week14

Mathi Manavalan

4/24/2020

Libraries

```
library(RMariaDB)
library(tidyverse)
```

Data Import and Cleaning

In this first section, I am creating a connection to the database and exploring what is available to me (commented out, as it is not necessary past exploration).

Here, I am using SQL to collect the data required to answer the question *Do people that use more social media platforms (i.e., a larger count of platforms) have greater acceptance of privacy intrusions (i.e., a higher mean score across the 4-point scale), and do these differences vary by age?* .

To do this, I first inner-joined the *responses* table with the *socialmedia* table using *responses.smu_code* and *socialmedia.code* to collect all the privacy questions' responses as well as social media usage information. Then, I inner-joined this with the *demos* table to be able to collect the ages of participants. (I then ordered by *responses.ident* as I discovered that that is what the tidyverse approach does by default.)

Finally, I saved the resulting table as a data frame in *sql_tbl*.

```
sql_tbl <- dbGetQuery(con, "SELECT ident, age, facebook, twitter, instagram,
                    youtube, snapchat, other, rec_events, rec_products,
                    rec_friends, rec_policial

            FROM responses AS r
            INNER JOIN socialmedia AS s
            ON r.smu_code = s.code

            INNER JOIN demos AS d
            ON r.ident = d.participant_num

            ORDER BY ident
            ")
```

Here, I am going to accomplish the same thing as above, but instead first import each table directly and then combine them using *tidyverse*.

```
demos <- dbGetQuery(con, "SELECT * FROM demos")

responses <- dbGetQuery(con, "SELECT * FROM responses")

socialmedia <- dbGetQuery(con, "SELECT * FROM socialmedia")


tidy_tbl <- responses %>%
  inner_join(socialmedia, by = c("smu_code" = "code")) %>%
  inner_join(demos, by = c("ident" = "participant_num")) %>%
  select(ident, age, facebook, twitter, instagram, youtube, snapchat, other, rec_events,
         rec_products, rec_friends, rec_policial)
```

Now, *sql_tbl* and *tidy_tbl* are essentially identical. I will now clean on *tidy_tbl* to prepare for the necessary analysis.

```
clean <-  tidy_tbl %>%
  select(-ident) %>%
  mutate_all(funs(
    str_replace(.,
                pattern = "NA|Refused",
                replacement = NA_character_))) %>%
  mutate_at(vars(matches("rec")), factor) %>%
  mutate_at(vars(matches("rec")), as.numeric) %>%
  mutate(meanPrivacyScore = rowMeans(select(., c(rec_events, rec_products, rec_friends,
                                                 rec_policial)))) %>%
  drop_na(meanPrivacyScore) %>%
  mutate_all(funs(
    str_replace(.,
              pattern = "Not selected",
              replacement = NA_character_))) %>%
  mutate(smCount = rowSums(!is.na(select(., c(facebook, twitter, instagram, youtube,
                                              snapchat, other))))) %>%
  mutate(age = as.factor(age)) %>%
  mutate(meanPrivacyScore = as.numeric(meanPrivacyScore)) %>%
  select(meanPrivacyScore, age, smCount)
```

Now, in my *clean* data, I just have the necessary variables to answer the research question. (For social media, I am including the response of 'Other' to mean that the participant is using a (1) social media that has not been listed.)

Analysis

I am now running an ordinary least squares regression on X = Number of Social Media Platforms Used and Y = Mean Privacy Score.

```
model <- lm(meanPrivacyScore ~ smCount, data = clean)
summary(model)
```

```
##
## Call:
```

```
## lm(formula = meanPrivacyScore ~ smCount, data = clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.87230 -0.50217  0.00108  0.49783  1.74459
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.132038   0.023795   89.60   <2e-16 ***
## smCount     0.123377   0.008924   13.82   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7304 on 4273 degrees of freedom
## Multiple R-squared:  0.04282,    Adjusted R-squared:  0.04259
## F-statistic: 191.1 on 1 and 4273 DF,  p-value: < 2.2e-16
```

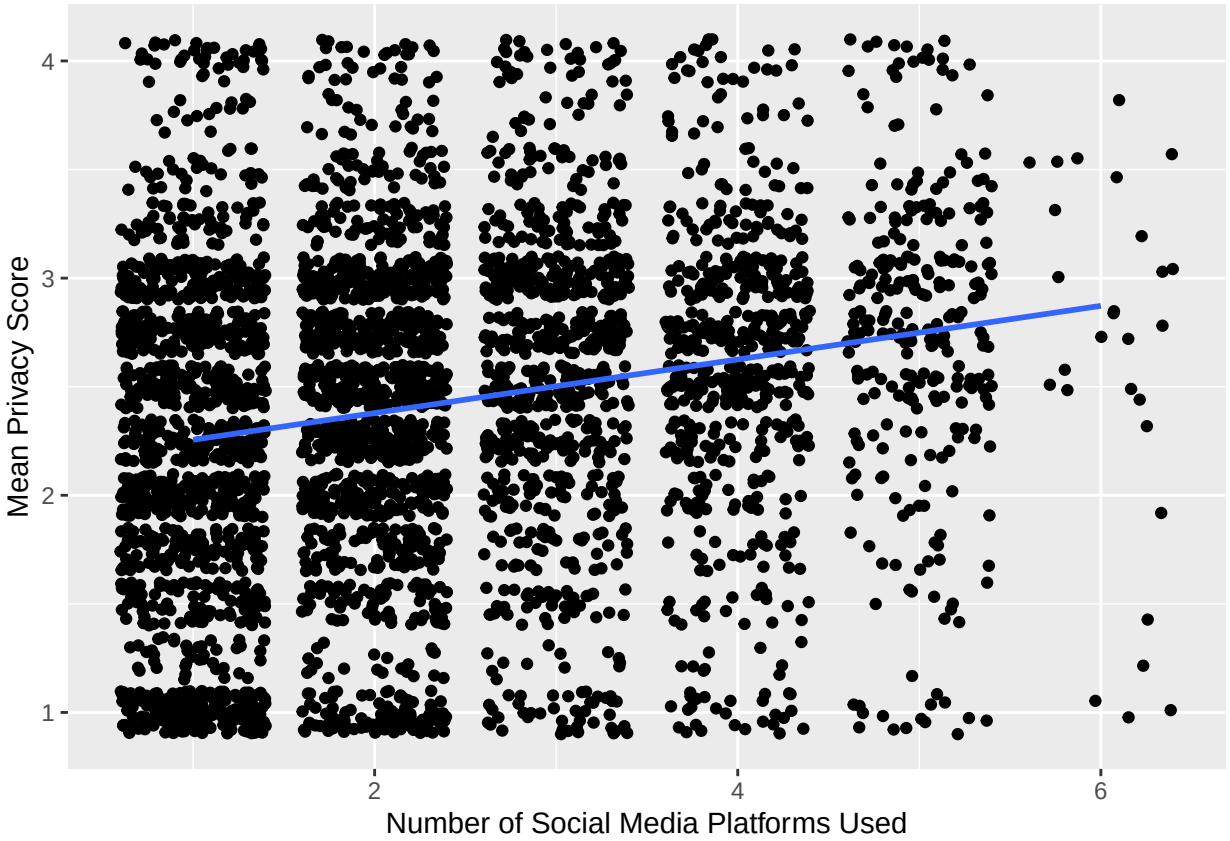
From the above output, we can see that the slope between our X and Y variables is 0.123377 and that it is highly significant as can be seen by the miniscule p-value. In other words, I would say that there is a strong linear relationship between the number of social media platforms used and mean privacy scores.

Visualization

Plot 1

```
ggplot(clean, aes(x = smCount, y = meanPrivacyScore)) +
  geom_jitter() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Number of Social Media Platforms Used", y = "Mean Privacy Score")
```

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



The above plot displays a basic scatterplot of number of social media platforms used vs. mean privacy scores, along with lm model regression line. It seems like a majority of the data lies around 4 social media platforms used or less, and around a mean privacy score of 3 and below.

Plot 2

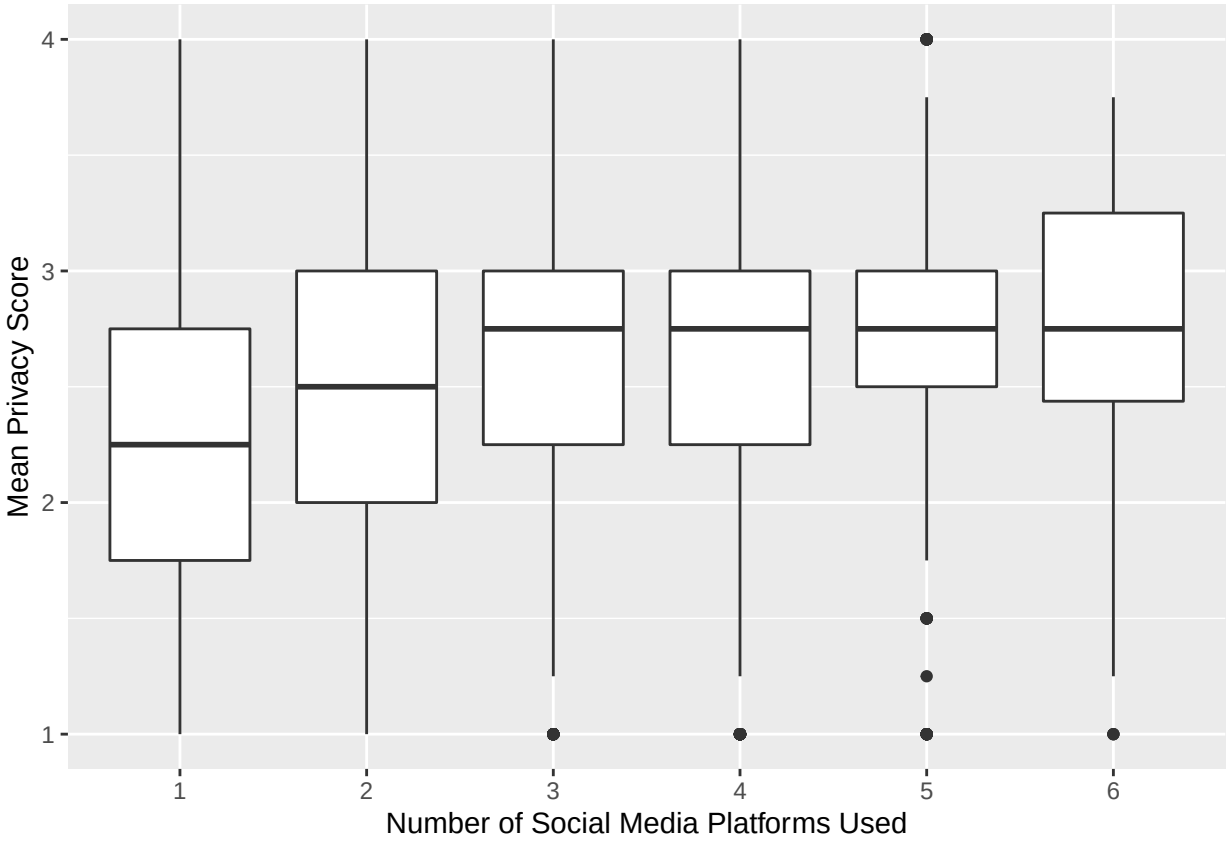
```
ggplot(clean, aes(x = smCount, y = meanPrivacyScore, col = age)) +
  geom_jitter() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Number of Social Media Platforms Used", y = "Mean Privacy Score")
```

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

From the above plot, we can see where the data lies for each category of ages with their corresponding colors and linear regression lines. It seems like the age groups are pretty similar in terms of the regression line slopes (ignoring the NA age group). So I would say that age is not a strong influencing factor here.

Plot 3

```
ggplot(clean, aes(x = as.factor(smCount), y = meanPrivacyScore)) +
  geom_boxplot() +
  labs(x = "Number of Social Media Platforms Used", y = "Mean Privacy Score")
```



Last but not least, the above plot displays a boxplot visualization portraying how mean privacy score changes as the number of social media platforms used increases (similar to what the regression line slope in Plot 1 shows). Here, I can see that the greatest increase in mean privacy scores occurs when the number of social media platforms used increases from 1 to 2 and to 3, but tapers off with the use of more than three social media platforms.