Project 1

Monica Colon Vargas

Introduction

Exposure to smoking during pregnancy (SDP) and exposure to environmental tobacco smoke (ETS) represent two environmental factors that affects children. 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. 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. The participants in this study were initially enlisted from a prior research project (n=738) and a subset was randomly selected to this study (n=100).

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" and "44989," which were both considered unrealistic and thus set to 'NA.' Additionally, "20-25" was adjusted to its mean value of 23.5, 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 dataset."

To start working on the dataset, first note that there were some issues on values of some variables. For example, on the income variable, there was one value that was changed from "250 000" to 250000, the variable mom_cig had one value of "40" which does not make sense based on the question asked for this variable so this was changed to NA, and mom_numcig got "2 black and miles a day"change to 2, 44989 to NA (this value is absurd), "20-25" change to its mean (23.5), and "none"to 0. Next, we changed the NA values of num_cigs_30, num_e_cigs_30, num_mj_30, and num_alc_30 to zero if they had a zero on the "X_ever" variable (where X is different for each). With these values we turn to observing missing data.

This data does not have a complete case which means that if we omit all the NA values the whole data set will be eliminated. Table 1, shows the number of missing data on each variable. We can see that there are some variables with the same number of missing values. Figure 1 shows a plot for the missing values on each individual per each patient. The patients are on the y-axis, we can see that there are almost 50 (the total number of patients is 49) . Note that many variables have missing values for the same patient and many patients have the same missing variables together. For example, the variables page to pethnic, pemploy to mom_numcig, mom_smoke_32wk, mom_smoke_pp1, bpm_att_p to tethnic, and cig_ever to pqm_paprental_control have missing values for the same 8 patients. This patterns may implies that the data is not MCAR but MAR. Since the are more variables than observations, multiple imputation may not be the efficient way to deal with them. However, for the purpose of this project, when studying certain variables' relationship, we will consider only the complete cases.

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
income	13	26.53	nidaalc	10	20.41
bpm__att__p	13	26.53	nidatob	10	20.41
tsex	13	26.53	nidaill	10	20.41
num__e__cigs__30	13	26.53	bpm__int__p	10	20.41
alc__ever	13	26.53	smoke__exposure__6mo	10	20.41
erq__cog	13	26.53	smoke__exposure__12mo	10	20.41
erq__exp	13	26.53	smoke__exposure__2yr	10	20.41
pmq__child__disclosure	13	26.53	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			

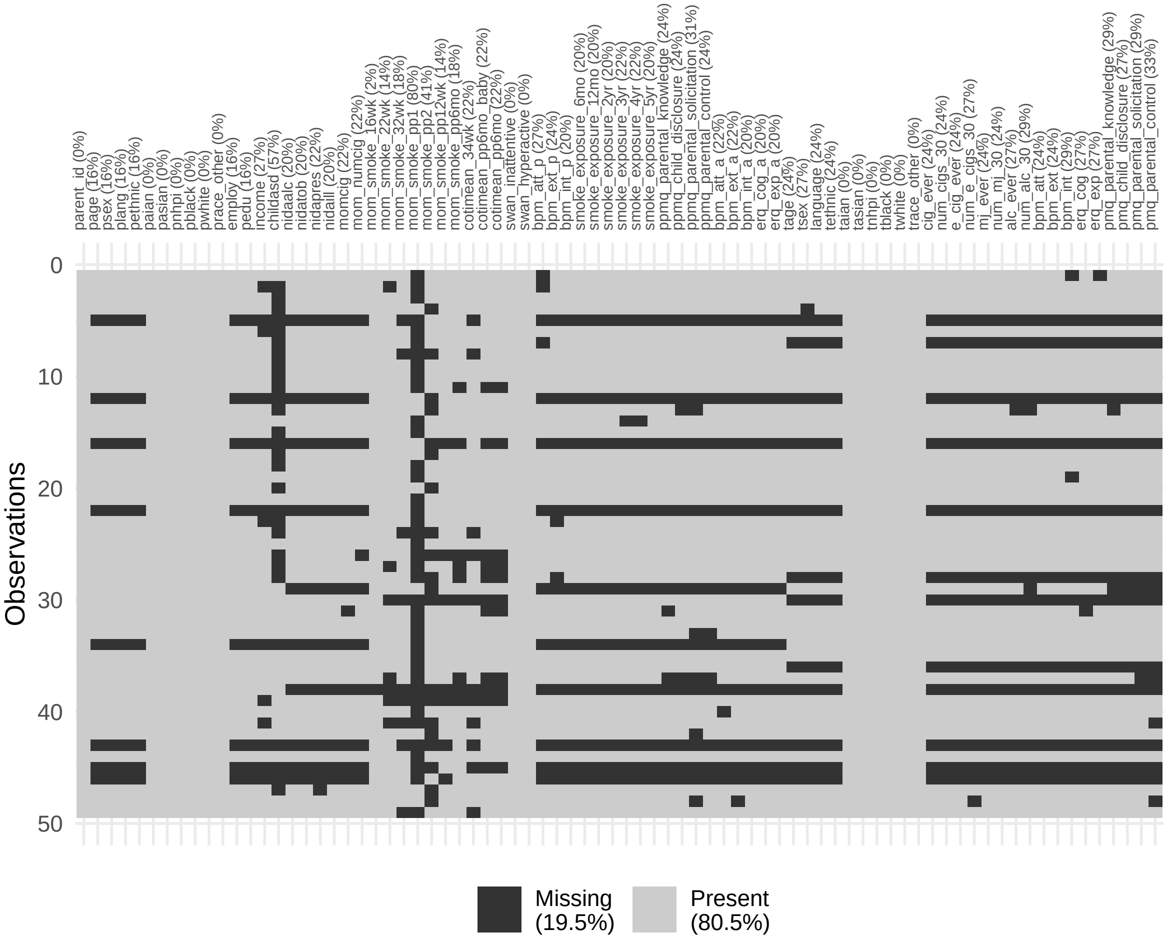

Figure 1: Missing Data Pattern

Demographics

We know provide plots for demographics of the parent and the child. Figure 2 shows the demographics of the parent. First, note that there is one male in the data-set which seems strange since the parents involved in the study experience pregnancy which implies it may be an error. The majority of the mothers are or race white, followed by native Hawaiian, other and Alaskan native and most are not considered as Hispanic or Latino ethnicity and the majority of age lies from 33 to 39 years old. Most mothers were employ full-time and the mean estimated household income was $63,138.05 which may be considered average and the majority of the mothers were had some college education level.

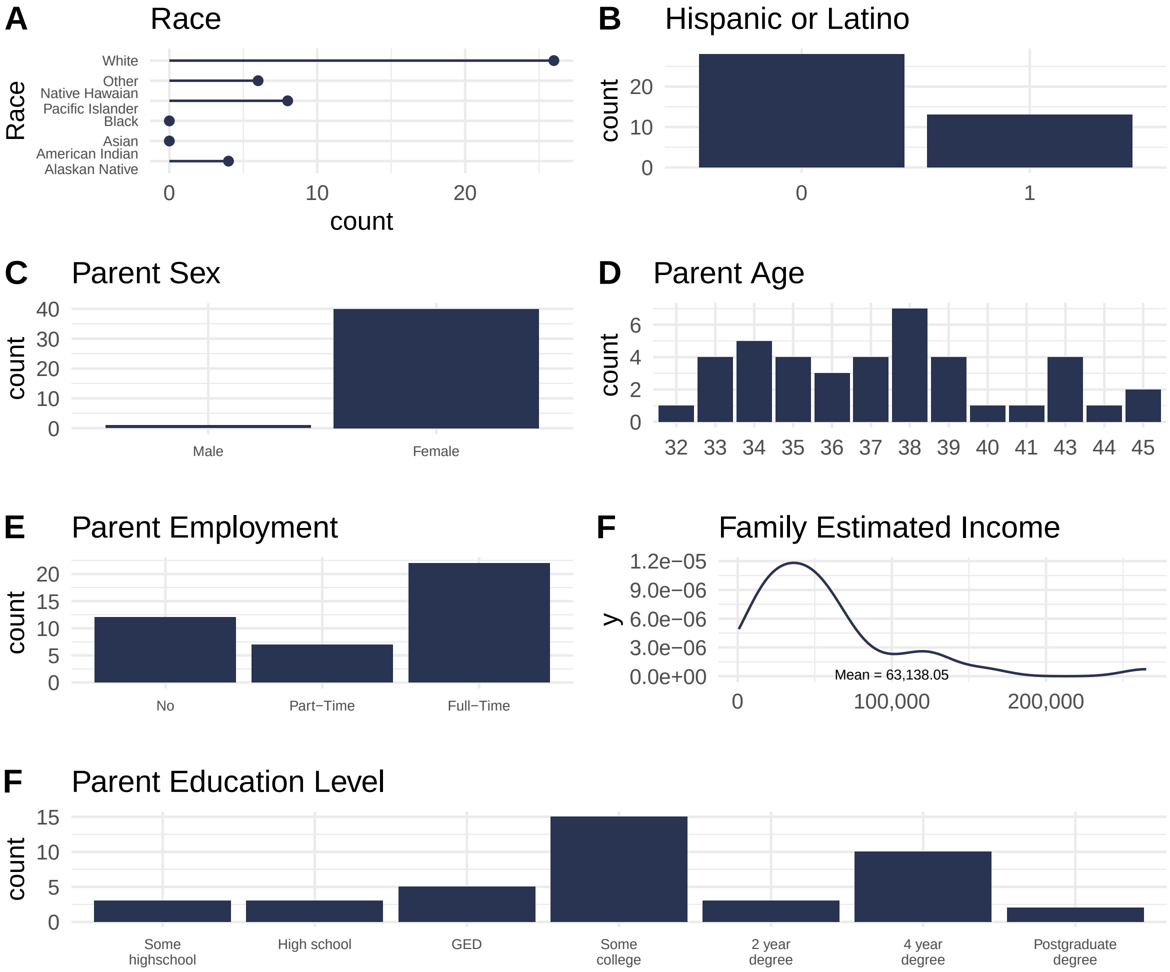

Figure 2: Parent Demographics

Figure 3 shows the demographics of the children in the study, with most being male and around 12-15 years old. Most children considered themselves as white race followed by black and the majority were not considered as Hispanic or Latino ethnicity. On figure 2A, note that no mothers were black race but many of the children were.

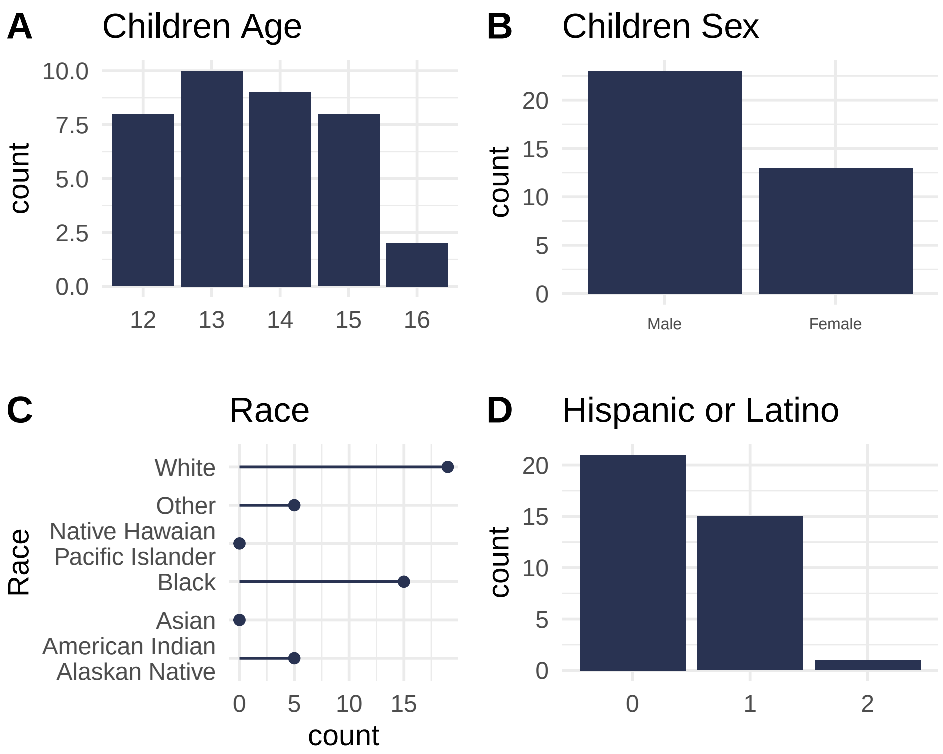

Figure 3: Child Demographics

Smoking During Pregnancy (SDP)

In order to evaluate the effects of SDP, the first thing to do is to create a class that separates a mother into three categories: Non-Smoker, Moderate-Smoker, and Heavy-Smoker. There are three variables that measures if the mom smoke during pregnancy, mom_smoke_16wk, mom_smoke_22wk, and mom_smoke_32wk which specifies if the mom smoke during the 16, 22, or 32 week of pregnancy respectively. Mothers who mark yes to all the variables were categorized as heavy-smokers, while mothers who mark all no were categorized as non-smokers. If the mother marked yes to at least one variable then she was categorized as a moderate-smoker. Table 2 shows the average of some testt regarding the child that were recorded. The erq variables consists of the emotional regulation questionnaire. Higher values implies the child a have worse tendency to regulate their emotions in cognitive reappraisal (erq_cog) and expressive suppression (erq_exp). It can be seen that children from a non-smoker mother overall have a smaller mean in both variables. Children from heavy-smokers mothers had a lower mean than from moderate smokers, although this could be because there are few observations in that group (n=3). Due to extreme small sample in moderate-smoker mothers, we will only be comparing heavy-smokers to non-smokers. The bpm variables consists of de Brief Problem Monitor where parents responded how true each statement was about their child rating attention (bpm_att_p), externalizing (bpm_ext_p) and internalizing (bpm_int_p) problems. Higher values indicate higher problems on their child. Note that non-smoker mothers' child have lower mean when compared to heavy smokers. Additionally, measures for SWAN test regarding ADHD-Hyperactive/Impulsive type (swan_hyperactive) and ADHD-Inattentive type (swan_hyperactive) were measured. In this two variables we can clearly see that children of heavy-smokers mothers have a higher score than of non-smoker mothers implying the child is likely to have ADHD of the type. Lastly we provided the mean for the urine nicotine levels of the child at 6 months old (cotimean_pp6mo_baby) which shows a clear difference between the children of heavy-smoker and non-smoker mothers. This results might implicate that indeed SDP have a negative effect on self-regulation and externalizing behavior issues in the child. Of course this is only exploratory, and this differences need to be statistically tested.

Table 2: 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 3: 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.

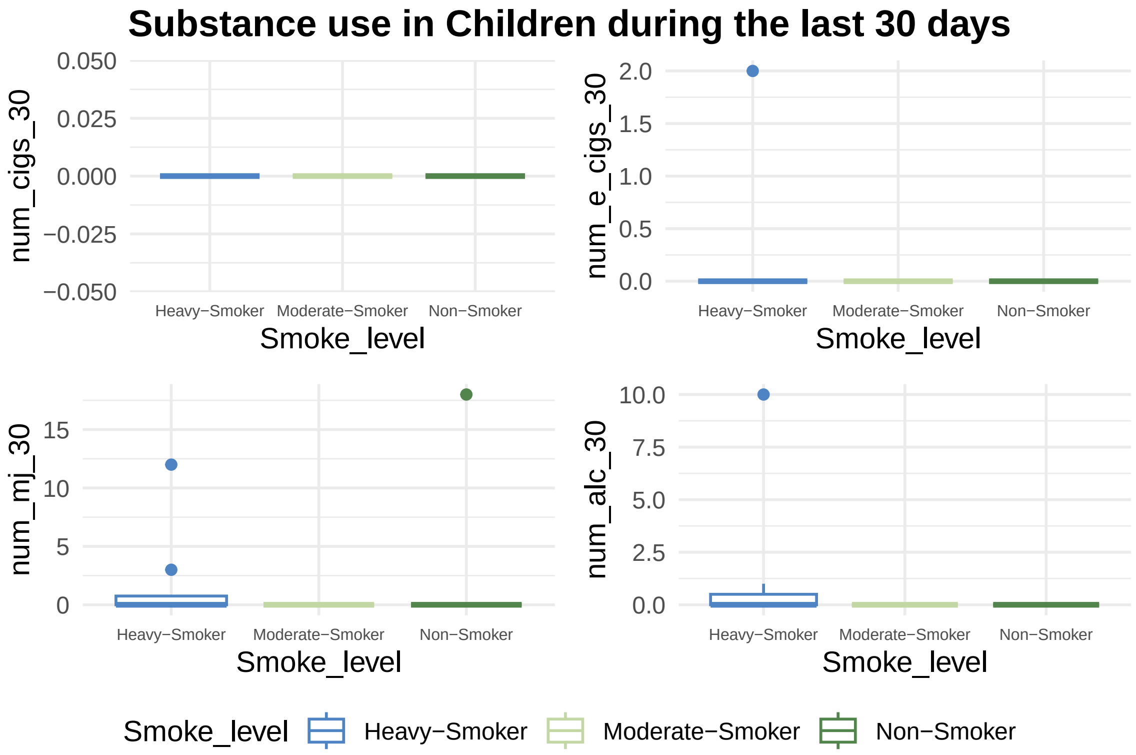

Figure 4: Child Substance use in the past 30 days

Environmental Tobacco Smoke (ETS)

We know turn to study the effect of (ETS) on the effect on self-regulation, externalizing behavior issues in the child, and substance use. In order to do so, a new variable called smoke_exposure_level was created based on the smoke exposure variables (smoke_exposure_6mo, smoke_exposure_12mo, smoke_exposure_2yr, smoke_exposure_3yr, smoke_exposure_4yr, and smoke_exposure_5yr) that were measured. These smoke exposure variables were 0 is the child was not exposed to smoke from mother or father during the time (at 6mo, 12mo, 1yr, 2yr, 3yr, 4yr, and 5yr) and 1 if the child was exposed to smoke. Each individual was categorized as heavily, moderately or not exposed. This was made by first calculating the mean for each individual across these variables. If the mean was equal to zero, then they were categorized as not exposed. If the mean was greater than 0 but less than or equal than .5, they were categorized as moderately exposed and if it was greater than .5, then they were categorized to extremely exposed. Then in order to study the effect of ETS, we selected only the children whose mother was a non-smoker during pregnancy. The reason for this is to study just the effect of ETS and not have both ETS and SDP at the same time.

Table 4 shows the mean and standard deviation for the same tests explained in the SDP section. Note that the Emotion Regulation Questionnaire on cognitive repraissal (erq_cog) and expressive suppressionn (erq_exp) for not exposed children have overall a smaller mean than children extremely and moderately exposed implying that ETS might have a negative effect for these questionnaires on children. The same conclusion hold for the Brief Monitor Problem (BPM) on externalizing and internalizing problems and the SWAN inattentive and hyperactive scores. However, we see that the mean of the BPM in attention problems is slightly higher for children not exposed and the mean of urine nicotine in the baby at six months. Note that these results need to be handled with care due to the very small sample size in children extremely and moderately exposed.

Table 4: 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 shows the number of children who have used a cigarette, e-cigarette, marijuana and/or alchohol 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). Due to the small number of children who have experienced substance use, the number of days during the past 30 days that they have used a substance is not showed,. However, it was noted that the same not exposed child who have experience with marijuana at least once was the only child had used substance on the past 30 days (18 days out of the 30 days) and this value was tested using Dixon's test and resulted in being an outline.

Table 5: 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's 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 datasets to corroborate our results.

Code Appendix

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

#Load data
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


#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)
```

```r
#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)
missing_cols[ nrow(missing_cols) + 1 , ] <- ""
missing_cols$Variable <- rownames(missing_cols)
missing_cols$Variable[64] <- ""

dd <- missing_cols %>% select(Variable, n, '%')
dd2 <- cbind(dd[1:32, ],dd[33:64,])

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
vis_miss(df)+theme(axis.text.x=element_text(size=rel(.72), angle = 90))
```

```r
#Race Variable
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))

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(.
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)))


#page <- ggplot(df,aes(x = page))+geom_density(color = '#293352')+theme_minimal()+
#  xlab("")+ggtitle("Parent Age")

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

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)))

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)

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)))


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")


#ggarrange(prace,page,pincome,ggarrange(pedu,pemploy))

p1 <- ggarrange(prace, pethnic, nrow = 1,ncol = 2,labels = c("A", "B"))
p2 <- ggarrange(psex,page, labels = c("C", "D"), nrow = 1,ncol = 2)
p3 <- ggarrange(pemploy,pincome, labels = c("E","F"), ncol = 2)
p4 <- ggarrange(pedu, labels = "F")
ggarrange(p1,p2,p3,p4, nrow = 4)
#Race Variable
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))

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")

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

```
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)))

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")


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)
#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))



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:

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

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)

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)

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)

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)

ggarrange(p1,p2,p3,p4,common.legend = T, legend = "bottom") %>% annotate_figure( text_grob

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)

df <- cbind(df,smoke_mean)

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"
                                          ))
```

```
df_exp <- df %>% filter(Smoke_level == "Non-Smoker")

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:



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)
```