Project One: Exploratory Data Analysis

Tara Amruthur

2023-10-08

Pregnant individuals are advised to refrain from smoking throughout their pregnancy. Studies have shown that smoking during pregnancy can be linked to several adverse birth outcomes, such as low birth weight, restricted head growth, placental problems, and other issues.

Pregnant individuals are advised to refrain from smoking throughout their pregnancy. This is due to the fact that smoking during pregnancy can be linked to several adverse birth outcomes, such as low birth weight, restricted head growth, placental problems, and other issues.

Unfortunately, smoking during pregnancy is a fairly common phenomenon. As of 2014, it has been reported that 8.4% of women smoked at any time during pregnancy (NIH 2021). Additionally, 7-15% of U.S. infants born per year are exposed to smoking during pregnancy (SDP). Not only does this have a negative effect on the children and mothers themselves, but it has an economical cost as well. SDP imposes a $4 billion annual burden on the U.S. economy due to health-care costs. As such, it is important to conduct research on this phenomenon.

A study published in January 2017 involved 800 pregnant people who were smoke exposed (either current smokers, smokers who quit on their own, or exposed to smoke of others), pregnant with only one baby, had access to a telephone and video player, and were randomized to either experimental or control conditions. This study aimed to compare the effectiveness of a tailored video intervention at reducing smoking and environmental tobacco exposure during and after pregnancy (Risica et al. 2017).

A subset of 100 mothers and children involved in this study make up the sample for the current study, conducted by Dr. Micalizzi. This study consists of three laboratory sessions: baseline, 6 month, and 12 month sessions. As part of this study, the parents and children complete self-regulatory assessments and provide self-reported information on substance use.

The data provided consists of 49 unique parents and their children. The demographic information of the parents has been provided in Tables 1 and 2 below.

From Table 1, we see that the mean age of individuals in this dataset is 37. Of the 41 individuals that provided their assigned sex at birth, there were 40 women and 1 man. However, this could be a data quality issue, and is worth looking into further. 61% of individuals in this dataset were white, 15% were Native Hawaiian/Pacific Islanders, 7.3% were biracial (meaning they identified as two races), and 1% were American Indian/Alaska Native. Additionally, 15% marked their race as other and 32% were Hispanic/Latino.

Table 2 provides information on the education and financial status of individuals in this dataset. We can see that 54% of individuals are employed, 29% lack any form of employment, and 17% have part-time employment. As such, it is not surprising that the median income is $46,848. We also have information on the highest level of education completed by individuals in this dataset, and we see that most have attended some college, with 24% having a 4 year degree, 7.3% having a 2 year degree, and 4.9% having gone on to have a post-graduate degree.

Among all variables in Table 1 and employment status in Table 2, we see that there are 8 unknown variables. Upon further investigation, all of these rows are missing values for the same individuals. That is, if there is an individual has a missing value in one of these columns, then they have a missing value in the other seven columns. As this could be due to any number of confounders, it is fair to say that this data is missing at

random or missing not at random. Whether the data is missing at random or missing not at random depends on the data that we have collected.

For the data to be missing not at random, we would need to have collected data on the confounder related to the reason for missingness in this demographic variables. However, the confounder could be one of several different things. These individuals might not have had time to complete the survey, might not have had access to a computer to complete the survey, or might not have felt comfortable revealing this information. There are several different reasons that could be the reason for these missing values, but they are not accounted for in our data. Therefore, I would say this data is missing at random.

Table 1: Data Demographics

Characteristic	**N = 49**[1]
Age	37.0 (35.0, 39.0)
Unknown	8
Sex	
F	40 (98%)
M	1 (2.4%)
Unknown	8
Race	
White	25 (61%)
Native Hawaiian/Pacific Islander	6 (15%)
Other	6 (15%)
Biracial	3 (7.3%)
American Indian/Alaska Native	1 (2.4%)
Unknown	8
Hispanic/Latino	13 (32%)
Unknown	8

[1] Median (IQR); n (%)

Table 2: Education & Financial Status

Characteristic	**N = 49**[1]
Employed	
Full-Time	22 (54%)
No	12 (29%)
Part-Time	7 (17%)
Unknown	8
Highest Level of Education	
Some College	15 (37%)
4-Year Degree	10 (24%)
GED	5 (12%)
2-Year Degree	3 (7.3%)
High School	3 (7.3%)
Some High School	3 (7.3%)
Post-Graduate Degree	2 (4.9%)
Unknown	8
Annual Income	46,848 (20,000, 70,000)
Unknown	12

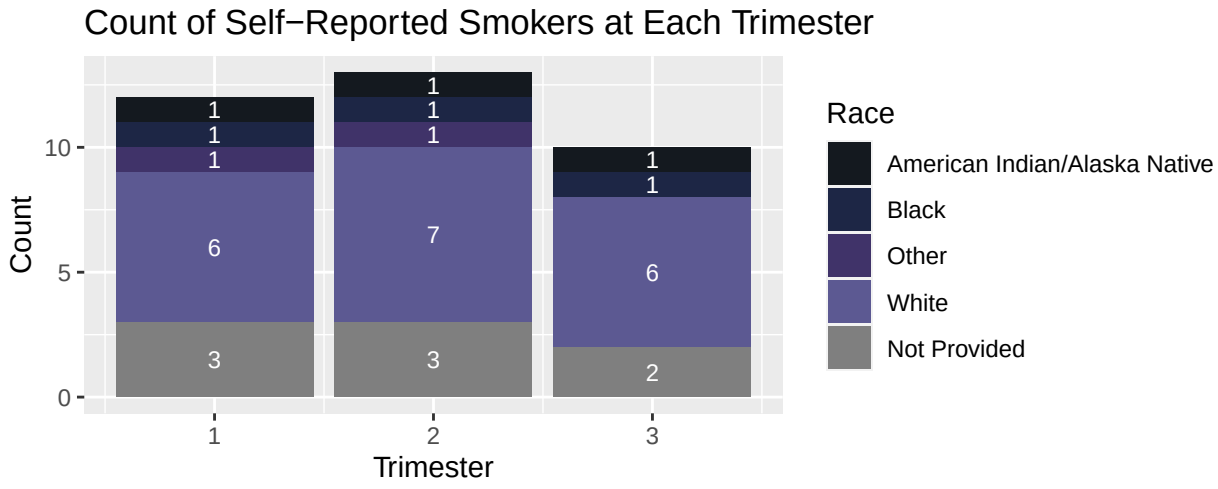
[1] n (%); Median (IQR)

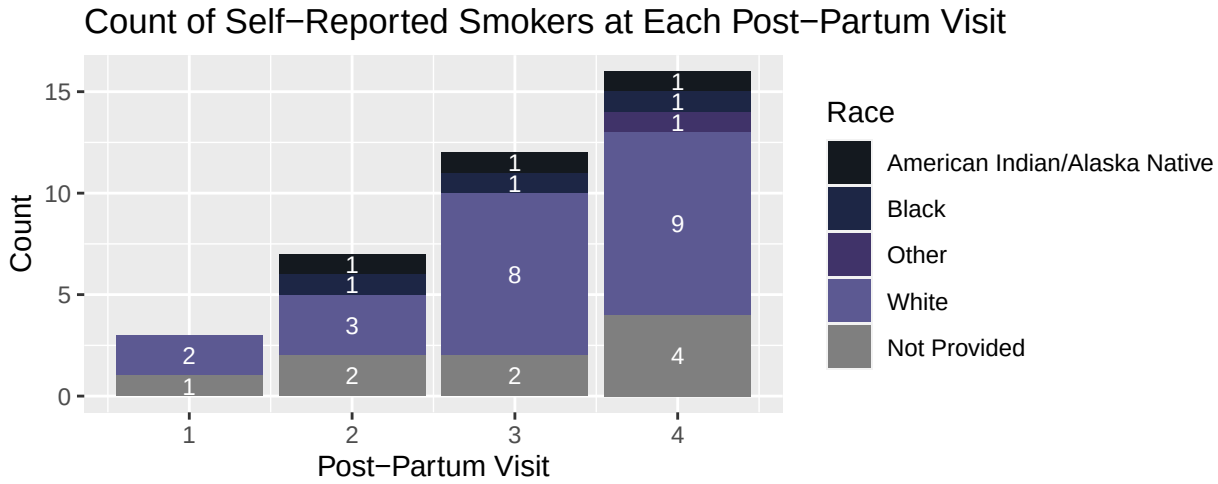
In the original study, data on the smoking habits of the individuals in the study was collected at three different time points during the pregnancy: 16 weeks, 22 weeks, and 32 weeks (the trimesters). Additionally, this data was collected at four different postpartum time points: the first postpartum visit, the second postpartum visit, 12 weeks postpartum, and 6 months postpartum.

In a study posted by the CDC in 2016, they found that Non-Hispanic American Indian or Alaska Native women had the highest prevalence of smoking during pregnancy, indicating potential for race-based confounding (Drake, Driscoll, and Mathews 2018). As such, I wanted to look at the prevalence of smoking at each of the time points by race. Our results seem to depart from the CDC study, as it appears that white women are the predominant smokers at each of the time points. However, as we are looking at a very small sample size, and there are few individuals who have not reported their race, it is hard to say if this is representative of the general population.

The plots do show that the total number of self-reported smokers at each trimester hovers around 10, with its peak during the second trimester. However, at the first postpartum visit, the numbers plummet, with only 3 individuals identifying as smokers during this time period. As time goes on, we see the numbers start to increase, with 16 total smokers by the fourth visit (at 6 months postpartum). However, the numbers by race appear to be largely constant: during the second postpartum visit, 1 American Indian/Alaska Native and 1 Black mother report being smokers, and this number doesn't change over the next two visits.

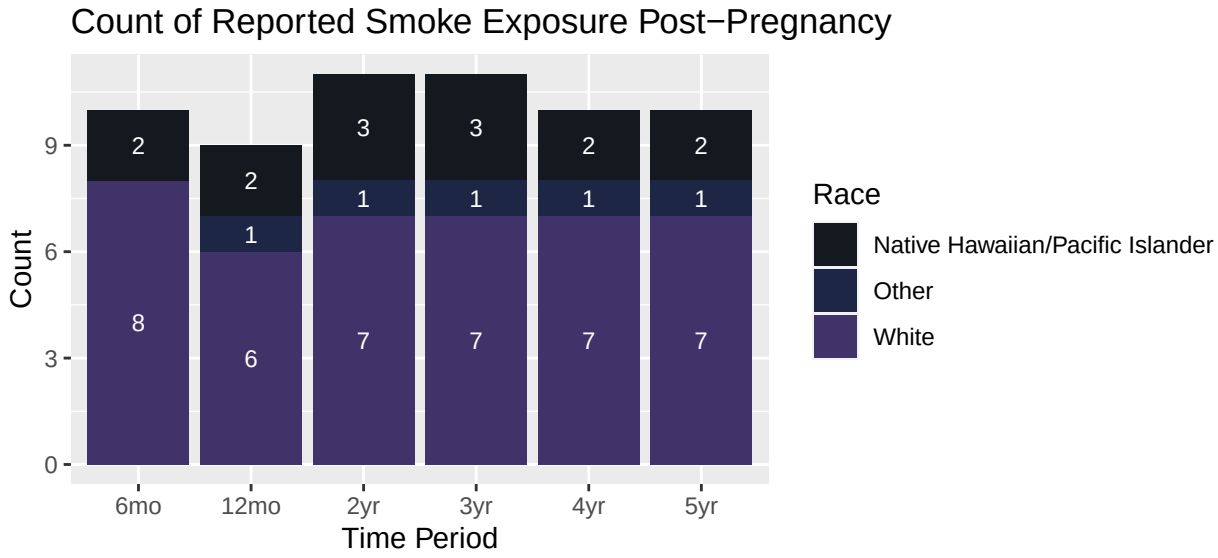
As these measures are self-reported, there is a chance that they might be incorrect; many mothers might not want to admit their smoking behaviors immediately following the birth of their child. Additionally, for the collection of race data, individuals are able to check off which race they identify with. As some individuals may have checked off multiple races, there may be a few cases of double counting.





These values from the original study unfortunately only provide information on SDP and smoke exposure up to 6 months postpartum. The new study has information on smoke exposure up to 5 years postpartum. The information has been shown in the visualization before, stratified by race. Only individuals who identify as Native Hawaiian/Pacific Islanders, White, or Other appear to have self-reported smoke exposure by either them or a partner. No American Indian or Alaskan Native women have self-reported smoke exposure post-pregnancy, which does not align what we would expect based on the CDC findings.

Smoke exposure results seem to fluctuate slightly, but for the most part, we have around 10 individuals reporting smoke exposure at each of these time points.



Though these self-reported measures provide information on whether or not an individual was smoking, it doesn't tell us how heavy of a smoker they are, whether they smoke frequently or smoke only in passing. The data collected includes information on urine cotinine values at 34 weeks into the pregnancy and 6 months postpartum. The cut-off values for the different levels of smokers have been a subject of lots of research, but for this analysis, I am using the following criteria: nonsmokers will have levels less than 100 ng/mL, passive smokers would have levels between 100 to 500 ng/mL, and active smokers would have levels greater than 500 ng/mL (Hellemons et al. 2015). Table 3 shows the numbers of women that fall in each of these categories both during their third trimester as well as 6 months postpartum.

It appears that at 34 weeks pregnant, most women in the study would be classified as non-smokers, due to their urine cotinine values. There are 8 that would be classified as passive smokers, and 11 that this information has not been recorded for. At 6 months postpartum, we see one individual classified as an active smoker, 26 classified as non-smokers, and 11 classified as passive smokers.

These numbers differ slightly with the values in the plots above, but as the cut-off numbers are only estimates, that could be the reason for this difference.

Table 3: Smoking Level by Urine Cotinine Pre- and Post- Pregnancy

Characteristic	**N = 49**[1]
Smoking Level (34 Week Gestation)	
nonsmoker	30 (79%)
passive	8 (21%)
Unknown	11
Smoking Level (6 Months Postpartum)	
active smoker	1 (2.6%)
nonsmoker	26 (68%)
passive	11 (29%)

Unknown	11

[1] n (%)

Some of the consequences of SDP exposure include externalizing behavior such as attention-deficit/hyperactivity disorder (ADHD), conduct disorder, or substance use. Therefore, I wanted to see if these patterns are prevalent in the data. To start, I looked at substance use among children of active or passive smokers and compared it to substance use among children of nonsmokers. Substance use is defined as any child that has tried e-cigarettes, cigarettes, alcohol, or marijuana. The results of this analysis are shown in Table 4.

It would appear that 25% of children of passive smokers at 34 weeks have engaged in some form of substance use, compared to 36% of children of passive smokers at 6 months postpartum. Additionally, we see that 13% of children of nonsmokers at 34 weeks have engaged in some form of substance use, compared to 11% of children of nonsmokers at 6 months postpartum. This would indicate that in the case of smoking, children who have been exposed to it post-birth are more likely to engage in substance use, while those who have not been exposed to it after birth are less likely. Again, this is a small sample size, so we are not able to draw any conclusions, but this could be indicative of a time effect of smoke exposure on a child's substance use. What is surprising, however, is that there is no substance use among the child of a heavy smoker at 6 months postpartum. As there is only one child of a heavy smoker, they could just be an outlier.

Table 4: Percent of Child Substance Use by Smoking Level at Different Periods

Smoking Level	Time Period	Percent of Child Substance Use
passive	34 weeks	25.00000
passive	6 months postpartum	36.36364
nonsmoker	34 weeks	13.33333
nonsmoker	6 months postpartum	11.53846
active smoker	6 months postpartum	0.00000

Furthermore, I wanted to see which substances make up the majority of total substance use. This can be seen in Table 5. Of children engaged in substance use, alcohol is the most popular, with 71% using alcohol, 42% each use marijuana and e-cigarettes, and only 14% use cigarettes.

Table 5 also provides information not just on the percentage of substance use by medium, but the average number of days that a substance has been used in the past 30 days. Marijuana appears to be, by far, the most frequently used substance, being used, on average, 11 of the past 30 days. Alcohol comes second, being used on average, 2.75 days of the past 30 days. E-cigarettes and cigarettes are not used as frequently on the other hand. These are self-reported measures by the children, which could be cause for concern, as they might not want to openly admit to using some of these substances.

Table 5: Substance Use by Medium

Substance	Percent of Substance Use	Average Days Used in Past 30 Days
E Cigarette	42.85714	1.00
Marijuana	42.85714	11.00
Alcohol	71.42857	2.75
Cigarette	14.28571	0.00

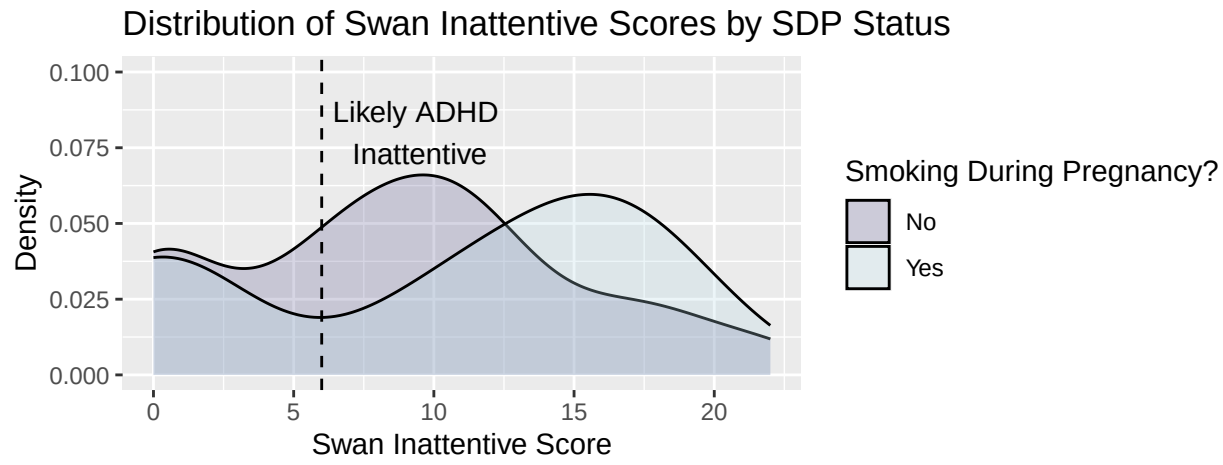
As mentioned earlier, there is a link between smoking during pregnancy and ADHD. In order to see if this trend is prevalent in our data, I created a variable for smoking during pregnancy. In order to determine the

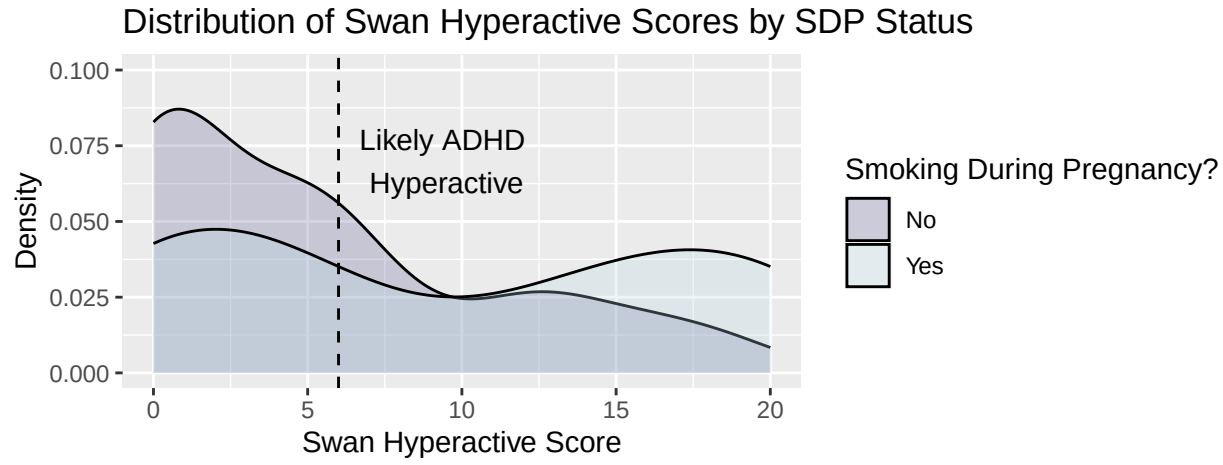
appropriate definition for smoking during pregnancy, I looked both at women who reported smoking behavior at any point during pregnancy and women who reported smoking behavior at all points during pregnancy. I found that 10 of the 14 (~71%) women who reported smoking at some point during pregnancy smoked throughout the entire pregnancy, so I chose to define SDP as smoking at any point during pregnancy.

To look at the relationship between SDP and ADHD, I chose to look at how the distribution of SWAN Inattentive and Hyperactive scores differs among children who were exposed to SDP and those who were not. It is worth nothing that if any of the SWAN scores are higher than 6, it indicates that the child is likely ADHD. This has been denoted in the plots below by the dotted line.

In both plots, we see that the distribution of SWAN scores of children exposed to SDP tend to peak more towards the higher values, while the SWAN scores for children not exposed to SDP peak more towards the lower values. This difference is most clear when looking at SWAN Hyperactive scores, as the peak for children not exposed to SDP is closer to 0, while the peak for children exposed to SDP is around 17.

In the plot of SWAN Inattentive scores, both distributions appear to peak at SWAN scores above 6, but we do see that the distribution of children exposed to SDP is shifted slightly more to the right than the distribution of children not exposed to SDP.



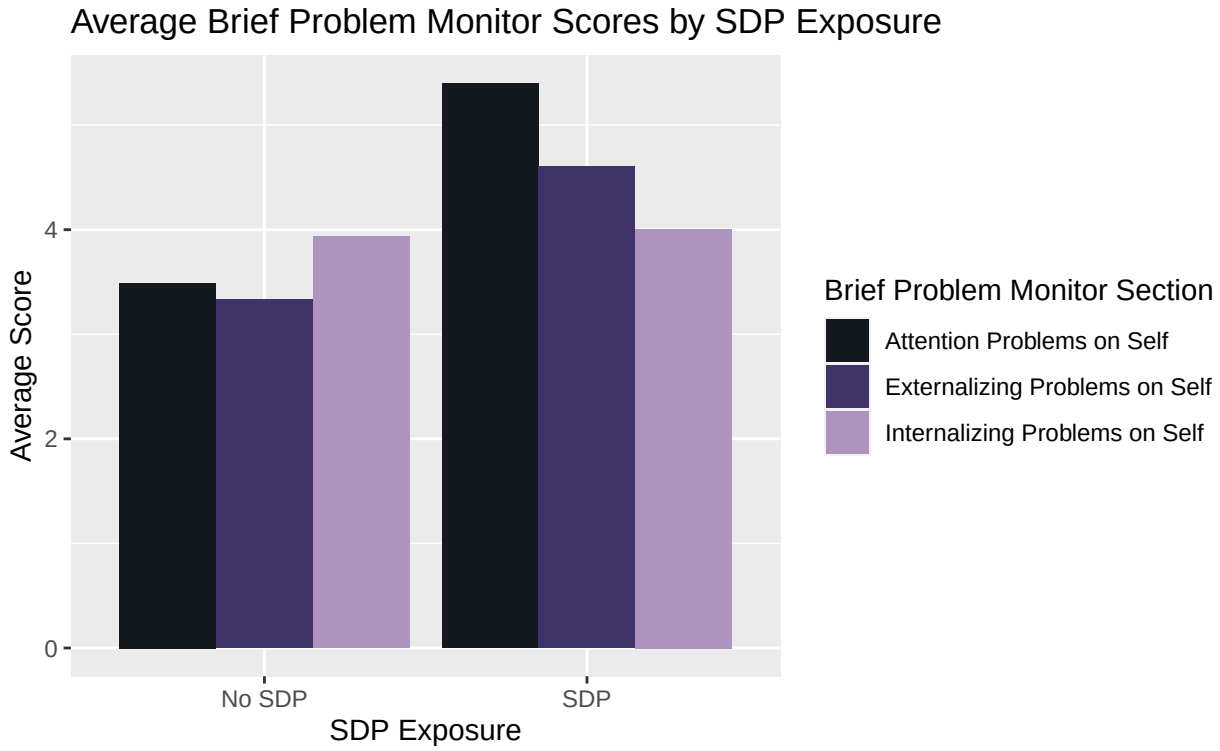


We also want to identify any differences in self-regulatory problems as a result of smoking during pregnancy. Contrary to ADHD, conduct disorder, and substance use, this is an internalizing behavior. For this study, these were measured through two different ways: the Brief Problem Monitor and Emotion Regulation Questionnaire.

The Brief Problem Monitor is used for normed multi-informant assessment of children's functioning & responses to interventions. In this study, it has been completed both by parents and children. There are

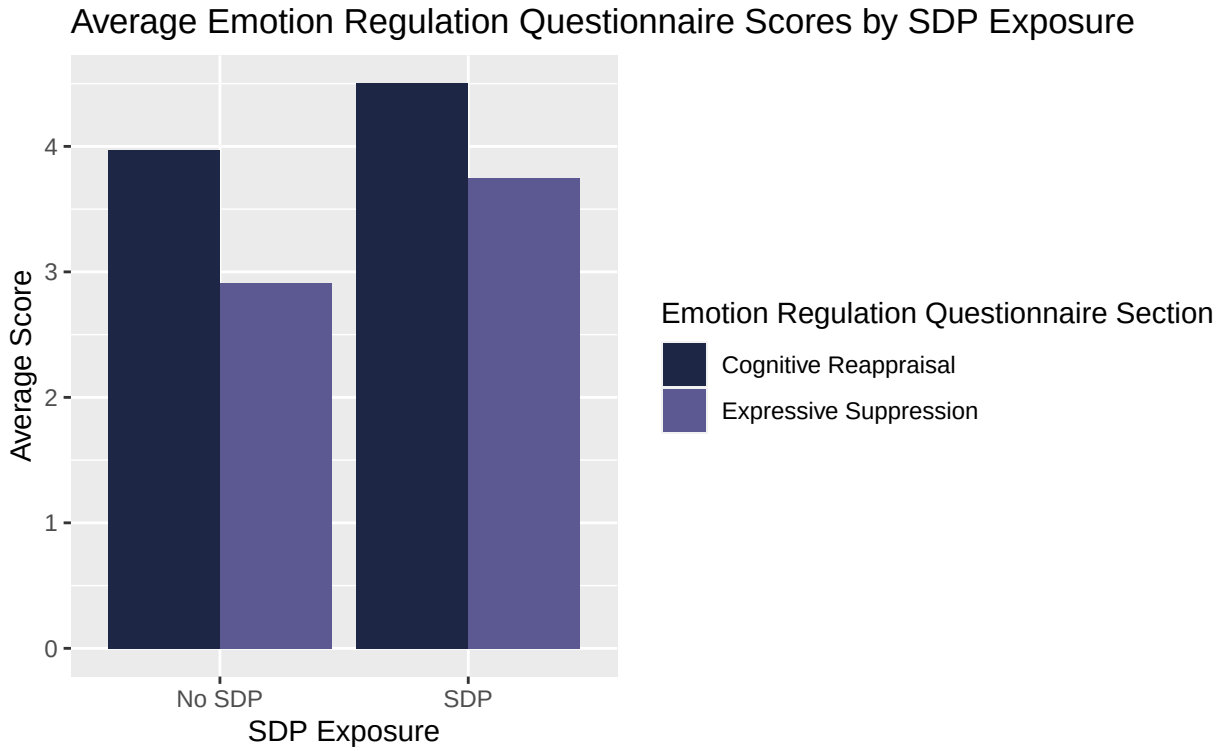
separate scales for attention problems, internalizing problems, and externalizing problems (ASEBA 2022).

The average BPM score for attention problems and externalizing problems on self appears to be higher for children exposed to SDP than children who have not been exposed to SDP. This could explain the higher SWAN scores and the increased substance use among children exposed to SDP. However, we do see that there does not appear to be much difference in the scores for internalizing problems on self between children exposed to SDP and those not exposed.



We have also been provided scores form the Emotion Regulation Questionnaire. The Emotion Regulation Questionnaire consists of 10 questions that involve two distinct aspects of an individual's emotional life: their emotional experience (i.e. what they feel inside) and their emotional expression (i.e. how they show their emotions in the way they talk, gesture, or behave). Each of the 10 questions are answered using a scale ranging from 1-7, with 1 indicating they strongly disagree with a statement and 7 indicating they strongly agree. The higher the score, the greater the use of cognitive reappraisal or emotion suppression (Gross and John 2003).

When comparing the responses to the emotion regulation questionnaire by smoking during pregnancy, we see that children who have been exposed to SDP use expressive suppression more than children who have not been exposed. Additionally, the average score for cognitive reappraisal appears to be higher for children exposed to SDP than those not exposed, although this difference appears to be relatively small, with only a difference of approximately 0.5.



This analysis has provided several points worth further research in analyzing the effects of pre- and post-natal smoke exposure to children. Not only that, but we have investigated the rates of smoking among the women in the study both during and after pregnancy. We have analyzed race-based differences, as it could be a potential confounder, while also determining how heavy of a smoker these individuals are.

Our analysis confirmed the belief that that smoking during pregnancy can be associated with ADHD. We have seen that smoking during pregnancy can lead to higher scores on the SWAN scores, indicating either Inattentive or Hyperactive ADHD. These differences seem more pronounced with Hyperactive ADHD however. We also see more problems with attention and externalizing behaviors among children exposed to SDP compared to those not exposed, though we do not see a similar pattern when it comes to internalizing behaviors. This could be an area for further research, as the sample size in this study is too small to make any generalizations.

We also find that children exposed to SDP are more likely to utilize expressive suppression and cognitive reappraisal than children without SDP. This could be indicative that children with SDP are more in-tune with their own emotions, as they are more practiced with controlling them and re-framing certain situations. However, there would need to be more data collected to confirm this hypothesis.

Ultimately, it would appear that there are negative effects of smoking during pregnancy on children, but in order to make any conclusive statements, we would need more data as well as measures of smoking behavior that are not self-reported, due to the imperfect nature of self-reporting. That being said, this analysis has highlighted associations with SDP that are worth looking into.

Code Appendix

```
knitr::opts_chunk$set(echo = FALSE)
suppressPackageStartupMessages(library(tidyverse, warn.conflicts = FALSE))
library(ggplot2)
library(gtsummary)
suppressPackageStartupMessages(library(kableExtra))
library(ghibli)
suppressPackageStartupMessages(library(gridExtra))

data <- read.csv("../Projects/project_1.csv")

# Data Pre-Processing (1) Turn race columns from
# dummy-coded into single race column.  Note: in
# cases where there are two races checked, store
# them as biracial.  (2) Convert sex, ethnicity,
# employment, and education columns from numbers
# into the actual values (taken from the data
# codebook).  (3) For income value 250,000 remove
# the extra space and column.  Convert income
# column to numeric.  (4) Convert smoking column
# values to 1 or 0 for smoking or no smoking.
# (5) Create new column for smoking level at 34
# weeks gestation period and six months
# postpartum. The smoking level is calculated
# based on the urine cotinine values. The cut-off
# values for each level have been taken from The
# Transplantation Journal. See citations for more
# details.
data <- data %>%
    mutate(parent_race = case_when(paian == 1 & pwhite ==
        1 | paian == 1 & pnhpi == 1 ~ "Biracial", paian ==
        1 ~ "American Indian/Alaska Native", pasian ==
        1 ~ "Asian", pnhpi == 1 ~ "Native Hawaiian/Pacific Islander",
        pblack == 1 ~ "Black", pwhite == 1 ~ "White",
        prace_other == 1 ~ "Other")) %>%
    mutate(parent_sex = case_when(psex == 1 ~ "F",
        psex == 0 ~ "M", TRUE ~ NA)) %>%
    mutate(ethnic = case_when(pethnic == 1 ~ "Yes",
        pethnic == 0 ~ "No", TRUE ~ NA)) %>%
    mutate(employed = case_when(employ == 0 ~ "No",
        employ == 1 ~ "Part-Time", employ == 2 ~ "Full-Time",
        TRUE ~ NA)) %>%
    mutate(education = case_when(pedu == 0 ~ "Some High School",
        pedu == 1 ~ "High School", pedu == 2 ~ "GED",
        pedu == 3 ~ "Some College", pedu == 4 ~ "2-Year Degree",
        pedu == 5 ~ "4-Year Degree", pedu == 6 ~ "Post-Graduate Degree")) %>%
    mutate(income = case_when(income == "250, 000" ~
        "250000", TRUE ~ income)) %>%
    mutate(income = as.numeric(income)) %>%
    mutate(mom_smoke_16wk = case_when(mom_smoke_16wk ==
        "1=Yes" ~ 1, mom_smoke_16wk == "2=No" ~ 0),
        mom_smoke_22wk = case_when(mom_smoke_22wk ==
            "1=Yes" ~ 1, mom_smoke_22wk == "2=No" ~
```

```
            0), mom_smoke_32wk = case_when(mom_smoke_32wk ==
            "1=Yes" ~ 1, mom_smoke_32wk == "2=No" ~
            0), mom_smoke_pp1 = case_when(mom_smoke_pp1 ==
            "1=Yes" ~ 1, mom_smoke_pp1 == "2=No" ~
            0), mom_smoke_pp2 = case_when(mom_smoke_pp2 ==
            "1=Yes" ~ 1, mom_smoke_pp2 == "2=No" ~
            0), mom_smoke_pp12wk = case_when(mom_smoke_pp12wk ==
            "1=Yes" ~ 1, mom_smoke_pp12wk == "2=No" ~
            0), mom_smoke_pp6mo = case_when(mom_smoke_pp6mo ==
            "1=Yes" ~ 1, mom_smoke_pp6mo == "2=No" ~
            0)) %>%
    mutate(smoking_level_34wk = case_when(is.na(cotimean_34wk) ~
        NA, cotimean_34wk < 100 ~ "nonsmoker", cotimean_34wk <
        500 ~ "passive", TRUE ~ "active")) %>%
    mutate(smoking_level_pp6mo = case_when(is.na(cotimean_pp6mo) ~
        NA, cotimean_pp6mo < 100 ~ "nonsmoker", cotimean_pp6mo <
        500 ~ "passive", TRUE ~ "active smoker"))

# Create Demographic Table
data %>%
    mutate(page = as.numeric(page)) %>%
    select(page, parent_sex, parent_race, ethnic) %>%
    tbl_summary(type = page ~ "continuous", label = list(page ~
        "Age", parent_sex ~ "Sex", parent_race ~ "Race",
        ethnic ~ "Hispanic/Latino"), sort = list(everything() ~
        "frequency")) %>%
    as_gt() %>%
    gt::tab_header(title = "Table 1: Data Demographics")

# Create Financial Table
data %>%
    select(employed, education, income) %>%
    tbl_summary(type = income ~ "continuous", label = list(employed ~
        "Employed", education ~ "Highest Level of Education",
        income ~ "Annual Income"), sort = list(everything() ~
        "frequency")) %>%
    as_gt() %>%
    gt::tab_header(title = "Table 2: Education & Financial Status")
# Creating a race lookup table for the parent
race_lookup <- data %>%
    gather(key = "race", value = "value", c(paian,
        pasian, pblack, pwhite, pnhpi, prace_other)) %>%
    filter(!is.na(value)) %>%
    filter(value == 1) %>%
    select(parent_id, race)

# Pivot the data to get whether or not the mother
# smoked during each trimester.  16 weeks
# represents trimester 1, 22 weeks represents
# trimester 2, and 32 weeks represents trimester
# 3. This is used to create the trimester column.
smoking_trimester <- data %>%
    gather(key = "time", value = "smoking", c(mom_smoke_16wk,
```

```
        mom_smoke_22wk, mom_smoke_32wk)) %>%
    mutate(trimester = case_when(time == "mom_smoke_16wk" ~
        1, time == "mom_smoke_22wk" ~ 2, time == "mom_smoke_32wk" ~
        3))

# Get the number of smokers at each trimester
# stratified by race.
smoking_by_race <- smoking_trimester %>%
    select(parent_id, trimester, smoking) %>%
    left_join(race_lookup, by = "parent_id", relationship = "many-to-many") %>%
    group_by(trimester, race) %>%
    summarize(count = sum(smoking, na.rm = TRUE), .groups = "drop_last") %>%
    filter(count > 0)

# Pivot the data to get whether or not the mother
# smoked during each postpartum visit.
post_partum_smoking <- data %>%
    gather(key = "time", value = "smoking", c(mom_smoke_pp1,
        mom_smoke_pp2, mom_smoke_pp12wk, mom_smoke_pp6mo)) %>%
    mutate(pp_visit = case_when(time == "mom_smoke_pp1" ~
        1, time == "mom_smoke_pp2" ~ 2, time == "mom_smoke_pp12wk" ~
        3, time == "mom_smoke_pp6mo" ~ 4))

# Get the number of smokers at each postpartum
# visit stratified by race.
pp_smoking_by_race <- post_partum_smoking %>%
    select(parent_id, pp_visit, smoking) %>%
    left_join(race_lookup, by = "parent_id", relationship = "many-to-many") %>%
    group_by(pp_visit, race) %>%
    summarize(count = sum(smoking, na.rm = TRUE), .groups = "drop_last") %>%
    filter(count > 0)

# Make two plots: one for smoking during
# pregnancy & one for postpartum smoking.
smoking_plot_1 <- ggplot(smoking_by_race, aes(x = trimester,
    y = count, fill = race, label = count)) + geom_bar(stat = "identity") +
    geom_text(size = 3, position = position_stack(vjust = 0.5),
        color = "white") + scale_fill_manual(name = "Race",
    labels = c("American Indian/Alaska Native", "Black",
        "Other", "White", "Not Provided"), values = ghibli_palette("LaputaMedium")) +
    labs(title = "Count of Self-Reported Smokers at Each Trimester",
        x = "Trimester", y = "Count")

smoking_plot_2 <- ggplot(pp_smoking_by_race, aes(x = pp_visit,
    y = count, fill = race, label = count)) + geom_bar(stat = "identity") +
    geom_text(size = 3, position = position_stack(vjust = 0.5),
        color = "white") + scale_fill_manual(name = "Race",
    labels = c("American Indian/Alaska Native", "Black",
        "Other", "White", "Not Provided"), values = ghibli_palette("LaputaMedium")) +
    labs(title = "Count of Self-Reported Smokers at Each Post-Partum Visit",
        x = "Post-Partum Visit", y = "Count")
grid.arrange(smoking_plot_1, smoking_plot_2, nrow = 2)
# Pivot the data to get values for smoke exposure
```

```
# at each time point postpartum.  This data comes
# from the newer study and includes time periods
# at 6 months, 12 months, 2 years, 3 years, 4
# years, and 5 years.
smoke_exp <- data %>%
    gather(key = "time", value = "smoke_exposure",
        c(smoke_exposure_6mo, smoke_exposure_12mo,
            smoke_exposure_2yr, smoke_exposure_3yr,
            smoke_exposure_4yr, smoke_exposure_5yr)) %>%
    select(parent_id, time, smoke_exposure) %>%
    mutate(time = substring(time, 16))

# Get the number of self-reported smoke exposures
# at each time point stratified by race.
smoke_exp_over_time <- smoke_exp %>%
    left_join(race_lookup, by = "parent_id", relationship = "many-to-many") %>%
    filter(!is.na(smoke_exposure)) %>%
    group_by(race, time) %>%
    summarize(count = sum(smoke_exposure, na.rm = TRUE),
        .groups = "drop_last") %>%
    filter(count > 0)

# Order the x-axis
level_order <- c("6mo", "12mo", "2yr", "3yr", "4yr",
    "5yr")
# Plot the smoke exposure postpartum over time by
# race.
smoke_exp_plot <- ggplot(smoke_exp_over_time, aes(x = factor(time,
    level = level_order), y = count, fill = race, label = count)) +
    geom_bar(stat = "identity") + geom_text(size = 3,
    position = position_stack(vjust = 0.5), color = "white") +
    scale_fill_manual(name = "Race", labels = c("Native Hawaiian/Pacific Islander",
        "Other", "White"), values = ghibli_palette("LaputaMedium")) +
    labs(title = "Count of Reported Smoke Exposure Post-Pregnancy",
        x = "Time Period", y = "Count")
grid.arrange(smoke_exp_plot, nrow = 1)
# Show the number of individuals that fall into
# each of the smoking levels at both time periods
# that urine cotinine levels are provided for.
data %>%
    select(smoking_level_34wk, smoking_level_pp6mo) %>%
    tbl_summary(label = list("smoking_level_34wk" ~
        "Smoking Level (34 Week Gestation)", "smoking_level_pp6mo" ~
        "Smoking Level (6 Months Postpartum)")) %>%
    as_gt() %>%
    gt::tab_header(title = "Table 3: Smoking Level by Urine Cotinine Pre- and Post- Pregnancy")
# Get the percentage of child substance use each
# substance contributes to based on the smoking
# level of the parent at 34 weeks gestation.
substance_use_34wk <- data %>%
    mutate(smoking_level = smoking_level_34wk) %>%
    group_by(smoking_level) %>%
    filter(!is.na(smoking_level)) %>%
```

```
    mutate(substance_use = ifelse(e_cig_ever == 1 |
        cig_ever == 1 | alc_ever == 1 | mj_ever ==
        1, 1, 0)) %>%
    summarize(sub_use_perc = sum(substance_use, na.rm = TRUE)/n() *
        100) %>%
    mutate(time = "34 weeks") %>%
    select(smoking_level, time, sub_use_perc)

# Get the percentage of child substance use each
# substance contributes to based on the smoking
# level of the parent at 6 months postpartum.
# Join it with the results from 34 weeks
# gestation.
substance_use_total <- data %>%
    mutate(smoking_level = smoking_level_pp6mo) %>%
    group_by(smoking_level) %>%
    filter(!is.na(smoking_level)) %>%
    mutate(substance_use = ifelse(e_cig_ever == 1 |
        cig_ever == 1 | alc_ever == 1 | mj_ever ==
        1, 1, 0)) %>%
    summarize(sub_use_perc = sum(substance_use, na.rm = TRUE)/n() *
        100) %>%
    mutate(time = "6 months postpartum") %>%
    select(smoking_level, time, sub_use_perc) %>%
    rbind(substance_use_34wk) %>%
    arrange(desc(smoking_level), time)

colnames(substance_use_total) <- c("Smoking Level",
    "Time Period", "Percent of Child Substance Use")

kable(substance_use_total, booktabs = T, escape = F,
    caption = "Percent of Child Substance Use by Smoking Level at Different Periods") %>%
    kable_styling(latex_options = "HOLD_position")
# Get the rate of total substance use by medium.
# Filter to include only individuals who have
# self-reported substance use and get the rate of
# use among each medium.
substance_use <- data %>%
    filter(e_cig_ever == 1 | cig_ever == 1 | alc_ever ==
        1 | mj_ever == 1) %>%
    select(e_cig_ever, mj_ever, alc_ever, cig_ever) %>%
    summarize(e_cig = sum(e_cig_ever, na.rm = TRUE)/n(),
        mj = sum(mj_ever, na.rm = TRUE)/n(), alc = sum(alc_ever,
            na.rm = TRUE)/n(), cig = sum(cig_ever,
            na.rm = TRUE)/n())

# Get the average number of days (out of the past
# 30 days) that children who have engaged in
# substance use have used a substance, organized
# by substance.  Join this with a vector of
# substance names using cbind for cleanliness.
sub_use_30 <- data %>%
    filter(e_cig_ever == 1 | cig_ever == 1 | alc_ever ==
```

```
        1 | mj_ever == 1) %>%
    select(num_cigs_30, num_e_cigs_30, num_alc_30,
        num_mj_30) %>%
    summarize(avg_e_cigs = mean(num_e_cigs_30, na.rm = TRUE),
        avg_mj = mean(num_mj_30, na.rm = TRUE), avg_alc = mean(num_alc_30,
            na.rm = TRUE), avg_cigs = mean(num_cigs_30,
            na.rm = TRUE)) %>%
    t() %>%
    as.data.frame() %>%
    mutate(days_used = V1) %>%
    cbind(substance = c("E Cigarette", "Marijuana",
        "Alcohol", "Cigarette")) %>%
    select(substance, days_used)

# Combine the two tables together to display
# using kable.
substance_use <- substance_use %>%
    t() %>%
    as.data.frame() %>%
    mutate(percentage = V1 * 100) %>%
    select(percentage) %>%
    cbind(substance = c("E Cigarette", "Marijuana",
        "Alcohol", "Cigarette")) %>%
    select(substance, percentage) %>%
    left_join(sub_use_30, by = "substance")

colnames(substance_use) <- c("Substance", "Percent of Substance Use",
    "Average Days Used in Past 30 Days")

kbl(substance_use, booktabs = T, escape = F, caption = "Substance Use by Medium") %>%
    kable_styling(latex_options = "HOLD_position")
# The SWAN scores for these individuals should be
# NA.  In the data pre-processing, they became 0.
# Changing them back to NA.
parent_ids <- c(50502, 51202, 51602, 52302, 53002,
    53502, 53902, 54402, 54602, 54702)

for (id in parent_ids) {
    data[data$parent_id == id, ]["swan_inattentive"] = NA
    data[data$parent_id == id, ]["swan_hyperactive"] = NA
}

# Smoking during pregnancy If they smoked at any
# point during pregnancy, set to SDP.  If not,
# set to No SDP.
sdp <- smoking_trimester %>%
    filter(!is.na(smoking)) %>%
    mutate(sdp = case_when(smoking == 1 ~ "SDP", TRUE ~
        "No SDP"))

# Smoking through pregnancy If they smoked at all
# points during pregnancy, set to 1.  If not, set
# to 1.  Get the percentage of individuals who
```

```
# smoked throughout pregnancy divided by the
# number of individuals who smoked during
# pregnancy.
percentage_stp <- sdp %>%
    filter(sdp == "SDP") %>%
    group_by(parent_id) %>%
    summarize(stp = case_when(sum(smoking, na.rm = TRUE) ==
        3 ~ 1, TRUE ~ 0)) %>%
    summarize(perc = sum(stp)/n())

# Make two plots: distribution of both types of
# SWAN scores by SDP status
sdp_plot_1 <- ggplot(sdp, aes(x = swan_inattentive,
    fill = sdp)) + geom_density(alpha = 0.25) + ylim(0,
    0.1) + geom_vline(xintercept = 6, linetype = "dashed",
    color = "black", linewidth = 0.5) + annotate(geom = "text",
    x = 9.5, y = 0.08, label = "Likely ADHD \nInattentive") +
    scale_fill_manual(name = "Smoking During Pregnancy?",
        labels = c("No", "Yes"), values = ghibli_palette("LaputaMedium")[c(4,
            6)]) + labs(title = "Distribution of Swan Inattentive Scores by SDP Status",
    x = "Swan Inattentive Score", y = "Density")

sdp_plot_2 <- ggplot(sdp, aes(x = swan_hyperactive,
    fill = sdp)) + geom_density(alpha = 0.25) + ylim(0,
    0.1) + geom_vline(xintercept = 6, linetype = "dashed",
    color = "black", linewidth = 0.5) + annotate(geom = "text",
    x = 9.5, y = 0.07, label = "Likely ADHD \nHyperactive") +
    scale_fill_manual(name = "Smoking During Pregnancy?",
        labels = c("No", "Yes"), values = ghibli_palette("LaputaMedium")[c(4,
            6)]) + labs(title = "Distribution of Swan Hyperactive Scores by SDP Status",
    x = "Swan Hyperactive Score", y = "Density")

grid.arrange(sdp_plot_1, sdp_plot_2, nrow = 2)
# Create a separate column to store the races of
# the children along with the SDP effects. There
# are two categories of biracial children
# (AIAN/Black and Black/ White) so I just defined
# these as biracial.
sdp_race <- sdp %>%
    mutate(child_race = case_when(taian == 1 & tblack ==
        "white" | twhite == 1 & tblack == 1 ~ "biracial",
        taian == 1 ~ "aian", tasian == 1 ~ "asian",
        tnhpi == 1 ~ "nhpi", tblack == 1 ~ "black",
        twhite == 1 ~ "white", trace_other == 1 ~ "other"))

# For each category of SDP and Race, calculate
# the average ERQ scores for expressive
# suppression and cognitive reappraisal.
erq <- sdp_race %>%
    filter(!is.na(child_race)) %>%
    group_by(sdp, child_race) %>%
    summarize(cog = mean(erq_cog, na.rm = TRUE), exp = mean(erq_exp,
        na.rm = TRUE), .groups = "drop_last") %>%
```

```
    gather(key = "erq", value = "value", c(cog, exp))

# For each category of SDP and Race, calculate
# the average BPM scores for attention,
# externalizing, and internalizing.
bpm <- sdp_race %>%
    filter(!is.na(child_race)) %>%
    group_by(sdp, child_race) %>%
    summarize(att = mean(bpm_att, na.rm = TRUE), ext = mean(bpm_ext,
        na.rm = TRUE), int = mean(bpm_int, na.rm = TRUE),
        .groups = "drop_last") %>%
    gather(key = "bpm", value = "value", c(att, ext,
        int))

# Plot average BPM values by SDP status
ggplot(bpm, aes(x = sdp, y = value, fill = bpm)) +
    geom_bar(stat = "identity", position = position_dodge()) +
    scale_fill_manual(name = "Brief Problem Monitor Section",
        labels = c("Attention Problems on Self", "Externalizing Problems on Self",
            "Internalizing Problems on Self"), values = ghibli_palette("LaputaMedium")[c(1,
            3, 5)]) + labs(title = "Average Brief Problem Monitor Scores by SDP Exposure",
    x = "SDP Exposure", y = "Average Score")
# Plot average ERQ values by SDP status
erq_plot <- ggplot(erq, aes(x = sdp, y = value, fill = erq)) +
    geom_bar(stat = "identity", position = position_dodge()) +
    scale_fill_manual(name = "Emotion Regulation Questionnaire Section",
        labels = c("Cognitive Reappraisal", "Expressive Suppression"),
        values = ghibli_palette("LaputaMedium")[c(2,
            4)]) + labs(title = "Average Emotion Regulation Questionnaire Scores by SDP Exposure",
    x = "SDP Exposure", y = "Average Score")

suppressWarnings(print(erq_plot))
```

Bibliography

ASEBA. 2022. *ASEBA*. https://aseba.org/school-age-bpm/#:~:text=The%https://aseba.org/school-age-bpm/#:~:text=The%20Brief%20Problem%20Monitor%20(BPM,youths%20(BPM%2DY).%20Problem%20Monitor%20(BPM,youths%20(BPM%2DY).

Drake, Patrick, Anne K Driscoll, and TJ Mathews. 2018. "Cigarette Smoking During Pregnancy: United States, 2016."

Gross, James J, and Oliver P John. 2003. "Individual Differences in Two Emotion Regulation Processes: Implications for Affect, Relationships, and Well-Being." *Journal of Personality and Social Psychology* 85 (2): 348.

Hellemons, Merel E, Jan-Stephan F Sanders, Marc AJ Seelen, Rijk OB Gans, Anneke C Muller Kobold, Willem J van Son, Douwe Postmus, Gerjan J Navis, and Stephan JL Bakker. 2015. "Assessment of Cotinine Reveals a Dose-Dependent Effect of Smoking Exposure on Long-Term Outcomes After Renal Transplantation." *Transplantation* 99 (9): 1926–32.

NIH. 2021. *National Institutes of Health*. U.S. Department of Health; Human Services. https://nida.nih.gov/publications/research-reports/tobacco-nicotine-e-cigarettes/what-are-risks-smoking-during-pregnancy.

Risica, Patricia Markham, Adam Gavarkovs, Donna R Parker, Ernestine Jennings, and Maureen Phipps. 2017. "A Tailored Video Intervention to Reduce Smoking and Environmental Tobacco Exposure During and After Pregnancy: Rationale, Design and Methods of Baby's Breath." *Contemporary Clinical Trials* 52: 1–9.