Investigating the effects of SDP and ETS in Children Develolpment: An exploratory data analysis

Monica Colon Vargas

Exposure to smoking during pregnancy (SDP) and exposure to environmental tobacco smoke (ETS) are two environmental factors that impact children. There's a belief that early exposure to smoke might be connected to higher instances of outward behaviors in children, such as Attention-Deficit/Hyperactivity Disorder, and an increased likelihood of substance use. Moreover, early exposure to smoke has been associated with difficulties in self-control, affecting physiological, emotional, behavioral, and cognitive aspects. This project aims to explore the relationship between smoking during pregnancy (SDP) and exposure to environmental tobacco smoke (ETS) and their influence on self-regulation, outward behavior, and substance use. Our study found SDP seems to negatively affect children's self-regulation and behavior, while heavy maternal smoking links to increased drug experimentation, with some outliers. ETS appears to impact emotional control and behavior, but results vary. Limited impact on substance use by ETS is suggested. Caution is urged due to no statistical testing, small sample size, and missing data; larger studies are needed for validation.

Introduction

Exposure to smoking during pregnancy (SDP) and exposure to environmental tobacco smoke (ETS) represent two environmental factors that affects children [1]. It has been speculated that early exposure to smoke is associated with elevated rates of externalizing behaviors among children, including conditions like Attention-Deficit/Hyperactivity Disorder, as well as an increased prevalence of substance. Furthermore, early smoke exposure has been linked to difficulties in self-regulation, encompassing challenges in maintaining control over physiological, emotional, behavioral, and cognitive aspects [2]. The aim of this project is to investigate the connection between smoking during pregnancy (SDP) and exposure to environmental tobacco smoke (ETS) and their impact on self-regulation, externalizing behavior, and substance use.

After pre-processing the data-set, we have a total of n=49 observations with 78 variables for the mother and the child. The 78 variables include demographics like age, sex, race, language, employment, education level, income. Additionally, we have variables measuring smoke exposure in children across different time points and variables that measures if the mother smoked across different stages during her pregnancy. Tests to study the children difficulties in self-regulation, encompassing challenges in maintaining control over physiological, emotional, behavioral, and cognitive aspects were recollected. Such tests include Brief Problem Monitor, Emotion Regulation Questionnaire,and SWAN scores for ADHD.

Missing Data

To begin working with the dataset, it's important to address some irregularities in certain variable values. For instance, in the `income` variable, one entry was changed from `250 000` to `250000`. Similarly, the `mom_cig` variable had a value of `40` which didn't align with the intended range for this variable, so it was replaced with `NA` In the `mom_numcig` variable, there were various inconsistent values, such as "2 black and miles a day" which was set to `2` and `44989` which was considered unrealistic and thus set to `NA`. Additionally, `20-25` was adjusted to its mean value of `22.5`, and `none` was converted to `0`. Following these adjustments, we then turned our attention to dealing with 'NA' values in the variables `num_e_cigs_30`, `num_mj_30,` and `num_alc_30`. We set these values to zero if the corresponding `X_ever` variable (where `X` varies for each) indicated zero usage. Having addressed these issues, our next step was to examine the patterns of missing data within the data-set.

This data-set lacks complete cases, indicating that omitting all the `NA` values would result in the entire data-set being eliminated. As depicted in Table 1, we can observe the quantity of missing data for each variable. It is noteworthy that certain variables exhibit identical counts of missing values. In Figure 1, a graphical representation of missing values per individual patient is presented, with patients depicted along the y-axis. It becomes apparent that nearly 50 patients (equivalent to the total count of 49 patients) exhibit missing data. Also, numerous variables share missing values for the same patient, and conversely, several patients share the same missing variables concurrently. For instance, variables spanning from `page` to `pethnic`, `pemploy` to `mom_numcig, mom_smoke_32wk` and `mom_smoke_pp1`, `bpm_att_p` to `tethnic` and `cig_ever` to `pqm_paprental_control` all display missing values for the same 8 patients. This pattern suggests that the data may conform to the Missing at Random (MAR) assumption rather than Missing Completely at Random (MCAR). Given that the number of variables exceeds the number of observations, multiple imputation may not be the most efficient approach for handling these missing data. However, for the purposes of this project and when investigating relationships involving specific variables, we will focus exclusively on the complete cases of the variables used at the moment.

Table 1: Missing Data Pattern

Variable	n	%	Variable	n	%
mom_smoke_pp1	39	79.59	momcig	11	22.45
childasd	28	57.14	mom_numcig	11	22.45
mom_smoke_pp2	20	40.82	cotimean_34wk	11	22.45
pmq_parental_control	16	32.65	cotimean_pp6mo_baby	11	22.45
ppmq_parental_solicitation	15	30.61	cotimean_pp6mo	11	22.45
num_alc_30	14	28.57	smoke_exposure_3yr	11	22.45
bpm_int	14	28.57	smoke_exposure_4yr	11	22.45
pmq_parental_knowledge	14	28.57	bpm_att_a	11	22.45
pmq_parental_solicitation	14	28.57	bpm_ext_a	11	22.45
bpm_att_p	13	26.53	nidaalc	10	20.41
tsex	13	26.53	nidatob	10	20.41
num_e_cigs_30	13	26.53	nidaill	10	20.41
alc_ever	13	26.53	bpm_int_p	10	20.41
erq_cog	13	26.53	smoke_exposure_6mo	10	20.41
erq_exp	13	26.53	smoke_exposure_12mo	10	20.41
pmq_child_disclosure	13	26.53	smoke_exposure_2yr	10	20.41
income	12	24.49	smoke_exposure_5yr	10	20.41
bpm_ext_p	12	24.49	bpm_int_a	10	20.41
ppmq_parental_knowledge	12	24.49	erq_cog_a	10	20.41
ppmq_child_disclosure	12	24.49	erq_exp_a	10	20.41
ppmq_parental_control	12	24.49	mom_smoke_32wk	9	18.37
tage	12	24.49	mom_smoke_pp6mo	9	18.37
language	12	24.49	page	8	16.33
tethnic	12	24.49	psex	8	16.33
cig_ever	12	24.49	plang	8	16.33
num_cigs_30	12	24.49	pethnic	8	16.33
e_cig_ever	12	24.49	employ	8	16.33
mj_ever	12	24.49	pedu	8	16.33
num_mj_30	12	24.49	mom_smoke_22wk	7	14.29
bpm_att	12	24.49	mom_smoke_pp12wk	7	14.29
bpm_ext	12	24.49	mom_smoke_16wk	1	2.04
nidapres	11	22.45			

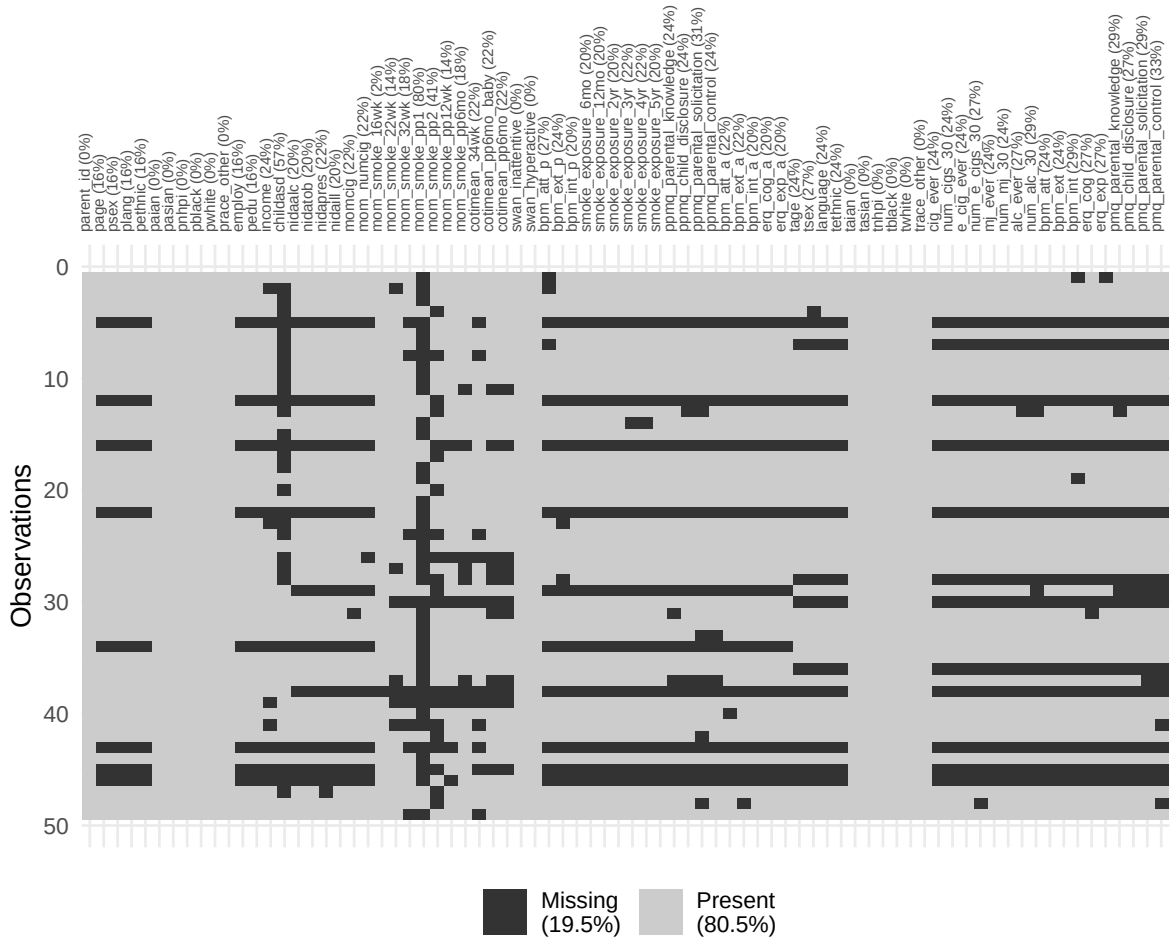

Figure 1: Missing Data Pattern

Demographics

We now provide graphical representations of the demographic characteristics of both the parents and their children. Figure 2 provides an overview of the parent's demographics. It is noteworthy to observe a singular male entry within the data-set, which requires further investigation, given the predominant focus of the study on pregnant individuals. Regarding racial demographics, the data-set predominantly comprises mothers of White ethnicity, followed by individuals of Native Hawaiian, other, and Alaskan Native backgrounds. Moreover, the majority of mothers do not identify as Hispanic or Latino in terms of ethnicity. Age distribution reveals that the majority falls within the range of 33 to 39 years, indicative of a central tendency within this age bracket. In terms of employment, a significant proportion of moth-

ers were employed full-time, and the dataset's estimated mean household income stands at $63,138.05, signifying an average income level. Furthermore, the majority of mothers possess an educational background that includes at least some college-level education.

Parent Demographics

Figure 3 offers a comprehensive presentation of the demographic characteristics of the children participating in the study. The data reveals that the majority of children are male, with ages predominantly falling within the 12 to 15-year-old range. It is noteworthy that a substantial proportion of these children identify themselves as belonging to the White racial group, followed by those who consider themselves as Black. Additionally, the majority of children do not identify as Hispanic or Latino in terms of ethnicity. Of particular interest is the observation made in Figure 2A, which indicates a discrepancy in racial demographics between the mothers and their children. Notably, while no mothers in the data-set self-identify as belonging to the Black racial group, there is a significant presence of children identifying as such.

Child Demographics

Smoking During Pregnancy (SDP)

To assess the impact of (SDP), the initial step involves categorizing mothers into three distinct groups: Non-Smokers, Moderate-Smokers, and Heavy-Smokers. This classification is predicated upon the presence or absence of smoking behavior during specific gestational periods, as indicated by three variables: `mom_smoke_16wk`, `mom_smoke_22wk`, and `mom_smoke_32wk`, which correspond to the 16th, 22nd, and 32nd weeks of pregnancy, respectively. Mothers affirming a positive response to all three variables are designated as Heavy-Smokers, while those indicating no for all three are classified as Non-Smokers. Mothers acknowledging a yes response to at least one variable fall into the Moderate-Smoker category.

Table 2 presents the average scores for selected child assessments. The `erq` variables encompass the Emotional Regulation Questionnaire, where higher scores suggest a greater propensity for the child to encounter challenges in regulating emotions, both in cognitive reappraisal `erq_cog` and expressive suppression `erq_exp`. Notably, children born to Non-Smoker mothers exhibit lower mean scores in both variables. Conversely, children of Heavy-Smoker mothers tend to have lower mean scores compared to Moderate Smokers, although this trend may be influenced by the relatively limited number of observations in the Heavy-Smoker group (n=3). Due to the notably small sample of Moderate-Smoker mothers, our subsequent analysis will primarily focus on a comparison between Heavy-Smokers and Non-Smokers.

Furthermore, the `bpm` variables pertain to the Brief Problem Monitor, where parents assess the veracity of statements regarding their child's attributes, particularly attention `bpm_att_p`, externalizing behavior `bpm_ext_p`, and internalizing issues `bpm_int_p`. Elevated values on these variables imply more significant problems in the respective domains. Note, children of Non-Smoker mothers exhibit lower mean scores when compared to those of Heavy Smokers.

Table 2: Parent Demographics

Characteristic	**N = 49**
___prace___	NA
AIAN	4 (9.8%)
NHPI	6 (15%)
Other	6 (15%)
White	25 (61%)
Unknown	8
___income___	46,848 (20,000, 70,000)
Unknown	12
___pethnic___	13 (32%)
Unknown	8
___psex___	NA
Male	1 (2.4%)
Female	40 (98%)
Unknown	8
___employ___	NA
No	12 (29%)
Part-Time	7 (17%)
Full- Time	22 (54%)
Unknown	8
___pedu___	NA
Some HS	3 (7.3%)
HS	3 (7.3%)
GED	5 (12%)
Some College	15 (37%)
Two-yr dg	3 (7.3%)
Four-yr dg	10 (24%)
Postgraduate dg	2 (4.9%)
Unknown	8

Table 3: Children Demographics

Characteristic	**N = 49**
___trace___	NA
AIAN	5 (14%)
Black	12 (33%)
Other	5 (14%)
White	14 (39%)
Unknown	13
___tsex___	NA
Male	23 (64%)
Female	13 (36%)
Unknown	13
___tethnic___	15 (42%)
Unknown	13
___tage___	NA
12	8 (22%)
13	10 (27%)
14	9 (24%)
15	8 (22%)
16	2 (5.4%)
Unknown	12

Additionally, we evaluated measures from the SWAN test, specifically addressing ADHD-Hyperactive/Impulsive `swan_hyperactive` and ADHD-Inattentive `swan_inattentive` types. In both instances, children born to Heavy-Smoker mothers tend to score higher, suggestive of a heightened propensity for ADHD-like symptoms. Finally, we report the mean levels of urinary nicotine for children at six months of age `cotimean_pp6mo_baby`, which reveals a conspicuous disparity between the offspring of Heavy-Smoker and Non-Smoker mothers. While these findings suggest a potential deleterious effect of SDP on child self-regulation and externalizing behavior, it is essential to underscore that this analysis serves as an exploratory step, and these distinctions necessitate statistical evaluation.

Table 4: Mean and SD for tests regarding Child

Characteristic	**Heavy-Smoker**, N = 10	**Moderate-Smoker**, N = 3	**Non-Smoker**, N = 24
erq__cog	3.36 (0.73)	3.67 (0.60)	2.97 (1.13)
erq__exp	2.94 (0.62)	3.67 (0.72)	2.54 (0.84)
bpm__att__p	3.75 (2.82)	1.00 (1.00)	1.71 (2.08)
bpm__ext__p	2.63 (2.62)	0.33 (0.58)	2.00 (2.92)
bpm__int__p	3.50 (2.88)	1.33 (1.15)	2.37 (2.75)
swan__inattentive	10.80 (7.63)	10.33 (1.53)	9.13 (6.40)
swan__hyperactive	9.90 (8.33)	10.33 (3.79)	4.96 (5.77)
cotimean__pp6mo__baby	9.32 (13.37)	2.15 (1.67)	3.28 (5.49)

We now study the effect of SDP in substance use on children. Table 3 shows the number of children that have used a cigarette, E-cigarette, Marijuana and Alcohol at least once in their life on based on their mother's category. For example, only 1 children had experience using a cigarette at least once, and this individual comes from a heavy-smoker mother. No children with a non-smoker mother have ever experience using a cigarette or an e-cigarette before. Only one child with a non-smoker mother have used marijuana and two from heavy-smoker mothers. Moreover, we can see that everyone has experienced alcohol before which is not surprisingly.

Table 5: Number of Children who have used substance at least once

Characteristic	**Heavy-Smoker**, N = 10	**Moderate-Smoker**, N = 3	**Non-Smoker**, N = 24
cig_ever	1.00 (12.50%)	0.00 (0.00%)	0.00 (0.00%)
e_cig_ever	1.00 (12.50%)	1.00 (33.33%)	0.00 (0.00%)
mj_ever	2.00 (25.00%)	0.00 (0.00%)	1.00 (5.56%)
alc_ever	2.00 (28.57%)	1.00 (33.33%)	2.00 (11.11%)

Additionally, the amount of substance use in the past 30 days in these children. Since only few of them had experience use of this substance before, we expect these numbers to be even lower due to the fact that maybe that first experience was not during the last 30 days. The purpose of Figure 4 is to show that the majority of these children may not be considered substance users (or at least during the past 30 days) due to the fact that only very few had used a substance. For instance, only one individual with a heavy-smoker mother used e-cigarettes on two days and alcohol on ten days (note that this is not necessarily the same individual) in the past 30 days. On the marijuana plot, we can see more individuals with heavy-smoker mothers that have used the substance in more days than the rest. However note that there is one individual with a non-smoker mother that have used marijuana around 18 days out of the past 30 days. However, note that all of these high values are outliers and the majority of the individuals have not used substances in the past 30 days. This variable would be more efficient if converted to a binary outcome.

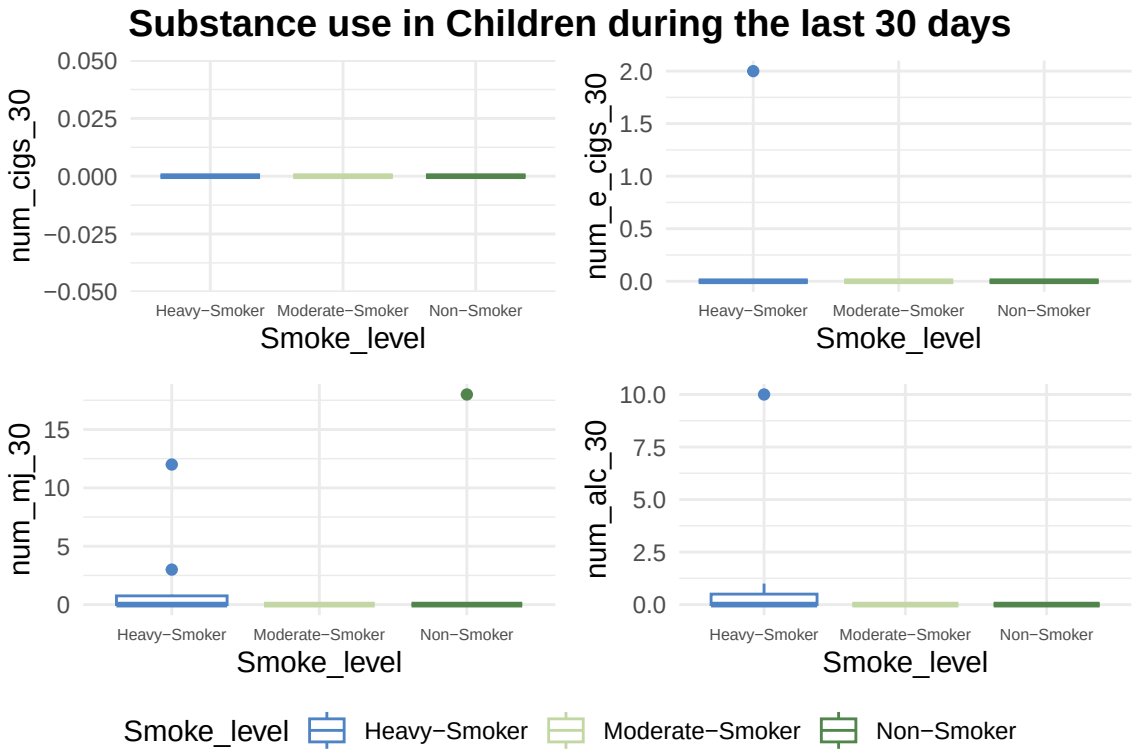

Figure 2: Child Substance use in the past 30 days

Environmental Tobacco Smoke (ETS)

We now focus to the effects of ETS on child self-regulation, externalizing behavior concerns, and substance use. To facilitate this exploration, we introduced a new variable known as `smoke_exposure_level`. This variable was created based on the presence or absence of exposure to smoke, as indicated by several exposure variables (specifically, `smoke_exposure_6mo`, `smoke_exposure_12mo`, `smoke_exposure_2yr`, `smoke_exposure_3yr`, `smoke_exposure_4yr`, and `smoke_exposure_5yr`). In this context, a value of `0` signified that the child had not been exposed to smoke from either the mother or the father during distinct time intervals, namely at 6 months, 12 months, 1 year, 2 years, 3 years, 4 years, and 5 years. Conversely, a value of `1` denoted exposure to smoke. Each individual within the data-set was categorized into one of three distinct exposure levels: heavily-exposed, moderately-exposed, or not-exposed. This classification process involved calculating the mean across the aforementioned exposure variables for each individual. If this mean equated to zero, the individual was classified as not-exposed. For individuals with a mean greater than zero but less than or equal to 0.5 they were designated as moderately-exposed. In contrast, if the mean exceeded 0.5 the individual was categorized as extremely-exposed.

To investigate the effect of ETS, we selectively focused on children whose mothers had been classified as non-smokers during pregnancy. This choice was deliberate, intended to isolate the influence of ETS independently from the impact of SDP.

Table 4 shows an overview of the mean and standard deviation of the tests previously showend in the SDP section. The data reveals that children categorized as not exposed to smoke tend to exhibit a lower mean score in the Emotional Regulation Questionnaire, encompassing cognitive reappraisal `erq_cog` and expressive suppression `erq_exp`. This implies that exposure to ETS may potentially have an negative effect on these aspects of children emotional regulation. A similar trend emerges when scrutinizing the BPM assessments of externalizing and internalizing issues, as well as the SWAN assessments of inattentive and hyperactive scores. In these cases, children exposed to ETS tend to register higher scores, suggesting an increased propensity for concerns in these domains. However, it's worth noting that the mean for the BPM attention problems is marginally higher for children classified as not exposed as is the mean for urine nicotine levels in infants at six months. It is imperative to approach these findings with caution due to the limited sample size of children categorized as extremely or moderately exposed to ETS.

Table 6: Mean and SD for tests regarding Child

Characteristic	**Extremely exposed**, N = 3	**Moderately Exposed**, N = 2	**Not exposed**, N = 14
erq_cog	3.83 (1.04)	3.17 (0.24)	2.97 (1.11)
erq_exp	3.13 (1.24)	2.75 (0.35)	2.52 (0.85)
bpm_att_p	1.50 (0.71)	1.00 (1.41)	1.85 (2.34)
bpm_ext_p	5.00 (5.29)	1.00 (0.00)	1.42 (2.07)
bpm_int_p	4.00 (3.00)	4.00 (5.66)	1.79 (2.29)
swan_inattentive	13.67 (3.79)	12.50 (7.78)	10.93 (4.84)
swan_hyperactive	8.00 (8.19)	8.50 (7.78)	5.57 (5.49)
cotimean_pp6mo_baby	1.48 (0.09)	1.07 (1.17)	3.93 (6.70)

Now, the effect of ETS on substance use is studied. Table 5, presents the number of children who have experimented with substances such as cigarettes, e-cigarettes, marijuana, and alcohol, each at least once. Note that almost none of the children have experienced substance use with the exception of two children (one not exposed have experience with marijuana before and one moderately exposed have experience with alcohol before). Given the limited number of children who have engaged in substance use, we have opted not to present data on the frequency of substance use within the preceding 30 days. Nevertheless, it is noteworthy that the same child categorized as not exposed who reported prior marijuana use, is also the sole child who has indicated recent substance use, spanning a substantial duration of 18 out of the past 30 days. It is essential to emphasize that this specific data point was identified as an outlier using Dixon's test.

Table 7: Number of Children who have used substance at least once

Characteristic	**Extremely exposed**, N = 3	**Moderately Exposed**, N = 2	**Not exposed**, N = 14
cig_ever	0.00 (0.00%)	0.00 (0.00%)	0.00 (0.00%)
e_cig_ever	0.00 (0.00%)	0.00 (0.00%)	0.00 (0.00%)
mj_ever	0.00 (0.00%)	0.00 (0.00%)	1.00 (9.09%)
alc_ever	0.00 (0.00%)	1.00 (50.00%)	0.00 (0.00%)

Conclusion

An exploratory analysis was used in this experiment to investigate the effects of SDP and ETS on substance use, externalizing behavioral problems in children, and self-regulation. The results of the child's ERQ, BPM, and SWAN tests indicate that SDP has a detrimental impact on the child's self-regulation and externalizing behaviors. In terms of drug usage, we saw that more kids whose mothers were heavy smokers had in fact experimented more, but the box plots demonstrate that the number of days during the previous 30 days on which the child had used the substance were regarded as outliers. On the other side, the results of the child's ERQ, BPM, and SWAN tests also imply that ETS may have a detrimental impact on the child's ability to control their own emotions and their tendency to externalize their conduct, however not all tests coincide. In the case of attention, for instance, BPM revealed that children who were not exposed had a slightly higher mean than the others, and this merits more investigation. Furthermore, the information suggests that ETS has little effect on children substance usage.

It's important to emphasize, though, that none of these conclusions were reached by statistical testing; instead, they were all reached by examining the data alone. As a result, no statistical inferences can be drawn; instead, the data must be analyzed. This study has some limitations that are largely due to the study's limited sample size and large amount of missing data. These findings should be interpreted with caution, and we encourage future studies with larger and more complete data-set to corroborate our results.

This project is under the repository: https://github.com/MCV20/PHP2050-Project1.git

References

[1] DiFranza JR, Aligne CA, Weitzman M. Prenatal and postnatal environmental tobacco smoke exposure and children's health. Pediatrics. 2004 Apr;113(4 Suppl):1007-15. PMID: 15060193.

[2] Yolton K, Khoury J, Hornung R, Dietrich K, Succop P, Lanphear B. Environmental tobacco smoke exposure and child behaviors. J Dev Behav Pediatr. 2008 Dec;29(6):450-7. doi: 10.1097/dbp.0b013e31818d0c21. PMID: 19093323; PMCID: PMC8875281.

Code Appendix

```
#Load Libraries
library(ggplot2)
library(dplyr)
library(gt)
library(naniar)
library(ggpubr)
library(kableExtra)
library(gtsummary)
library(outliers)
library(forcats)

library(readr)
library(gtsummary)
library(tableone)
library(dplyr)
library(ggplot2)
library(gtable)
library(kableExtra)
library(mice)
library(corrplot)
library(ggpubr)
library(pROC)
library(glmnet)

#Load data
#df <- read.csv("Data/project1.csv",na.strings=c("","NA"))
df <- read.csv("../Data/project1.csv",na.strings=c("","NA"))

#Set colors
cbp2 <- c( "#4E84C4", "#C3D7A4" ,"#52854C", "#F4EDCA","#293352")
#Change NA from num_X_30 from NA to 0 if they had never used X before
df <- df %>% mutate(num_cigs_30 = case_when(cig_ever == 0 ~ 0,
                        cig_ever == 1 ~ num_cigs_30,
                        TRUE ~ NA),
                    num_e_cigs_30 = case_when(e_cig_ever == 0 ~ 0,
                                              e_cig_ever == 1 ~ num_e_cigs_30,
                                              TRUE ~ NA),
                    num_mj_30 = case_when(mj_ever == 0 ~ 0,
                                          mj_ever == 1 ~ num_mj_30,
```

```
                                          TRUE ~ NA),
                    num_alc_30 = case_when(alc_ever == 0 ~ 0,
                                           alc_ever == 1 ~ num_alc_30,
                                           TRUE ~ NA))
#Change 1=Yes to 1, and 2=No to 0
df <- df %>% mutate(mom_smoke_16wk = case_when(mom_smoke_16wk == "1=Yes" ~ 1,
                                               mom_smoke_16wk == "2=No" ~ 0,
                                               TRUE ~ NA),
                    mom_smoke_22wk = case_when(mom_smoke_22wk == "1=Yes" ~ 1,
                                               mom_smoke_22wk == "2=No" ~ 0,
                                               TRUE ~ NA),
                    mom_smoke_32wk = case_when(mom_smoke_32wk == "1=Yes" ~ 1,
                                               mom_smoke_32wk == "2=No" ~ 0,
                                               TRUE ~ NA),
                    mom_smoke_pp1 = case_when(mom_smoke_pp1 == "1=Yes" ~ 1,
                                              mom_smoke_pp1 == "2=No" ~ 0,
                                               TRUE ~ NA),
                    mom_smoke_pp2 = case_when(mom_smoke_pp2 == "1=Yes" ~ 1,
                                              mom_smoke_pp2 == "2=No" ~ 0,
                                              TRUE ~ NA),
                    mom_smoke_pp12wk = case_when(mom_smoke_pp12wk == "1=Yes" ~ 1,
                                              mom_smoke_pp12wk == "2=No" ~ 0,
                                              TRUE ~ NA),
                    mom_smoke_pp6mo = case_when(mom_smoke_pp6mo == "1=Yes" ~ 1,
                                              mom_smoke_pp6mo == "2=No" ~ 0,
                                              TRUE ~ NA))
#issues with income
df$income[6] <- 250000 #value had one space
#range(as.numeric(df$income), na.rm = T)
#sort(as.numeric(df$income), decreasing = F) maybe outlier 760?? Or just incorrect

#issues with momcig
#range(df$momcig, na.rm = T) #40 does not make sense. Typo? Was it 4? Was it 30 the max?
df[which(df$momcig == 40),]$momcig <- NA

#issues which mom_numcig
df[which(df$mom_numcig == "2 black and miles a day"),]$mom_numcig <- 2
df[which(df$mom_numcig == "44989"),]$mom_numcig <- NA
df[which(df$mom_numcig == "20-25"),]$mom_numcig <- mean(20:25)
df[which(df$mom_numcig == "None"),]$mom_numcig <- 0
```

```r
#Change to factor levels and numerical variables
df <- df %>%  mutate_at(c('psex', 'plang', 'pethnic','paian', 'pasian', 'pnhpi', 'pblack',
                          'pwhite','prace_other', 'employ', 'pedu', 'childasd',
                          'nidaalc', 'nidatob', 'nidaill', "nidapres",
                          'mom_smoke_16wk','mom_smoke_22wk','mom_smoke_32wk',
                          'mom_smoke_pp1', 'mom_smoke_pp2', 'mom_smoke_pp12wk',
                          'mom_smoke_pp6mo', 'smoke_exposure_6mo',
                          'smoke_exposure_12mo', 'smoke_exposure_2yr',
                          'smoke_exposure_3yr', 'smoke_exposure_4yr',
                          'smoke_exposure_5yr', 'tsex', 'language', 'tethnic',
                          'taian', 'tasian', 'tnhpi', 'tblack', 'twhite',
                          'trace_other', 'parent_id'),as.factor)
df <- df %>% mutate_if(is.character, as.numeric)
df <- df %>% mutate_if(is.integer, as.numeric)

#Creating Missing Values Table
missing_table <- df %>%
  summarize(across(everything(), ~ sum(is.na(.x)))) %>%
  t() %>%
  as.data.frame() %>%
  mutate(n=V1) %>%
  select(n)

missing_cols <- missing_table %>%
  filter(n > 0) %>%
  arrange(desc(n)) %>% mutate("%" = round(n/dim(df)[1],4)*100)

#adding extra blank space to separate into two columns of equal space
missing_cols[ nrow(missing_cols) + 1 , ] <- ""
missing_cols$Variable <- rownames(missing_cols)
missing_cols$Variable[64] <- "" #putting empty space

dd <- missing_cols %>% select(Variable, n, '%') #taking what we are going to separate
dd2 <- cbind(dd[1:32, ],dd[33:64,])#separating

#table
kable(dd2,
      caption = "Missing Data Pattern",booktabs=T,  row.names = FALSE,
      align = "lrr") %>%
  kable_styling(full_width=T, font_size = 10,latex_options = c('scale_down'))
```

```r
#plot for missing data pattern using naniar
vis_miss(df)+theme(axis.text.x=element_text(size=rel(.72), angle = 90))
```

```r
df$psex <- factor(df$psex, levels = c(0, 1), labels = c("Male", "Female"))
df$plang <- factor(df$plang, levels = c(0, 1), labels = c("No", "Yes"))
df$pethnic <- factor(df$pethnic, levels = c(0, 1), labels = c("No", "Yes"))
df$employ <- factor(df$employ, levels = c(0, 1, 2), labels = c("No", "Part-Time", "Full- T
df$pedu <- factor(df$pedu, levels = c(0,1,2,3,4,5,6), labels = c("Some HS", "HS", "GED", "

df <- df %>%
  mutate(prace = case_when(
    paian == 1 ~ "AIAN",
    pasian == 1 ~ "Asian",
    pblack == 1 ~ "Black",
    pwhite == 1 ~ "White",
    pnhpi == 1 ~ "NHPI",
    prace_other == 1 ~ "Other"
  ))
##CHildren

df$tsex <- factor(df$tsex, levels = c(0, 1), labels = c("Male", "Female"))
df$tethnic <- factor(df$tethnic, levels = c(0, 1), labels = c("No", "Yes"))
df <- df %>%
  mutate(trace = case_when(
    taian == 1 ~ "AIAN",
    tasian == 1 ~ "Asian",
    tblack == 1 ~ "Black",
    twhite == 1 ~ "White",
    tnhpi == 1 ~ "NHPI",
    trace_other == 1 ~ "Other"
  ))
```

```r
table1 <- df %>%
  dplyr::select(prace, income, pethnic, psex, employ, pedu) %>%
```

```
  tbl_summary() %>%
  bold_labels()

# Summary table for the second set of columns
table2 <- df %>%
  dplyr::select(trace, tsex, tethnic, tage) %>%
  tbl_summary() %>%
  bold_labels()



#PARENT
#Race Variable
#creating counts to plot
counts <- df %>% select(paian, pasian, pnhpi, pblack, pwhite, prace_other) %>%
  mutate_if(is.factor, as.numeric) %>%
  mutate(across(everything(), ~  . - 1)) %>%
  colSums() %>%
  as.data.frame()

colnames(counts) <- "count"
counts <- counts %>% mutate(Race = rownames(counts))

#plot
prace <- ggplot(counts, aes(y = Race, x = count))+geom_point(color = '#293352')+
  geom_segment(aes(x = rep(0,6), y = 1:6, xend = c(4,0,0,8,6,26), yend = 1:6),
               color = '#293352')+
  theme_minimal()+
  scale_y_discrete(labels = c("paian" = "American Indian\nAlaskan Native",
                              "pasian" = "Asian",
                              "pnhpi" = "Native Hawaian\nPacific Islander",
                              "pblack" = "Black",
                              "pwhite" = "White",
                              "prace_other" = "Other"
                              ))+ggtitle("Race")+theme(axis.text.y=element_text(size=rel(.

#education plot
df1 <- df %>% select(pedu) %>% na.omit()
pedu <- ggplot(df1,aes(x = as.factor(pedu)))+geom_bar(fill='#293352')+theme_minimal()+
  scale_x_discrete(labels=c("0" = "Some\nhighschool", "1" = "High school",
                            "2" = "GED","3" = "Some\ncollege",
```

```
                              "4" = "2 year\ndegree","5" = "4 year\ndegree",
                              "6" = "Postgraduate\ndegree"))+
  xlab("")+
  ggtitle("Parent Education Level")+theme(axis.text.x=element_text(size=rel(.7)))


#age plot
df1 <- df %>% select(page) %>% na.omit()
page <- ggplot(df1,aes(x = as.factor(page)))+geom_bar(fill = '#293352')+theme_minimal()+
  xlab("")+ggtitle("Parent Age")

#employment plot
df1 <- df %>% select(employ) %>% na.omit()
pemploy <- ggplot(df1, aes(x = employ))+geom_bar(fill = '#293352')+theme_minimal()+
  scale_x_discrete(labels = c("0" = "No", "1" = "Part-Time",
                              "2" = "Full-Time"))+
  xlab("")+ggtitle("Parent Employment")+theme(axis.text.x=element_text(size=rel(.7)))

#income plot
df1 <- df %>% select(income) %>% na.omit()
pincome <- ggplot(df1,aes(x = income))+geom_density(color = '#293352')+theme_minimal()+
  xlab("")+ggtitle("Family Estimated Income")+ scale_x_continuous(labels = scales::comma)

#sex plot
df1 <- df %>% select(psex) %>% na.omit()
psex <- ggplot(df1,aes(x = as.factor(psex)))+geom_bar(fill = '#293352')+theme_minimal()+
  xlab("")+ggtitle("Parent Sex")+
  scale_x_discrete(labels = c("0" = "Male",
                              "1" = "Female"))+
  theme(axis.text.x=element_text(size=rel(.7)))

#ethnicity plot
df1 <- df %>% select(pethnic) %>% na.omit()
pethnic <- ggplot(df1,aes(x = as.factor(pethnic)))+geom_bar(fill = '#293352')+theme_minima
  xlab("")+ggtitle("Hispanic or Latino")



#join the plots togeter
p1 <- ggarrange(prace, pethnic, nrow = 1,ncol = 2,labels = c("A", "B"))
p2 <- ggarrange(psex,page, labels = c("C", "D"), nrow = 1,ncol = 2)
```

```r
p3 <- ggarrange(pemploy,pincome, labels = c("E","F"), ncol = 2)
p4 <- ggarrange(pedu, labels = "F")
#ggarrange(p1,p2,p3,p4, nrow = 4)

 df %>%
  dplyr::select(prace, income, pethnic, psex, employ, pedu) %>%
  tbl_summary() %>%
  bold_labels() %>%
   kable(caption = "Parent Demographics",booktabs=T,row.names = F) %>%
   kable_styling(full_width=T, font_size = 10,latex_options = c('scale_down'))



#CHILDREN
#Race Variable
#creating count for race to make plot
counts <- df %>% select(taian, tasian, tnhpi, tblack, twhite, trace_other) %>%
  mutate_if(is.factor, as.numeric) %>%
  mutate(across(everything(), ~  . - 1)) %>%
  colSums() %>%
  as.data.frame()

colnames(counts) <- "count"
counts <- counts %>% mutate(Race = rownames(counts))

#race plot
trace <-ggplot(counts, aes(y = Race, x = count))+geom_point(color = '#293352')+
  geom_segment(aes(x = rep(0,6), y = 1:6, xend = c(5,0,15,0,5,19), yend = 1:6),
               color = '#293352')+
  theme_minimal()+
  scale_y_discrete(labels = c("taian" = "American Indian\nAlaskan Native",
                              "tasian" = "Asian",
                              "tnhpi" = "Native Hawaian\nPacific Islander",
                              "tblack" = "Black",
                              "twhite" = "White",
                              "trace_other" = "Other"
  ))+ggtitle("Race")

#age plot
df1 <- df %>% select(tage) %>% na.omit()
```

```r
tage <- ggplot(df1,aes(x = as.factor(tage)))+geom_bar(fill = '#293352')+theme_minimal()+
  xlab("")+ggtitle("Children Age")

#sex plot
df1 <- df %>% select(tsex) %>% na.omit()
tsex <- ggplot(df1,aes(x = tsex))+geom_bar(fill = '#293352')+theme_minimal()+
  xlab("")+ggtitle("Children Sex")+
  scale_x_discrete(labels = c("0" = "Male",
                              "1" = "Female"))+
  theme(axis.text.x=element_text(size=rel(.7)))

#ethnicity plot
df1 <- df %>% select(tethnic) %>% na.omit()
tethnic <- ggplot(df1, aes(x = as.factor(tethnic)))+geom_bar(fill = '#293352')+
  theme_minimal()+
  xlab("")+ggtitle("Hispanic or Latino")

#join the plots together
p1 <- ggarrange(tage,tsex,labels = c("A", "B"), ncol = 2)
p2 <- ggarrange(trace,tethnic,labels = c("C", "D"), ncol = 2)
#ggarrange(p1,p2, nrow = 2)
df %>%
  dplyr::select(trace, tsex, tethnic, tage) %>%
  tbl_summary() %>%
  bold_labels()%>%
kable(caption = "Children Demographics",booktabs=T,row.names = F) %>%
   kable_styling(full_width=T, font_size = 10,latex_options = c('scale_down'))



#Creating Variable for Category Smoker
 df <- df %>%
  mutate(Smoke_level = case_when((as.numeric(mom_smoke_16wk)-1+as.numeric(mom_smoke_22wk)-
                                  (as.numeric(mom_smoke_16wk)-1+as.numeric(mom_smoke_22wk)-
                                  (as.numeric(mom_smoke_16wk)-1+as.numeric(mom_smoke_22wk)-
                                  (as.numeric(mom_smoke_16wk)-1+as.numeric(mom_smoke_22wk)-
                                                                  TRUE ~ NA))


#creating summary table
df %>%
```

```
  tbl_summary(include = c(erq_cog,erq_exp, bpm_att_p, bpm_ext_p, bpm_int_p,
                          swan_inattentive, swan_hyperactive, cotimean_pp6mo_baby),
              type = list(everything() ~ 'continuous'),
                     digits = list(everything() ~ c(2)),
                     statistic = list(~ "{mean} ({sd})"),
                     by = Smoke_level,
                     missing = "no") %>%
  kable(booktabs = TRUE, caption = "Mean and SD for tests regarding Child") %>% kableExtra
#creating summary table for susbtance use at least once
df %>% tbl_summary(include = c(cig_ever, e_cig_ever, mj_ever, alc_ever),
                   digits = list(everything() ~ c(2)),
                   statistic = list(all_continuous() ~ "{mean} ({sd})"),
                   by = Smoke_level,
                   missing = "no") %>%
  kable(booktabs = TRUE, caption = "Number of Children who have used substance at least on


#Plots for number of days of substance use
#cig plot
df1 <- df %>% select(num_cigs_30, Smoke_level) %>% na.omit()
p1 <- ggplot(df1, aes(x = Smoke_level,y = num_cigs_30, color = Smoke_level))+
  geom_boxplot()+theme_minimal()+theme(axis.text.x=element_text(size=rel(.7)))+
  scale_color_manual(values = cbp2)

#ecig plot
df1 <- df %>% select(num_e_cigs_30, Smoke_level) %>% na.omit()
p2 <- ggplot(df1, aes(x = Smoke_level,y = num_e_cigs_30, color = Smoke_level))+
  geom_boxplot()+theme_minimal()+theme(axis.text.x=element_text(size=rel(.7)))+
  scale_color_manual(values = cbp2)

#mj plot
df1 <- df %>% select(num_mj_30, Smoke_level) %>% na.omit()
p3 <- ggplot(df1, aes(x = Smoke_level,y = num_mj_30, color = Smoke_level))+
  geom_boxplot()+theme_minimal()+theme(axis.text.x=element_text(size=rel(.7)))+
  scale_color_manual(values = cbp2)

#alc plot
df1 <- df %>% select(num_alc_30, Smoke_level) %>% na.omit()
p4<- ggplot(df1, aes(x = Smoke_level,y = num_alc_30, color = Smoke_level))+
  geom_boxplot()+theme_minimal()+theme(axis.text.x=element_text(size=rel(.7)))+
  scale_color_manual(values = cbp2)
```

```r
#join them together
ggarrange(p1,p2,p3,p4,common.legend = T, legend = "bottom") %>% annotate_figure( text_grob


#creating new smoke exposure level variable

#calcualte the mean accros all exposure
smoke_mean <- df %>%
  select(smoke_exposure_6mo, smoke_exposure_12mo, smoke_exposure_2yr,
         smoke_exposure_3yr, smoke_exposure_4yr,
              smoke_exposure_5yr) %>%
  mutate_if(is.factor, as.numeric) %>%
  mutate_all(~ . - 1) %>%
  rowMeans(na.rm = T)

#join them
df <- cbind(df,smoke_mean)

#create the variable
df <- df %>%
  mutate(smoke_exposure_level = case_when(smoke_mean == 0 ~ "Not exposed",
                                          (smoke_mean > 0 & smoke_mean <= .5) ~ "Moderatel
                                          smoke_mean > .5 ~ "Extremely exposed"
                                          ))
#take only non-smoker to evaluate only the effect of ETS
df_exp <- df %>% filter(Smoke_level == "Non-Smoker")
#table for summary of tests
df_exp %>%
  tbl_summary(include = c(erq_cog,erq_exp, bpm_att_p, bpm_ext_p, bpm_int_p,
                          swan_inattentive, swan_hyperactive, cotimean_pp6mo_baby),
              type = list(everything() ~ 'continuous'),
                     digits = list(everything() ~ c(2)),
                     statistic = list(~ "{mean} ({sd})"),
                     by = smoke_exposure_level,
                     missing = "no") %>%
  kable(booktabs = TRUE, caption = "Mean and SD for tests regarding Child") %>%kableExtra:

#table for summary of substance use at least once
df_exp %>% tbl_summary(include = c(cig_ever, e_cig_ever, mj_ever, alc_ever),
                    digits = list(everything() ~ c(2)),
                    statistic = list(all_continuous() ~ "{mean} ({sd})"),
```

```
                    by = smoke_exposure_level,
                    missing = "no") %>%
  kable(booktabs = TRUE, caption = "Number of Children who have used substance at least on


#grubbs.test(df_exp$num_mj_30)
```