Grade: 95
Well done! Super clear! Just need to add the unit of observation for the dataset (-5pts)

Empirical Research Project

Alexia Ge, Bianca Lewis, Larry Lo, Miles Ma, Nick Araya

Project title: The effect of caffeine consumption on students' school success measured by grade point average.

Research question: The research question that we want to investigate is the relationship between student caffeine consumption (intensive and extensive consumption) and their educational success, measured by grade point average.

Motivation: This question is interesting for us to investigate because we know that a large majority of students consume caffeine on a daily basis, which is why we want to understand the relationship that may exist (if any) between caffeine and educational success.

Sample: Our variables are listed below, all of which the data came from the "High School Risk Survey (2006)" with high school students as the unit. Units for each sample and number of observations are individually addressed by variable.

Addressing feedback: For the time ordering/reverse causality issue, we think that it has potential effects on causality, and we intend to address this issue by performing an additional regression analysis with caffeine consumption as the dependent variable against GPA. We will compare the coefficients of both regressions and investigate whether caffeine consumption has a stronger effect on GPA or vice versa.

Continuous Variables:

- **Q6: Grade average, main continuous variable named *gpa***

With *q6* being categorical, we transformed the variable (according to feedback) into a continuous variable by assigning 4.0 scale values to each option given in the questionnaire. Specifically, 4.0 is assigned to "mostly A's", 3.0 to "mostly B's", 2.5 to "B's and C's", 1.5 to "C's and D's", and 0.5 to "D's or lower". We choose this variable as our dependent variable to investigate our question- the effect of caffeine consumption (extensive and intensive) on high school students' grade average. Due to the large sample size that we have, the grade average point is roughly normally distributed as shown by the histogram in Graph 1.

Below is the summary of *gpa's* mean and standard deviation:

Variable	Obs	Mean	Std.	Min	Max
gpa	3,294	2.518519	0.6088901	0.5	3

Below is the tabulation of *gpa*:

gpa	Frequency	Percentage	Cumulative Percentage
0.5	109	2.49	2.49
1.5	431	9.84	12.33
2.5	1,334	30.45	42.78
3	1,420	32.41	75.19
4	1,087	24.81	100
Total	4,381	100	

Graph 2 is a compared histogram of *gpa* conditional on *genderS*, where male students are on the left and females on the right. Both distributions appear to be roughly normally distributed.

- **Q11: Height in inches, continuous variable named *heightinch***

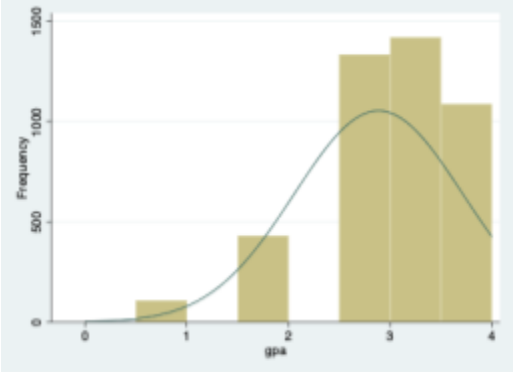
This variable was converted to inches using feet + inches*12; and was chosen to see if height, along with weight, are confounding variables that affect student's metabolism and therefore caffeine consumption. Graph 3 is a histogram of height. The graph shows it is approximately normally distributed. Below is a summary of *heightinch*:

Variable	Obs	Mean	Std.	Min	Max
heightinch	4,284	66.4064	4.491761	12	106

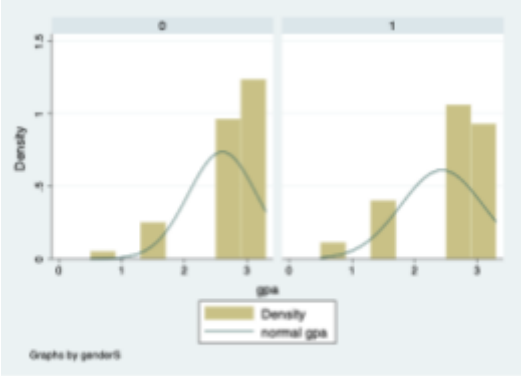
- **Q12: Weight in pounds, continuous variable named *weightpds***

This variable is measured in pounds and we do not have to do transformations. The variable was chosen for the same reason as height: to better understand student's metabolism and caffeine consumption. Graph 4 is a histogram of weight. The distribution of *weightpds* is shown in the tabulated table below:

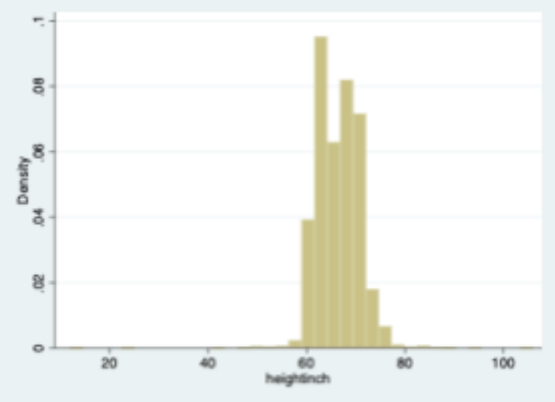
Variable	Obs	Mean	Std.	Min	Max
weightpds	4,187	144.5703	39.42241	55	971



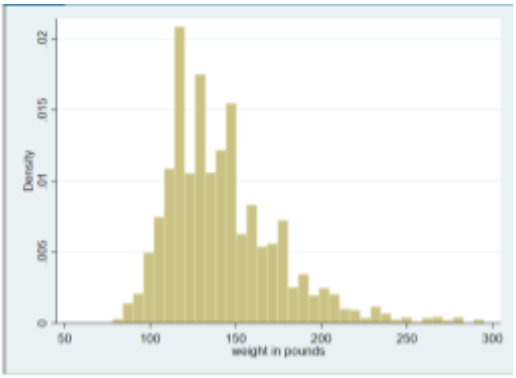
Graph 1 - GPA



Graph 2 - GPA conditional on gender



Graph 3 - height



Graph 4 - Weight

Categorical Variables:

- **Q1: Age, categorical variable named *ageyrs***

We assigned each category of age to an indicator variable, and created seven new indicator variables for different age groups. These variables include *ageyrs_less14, ageyrs_14, ageyrs_15, ageyrs_16, ageyrs_17, ageyrs_18,* and *ageyrs_more19.* The motivation for choosing this variable is for both geographic reasons and regression analysis. We suspect that age may share a positive relationship with caffeine consumption, because caffeine is typically consumed more by older students who have more school work. The distribution of *ageyrs* is shown in the tabulated table below:

ageyrs	Frequency	Percentage	Cumulative Percentage
<14	114	2.52	2.52

14	506	11.19	13.71
15	1,200	26.53	40.24
16	1,249	27.61	67.85
17	978	21.62	89.48
18	406	8.98	98.45
>19	58	1.28	99.73
.	12	0.27	100.00
Total	4,523	100.00	

Below is a summary of *ageyrs*' mean and standard deviation:

Variable	Obs	Mean	Std.	Min	Max
ageyrs	4,511	3.869209	1.271976	1	7

- **Q10: Income Level, categorical variable named *dincome***

This variable measures the income of the family of each questionnaire respondent. We assigned each category of income to an indicator variable, creating the following four new indicator variables: *dincome_welf, dincome_low, dincome_middle, dincome_high*. There was also the possibility for respondents to answer that they were unsure, represented by the data simply being left out of the data set instead of being assigned a value. These variables represent respondents with families who are on welfare, or have low, medium, and high incomes respectively. We expect income to have a positive association with both caffeine consumption as well as GPA, as families with higher income can afford more coffee, and can afford tutoring and other means that would help increase a student's GPA. The distribution of *dincome* is shown in the table below:

dincome	Frequency	Percentage	Cumulative Percentage
welf	90	2.06	2.06
low	218	5.00	7.06
middle	2,190	50.23	57.29
high	525	12.04	69.33
unsure	1,337	30.67	100.00
Total	4,360	100.00	

As you can see, 30.67% of respondents indicated that they were unsure of their family's income. It is important to note this information, because if a majority of respondents had this response then it would weaken the significance of the data analysis.

~~Below is a summary of *dincome*'s mean and standard deviation~~:

Variable	Obs	Mean	Std.	Min	Max
dincome	4,360	3.642431	1.033012	1	5

- **Q91: Caffeine Consumption: Intensive consumption, categorical variable named *caf_inten***

The variable *caf_inten* (added according to feedback) measures the number of standard caffeine drinks consumed per day. No drinks = 1, One to two caffeinated drinks per day = 2, three to four =3, five to six = 4, and more than six drinks per day is represented by 5. The intensity of caffeine consumption variable will be used to compare how increasing caffeine consumption is correlated with other variables such as GPA. Below is a summary of *caf_inten*:

~~Variable~~	~~Obs~~	~~Mean~~	~~Std.~~	~~Min~~	~~Max~~
caf_inten	4,168	2.205854	.9957164	1	5

Below is a tabulation of *caf_inten:*

caf_inten	Frequency	Percentage	Cumulative Percentage

1	855	20.51	20.51
2	2,239	53.72	74.23
3	666	15.98	90.21
4	177	4.25	94.46
5	231	5.54	100.00
Total	4,168	100.00	

- **Q66: Other Substances (Marijuana, Alcohol, Other Drugs)**
    - **Frequency of Marijuana usage is a categorical variable (*dmarjfreq*)**, representing the number of days a week where the student uses Marijuana. (Never=1, once/week=2, twice=3, 3-6= 4, daily= 5)

dmarjfreq	Frequency	Percentage	Cumulative Percentage
1	885	46.78	46.78
2	225	11.89	58.67
3	191	10.1	68.76
4	234	12.37	81.13
5	357	18.87	100
Total	1892	100	

    - **Q74: Frequency of Alcohol consumption is a categorical variable (*dalcfreq*)**, representing the number of days a month where the student drinks at least 1 drink of alcohol. (0= 1, 1-2= 2, 3-5= 3, 6-9= 4, 10-19= 5, 20-29= 6, all 30= 7)

dalcfreq	Frequency	Percentage	Cumulative Percentage
1	1097	35.18	35.18
2	804	25.79	60.97
3	457	14.66	75.63
4	325	10.42	86.05
5	269	8.63	94.68
6	61	1.96	96.63
7	105	3.37	100
Total	3118	100	

    - **Q75: Intensity of Alcohol consumption is a categorical variable (*dalcinten*)**, representing the number of days in a month where the student drinks 5 or more drinks in a row. (0= 1, 1-2= 2, 3-5= 3, 6-9= 4, 10-19= 5, 20-29= 6, every day= 7)

dalcinten	Frequency	Percentage	Cumulative Percentage
1	1769	57.23	57.23
2	584	18.89	76.12
3	310	10.03	86.15
4	209	6.76	92.91
5	118	3.82	96.73
6	35	1.13	97.86
7	66	2.14	100
Total	3091	100	

    - **Q82: Frequency of Designer drug usage and other drugs is a categorical variable (*dothdrug*)**, representing the number of days a week where the student uses some form of drug. (never= 1, once/week= 2, twice= 3, 3-6= 4, daily= 5)

dothdrug	Frequency	Percentage	Cumulative Percentage
1	378	64.73	64.73
2	47	8.05	72.77
3	42	7.19	79.97
4	32	5.48	85.45
5	85	14.55	100
Total	584	100	

The 4 variables in “Other Substances" can be lumped together since they are used to measure the frequency/intensity of some form of substance use. Our reason for choosing to include these variables is because there is a possibility that both student GPA and caffeine consumption are tied to the usage of these other substances, and analyzing them will allow us to account for confounders. For example, frequent consumption of alcohol and other drugs may result in a lower GPA, or perhaps substance use and caffeine consumption is correlated, due to the similar way these stimulants affect the brain and cause addiction. It should be noted that the survey sample for Designer Drug use was rather small; 584 compared to 2000-3000 for similar variables, likely because of students being reluctant to answer due to the controversial nature of these substances. Thus, the results from this variable’s (*dothdrug*) tests should be taken with a grain of salt.

- **Q130: Number of hours playing video computer games, categorical variable named *dvidgam***

We assigned each category of time length playing video games to an indicator variable. Thus, we now have five new indicator variables representing different time lengths: *dvidgam_no, dvidgam_less7, dvidgam_7to14, dvidgam_15to20, a*nd *dvidgam_more21.* We suspect that the number of hours playing video computer games may share a positive relationship with the amount of caffeine consumption, because more game time may reduce total sleep time, and so the student may need caffeine to stay focused. The distribution of *dvidgam* is shown in the tabulated table below:

dvidgam	Frequency	Percentage	Cumulative Percentage
1	1964	48.76	48.76
2	1262	31.33	80.09
3	385	9.56	89.65
4	192	4.77	94.41
5	225	5.59	100
Total	4028	100	

- **Q144: Growth and Maturation**, **categorical variable named *dgrowth***

This variable measures the growth and maturation of the respondents. We assigned each category of growth to an indicator variable: *dgrowth_no, dgrowth_bstart, dgrowth_udwy, dgrowth_compl*. These variables represent respondents whose growth spurt has not yet started, has barely started, is underway, and is completed. We have chosen this variable to help us understand the relationship between growth and caffeine consumption; we predict a positive correlation because caffeine is more often consumed by adults than adolescents. The distribution of *dgrowth* is shown in the table below:

dgrowth	Frequency	Percentage	Cumulative Percentage
not started	166	5.21	5.21
barely started	260	8.16	13.37
underway	915	28.71	42.08
completed	1,846	57.92	100.00
Total	3,187	100.00	

Below is a summary of *dgrowth*'s mean and standard deviation:

Variable	Obs	Mean	Std.	Min	Max
dgrowth	3,187	3.393473	0.8453158	1	4

Indicator Variables:

- **Q2: Gender, indicator variable named *genderS***

This variable equals 1 if the student is male and equals 0 if otherwise. This variable is chosen for both geographical reasons to ensure that the sample that we have does not constitute selection bias, and regression analysis on the potential correlation among gender, grade average, and caffeine consumption.

Variable	Observations	Mean	Std.	Min	Max
genderS	4,469	0 .4752741	0.4994441	0	1

- **Q91: Caffeine Consumption: Extensive consumption, indicator variable named *caf_exten***

This extensive measure of consumption was added according to feedback, where c*af_exten* =1 if the person drinks any caffeine in a day, and =0 otherwise. The purpose of having this extensive measure of caffeine consumption helps when comparing to other variables to see if the presence of any caffeine consumption is correlated with GPA.

Below is a summary of *caf_exten:*

Variable	Obs	Mean	Std.	Min	Max
caf_exten	4,523	.8109662	.3915788	0	1

- **Q8: Paid part time job, indicator variable named *ptj***

This dataset originally assigned 1 to yes, and 2 to no, but we make it 0 when it equals to no, as it is now transformed into a dummy variable and is easier to analyse. We suspect that students doing part-time jobs may have much heavier workloads, and they may have to consume more caffeine to concentrate. This variable is chosen to analyse the potential correlation between paid part time job and caffeine consumption. The summary of *ptj* is shown below:

Variable	Obs	Mean	Std.	Min	Max
ptj	4419	.4039375	.4907408	0	1

Statistical Tests

- **GPA v. Income Level**

In this statistical test, we compare the difference in GPA among students with different income levels (welfare, low, middle, and high). In each of the following tables, 1 denotes GPA of groups within the specific income level, and 0 denotes GPA of everyone in other income levels and those who answered that they are not certain about their family income. In all four cases, p-value for the null hypothesis that the difference in GPA equals to 0 is approximately 0, indicating that we can reject the null hypothesis and there is a statistically significant difference between these two groups' GPA at 95% confidence level.

dincome_wel	Obs	Mean	SE	Std.	[95% CI]	
0	3,224	2.532723	0.0104112	0.5911483	2.51231	2.553136
1	70	1.864286	0.1150756	0.9627912	1.634716	2.093855
combined	3,294	2.518519	0.0106091	0.6088901	2.497717	2.53932
diff		0.6684376	0.072645		0.5260036	0.8108716
diff = mean(0) - mean(1)			t = 9.2014			
Ho: diff = 0			degrees of freedom = 3292			
Ha: diff < 0		Ha: diff != 0		Ha: diff > 0		
Pr(T < t) =1.0000		Pr(|T| > |t|) = 0.0000		Pr(T > t) = 0.0000		

dincome_low	Obs	Mean	SE	Std.	[95% CI]	
0	3,126	2.528151	0.0107124	0.5989385	2.507147	2.549155
1	168	2.339286	0.0579325	0.7508905	2.224911	2.45366
combined	3,294	2.518519	0.0106091	0.6088901	2.497717	2.53932
diff		0.1888653	0.0481176		0.0945219	0.2832087
diff = mean(0) - mean(1)			t = 3.9251			
Ho: diff = 0			degrees of freedom = 3292			
Ha: diff < 0		Ha: diff != 0		Ha: diff > 0		
Pr(T < t) =1.0000		Pr(|T| > |t|) = 0.0001		Pr(T > t) = 0.0000		

dincome_middle	Obs	Mean	SE	Std.	[95% CI]	
0	1,746	2.447308	0.0157954	0.6600132	2.416328	2.478288
1	1,548	2.598837	0.0135838	0.5344487	2.572193	2.625482
combined	3,294	2.518519	0.0106091	0.6088901	2.497717	2.53932
diff		-0.1515291	0.0210951		-0.19289	-0.1101682
diff = mean(0) - mean(1)			t = -7.1831			
Ho: diff = 0			degrees of freedom = 3292			
Ha: diff < 0		Ha: diff != 0		Ha: diff > 0		
Pr(T < t) =0.0000		Pr(|T| > |t|) = 0.0000		Pr(T > t) = 1.0000		

dincome_high	Obs	Mean	SE	Std.	[95% CI]	
0	2,937	2.508001	0.0113307	0.6140559	2.485784	2.530218
1	357	2.605042	0.0295279	0.5579128	2.546971	2.663113
combined	3,294	2.518519	0.0106091	0.6088901	2.497717	2.53932
diff		-0.0970407	0.0340916		-0.1638834	-0.0301979
diff = mean(0) - mean(1)			t = -2.8465			
Ho: diff = 0			degrees of freedom = 3292			
Ha: diff < 0		Ha: diff != 0		Ha: diff > 0		
Pr(T < t) =0.0022		Pr(|T| > |t|) = 0.0044		Pr(T > t) =0.9978		

- **GPA v. Video Games**

In this statistical test, we compare the difference in GPA between students who play video games and those who do not. As shown in the following t-test table, the average GPA for those who played video games (*dvidgam_no*=1) is 2.57, and those who do not play video games (*dvidgam_no*=0) is 2.48. P-value for the null hypothesis that the difference in GPA equals to 0 is 0, indicating that we can reject the null hypothesis and there is a statistically significant difference between these two groups' GPA.

Group	Obs	Mean	SE	Std.	[95% CI]	
0	1,925	2.481818	0.0144315	0.6331798	2.453515	2.510121
1	1,369	2.570124	0.0153847	0.5692331	2.539944	2.600304
combined	3,294	2.518519	0.0106091	0.6088901	2.497717	2.53932
diff		-0.088306	0.0214752		-0.1304121	-0.0461999
diff = mean(0) - mean(1)			t = -4.1120			
Ho: diff = 0			degrees of freedom = 3292			
Ha: diff < 0		Ha: diff != 0		Ha: diff > 0		
Pr(T < t) = 0.0000		Pr(|T| > |t|) = 0.0000		Pr(T > t) = 1.0000		