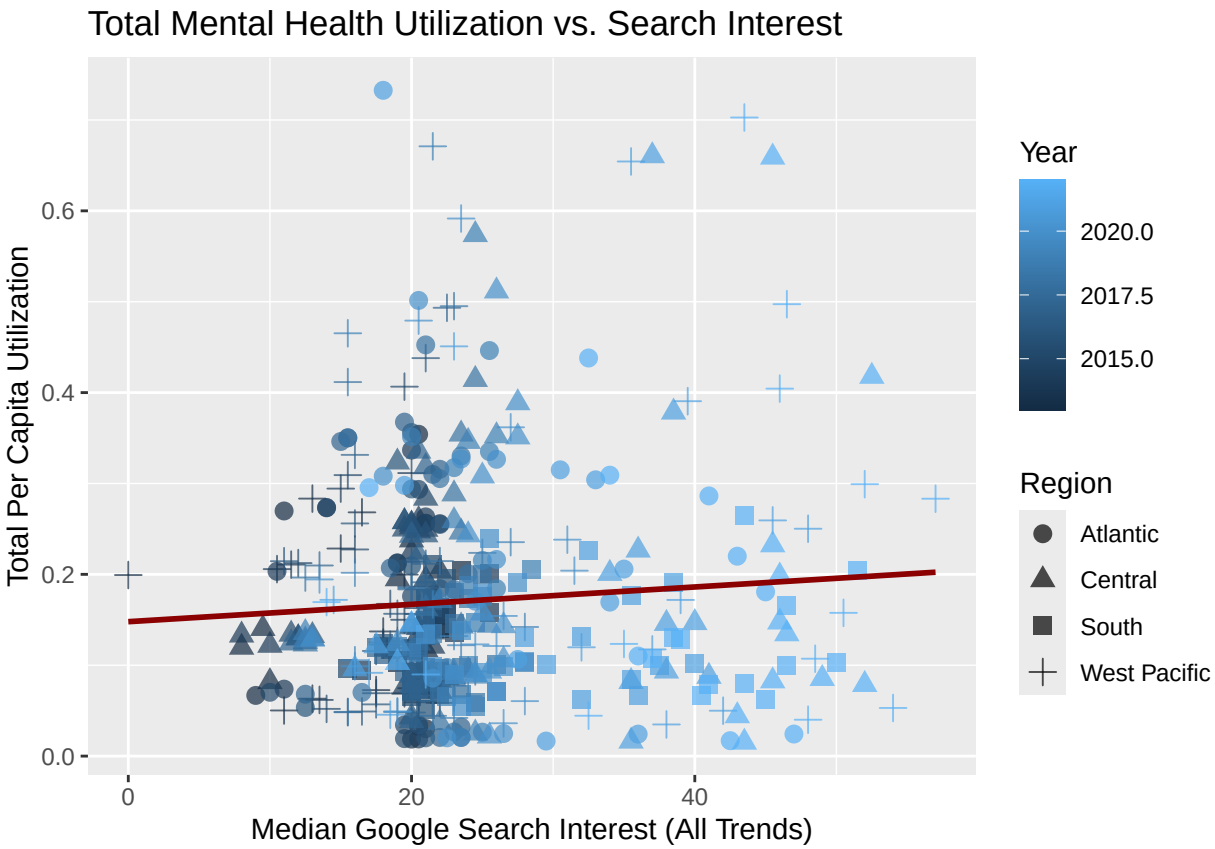
```
## # A tibble: 10 x 6
##    state  year total_util outpatient_util inpatient_util median_all_trends
##    <chr> <dbl>      <dbl>           <dbl>          <dbl>             <dbl>
##  1 DC     2021      0.733          0.106           0.627              18
##  2 NM     2022      0.703          0.102           0.601              43.5
##  3 NM     2019      0.671          0.0972          0.574              21.5
##  4 IA     2021      0.661          0.0972          0.563              37
##  5 IA     2022      0.659          0.0985          0.561              45.5
##  6 NM     2021      0.654          0.0946          0.560              35.5
##  7 NM     2020      0.592          0.0856          0.506              23.5
##  8 IA     2019      0.574          0.0849          0.489              24.5
##  9 IA     2020      0.512          0.0774          0.434              26
## 10 DC     2020      0.501          0.0724          0.429              20.5
```

```r
# calculate the private and public utilization rates
data <- data %>%
  mutate(state_util = (state_psych_care/population_est),
         private_util = (private_psych_care/population_est),
         diff_util = (total_util-(state_util + private_util))
  )

table3 <- data %>%
  arrange(desc(total_util)) %>%
  select(state, year, total_util, state_util, private_util, diff_util) %>%
  head(10)
print(table3)
```

```
## # A tibble: 10 x 6
##    state  year total_util state_util private_util diff_util
##    <chr> <dbl>      <dbl>      <dbl>        <dbl>     <dbl>
##  1 DC     2021      0.733      0.208        0.209     0.315
##  2 NM     2022      0.703      0.203        0.196     0.304
##  3 NM     2019      0.671      0.193        0.187     0.291
##  4 IA     2021      0.661      0.191        0.186     0.283
##  5 IA     2022      0.659      0.193        0.179     0.287
##  6 NM     2021      0.654      0.189        0.182     0.283
##  7 NM     2020      0.592      0.170        0.166     0.255
##  8 IA     2019      0.574      0.166        0.161     0.247
##  9 IA     2020      0.512      0.149        0.141     0.222
## 10 DC     2020      0.501      0.143        0.143     0.216
```

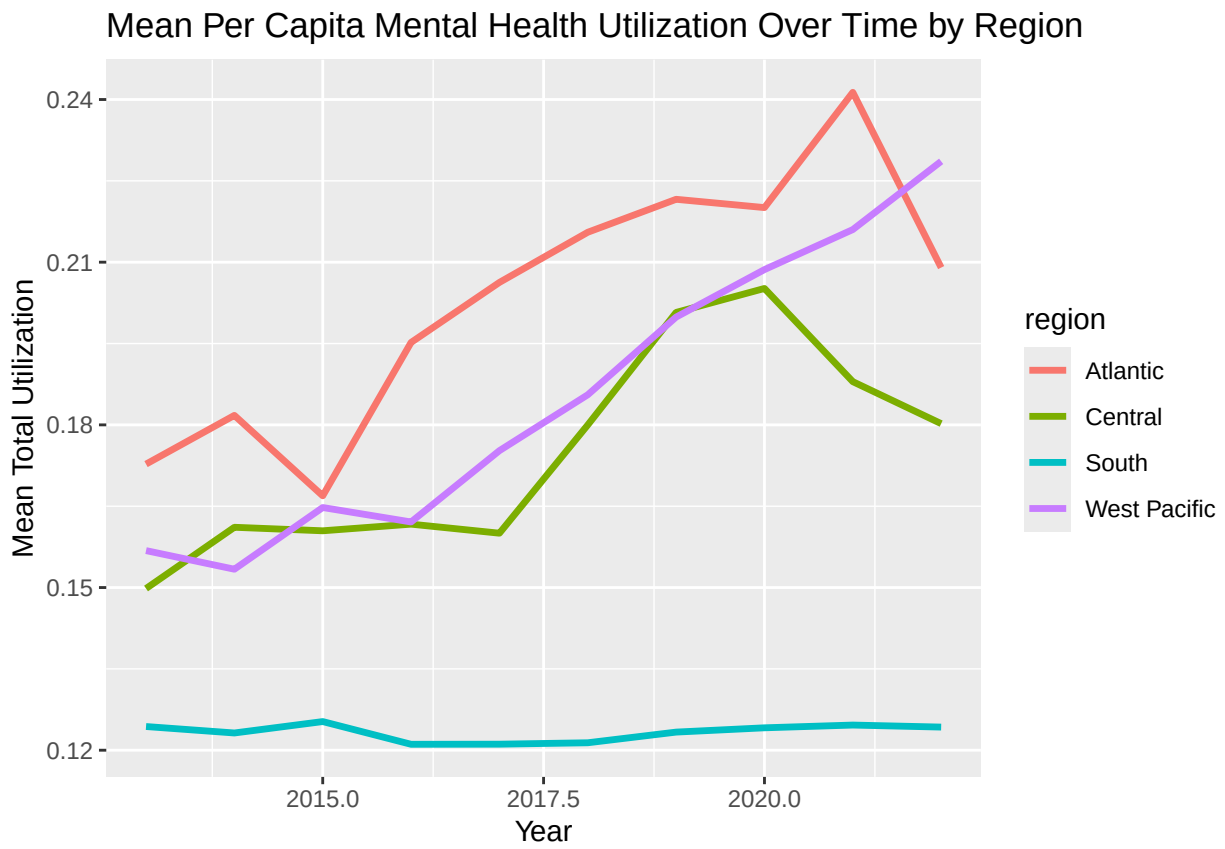
```r
ggplot(data, aes(x = median_all_trends, y = total_util)) +
  geom_point(aes(color = year, shape = region), size = 3, alpha = 0.7) +
  geom_smooth(method = "lm", se = FALSE, color = "darkred") +
  labs(title = "Total Mental Health Utilization vs. Search Interest",
       x = "Median Google Search Interest (All Trends)",
       y = "Total Per Capita Utilization",
       color = "Year",
       shape = "Region")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

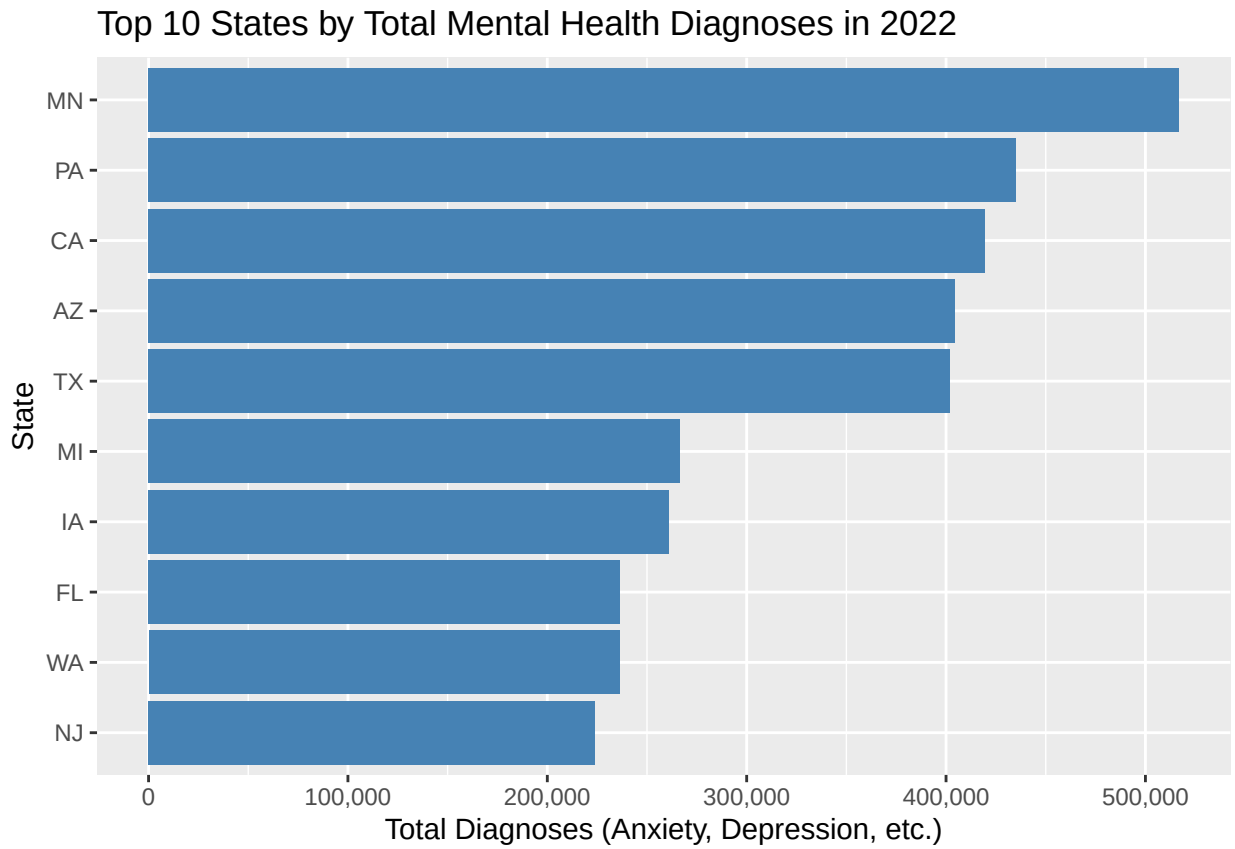



```
ggplot(data, aes(x = year, y = total_util, color = region)) +
  geom_line(stat = "summary", fun = "mean", size = 1.2) +
  labs(title = "Mean Per Capita Mental Health Utilization Over Time by Region",
       x = "Year", y = "Mean Total Utilization")
```

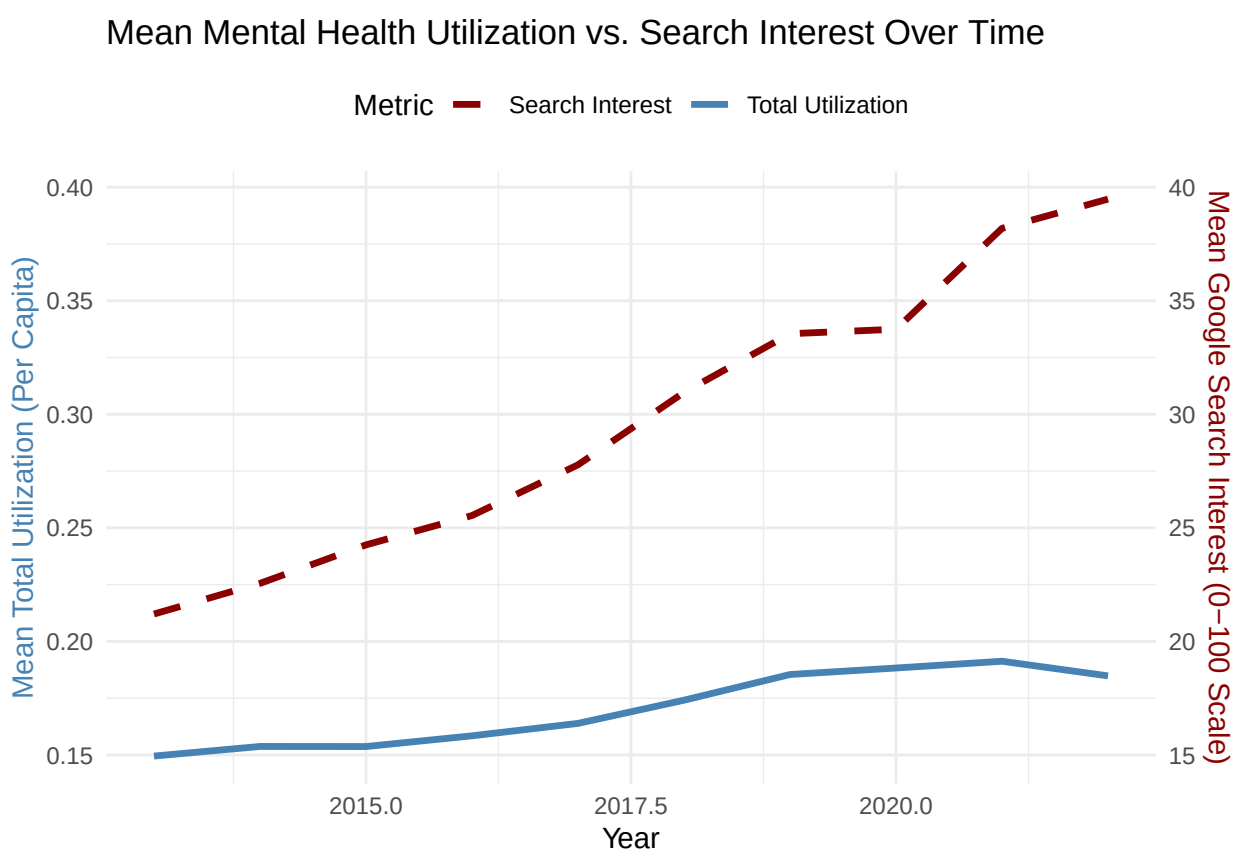
```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



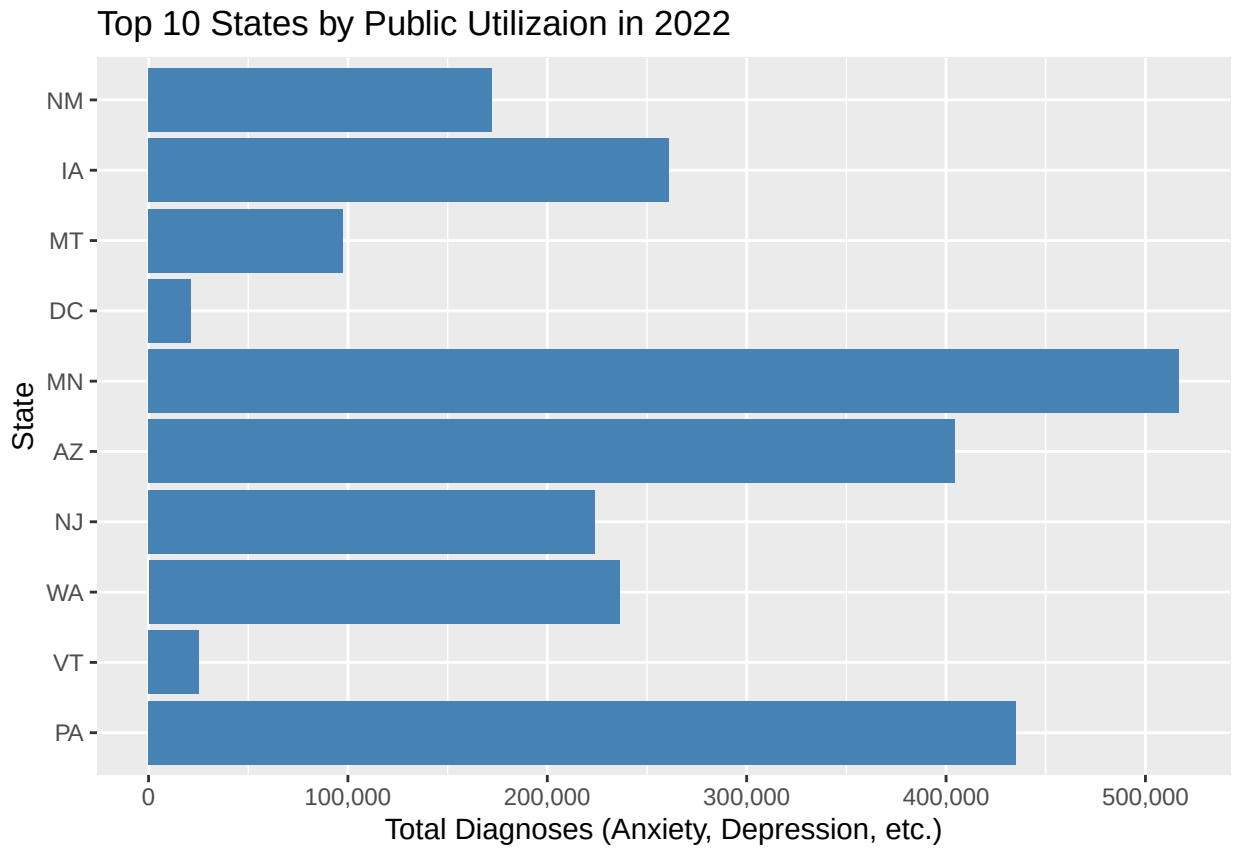

```r
latest_year <- max(data$year, na.rm = TRUE)

data_latest <- filter(data, year == latest_year)

data_latest$diagnoses_total <- rowSums(data_latest[, c("anxiety_ct", "depression_ct", "adhd_ct", "bipola

top_diagnosis_states <- data_latest %>%
  arrange(desc(diagnoses_total)) %>%
  head(10)

ggplot(top_diagnosis_states, aes(x = reorder(state, diagnoses_total), y = diagnoses_total)) +
  geom_col(fill = "steelblue") +
  coord_flip() +
  labs(title = paste("Top 10 States by Total Mental Health Diagnoses in", latest_year),
       x = "State", y = "Total Diagnoses (Anxiety, Depression, etc.)") +
  scale_y_continuous(labels = label_comma())
```



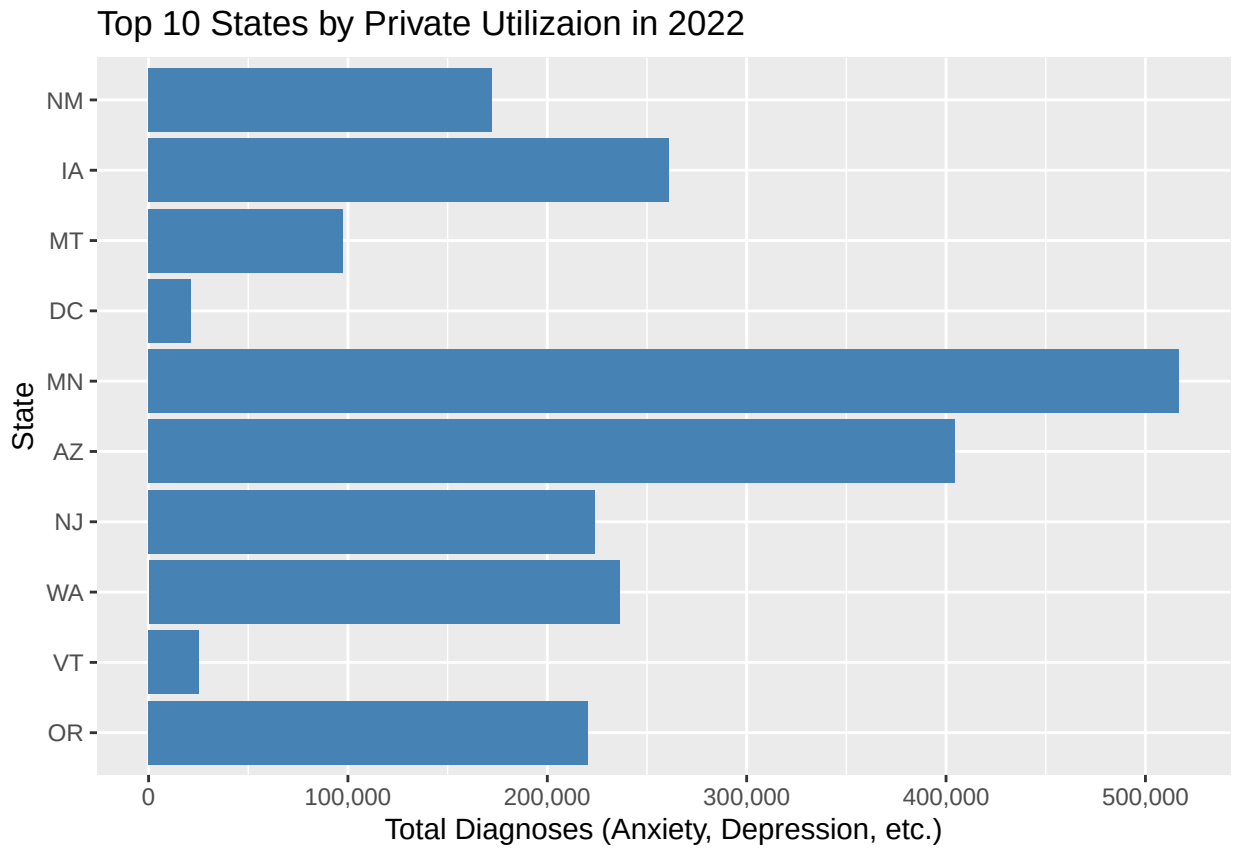

```r
summary_by_year <- data %>%
  group_by(year) %>%
  summarise(
    mean_total_util = mean(total_util, na.rm = TRUE),
    mean_search_interest = mean(mean_all_trends, na.rm = TRUE)
  )

ggplot(summary_by_year, aes(x = year)) +
  geom_line(aes(y = mean_total_util, color = "Total Utilization"), size = 1.2) +
  geom_line(aes(y = mean_search_interest / 100, color = "Search Interest"), size = 1.2, linetype = "dash
  scale_y_continuous(
    name = "Mean Total Utilization (Per Capita)",
    sec.axis = sec_axis(~ . * 100, name = "Mean Google Search Interest (0-100 Scale)")
  ) +
  scale_color_manual(values = c("Total Utilization" = "steelblue", "Search Interest" = "darkred")) +
  labs(
    title = "Mean Mental Health Utilization vs. Search Interest Over Time",
    x = "Year",
    color = "Metric"
  ) +
  theme_minimal() +
  theme(
    axis.title.y.left = element_text(color = "steelblue"),
    axis.title.y.right = element_text(color = "darkred"),
    legend.position = "top")
```




```
data_latest$diagnoses_total <- rowSums(data_latest[, c("anxiety_ct", "depression_ct", "adhd_ct", "bipola

top_public_utillization_states <- data_latest %>%
  arrange(desc(state_util)) %>%
  head(10)

ggplot(top_public_utillization_states, aes(x = reorder(state, state_util), y = diagnoses_total)) +
  geom_col(fill = "steelblue") +
  coord_flip() +
  labs(title = paste("Top 10 States by Public Utilizaion in", latest_year),
       x = "State", y = "Total Diagnoses (Anxiety, Depression, etc.)") +
  scale_y_continuous(labels = label_comma())
```

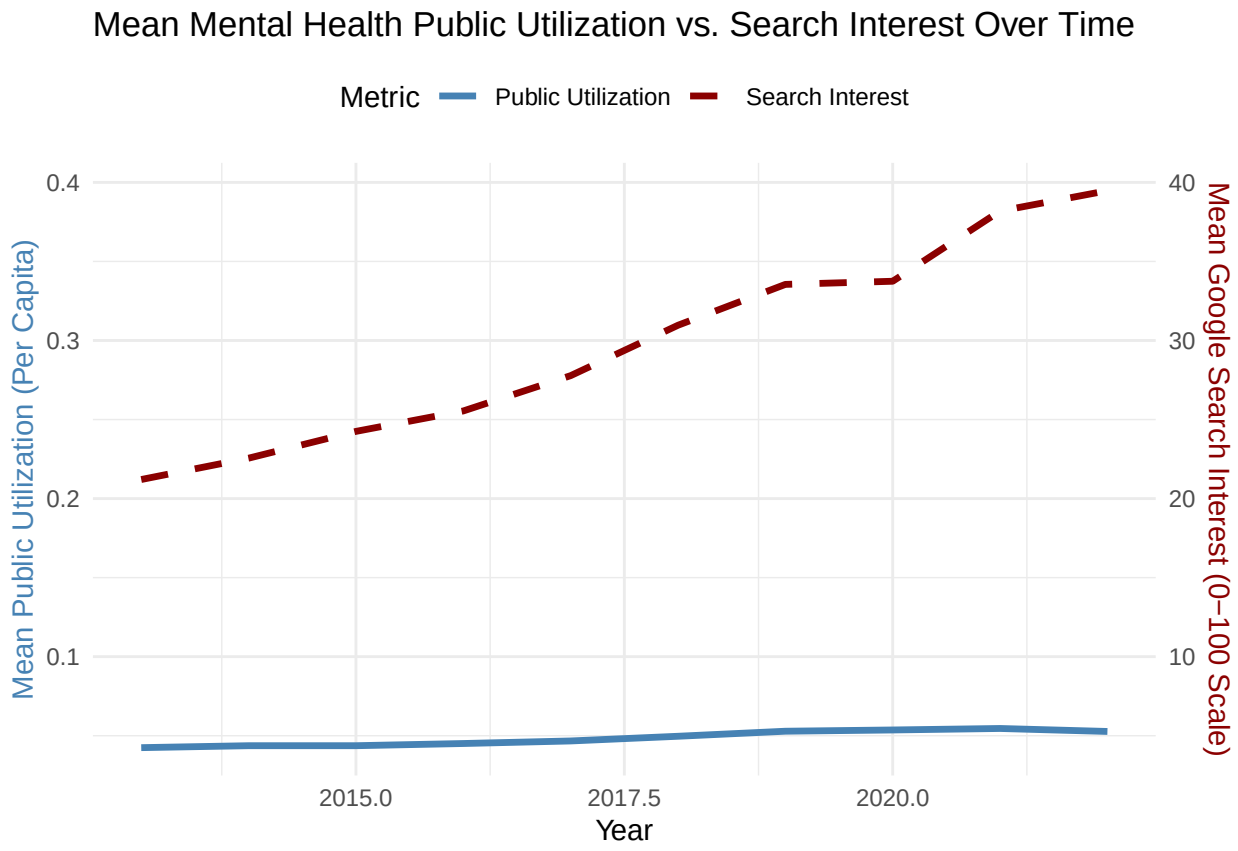



```r
data_latest$diagnoses_total <- rowSums(data_latest[, c("anxiety_ct", "depression_ct", "adhd_ct", "bipola

top_private_utillization_states <- data_latest %>%
  arrange(desc(private_util)) %>%
  head(10)

ggplot(top_private_utillization_states, aes(x = reorder(state, private_util), y = diagnoses_total)) +
  geom_col(fill = "steelblue") +
  coord_flip() +
  labs(title = paste("Top 10 States by Private Utilizaion in", latest_year),
       x = "State", y = "Total Diagnoses (Anxiety, Depression, etc.)") +
  scale_y_continuous(labels = label_comma())
```




```r
summary_by_year <- data %>%
  group_by(year) %>%
  summarise(
    mean_state_util = mean(state_util, na.rm = TRUE),
    mean_search_interest = mean(mean_all_trends, na.rm = TRUE)
  )

ggplot(summary_by_year, aes(x = year)) +
  geom_line(aes(y = mean_state_util, color = "Public Utilization"), size = 1.2) +
  geom_line(aes(y = mean_search_interest / 100, color = "Search Interest"), size = 1.2, linetype = "dash
  scale_y_continuous(
    name = "Mean Public Utilization (Per Capita)",
    sec.axis = sec_axis(~ . * 100, name = "Mean Google Search Interest (0-100 Scale)")
  ) +
  scale_color_manual(values = c(
    "Public Utilization" = "steelblue",
    "Search Interest" = "darkred"
  )) +
  labs(
    title = "Mean Mental Health Public Utilization vs. Search Interest Over Time",
    x = "Year",
    color = "Metric"
  ) +
  theme_minimal() +
  theme(
    axis.title.y.left = element_text(color = "steelblue"),
```

```
    axis.title.y.right = element_text(color = "darkred"),
    legend.position = "top"
  )
```




```
summary_by_year <- data %>%
  group_by(year) %>%
  summarise(
    mean_private_util = mean(private_util, na.rm = TRUE),
    mean_search_interest = mean(mean_all_trends, na.rm = TRUE)
  )

ggplot(summary_by_year, aes(x = year)) +
  geom_line(aes(y = mean_private_util, color = "Private Utilization"), size = 1.2) +
  geom_line(aes(y = mean_search_interest / 100, color = "Search Interest"), size = 1.2, linetype = "dash
  scale_y_continuous(
    name = "Mean Private Utilization (Per Capita)",
    sec.axis = sec_axis(~ . * 100, name = "Mean Google Search Interest (0-100 Scale)")
  ) +
  scale_color_manual(values = c(
    "Private Utilization" = "darkgreen",
    "Search Interest" = "darkred"
  )) +
  labs(
    title = "Mean Mental Health Private Utilization vs. Search Interest Over Time",
    x = "Year",
    color = "Metric"
```

```
) +
theme_minimal() +
theme(
  axis.title.y.left = element_text(color = "steelblue"),
  axis.title.y.right = element_text(color = "darkred"),
  legend.position = "top"
)
```