Project1

Kerry Ye

2023-10-02

Introduction

In collaboration with Dr. Lauren Micalizzi, this project aims to conduct an Exploratory Data Analysis to investigate the impacts of smoking during pregnancy (SDP) and environmental tobacco smoke (ETS) exposure on children's substance use, externalizing behaviors, and self-regulation. Originating from a smoke avoidance program for low-income pregnant women on reducing both SDP and ETS exposure. A select group of adolescents and their mothers were chosen for further research. Currently, we have the baseline data, with two longitudinal follow-up assessments for 6 and 12 months post-baseline. This collected dataset comprises 49 pairs of adolescents and mothers, with 78 variables from both the current and prior studies.

Pre-processing Data

In the process of data pre-processing, we identified multiple issues within the dataset. Firstly, the *mom_numcig* column, which represents the number of cigarettes a mother smokes per day, contained several incorrect entries. For example, the entry "2 black and miles a day" was corrected to 2. The value "44989", which is implausibly high, was replaced with "NA". The term "None" was converted to the number 0. For the entry "20-25", we used its approximate mean, updating the value to 23. Additionally, in the *income* column, an incorrect entry "250, 000" was corrected by removing the comma, resulting in the integer 250000. For consistency, we converted all entries of "2=No" in the dataset to 0 and "1=Yes" to 1, while empty entries were replaced with "NA". In the *swan_inattentive* and *swan_hyperactive* columns, values of 0 were correctly modified to "NA". Lastly, we changed various categorical variables into the factor type when their entries represented different levels. All columns of character type that contained numeric values were transformed into numeric type.

Missing Data

A significant issue researchers should be aware of is the considerable amount of missing values in this dataset. The presence of missing data can reduce the statistical power of the analysis, making it more challenging to detect significant differences in relationships and potentially leading to less accurate and robust estimations in subsequent modeling. Out of the 78 variables in this dataset, 65 variables have missing values.

Table 1: Top 5 and Bottom 5 Variables by Missingness in Tobacco Dataset

variable	n_miss	pct_miss	variable	n_miss	pct_miss
num_cigs_30	48	97.95918	employ	8	16.326531
num_e_cigs_30	47	95.91837	pedu	8	16.326531
num_mj_30	46	93.87755	mom_smoke_22wk	7	14.285714
num_alc_30	45	91.83673	mom_smoke_pp12wk	7	14.285714
mom_smoke_pp1	39	79.59184	mom_smoke_16wk	1	2.040816

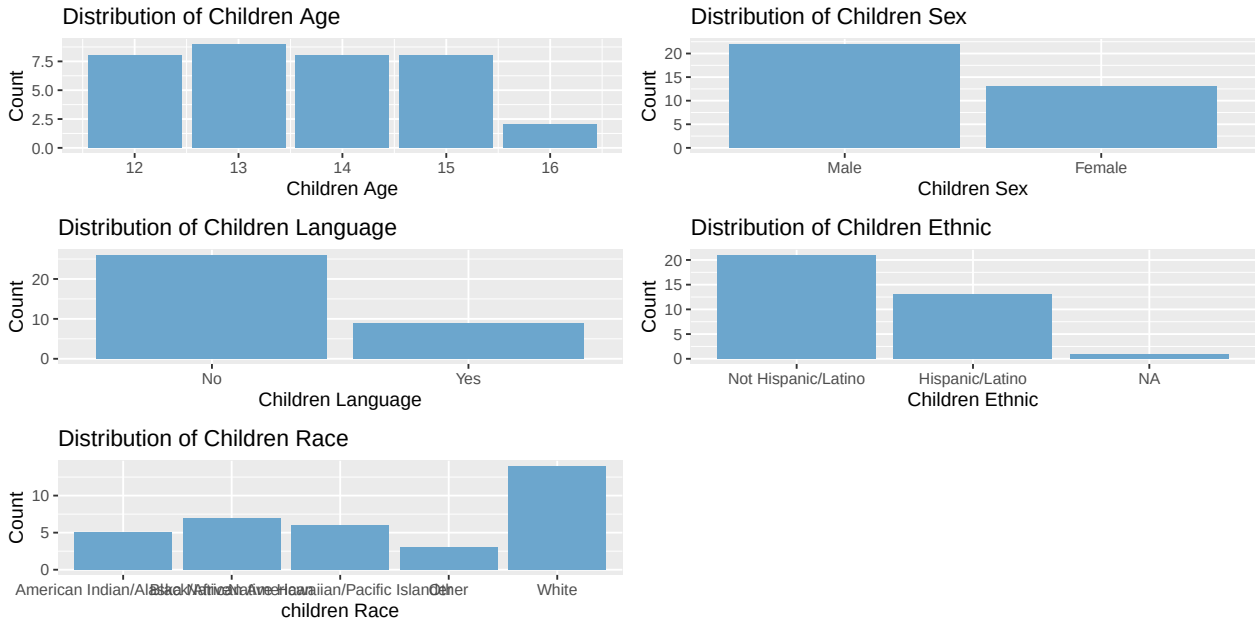
Examining the above "Top 5 and Bottom 5 Variables by Missingness in Tobacco Dataset" table reveals that the five variables with the highest percentages of missingness are *num_cigs_30*, *num_e_cigs_30*, *num_mj_30*, *num_alc_30*, and *mom_smoke_pp1*. Their missingness percentages range from 97.95918% to

79.59184%. Relying on these columns for analysis could skew results given their substantial missing data. Specifically, the *num_cigs_30* variable, which has the highest percentage of missing data at over 97.95918%, has 48 missing values. On the other hand, the variables with the smallest percentages of missingness are *employ*, *pedu*, *mom_smoke_22wk*, *mom_smoke_pp12wk*, and *mom_smoke_16wk*, missingness range from 16.326531% to 2.040816%.

Demographic

One crucial aspect of Exploratory Data Analysis (EDA) is assessing the demographics of the patients in the study, as it helps in determining the generalizability of the study results. If a study only includes a specific demographic group, its results might not be applicable to other groups. In this study, we examine the demographics of both adolescents and parents. We present two graphs that depict the distinct univariate demographic distributions for these two groups.

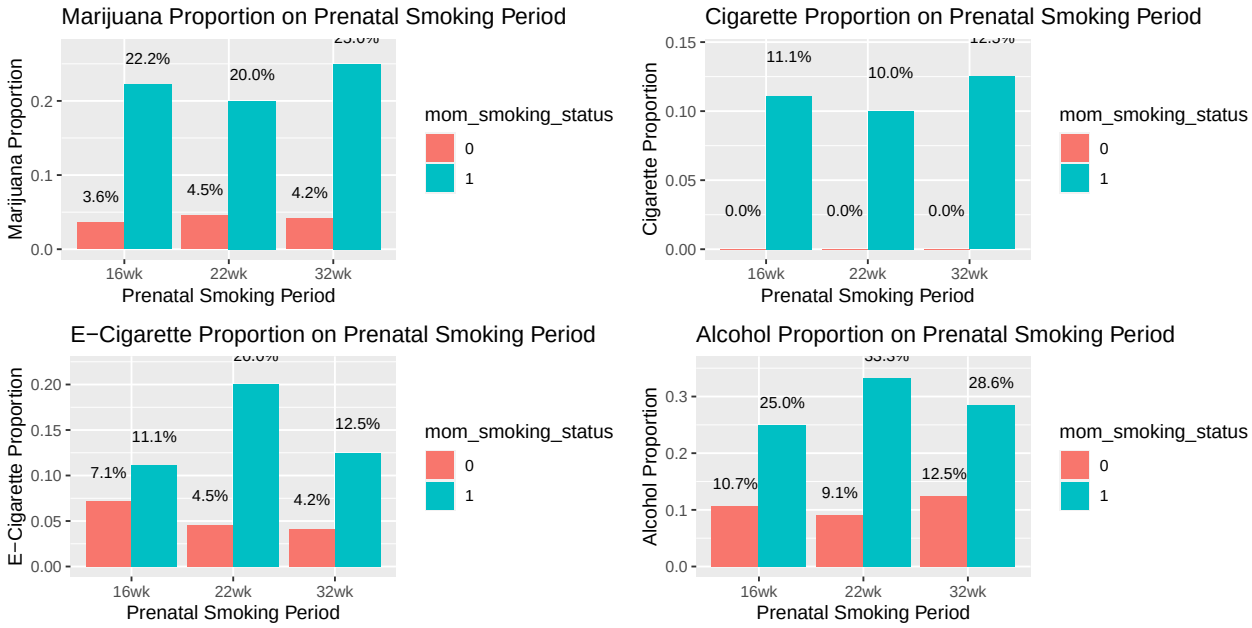
The graph above displays various distributions of demographic variables for parents. The age distribution reveals that the majority of parents in this study fall between the ages of 33 to 39. Fewer are distributed between the ages of 40 to 45, and none are below 33 or exceed 45. The gender distribution indicates that out of 49 parents, 48 are female, with only one male present in the dataset. This lone male entry might have been inadvertently included or could represent an erroneous data entry. The language distribution suggests that most participants speak the same language at home. In terms of ethnicity, a larger proportion identify as not Hispanic or Latino compared to those who identify as Hispanic/Latino. The employment distribution reveals that most parents work full-time. The second-largest group is unemployed, and the smallest group works part-time. Regarding education, the majority of parents have attended some college, followed by those who have completed four years of college, and the fewest have only a high school education. The income distribution highlights that the majority of parents earn between 0 to 100,000, with a concentration around 50,000 and a few earning more than 100,000. In terms of race, most parents are White. Fewer identify as American Indian/Alaska Native, Hawaiian/Pacific Islander, or other, and there are no participants identifying as Asian or African American. Researchers should be mindful of the parents' racial composition in this study, as minority groups are underrepresented. Additionally, attention should be given to the education level, as only a few participants have education levels below college.



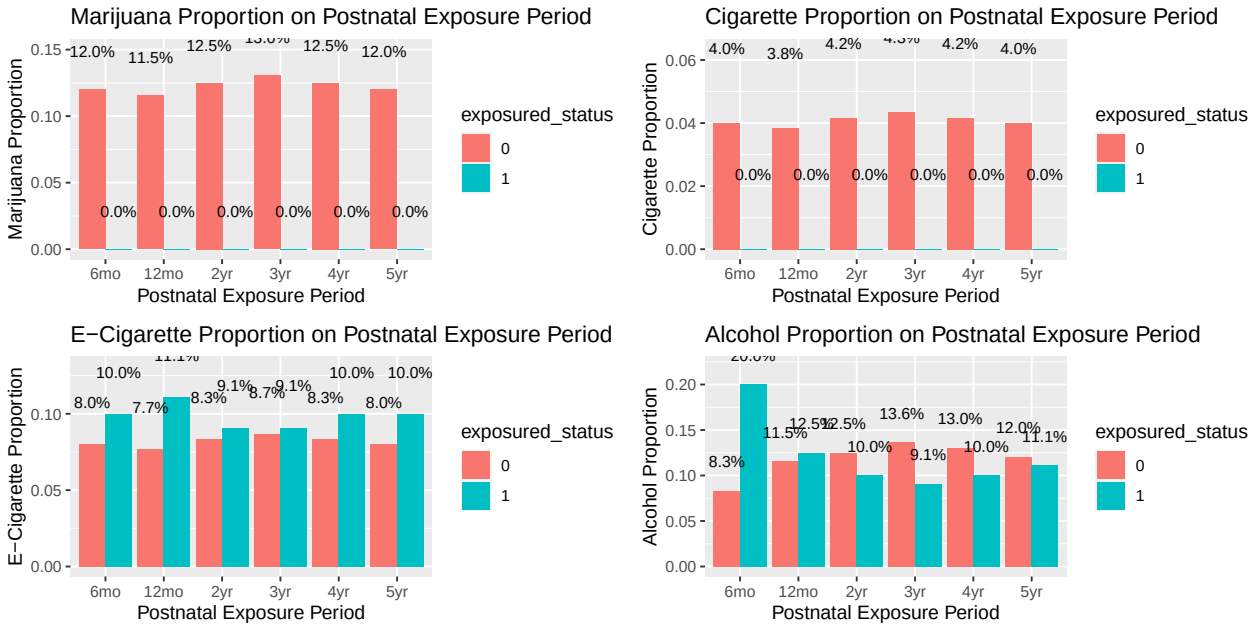
The graph above presents the demographic distributions for children. The age distribution indicates that most children in this study are aged between 12 and 15 years. Few are 16 years old, and none are younger than 12 or older than 16. In terms of gender, there are more male children than females. Regarding language, most children speak the same language at home, with fewer speaking other languages. The ethnicity distribution reveals that the majority are not of Hispanic or Latino descent. Racially, most children are White, with other minorities being less represented. The similarities in language, ethnicity, and race distributions between children and parents suggest that children often inherit or are influenced by their parents' characteristics in these areas.

Prenatal and Postnatal on Substance Use

In this section, we aim to evaluate the effects of Smoke During Pregnancy (SDP) and Environmental Tobacco Smoke (ETS) on adolescent substance use. We first identified the SDP variables as "mom_smoke_16wk", "mom_smoke_22wk", and "mom_smoke_32wk". The ETS variables are identified as "smoke_exposure_6mo", "smoke_exposure_12mo", "smoke_exposure_2yr", "smoke_exposure_3yr", "smoke_exposure_4yr", and "smoke_exposure_5yr". For children's substance use, we selected "cig_ever", "e_cig_ever", "mj_ever", and "alc_ever". We opted to use these categorical variables instead of those recording the number of substance uses in the past 30 days due to the significant amount of missing data.



To explore the impact of Smoking During Pregnancy on children's substance use, we created bar plots depicting the proportion of different substance use across all prenatal smoking periods, from 16 wks to 32wks. In the "Marijuana Proportion on Prenatal Smoking Period" graph, we observe that children whose mothers smoked during pregnancy have a substantially higher proportion of marijuana use than those whose mothers did not smoke. This trend is consistent across all prenatal smoking periods. The percentages for mothers who smoked versus those who did not are as follows: 22.2% vs. 3.6% at 16 weeks, 20.0% vs. 4.5% at 22 weeks, and 23.0% vs. 4.2% at 32 weeks. In the "Cigarette Proportion on Prenatal Smoking Period" graph, children of mothers who smoked during pregnancy exhibit high proportions of cigarette use. However, a comparison with children of non-smoking mothers is not feasible due to missing data in this dataset. The "E-Cigarette Proportion on Prenatal Smoking Period" graph shows that the proportion of e-cigarette use is higher for children of mothers who smoked during pregnancy across all periods. Specifically, the percentages are 11.1% vs. 7.1% at 16 weeks, 20.0% vs. 4.5% at 22 weeks, and 12.5% vs. 4.2% at 32 weeks. Finally, in the "Alcohol Proportion on Prenatal Smoking Period" graph, children of mothers who smoked during pregnancy consistently show a higher proportion of alcohol use than their counterparts. The percentages are 25.0% vs. 10.7% at 16 weeks, 33.3% vs. 9.1% at 22 weeks, and 28.6% vs. 12.5% at 32 weeks. In summary, given that the proportions of children's use of marijuana, e-cigarettes, and alcohol are substantially higher for those whose mothers smoked during pregnancy compared to those whose mothers did not, we can infer that maternal smoking during pregnancy is positively associated with children's substance use. However, due to the absence of data on children's cigarette use whose mothers did not smoke during pregnancy, this conclusion requires more rigorous examination in future statistical analyses.

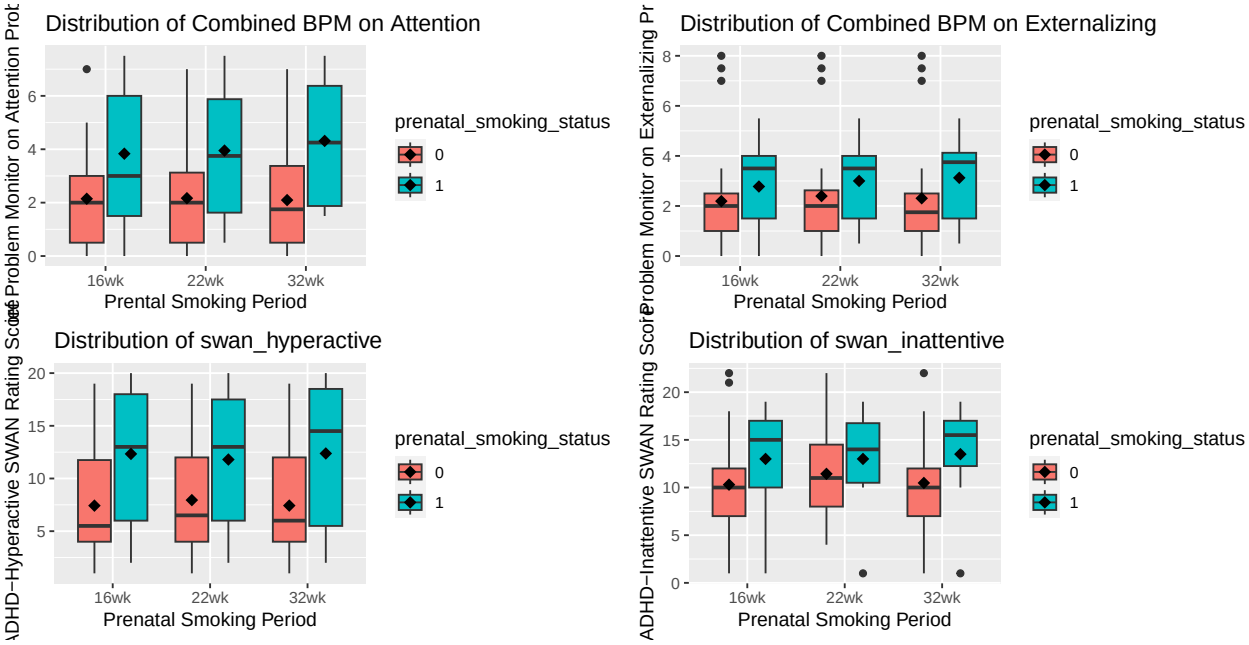


To explore the impact of Environmental Tobacco Smoke (ETS) exposure on children's substance use, we created bar plots depicting the proportion of different substance use across all postnatal exposure periods, from 6 months to 5 years. For both the "Marijuana Proportion based on Postnatal Exposure Period" and the "Cigarette Proportion based on Postnatal Exposure Period" graphs, all data points representing children's marijuana that were exposed to use Environmental Tobacco Smoke are missing. Consequently, we cannot compare the marijuana use proportions between children exposed to maternal smoking during pregnancy and those not exposed across all postnatal periods. In the "E-cigarette Proportion based on Postnatal Exposure Period" graph, children's e-cigarette use proportions during periods of ETS exposure are slightly higher than those during periods without exposure. The "Alcohol Proportion based on Postnatal Exposure Period" graph reveals some interesting patterns: During the 6-month and 12-month exposure periods, children exposed to ETS have a higher alcohol use proportion than those not exposed. However, for the 2-year, 3-year, 4-year, and 5-year exposure periods, children exposed to ETS have a lower alcohol use proportion than those not exposed. In summary, given the missing values in the marijuana and cigarette graphs, the marginally higher e-cigarette proportions among exposed children, and the alternating patterns in alcohol proportions, we cannot conclusively determine any relationship between ETS exposure and children's substance use.

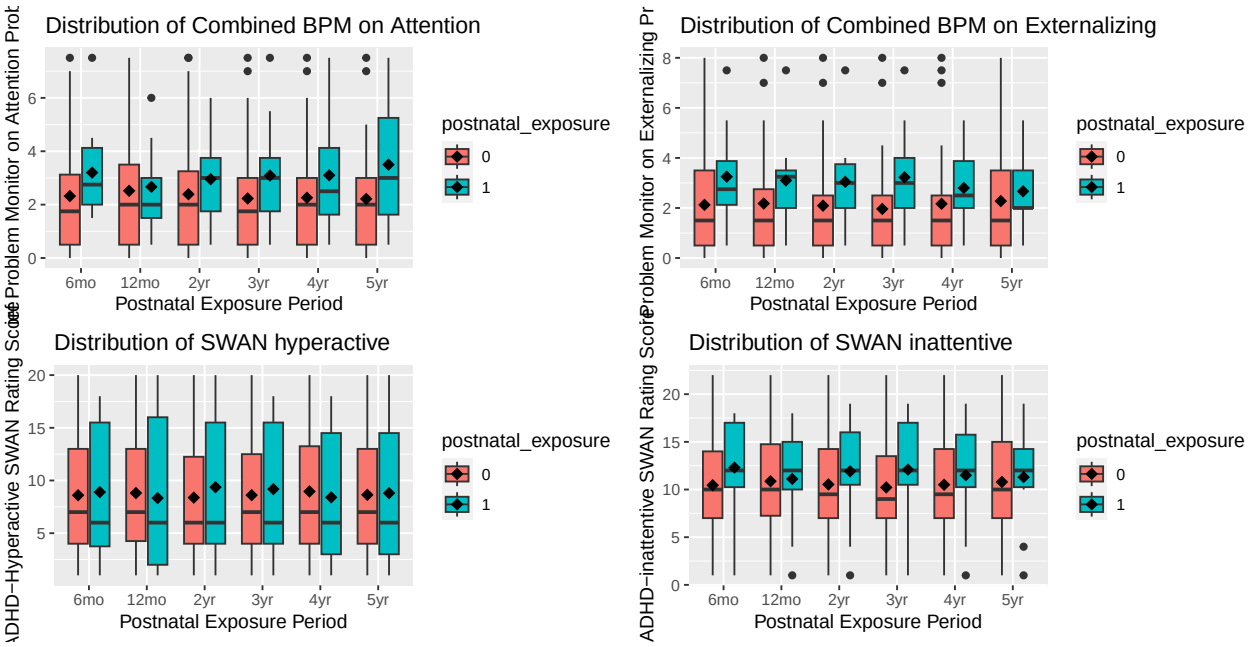
Prenatal and Environmental Tobacco Smoke on Children's Externalizing Behavior

In this section, we aimed to examine the relationship between Smoking During Pregnancy and Environmental Tobacco Smoke on Children's Externalizing Behavior. First, we identified those Externalizing behaviors: bpm_att", "bpm_ext", "bpm_att_p", "bpm_ext_p", "swan_hyperactive" and "swan_inattentive".

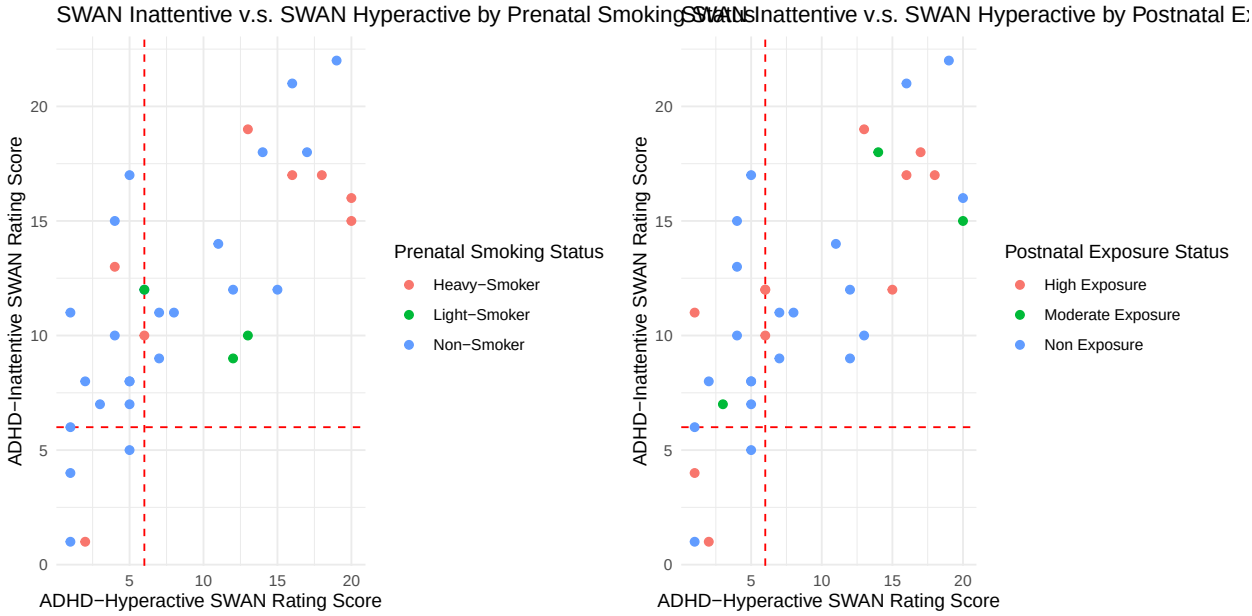
When comparing the bmp scores from parents and children in the dataset, we observed that parents typically report lower scores when assessing their children's attention and externalizing problems than the scores children report for themselves. To account for this discrepancy, we introduced two new variables: **bmp_ext_combined** and **bmp_att_combined**. Given that the dataset includes variables that record the Brief Problem Monitor for Attention Problems and the Brief Problem Monitor for Externalizing Problems from both parents and adolescents, we calculated the average scores of bmp_ext and bmp_att for both children and mothers. These averaged scores were defined as **bmp_ext_combined** and **bmp_att_combined**.



To assess the impact of Smoking During Pregnancy (SDP) on Children's Externalizing Behavior, we used boxplots to depict the distributions of BMP scores for attention and externalizing problems, and SWAN scores for hyperactive and inattentive issues. In the distribution of the combined BMP on attention problems, both the median and mean scores for children whose mothers smoked during pregnancy are noticeably higher than for those whose mothers did not smoke. For the distribution of the combined BMP on externalizing problems, the median score for children of mothers who smoked during pregnancy is visibly higher, although the mean is only slightly elevated compared to those whose mothers did not smoke. In the distribution of the SWAN score for ADHD Hyperactivity, both the median and mean scores for children of smoking mothers are noticeably higher than for those of non-smoking mothers. For the SWAN score on ADHD Inattentiveness, the median for children of smokers is visibly higher, but the mean differs only slightly from children of non-smokers. Overall, the data suggests that SDP is positively associated with children's externalizing behavior. However, given the similar means in the combined BMP for externalizing problems and SWAN for inattentiveness between the two groups, this positive association needs further validation in subsequent studies.



To assess the impact of Environmental Tobacco Smoke (ETS) on Children's Externalizing Behavior, we used boxplots to illustrate the distributions of BMP scores for attention and externalizing problems, as well as SWAN scores for hyperactivity and inattentiveness.In the distribution of the combined BMP for attention problems during postnatal exposure periods, both the median and mean scores for children exposed to environmental smoke were slightly higher than those not exposed, notably during the 6-month, 2-year, 3-year, 4-year, and 5-year periods. However, during the 12-month period, both metrics were similar for both groups. For the distribution of the combined BMP on externalizing problems, both the median and mean scores for children exposed to environmental smoke were somewhat elevated compared to those who weren't, across all postnatal exposure periods. In the distribution of the SWAN score for ADHD Hyperactivity, the median and mean scores for children exposed to environmental smoke alternated in similarity with those not exposed. Similarly, for the SWAN scores on ADHD Inattentiveness, the metrics for children exposed to smoke were intermittently similar to those not exposed. Overall, the data does not indicate a definitive relationship between ETS and children's externalizing behavior, given the similarities in the mean and median scores of BMP for attention and externalizing problems, and SWAN scores for hyperactivity and inattentiveness, between the exposed and non-exposed groups across various postnatal periods.

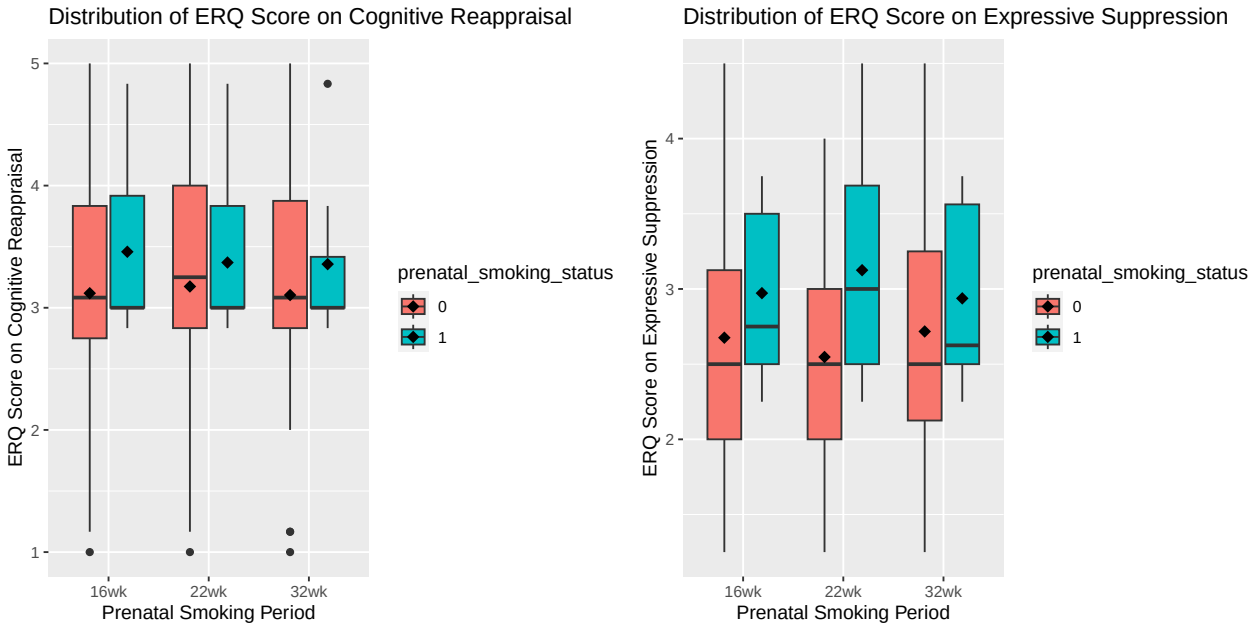


Before delving into the analysis, we introduced two severity variables. The first one, **prenatal_smoker**, is categorized as follows: *Non-Smoker* if the parent did not smoke at any point during the recorded period from 16 weeks to 32 weeks; *Light-Smoker* if their accumulated responses from 16 weeks to 32 weeks fall within the open interval (0,3); *Heavy Smoker* if they self-reported smoking during all three periods." The second one, *postnatal_exposure* is categorized as follows: *Non Exposure* if the child did not exposure to smoke from parents or other people during the recorded period from 6 months to 5 years; *Moderate Exposure* if their accumulated responses from 6 months to 5 years fall within the open interval (0,3); *High Exposure* if the Child exposed more than 3 times during these period.

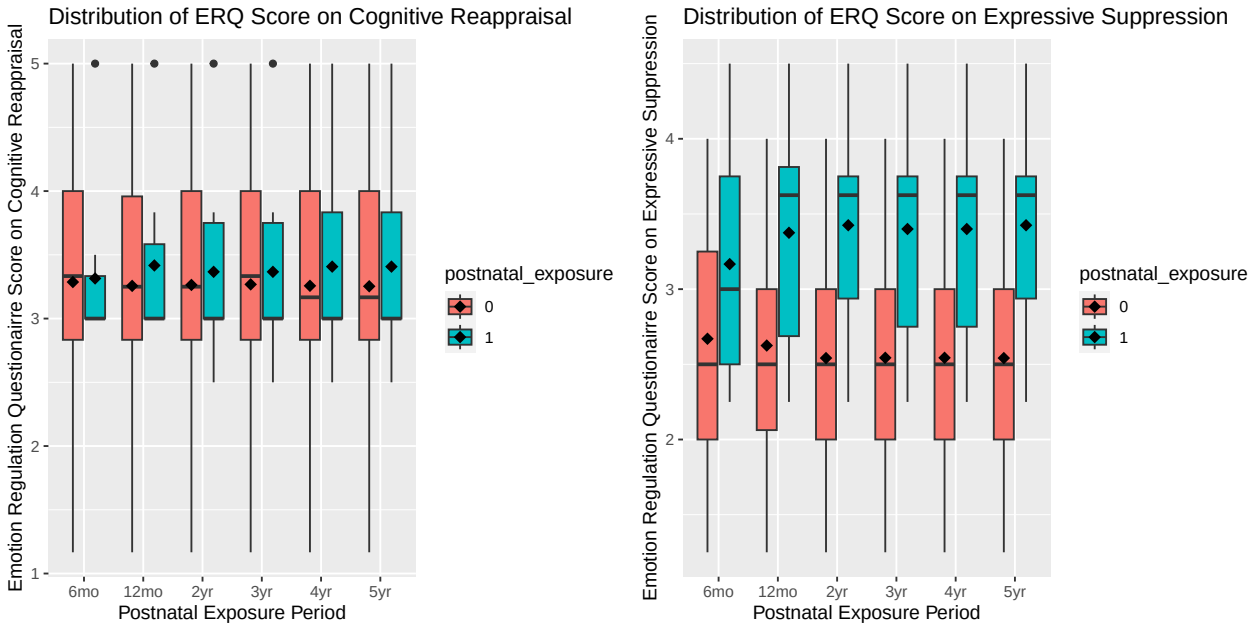
Note that a score of 6 or greater indicates the child is likely ADHD-Hyperactive/Impulsive type, and a score of 6 or greater suggests the child is likely ADHD-Inattentive type. In the first scatterplot during prenatal period, between SWAN scores of ADHD hyperactive and ADHD inattentive, we observe that most of the "Heavy Smokers" cluster above both the Hyperactive and Inattentive score thresholds of 6. All "Light Smokers" cluster at or above the Hyperactive and Inattentive score of 6, while "Non-Smokers" are randomly distributed across all scores. In the second graph during postnatal period, it is challenging to distinguish between "Non Exposure", "Moderate Exposure", and "High Exposure" as the points are distributed across all scores.

Prenatal and Environmental Tobacco Smoke on Children's Self-Regulation

In this section, we aimed to examine the relationship between Smoking During Pregnancy and Environmental Tobacco Smoke on Children's Self-Regulation. We first identified the Self-Regulation Variables as:"erq_cog" and "erq_exp". Just note that a higher ERQ score on Cognitive Reappraisal means the children have better emotional well-being, more positive interpersonal functioning, and greater overall psychological health, while a higher ERQ score on Expressive Suppression linked to poorer emotional well-being, reduced personal relationships quality, and other negative psychological outcomes.



To assess the influence of Smoking During Pregnancy (SDP) on Children's Self-Regulation, we utilized box plots to depict the distribution of ERQ scores for Cognitive Reappraisal and Expressive Suppression. In the distribution for the ERQ score on Cognitive Reappraisal, we observed that the median score for children whose mothers smoked during pregnancy was consistently lower than for those whose mothers did not smoke across all prenatal periods. Conversely, the mean score for children of smoking mothers was consistently higher than for those of non-smoking mothers throughout all prenatal periods. When examining the distribution of the ERQ Score on Expressive Suppression, both the median and mean scores for children of smoking mothers were consistently higher compared to children of non-smoking mothers across all prenatal periods. Given these observations, it is challenging to confirm a relationship between SDP and Children's Self-Regulation based on this graph, especially considering the contrasting patterns in the mean and median scores for Cognitive Reappraisal across prenatal periods.



To assess the influence of Environmental Tobacco Smoke (ETS) on Children's Self-Regulation, we utilized box plots to illustrate the distribution of ERQ scores for Cognitive Reappraisal and Expressive Suppression. In the distribution of the ERQ score on Cognitive Reappraisal, we noted that the median score for children

exposed to environmental smoke was consistently lower than for those not exposed across all postnatal exposure periods. However, the mean score for children exposed to environmental smoke was consistently higher than for those not exposed throughout all postnatal periods. When examining the distribution of the ERQ Score on Expressive Suppression, both the median and mean scores for children exposed to environmental smoke were consistently higher compared to children not exposed across all postnatal periods. Given these findings, it is challenging to establish a definitive relationship between ETS and Children's Self-Regulation based on this graph, particularly given the contrasting trends observed in the mean and median scores for Cognitive Reappraisal across postnatal exposure periods.

Conclusion

On the prenatal side, children whose mothers smoked during pregnancy exhibited higher rates of marijuana, e-cigarette, and alcohol use. This suggests a positive association between maternal smoking and children's substance use. However, the absence of comprehensive data on children's cigarette use requires the need for further rigorous analysis. Additionally, while prenatal smoking seems to influence children's externalizing behaviors, certain measured variables, such as the combined BMP for externalizing problems and SWAN for inattentiveness, show very small differences. These patterns highlight the need for more in-depth studies. On the postnatal side, the data does not conclusively demonstrate a consistent impact of ETS on children's substance use, externalizing behaviors, or self-regulation. This is especially when given the mixed trends in scores across various measured variables and periods, as well as significant missing in the data. The dataset in this study presents several limitations that may impact the robustness and generalizability of the findings. One major limitation is the small sample size. Smaller datasets can reduce the statistical power of tests, making it difficult to detect true effect. This essentially means that even if there is a real effect or association, the limited data might not capture it. Another significant concern is the substantial amount of missingness in the variables. Missing data can introduce bias, especially if the missingness is not random. Researchers should be cautious of these two limitations when using this data for statistical analysis.