# EDA Report

Jessica Riedy

2025-04-05

```
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")
```

```
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
```

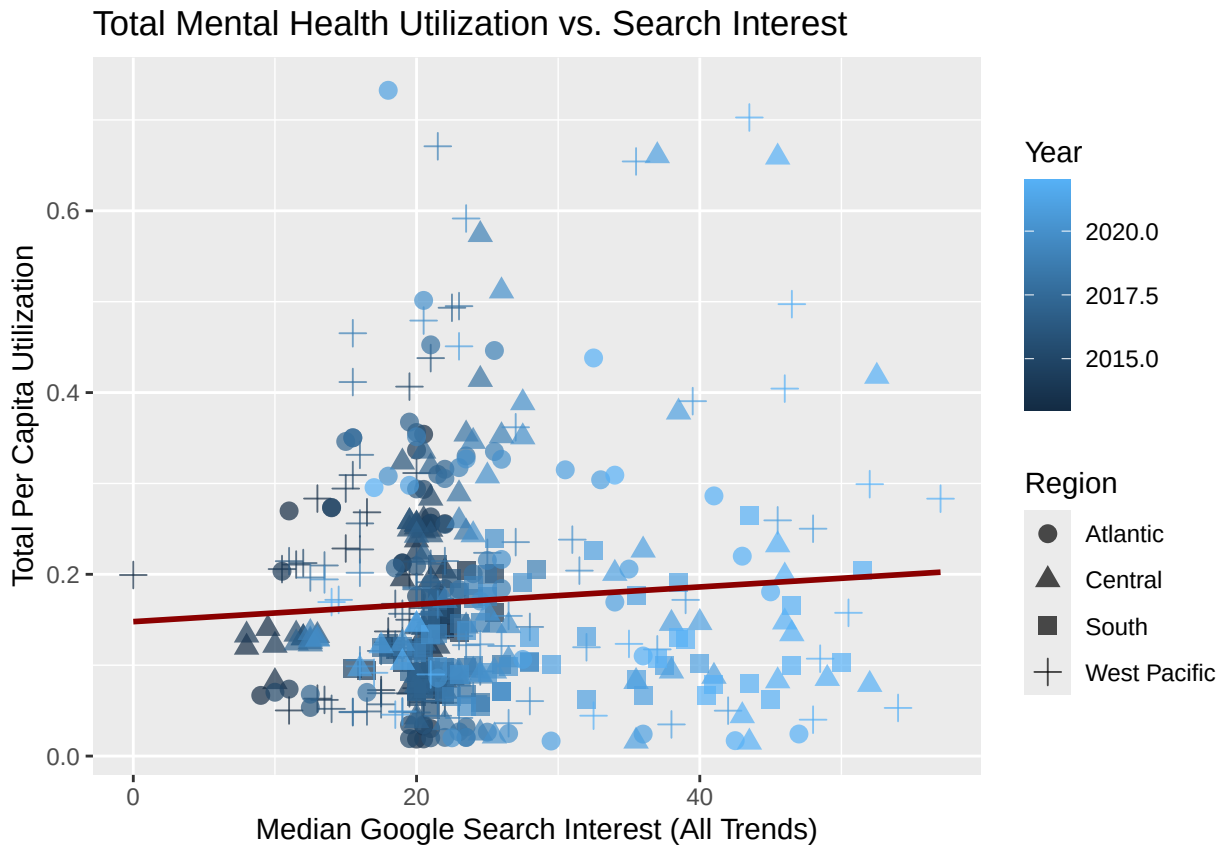
```
table2 <- data %>%
  arrange(desc(total_util)) %>%
  select(state, year, total_util, outpatient_util, inpatient_util, median_all_trends) %>%
  head(10)
print(table2)
```

```
## # 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
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")
```

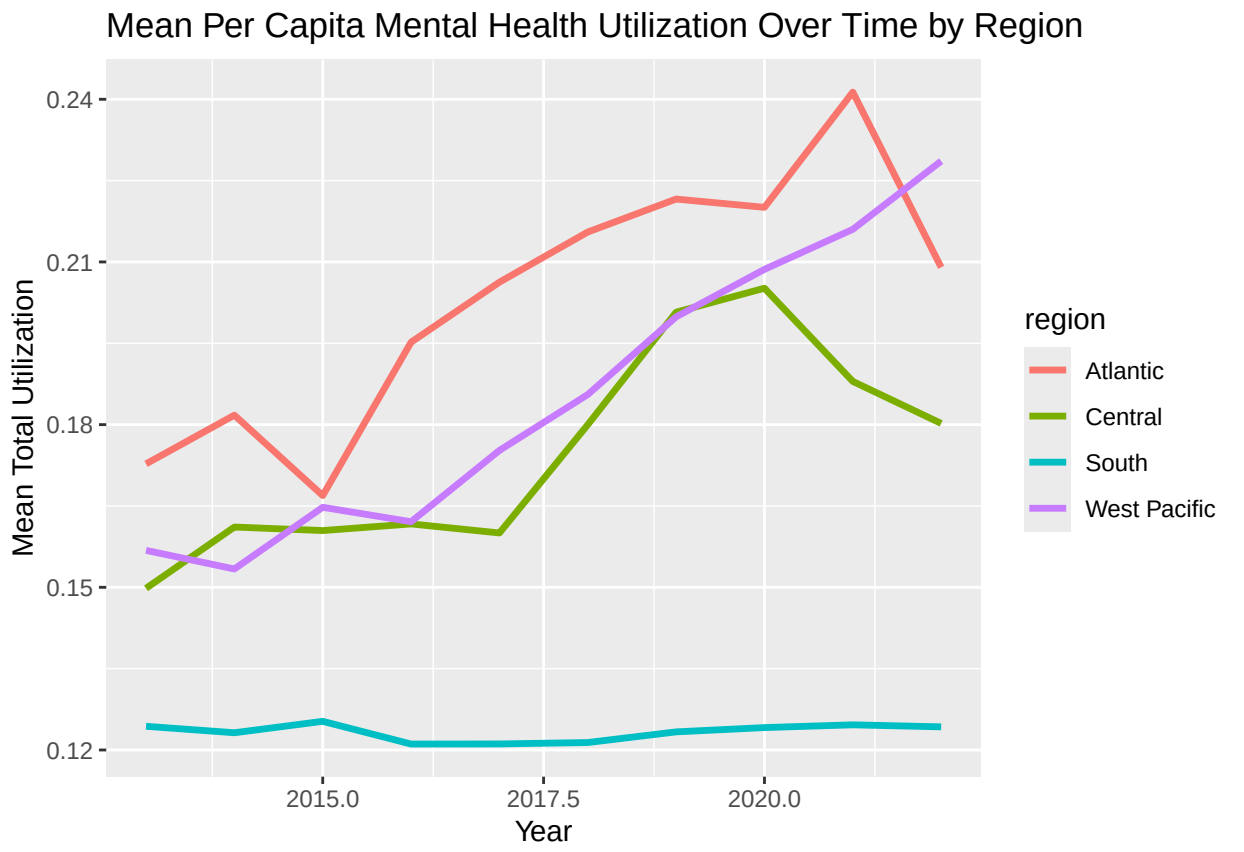
```
## `geom_smooth()` using formula = 'y ~ x'
```



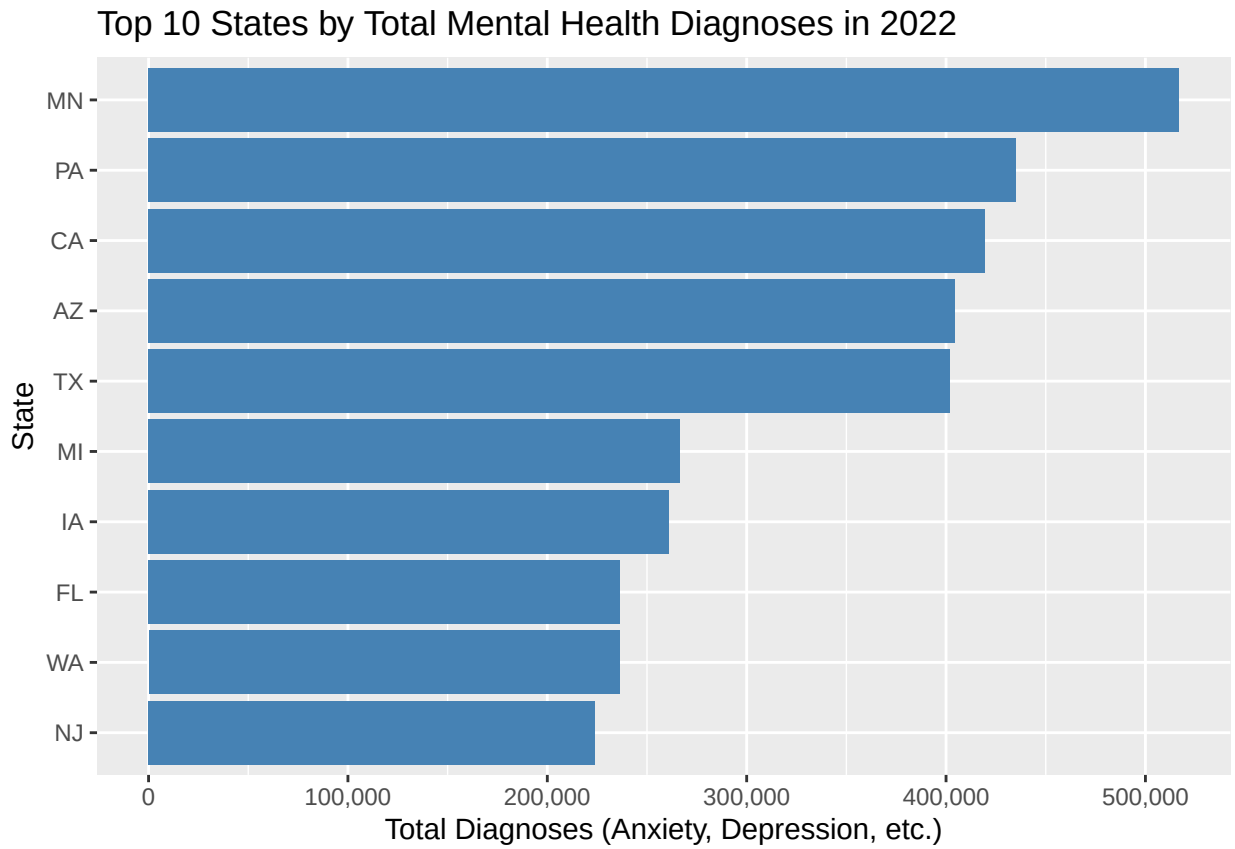

```r
ggplot(data, aes(x = year, y = total_util, color = region)) +
  geom_line(stat = "summary", fun = "mean", size = 1.2) +
  labs(title = "Mean Per Capita Mental Health Utilization Over Time by Region",
```

```
       x = "Year", y = "Mean Total Utilization")
```

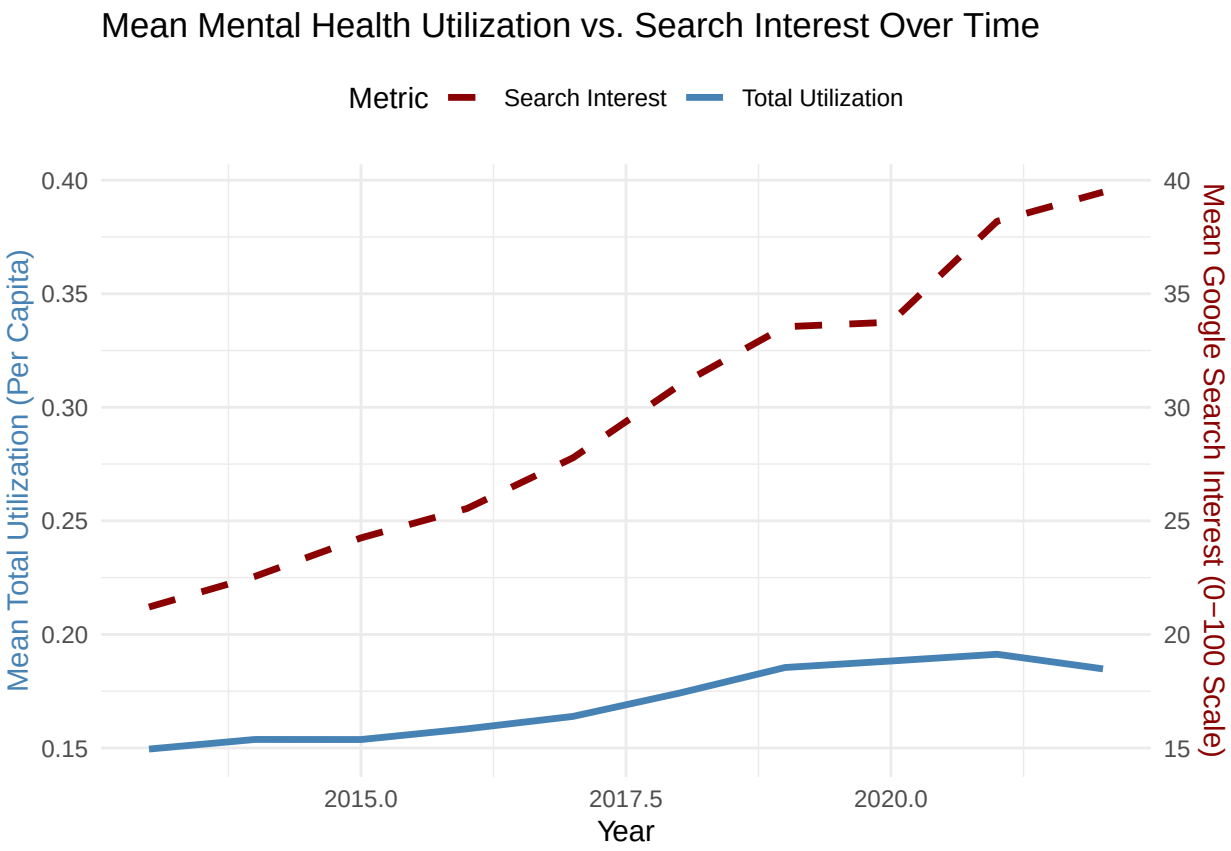
```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```




```
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")
```


Mean Mental Health Utilization vs. Search Interest Over Time
Metric
Search Interest
Total Utilization
Mean Total Utilization (Per Capita)
Mean Google Search Interest (0–100 Scale)
0.40
0.35
0.30
0.25
0.20
0.15
40
35
30
25
20
15
2015.0
2017.5
2020.0
Year