PHP 2550: Project 1

Due: October 8th at 11:59pm

1. Introduction

Exposure to Smoke During Pregnancy (SDP) and Environmental Tobacco Smoke (ETS) poses significant health risks to children. Dr. Lauren Micalizzi's study seeks to understand the effects of SDP and ETS on children's behavior. Participants were sourced from a prior smoke avoidance intervention study involving low-income women, with a subset of adolescents and their mothers selected for this research.

The primary objectives for this report include analyzing descriptive statistics for outliers and missing data, and formulating exposure variables for SDP and ETS while exploring their interrelationship. As well as delving into the connections among outcome variables concerning self-regulation, externalizing, and substance use, and conducting initial regressions to evaluate the influence of prenatal and postnatal exposure on these outcomes.

Github Project Repository: github.com/Kdossal/2550_Project1

2. Data Cleanliness and Quality

After cleaning and preparation, the original dataset was reduced to consist of 78 variables from 49 participants. This cleaned dataset is the main focus of this Exploratory Data Analysis Report. A detailed code book, providing insights into the variable names and related data, is available on the project's Github under "Codebook".

2.1 Data Cleaning

A clean dataset, free from errors and inconsistencies, provides a solid foundation for analysis. To begin the data cleaning process, various inconsistencies were addressed to ensure the data's reliability and usability. Income data, which was originally recorded in a mix of formats, underwent standardization. Income values were converted to numeric forms, and an outlier value of "760" was adjusted to "76000", assuming it to be a mistyped value. The column

detailing the mother's daily cigarette consumption, `mom_numcig`, had several ambiguous entries like "None", ranges such as "20-25", and erroneous data like "2 black and miles a day" which were adjusted to numerical values of 0, 22, and 2 respectively. An improbable value of "44989" within the same column was marked as NA.

Further cleaning included the binary conversion of certain columns, specifically those related to child ETS exposure. These columns were transformed from textual formats "1=Yes" and "2=No" to a binary system: 1 for "Yes" and 0 for "No". When examining the SWAN scores, entries with scores of 0 in both hyperactive and inattentive tests were marked as NA, based on patterns observed with other missing data. Finally, to simplify the dataset binary race indicator variables were merged into singular columns, `p_race`, and `t_race`.

2.2 Missing Data

Variables with More Than 25% Missing Data

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

From the table above, one notable limitation of the dataset is the prevalence of missing data. Several columns possess more than 25% missing entries with all variables missing at least a few entries. This is a substantial concern given the sample size of only 49. Furthermore, the pattern of missing data across certain variables suggests the missingness is not at random, caused by

unobserved factors, such as the method of data collection or the population's characteristics. One example is missingess about variables concerning a child's substance use which could be caused by response bias.

Additionally, `mom_smoke_pp1`, indicating if the mother smoked postpartum, has 79.59% missing data. While not as extreme, this is consistent with other smoking-related columns and raises concerns about reliability. Considering the dataset's small sample size, imputation is not recommended. As seen in Figure 1, missing data is particularly concerning in columns on smoking exposure, children's autism spectrum, and parental management. Given the substantial proportion of missing data, imputation could introduce more bias, making the results less reliable.

3. Preliminary Analysis

In this section, we delve into organizing all of the 78 variables within the dataset. These variables have been systematically grouped into four broad categories: outcomes, exposure, demographic, and Not of Interest.

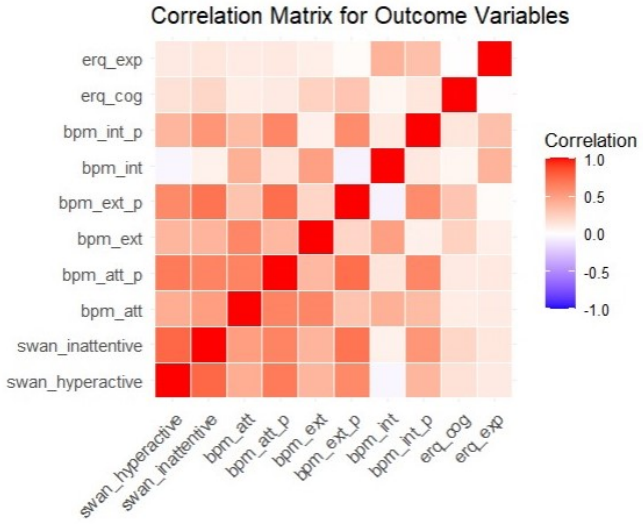
Outcomes capture variables related to Self-regulation, Externalizing, and Substance Use. This includes parent and child responses to the Brief Problem Monitor (BPM), Emotional Regulation Questionnaire (ERQ), Strengths and Weaknesses of Attention-Deficit/Hyperactivity Disorder Symptoms and Normal Behavior Scale (SWAN), as well as Child Substance use variables.

Exposure contains all the variables directly related to SDP and ETS. This included prenatal and postnatal exposure as well as Urine cotinine amounts.

Demographics includes all variables related to parent and child backgrounds. This included variables related to income, race, education, ethnicity, language spoken, and employment status. This is crucial for identifying potential confounding factors during regression.

Variables Not of Interest all seem to have limited direct relevance to the main objectives of the study. They include the Parental Monitoring Questionnaire (excluded due to its lack of connection to the study outcomes for self-regulation) and variables concerning parental substance use in the past six months (excluded for their lack of connection to ETS and SDP variables which were recorded in the latest 5 years postpartum).

3.1 Outcome Variables



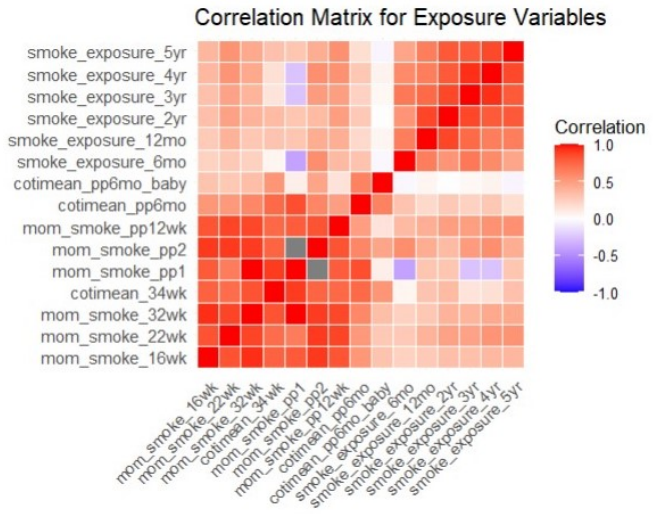
As seen in Figure 2 above, there is a strong correlation among the SWAN and BPM variables related to Attention (att) and Externalizing (ext). After exploring the questionnaires these variables seemed to all be related to a child's externalizing behavior tendencies. Meanwhile, although variables for BPM Internalizing (int) and ERQ data are not as tightly correlated, the questionnaires all focus on a child's ability to emotionally regulate. These two subgroups of externalizing and self-regulation will be used when creating our composite variables for these metrics.

Outcome Variables Summary Statistics

Variable	Mean	SD	Min	Max	Missing %
swan_hyperactive	7.79	6.38	0.00	20.0	20
swan_inattentive	10.92	5.44	1.00	22.0	20
bpm_att	3.00	2.62	0.00	9.0	24
bpm_att_p	2.06	2.22	0.00	8.0	27
bpm_ext	2.81	2.01	0.00	8.0	24
bpm_ext_p	1.68	2.50	0.00	11.0	24
bpm_int	2.71	2.73	0.00	12.0	29
bpm_int_p	2.21	2.48	0.00	9.0	20
erq_cog	3.19	0.97	1.00	5.0	27
erq_exp	2.75	0.80	1.25	4.5	27

Above are the summary statistics for the outcome variables, an immediate observation from the table is the absence of extreme outliers. However, there's one red flag concerning data quality—each of these variables has roughly 20-30% missing data. As stated earlier the significant missing proportion may introduce biases and impact the analysis, warranting further scrutiny of any findings.

3.2 Exposure Variables



From the above Figure, we can make out the strong inter-correlation between the prenatal and postnatal smoking exposure variables and cotinine measures. One outlier however is the variable `mom_smoke_pp1`, which showcases a strange negative correlation with some exposure variables, this is most likely due to substantial missing data of about 80% from the column. Additionally, `cotimean_pp6mo_baby`'s weak correlation with other ETS variables suggests it could be capturing different dimensions or aspects of exposure or could be influenced by external factors.

The most alarming finding, however, is the correlation patterns observed within postpartum ETS exposure surveys from studies one and two. There seems to be a clear disjunction between the two studies, likely because parents, in the second study, are trying to recall exposures almost a decade ago. This could introduce a recall bias, which is a grave concern, as this can skew our analysis.

ETS/SDP Exposure Summary Statistics

Variable	Yes	No	Missing %
mom_smoke_16wk	36	12	2
mom_smoke_22wk	29	13	14
mom_smoke_32wk	30	10	18
mom_smoke_pp1	7	3	80
mom_smoke_pp2	22	7	41
mom_smoke_pp12wk	30	12	14
smoke_exposure_6mo	29	10	20
smoke_exposure_12mo	30	9	20
smoke_exposure_2yr	28	11	20
smoke_exposure_3yr	27	11	22
smoke_exposure_4yr	28	10	22
smoke_exposure_5yr	29	10	20

Cotinine Variable Summary Statistics

Variable	Mean	SD	Min	Max	Missing %
cotimean_34wk	49.75	97.69	0	382.77	22
cotimean_pp6mo_baby	4.04	7.62	0	41.68	22
cotimean_pp6mo	100.49	179.34	0	878.96	22

Upon inspection of the summary statistics for exposure variables, everything seems standard, except for the maximum values for the Cotinine-related variables. A deeper dive into these variables revealed a single extreme outlier, seen in the maximum value for each variable. This child's Cotinine values both ETS and SDP were considerably higher than the rest—almost double the second highest. Notably, the same child also had missing data for various pivotal variables concerning outcomes and demographics, as such they were removed from the dataset.

3.3 Demographic Variables

A thorough examination of demographic variables was conducted, in the interest of brevity and to align with the report's focus, here are the most salient findings. The study's parents

had a median age of 37, with an IQR of 35 to 39, and 98% (40 individuals) belonged were of female sex. While 37% (15 individuals) identified with a speaking another language at home, the racial distribution was dominated by White parents at 61%, with NHPI and Other both at 15%. Employment varied, with 54% holding status '2' for Full-Time and parental education spread across categories, with 37% at '3' for some college, additionally the median household income stood at $48,424. Teenagers aged 12 to 15 were almost equally represented, though about 5% were 16 years old. For teens, 30% identified with speaking another language at home and 42% identified with being Hispanic/Latino, while the race distribution saw White teens at 39%, Biracial/Multiracial at 22%, and Black at 19%. Overall demographic data presented no large issues, however one note is that all variables contained missing data of about 12-25%.

4. Composite Variable Creation

To simplify the large amount of variables within the dataset, composite variables were developed. These variables amalgamate several related metrics, allowing for an easier analysis. Drawing from the previously discussed groupings in Section 3, composite variables were established for Exposures: SDP and ETS, as well as Outcomes: Self Regulation (SR), Externalizing (EXT), and Substance Use (SU).

Since many of the original variables were a mix of binary and continuous data, various strategies were used to standardize the different metrics. To begin a log transformation was applied to all variables related to Cotinine levels, this was done to reduce the positive skew of the data as many of the values were close to 0 with others being in the several hundred. After this, these values underwent Z-score normalization and then were all made positive with the lowest Z-score given a value of 0. This was done since the next step involved aggregating these new standardized values with the binary data concerning gestational and postpartum smoke exposure variables, with values of 0 signifying no exposure. After taking the average across these exposure values the composite variables for ETS and SDP were complete.

The outcome composite variables follow a similar suit, utilizing Z-score normalization of BPM, ERQ, and SWAN related variables to standardize the different scoring systems used. As said in Section 3, these values were then grouped into EXT and SR to create the composite externalizing and self-regulation metrics. Finally, a composite variable for Substance Use (SU) was generated based on whether the child had ever used marijuana, alcohol, cigarettes, or e-cigarettes before.

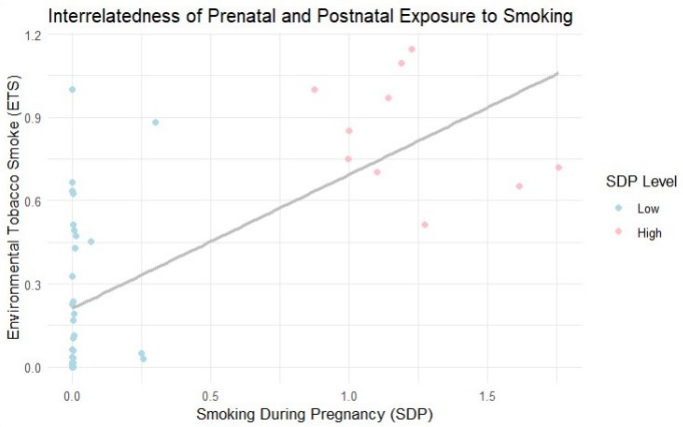
4.1 Composite Variables: Summary Statistics

Composite Variables Summary Statistics

Variable	Mean	SD	Min	Max	Missing %
SDP	0.28	0.51	0.00	1.76	2
ETS	0.35	0.37	0.00	1.15	2
SR	-0.01	0.67	-1.72	1.33	15
EXT	0.00	0.80	-1.11	2.03	15
SU	0.19	0.40	0.00	1.00	25

Upon initial observation, the new composite variables were successfully standardized with no extreme outliers. One note however is the striking disparity in missing data percentages between the exposure and outcome groups. While the exposure variables (SDP and ETS) exhibit minimal missing data (both at 2%), the outcome variables present considerably more, ranging from 15% to 25%. This was mainly caused by missing values for one behavioral test being linked to that Child also having missing data for the other behavioral tests. This could certainly impact analyses and should be considered when discussing the limitations of this study.

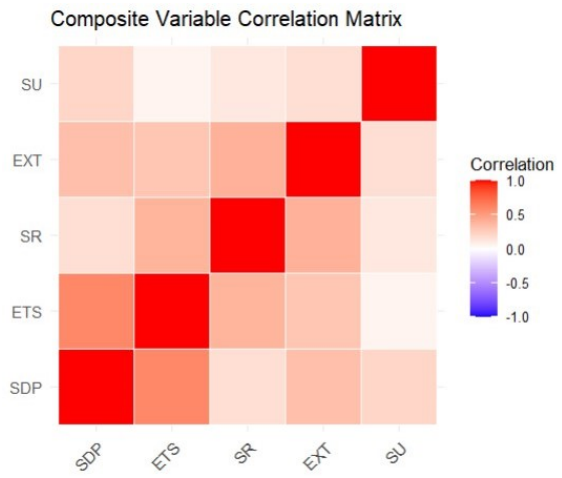
4.2 Interrelation of SDP/ETS



The figure above illustrates the interrelation between SDP and ETS. Interestingly, there's a discernible bimodal distribution within childrens SDP exposure, suggesting two subgroups - low and high. For visualization purposes these groups will be separated out in following plots. Notably, high levels of SDP exposure appear to coincide with heightened ETS levels. In contrast, low SDP levels don't exhibit a strong association with ETS. This suggests that

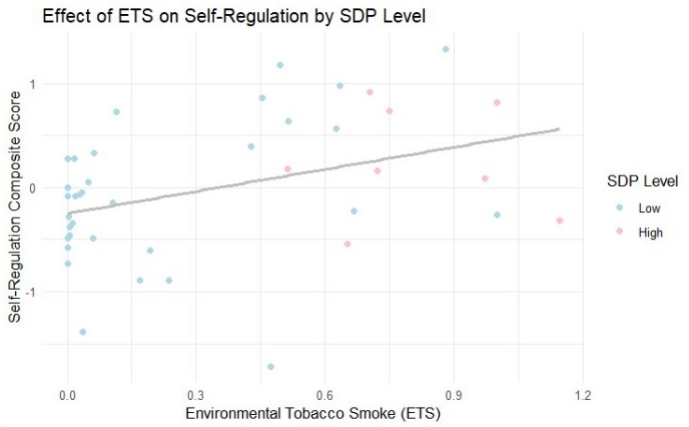
mothers who smoke during pregnancy (SDP) are potentially more likely to also expose their children to secondhand smoke (ETS) postnatally, whereas mothers who refrain from smoking during pregnancy do not necessarily guarantee an ETS-free environment for their children post-birth.

5. Interrelation Between Exposure and Outcomes



The Figure above illustrates the correlation matrix between the exposure and outcome composite variables. From this, we can see the previously discussed strong association between ETS and SDP. Additionally, while both exposure variables are connected to EXT and ER, the correlation with ETS is much stronger. Interestingly, SU does not share a strong correlation with exposure or outcome data, this could be attributed to the 25% missing data and potential issues stemming from response bias.

5.1 Effect of ETS and SDP on Self-Regulation

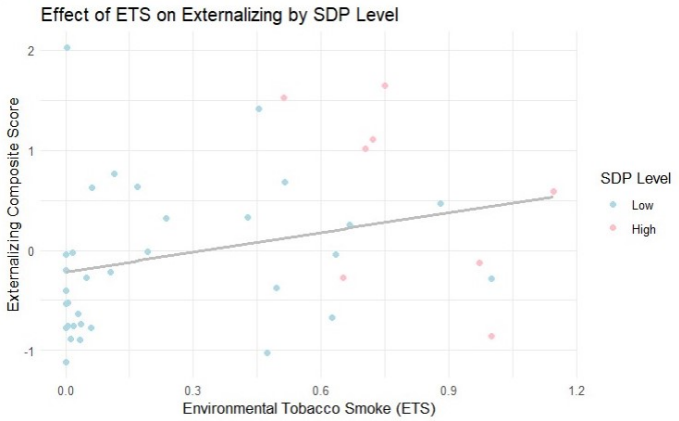


The visualization above highlights the relationship that ETS exposure increases SR also increases signaling worse child self-regulation capabilities. This relation is similar between SDP and SR, albeit with a milder trend.

A rudimentary linear regression model was then created to examine the effects of ETS and SDP on self-regulation. It's vital to note that this model's variable selection was rudimentary. The initial model encompassed all demographic variables, which were then pruned based on their significance. Models were juxtaposed before and after this pruning using ANOVA.

The model was formulated as: `SR ~ SDP + ETS + page + pethnic + pwhite + prace_other + tage + language + tethnic + taian` with covariates being selected from parent and child demographic data. Notably, ETS was found to be statistically significant with a p-value of 0.041 and an estimated coefficient of 0.710, with a standard error of 0.329 meaning controlling for all other factors for every point increase in ETS we would expect a .710 point increase in child SR levels. Important to note is that this model has an Adjusted R-squared: 0.3734, indicating that there remains a significant proportion of variance in SR that is not accounted for. Lastly, pethnic and prace_other were also found to be statistically significant in this model.

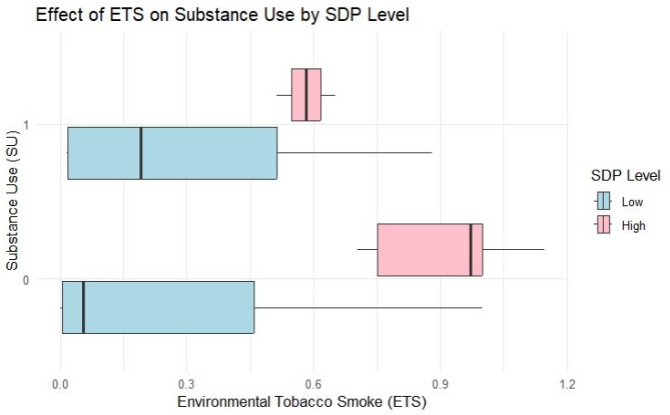
5.2 Effect of ETS and SDP on Externalizing Behavior



Similar to the effects of ETS on Self-regulation, as a child's ETS score increases so does the degree of externalizing behaviors. The relationship of SDP to externalizing behaviors, on the other hand, seems relatively indistinct, although the previous correlation analysis did indicate a mild positive relationship.

Applying a linear regression model, using a similar variable selection approach used in the self-regulation model, was formulated as: `EXT ~ SDP + ETS + pethnic + tage + language + taian + tblack`. Interestingly, this model did not achieve statistical significance, evidenced by an Adjusted R-squared value of 0.1539. One explanation for this might be the weak correlation between BPM Internalizing and ERQ variables, potentially rendering the composite Externalizing variable less reliable in this context.

5.3 Effect of ETS and SDP on Substance Use



As seen in the figure above for children exhibiting low SDP levels, there's a noticeable pattern where increased substance use aligns with heightened ETS exposure. Surprisingly, within the

high SDP level bracket, however, an increase in SU seems to be paired with diminished ETS. This divergence could potentially be attributed to the constrained sample size, with only 2 cases of children at a high SDP level registering an SU of 1, and only 5 cases at an SU of 0.

Utilizing a logistic regression model, echoing the variable selection methodology from previous models, was formulated as: `SU ~ SDP + ETS + pethnic + employ + pedu + income + tage + language + tethnic`. Interestingly, this exploration found neither SDP nor ETS to be statistically relevant. The sole significant determinant emerged as tage, bearing a p-value of 0.0468. Additionally, deviance analysis presented a null deviance of 32.055 (with 28 degrees of freedom) and a residual deviance of 16.307 (with 19 degrees of freedom). The reduction in deviance indicates the model with predictors is a better fit than the null model, but it still has an unexplained variance, pointing to potential missing factors or complexities.

6. Limitations

During this analysis, several potential limitations to using this dataset were discovered. Firstly, a striking concern arose from the substantial missing data in crucial columns, which could potentially distort findings. Given the modest sample size of only 49 participants, it's debatable how effectively observations can be generalized. With several notable variables such as child substance use and postnatal smoke exposure having large amounts of data missing, the robustness of any conclusions made will certainly be effected.

Another challenge was found in the form of recall biases; this was particularly evident from the disparities between postpartum ETS exposure surveys across two distinct studies. In the latter study, parents were tasked with recalling exposures from nearly a decade ago, possibly compromising the accuracy of their responses. Lastly, the regression models, despite shedding light on significant relationships, were rooted in a rather basic variable selection process. This rudimentary approach might have either missed key confounding factors or tailored the models too closely to our specific sample.

7. Conclusion

In this report preliminary exploratory data analysis on the potential effects of SDP and ETS on children's self-regulation, externalizing behavior, and substance use, early findings suggest possible negative implications of both SDP and ETS. ETS, in particular, appeared to have a significant impact on children's self-regulation. Our EDA also indicated a potential relationship between SDP and ETS, suggesting that mothers who smoked during pregnancy might also expose their children to secondhand smoke postnatally. However, these are initial insights, and the connection between these exposures and outcomes requires further examination. This EDA serves as a starting point for guiding more in-depth work in the future.

Appendix

Github

For full Code Appendix go to this Project Github Repo: github.com/Kdossal/2550_Project1 here you will find the project1.qmd file used to generate the report and the code used for my analyses.

Code

```
# Loading in data and Libraries
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.3     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.3     v tibble    3.2.1
v lubridate 1.9.2     v tidyr     1.3.0
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(purrr)
library(broom)
library(ggplot2)
library(knitr)
library(kableExtra)
```

```
Attaching package: 'kableExtra'

The following object is masked from 'package:dplyr':

    group_rows
```

```
library(summarytools)
```

```
Attaching package: 'summarytools'

The following object is masked from 'package:tibble':

    view
```

```
library(reshape2)
```

```
Attaching package: 'reshape2'

The following object is masked from 'package:tidyr':

    smiths
```

```
library(gtsummary)

df <- read.csv("Data/project1.csv")
child_df <- df
```

```
### Cleaning income

# 1. Handle the mistypo
child_df$income[child_df$income == "760"] <- "76000"

# 2. Remove commas
child_df$income <- gsub(",", "", child_df$income)

# 3. Convert empty strings to NA
child_df$income[child_df$income == ""] <- NA

# 4. Convert the character column to numeric
child_df$income <- as.numeric(child_df$income)
```

```
Warning: NAs introduced by coercion
```

```r
### Cleaning mom_numcig

# 1. Replace "None" with "0"
child_df$mom_numcig[child_df$mom_numcig == "None"] <- "0"

# 2. Handle the range "20-25"
child_df$mom_numcig[child_df$mom_numcig == "20-25"] <- "22"

# 3. Convert "2 black and miles a day" to "2"
child_df$mom_numcig[grep("black and miles", child_df$mom_numcig)] <- "2"

# 4. Handle the value "44989"
child_df$mom_numcig[child_df$mom_numcig == "44989"] <- NA

# 5. Convert the character column to numeric after replacing empty strings with NA
child_df$mom_numcig[child_df$mom_numcig == ""] <- NA
child_df$mom_numcig <- as.numeric(child_df$mom_numcig)

### Convert columns to Binary 0/1
columns_to_convert <- c("mom_smoke_16wk", "mom_smoke_22wk",
                        "mom_smoke_32wk","mom_smoke_pp1", "mom_smoke_pp2",
                        "mom_smoke_pp12wk", "mom_smoke_pp6mo")

for (column in columns_to_convert) {
  child_df[[column]][child_df[[column]] == "1=Yes"] <- 1
  child_df[[column]][child_df[[column]] == "2=No"] <- 0

  # Convert column to numeric
  child_df[[column]] <- as.numeric(child_df[[column]])
}

### Fill in 30 days previous columns
sub_ever <- c("cig_ever", "e_cig_ever", "mj_ever", "alc_ever")
sub_30 <- c('num_cigs_30', 'num_e_cigs_30', 'num_mj_30', 'num_alc_30')

# Loop over each substance
for (i in 1:4) {

  # Set the number column to 0 where the ever column is NA or 0
  child_df[is.na(child_df[[sub_ever[i]]]) |
             child_df[[sub_ever[i]]] == 0, sub_30[i]] <- 0
```

```
}

# Clean SWAN variables, when both tests have scores of 0, these should be NA
child_df$swan_hyperactive[with(df, swan_hyperactive == 0 &
                                 swan_inattentive == 0)] <- NA
child_df$swan_inattentive[with(df, swan_hyperactive == 0 &
                                 swan_inattentive == 0)] <- NA

# Replace prefer not to answer with NA
child_df$tethnic[child_df$tethnic == 2] <- NA
```

```
# Cleaning Race column from several binary indicator variables to one prace
# and trace variable (accounts for ind. indicating multiple races by
# grouping them into biracial)

# Change prace columns
child_df <- child_df %>%
  mutate(
    race_count = rowSums(select(.,
                                paian, pasian, pnhpi, pblack, pwhite,
                                prace_other), na.rm = TRUE),
    p_race = case_when(
      race_count > 1 ~ "Biracial/Multiracial",
      paian == 1 ~ "AIAN",
      pasian == 1 ~ "Asian",
      pnhpi == 1 ~ "NHPI",
      pblack == 1 ~ "Black",
      pwhite == 1 ~ "White",
      prace_other == 1 ~ "Other",
      TRUE ~ NA_character_  # default case when all 0
    )
  ) %>%
  select(-race_count) # remove the temporary race_count column

# Change trace columns
child_df <- child_df %>%
  mutate(
    race_count = rowSums(select(.,
                                taian, tasian, tnhpi, tblack, twhite,
                                trace_other), na.rm = TRUE),
    t_race = case_when(
```