Project 1

Due: October 8th at 11:59pm

Tabib Chowdhury

Introduction

In this report, we conduct an exploratory data analysis of this data to examine the association between smoking during pregnancy (SDP) and environmental tobacco smoke (ETS) exposure and self-regulation, externalizing behavior, and substance use. The women in this study were recruited from a previous study on smoke avoidance intervention to reduce low-income women's (N=738) smoking and ETS exposure during pregnancy and children's exposure to ETS in the immediate postpartum period. A subset of adolescents (N=100) and their mothers are randomly selected for recruitment into this study.

The data I work with has been cleaned and preprocessed by previous researchers.

In the first section we will analyze missing data. This is followed-up by looking at some demographic statistics of the participants of the study. Next we attempt to find patterns between ETS and SDP and their effects on adolescent self-regulation, substance use, and externalizing.

Missing data

Table 1: Frequencies of Total Variables Missing for Each Patient

# Variables Missing for Each Participant																						
1	4	5	6	7	8	9	10	11	12	14	15	22	24	28	30	33	37	54	55	56	58	
1	1	6	8	6	3	4	2	1	1	1	1	1	1	1	1	1	1	2	2	1	3	# Participants Miss

Table 2: Frequencies of Total Observations Missing for Each Variable

# Observations Missing for Each Variable																			
0	1	7	8	9	10	11	12	13	14	15	16	20	28	39	45	46	47	48	
15	1	2	6	2	13	8	13	6	3	1	1	1	1	1	1	1	1	1	# Variables

Looking at the raw dataset we see that missing values come in the form of empty strings and NA values, we want everything to be consistent to analyze missing data so we replaced the empty strings with NA objects. To find major patterns in the missing data, we can first see if there are variables and participants with the same number of missing units.

When looking at the frequency of total variables missing for each participant(table 1), we can see that the primary pattern is that there are 8 participants missing at least 54 of the 59 variables(>91% of the columns). Looking at the frequency of participants with missing values for each variable(table 2), it is apparent that there are atleast 4 variables missing 45/49(91.8%) observations. These variables are num_cigs_30, num_e_cigs_30, num_mj_30 and num_alc_30. These record observations of the child when asked if they smoked cigarettes, e-cigaretttes or vape, marijuana or drink alcohol in the past 30 days. The reason there are so many missing observations are because these questions were asked only if the child had EVER done the activities described and most of the children had not. So we should place a zero in the where the child was asked if they smoked cigarettes, e-cigaretttes or vape, marijuana or drink alcohol in the past 30 days if the child that had never smoked cigarettes, e-cigaretttes or vape, marijuana or drink alcohol.

After doing this, we can check to see the over all missing patterns among the whole dataset. From table 3, we can see that more than 50% of the observations are missing from 5 variables:childasd, num_alc_30, num_mj_30, num_ecigs_30 and mom_smoke_pp1.

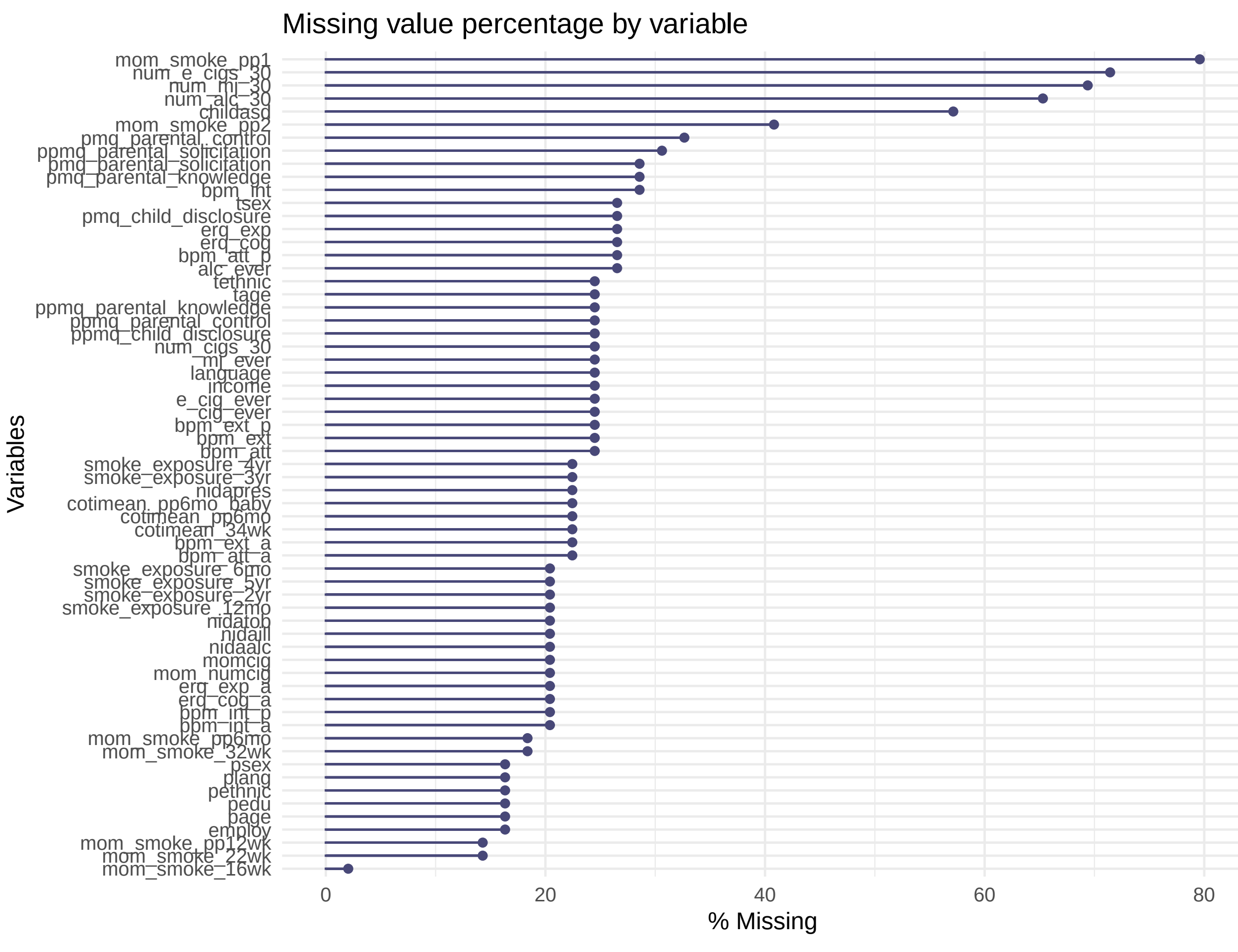
Table 3: Missing Data Proportion for Each Variable

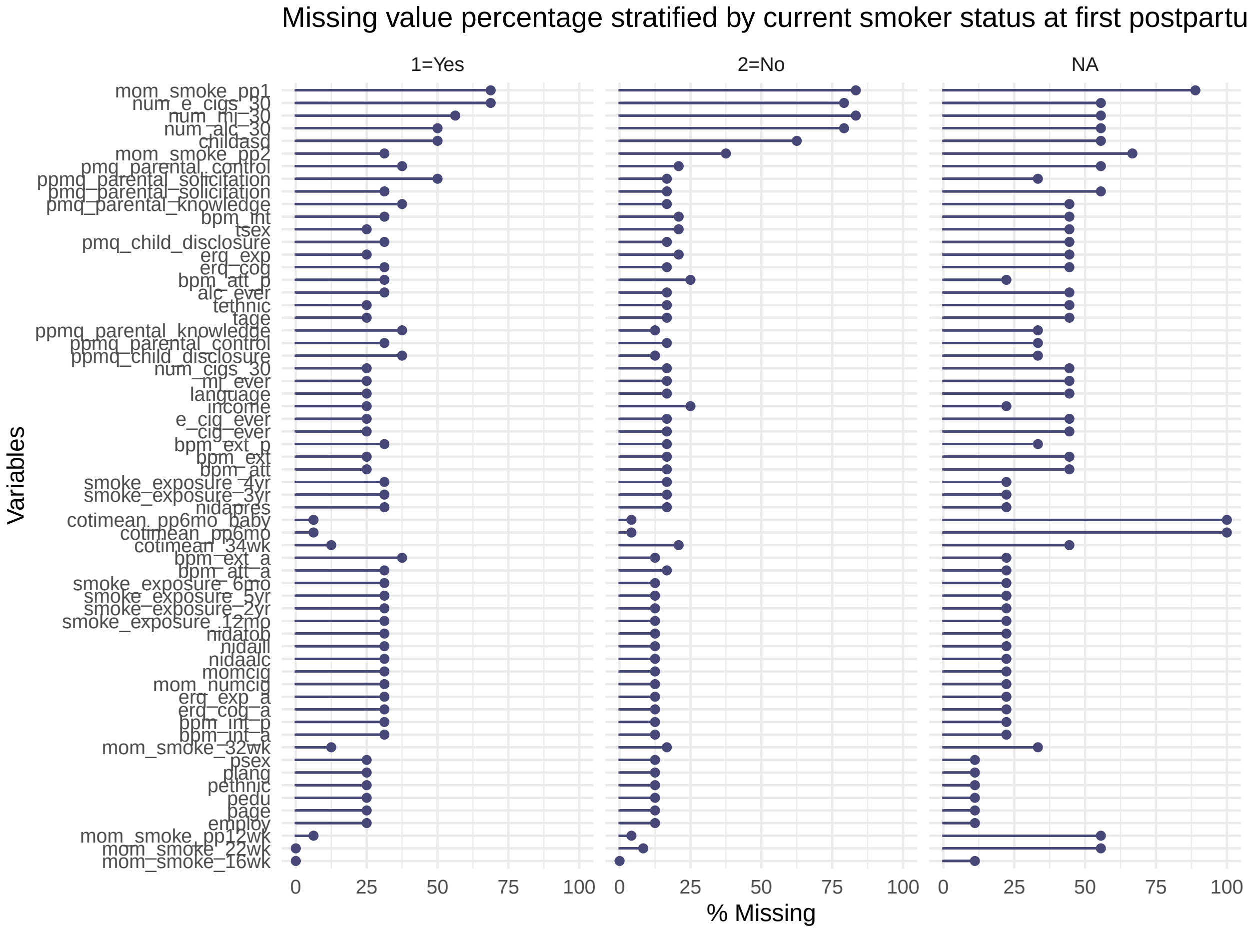
Variable	Observation Missing	Proportion Missing
mom_smoke_pp1	39	79.591837
num_e_cigs_30	35	71.428571
num_mj_30	34	69.387755
num_alc_30	32	65.306122
childasd	28	57.142857
mom_smoke_pp2	20	40.816327
pmq_parental_control	16	32.653061
ppmq_parental_solicitation	15	30.612245
bpm_int	14	28.571429
pmq_parental_knowledge	14	28.571429
pmq_parental_solicitation	14	28.571429
bpm_att_p	13	26.530612
tsex	13	26.530612
alc_ever	13	26.530612
erq_cog	13	26.530612
erq_exp	13	26.530612
pmq_child_disclosure	13	26.530612
income	12	24.489796
bpm_ext_p	12	24.489796
ppmq_parental_knowledge	12	24.489796
ppmq_child_disclosure	12	24.489796
ppmq_parental_control	12	24.489796
tage	12	24.489796
language	12	24.489796
tethnic	12	24.489796
cig_ever	12	24.489796
num_cigs_30	12	24.489796
e_cig_ever	12	24.489796
mj_ever	12	24.489796
bpm_att	12	24.489796
bpm_ext	12	24.489796
nidapres	11	22.448980
cotimean_34wk	11	22.448980
cotimean_pp6mo_baby	11	22.448980
cotimean_pp6mo	11	22.448980
smoke_exposure_3yr	11	22.448980
smoke_exposure_4yr	11	22.448980
bpm_att_a	11	22.448980
bpm_ext_a	11	22.448980
nidaalc	10	20.408163
nidatob	10	20.408163
nidaill	10	20.408163
momcig	10	20.408163
mom_numcig	10	20.408163
bpm_int_p	10	20.408163
smoke_exposure_6mo	10	20.408163
smoke_exposure_12mo	10	20.408163
smoke_exposure_2yr	10	20.408163
smoke_exposure_5yr	10	20.408163
bpm_int_a	10	20.408163
erq_cog_a	10	20.408163
erq_exp_a	10	20.408163

To visualize the missing data, the plot showing missing value percentage by variable displays missing values with atleast one NA as an observation. Those with no missing values(race ethnicity) were excluded. It can be said that there is some patterns to it as the variables look grouped together by how many observations are missing per variable.

Another example of grouped missing variables come from the variables describing the average response on the Parental Knowledge Questionnaire and of responses on the Brief Problem Monitor. Variables describing children who responded to the Parental Knowledge Questionnaire have a missing value percentage of between 26.5-32.7%, while parents who responded to the Parental Knowledge Questionnaire have a missing value percentage of between 24.5-30.6%. Variables describing children's record on the Brief Problem Monitor have a missing value rate of 24.5-28.6%

Next we can stratify the missing value percentages by one of the independent variables of ETS(Environmental Tobacco Smoke): current smoker status at first postpartum visit of parent. It can be said that we are more likely to have more missing values across most of the variables if the parent is a smoker at the first postpartum visit than if they are not. It is also more likely to have missing values across all variables if the parent did not respond or follow up for the first postpartum visit.





Given all these patters we can rule out the fact that our data is MCAR(missing completely at random). To conclude that the data is MNAR(missing not at random), we would need to investigate more as we do not know what outside factors can affect the missing data patterns. Thus the data is MAR(missing at random) because we have found significant missing data patterns though the observational data.

It is important to note that the data's dimention is (49 rows, 78 cols). Because there are only 49 observations it would not be smart to perform data imputations to replace the missing values for the data set. Additionally the number of variables being larger than the number of observations does not help the case for imputaion as well.

Exploratory Data Analysis

To start the main exploratory data analysis, it is important to look at the demographics of the children and parents involved in the study. This is shown in table 4 and 5. One main concern is the imbalance of characteristics in the data. Group demographics are not proportionate across subject.

In table 6, we perform Exploratory Data Analysis on the SDP variable which indicated if the mom had smoked in weeks 16,22 or 32. I created a new variable which counted the number of times the mom had said yes to this question to guage whether there was any specific relationship between smoking in weeks 16,22 or 32 and the dependent variables in the study. When we look at the urine cotinine levels for the mom, we can see that as the number of times the parent responded yes to smoking at those times increases the mean urine cotinine levels increase at both 34 weeks gestation and 6 months post-partinum. Similarly, the baby's urine cotinine levels increase as the smoking frequency during pregnancy increases. Additionally this relationship is statistically different between the 3 smoking groups.

Table 4: Demographic summary of parents

Characteristic	**N = 49**
Age	37.5 (3.6)
(Missing)	8
Race	
American Indian/Alaska Native	4 / 41 (9.8%)
Native Hawaiin or Pacific Islander	6 / 41 (15%)
Other	6 / 41 (15%)
White	25 / 41 (61%)
(Missing)	8
income	63,138 (59,885)
(Missing)	12
Employment Status	
No	12 / 41 (29%)
Part-Time	7 / 41 (17%)
Full-Time	22 / 41 (54%)
(Missing)	8
Sex	
Male	1 / 41 (2.4%)
Female	40 / 41 (98%)
Intersex	0 / 41 (0%)
(Missing)	8
Language Spoken	
Did not speak another language at home	26 / 41 (63%)
Spoke another language at home	15 / 41 (37%)
(Missing)	8

[1] Mean (SD); n / N (%)

Table 5: Demographic summary of children

Characteristic	N = 49
Age	13.62 (1.21)
(Missing)	12
Sex	
0	23 / 36 (64%)
1	13 / 36 (36%)
(Missing)	13
Language Spoken	
0	26 / 37 (70%)
1	11 / 37 (30%)
(Missing)	12
Race	
American Indian/Alaska Native	5 / 29 (17%)
Other	5 / 29 (17%)
White	19 / 29 (66%)
(Missing)	20
Sex	
Male	23 / 36 (64%)
Female	13 / 36 (36%)
Intersex	0 / 36 (0%)
(Missing)	13

[1] Mean (SD); n / N (%)

Table 6: **Self Regulation and externalizing factors stratified by SDP scores**

Variable	N	**Overall**, N = 48	Number of times mother said 'yes' to smoking at 16,22, and 32 weeks pregnant				p-value
			0, N = 34	**1**, N = 3	**2**, N = 1	**3**, N = 10	
cotimean_34wk	38	50 (98)	2 (4)	7 (9)	NA (NA)	183 (112)	<0.001
Unknown		10	9	0	1	0	
cotimean_pp6mo_baby	38	4.0 (7.6)	2.7 (4.9)	2.1 (1.7)	3.2 (NA)	9.3 (13.4)	0.031
Unknown		10	8	0	0	2	
cotimean_pp6mo	38	100 (179)	42 (103)	47 (57)	83 (NA)	313 (256)	0.001
Unknown		10	8	0	0	2	
swan_inattentive	48	8.9 (6.5)	8.4 (6.4)	10.3 (1.5)	0.0 (NA)	10.8 (7.6)	0.31
swan_hyperactive	48	6.3 (6.5)	5.1 (5.7)	10.3 (3.8)	0.0 (NA)	9.9 (8.3)	0.11
bpm_att_p	36	2.06 (2.22)	1.64 (1.87)	1.00 (1.00)	NA (NA)	3.75 (2.82)	0.094
Unknown		12	9	0	1	2	
bpm_ext_p	37	1.68 (2.50)	1.54 (2.55)	0.33 (0.58)	NA (NA)	2.63 (2.62)	0.29
Unknown		11	8	0	1	2	
bpm_int_p	39	2.21 (2.48)	1.93 (2.40)	1.33 (1.15)	NA (NA)	3.50 (2.88)	0.18
Unknown		9	6	0	1	2	
ppmq_parental_knowledge	37	4.26 (0.58)	4.24 (0.56)	4.33 (0.40)	NA (NA)	4.30 (0.76)	0.71
Unknown		11	7	0	1	3	
ppmq_child_disclosure	37	3.68 (0.67)	3.76 (0.57)	3.87 (0.50)	NA (NA)	3.29 (1.01)	0.42
Unknown		11	7	0	1	3	
ppmq_parental_solicitation	34	4.18 (0.73)	4.22 (0.62)	5.00 (0.00)	NA (NA)	3.73 (1.03)	0.075
Unknown		14	8	1	1	4	
ppmq_parental_control	37	4.58 (0.95)	4.65 (0.85)	4.40 (0.53)	NA (NA)	4.45 (1.40)	0.39
Unknown		11	8	0	1	2	
bpm_att_a	38	1.47 (1.96)	1.07 (1.30)	2.67 (4.62)	NA (NA)	2.38 (2.39)	0.33
Unknown		10	7	0	1	2	
bpm_ext_a	38	1.24 (1.57)	1.11 (1.64)	1.00 (0.00)	NA (NA)	1.75 (1.49)	0.26
Unknown		10	6	1	1	2	
bpm_int_a	39	1.54 (1.86)	1.14 (1.48)	1.67 (1.53)	NA (NA)	2.88 (2.64)	0.17
Unknown		9	6	0	1	2	
erq_cog_a	39	5.38 (1.30)	5.50 (1.35)	4.78 (2.10)	NA (NA)	5.21 (0.83)	0.39
Unknown		9	6	0	1	2	
erq_exp_a	39	3.46 (1.58)	3.09 (1.24)	5.00 (2.14)	NA (NA)	4.16 (2.05)	0.20
Unknown		9	6	0	1	2	
bpm_att	37	3.00 (2.62)	2.50 (2.27)	2.33 (4.04)	NA (NA)	4.88 (2.70)	0.084
Unknown		11	8	0	1	2	
bpm_ext	37	2.81 (2.01)	2.54 (1.79)	3.00 (3.61)	NA (NA)	3.63 (2.13)	0.45
Unknown		11	8	0	1	2	
bpm_int	35	2.71 (2.73)	2.54 (2.15)	4.00 (6.93)	NA (NA)	2.75 (2.55)	0.82
Unknown		13	10	0	1	2	
erq_cog	36	3.19 (0.97)	3.10 (1.05)	3.67 (0.60)	NA (NA)	3.36 (0.73)	0.53
Unknown		12	8	0	1	3	
erq_exp	36	2.75 (0.80)	2.58 (0.79)	3.67 (0.72)	NA (NA)	2.94 (0.62)	0.081
Unknown		12	9	0	1	2	
pmq_parental_knowledge	35	3.99 (0.79)	4.13 (0.62)	3.67 (0.78)	NA (NA)	3.62 (1.22)	0.42
Unknown		13	9	0	1	3	
pmq_child_disclosure	36	3.43 (1.00)	3.62 (0.99)	3.13 (0.12)	NA (NA)	2.98 (1.10)	0.29
Unknown		12	9	0	1	2	
pmq_parental_solicitation	35	2.98 (1.34)	3.19 (1.29)	2.33 (0.83)	NA (NA)	2.58 (1.61)	0.37
Unknown		13	10	0	1	2	
pmq_parental_control	33	4.35 (0.93)	4.39 (0.88)	4.70 (0.42)	NA (NA)	4.13 (1.19)	0.78
Unknown		15	11	1	1	2	
Autism Diagnosis Status	20						>0.99
No		18 / 20 (90%)	14 / 16 (88%)	1 / 1 (100%)	0 / 0 (NA%)	3 / 3 (100%)	
Diagnosed		1 / 20 (5.0%)	1 / 16 (6.3%)	0 / 1 (0%)	0 / 0 (NA%)	0 / 3 (0%)	
Suspected		1 / 20 (5.0%)	1 / 16 (6.3%)	0 / 1 (0%)	0 / 0 (NA%)	0 / 3 (0%)	
Unknown		28	18	2	1	7	

[1] Mean (SD); n / N (%)

[2] Kruskal-Wallis rank sum test; Fisher's exact test

Table 7: **Substance Use stratified by SDP Score**

Variable	N	Overall, N = 48	SDP Score: 0, N = 34	1, N = 3	2, N = 1	3, N = 10	p-value
cig_ever	37						0.30
0		36 / 37 (97%)	26 / 26 (100%)	3 / 3 (100%)	0 / 0 (NA%)	7 / 8 (88%)	
1		1 / 37 (2.7%)	0 / 26 (0%)	0 / 3 (0%)	0 / 0 (NA%)	1 / 8 (13%)	
Unknown		11	8	0	1	2	
e_cig_ever	37						0.11
0		34 / 37 (92%)	25 / 26 (96%)	2 / 3 (67%)	0 / 0 (NA%)	7 / 8 (88%)	
1		3 / 37 (8.1%)	1 / 26 (3.8%)	1 / 3 (33%)	0 / 0 (NA%)	1 / 8 (13%)	
Unknown		11	8	0	1	2	
mj_ever	37						0.21
0		34 / 37 (92%)	25 / 26 (96%)	3 / 3 (100%)	0 / 0 (NA%)	6 / 8 (75%)	
1		3 / 37 (8.1%)	1 / 26 (3.8%)	0 / 3 (0%)	0 / 0 (NA%)	2 / 8 (25%)	
Unknown		11	8	0	1	2	
alc_ever	36						0.14
0		31 / 36 (86%)	24 / 26 (92%)	2 / 3 (67%)	0 / 0 (NA%)	5 / 7 (71%)	
1		5 / 36 (14%)	2 / 26 (7.7%)	1 / 3 (33%)	0 / 0 (NA%)	2 / 7 (29%)	
Unknown		12	8	0	1	3	

[1] n / N (%)

[2] Fisher's exact test

To study the association between SDP and externalizing factors, we can look at the bpm scores of both the children and parents. The mean scores have a general increasing trend for internal, external and attention problems for both the parent and the children. Thus the bpm scores related to attention, externalizing and internalizing problems on self and child increase as the smoking frequency during pregnancy increases for the parent. More importantly, the same relationship exists for the child where the child's bpm scores related to attention, externalizing and internalizing problems increase as the smoking during pregnancy frequency increase. This is once again detailed in table 6.

Looking at the relationship between SDP and substance use in table 7, we don't observe any major statistical differences however we can see how the use of e_cigs, cigarettes, alcohol and marijuana have a general increase as SDP scores increase.

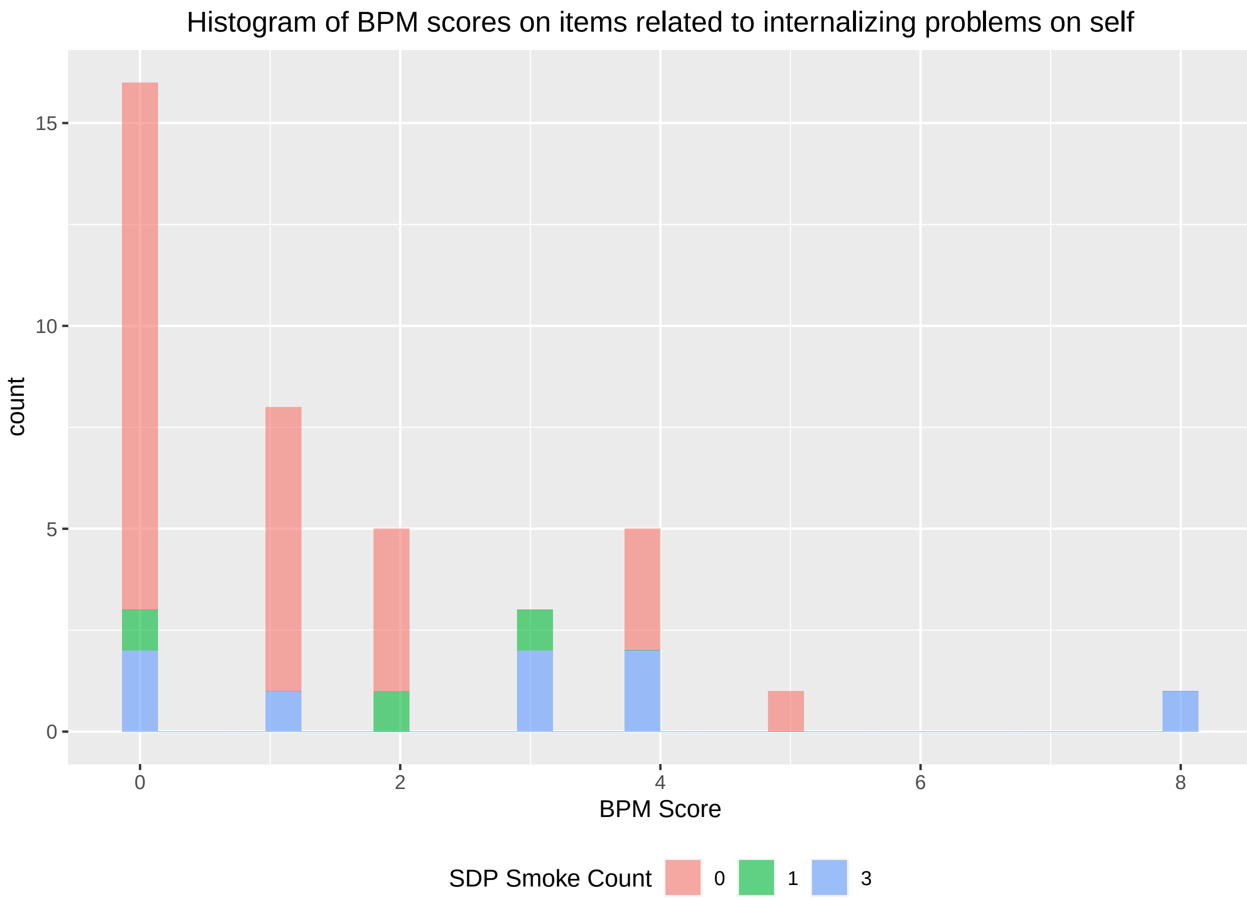


Figure 3. Association between SDP and BPM scores

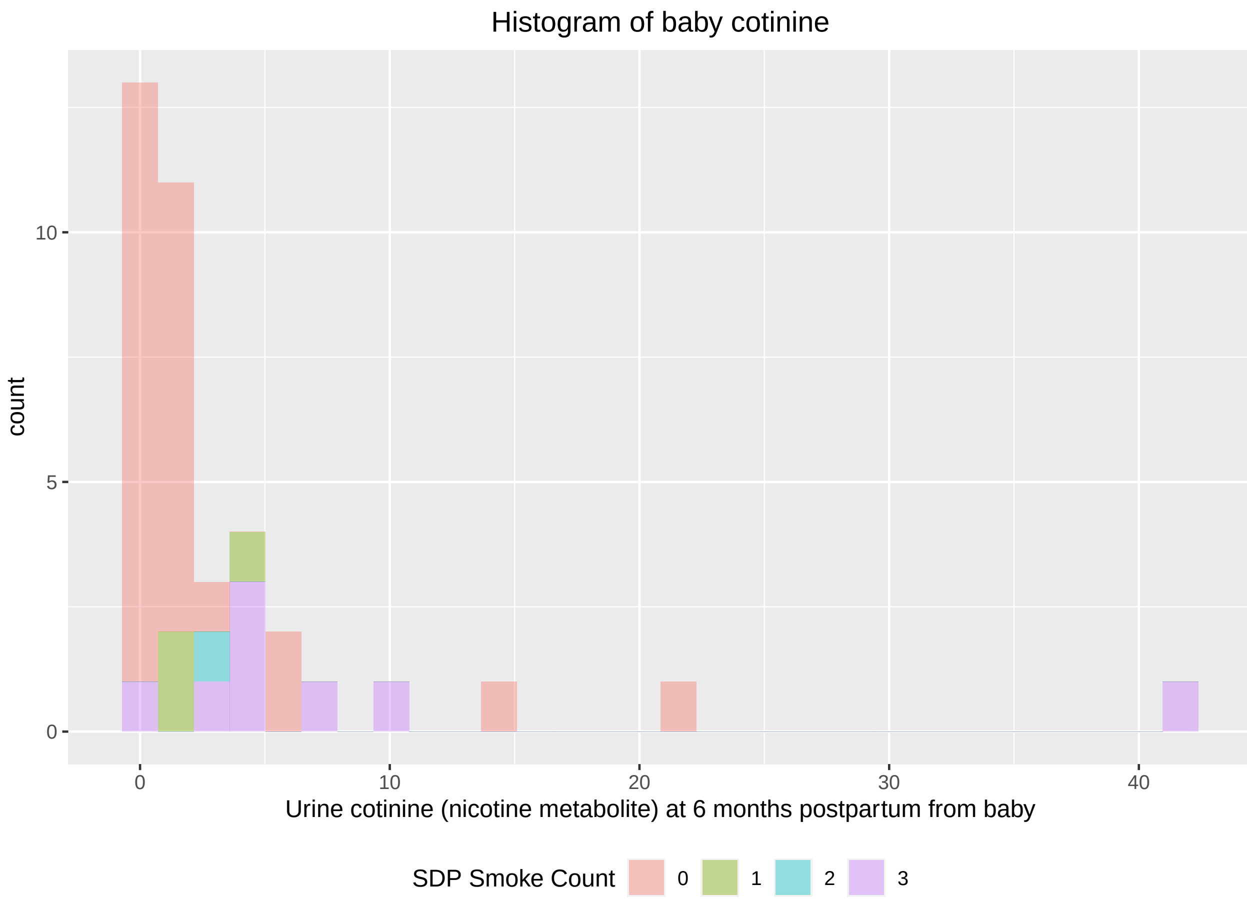


Figure 3. Association between SDP and baby cotinean levels

To confirm the increasing relationship between the SDP scores and BPM scores, we construct a histogram stratified by the SDP scores where we see that low bpm scores are positively correlated with low sdp scores and high bpm scores are positively correlated with high sdp scores. The same relationship is apparent for the histogram displaying the association between baby cotinine levels and BPM scores

Similar to the SDP score, I created a new variable to score ETS exposure of each participant. I summed up the rows of recorded smoke exposure counts across the different follow-up times. Using this column, I stratified the data to find any possible relationship between smoke exposure(ETS) and adolescent self-regulation, substance use, and externalizing. Table 8 details the relationship between ETS and self-regulation and externalizing. We don't observe any statistically significant relationship between substance use and ETS smoke exposure, except for the variable describing the bpm scores related to externalizing problems on child as shown in table 8.

However in table 9, we do see that there is a generalized increasing relationship between smoke exposure and children's alcoholic and cigarette use. More specifically, it can be said that the alcohol use increased for adolescents with higher smoke exposure. 20% of kids who drank aclohol before have smoke exposure score of 6 while only 10% of kids who drank alcohol have smoke exposure score of 1.

Table 8: **self-regulation, substance use, and externalizing stratified by smoke exposure score**

Variable	N	**Overall**, N = 38	Smoke Exposure Score: **0**, N = 24	**1**, N = 2	**3**, N = 1	**4**, N = 4	**5**, N = 2	**6**, N = 5	p-value
cotimean__34wk	31	44 (86)	24 (74)	1 (0)	329 (NA)	72 (47)	78 (110)	46 (57)	0.18
Unknown		7	5	0	0	1	0	1	
cotimean__pp6mo__baby	30	2.59 (4.28)	2.59 (4.97)	1.07 (1.17)	3.78 (NA)	1.06 (0.68)	7.21 (NA)	2.70 (1.34)	0.37
Unknown		8	3	0	0	2	1	2	
cotimean__pp6mo	30	82 (137)	45 (110)	233 (329)	170 (NA)	131 (152)	354 (NA)	85 (73)	0.20
Unknown		8	3	0	0	2	1	2	
swan__inattentive	38	11 (5)	10 (5)	13 (8)	15 (NA)	17 (3)	11 (9)	9 (5)	0.22
swan__hyperactive	38	8 (6)	7 (6)	9 (8)	20 (NA)	16 (2)	9 (11)	4 (2)	0.094
bpm__att__p	35	2 (2)	2 (2)	1 (1)	5 (NA)	6 (2)	2 (1)	1 (1)	0.10
Unknown		3	2	0	0	1	0	0	
bpm__ext__p	36	2 (3)	1 (2)	1 (0)	3 (NA)	6 (4)	2 (3)	1 (1)	0.046
Unknown		2	2	0	0	0	0	0	
bpm__int__p	38	2 (2)	1 (2)	4 (6)	4 (NA)	6 (3)	1 (0)	2 (2)	0.10
ppmq__parental__knowledge	36	4.26 (0.58)	4.39 (0.50)	3.67 (0.16)	3.78 (NA)	3.81 (0.98)	4.89 (NA)	4.19 (0.46)	0.11
Unknown		2	0	0	0	0	1	1	
ppmq__child__disclosure	36	3.69 (0.68)	3.83 (0.68)	3.40 (0.00)	NA (NA)	3.20 (0.16)	2.60 (NA)	3.72 (0.77)	0.094
Unknown		2	0	0	1	0	1	0	
ppmq__parental__solicitation	33	4.19 (0.74)	4.30 (0.69)	2.90 (0.71)	NA (NA)	4.15 (0.41)	5.00 (NA)	4.05 (0.82)	0.15
Unknown		5	2	0	1	0	1	1	
ppmq__parental__control	36	4.57 (0.96)	4.45 (1.16)	5.00 (0.00)	5.00 (NA)	4.60 (0.57)	5.00 (NA)	4.76 (0.43)	0.82
Unknown		2	1	0	0	0	1	0	
bpm__att__a	37	2 (2)	1 (2)	2 (NA)	6 (NA)	3 (1)	1 (0)	2 (3)	0.12
Unknown		1	0	1	0	0	0	0	
bpm__ext__a	37	1 (2)	1 (2)	0 (0)	1 (NA)	1 (1)	1 (1)	2 (1)	0.42
Unknown		1	0	0	0	0	0	1	
bpm__int__a	38	2 (2)	1 (2)	3 (2)	8 (NA)	3 (2)	2 (1)	1 (1)	0.16
erq__cog__a	38	5.38 (1.32)	5.39 (1.39)	4.33 (2.12)	5.17 (NA)	5.46 (1.30)	6.17 (1.18)	5.43 (1.16)	0.86
erq__exp__a	38	3.46 (1.60)	3.22 (1.36)	2.25 (0.00)	7.00 (NA)	4.06 (1.86)	2.50 (1.06)	4.30 (2.06)	0.25
bpm__att	34	3 (3)	2 (3)	6 (1)	7 (NA)	4 (2)	1 (1)	4 (2)	0.10
Unknown		4	4	0	0	0	0	0	
bpm__ext	34	3 (2)	3 (2)	4 (0)	4 (NA)	3 (2)	4 (0)	3 (2)	0.41
Unknown		4	4	0	0	0	0	0	
bpm__int	32	3 (3)	2 (2)	4 (2)	3 (NA)	3 (4)	4 (2)	5 (4)	0.47
Unknown		6	5	0	0	1	0	0	
erq__cog	33	3.30 (0.90)	3.29 (1.06)	3.17 (0.24)	3.00 (NA)	3.08 (0.55)	3.42 (0.59)	3.63 (0.95)	0.96
Unknown		5	4	0	0	0	0	1	
erq__exp	33	2.80 (0.79)	2.51 (0.70)	2.75 (0.35)	2.75 (NA)	3.25 (0.66)	3.13 (0.88)	3.60 (0.84)	0.15
Unknown		5	4	0	0	1	0	0	
pmq__parental__knowledge	33	3.99 (0.81)	4.19 (0.80)	3.11 (0.63)	NA (NA)	3.78 (0.62)	3.06 (0.86)	4.07 (0.74)	0.077
Unknown		5	4	0	1	0	0	0	
pmq__child__disclosure	34	3.45 (1.02)	3.84 (0.98)	2.10 (0.99)	3.20 (NA)	2.40 (0.69)	3.00 (0.57)	3.48 (0.61)	0.027
Unknown		4	4	0	0	0	0	0	
pmq__parental__solicitation	33	2.97 (1.30)	3.44 (1.23)	1.40 (0.28)	2.00 (NA)	2.90 (1.57)	1.40 (NA)	2.28 (0.78)	0.10
Unknown		5	4	0	0	0	1	0	
pmq__parental__control	31	4.30 (0.95)	4.46 (0.99)	3.30 (0.71)	5.00 (NA)	4.05 (1.00)	3.40 (NA)	4.35 (0.77)	0.22
Unknown		7	5	0	0	0	1	1	
Autism Diagnosis Status	18								>0.99
No		17 / 18 (94%)	10 / 11 (91%)	1 / 1 (100%)	0 / 0 (NA%)	2 / 2 (100%)	1 / 1 (100%)	3 / 3 (100%)	
Diagnosed		0 / 18 (0%)	0 / 11 (0%)	0 / 1 (0%)	0 / 0 (NA%)	0 / 2 (0%)	0 / 1 (0%)	0 / 3 (0%)	
Suspected		1 / 18 (5.6%)	1 / 11 (9.1%)	0 / 1 (0%)	0 / 0 (NA%)	0 / 2 (0%)	0 / 1 (0%)	0 / 3 (0%)	
Unknown		20	13	1	1	2	1	2	

[1] Mean (SD); n / N (%)

[2] Kruskal-Wallis rank sum test; Fisher's exact test

Table 9: **Substance Use stratified by smoke exposure score**

Variable	N	**Overall**, N = 38	**Smoke Exposure Score** **0**, N = 24	**1**, N = 2	**3**, N = 1	**4**, N = 4	**5**, N = 2	**6**, N = 5	p-value
cig_ever	34								>0.99
0		33 / 34 (97%)	19 / 20 (95%)	2 / 2 (100%)	1 / 1 (100%)	4 / 4 (100%)	2 / 2 (100%)	5 / 5 (100%)	
1		1 / 34 (2.9%)	1 / 20 (5.0%)	0 / 2 (0%)	0 / 1 (0%)	0 / 4 (0%)	0 / 2 (0%)	0 / 5 (0%)	
Unknown		4	4	0	0	0	0	0	
e_cig_ever	34								0.81
0		31 / 34 (91%)	18 / 20 (90%)	2 / 2 (100%)	1 / 1 (100%)	4 / 4 (100%)	2 / 2 (100%)	4 / 5 (80%)	
1		3 / 34 (8.8%)	2 / 20 (10%)	0 / 2 (0%)	0 / 1 (0%)	0 / 4 (0%)	0 / 2 (0%)	1 / 5 (20%)	
Unknown		4	4	0	0	0	0	0	
mj_ever	34								>0.99
0		31 / 34 (91%)	17 / 20 (85%)	2 / 2 (100%)	1 / 1 (100%)	4 / 4 (100%)	2 / 2 (100%)	5 / 5 (100%)	
1		3 / 34 (8.8%)	3 / 20 (15%)	0 / 2 (0%)	0 / 1 (0%)	0 / 4 (0%)	0 / 2 (0%)	0 / 5 (0%)	
Unknown		4	4	0	0	0	0	0	
alc_ever	33								0.43
0		29 / 33 (88%)	18 / 20 (90%)	1 / 2 (50%)	0 / 0 (NA%)	4 / 4 (100%)	2 / 2 (100%)	4 / 5 (80%)	
1		4 / 33 (12%)	2 / 20 (10%)	1 / 2 (50%)	0 / 0 (NA%)	0 / 4 (0%)	0 / 2 (0%)	1 / 5 (20%)	
Unknown		5	4	0	1	0	0	0	

[1] n / N (%)

[2] Fisher's exact test

Conclusion

In this report, we conducted missing and exploratory data analysis of the data provided. We found that the data is MAR(missing at random). We showed this thru visuals, examples and tables. We also found significant relationships between SDP and adolescent self-regulation, substance use, and externalizing. This relationship was not as easily visible for ETS, however we did find some relationships between ETS and adolescent substance use.

The main set back of the study was the low number of observations. This could lead to high bias among our results. More importantly, the low number of observations decreased the statistical power of the study. Another set back of the study is having more observations than variables. Often this leads to high data sparsity, which ddidn't allow us to deal with missing data through imputaions.

Code Appendix

Github repo: https://github.com/TabibC0807/Project-1

```
#############
### Setup ###
#############

library(formatR)

knitr::opts_chunk$set(echo = TRUE)
knitr::opts_chunk$set(message = F)
knitr::opts_chunk$set(warning = F)
knitr::opts_chunk$set(fig.align="center")
knitr::opts_chunk$set(fig.width=8, fig.height=6)

###############
### Library ###
###############

library(HDSinRdata)
library(tidyverse)
library(ggplot2)
library(tableone)
library(mice)
library(naniar)
library(gt)
library(gtsummary)
library(kableExtra)
library(tidyverse)

K01BB <- read.csv("~/Downloads/K01BB.csv")
df <- read.csv("~/Downloads/project1 (1).csv") #49 rows, 78 cols
df[df == ''] <- NA

#frequency of variables with same number NA values
missingVar = table(rowSums(sapply(df, is.na)))
as.data.frame(t(as.matrix(missingVar))) %>%
  mutate(' ' = c('# Participants Missing')) %>%
  kableExtra::kbl(caption = 'Frequencies of Total Variables Missing for
                  Each Patient'
                  , booktabs = T
                  , escape = T
                  , align = 'c') %>%
  kableExtra::kable_classic(full_width = F
                            , html_font = 'Cambria'
                            , latex_options = 'HOLD_position') %>%
  add_header_above(c(' ' = 1, "# Variables Missing for Each Participant" = 22)
                   , escape = T)


# Getting the frequency table of how many missing data each person has
missingObs = table(colSums(sapply(df, is.na)))
as.data.frame(t(as.matrix(missingObs))) %>%
```

```r
  mutate(' ' = c('# Variables')) %>%
  kableExtra::kbl(caption = 'Frequencies of Total Observations Missing for
                Each Variable'
                , booktabs = T
                , escape = T
                , align = 'c') %>%
  kableExtra::kable_classic(full_width = F
                          , html_font = 'Cambria'
                          , latex_options = 'HOLD_position') %>%
  add_header_above(c(' ' = 1, '# Observations Missing for Each Variable' = 19)
                 , escape = T)

#df$mom_numcig[1] = 2 #"2 black and miles a day-> smoke every day"
#df$mom_numcig[47] = 0 #none
#df$mom_numcig[37] = 22  #20-25
#df$mom_numcig[5] = NA


#some NA should be 0:

#num_cigs_30
df$num_cigs_30[df$cig_ever == 0] <- 0
#num_e_cigs
df$num_e_cigs_30[is.na(df$e_cig_ever == 0)] <- 0
#num_mj_30
df$num_mj_30[is.na(df$mj_ever == 0)] <- 0
#num_alc_30
df$num_alc_30[is.na(df$alc_ever == 0)] <- 0
#missing data
varMissingProp = miss_var_summary(df)
varMissingProp %>%
  filter(n_miss > 0) %>%
  kableExtra::kbl(caption = 'Missing Data Proportion for Each Variable'
                , booktabs = T
                , escape = T
                , align = 'c'
                , col.names = c('Variable','Observation Missing','Proportion
                                Missing')) %>%
  kableExtra::kable_classic(full_width = F
                          , html_font = 'Cambria'
                          , latex_options = 'HOLD_position')

drop.cols = colnames(df[,colSums(is.na(df))==0])

df %>%
  select(-one_of(drop.cols)) %>%
  gg_miss_var(show_pct = TRUE) + ggtitle("Missing value percentage by variable")

df %>%
  select(-one_of(drop.cols)) %>%
  gg_miss_var(show_pct = TRUE, facet = mom_smoke_pp6mo)  + ggtitle("Missing value percentage stratified
```

```r
#outcome variables: erq's, swans, child's autism, bpm
#explanatory variables: mom's smoking behavior during pregnancy
#split data into demographics(age, sex, race)
df.cat = df
col.cat = c(3:13, 15:19, 22:28, 37:42, 53:62,64,66,68)


col.cat = c(3:13, 15:19, 22:28, 37:42, 53:62,64,66,68)
df.cat[,col.cat] <- lapply(df.cat[,col.cat] , factor)
df.cat$income[6]=250000
df.cat$income = as.numeric(df.cat$income)
df.cat$page = as.numeric(df.cat$page)
df.cat$tage = as.numeric(df.cat$tage)


df.dems =  df.cat %>%
     mutate(
         parent_race = case_when(
             pasian == 1 ~ "Asian",
             paian == 1 ~ "American Indian/Alaska Native",
             pwhite == 1 ~ "White",
             pnhpi ==  1 ~ "Native Hawaiin or Pacific Islander",
             prace_other == 1 ~ "Other"
         )
     ) %>%
  mutate(
         child_race = case_when(
             tasian == 1 ~ "Asian",
             taian == 1 ~ "American Indian/Alaska Native",
             twhite == 1 ~ "White",
             tnhpi ==  1 ~ "Native Hawaiin or Pacific Islander",
             trace_other == 1 ~ "Other"
         )
     )

df.dems.parents = df.dems %>%
  select(c(page, psex,plang, parent_race, employ, income))
df.dems.child = df.dems %>%
  select(c(tage, tsex,language, child_race))

#summary of parent and child demographics
df.dems.parents %>% mutate(Employment_Status = factor(employ, levels = c(0, 1, 2), labels = c("No", "Pa
  mutate(lang = factor(plang, levels = c(0, 1), labels = c("Did not speak another language at home", "Sp
      all_continuous() ~ "{mean} ({sd})",
      all_categorical() ~ "{n} / {N} ({p}%)"
    ), missing_text = "(Missing)") %>% modify_caption("Demographic summary of parents") %>%   bold_labe
  as_kable_extra(booktabs = TRUE) %>%
  kableExtra::kable_styling(latex_options = "scale_down")



df.dems.child %>%  mutate(Sex = factor(tsex, levels = c(0, 1, 2), labels = c("Male", "Female", "Interse
  mutate(lang = factor(language, levels = c(0, 1), labels = c("Did not speak another language at home",
```

```r
      all_continuous() ~ "{mean} ({sd})",
      all_categorical() ~ "{n} / {N} ({p}%)"
    ), type = list(tage  ~ "continuous"),
   missing_text = "(Missing)") %>% modify_caption("Demographic summary of children") %>%    bold_labels
  as_kable_extra(booktabs = TRUE) %>% kableExtra::kable_styling(latex_options = "scale_down")


#SDP(smoking during pregnancy): mom_smoke_16wk, mom_smoke_22wk, mom_smoke_32wk, cotimean_34wk,
#ETS(environmental tobacco smoke): cotimean_pp6mo, cotimean_pp6mo_baby, smoke_exposure_
#self-regulation: erq
#externalizing: bpm
#ADHD: SWAN
#Autism

#SDP

df.two = df.cat
df.two$mom_smoke_16wk <- gsub("^.{0,2}", "", df$mom_smoke_16wk)
df.two$mom_smoke_22wk <- gsub("^.{0,2}", "", df$mom_smoke_22wk)
df.two$mom_smoke_32wk <- gsub("^.{0,2}", "", df$mom_smoke_32wk)
df.smoke =  data.frame(df.two$mom_smoke_16wk, df.two$mom_smoke_22wk, df.two$mom_smoke_32wk)
df.two$smoke_count <- rowSums(df.smoke == "Yes")

df.two$smoke_count <- rowSums(df.smoke == "Yes", na.rm = TRUE)
df.two$smoke_count[38] = NA

comp.df = df.two %>% select(c(smoke_count, childasd,cotimean_34wk:bpm_int_p, ppmq_parental_knowledge:er
comp.df[, 3:ncol(comp.df)] <- lapply(comp.df[, 3:ncol(comp.df)], as.numeric)

comp.table = comp.df %>% mutate(child_asd = factor(childasd, levels = c(0, 1, 2), labels = c("No", "Dia
  tbl_summary(label = list(child_asd ~ "Autism Diagnosis Status"), by = smoke_count,   statistic = list
      all_continuous() ~ "{mean} ({sd})",
      all_categorical() ~ "{n} / {N} ({p}%)"
    ), type = list(ppmq_parental_control     ~ "continuous", bpm_att_a ~ "continuous",  bpm_att_p ~ "co
  add_p(pvalue_fun = ~ style_pvalue(.x, digits = 2)) %>%
  add_overall() %>%
  add_n() %>%
  modify_header(label ~ "**Variable**") %>%
  modify_spanning_header(c("stat_1", "stat_2", "stat_3") ~ "**Number of times mother said 'yes' to smok
  modify_caption("**Self Regulation and externalizing factors stratified by SDP scores**") %>%
  bold_labels() %>%
  as_kable_extra(booktabs = TRUE) %>%
  kableExtra::kable_styling(latex_options = "scale_down")
comp.table
#SDP on substance abuse
sub.sdp = df.two %>% select(c(smoke_count,cig_ever,e_cig_ever,mj_ever,alc_ever))
comp.sub.sdp= sub.sdp %>%
  tbl_summary( by = smoke_count,   statistic = list(
      all_continuous() ~ "{mean} ({sd})",
      all_categorical() ~ "{n} / {N} ({p}%)"
    )) %>%
  add_p(pvalue_fun = ~ style_pvalue(.x, digits = 2)) %>%
  add_overall() %>%
```

```
  add_n() %>%
  modify_header(label ~ "**Variable**") %>%
  modify_spanning_header(c("stat_1", "stat_2", "stat_3") ~ "**SDP Score**") %>%
  modify_caption("**Substance Use stratified by SDP Score**") %>%
  bold_labels() %>%
  as_kable_extra(booktabs = TRUE) %>%
  kableExtra::kable_styling(latex_options = "scale_down")
comp.sub.sdp

ggplot(comp.df, aes(x = bpm_int_a, fill = as.factor(smoke_count))) +
geom_histogram(alpha = 0.6) +  labs(title = 'Histogram of BPM scores on items related to internalizing p
       , x = 'BPM Score'
       , color = 'SDP Smoke Count'
       , fill = 'SDP Smoke Count'
       , caption = 'Figure 3. Association between SDP and BPM scores')+  theme(legend.position = 'bottom
        , plot.title = element_text(hjust = 0.5)
        , plot.caption = element_text(hjust = 0.5))

ggplot(comp.df, aes(x = cotimean_pp6mo_baby, fill = as.factor(smoke_count))) +
geom_histogram(alpha = 0.4) +  labs(title = 'Histogram of baby cotinine'
       , x = 'Urine cotinine (nicotine metabolite) at 6 months postpartum from baby'
       , color = 'SDP Smoke Count'
       , fill = 'SDP Smoke Count'
       , caption = 'Figure 3. Association between SDP and baby cotinean levels')+  theme(legend.position
        , plot.title = element_text(hjust = 0.5)
        , plot.caption = element_text(hjust = 0.5))

#Smoke Exposure(ETS)
df.smoke.exp =  data.frame(df.two$smoke_exposure_6mo, df.two$smoke_exposure_12mo, df.two$smoke_exposure_

df.two$smoke_count_exp <- rowSums(df.smoke.exp == 1, na.rm = TRUE)
df.two$smoke_count_exp[which(!complete.cases(df.smoke.exp))] = NA

comp.df.exp = df.two %>% select(c(smoke_count_exp, childasd,cotimean_34wk:bpm_int_p, ppmq_parental_knowl

comp.table.exp = comp.df.exp %>% mutate(child_asd = factor(childasd, levels = c(0, 1, 2), labels = c("No
  tbl_summary(label = list(child_asd ~ "Autism Diagnosis Status"), by = smoke_count_exp,   statistic = l
      all_continuous() ~ "{mean} ({sd})",
      all_categorical() ~ "{n} / {N} ({p}%)"
    ), type = list(ppmq_parental_control     ~ "continuous", bpm_att_a ~ "continuous",  bpm_att_p ~ "con
  add_p(pvalue_fun = ~ style_pvalue(.x, digits = 2)) %>%
  add_overall() %>%
  add_n() %>%
  modify_header(label ~ "**Variable**") %>%
  modify_spanning_header(c("stat_1", "stat_2", "stat_3") ~ "**Smoke Exposure Score**") %>%
  modify_caption("**self-regulation, substance use, and externalizing stratified by smoke exposure score
  bold_labels() %>%
  as_kable_extra(booktabs = TRUE) %>%
  kableExtra::kable_styling(latex_options = "scale_down")

comp.table.exp

#smoke exposure(ETS) on substance abuse
```

```r
sub.exp = df.two %>% select(c(smoke_count_exp,cig_ever,e_cig_ever,mj_ever,alc_ever))
comp.sub.exp = sub.exp %>%
  tbl_summary( by = smoke_count_exp,   statistic = list(
      all_continuous() ~ "{mean} ({sd})",
      all_categorical() ~ "{n} / {N} ({p}%)"
    )) %>%
  add_p(pvalue_fun = ~ style_pvalue(.x, digits = 2)) %>%
  add_overall() %>%
  add_n() %>%
  modify_header(label ~ "**Variable**") %>%
  modify_spanning_header(c("stat_1", "stat_2", "stat_3") ~ "**Smoke Exposure Score**") %>%
  modify_caption("**Substance Use stratified by smoke exposure score**") %>%
  bold_labels() %>%
  as_kable_extra(booktabs = TRUE) %>%
  kableExtra::kable_styling(latex_options = "scale_down")
comp.sub.exp
```