Fanzhong Cao, Yaning Ling, Han Shi
STAT:2010:0AAA Spr17
4/29/2017

What affects our grades in high school?

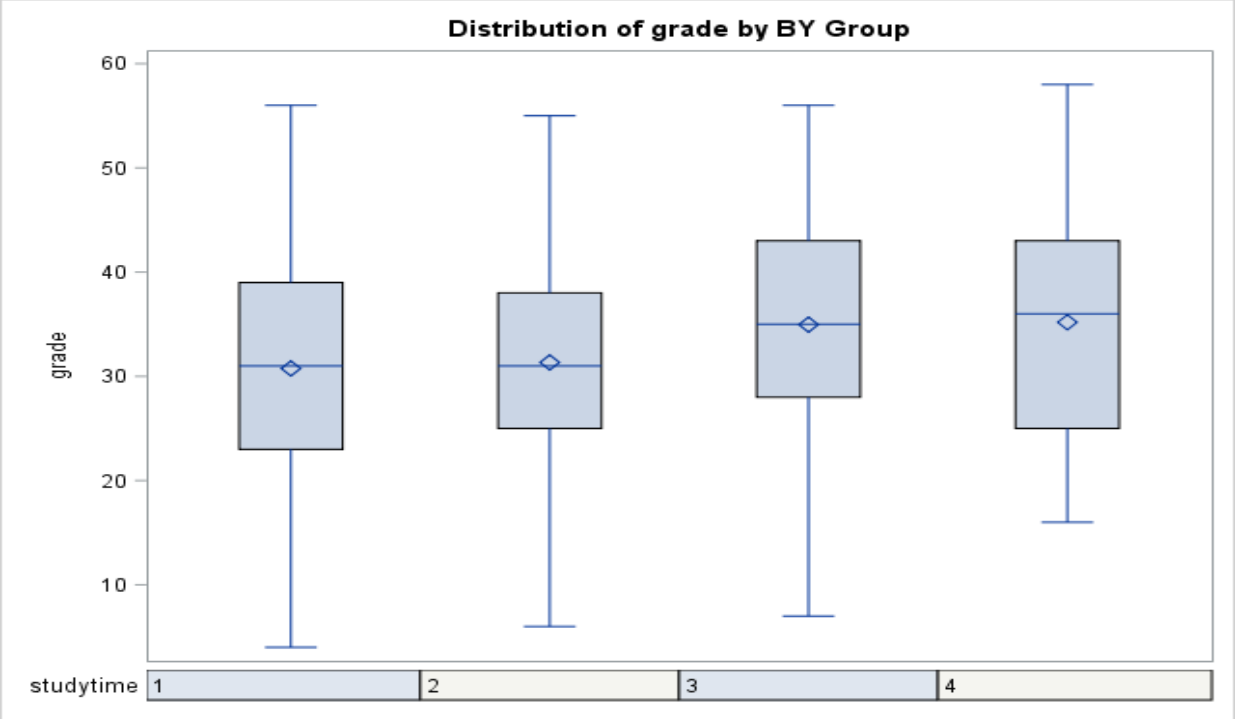
In our project, we analyzed a dataset containing information about 395 high school aged students. The data were obtained in a survey of students in math classes, and it contains information about their study habits, alcohol consumption, and attendance consistency. We found the dataset on kaggle.com and we modified it to fit our study. Our research question is: "How do study time, alcohol consumption and absences affect academic performance in high school?".

- ***Grades and study time***

SAS Code:

```
data study ;
input grade studytime absence au ;
datalines ;
*datalines here
;
run ;

proc sort data = study ;
by studytime ;
run ;

proc univariate plot data = study ;
var grade ;
by studytime ;
run ;

proc means data = study ;
var grade ;
by studytime ;
run ;

proc anova data = study ;
class studytime ;
model grade = studytime ;
run ;
```

SAS Results:



studytime=1

Analysis Variable : grade				
N	Mean	Std Dev	Minimum	Maximum
105	30.7619048	12.4077549	4.0000000	56.0000000

studytime=2

Analysis Variable : grade				
N	Mean	Std Dev	Minimum	Maximum
198	31.3282828	10.1400502	6.0000000	55.0000000

studytime=3

Analysis Variable : grade				
N	Mean	Std Dev	Minimum	Maximum
65	34.9538462	10.9664699	7.0000000	56.0000000

studytime=4

Analysis Variable : grade				
N	Mean	Std Dev	Minimum	Maximum
27	35.1851852	11.4456476	16.0000000	58.0000000

The ANOVA Procedure

Class Level Information		
Class	Levels	Values
studytime	4	1 2 3 4

Number of Observations Read	395
Number of Observations Used	395

The ANOVA Procedure

Dependent Variable: grade

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1090.78553	363.59518	3.00	0.0305
Error	391	47369.64485	121.14999		
Corrected Total	394	48460.43038			

R-Square	Coeff Var	Root MSE	grade Mean
0.022509	34.35553	11.00682	32.03797

Source	DF	Anova SS	Mean Square	F Value	Pr > F
studytime	3	1090.785532	363.595177	3.00	0.0305

Bonferroni (Dunn) t Tests for grade

Note: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than Tukey's for all pairwise comparisons.

Alpha	0.05
Error Degrees of Freedom	391
Error Mean Square	121.15
Critical Value of t	2.65175

Comparisons significant at the 0.05 level are indicated by ***.

studytime Comparison	Difference Between Means	Simultaneous 95% Confidence Limits	
4 - 3	0.231	-6.451	6.914
4 - 2	3.857	-2.131	9.845
4 - 1	4.423	-1.875	10.721
3 - 4	-0.231	-6.914	6.451
3 - 2	3.626	-0.547	7.798
3 - 1	4.192	-0.415	8.798
2 - 4	-3.857	-9.845	2.131
2 - 3	-3.626	-7.798	0.547
2 - 1	0.566	-2.957	4.090
1 - 4	-4.423	-10.721	1.875
1 - 3	-4.192	-8.798	0.415
1 - 2	-0.566	-4.090	2.957

Interpretation: The numbers 1 to 4 represent different levels of study time: 1-<2 hours, 2-2 to 5 hours, 3-5 to 10 hours, and 4->10 hours. From the side-by-side boxplots of grade by study time, we can observe that the median grades increase with more study time. The minimum grade of each study time level increases with study time, and maximum grade of the fourth level of study time is the highest of the whole dataset. There is some evidence suggesting that students who study for more hours tend to get higher grades.

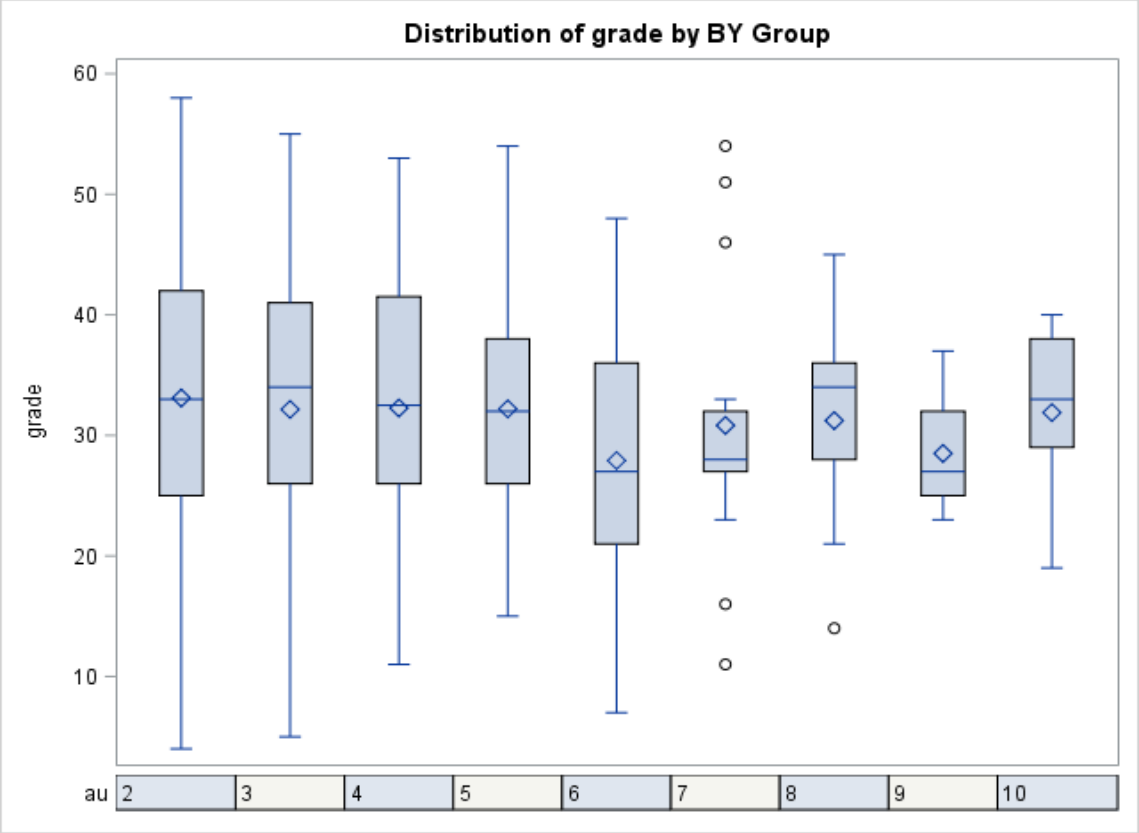
We did the ANOVA test in SAS, and the P-value is 0.03. Therefore, we can conclude that the mean grades of each study time level differs at a 0.05 significance level. By observing the difference between mean grades for each level of study time, we conclude that on average, students' grades increase as study time increases.

- ***Grades and alcohol consumption***

SAS Code:

```
proc sort data = study ;
by au ;
run ;

proc univariate plot data = study ;
var grade ;
by au ;
run ;

proc means data = study ;
var grade ;
by au ;
run ;

proc anova data = study ;
class au ;
model grade = au ;
run ;
```

SAS Results:



The MEANS Procedure

alcohol=2

Analysis Variable : grade				
N	Mean	Std Dev	Minimum	Maximum
150	33.1000000	12.1460630	4.0000000	58.0000000

alcohol=3

Analysis Variable : grade				
N	Mean	Std Dev	Minimum	Maximum
66	32.1515152	11.9463447	5.0000000	55.0000000

alcohol=4

Analysis Variable : grade				
N	Mean	Std Dev	Minimum	Maximum
60	32.2666667	10.3102165	11.0000000	53.0000000

alcohol=5

Analysis Variable : grade				
N	Mean	Std Dev	Minimum	Maximum
45	32.2000000	8.7791074	15.0000000	54.0000000

alcohol=6

Analysis Variable : grade				
N	Mean	Std Dev	Minimum	Maximum
35	27.9142857	10.1209492	7.0000000	48.0000000

alcohol=7

Analysis Variable : grade				
N	Mean	Std Dev	Minimum	Maximum
17	30.8235294	10.9899687	11.0000000	54.0000000

alcohol=8

Analysis Variable : grade				
N	Mean	Std Dev	Minimum	Maximum
9	31.2222222	9.5233631	14.0000000	45.0000000

alcohol=9

Analysis Variable : grade				
N	Mean	Std Dev	Minimum	Maximum
4	28.5000000	5.9721576	23.0000000	37.0000000

alcohol=10

Analysis Variable : grade				
N	Mean	Std Dev	Minimum	Maximum
9	31.8888889	7.4236858	19.0000000	40.0000000

The SAS System

The ANOVA Procedure

Class Level Information		
Class	Levels	Values
alcohol	9	2 3 4 5 6 7 8 9 10

Number of Observations Read	395
Number of Observations Used	395

The SAS System

The ANOVA Procedure

Dependent Variable: grade

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	850.85431	106.35679	0.86	0.5485
Error	386	47609.57607	123.34087		
Corrected Total	394	48460.43038			

R-Square	Coeff Var	Root MSE	grade Mean
0.017558	34.66478	11.10589	32.03797

Source	DF	Anova SS	Mean Square	F Value	Pr > F
alcohol	8	850.8543081	106.3567885	0.86	0.5485

Interpretation: The numbers 2 to 10 represent different levels of alcohol consumption: 2-very low to 10-very high. From the side-by-side boxplots of grade by alcohol consumption, there are outliers of the seventh level of alcohol use. If we ignore the seventh level, we can see the maximum grade of each box is roughly decreasing as alcohol consumption increases. The 75% quartile of boxes with higher alcohol consumption tend to have lower grades. Therefore, we can say that there is some evidence suggesting that students with less alcohol consumption are more likely to achieve better grades.

We also did the ANOVA test in SAS, and the P-value is 0.54. Therefore, there is not enough evidence to state that the mean grades of each alcohol consumption level differ significantly.

❖ *Grades and absences*

SAS Code:

```
proc corr data = study ;
var absence grade ;
run ;

proc reg data = study ;
model grade = absence ;
run ;

proc means data = study ;
var grade ;
by absence ;
run ;

proc anova data = study ;
class absence ;
model grade = absence ;
run ;
```

SAS Results:

The SAS System

The CORR Procedure

2 Variables:	absence grade

Simple Statistics

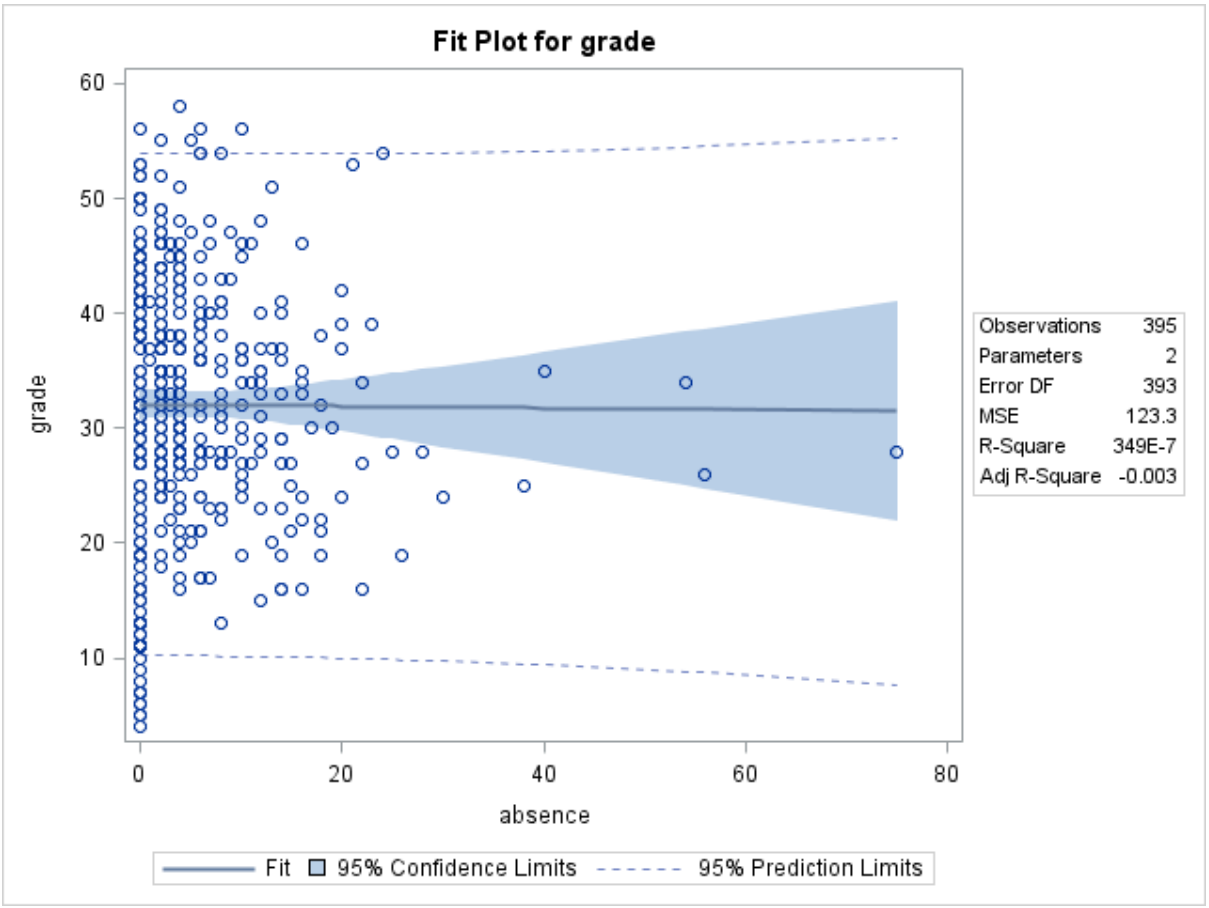
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
absence	395	5.70886	8.00310	2255	0	75.00000
grade	395	32.03797	11.09036	12655	4.00000	58.00000

Pearson Correlation Coefficients, N = 395
Prob > |r| under H0: Rho=0

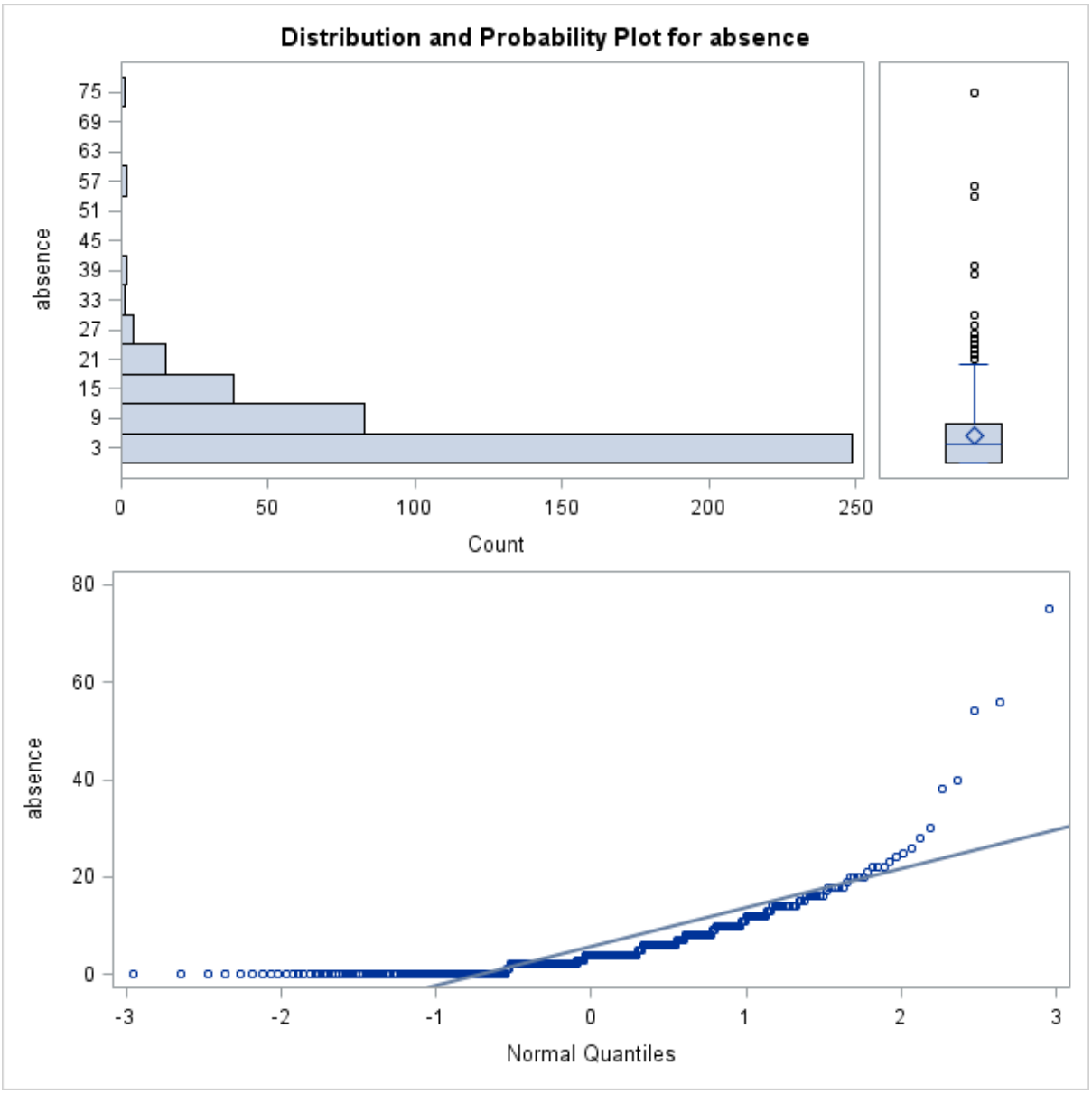
	absence	grade
absence	1.00000	-0.00591 0.9068
grade	-0.00591 0.9068	1.00000

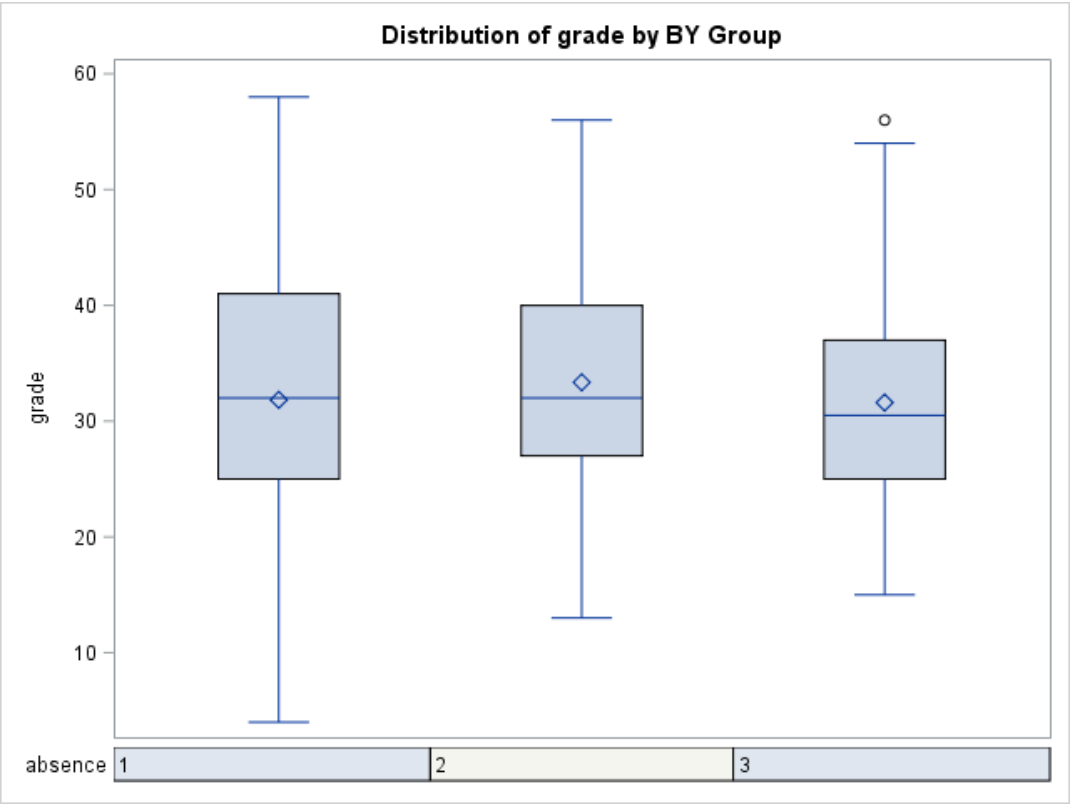
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	32.08472	0.68659	46.73	<.0001
absence	1	-0.00819	0.06990	-0.12	0.9068



Quantiles (Definition 5)	
Level	Quantile
100% Max	75
99%	40
95%	19
90%	14
75% Q3	8
50% Median	4
25% Q1	0
10%	0
5%	0
1%	0
0% Min	0


Distribution and Probability Plot for absence
absence
75
69
63
57
51
45
39
33
27
21
15
9
3
0
50
100
150
200
250
Count
absence
80
60
40
20
0
-3
-2
-1
0
1
2
3
Normal Quantiles



The MEANS Procedure

absence=1

Analysis Variable : grade				
N	Mean	Std Dev	Minimum	Maximum
244	31.8401639	11.8704888	4.0000000	58.0000000

absence=2

Analysis Variable : grade				
N	Mean	Std Dev	Minimum	Maximum
65	33.3538462	10.1971668	13.0000000	56.0000000

absence=3

Analysis Variable : grade				
N	Mean	Std Dev	Minimum	Maximum
86	31.6046512	9.3472601	15.0000000	56.0000000

The ANOVA Procedure

Class Level Information		
Class	Levels	Values
absence	3	1 2 3

Number of Observations Read	395
Number of Observations Used	395

The ANOVA Procedure

Dependent Variable: grade

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	138.24431	69.12215	0.56	0.5712
Error	392	48322.18607	123.27088		
Corrected Total	394	48460.43038			

R-Square	Coeff Var	Root MSE	grade Mean
0.002853	34.65494	11.10274	32.03797

Source	DF	Anova SS	Mean Square	F Value	Pr > F
absence	2	138.2443083	69.1221542	0.56	0.5712

Interpretation: After doing the correlation in SAS, we find that there is a weak negative relationship between absences and grades, so we divided the number of absences into three levels(1- 0 to 4 days, 2- 5 to 8 days, 3- more than 8 days) according to the quantiles of absence data. From the side-by-side boxplots of grade by absence, the maximum grade of each box decreases with the increase of absences. Also, the 75% quartile of higher level of absences decreases. It seems that students with fewer absence days are more likely to get good grades.However, after doing the ANOVA test in SAS, we find that the P-value is 0.57. Moreover, we observe that the mean grade for absence=1, the mean grade for absence=2, and the mean grade for absence=3 are respectively 31.84, 33.35, and 31.60: they do not differ significantly. Therefore, we can conclude that there is not enough evidence to state that the amount of absences affect the grades received by the students.

Link to the original dataset: https://www.kaggle.com/uciml/student-alcohol-consumption