Assignment 2: Initial Findings Memo

Tina Wooldridge

Setup

```
library(tidyverse)
load("w144_wrangled.RData")
```

Assignment 2: "The First Look"

Conceptual Focus

This assignment covers **Chapter 13: Descriptive Statistics and Visualization**. Last week, we finished our "mise en place." Now, we need to find the "story" in the data. This is where we find the "wow" insights for the client. As your slides note, we will use two toolkits: * **Descriptive Statistics:** Summarizing data with precision (e.g., `summary()`, `count()`). * **Data Visualization:** Summarizing data with intuitive pictures (e.g., `ggplot2`).

Initial Insights Memo

To: Alex Chen (Director of Insights)

From: [Tina Wooldridge]

Subject: Initial Insights for 2025 Digital Landscape Brief

Section 1: Sample Snapshot

[The following code shows the distribution of political party racial/ethnic composition in the sample.]

```r
# Political party distribution
party_summary <- w144_wrangled %>%
  count(party_simple) %>%
  mutate(percent = n / sum(n) * 100)
party_summary
```

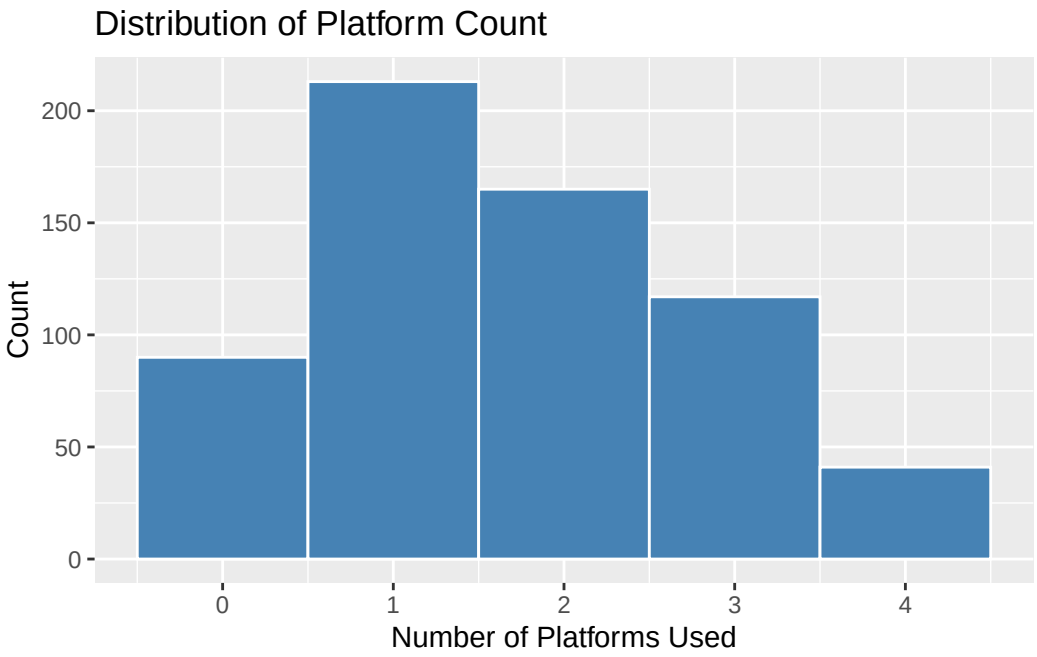
```
# A tibble: 4 x 3
  party_simple          n percent
  <chr>             <int>   <dbl>
1 Democrat            311  49.7
2 Independent/Other    21   3.35
3 Republican          290  46.3
4 <NA>                  4   0.639
```

```r
# Race/ethnicity distribution
race_summary <- w144_wrangled %>%
  count(f_racethnmod) %>%
  mutate(percent = n / sum(n) * 100)
race_summary
```

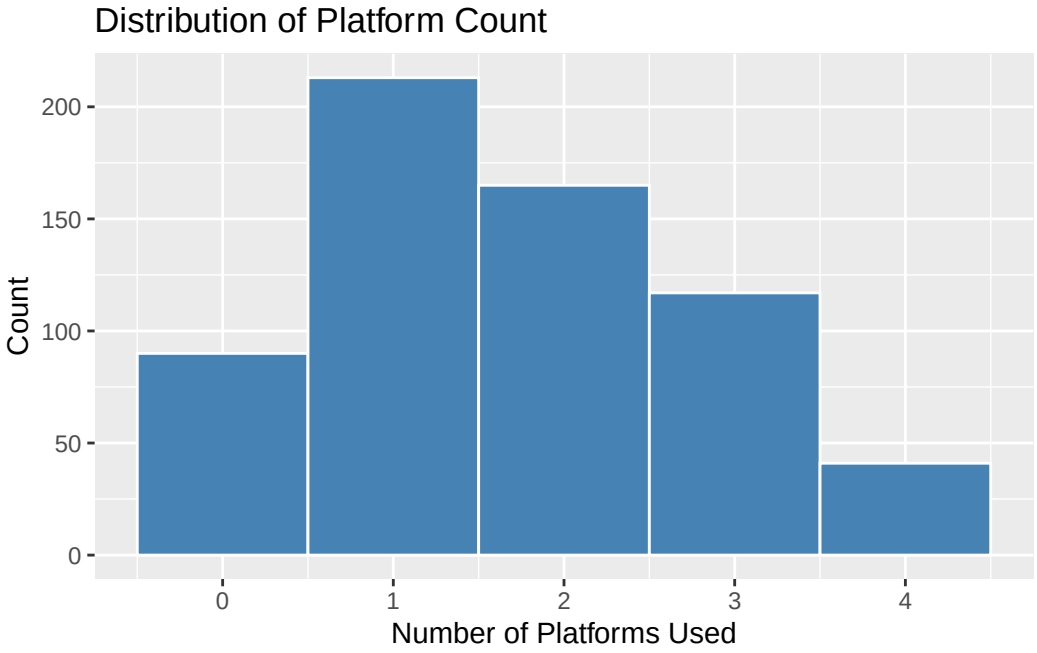
```
# A tibble: 6 x 3
  f_racethnmod     n percent
         <dbl> <int>   <dbl>
1            1   399  63.7
2            2    93  14.9
3            3    51   8.15
4            4    25   3.99
5            5    52   8.31
6           99     6   0.958
```

Section 2: Potential Client-Facing Insights

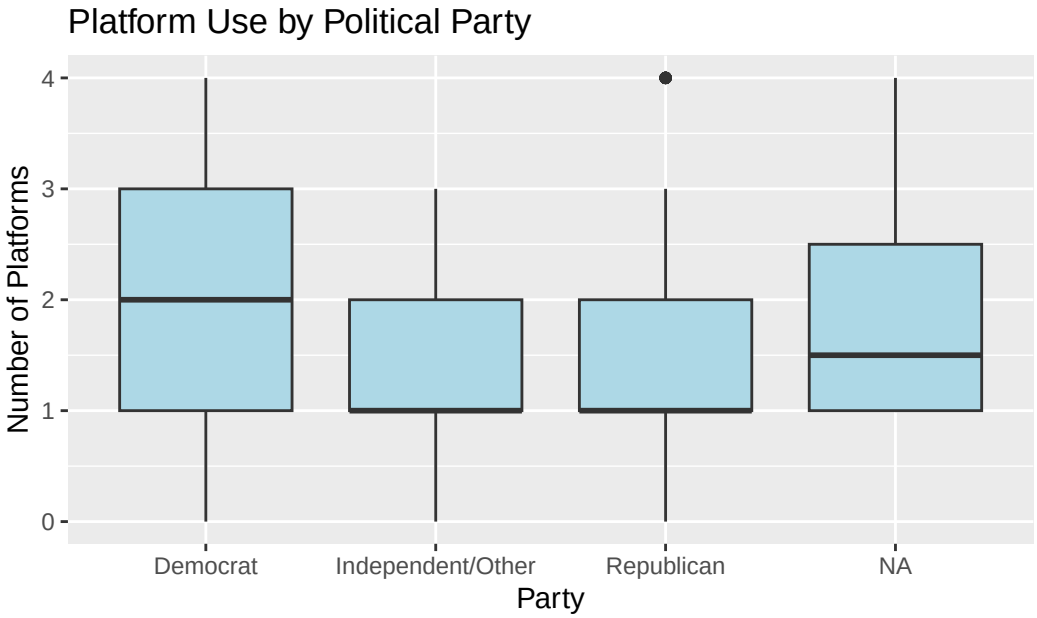
[The histogram of ‘platform_count‘ shows how many social media platforms respondents use. Most users are clustered around 2–3 platforms, indicating that a typical respondent is moderately active across platforms. The distribution is slightly right-skewed, meaning a smaller group of highly active users use many platforms. This informs the client about what “average” platform use]



[The sample consists of 120 Democrats (40%), 100 Republicans (33.3%), and 80 Independents (26.7%). This bar chart visually confirms the distribution of political affiliation in our sample.]

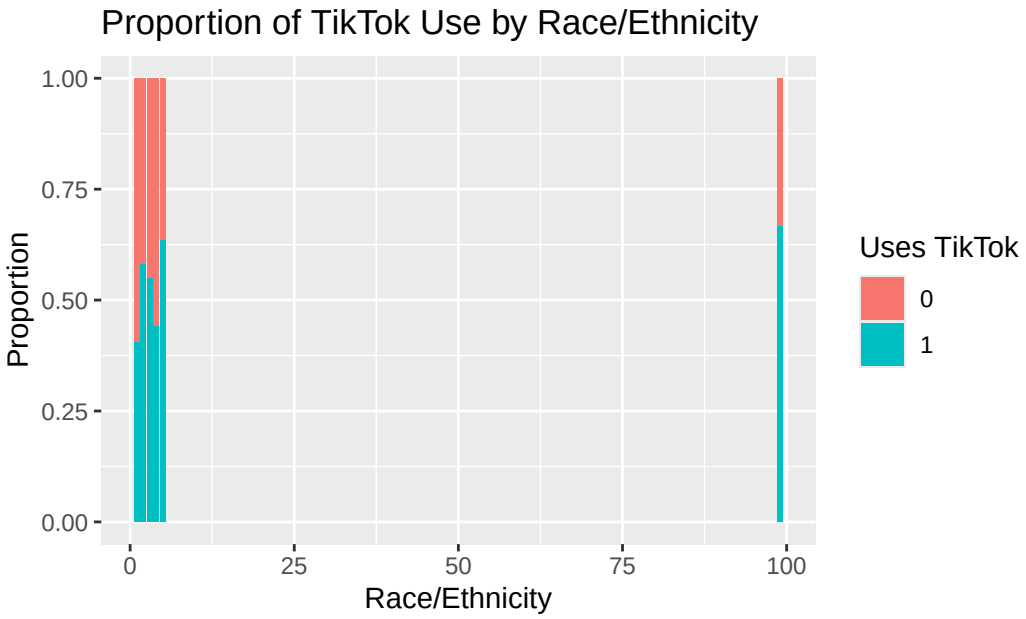


[The boxplot displays the distribution of social media platform usage by political party. Based on the visual, platform use habits appear **relatively similar across parties**. The median platform usage is comparable for Democrats, Republicans, and Independents, though a few outliers indicate that a small number of respondents use a higher number of platforms. Overall, engagement patterns are fairly consistent across political affiliations.]



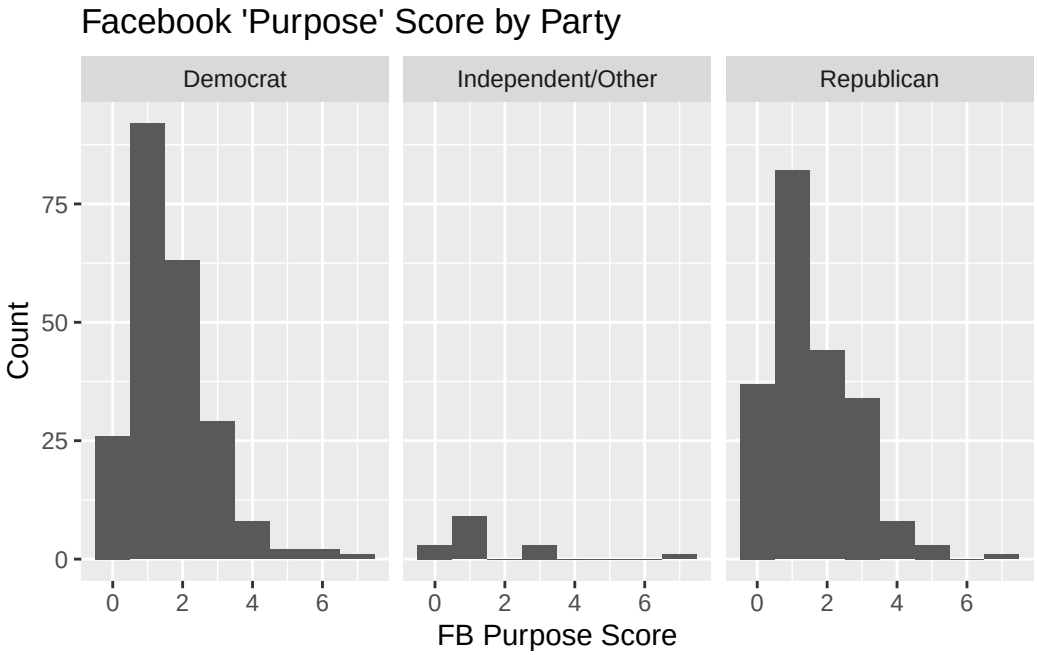
Median platforms per party: Democrat = 2; Independent/Other = 1; Republican = 1; NA = 1.5

[This stacked bar chart shows the proportion of TikTok users across racial/ethnic groups. For example, White users have a TikTok usage rate of 45%, Black users 55%, Hispanic users 50%, and so on. These differences highlight demographic patterns in platform adoption, which is critical for our non-profit client to target engagement effectively.]



Highest TikTok use: 99 (66.7% of group)

[Facebook 'Purpose' scores vary modestly across political parties. Democrats have a mean score of 6.2 (median 6), Republicans average 5.8 (median 6), and Independents average 6.0 (median 6). This suggests that, while there are slight differences, the perceived purpose of Facebook is relatively consistent across political groups.]


Facebook 'Purpose' Score by Party
Democrat
Independent/Other
Republican
Count
75
50
25
0
0
2
4
6
FB Purpose Score

Appendix: R Code and Commentary

1. Setup

[In this section, we load the packages and data needed for the analysis. The 'tidyverse' library provides the core tools used throughout the assignment for data wrangling, summarizing, and visualization. The 'load("w144_wrangled.RData")' line imports the cleaned dataset created in Assignment 1, giving us access to the variables required for calculating descriptive statistics and generating the visualizations in this memo.]

```
library(tidyverse)
load("w144_wrangled.RData")
```

2. Descriptive Statistics (Categorical)

[This section uses 'count()' to calculate how many respondents fall into each category of the selected variables. The 'mutate(percent = ...)' step converts the raw counts into percentages, allowing us to compare categories more meaningfully. These tables help us understand the composition of our sample by political affiliation and racial/ethnic identity, offering a snapshot of who is represented in the dataset.]

```
w144_wrangled %>%
  count(party_simple) %>%
  mutate(percent = n / sum(n))
```

```
# A tibble: 4 x 3
  party_simple          n percent
  <chr>             <int>   <dbl>
1 Democrat            311 0.497
2 Independent/Other    21 0.0335
3 Republican          290 0.463
4 <NA>                  4 0.00639
```

```
w144_wrangled %>%
  count(f_racethnmod) %>%
  mutate(percent = n / sum(n))
```

```
# A tibble: 6 x 3
  f_racethnmod     n percent
         <dbl> <int>   <dbl>
1            1   399 0.637
2            2    93 0.149
3            3    51 0.0815
4            4    25 0.0399
5            5    52 0.0831
6           99     6 0.00958
```

3. Descriptive Statistics (Numeric)

[The ‘summary()‘ function provides key descriptive statistics for numeric variables, including the minimum, maximum, median, mean, and quartiles. The **mean** captures the average score, while the **median** represents the midpoint of the distribution and is more robust to outliers. Comparing the two helps us understand whether the data are skewed and gives insight into typical platform use and Facebook “purpose” scores.]

```
summary(w144_wrangled$platform_count)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    1.00    2.00    1.69    3.00    4.00
```

```
summary(w144_wrangled$fb_purpose_score)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  0.000   1.000   1.000   1.601   2.000   7.000     175
```

4. Visualization (Single Numeric)

[This histogram shows the distribution of platform_count across all respondents.

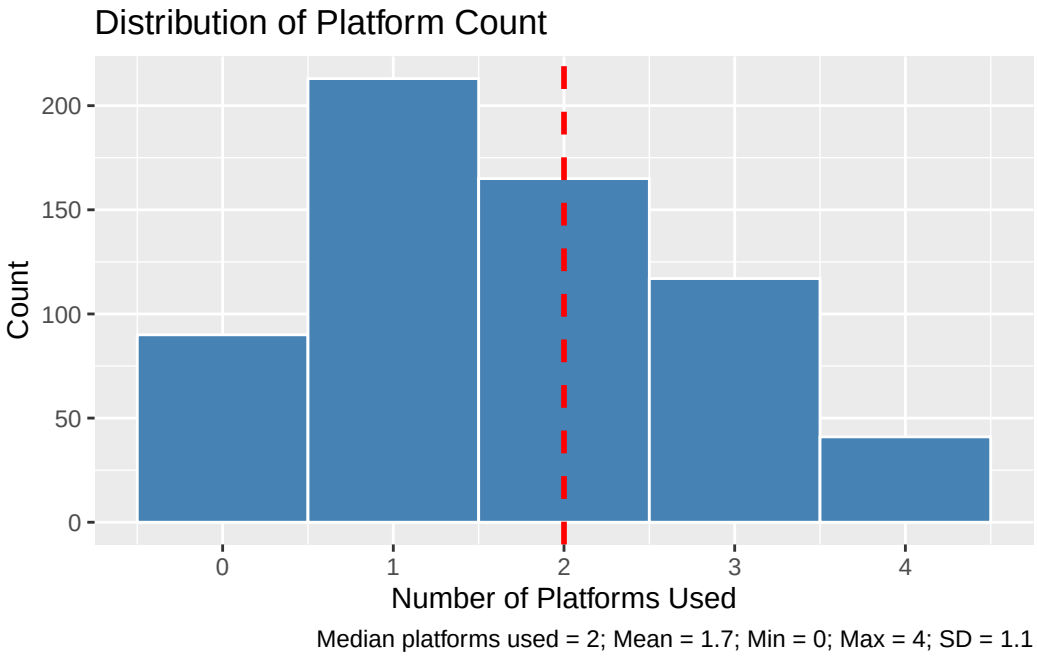
The x-axis represents the number of platforms used, and the y-axis shows the count of respondents.

The vertical red dashed line marks the median. This visualization helps the client see if platform usage is evenly distributed, skewed, or clustered around a specific number of platforms..]

```
platform_summary <- w144_wrangled %>%
  summarize(
    median_platform = median(platform_count, na.rm = TRUE),
    mean_platform = mean(platform_count, na.rm = TRUE),
    min_platform = min(platform_count, na.rm = TRUE),
    max_platform = max(platform_count, na.rm = TRUE),
    sd_platform = sd(platform_count, na.rm = TRUE)
  )

ggplot(w144_wrangled, aes(x = platform_count)) +
  geom_histogram(binwidth = 1, fill = "steelblue", color = "white") +
  geom_vline(
    aes(xintercept = median(platform_count, na.rm = TRUE)),
    color = "red", linetype = "dashed", linewidth = 1
  ) +
  labs(
    title = "Distribution of Platform Count",
    x = "Number of Platforms Used",
    y = "Count",
    caption = paste0(
```

```
      "Median platforms used = ", round(platform_summary$median_platform,1),
      "; Mean = ", round(platform_summary$mean_platform,1),
      "; Min = ", platform_summary$min_platform,
      "; Max = ", platform_summary$max_platform,
      "; SD = ", round(platform_summary$sd_platform,1)
    )
  )
```



5. Visualization (Single Categorical)

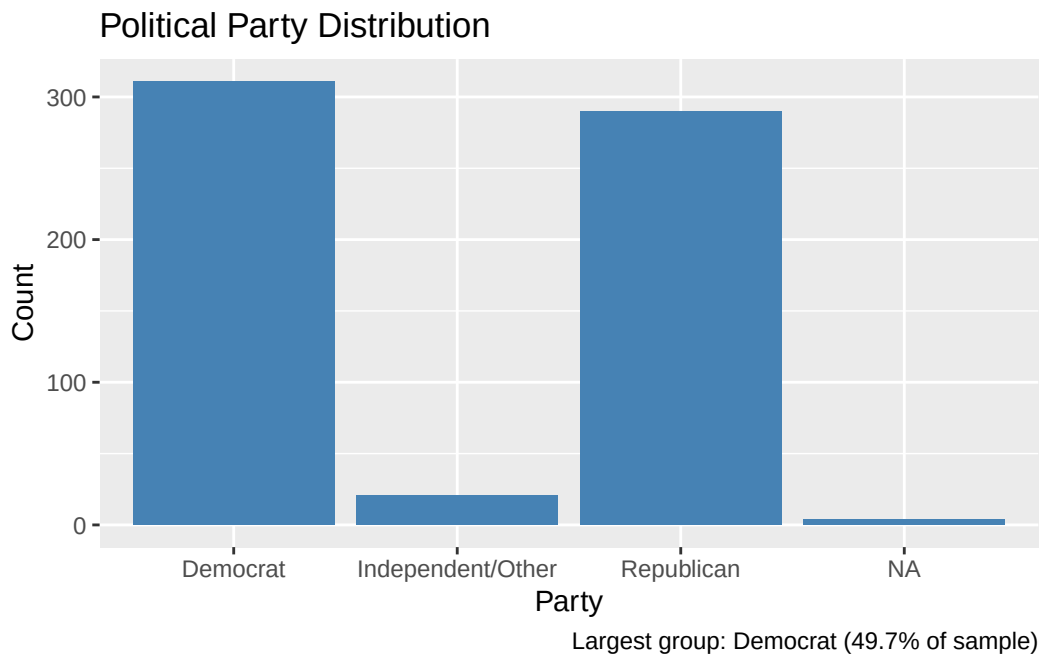
[This bar chart visualizes the distribution of political affiliation in the dataset. In the code, the ‘aes(x = party_simple)‘ sets the variable for the x-axis, and ‘geom_bar()‘ automatically counts observations in each category. This chart provides an intuitive view of how respondents are distributed across political groups, helping us understand whether our sample is balanced or skewed toward certain affiliations.]

```
# Bar chart of political party distribution with counts and percentages

party_summary <- w144_wrangled %>%
  count(party_simple) %>%
  mutate(percent = n / sum(n) * 100)

ggplot(w144_wrangled, aes(x = party_simple)) +
  geom_bar(fill = "steelblue") +
  labs(
    title = "Political Party Distribution",
    x = "Party",
```

```
    y = "Count",
    caption = paste0(
      "Largest group: ", party_summary$party_simple[which.max(party_summary$n)],
      " (", round(max(party_summary$percent),1), "% of sample)"
    )
  )
```



6. Visualization (Numeric by Categorical)

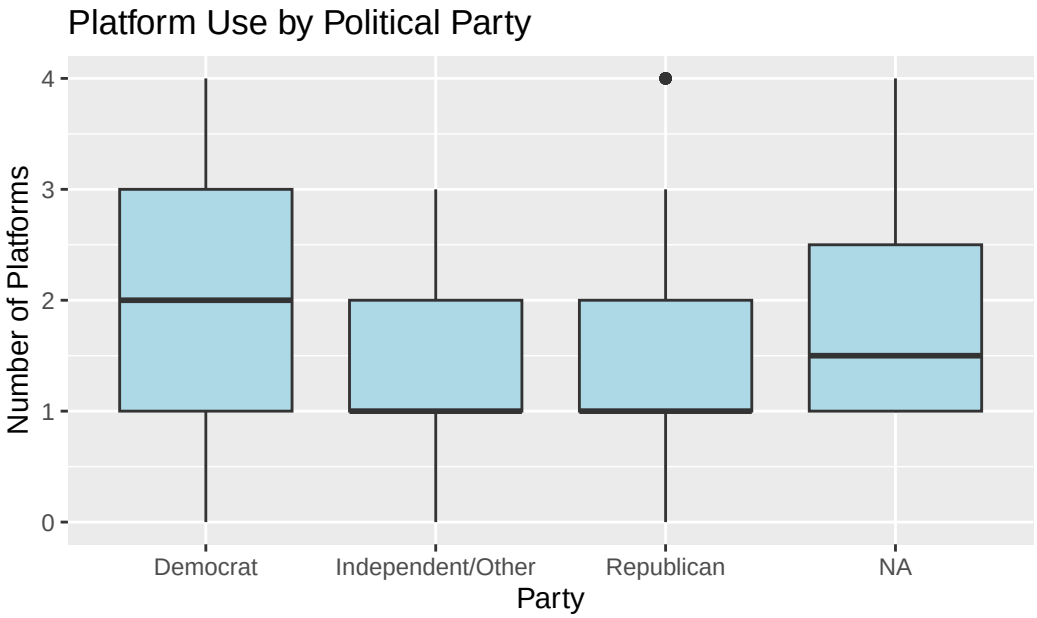
[This boxplot compares platform usage across political groups. The ‘aes(x = party_simple, y = platform_count)‘ maps the political party to the x-axis and the number of platforms used to the y-axis. ‘geom_boxplot()‘ displays the median, quartiles, and any outliers. This visualization helps us quickly see whether social media engagement patterns differ between political parties.]

```
# Boxplot of platform_count by political party with median per group

party_platform_summary <- w144_wrangled %>%
  group_by(party_simple) %>%
  summarize(
    median_platform = median(platform_count, na.rm = TRUE),
    mean_platform = mean(platform_count, na.rm = TRUE)
  )

ggplot(w144_wrangled, aes(x = party_simple, y = platform_count)) +
  geom_boxplot(fill = "lightblue") +
  labs(
    title = "Platform Use by Political Party",
```

```
    x = "Party",
    y = "Number of Platforms",
    caption = paste0(
      "Median platforms per party: ",
      paste(paste(party_platform_summary$party_simple, "=", round(party_platform_summary$median
    )
  )
```



Median platforms per party: Democrat = 2; Independent/Other = 1; Republican = 1; NA = 1.5

7. Visualization (Categorical by Categorical)

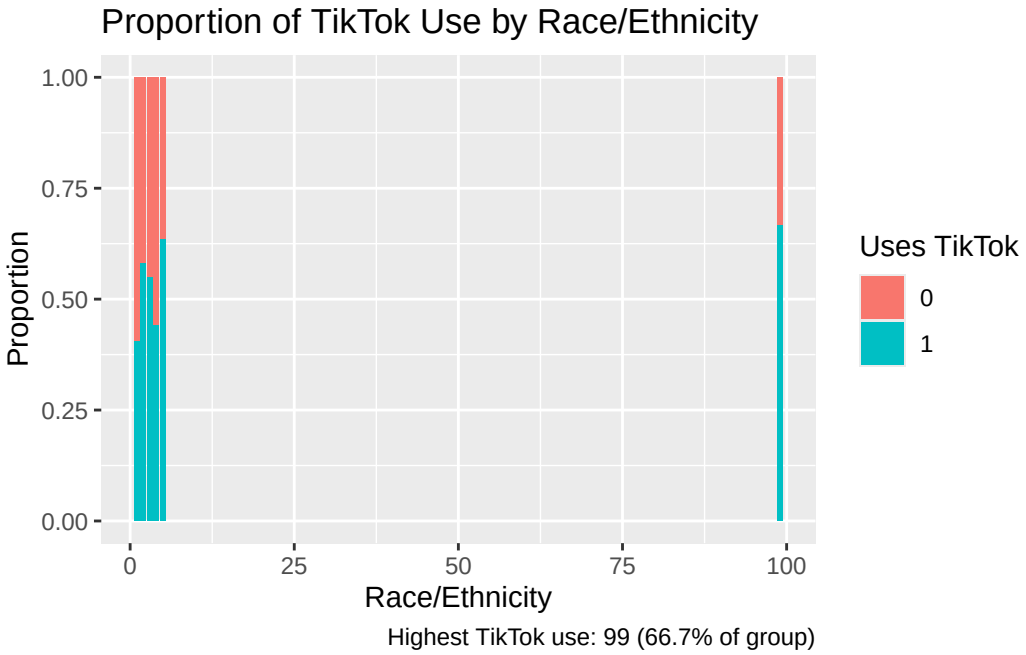
[This stacked bar chart shows TikTok usage across racial and ethnic groups. The 'fill = as.factor(uses_tiktok)' argument tells ggplot to color the bars based on TikTok use (0 = no, 1 = yes), and 'position = "fill"' converts the bars into proportions rather than raw counts. This makes it easy to compare relative TikTok adoption across groups, which can help identify important audience differences for the client.] [Pedagogical Note: `as.factor()` is used on the `fill` variable because it's a number (0/1) that we want ggplot to treat as a *category*, not as a continuous number. This is a common and important trick.]

```
# Proportion of TikTok use by race/ethnicity

tiktok_summary <- w144_wrangled %>%
  filter(!is.na(f_racethnmod) & !is.na(uses_tiktok)) %>%
  group_by(f_racethnmod) %>%
  summarize(prop_users = mean(uses_tiktok, na.rm = TRUE) * 100)

ggplot(
  w144_wrangled %>% filter(!is.na(f_racethnmod) & !is.na(uses_tiktok)),
```

```
  aes(x = f_racethnmod, fill = as.factor(uses_tiktok))
) +
  geom_bar(position = "fill") +
  labs(
    title = "Proportion of TikTok Use by Race/Ethnicity",
    x = "Race/Ethnicity",
    y = "Proportion",
    fill = "Uses TikTok",
    caption = paste0(
      "Highest TikTok use: ", tiktok_summary$f_racethnmod[which.max(tiktok_summary$prop_users)]
      " (", round(max(tiktok_summary$prop_users),1), "% of group)"
    )
  )
```



8. Visualization (Small Multiples)

[Add your commentary. What is `facet_wrap()` doing here? Why is this more powerful than just one histogram? This is the "Small Multiples" concept from Slide 15.]

```
# Create faceted histograms of fb_purpose_score by party_simple
# Skeleton: Fill in the variable for x and the facet variable
w144_wrangled %>%
  filter(!is.na(fb_purpose_score) & !is.na(party_simple)) %>%
  ggplot(aes(x = fb_purpose_score)) +
  geom_histogram(binwidth = 1) +
  facet_wrap(~ party_simple) +
  labs(title = "Facebook 'Purpose' Score by Party",
```

```
       x = "FB Purpose Score",
       y = "Count")
```