Project 1 Practical Data Analysis

Exploratory Data Analysis

2023-10-08

Prenatal Tobacco Exposure and Child Outcomes

Github Repository: https://github.com/kasigi234/Exploratory-Data-Analysis

1. Background

Exposure to smoking during pregnancy (SDP) and exposure to environmental tobacco smoke (ETS) are among the most widespread and harmful environmental factors that affect children. Previous studies have shown that approximately 7% to 15% of newborns each year are exposed to SDP, and over 25% of children are exposed to ETS within their homes. Dr Lauren Micalizzi's study seeks to examine self-regulation and substance use/externalizing behavior in youth who were prenatally exposed to smoking during pregnancy. It is hypothesized that early smoke exposure is associated with difficulties in self-regulation, early substance use and externalizing problems among children which affects their control of physiological, emotional, behavioral, and cognitive aspects. The primary objective for this analysis is to examine the association between smoking during pregnancy (SDP) and environmental tobacco smoking (ETS) exposure and self-regulation, externalizing behavior, and substance use.

2.Data

We use data collected from a randomly selected sub-population (N = 100) of mothers and their (12-16 years) adolescent children from a previous study on smoke avoidance intervention to reduce smoking and environmental tobacco smoking (ETS) exposure during pregnancy and exposure to ETS in the immediate postpartum period for children among low-income women. The data comprises mothers' and children responses on both prenatal and postnatal smoking exposures , self-regulation, and substance use for children. The data was earlier preprocessed and reduced to 49 observations for both mothers and their children measured over 78 variables on child exposure to smoke during and after pregnancy, as well as measurements for child behavior on externalizing, self-regulation, and substance. The independent and dependent variables to be used in the analysis are created based on the maternal report on smoking during pregnancy (SDP) and environmental tobacco smoking (ETS) while the dependent variables are substance use (SU), externalizing problems (EXT), and self-regulation (SR).

2.1 Data cleaning

Before exploring the data further additional reprocessing was done to ensure that the data was clean from any anomalies and missingness handled appropriately. This involved cleaning the outliers, handling inconsistencies as well as missing data to ensure quality and reliability of the data. All blank entries were re-recorded as NA values. **'mom_numcig'** which collects information on the mother's daily cigarettes consumption had several records that were abnormally different; "2 black and miles a day","20-25", "None" and "44989". These records were adjusted accordingly on the assumption that it meant the number and the average of cigarettes smoked while none simply meant 0 whereas 44989 as NA. **Income** data had different formats for the entries and this was standardized. **'momcig'** had an outlier value of 40 which was adjusted to 4 assuming that it was an error during data entry. Further cleaning involved data type transformation for continuous variables and binary records. **'num_cigs_30'**, **'num_e_cigs_30'**, **'num_mj_30'**, **'num_alc_30'** which detailed child's substance use were originally recorded as NAs and this was adjusted to 0 if the child reported to be a non cigarette, e-cigarette, marijuana or alcohol user. This reduced the amount of missingness that were initially observed in the original data.

2.2 Missing Data

\begin{table}

\caption{Variables with above 25% Missing Data}

Variable	Proportion (%)	n
mom_smoke_pp1	39	79.59
childasd	28	57.14
mom_smoke_pp2	20	40.82
pmq_parental_control	16	32.65
ppmq_parental_solicitation	15	30.61
num_alc_30	14	28.57
bpm_int	14	28.57
pmq_parental_knowledge	14	28.57
pmq_parental_solicitation	14	28.57
bpm_att_p	13	26.53
tsex	13	26.53
num_e_cigs_30	13	26.53
alc_ever	13	26.53
erq_cog	13	26.53
erq_exp	13	26.53
pmq_child_disclosure	13	26.53

\end{table} From the above table ,we note that the '**mom_smoke_pp1**' which indicates whether the mother was smoking during the first postpartum visit has 79.6% of missingness, indicating a substantial amount of missing data which seems inconsistent with the other postpartum visits. '**childasd**' also exhibit relatively high percentage of missingness at 57% while the other variables have less that 50% varying degrees of missingness. Certain variables, such as '**num_alc_30**', '**bpm_int**', '**pmq_parental_knowledge**','**pmq_parental_solicitation**' are notable seen to be missing together at 28.57%. Similarly, '**bpm_att_p**', '**tsex**', '**nnum_e_cigs_30**','**alc_ever**','**erq_cog**','**erq_exp**' and '**pmq_child_disclosure**' variables are also missing together at 26.53. These findings underscore the importance of addressing the missing data but the study's limitation is on the small amount of data, therefore missing data methods such as imputation were not be plausible as this was likely to induce bias to the results estimates. In subsequent analysis missing data for the variables of interest were dropped.

Also, there appears to be a trend in the missingness across the variables. Most variables appear to be missing together. This consistent pattern suggests that the missingness is not at random (MNAR) indicating that there is a common factor influencing the data and these could be due to the unobserved factors which may be related to the nature of these variables and how they were measured during data collection or unobserved characteristics of the mothers and the children in this study that were not measured. For instance mothers with extremely high smoking habits are likely to be biased in their response as they might be less likely to disclose their true smoking behavior.

Table 1: Demographic Characteristics

Variable	N = 49
Mom Age	37 (35, 39)
Mom Education	NA
0	3 (7.3%)
1	3 (7.3%)
2	5 (12%)
3	15 (37%)
4	3 (7.3%)
5	10 (24%)
6	2 (4.9%)
Mom Employed	NA
0	12 (29%)
1	7 (17%)
2	22 (54%)
Child Age	NA
12	8 (22%)
13	10 (27%)
14	9 (24%)
15	8 (22%)
16	2 (5.4%)
Child Sex	NA
0	23 (64%)
1	13 (36%)
Mom Other Race	NA
0	43 (88%)
1	6 (12%)
Child Other Race	NA
0	44 (90%)
1	5 (10%)

3. Analysis

3.1 Demographic summary statistics

From the table, the average child age is 13 years with 36% being females. The mother's mean age is 37 years with interquartile range of 35-39 years with 37% having some college education, 24% had 4 year degree education level while about 5% have a postgraduate degree. Slightly above half the mothers who participated in the study had full time jobs (54%), while about 29% were unemployed. Further analysis will be guided by an understanding of the participants characteristics which may be potentially confound the results.

To further the analysis, variables were grouped into independent (exposure) and dependent variables (outcome). The variables were based on the information on the questionnaires, codebook and how correlated the variables seemed to be based on correlation matrices.

3.2 Aggregate Variable for Exposure

The aggregate exposures incorporated both trimester timing and dosage effect from prenatal cotinine levels which were informed by prospective maternal report on the questionnaires of smoking during and after pregnancy and the environmental tobacco smoking to come up with smoking during pregnancy (SDP) and environmental tobacco smoking (ETS) variables. SDP was a composite for all the variables that relate to mothers smoking status during pregnancy as well as urine cotinine levels during that period whereas ETS comprised of all variables for smoke exposure post pregnancy and the mother and baby's urine cotinine levels. All these were put together aggregated exposures. Mother and baby's cotinine data were first transformed using the logarithmic scale and normalized on the z-scale before their inclusion in the composite exposure variable to be used in the analysis.

3.2.1 Interrelatedness of SDP and ETS

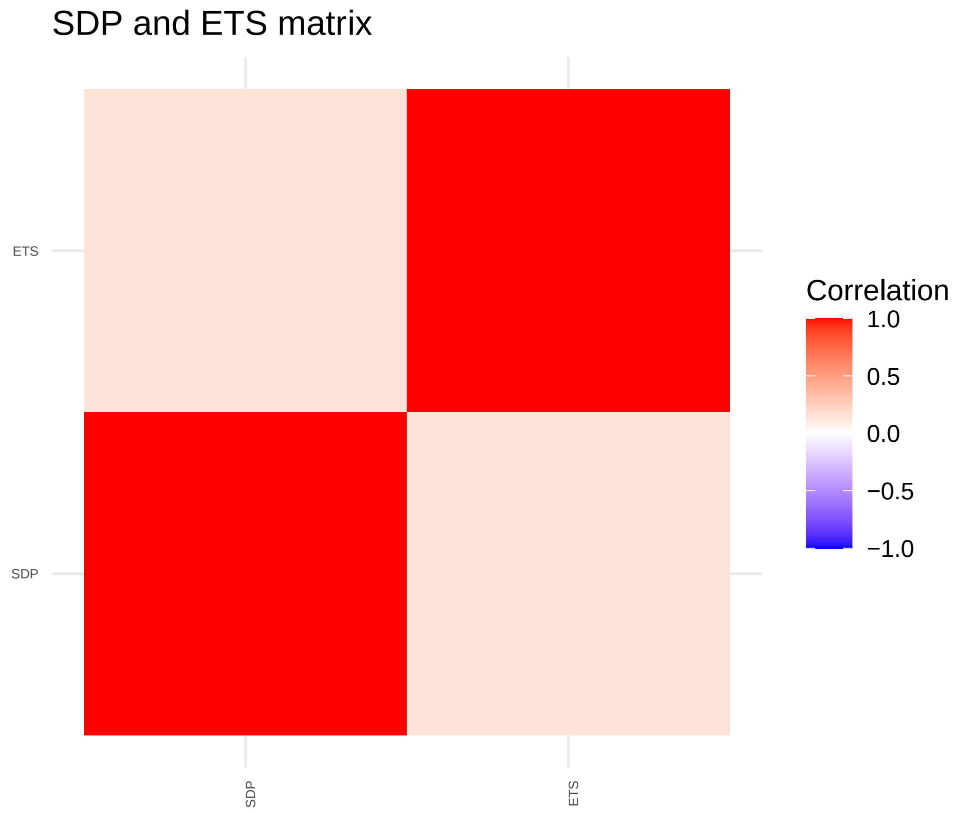

Figure 1: SDP vs ETS Correlation Matrix

Figure 1 above shows that smoke exposure during pregnancy (SDP) and environmental tobacco exposure (ETS) have a fairly positive correlation with each other indicating a weak interrelatedness of the smoking during pregnancy and environmental tobacco smoking composite variables for SDP and ETS were fairly correlated with each other.

3.3 Outcome Variables

Also, the outcome variables were created and they included self regulation (SR), externalizing (Ext) problem and substance use. EXT and SR use were based on the responses to questions relating to emotional regulation, attention deficit and problem monitor from both the mother and the child. Potential correlations for all these variables were assessed using a correlation matrix before the creation of the EXT and SR sub-groups. Substance use (SU) was created based on the child's response on any of the substances being studied.

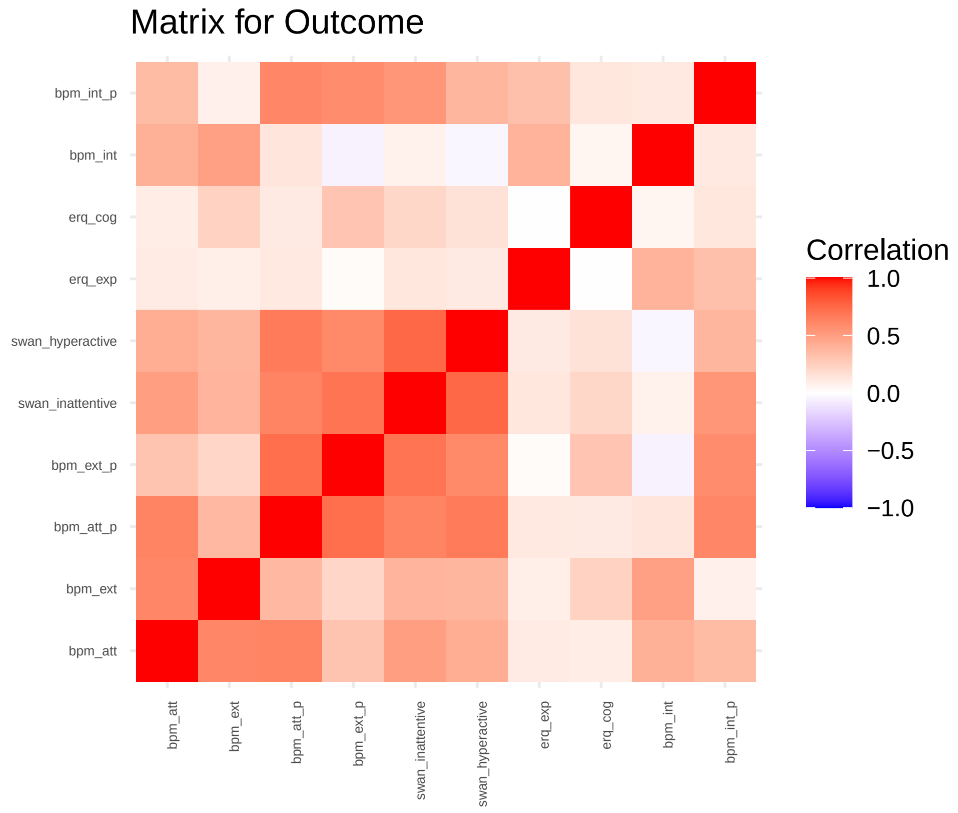

Figure 2: Correlation Matrix

Figure 2 above, shows that there is an evident correlation between the variables for externalizing and self regulation variables. This aligns with the descriptions on the codebook and the questionnaires relating to child externalizing and attention problem. Some other variables exhibit a fair correlation. These variables will then be categorized as either self regulation (SR) or externalizing (EXT).

3.3.1 Aggregate scores for EXT and SR

Before embarking on the primary objective for the analysis, an aggregate variable for the outcome was created. Basing on the correlation matrix for the outcome, the two sub-groups were created for the EXT and SR. As seen earlier in the correlation matrix for the outcomes most of the variables were seen to be highly correlated. This enabled the creation of an aggregate variable for externalizing and self regulate variables separately while also basing on the provided code book. The variables relating to attention and regulation scores were first normalized using Z-score to standardize the values to account for the different systems utilized during generation of the scores.

3.3.2 Interrelatedness of the externalizing and self regulate behaviors in children

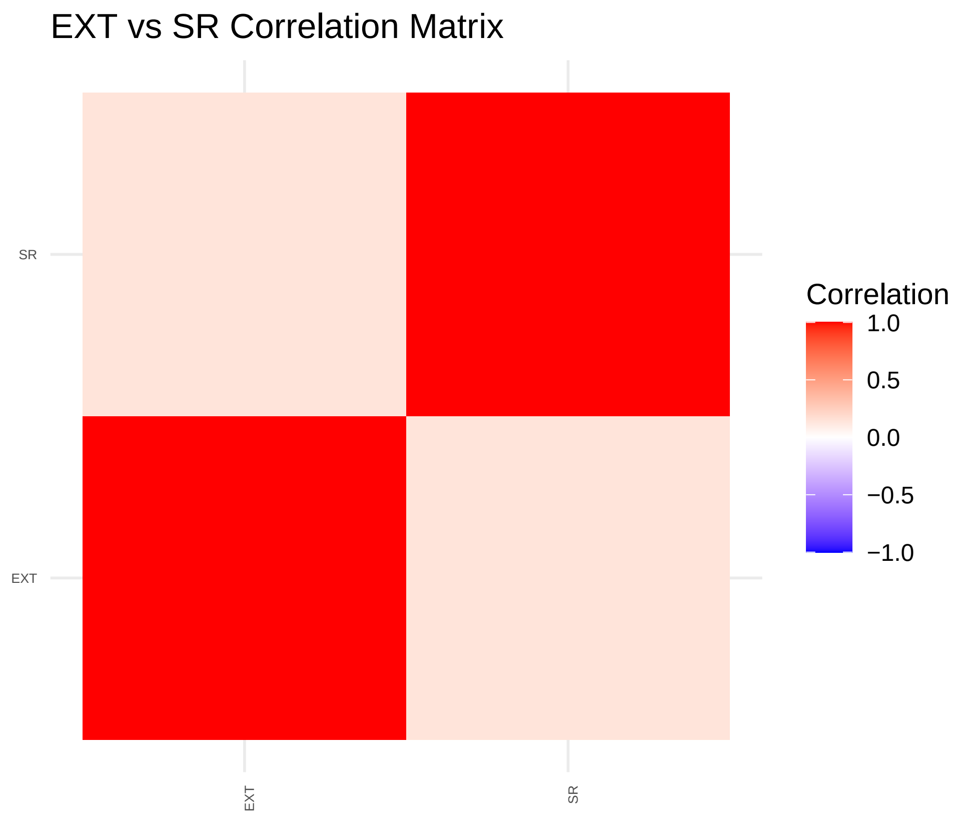

Figure 3: EXT/SR Matrix

Figure 3 shows that the externalizing and self regulate problems among the children are correlated suggesting that a child who exhibits externalizing problems is likely to experience self regulation problems.

4. Effect of ETS on Externalizing problems

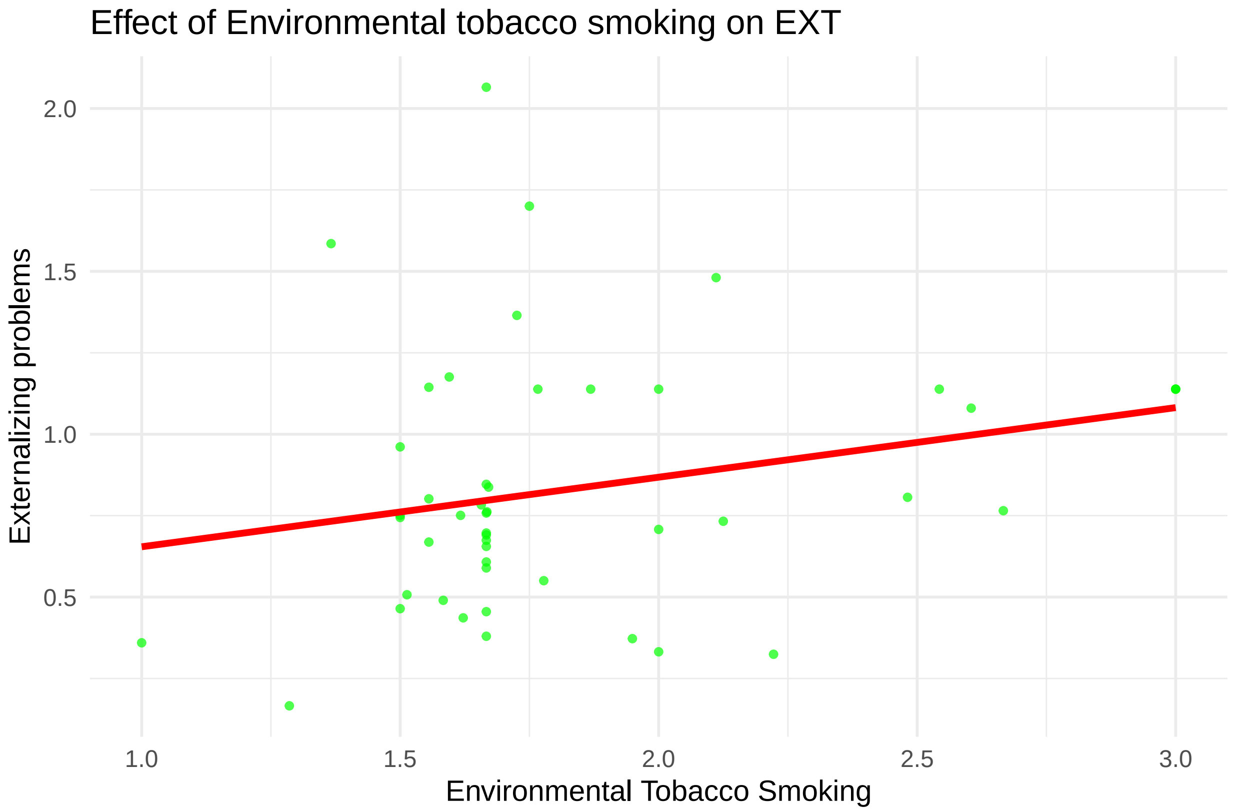

Figure 4: Effect of Environmental Tobacco Smoking on Externalizing problems

Figure 4 shows externalizing problems in children increases with increasing environmental tobacco smoking suggesting that children exposed to environmental tobacco smoking have capabilities to externalizing problems, self regulation and escalated substance use. Although no effect was notable for smoking during pregnancy on the outcomes. Linear models were later used to examine whether SDP had an effect on the outcomes with adjusted demographic characteristics.

To wrap the exploratory analysis process, logistics for binary outcome and linear regression models were utilized to further examine the primary objective. During model fitting we recognized the critical importance of model selection to ensure that the statistical analysis accurately captures the relationship between environmental tobacco smoking (ETS) and externalizing problems (EXT). To achieve this, both linear and logistics regression models were evaluated.Each model was assessed based on their ability to fit the data. Goodness of fit was considered for the presence of statistically significant pvalues for the predictors. After a comprehensive model selection process the best model for evaluating SDP/ETS on Externalizing problems, the best model with race, employment, education controlling for parents age and SDP had significant pvalue of 0.01457. The logistics regression model for Substance use as the outcome did not show any significant results even after controlling for SDP and other baseline covariates.

In conclusion the analysis showed that environmental tobacco smoking is associated with self regulation and externalizing behavior among children. Further analysis is required to further explore these effects and to extensively account for any potential confounding factors.

Appendix:

Github

For a full code appendix go to this Project Github Repo:
https://github.com/kasigi234/Exploratory-Data-Analysis

Code

```r
knitr::opts_chunk$set(echo = FALSE)

# libraries to be used
#install.packages("kableExtra")
library(gtsummary)
library(tidyverse)
library(kableExtra)
library(knitr)
library(tidyr)
library(naniar)
library(tidyverse)
library(dplyr)
library(ggplot2)
library(tableone)
library(dplyr)
library(lattice)
library(reshape2)
library(formatR)
#library(summarytools)
#library(broom)

knitr::opts_chunk$set(echo = FALSE)
knitr::opts_chunk$set(warning = FALSE, message = FALSE, echo = FALSE, fig.align = "center")

#setwd("/Users/kasigi/OneDrive - Brown #University/Desktop/Fall_2023/PHP2550/Project1")

#setwd("C:\\Users\\kasigi\\OneDrive - Brown #University\\Documents\\GitHub\\Exploratory-Data-Analysis\\

# load the data
project_data <- read.csv("/Users/kasigi/OneDrive - Brown University/Desktop/Fall_2023/PHP2550/Project1/


# Data cleaning

# variable transformation
project_data$psex <- as.factor(project_data$psex)
project_data$plang <- as.factor(project_data$plang)
project_data$pethnic <- as.factor(project_data$pethnic)
project_data$paian <- as.factor(project_data$paian)
project_data$pasian <- as.factor(project_data$pasian)
project_data$pnhpi <- as.factor(project_data$pnhpi)
project_data$pblack <- as.factor(project_data$pblack)
project_data$pwhite <- as.factor(project_data$pwhite)
project_data$prace_other <- as.factor(project_data$prace_other)
project_data$employ <- as.factor(project_data$employ)
project_data$pedu <- as.factor(project_data$pedu)
project_data$childasd <- as.factor(project_data$childasd)
project_data$nidaalc <- as.factor(project_data$nidaalc)
project_data$nidatob <- as.factor(project_data$nidatob)
project_data$nidapres  <- as.factor(project_data$nidapres)
project_data$nidaill <- as.factor(project_data$nidaill)
project_data$mom_smoke_16wk <- as.factor(project_data$mom_smoke_16wk)
project_data$mom_smoke_22wk <- as.factor(project_data$mom_smoke_22wk)
```

```r
project_data$mom_smoke_32wk <- as.factor(project_data$mom_smoke_32wk)
project_data$mom_smoke_pp1 <- as.factor(project_data$mom_smoke_pp1)
project_data$mom_smoke_pp2 <- as.factor(project_data$mom_smoke_pp2)
project_data$mom_smoke_pp12wk <- as.factor(project_data$mom_smoke_pp12wk)
project_data$mom_smoke_pp6mo <- as.factor(project_data$mom_smoke_pp6mo)
project_data$smoke_exposure_6mo <- as.factor(project_data$smoke_exposure_6mo)
project_data$smoke_exposure_12mo <- as.factor(project_data$smoke_exposure_12mo)
project_data$smoke_exposure_2yr <- as.factor(project_data$smoke_exposure_2yr)
project_data$smoke_exposure_3yr <- as.factor(project_data$smoke_exposure_3yr)
project_data$smoke_exposure_4yr <- as.factor(project_data$smoke_exposure_4yr)
project_data$tsex <- as.factor(project_data$tsex)
project_data$language <- as.factor(project_data$language)
project_data$tethnic <- as.factor(project_data$tethnic)
project_data$alc_ever <- as.factor(project_data$alc_ever)
project_data$mj_ever <- as.factor(project_data$mj_ever)
project_data$e_cig_ever <- as.factor(project_data$e_cig_ever)
project_data$cig_ever <- as.factor(project_data$cig_ever)
project_data$taian <- as.factor(project_data$taian)
project_data$tasian <- as.factor(project_data$tasian)
project_data$tnhpi <- as.factor(project_data$tnhpi)
project_data$tblack <- as.factor(project_data$tblack)
project_data$twhite <- as.factor(project_data$twhite)
project_data$trace_other <- as.factor(project_data$trace_other)
project_data$smoke_exposure_5yr <- as.factor(project_data$smoke_exposure_5yr)
project_data$momcig <- as.numeric(project_data$momcig)
project_data$mom_numcig <- as.numeric(project_data$mom_numcig)
project_data$income <- as.numeric(project_data$income)
project_data$cotimean_34wk <- as.numeric(project_data$cotimean_34wk)
project_data$cotimean_pp6mo <- as.numeric(project_data$cotimean_pp6mo)
project_data$cotimean_pp6mo_baby <- as.numeric(project_data$cotimean_pp6mo_baby)

#project_data$mom_smoke_16wk <- as.numeric(project_data$mom_smoke_16wk)
#project_data$cmom_smoke_22wk <- as.numeric(project_data$mom_smoke_22wk)
#project_data$mom_smoke_32wk <- as.numeric(project_data$mom_smoke_32wk)


# blank values to NA
project_data[project_data == ""] <- NA

# change the income input
project_data$income[6] <- 250000
project_data$income[1] <- 76000

# handle the outlier value on momcig
project_data$momcig[31] <- 4

# replace the erronously recorded value with NA
project_data[1, "mom_numcig"] <- 2
project_data[26, "mom_numcig"] <- NA # 44489
project_data[37, "mom_numcig"] <- 23  # 20-25
project_data[47, "mom_numcig"] <- NA  # none

# update 'num_cig_30' column to show true values
```

```
project_data <- project_data %>% mutate(num_cigs_30 =
                                          case_when(cig_ever == 0 ~ 0,
                                          cig_ever == 1 ~ num_cigs_30,
                                          TRUE ~ NA))
# update 'num_e_cig_30' column
project_data <- project_data %>% mutate(num_e_cigs_30 = case_when(e_cig_ever == 0 ~ 0,
                                        e_cig_ever == 1 ~ num_e_cigs_30,
                                        TRUE ~ NA))
# update 'num_mj_30' column
project_data <- project_data %>% mutate(num_mj_30 = case_when(mj_ever == 0 ~ 0,
                                        mj_ever == 1 ~ num_mj_30,
                                        TRUE ~ NA))
# update 'num_alc_30' column
project_data <- project_data %>% mutate(num_alc_30 = case_when(alc_ever == 0 ~ 0,
                                        alc_ever == 1 ~ num_alc_30,
                                        TRUE ~ NA))

# summary statistics for missing data
missing_data <- data.frame(
  Variable = names(project_data),
  Missing_Count = sapply(project_data, function(x) sum(is.na(x)))
)
missing_data$Percentage_Missing <- missing_data$Missing_Count / nrow(project_data) * 100
missing_data <- missing_data %>%
  arrange(desc(Percentage_Missing))

# Remove row names
rownames(missing_data) <- NULL

# variables with missing values above 25%
missing_data_only <- missing_data %>%
  filter(Percentage_Missing > 25)
project_missing <- naniar::miss_var_summary(project_data, order = FALSE)
project_missing$pct_miss <- round(project_missing$pct_miss, 2)
missing_data_only$Missing_Count <- round(missing_data_only$Missing_Count, 2)

missing_data_only$Percentage_Missing <- round(missing_data_only$Percentage_Missing, 2)

missing_data_only %>%
  kbl(caption = "Variables with above 25% Missing Data",
col.names = linebreak(c("Variable", "Proportion (%)", "n"))) %>%
kable_styling(bootstrap_options = c("hover"), full_width = F) %>%
    row_spec(0, background = "black", color = "white")


# summary table for demographic characteristics
tbl_summary_demographic <-  project_data %>%
  select(page,pedu,employ,tage , tsex, prace_other, trace_other) %>%
  tbl_summary(missing = "no",
              label = list(page = "Mom Age", pedu = "Mom Education", employ = "Mom Employed",tage = "Chi

# change tbl_summary to a data frame
data_frame <- tbl_summary_demographic %>%
```

```r
  as_tibble()
column_names <- c("Variable", "N = 49")
colnames(data_frame) <- column_names

# for nicer display
kable_tbl <- data_frame %>%
  kbl(caption = "Demographic Characteristics",
col.names = linebreak(c("Variable", "N = 49"))) %>%
kable_styling(bootstrap_options = c("hover"), full_width = F) %>%
    row_spec(0, background = "black", color = "white")

kable_tbl

# Standardize cotinine variable for mom and child
mom_cotimean_34_z <- abs(log(scale(project_data$cotimean_34wk)))
mom_cotimean_6mo_z <- abs(log(scale(project_data$cotimean_pp6mo)))
child_cotimean_6mo_z <-abs(log(scale(project_data$cotimean_pp6mo_baby)))

# calculate composite scores for SDP variable
project_data$SDP <- with(project_data,
                         rowMeans(cbind(mom_numcig, mom_smoke_16wk,
                                        mom_smoke_22wk, mom_smoke_32wk,
                                        mom_cotimean_34_z),na.rm = TRUE))

# calculate composite scores for ETS variable
project_data$ETS <- with(project_data,
                         rowMeans(cbind(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,mom_cotimean_6mo_z,
                                        child_cotimean_6mo_z),na.rm = TRUE))


# correlation matrix for SDP  and ETS
correlation_matrix_b <- cor(project_data[c("SDP", "ETS")], use ="complete.obs")

# reshape the correlation matrix for ggplot
melted_cor_matrix_b <- melt(correlation_matrix_b)

# correlation matrix visual
ggplot(melted_cor_matrix_b, aes(Var1, Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1, 1), space = "Lab",
                        name = "Correlation") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 1,
                                   size = 5, hjust = 1),
        axis.text.y = element_text(size = 5)) +
  labs(x = NULL, y = NULL, fill = "Correlation") +
  coord_fixed() +
  ggtitle("SDP and ETS matrix")+
  labs(caption = "Figure 1: SDP vs ETS Correlation Matrix")
# correlation matrix for SR/EXT
correlation_matrix_out <- cor(project_data[c("bpm_att", "bpm_ext", "bpm_att_p", "bpm_ext_p", "swan_inat

# reshape the correlation matrix for ggplot
melted_cor_matrix_out <- melt(correlation_matrix_out)

# correlation matrix visual
ggplot(melted_cor_matrix_out, aes(Var1, Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1, 1), space = "Lab",
                       #breaks = seq(-1, 1, by = 0.5)) +
                       name = "Correlation") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 1,
                                   size = 5, hjust = 1),
        axis.text.y = element_text(size = 5)) +
  labs(x = NULL, y = NULL, fill = "Correlation") +
  coord_fixed() +
  ggtitle("Matrix for Outcome ")+
  labs(caption = "Figure 2: Correlation Matrix")

# Standardize SR, EXT variables for child
bpm_att_z = abs(scale(project_data$bpm_att))
bpm_ext_z = abs(scale(project_data$bpm_ext))
bpm_int_z = abs(scale(project_data$bpm_int))
erq_cog_z = abs(scale(project_data$erq_cog))
erq_exp_z = abs(scale(project_data$erq_exp))
erq_cog_a_z = abs(scale(project_data$erq_cog_a)) # exclude parent
erq_exp_a_z = abs(scale(project_data$erq_exp_a)) # exclude parent
bpm_att_p_z = abs(scale(project_data$bpm_att_p))
bpm_ext_p_z = abs(scale(project_data$bpm_ext_p))
bpm_int_p_z = abs(scale(project_data$bpm_int_p))
bpm_att_a_z = abs(scale(project_data$bpm_att_a)) # exclude parent
bpm_ext_a_z = abs(scale(project_data$bpm_ext_a)) # exclude parent
bpm_int_a_z = abs(scale(project_data$bpm_int_a)) # exclude parent
swan_hyperactive_z = abs(scale(project_data$swan_hyperactive))
swan_inattentive_z = abs(scale(project_data$swan_inattentive))

# calculate composite scores for self regulation variable
project_data$SR <- with(project_data,
                        rowMeans(cbind(erq_cog_z, erq_exp_z, bpm_int_p_z,
                                       bpm_int_z),na.rm = TRUE))

# calculate composite scores for externalizing variable
project_data$EXT <- with(project_data,
                         rowMeans(cbind(bpm_att_z, bpm_ext_z, bpm_att_p_z,
                                        bpm_ext_p_z, swan_inattentive_z, swan_hyperactive_z),na.rm = TRU

# composite for substance use variable
project_data$SU <- with(project_data,
                        ifelse(cig_ever == 1 | e_cig_ever == 1 |
                                 mj_ever == 1 | alc_ever == 1, 1, 0))


## variables for the outcome variables
df_outcomes = project_data %>%
  select(bpm_att, bpm_ext, bpm_att_p, bpm_ext_p, swan_inattentive,erq_cog, erq_exp, swan_hyperactive)

# correlation matrix for interrelatedness externalizing variables
correlation_matrix <- cor(project_data[c("bpm_att", "bpm_ext",
                                         "bpm_att_p","bpm_ext_p", "swan_inattentive",
                                         "swan_hyperactive")], use = "pairwise.complete.obs")

# correlation matrix for interrelatedness self regulation
correlation_matrix2 <- cor(project_data[c("erq_exp", "erq_cog", "bpm_int", "bpm_int_p")],
                           use = "pairwise.complete.obs")


# correlation matrix for SR/EXT
#correlation_matrix <- cor(project_data[c("EXT", "SU", "SR",
                            # "SDP", "EXT")], use ="complete.obs")
# correlation matrix for SR/EXT
correlation_matrix_c <- cor(project_data[c("EXT", "SR")], use ="complete.obs")

# reshape the correlation matrix for ggplot
melted_cor_matrix_c <- melt(correlation_matrix_c)

# correlation matrix visual
ggplot(melted_cor_matrix_c, aes(Var1, Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1, 1), space = "Lab",
                       #breaks = seq(-1, 1, by = 0.5)) +
                       name = "Correlation") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 1,
                                   size = 5, hjust = 1),
        axis.text.y = element_text(size = 5)) +
  labs(x = NULL, y = NULL, fill = "Correlation") +
  coord_fixed() +
  ggtitle("EXT vs SR Correlation Matrix")+
  labs(caption = "Figure 3: EXT/SR Matrix")

# ETS on Externalizing problems
ggplot(project_data, aes(x = ETS, y = EXT)) +
  geom_point(size = 1, alpha = 0.7, color = "green") +
  geom_smooth(method = "lm", se = FALSE, size = 1.2,
              linetype = "solid", color = "red") +
  labs(title = "Effect of Environmental tobacco smoking on EXT",
       x = "Environmental Tobacco Smoking",
       y = "Externalizing problems") +
  theme_minimal() +
  labs(caption = "Figure 4: Effect of Environmental Tobacco Smoking on Externalizing problems")
```

```
# model selection for externalizing problem as the outcome
# fit first model with all covariates
model3 = lm(EXT ~ SDP +  ETS + page + pethnic + paian + pnhpi + pwhite + prace_other + employ +
              pedu + income + tage + tethnic + taian + tblack + twhite + trace_other, data = project_da


# drop 'tethnic'
model3.1 = lm(EXT ~ SDP +  ETS + page + pethnic + paian + pnhpi + pwhite + prace_other + employ + pedu +


# drop 'trace_other'
model3.2 = lm(EXT ~ SDP +  ETS + page + pethnic + paian + pnhpi + pwhite + prace_other + employ +
                pedu + income + tage  + taian + tblack + twhite, data = project_data)


# drop 'pwhite'
model3.3 = lm(EXT ~ SDP +  ETS + page + pethnic + paian + pnhpi + twhite + prace_other + employ + pedu +

# drop 'taian'
model3.4 = lm(EXT ~ SDP +  ETS + page + pethnic + paian + pnhpi + pwhite + prace_other + employ + pedu +

# drop 'pnhpi'
model3.5 = lm(EXT ~ SDP +  ETS + page + pethnic + paian + pedu + pwhite + prace_other + employ + income


# drop 'income'
model3.6 = lm(EXT ~ SDP +  ETS + page+ pethnic+ paian + pedu + pwhite + prace_other + employ +
                tage  + tblack , data = project_data)


# final model # adjust for page
model3.6 = lm(EXT ~ SDP +  ETS + pethnic+ paian + pedu + pwhite + prace_other + employ +
                tage  + tblack , data = project_data)
#summary(model3.6)

logm1 = glm(as.factor(SU) ~ SDP +  ETS + page + pethnic + paian + pnhpi + pwhite + prace_other + employ
#summary(logm)

logm2 = glm(as.factor(SU) ~ SDP +  ETS  + prace_other +tblack   ,family = binomial,  data = project_dat
#summary(logm2)
```