Project_1_EDA

Yu Yan

2023-10-05

Report Introduction

In this report, it addresses preliminary exploratory data analysis to Dr. Lauren Micalizzi's project from the Department Behavioral and Social Sciences. The project is about is examining the association between smoking during pregnancy (SDP) and environmental tobacco smoke (ETS) exposure and self-regulation, externalizing behavior, and substance use. The whole project was divided to two parts as a follow-up study, the original study was conducted around in 2017 on about 800 pregnant mother and a new study incorporating a subset of about 100 pairs mother and children from the original study. Our data contains information from both study. This report aims to help answer one of the three primary aims that Dr. Lauren have, in more detail: Examine effects of SDP/ETS on adolescent self-regulation, substance use, and externalizing.

After looking at the data dictionary, the plan of this exploratory data analysis is as follows: after conducting data quality checks, create both timing and intensity variables from the existing columns. The meaning of both timing and intensity columns and how they are created will be explained in more detail in the later section. As the goal is to examine effects of SDP/ETS, those timing and intensity variables might be a method of quantifying the effects brought about by SDP/ETS. The main analysis is to explore the relationship of those timing and intensity variables with the scores from different questionnaires corresponding to each of the three dimensions of adolescent behavior of interest (self-regulation, substance use, and externalizing). We establish our hypothesis as follows: Children with earlier or greater exposure to smoke will demonstrate poorer self-regulation, earlier SU initiation, faster SU escalation, and more EXT than children with later or lighter exposure. As we are all aware of the bad effects of SDP/EXT, we would like to take a step further here trying to understand are there a difference as to earlier vs later, lighter vs intense and potentially quantify how much the differences are if there are any.

Data Dimention

```
## [1] "The Dimension of Data: 49 rows, and 78 columns"
```

```
## [1] "There are 49 Patients"
```

Missing Data and Outlier

We started doing missingness pattern checking of the data. By looking at the 10 most missing columns, we found out one interesting thing that almost all observations (49 observation in total), are lacking four particular columns, num_cigs_30 (missing 48), num_e_cigs_30(missing 47),num_mj_30(missing 46),num_alc_30(missing 45). By referring to the code book, these are the number of days in the last 30 days for teenagers to use cigratte, e_cigrattes, marijuana, and alcohol respectively. As the only four direct

Table 1: Missingness Count and Percentage of columns

	Missing_num	Missing_Pct
num_cigs_30	48	0.9795918
num_e_cigs_30	47	0.9591837
num_mj_30	46	0.9387755
num_alc_30	45	0.9183673
mom_smoke_pp1	39	0.7959184
childasd	28	0.5714286
mom_smoke_pp2	20	0.4081633
pmq_parental_control	16	0.3265306
ppmq_parental_solicitation	15	0.3061224
bpm_int	14	0.2857143

Table 2: Number of Teenagers answered YES Substance Use related questions

	Number
cig_ever	1
e_cig_ever	3
mj_ever	3
alc_ever	5

variables that we have relating to teenagers' substance use behavior, this implies that either the majority of teenagers didn't perform SU related activities, or the data is simply missing those variables.

And as we calculated the number of teenagers that answered YES in those preceding columns, only a small amount of them answered YES. This may impose bias if we were to look into hypothesis of whether early exposure to SDP lead to earlier SU initiation or faster SU escalation. With this initial finding bearing in mind, we keep the analysis forward by identifying outliers.

We identified that in the column of mom_cig, we have two very obvious outliers probably due to bad reporting of data.

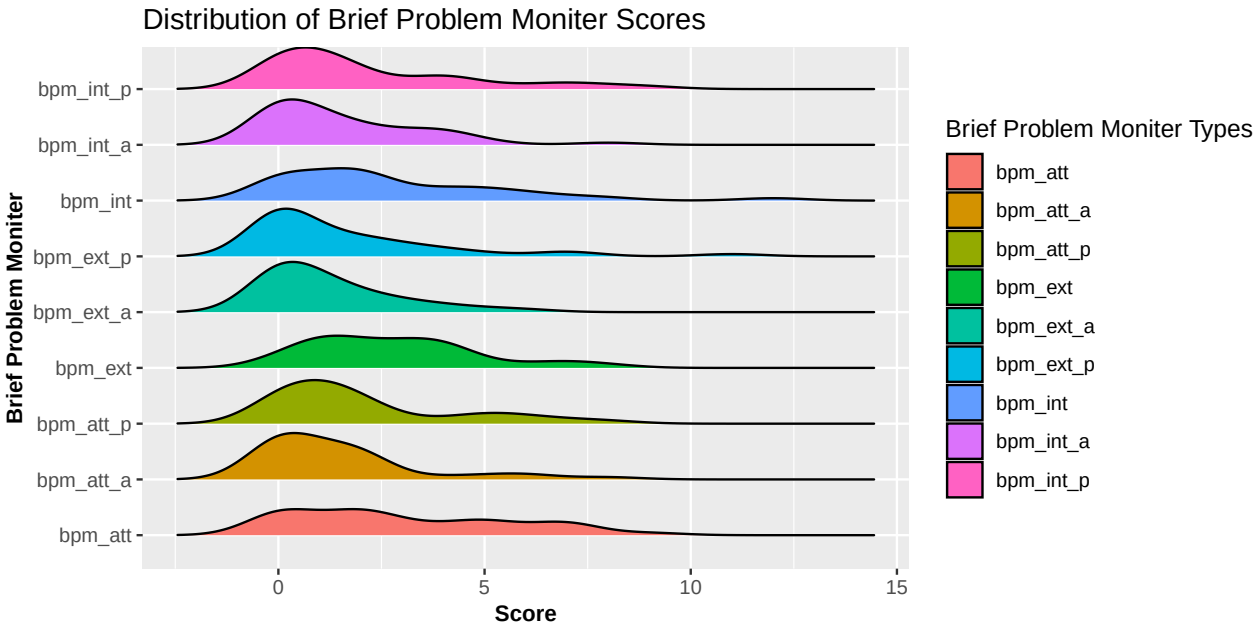
Table 3: Irregular reporting of mom numcig

parent_id	mom_numcig
50102	2 black and miles a day
52702	44989
53802	20-25

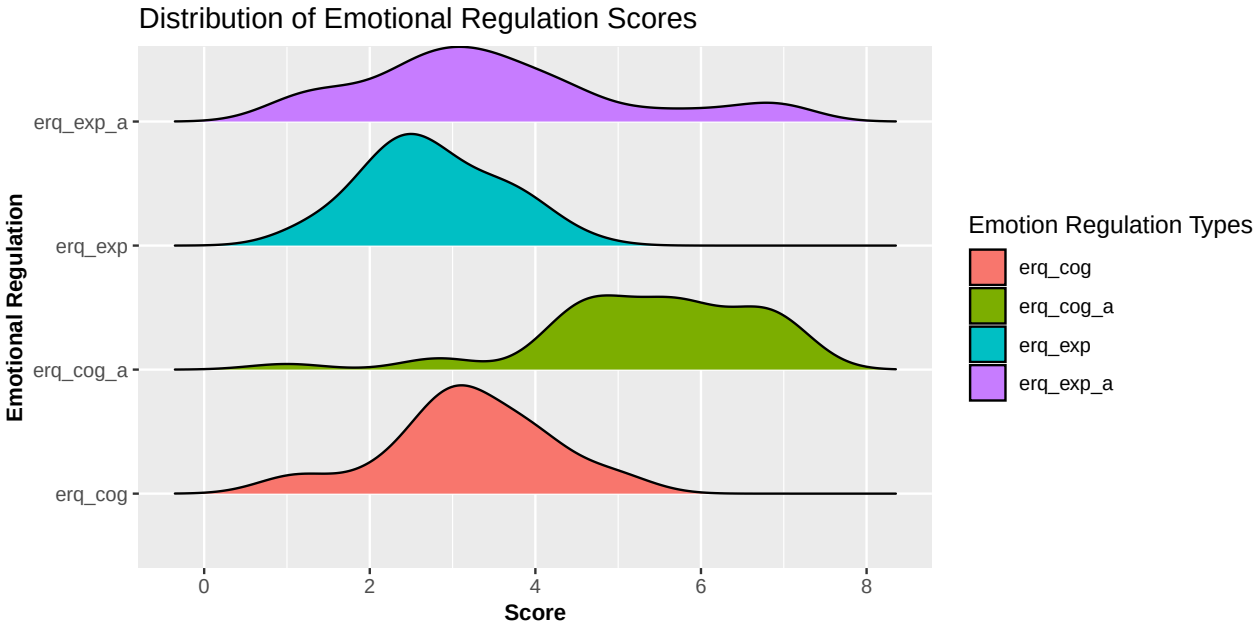
For other continuous variables like composite score from the three types of questionaire (Brief Problem Moniter, Parental Knowledge, and Emotional Regulation), we perform univariate analysis looking at their distribution.

Univariate Analysis

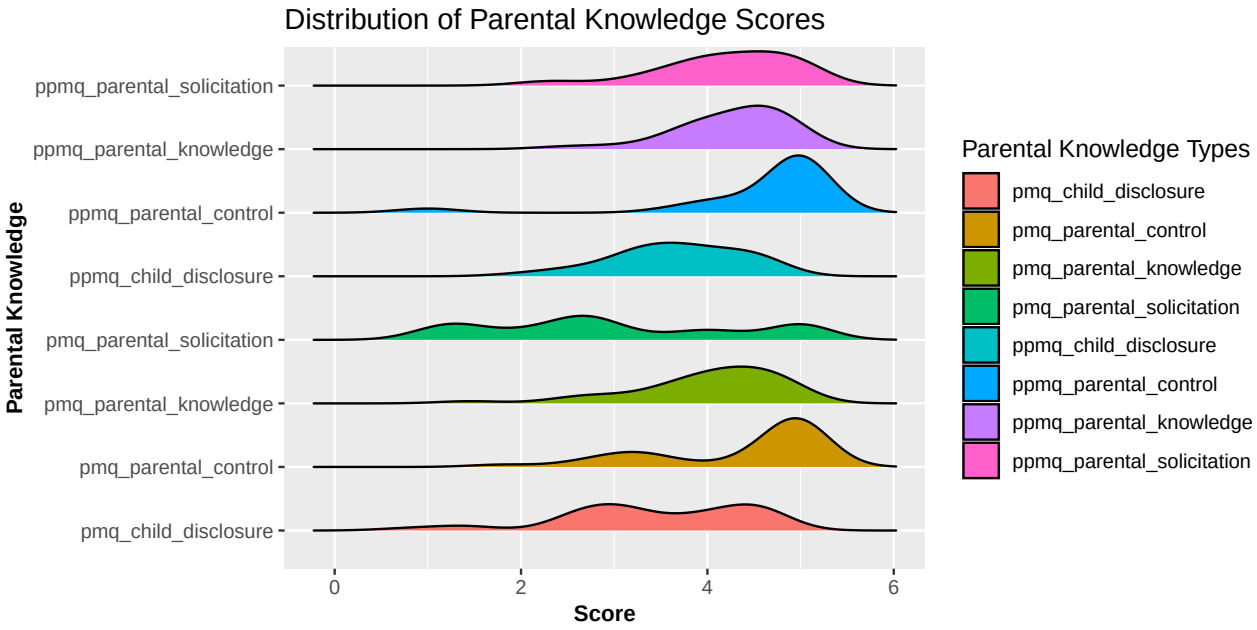
By plotting univariate distribution of all the different bpm scores, we can see that they have approximately similar distribution of left-skewed.



By plotting univariate distribution of all the different emotion regulation scores, we can see that they have approximately similar normal distribution and erq__cog__a is a bit right skewed.



By plotting univariate distribution of all the different bpm scores, we can see that they have approximately similar distribution of right-skewed.



In conclusion, by looking at univariate distribution of each subcategory of the questionaire, we do not observe any irregular patterns of the scores. This ensures the next step of our exploratory analysis.

Data Tranformation

One of the main process of data transformation follows this section. As explained at the analysis plan in the beginning, we are creating both timing and intensity variables to evaluate the effects of SDP/EXT in a quantitative and visualizing way. Based off the variables that we have, we plan to create timing variables in two ways. The first one is a binary variable divided in terms of prenatal vs postnatal. Using this variable we can see how in a big picture exposure first occurred at prenatal period or postnatal would impose effects. For prenatal, we have variables including mom_smoke from 16 weeks pregnant to 32 weeks, as well as lab recording of Urine cotinine in mothers at 34 weeks gestation. So any observation whose first exposure happened in these four period would be given a label of 'prenatal'. For postnatal, we have variables including mom smoke postpartum from visit 1 to 12 weeks, lab recording of Urine cotinine in both mothers and children at 6 month postpartum, and smoke exposure from mom or partner from begin of postpartum til 5 years. As a note, the exposure variables are retrospective of mothers at the new study. So any observation whose first exposure happened in these four period would be given a label of 'postnatal'.

The second way is to look deeper into each of the periods. In addition, as the aim is targeting, we are primarily interested in the effects of SDP and ETS, we could also incorporate this dividing in the second way of creating timing and intensity variables. Basically, for both natal periods, we subdivided their respective first exposure within that period and intensity as well. For intensity, we coded as sum of presense of all the related smoke or exposure variables mentioned above. So for prenatal, we named the timing variable as 'trimester', since we set the levels of 1,2,3 to indicate first,second, and third trimester for which the baby's initial smoke exposure. And intensity as 'trimester_int' for the sum of presense of exposure in the variables. For the lab Urine cotinine values, we set a level of bigger than 0 as a presense of smoke exposure for baby and bigger than 10 for mother as indication of smoker or exposed to second hand smoke.

For postnatal, transformation is quite similar except for the time point of 6 month postpartum. For this time point, we have all three of self-reported smoke, lab Urine cotinine and smoke exposure from mom or partner. Due to the nature of self-reported survey typed questions and its validity problems, when coding for postpartum timing, in the level of 6-momth, we primarily look at the presense from lab Urine cotinine

values and mask the other two variables by it. The postpartum timing variable is named 'postpartum' and inteisity 'postpartum_int'.

Moreover, considering the SDP vs ETS, we can say that prenatal variables (trimester,trimester_int) corresponds to smoke during pregnancy since at thses stages, the mother is still in pregnancy and bay hasn't born yet. For ETS, we can say that postnatal variables (postpartum,postpartum_int) corresponds to environmental tobacco smoke since the baby is currently out of the mother's uterus and any surrounding environmental expsure to tobacco would be considered exposure. This is also why we set the level of baby lab Urine cotinine to be if it is bigger than 0.

Here's the summary of the variables that we generated at this stage.

Table 4: Summary of Prenatal Variable

	0	1	2	4
2	0	0	4	10
3	0	24	0	0

[a] Row is time, column is intensity

Table 5: Summary of Postpartum Variable

	0	1	2	3	4	5	6	7	8
1	0	0	0	2	0	0	1	0	0
2	0	0	0	3	1	0	1	1	1
3	0	0	1	0	0	0	0	1	1
4	0	14	8	0	0	1	1	0	0
5	0	0	0	0	0	2	0	0	0
6	0	0	0	0	1	0	0	0	0

[a] Row is time, column is intensity

From the summary and distribution of the variables that we created, we can see that our data is highly unbalanced. There are a lot of cells that do not have any observations. We may continue for the purpose of conducting exploratory analysis to identify potential patterns, but the current data quality may not be a good choice for performing any statisticl analysis.

Prenatal VS Postnatal

We then compute this table looking at how all the scores differ in terms of prenatal vs postnatal to look for general patterns.

Characteristic	**N**	**Postnatal**, N = 8[1]	**Prenatal**, N = 38[1]	**p-value**[2]
bpm_att	37	2.29	3.17	0.2
NA		1	8	
bpm_ext	37	2.29	2.93	0.4
NA		1	8	
bpm_int	35	2.57	2.75	0.8
NA		1	10	
erq_cog	36	3.64	3.09	0.2
NA		1	9	

erq_exp	36	2.75	2.75	>0.9
NA		1	9	
pmq_parental_knowledge	35	4.11	3.96	0.6
NA		1	10	
pmq_child_disclosure	36	3.63	3.39	0.6
NA		1	9	
pmq_parental_solicitation	35	3.00	2.97	>0.9
NA		1	10	
pmq_parental_control	33	3.87	4.45	0.2
NA		2	11	
erq_cog_a	38	6.07	5.26	0.085
NA		1	7	
erq_exp_a	38	3.25	3.50	0.6
NA		1	7	
bpm_att_p	35	1.86	2.14	0.7
NA		1	10	
bpm_ext_p	36	1.43	1.79	0.7
NA		1	9	
bpm_int_p	38	1.71	2.35	0.4
NA		1	7	
ppmq_parental_knowledge	36	4.17	4.28	0.6
NA		1	9	
ppmq_child_disclosure	36	3.94	3.61	0.2
NA		1	9	
ppmq_parental_solicitation	33	4.31	4.12	0.4
NA		1	12	
ppmq_parental_control	36	4.74	4.53	0.4
NA		1	9	
bpm_att_a	37	1.29	1.50	0.7
NA		1	8	
bpm_ext_a	37	1.86	1.13	0.4
NA		1	8	
bpm_int_a	38	0.86	1.74	0.088
NA		1	7	
swan_hyperactive	38	7.3	8.0	0.8
NA		1	7	
swan_inattentive	38	10.1	11.2	0.7
NA		1	7	

[1] Mean
[2] Welch Two Sample t-test

Sadly we do not obtain any significant P-values and the score differences vary by type: some have larger score prenatal, some larger at postnatal. It does not seem to have any meaningful patterns at this stage, we shall proceed by looking at their respective sub timing and intensity in terms of bivariate comparisons.

Bivariate comparison

Self-Regulation

Based off PPT and resources, Self-regulation means the ability to understand and manage your own behavior and reactions, and it contains behaviors in the following four dimensions: executive function, emotion

Table 7: **Self Regulation Summary for Prenatal(SDP)**

		2			3			NA		
Score	**N**	**2**, N = 4	**4**, N = 10	**p-value**	**N**	**1**, N = 24	**p-value**	**N**	**0**, N = 11	**p-value**
erq_cog	10	3.67	3.36	0.5	19	2.89		7	3.64	
NA		1	3			5			4	
erq_exp	11	3.67	2.94	0.2	18	2.51		7	2.75	
NA		1	2			6			4	
bpm_att	11	2.33	4.88	0.4	19	2.58		7	2.29	
NA		1	2			5			4	
bpm_att_p	11	1.00	3.75	0.041	17	1.59		8	1.75	
NA		1	2			7			3	
swan_hyperactive	11	10.3	12.4	0.6	20	5.9		8	7.0	
NA		1	2			4			3	
swan_inattentive	11	10.3	13.5	0.2	20	10.4		8	9.9	
NA		1	2			4			3	

[1] Mean

[2] Welch Two Sample t-test

Table 8: **Self Regulation Summary for Postnatal(ETS)**

	1				2							3					4						5		6		NA	
Score	**N**	**3**, N = 2	**6**, N = 1	**p-value**	**N**	**3**, N = 3	**4**, N = 1	**6**, N = 1	**7**, N = 1	**8**, N = 1	**p-value**	**N**	**2**, N = 1	**7**, N = 1	**8**, N = 1	**p-value**	**N**	**1**, N = 14	**2**, N = 8	**5**, N = 1	**6**, N = 1	**p-value**	**N**	**5**, N = 2	**p-value**	**N**	**4**, N = 1	**p-value**	**N**	**0**, N = 9	**p-value**
erq_cog	2	4.83	3.00	>0.9	4	2.83	3.00	NA	3.00	3.00	0.4	2	NA	3.83	3.00	0.3	21	2.92	2.89	3.00	5.00	0.5	2	3.67		1	2.5000		4	3.83	
NA		1	0			2	0	1	0	0			1	0	0			1	2	0	0								5	
erq_exp	2	2.25	2.75	>0.9	5	2.50	2.50	3.75	2.50	3.50	0.4	2	NA	3.75	4.50	0.3	20	2.69	2.38	NA	4.00	0.2	2	3.00		1	3.5000		4	2.06	
NA		1	0			2	0	0	0	0			1	0	0			1	2	1	0								5	
bpm_att	2	9.00	7.00	>0.9	5	2.00	7.00	2.00	2.00	5.00	0.4	2	NA	5.00	7.00	0.3	21	2.31	3.00	3.00	1.00	0.9	2	1.50		1	2.0000		4	2.00	
NA		1	0			2	0	0	0	0			1	0	0			1	2	0	0								5	
bpm_att_p	2	6.00	5.00	>0.9	5	1.00	8.00	1.00	2.00	1.00	0.4	2	NA	6.00	2.00	0.3	18	1.50	2.40	NA	2.00	0.2	2	1.0000		1	4.0000		6	0.67	
NA		1	0			2	0	0	0	0			1	0	0			2	3	1	0								3	
swan_hyperactive	2	20.0000	20.0000		5	4.0	18.0	2.0	16.0	6.0	0.4	3	4.00	13.00	6.00	0.4	20	5.7	13.2	17.0	6.0	0.036	2	1.0000		1	15.0000		6	2.50	
NA		1	0			2	0	0	0	0								1	3	0	0								3	
swan_inattentive	2	16.00	15.00	>0.9	5	13.0	17.0	1.0	17.0	10.0	0.4	3	15.00	19.00	12.00	0.4	20	8.9	15.0	18.0	12.0	0.069	2	7.50		1	12.0000		6	7.17	
NA		1	0			2	0	0	0	0								1	3	0	0								3	

[1] Mean

[2] Wilcoxon rank sum exact test; Wilcoxon rank sum test

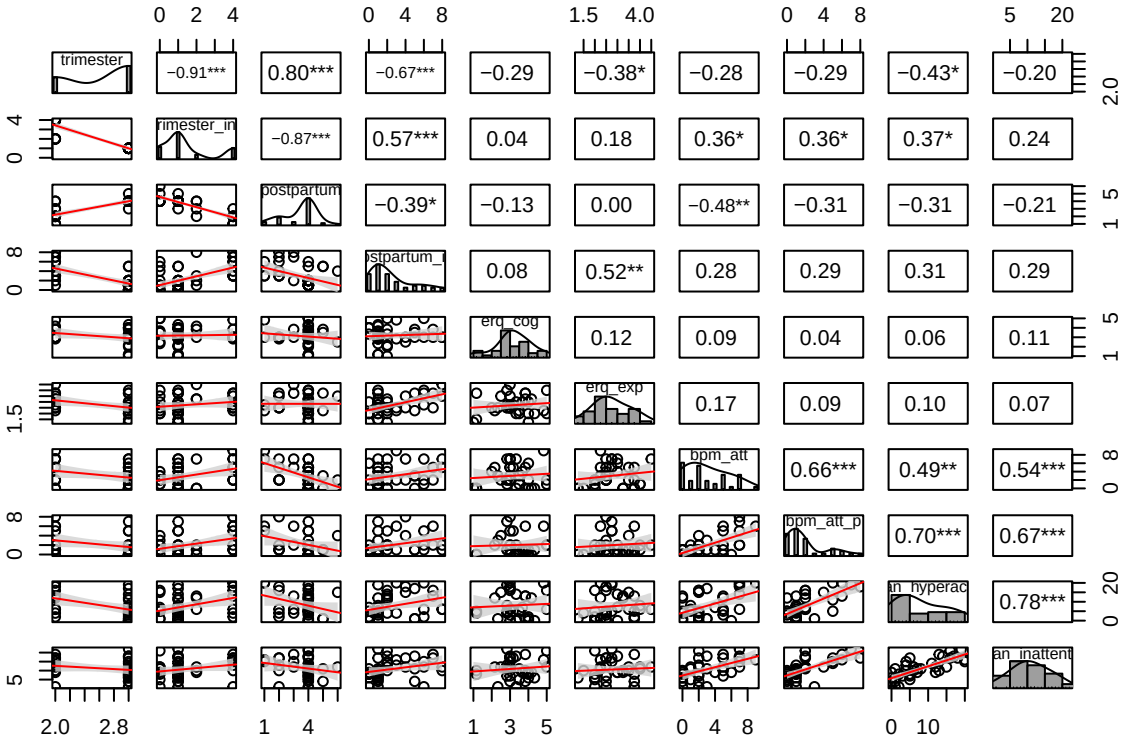
[3] Kruskal-Wallis rank sum test

regulation, effortful control, vagal tone. Given what we have in the data set, we decided that the following variables are related to adolescent self-regulation issues: emotional regulation scores of the adolescent themselves; attention-relating scores on the Brief Problem Monitor questionnaire of both the adolescent own answering, and parents evaluation on their child; lastly, we think the SWAN response would also be relevant as the conditions of ADHD reasonably indicate that the child has self-regulation problems such as easily get distracted. Here's the finding that we have:

The following two table (TABLE 7,8) displays the summary of the selected scores stratified by timing and intensity, and each period (prenatal and postnatal) has one table. The first row of grouping indicate timing and the second indicate intensity. This table is intended to show two things: 1. under the same first exposure timing, what the score would change as intensity changes; 2. for the same intensity level, what would the effects of first exposure timing be.

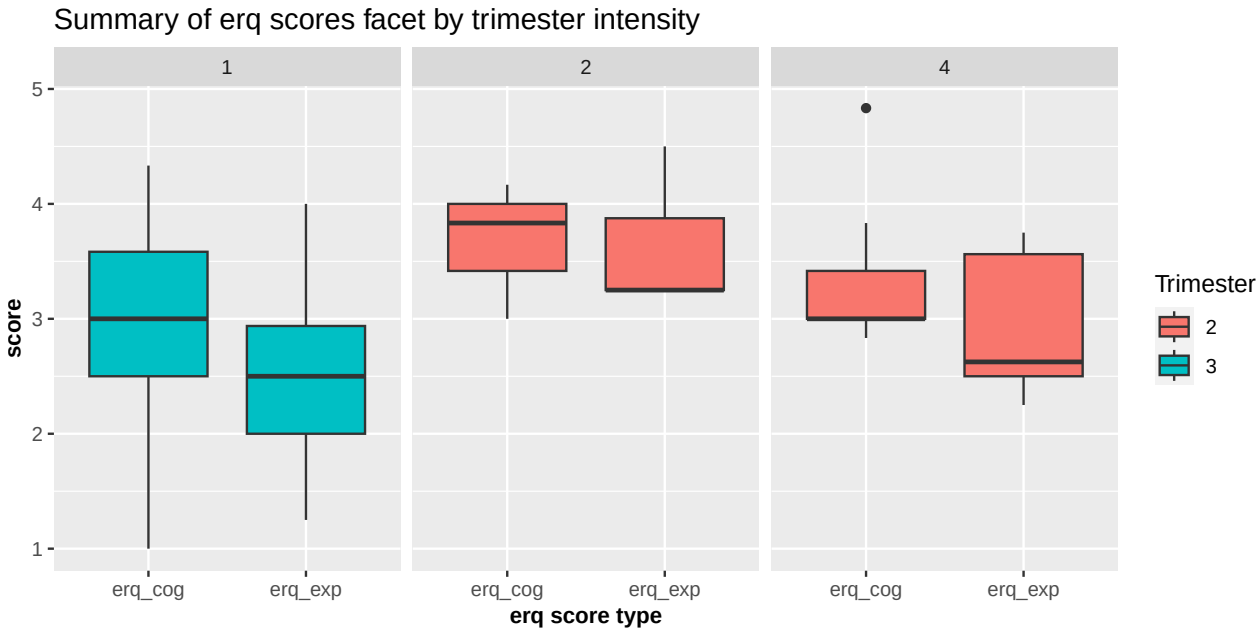
From the two summary table (TABLE 7,8), we can only see that for second trimester, higher intensity is associated with lower erq scores and higher bpm and swan scores. For postnatal, we see different patterns: for postnatal timing 1, higher intensity is associated with lower erq scores while timing 2, higher intensity is associated with slightly higher erq scores. However, we should not rely ourselves too much on such observations since we can also tell from the table that either the P-values associated to almost all comparisons are non significant or some groupings only have 1 observation to give a statistic.

We would also to look at the correalations between those variables.

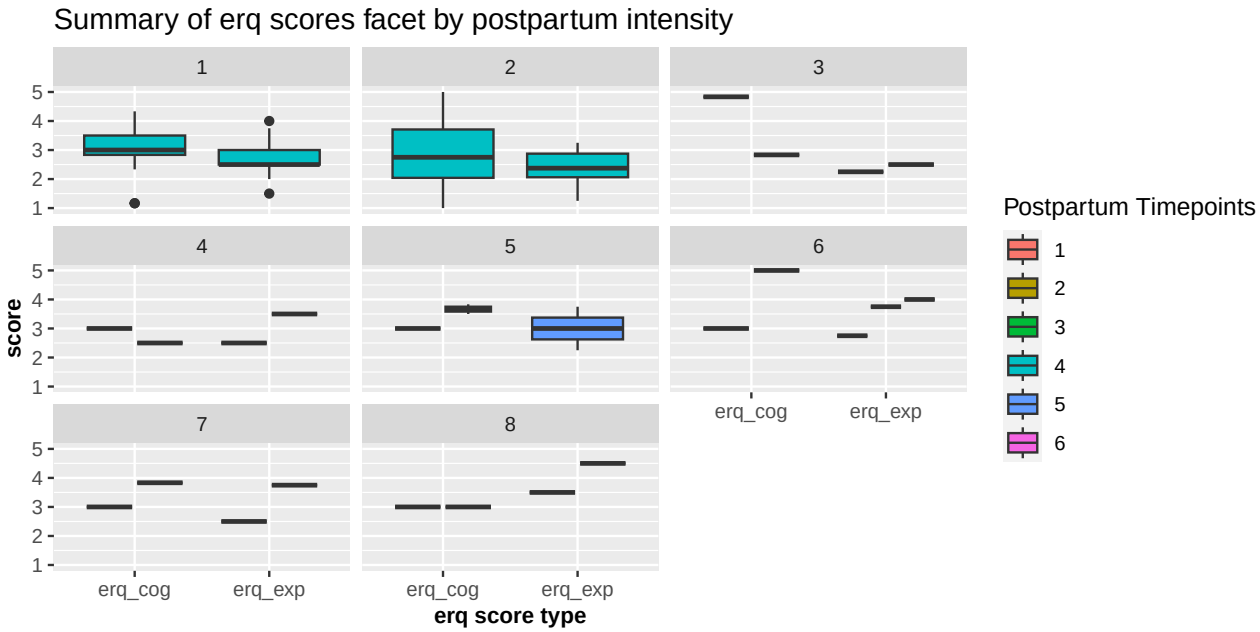


From the panel, we can see that for timing both prenatal(trimester) and postnatal(postpartum), there are generally negative association between each of the score that we selected. And also by observing the regression line we can see a similar down warding trend. This implies that earlier trimester, which is indication of earlier first exposure of SDP would lead to higher scores. A higher score in the selected questions means more potential problems with self_regulation. The intensity variable has positive correlation with the scores. Through interpretation, we can conclude that it implies as the intensity of SDP exposure increases, there may be an up warding trend of the scores, meaning more problems with self_regulation behavior. The respective relationships between each bivariate scores are positively correlated with different slopes. This is also intuitive that higher scores of one question would normally mean higher score on another related questions, indicating consistency of the data.

Before coming to the final conclusion in this section, we want to highlight the differences between erq_cog and erq_exp scores. cog correspond to Cognitive Reappraisal while exp correspond to Expressive Suppression. By looking carefully through the original questions that was answered, we realize that their trends should be considered separately. Higher cog scores means better regulation while lower exp means better regulation. So we decided to explicitely look at these two variable in particular.



To compare, we can see that both mean erq score for trimester 2 is higher for intensity 2 in comparion to 4 and we do not have more data to compare the thrid trimester.



The trend for postpartum is more obscure and we are not able to tell anything form this graph.

So our general conclusion in this section is as follows: Given the condition of data that we have, we focused on interpreting the effects of pre-natal SDP exposure in terms of time and intensity on adolescents' self-regulation problems. Although we do not obtain all statistical significant outcomes, due to the limited size of our data, we observe some consistent and genearlized patterns: earlier first-time exposure and higher cumulative intensity of exposure may be potential risk factors of worse condition of adolescent's self-regulation behavior.

Table 9: **Externalizing Summary for Prenatal(SDP)**

	2				**3**			**NA**		
Score	**N**	**2**, N = 4	**4**, N = 10	**p-value**	**N**	**1**, N = 24	**p-value**	**N**	**0**, N = 11	**p-value**
bpm_ext	11	3.00	3.63	0.8	19	2.63		7	2.29	
NA		1	2			5			4	
bpm_int	11	4.0	2.8	0.8	17	2.53		7	2.57	
NA		1	2			7			4	
bpm_ext_p	11	0.33	2.63	0.046	18	1.67		8	1.25	
NA		1	2			6			3	
bpm_int_p	11	1.33	3.50	0.11	20	2.05		8	1.63	
NA		1	2			4			3	

[1] Mean

[2] Welch Two Sample t-test

Table 10: **Externalizing Summary for Postnatl(ETS)**

	1				**2**						**3**					**4**						**5**			**6**		**NA**		
Score	**N**	**3**, N = 2	**6**, N = 1	**p-value**	**N**	**3**, N = 3	**4**, N = 1	**6**, N = 1	**7**, N = 1	**8**, N = 1	**p-value**	**N**	**2**, N = 1	**7**, N = 1	**8**, N = 1	**p-value**	**N**	**1**, N = 14	**2**, N = 8	**5**, N = 1	**6**, N = 1	**p-value**	**N**	**5**, N = 2	**p-value**	**N**	**4**, N = 1	**p-value**	**N**	**0**, N = 9	**p-value**
bpm_ext	2	7.00	4.00	>0.9	5	3.00	6.00	1.00	4.00	3.00	0.4	2	NA	1.00	7.00	0.3	21	2.00	2.67	4.00	3.00	0.3	2	3.50		1	2.0000		4	2.50	
NA		1	0			2	0	0	0	0			1	0	0			1	2	0	0								5	
bpm_int	2	2.00	3.00	>0.9	5	0.00	8.00	4.00	2.00	3.00	0.4	2	NA	0.0	12.0	0.3	19	3.00	1.33	NA	2.00	0.4	2	3.50		1	1.0000		4	1.75	
NA		1	0			2	0	0	0	0			1	0	0			2	2	1	0								5	
bpm_ext_p	2	2.00	3.00	>0.9	5	0.00	5.00	0.00	4.00	0.00	0.4	3	7.00	7.00	0.00	0.4	19	0.42	1.80	11.00	3.00	0.017	2	0.50		1	2.0000		5	0.60	
NA		1	0			2	0	0	0	0								2	3	0	0								4	
bpm_int_p	2	2.00	4.00	>0.9	5	1.00	6.00	4.00	1.00	1.00	0.4	3	4.00	9.00	2.00	0.4	20	1.46	2.40	7.00	4.00	0.3	2	1.0000		1	1.0000		6	1.17	
NA		1	0			2	0	0	0	0								1	3	0	0								3	

[1] Mean

[2] Wilcoxon rank sum exact test

[3] Kruskal-Wallis rank sum test

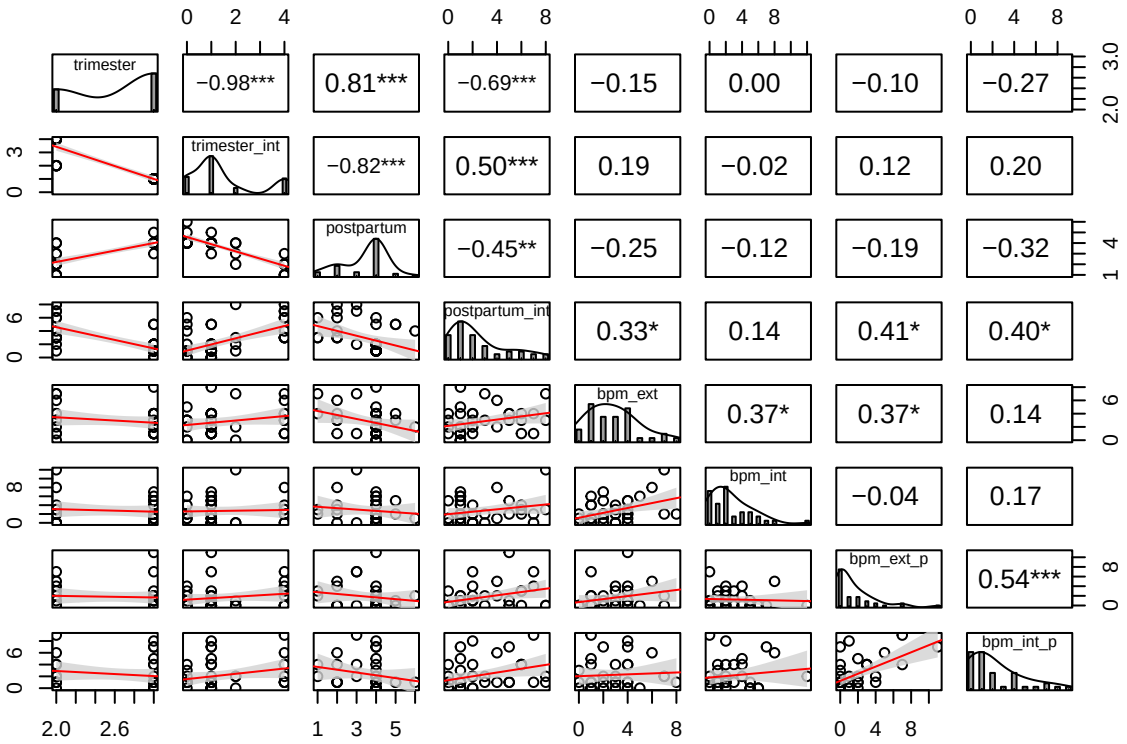
Externalizing Problems

Then we would like to focus of externalizing problems. The most intuitive indicator of externalizing problems would be bpm_score in terms of externalizing. Since we also have records about internalizing problems, we would like to also include that in this section for an additional reference. We follow the simillar pattern of exploring as above: first see the general trend of prenatal exposure vs postnatal and then look deeper into how, in each phases, exposure time and intensity may impact adolescent externalizing behavior.

Through looking at the bmp scores of externalizing and internalizing between prenatal vs posnatal from the above table(bpm_ext,bpm_int,bpm_ext_p,bpm_int_p), although the P-value does not indicate statistical significance, again, this may be due to the quality of our data, we can see similar numerical pattern in all the scores, postnatal has generally lower scores than prenatal. We would like to expolre the relationship between time and intensity.

Similar Conclusion as to the self_regulation section.(TABLE 9,10) For prenatal, the groupings are not limited and for postnatal comparison, the number of observations in each group is limited to draw more robust conclusion as well as the P-values.

We then move on to look at their correlations between each other.



By plotting time and intensity together, we can observe that for timing, both prenatal(trimester) and postnatal(postpartum) shows negative correlation with the externalizing and internalizing scores, meaning as the timing of first exposure to SDP decreases, the score decreases. This implies earlier exposure to SDP may be a risk factor to adolescent externalizing behavior as well as internalizing. On the other hand, for intensity variables, both prenatal and postnatal have positive correlation to the scores in general. This implies higher intensity of exposure to SDP may be a risk factor to adolescent externalizing behavior.

Substance Use Problem

The substance use section should contain analysis of how SDP/EST would impact adolescent's substanse use behavior. From the data quality check section, we have found out that variables related to this problem is almost entirely missing, not to mention if we startify them by timing and intensity. Here's the output

Table 11: Summary table of Substance use

	Number of YES	Cases with reporting
Cigarette	1	1
E-cigarette	3	2
Marijuana	3	3
Alcohol	5	4

From the summary table, we can see the among the total of 49 observations in our data, only 1,3,3,5 has reported YES to questions 'Have you ever' used cigarette, E-cigarette, Marijuana, and alcohol respectively and 1,2,3,4 observations reported a number of substance that they used. And this does not match to previous question, for example E-cigratte and alcohol. So we argue that with the current condition of data that we have, there's no point conducting analysis with substance use behavior and the main focus should be acquiring better data.

Conclusion and Limitation

Based on the data and variables we have, we conducted an exploratory data analysis aiming to examine effects of SDP/ETS on adolescent self-regulation, substance use, and externalizing. We created timing and intensity of effects of SDP and ETS to specify the effects. We conducted analysis in each of the three behavior dimensions of interest and used summary tables and association panels to identify their effects. Due to the limitation and quality of data that we have, we have identified patterns as follows that may serve as directions to more prudent analysis.

For self-regulation, we found out that earlier exposure to SDP/ETS may be associated to less regulation and higher cumulative intensity of exposure is associated with less regulation. Such pattern is similar for both prenatal and postnatal periods.

For externalizing, we found out that earlier exposure to SDP/ETS may be associated to more externalizing behavior and higher cumulative intensity of exposure is associated with more externalizing behavior. Such pattern is similar for both prenatal and postnatal periods.

Those findings in line with the general belief that smoke exposure is a risk factor to baby's health especially during the pregnancy phase. Take into note that we still need to have more robus statistical testings and should certain not treat such correalations as causations.

There are several crucial limitations to the data that may require further analysis if better quality of data are presented. First of all, the data size is very limited with a lot of NAs and missing. It is really hard to capture trends with different sub categories since many only contains 1 or non observations. And the data is highly unbalanced making future statistical analysis and models hard to implement. Secondly, the data contains primarily self-reported question scores and the only 'real' quantitative measure is Urine cotinine. This may suffer a great deal of bias in terms of the validity of outcomes that we want to measure. Moreover, the substance use related variables are lacking for the adolescent as most values are NA or not useful. Again this may also due to the nature of self-reporting surveys and future analysis may create methods to work around with it such as weighting and matching.

Code Appendix:

```
knitr::opts_chunk$set(echo = F)
knitr::opts_chunk$set(error = F)
knitr::opts_chunk$set(warning = F)
knitr::opts_chunk$set(message = F)
#knitr::opts_chunk$set(fig.width=8, fig.height=4)
library(tidyverse)
library(kableExtra)
library(mice)
library(gtsummary)
library(psych)
library(ggridges)
# Dimension checking
df <- read.csv('project1.csv')
print(paste0('The Dimension of Data: ',nrow(df), ' rows, and ', ncol(df), ' columns'))


# Check how many unique patients
print(paste0('There are ',length(unique(df$parent_id)),' Patients'))

# Missing checking
# Some Missing in the data is '' convert all to be NA
df = as.data.frame(apply(df, 2, function(x) ifelse(x=='',NA,x)))

# Compute missing summary table
missing_sum <- as.data.frame(apply(df[,-1], 2, function(x) sum(is.na(x)))) %>% rename('Missing_num'= "ap
  filter(Missing_num!=0) %>%
  arrange(desc(Missing_num)) %>%
  mutate('Missing_Pct'= Missing_num / nrow(df) )

# Report missing summary table
missing_sum[1:10,] %>%
  kable(booktabs = TRUE, caption = "Missingness Count and Percentage of columns") %>%
  kable_styling(full_width = TRUE, latex_options = "hold_position")

ever_sum <- apply(df[,c("cig_ever","e_cig_ever","mj_ever","alc_ever")], 2, function(x) x==' 1') %>% col

# Report summary of substance related questions
ever_sum %>%
  kable(booktabs = TRUE, caption = "Number of Teenagers answered YES Substance Use related questions",
  kable_styling(full_width = TRUE, latex_options = "hold_position")
# Identify outliers
ol = as.data.frame(as.matrix(cbind(df$parent_id[c(1,26,37)],df$mom_numcig[c(1,26,37)])))
names(ol) = c('parent_id','mom_numcig')

# Report outliers
ol %>% kable(booktabs = TRUE, caption = "Irregular reporting of mom numcig") %>%
  kable_styling(full_width = TRUE, latex_options = "hold_position")
# Univariate analysis with Brief problem moniter
df %>% pivot_longer(c( "bpm_att","bpm_ext","bpm_int","bpm_att_p","bpm_ext_p","bpm_int_p","bpm_att_a","bp
ggplot(aes(x = as.numeric(score), y = as.factor(score_type), fill=as.factor(score_type))) + geom_density
       labs(x='Score',y='Brief Problem Moniter',title='Distribution of Brief Problem Moniter Scores',fi
       theme(
```

```
        axis.title=element_text(size=10,face="bold"))
# Univariate analysis with emotion regulation
df %>% pivot_longer(c("erq_cog_a","erq_exp_a","erq_cog","erq_exp"),names_to ='score_type',values_to = 's
ggplot(aes(x = as.numeric(score), y = as.factor(score_type), fill=as.factor(score_type))) + geom_density
       labs(x='Score',y='Emotional Regulation',title='Distribution of Emotional Regulation Scores',fill=
       theme(
        axis.title=element_text(size=10,face="bold"))
# Univariate analysis with parental knowledge
df %>% pivot_longer(c('pmq_parental_knowledge','pmq_child_disclosure','pmq_parental_solicitation','pmq_
ggplot(aes(x = as.numeric(score), y = as.factor(score_type), fill=as.factor(score_type))) + geom_density
       labs(x='Score',y='Parental Knowledge',title='Distribution of Parental Knowledge Scores',fill='Pa
       theme(
        axis.title=element_text(size=10,face="bold"))
# Transferring scores and lab test to be numeric

score_col = c('bpm_att','bpm_ext','bpm_int','erq_cog','erq_exp','pmq_parental_knowledge','pmq_child_dis

df = df %>% mutate_at(score_col, as.numeric)
# Correct swan vals
swan_na = which(df$parent_id%in%c(50502,51202,51602,52302,53002,53502,53902,54402,54602,54702))
df$swan_hyperactive[swan_na] <- NA
df$swan_inattentive[swan_na] <- NA
# coti indicator
df$cotimean_34wk_1 <- ifelse(df$cotimean_34wk != 0,"1","0")

# trimester timing, meaning the first trimester this observation was exposed to SDP
df <- df %>% mutate(trimester = ifelse(mom_smoke_16wk%in%'1=Yes'|mom_smoke_22wk%in%'1=Yes',2,
                                       ifelse(mom_smoke_32wk%in%'1=Yes'|cotimean_34wk_1%in%'1',3, NA)))

# trimester intensity
df <- df %>% mutate(trimester_int = rowSums(apply(select(.,c(mom_smoke_16wk,mom_smoke_22wk,mom_smoke_32
                                                  ,2, function(x) x %in% c("1",'1=Yes')), na.rm=TRUE))


#cbind(df$trimester,df$trimester_int)

# cotimean indicator
df$cotimean_pp6mo_baby_1 = ifelse(df$cotimean_pp6mo_baby != 0,"1","0")

# postpartum timing, meaning the first postpartum timepoint collected in the dataset, this observation
df <- df %>% mutate(postpartum = ifelse(mom_smoke_pp1%in%'1=Yes',1,
                                 ifelse(mom_smoke_pp2%in%'1=Yes',2,
                                 ifelse(mom_smoke_pp12wk%in%'1=Yes',3,
                                 #ifelse(mom_smoke_pp6mo=='1=Yes',4,
                                 ifelse(cotimean_pp6mo_baby_1%in%'1',4,
                                 ifelse(df$smoke_exposure_12mo%in%' 1',5,
                                 ifelse(smoke_exposure_2yr%in%' 1',6,
                                 ifelse(smoke_exposure_3yr%in%' 1',7,
                                 ifelse(smoke_exposure_4yr%in%' 1',8,
                                 ifelse(smoke_exposure_5yr%in%' 1',9, NA)))))))))))

# combine early pp for intensity coding
df <- df %>% mutate(early_post = ifelse(mom_smoke_pp1 %in% '1=Yes'|mom_smoke_pp2 %in% '1=Yes'|mom_smoke_
```

```r
# cotimean indicator
df$cotimean_pp6mo_1 = ifelse(df$cotimean_pp6mo > 10,"1","0")



df <- df %>% mutate(postpartum_int = rowSums(apply(select(.,c(early_post,cotimean_pp6mo_1,cotimean_pp6m



# Timing variable, in terms of first exposure occured at prenatal vs postnatal
df = df %>% mutate(natal = ifelse(mom_smoke_16wk%in%'1=Yes'|mom_smoke_22wk%in%'1=Yes'|
                                  mom_smoke_32wk%in%'1=Yes'|cotimean_34wk_1%in%'1','Prenatal',
                           ifelse(mom_smoke_pp1%in%'1=Yes'|mom_smoke_pp2%in%'1=Yes'|mom_smoke_pp12wk%in
                                  smoke_exposure_6mo%in%' 1'|smoke_exposure_12mo%in%' 1'|
                                  mom_smoke_pp6mo%in%'1=Yes'|cotimean_pp6mo_baby_1%in%'1'|cotimean_pp6m
                                  smoke_exposure_2yr%in%' 1'|smoke_exposure_3yr%in%' 1'|smoke_exposure_
                                  ,'Postnatal', NA )))

# Report summary of prenatal timing and intensity
table(df$trimester,df$trimester_int)%>%
  kable(booktabs = TRUE, caption = "Summary of Prenatal Variable",row.names = T ) %>%
kable_styling(full_width = F, latex_options = "hold_position") %>% add_footnote(c("Row is time, column

# Report summary of postnatal timing and intensity
table(df$postpartum,df$postpartum_int)%>%
  kable(booktabs = TRUE, caption = "Summary of Postpartum Variable",row.names = T ) %>%
kable_styling(full_width = F, latex_options = "hold_position") %>% add_footnote(c("Row is time, column
# Summary by natal
df %>% select(natal, bpm_att , bpm_ext , bpm_int , erq_cog , erq_exp , pmq_parental_knowledge , pmq_chi
      tbl_summary(by = natal,
        missing_text = "NA",
        type = list(everything() ~ 'continuous'
                    ),
    statistic = all_continuous() ~ "{mean}") %>%
      add_n() %>% add_p(test=all_continuous() ~ "t.test") %>% as_gt()
# Descriptive table of the self_regulation scores stratified by exposure and intensity
# Prenatal
t1 = df %>% select(c(trimester,trimester_int,erq_cog,erq_exp,bpm_att,bpm_att_p,swan_hyperactive,swan_ina
  tbl_strata(
    strata = trimester,
    .tbl_fun =
      ~ .x %>%
      tbl_summary(by = trimester_int,
        missing_text = "NA",
        type = list(erq_cog~'continuous',
                    erq_exp~'continuous',
                    bpm_att~'continuous',
                    bpm_att_p~'continuous',
                    swan_hyperactive~'continuous',
                    swan_inattentive~'continuous'
                    ),
    statistic = all_continuous() ~ "{mean}") %>%
      add_n() %>% add_p(test=all_continuous() ~ "t.test"),
```

```
    .header = "**{strata}**"
  )


 t1 %>%
  modify_header(label = "**Score**", p.value = "**P**") %>%
  modify_caption("**Self Regulation Summary for Prenatal(SDP)**")  %>%  as_kable_extra(booktabs = TRUE)
  kableExtra::kable_styling(font_size = 7)
# Descriptive table of the self_regulation scores stratified by exposure and intensity
# Postnatal
t2 = df %>% select(c(postpartum,postpartum_int,erq_cog,erq_exp,bpm_att,bpm_att_p,swan_hyperactive,swan_
  tbl_strata(
    strata = postpartum,
    .tbl_fun =
      ~ .x %>%
      tbl_summary(by = postpartum_int,
        missing_text = "NA",
        type = list(erq_cog~'continuous',
                    erq_exp~'continuous',
                    bpm_att~'continuous',
                    bpm_att_p~'continuous',
                    swan_hyperactive~'continuous',
                    swan_inattentive~'continuous'
                   ),
    statistic = all_continuous() ~ "{mean}") %>%
      add_n() %>% add_p(),
    .header = "**{strata}**"
  )


 t2 %>%
  modify_header(label = "**Score**", p.value = "**P**") %>%
  modify_caption("**Self Regulation Summary for Postnatal(ETS)**")  %>%  as_kable_extra(booktabs = TRUE)
  kableExtra::kable_styling(font_size = 7,latex_options = "scale_down")
# Correlation penal pair of timing and intensity of self_regulation
pairs.panels(df[,c("trimester","trimester_int",'postpartum','postpartum_int','erq_cog','erq_exp','bpm_a
             smooth = F,       # If TRUE, draws loess smooths
             scale = FALSE,       # If TRUE, scales the correlation text font
             density = TRUE,      # If TRUE, adds density plots and histograms
             ellipses = F,     # If TRUE, draws ellipses
             method = "pearson", # Correlation method (also "spearman" or "kendall")
             pch = 21,            # pch symbol
             lm = T,          # If TRUE, plots linear fit rather than the LOESS (smoothed) fit
             cor = T,          # If TRUE, reports correlations
             jiggle = FALSE,      # If TRUE, data points are jittered
             factor = 2,          # Jittering factor
             hist.col = 8,        # Histograms color
             stars = TRUE,        # If TRUE, adds significance level with stars
             ci = TRUE)           # If TRUE, adds confidence intervals
# analysis of erq score at prenatal period
df %>% pivot_longer(c(erq_cog,erq_exp),names_to = 'score_type',values_to = 'score') %>% filter(!is.na(t
  ggplot(aes(x=score_type,y=score,fill=as.factor(trimester))) + geom_boxplot() + facet_wrap(~as.factor(
       labs(x='erq score type',y='score',title='Summary of erq scores facet by trimester intensity',fil
```

```
       theme(
        axis.title=element_text(size=10,face="bold"))
# analysis with erq scores at postntal period
df %>% pivot_longer(c(erq_cog,erq_exp),names_to = 'score_type',values_to = 'score') %>% filter(!is.na(p
  ggplot(aes(x=score_type,y=score,fill=as.factor(postpartum))) + geom_boxplot() + facet_wrap(~as.factor
       labs(x='erq score type',y='score',title='Summary of erq scores facet by postpartum intensity',fi
       theme(
        axis.title=element_text(size=10,face="bold"))
# Descriptive table of the externalizing scores stratified by exposure and intensity
# Prenatal
t3 = df %>% select(c(trimester,trimester_int,bpm_ext,bpm_int,bpm_ext_p,bpm_int_p)) %>%
  tbl_strata(
    strata = trimester,
    .tbl_fun =
      ~ .x %>%
      tbl_summary(by = trimester_int,
        missing_text = "NA",
        type = list(everything() ~ 'continuous'
                    ),
    statistic = all_continuous() ~ "{mean}") %>%
      add_n() %>% add_p(test=all_continuous() ~ "t.test"),
    .header = "**{strata}**"
  )


 t3 %>%
  modify_header(label = "**Score**", p.value = "**P**") %>%
  modify_caption("**Externalizing Summary for Prenatal(SDP)**")  %>%  as_kable_extra(booktabs = TRUE) %
  kableExtra::kable_styling(font_size = 7)
# Descriptive table of the externalizing behavior scores stratified by exposure and intensity
# Postnatal
t4 = df %>% select(c(postpartum,postpartum_int,bpm_ext,bpm_int,bpm_ext_p,bpm_int_p)) %>%
  tbl_strata(
    strata = postpartum,
    .tbl_fun =
      ~ .x %>%
      tbl_summary(by = postpartum_int,
        missing_text = "NA",
        type = list(everything() ~ 'continuous'
                    ),
    statistic = all_continuous() ~ "{mean}") %>%
      add_n() %>% add_p(),
    .header = "**{strata}**"
  )


 t4 %>%
  modify_header(label = "**Score**", p.value = "**P**") %>%
  modify_caption("**Externalizing Summary for Postnatl(ETS)**")  %>%  as_kable_extra(booktabs = TRUE) %
  kableExtra::kable_styling(font_size = 7,latex_options = "scale_down")
# Correlation penal pair of timing and intensity of self_regulation
pairs.panels(df[,c("trimester","trimester_int","postpartum","postpartum_int",'bpm_ext','bpm_int','bpm_e
             smooth = F,       # If TRUE, draws loess smooths
```

```
             scale = FALSE,       # If TRUE, scales the correlation text font
             density = TRUE,      # If TRUE, adds density plots and histograms
             ellipses = F,    # If TRUE, draws ellipses
             method = "spearman", # Correlation method (also "spearman" or "kendall")
             pch = 21,            # pch symbol
             lm = T,         # If TRUE, plots linear fit rather than the LOESS (smoothed) fit
             cor = T,         # If TRUE, reports correlations
             jiggle = FALSE,      # If TRUE, data points are jittered
             factor = 2,          # Jittering factor
             hist.col = 8,        # Histograms color
             stars = TRUE,        # If TRUE, adds significance level with stars
             ci = TRUE)           # If TRUE, adds confidence intervals
# Compute substance use summary tables
cig_ever = sum(df$cig_ever %in% ' 1')
num_cig = sum(!is.na(df$num_cigs_30))
ecig_ever = sum(df$e_cig_ever%in%' 1')
num_ecig = sum(!is.na(df$num_e_cigs_30))
mj_ever = sum(df$mj_ever %in% ' 1')
num_mj = sum(!is.na(df$num_mj_30))
al_ever = sum(df$alc_ever %in% ' 1')
num_al = sum(!is.na(df$num_alc_30))

su_tab = as.data.frame(cbind(c(cig_ever,ecig_ever,mj_ever,al_ever),c(num_cig,num_ecig,num_mj,num_al)))
rownames(su_tab) = c('Cigarette','E-cigarette','Marijuana','Alcohol')
colnames(su_tab) = c('Number of YES','Cases with reporting')

# Report substance use summary table
su_tab %>% kable(booktabs = TRUE, caption = "Summary table of Substance use") %>%
  kable_styling(full_width = TRUE, latex_options = "hold_position")
```