CSL 524

Data Mining

Assignment 1: Statistics and Plotting

Task-1

The mean value and standard deviation for the attributes G1, G2, G3 are as follow:

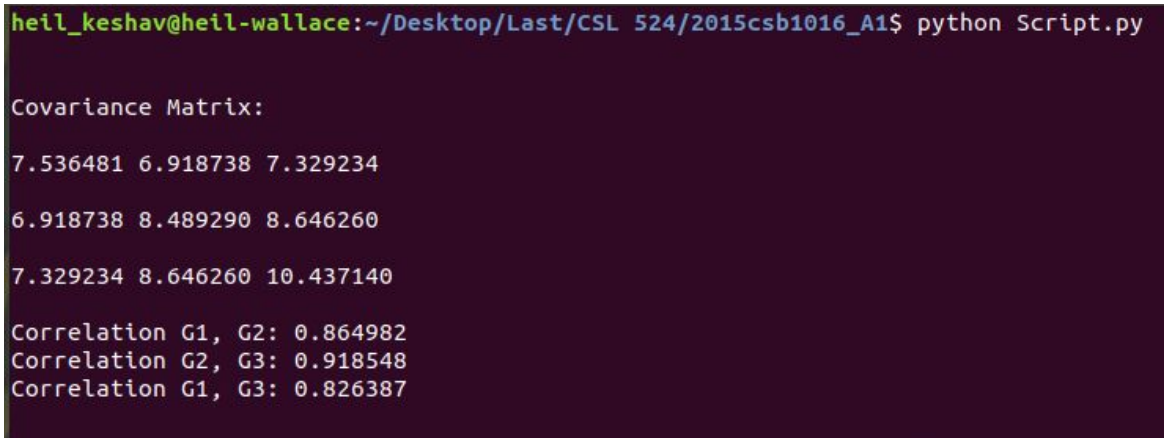
```
heil_keshav@heil-wallace:~/Desktop/Last/CSL 524/2015csb1016_A1$ python Script.py


Mean G1: 11.399076
Mean G2: 11.570108
Mean G3: 11.906009


Standard deviation G1: 2.745265
Standard deviation G2: 2.913639
Standard deviation G3: 3.230656
```

Mean represents the average and is computed as the sum of all the observed outcomes from the dataset divided by the total number of data entries (row).

The standard deviation is a description of the data's spread, how widely it is distributed about the mean.

Task-2

The covariance matrix for attributes G1, G2, and G3 and the correlation between each pair:



The upper half of covariance matrix is same as lower half, Since Cov(A, B) = Cov(B, A)

You can use the covariance to determine the direction of a linear relationship between two variables as follows:

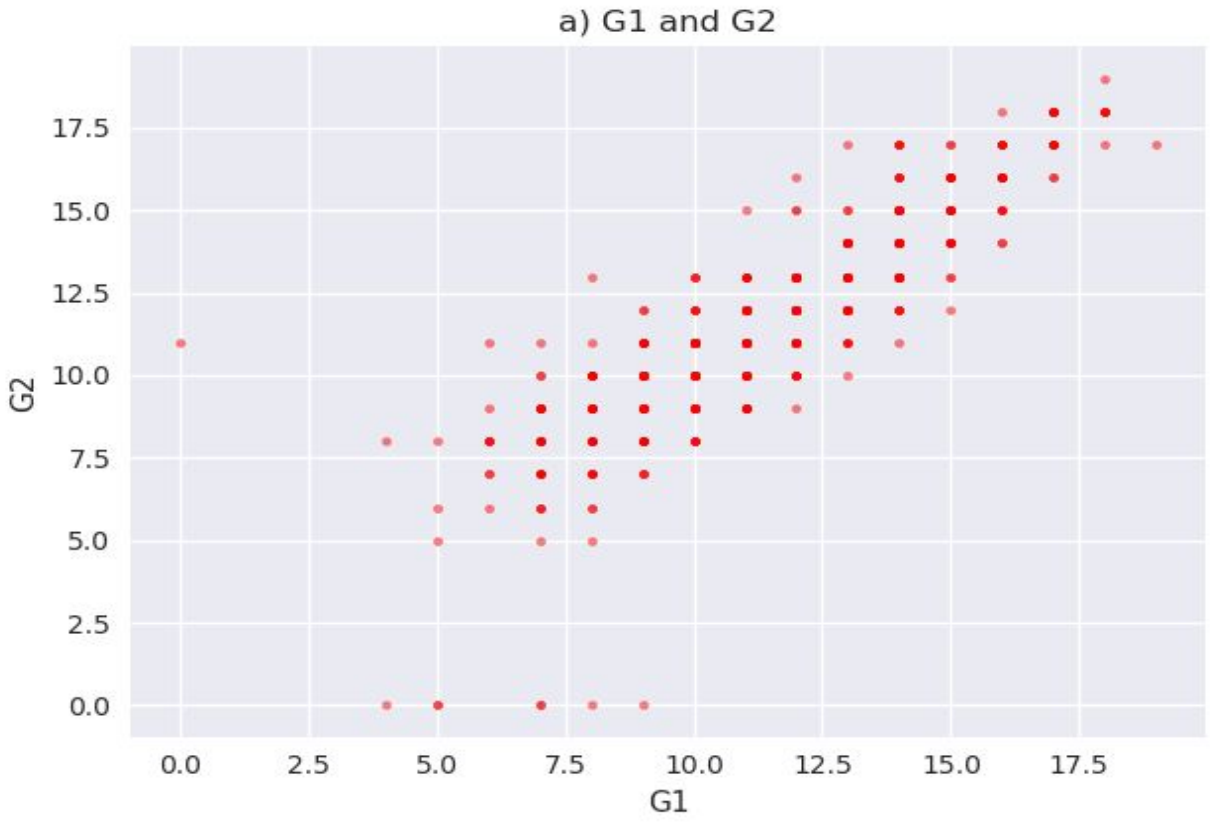
- If both variables tend to increase or decrease together, the coefficient is positive.
- If one variable tends to increase as the other decreases, the coefficient is negative.

To assess the strength of a relationship between two variables using a standardized scale of -1 to +1, use Correlation. Here, the values are positive and also correlation coefficient value is close to +1 which shows a strong uphill (positive) linear relationship. So, if someone has high grade in period G1, then there is strong probability that his second period/overall grade will be high too.

Task-3

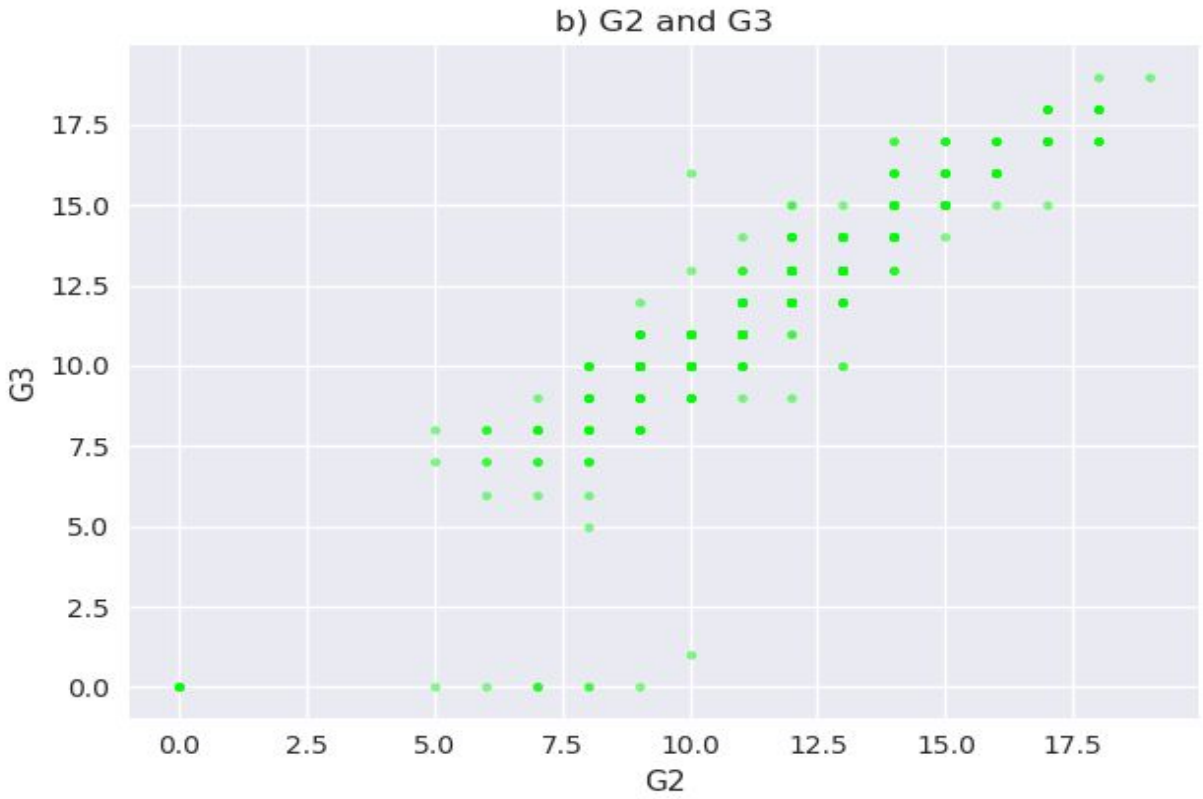
Scatter plots for the pairs of attributes:

a) G1 and G2,



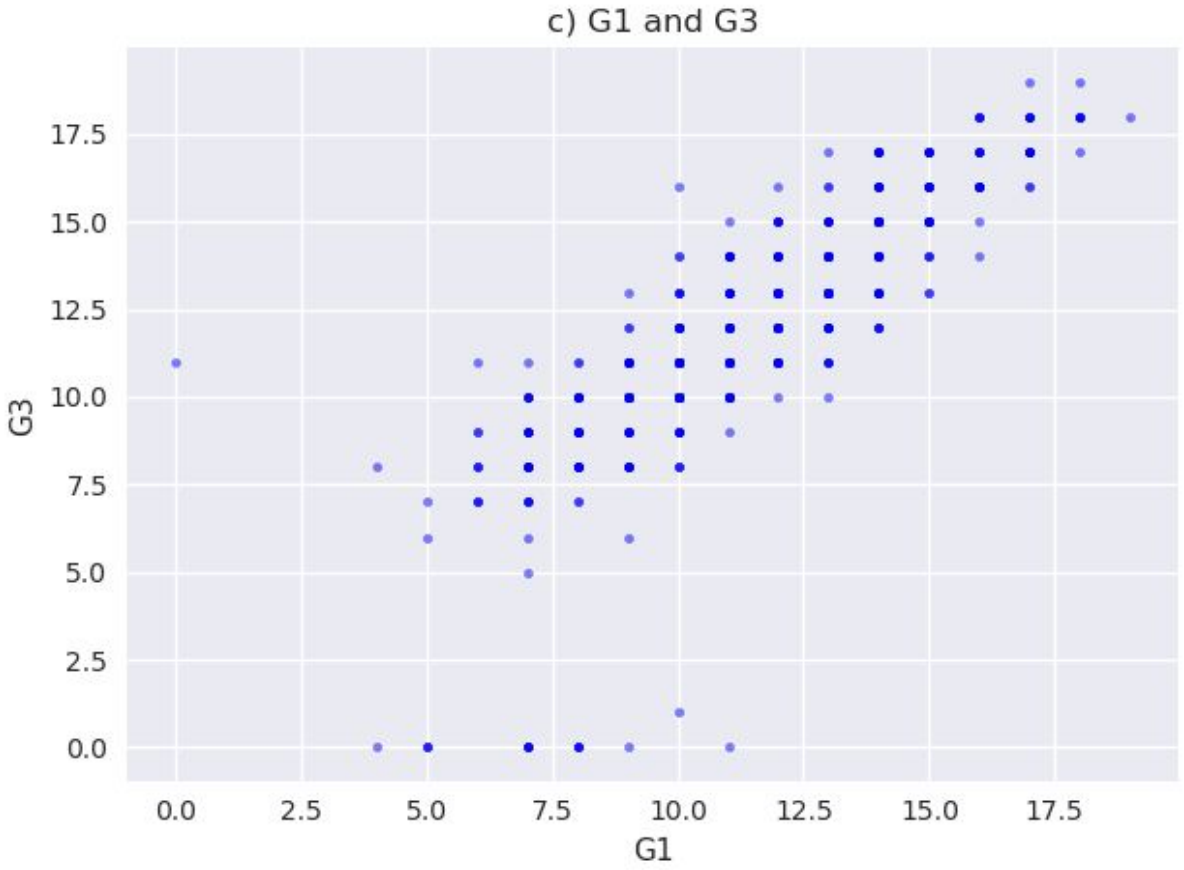
As evident from the results of Task 2, scatter plot of G1 and G2 shows a strong positive linear relationship. There are some outliers too. The dark dots represent overlapping points with same coordinates/values of (G1, G2).

Correlation value G1, G2 = 0.864982



As evident from the results of Task 2, scatter plot of G2 and G3 shows a strong positive linear relationship. There are some outliers too. The dark dots represent overlapping points with same coordinates/values of (G2, G3).

Correlation value G1, G2 = 0.918548



As evident from the results of Task 2, scatter plot of G1 and G3 shows a strong positive linear relationship. There are some outliers too. The dark dots represent overlapping points with same coordinates/values of (G1, G3).

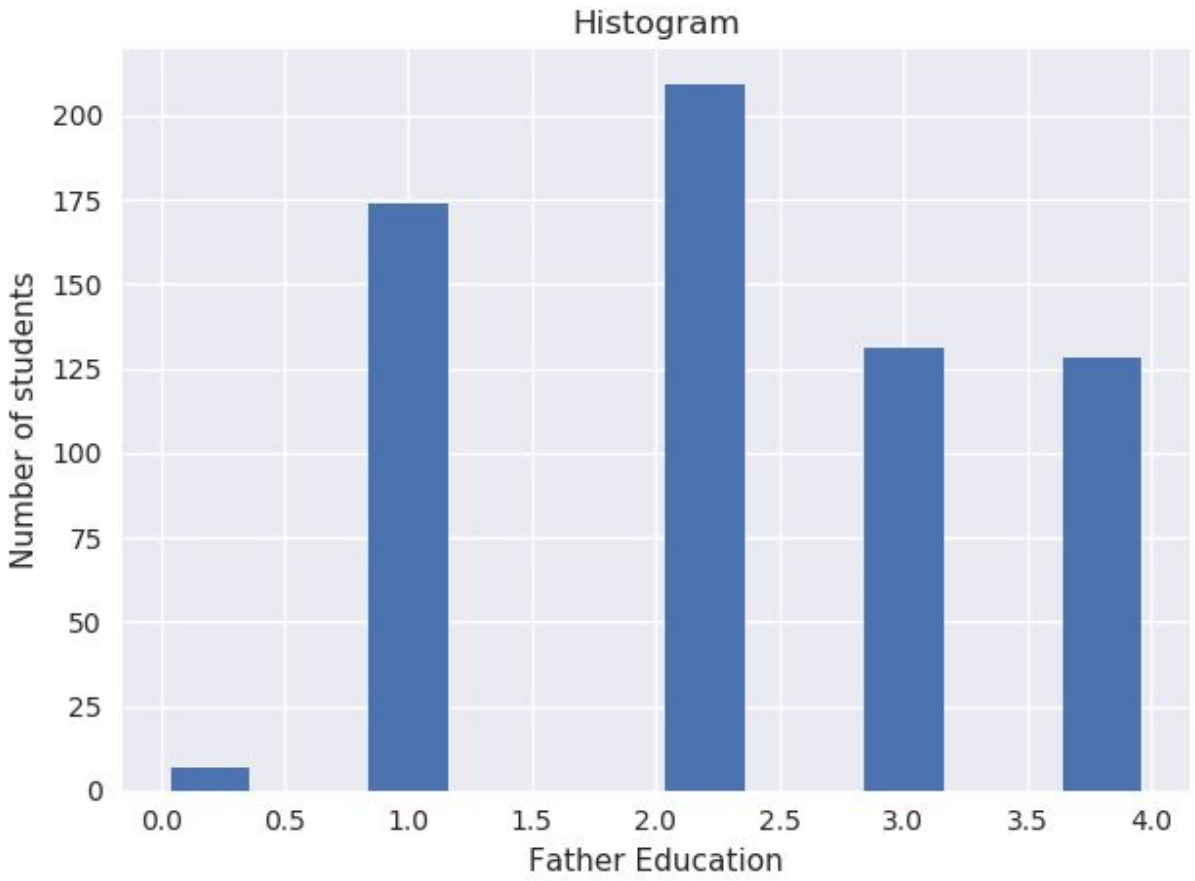
Correlation value G1, G3 = 0.826387

Task-4

Histograms for the attributes: a) Fedu, b) Mjob, c) studytime, d) failures,

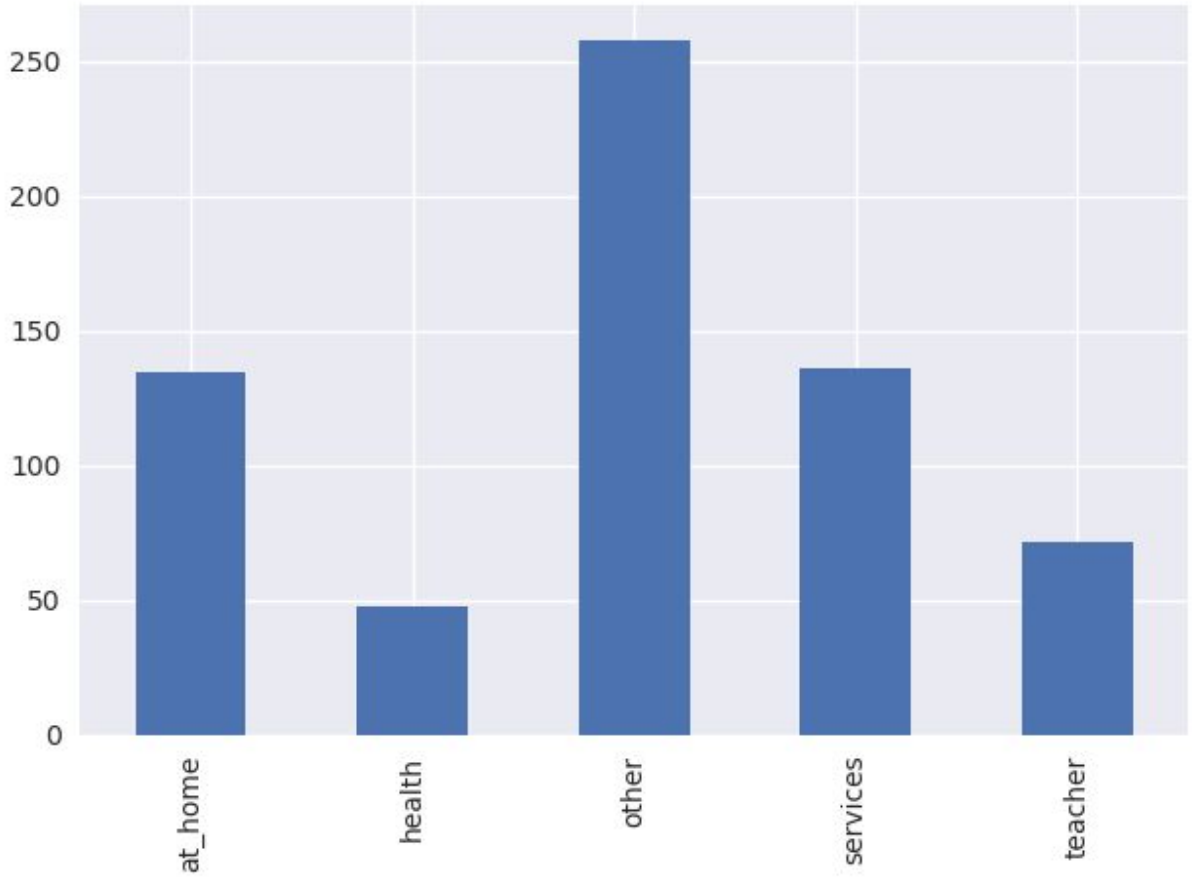
and e) goout

Histogram is a chart that shows frequency distribution of a variable/attribute.

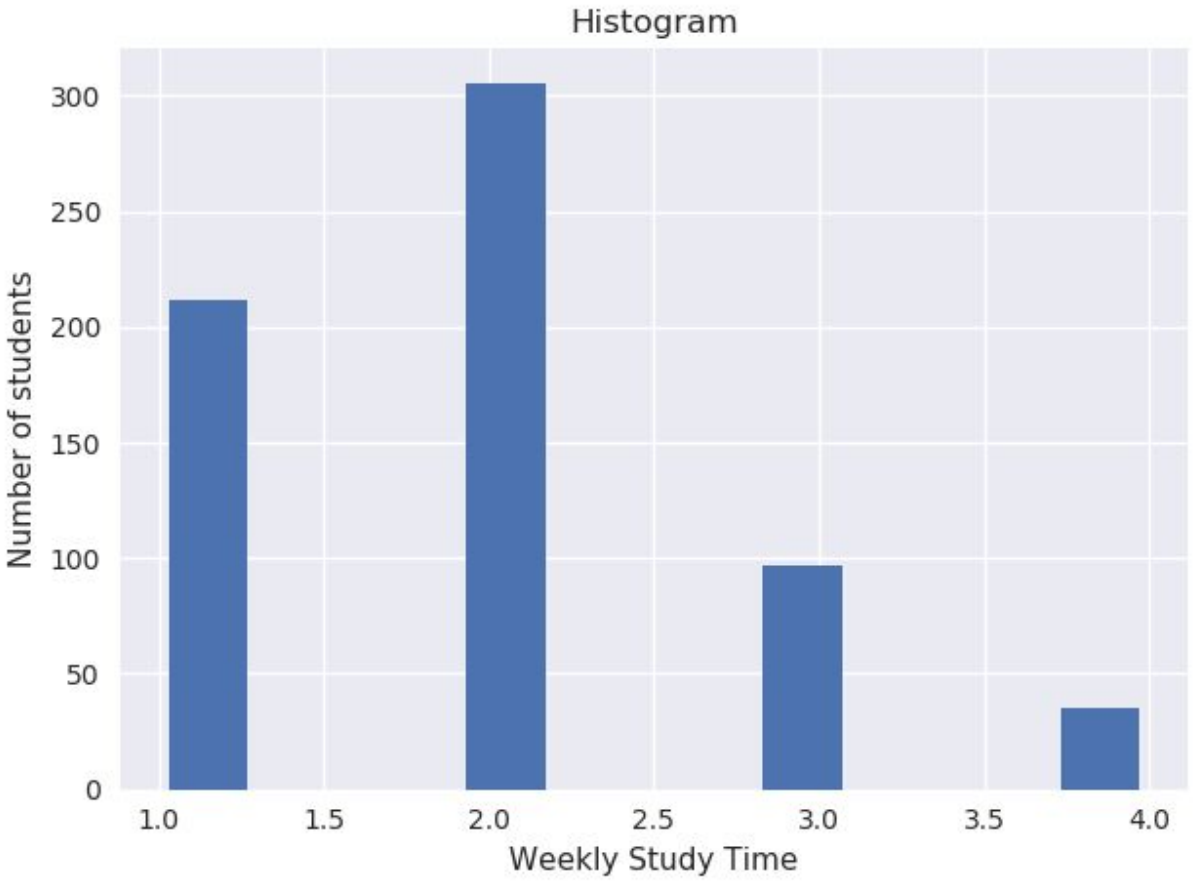


Here, father of most students have education of 2 (5th to 9th grade).

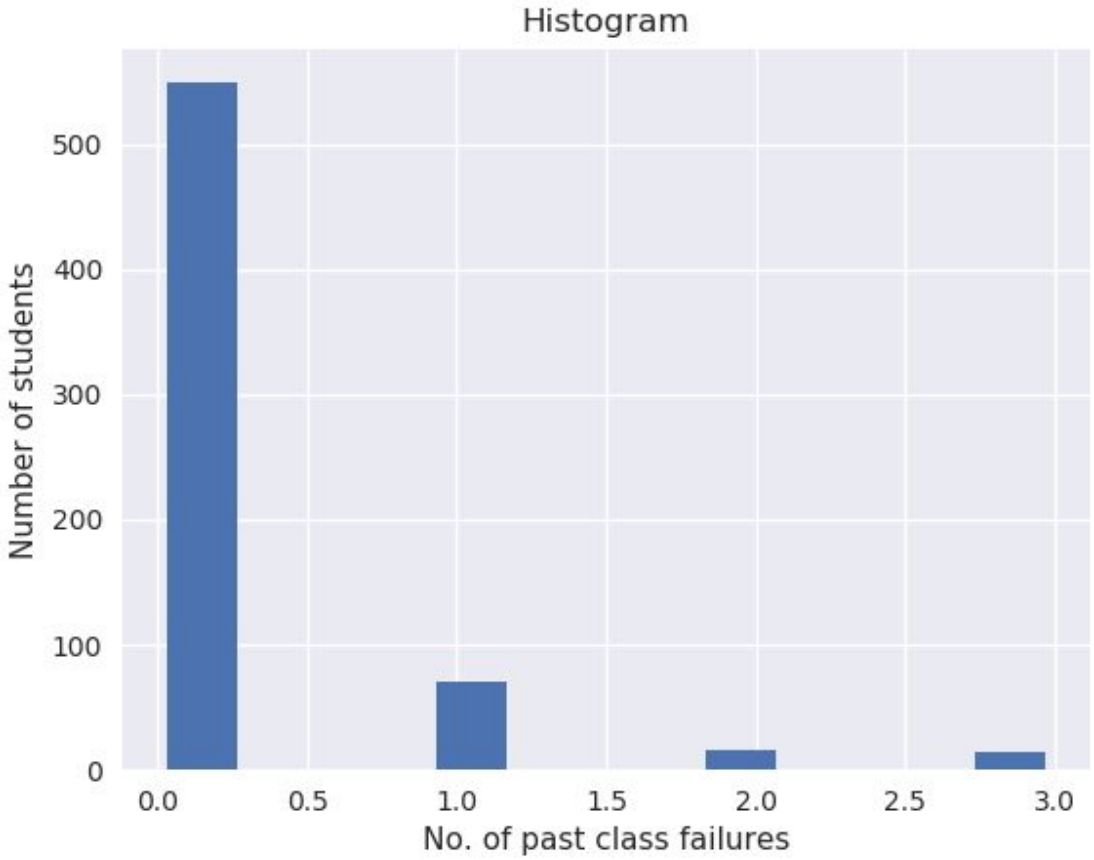
And only a few students father have no education.



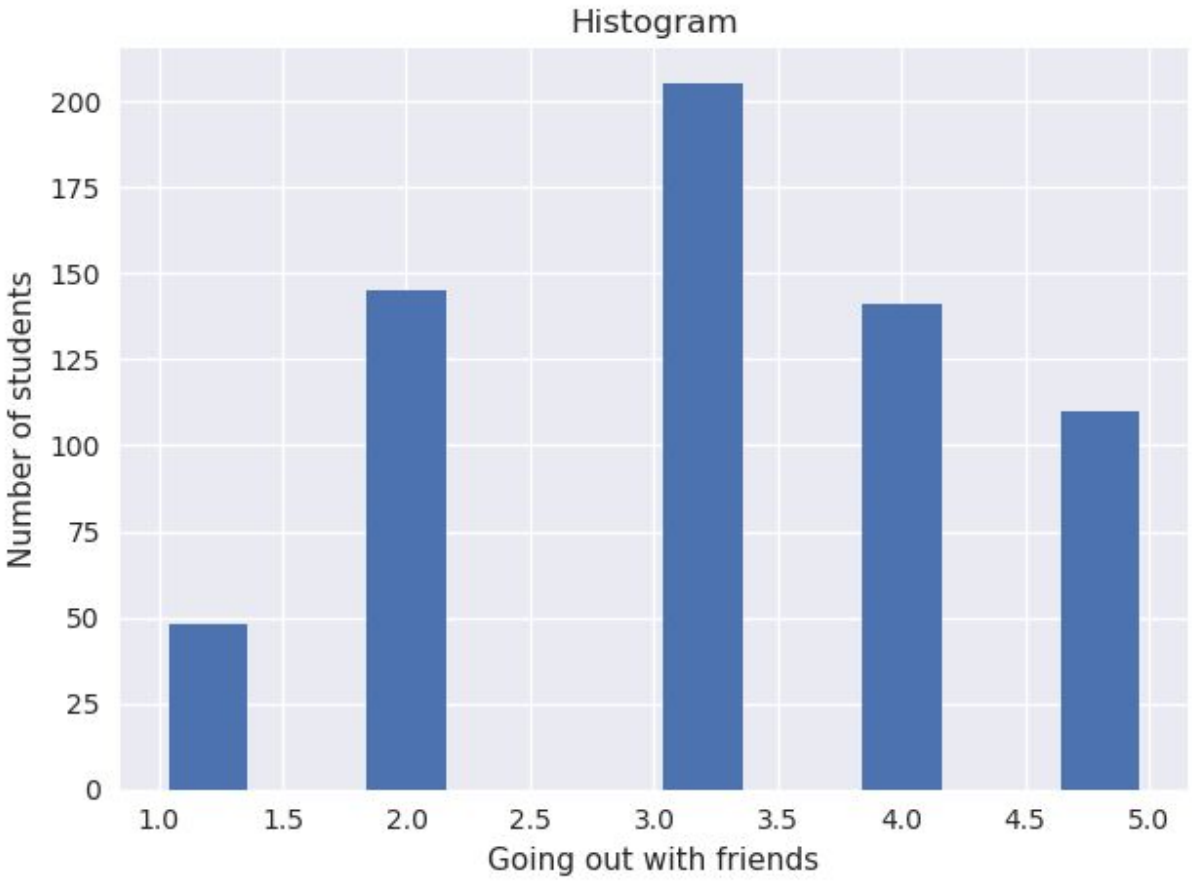
Most mothers have “other” jobs while least work in the health sector.



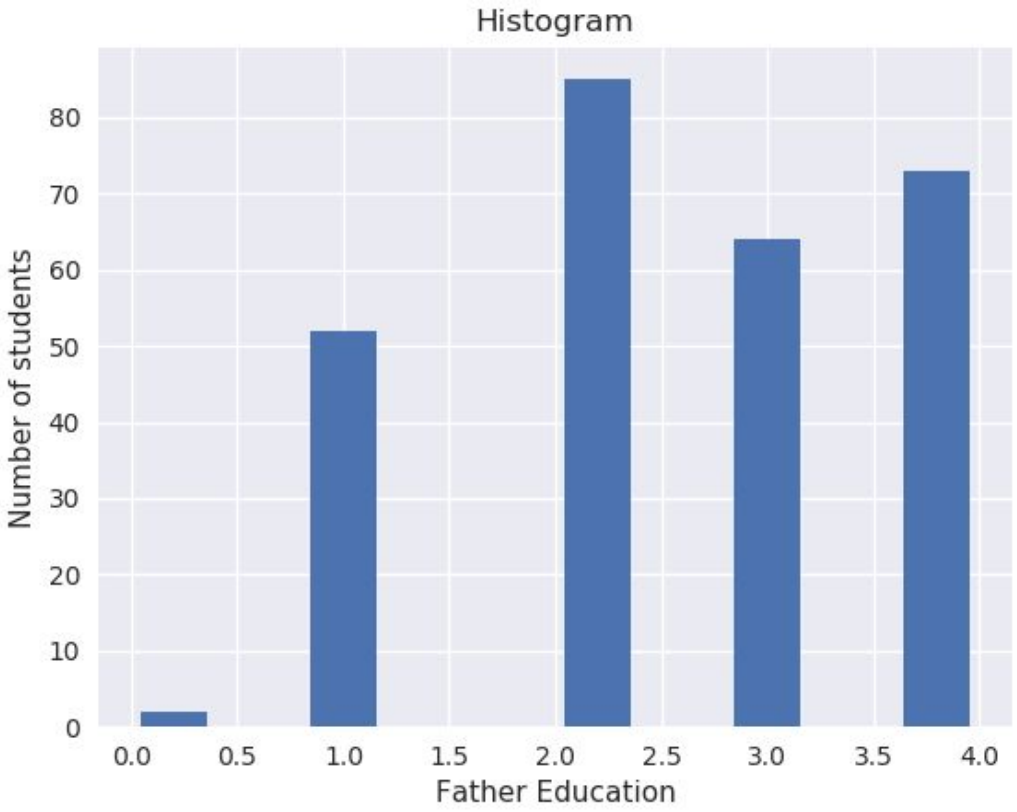
Most students(300+) study for 2 to 5 hours.



Most students(500+) have not failed in any past class.



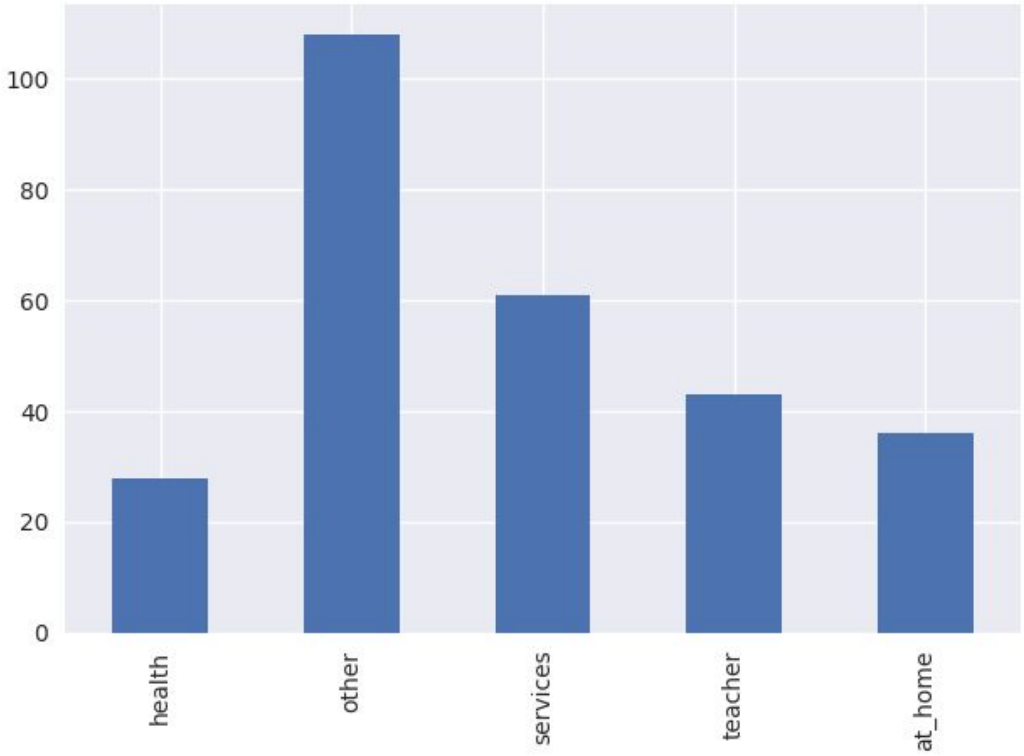
The frequency is high near the mid, so students prefer to go out modestly.



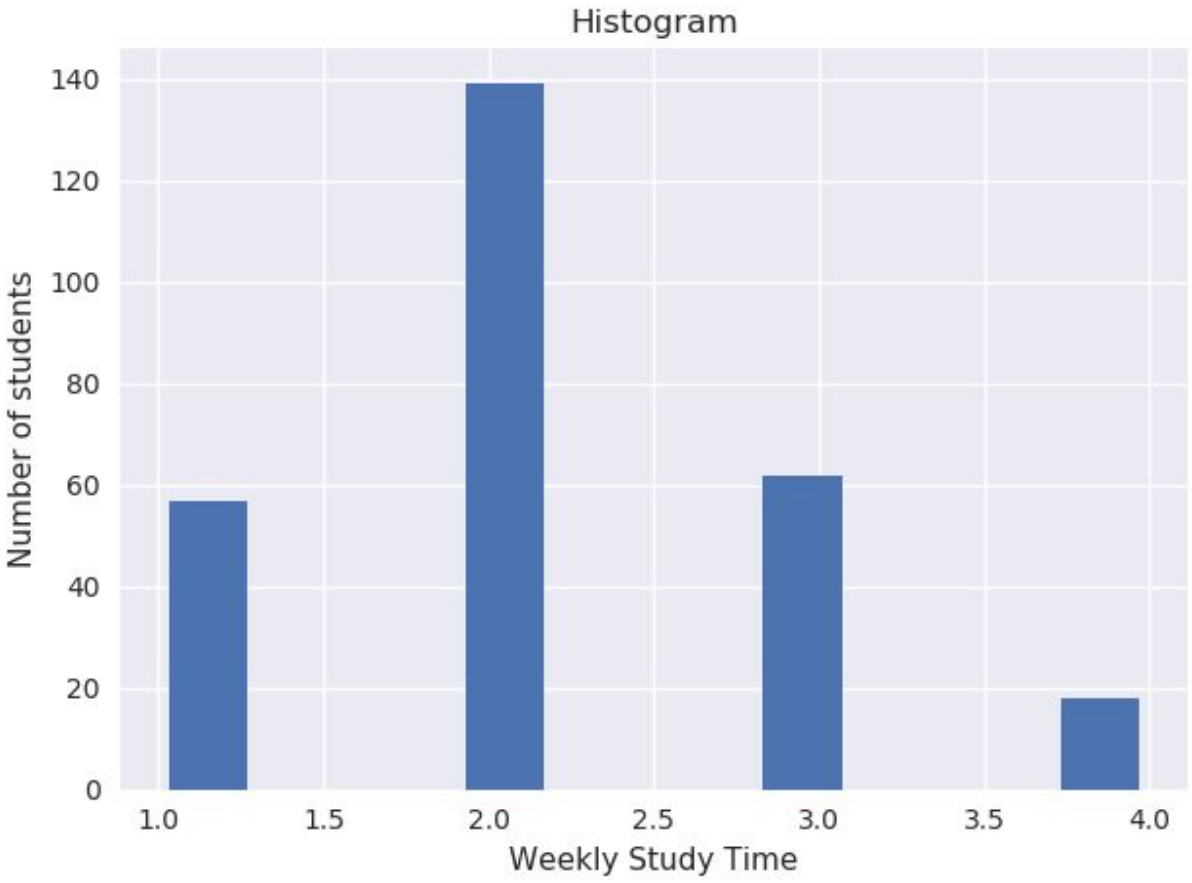
For students with final grade G3>12:

Histograms for the attributes: a) Fedu, b) Mjob, c) studytime, d) failures, and e) goout

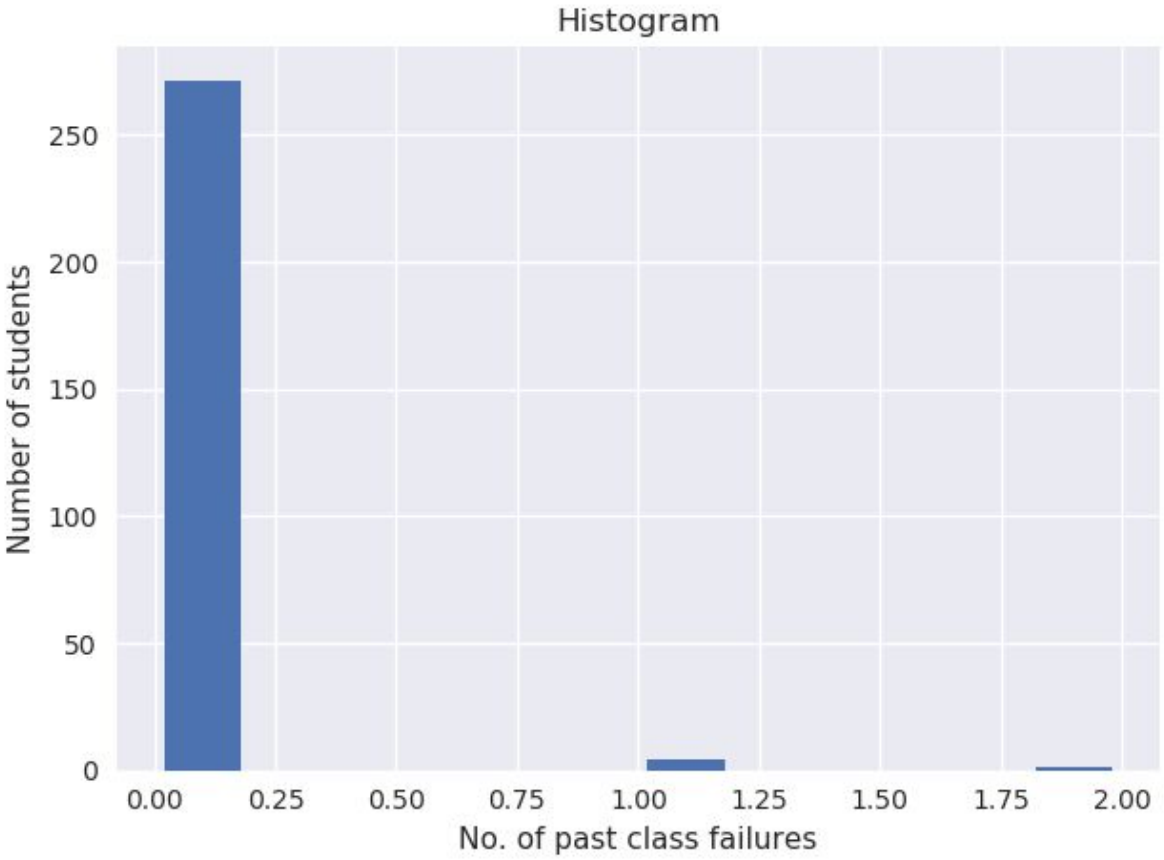
Students having good final grade generally have more educated fathers.



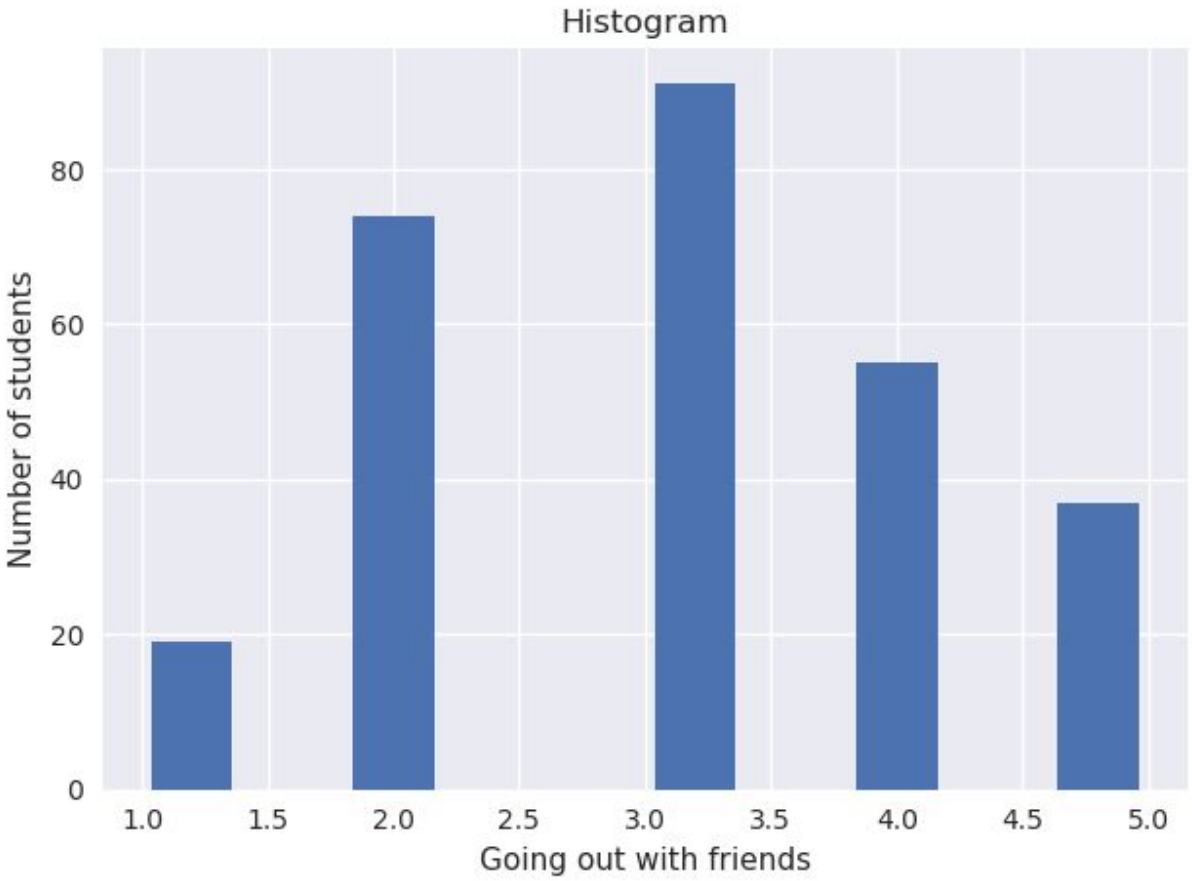
This is identical to the general view. So, mother of students with good grade also have “other” jobs.



This shows that students with good grade doesn't necessary study for long hours as the peak is at 2 (2 to 5 hours). Instead, they study with full concentration for average time.

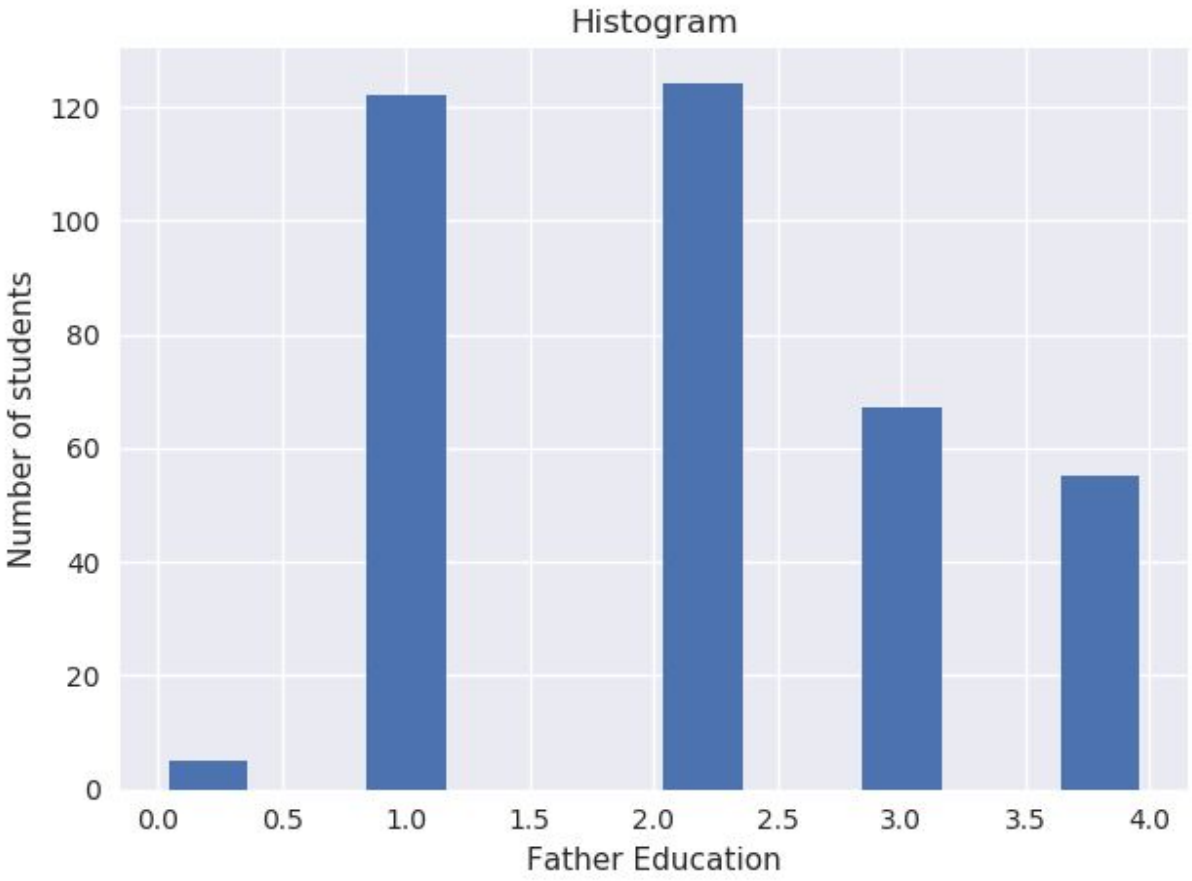


Students with high grade generally have never failed in a past class except for few.



This is also identical to the overall/general data.

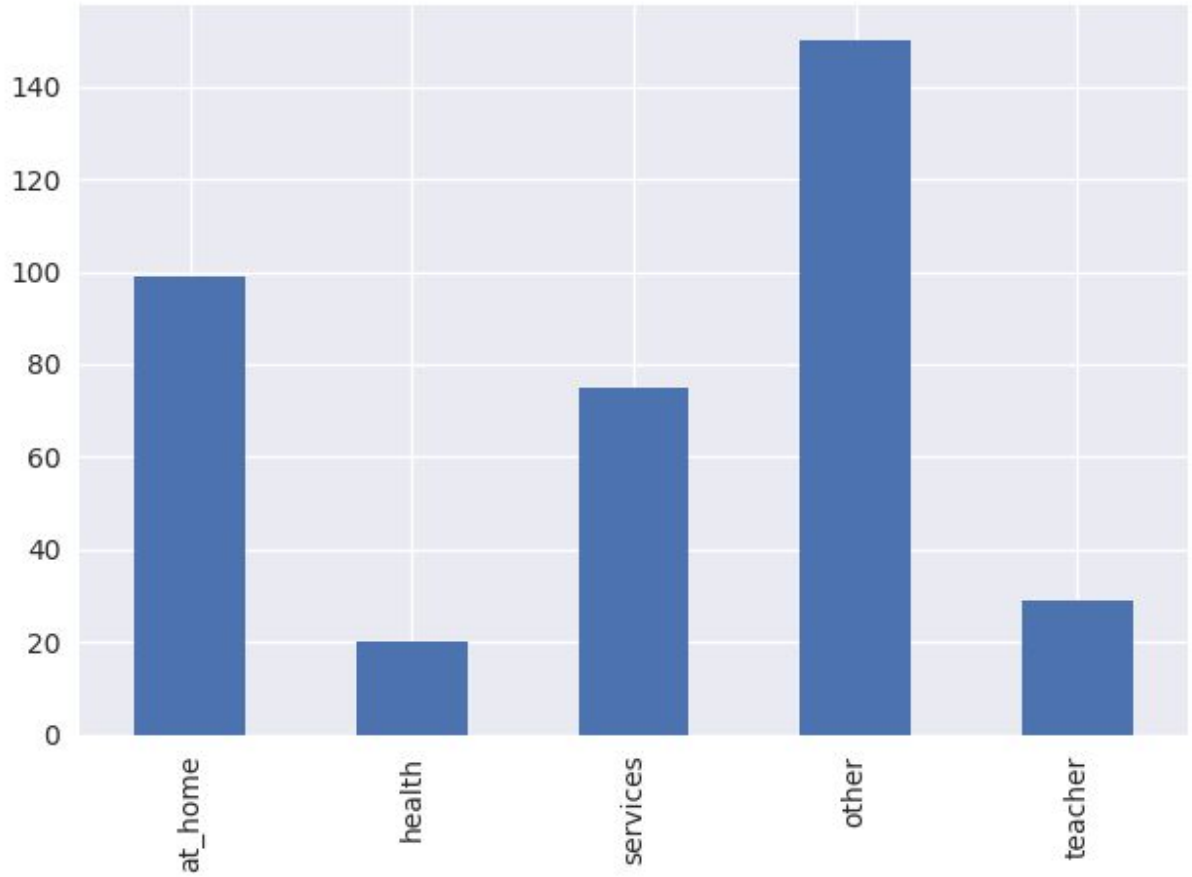
Students having high grade do go out with friends.



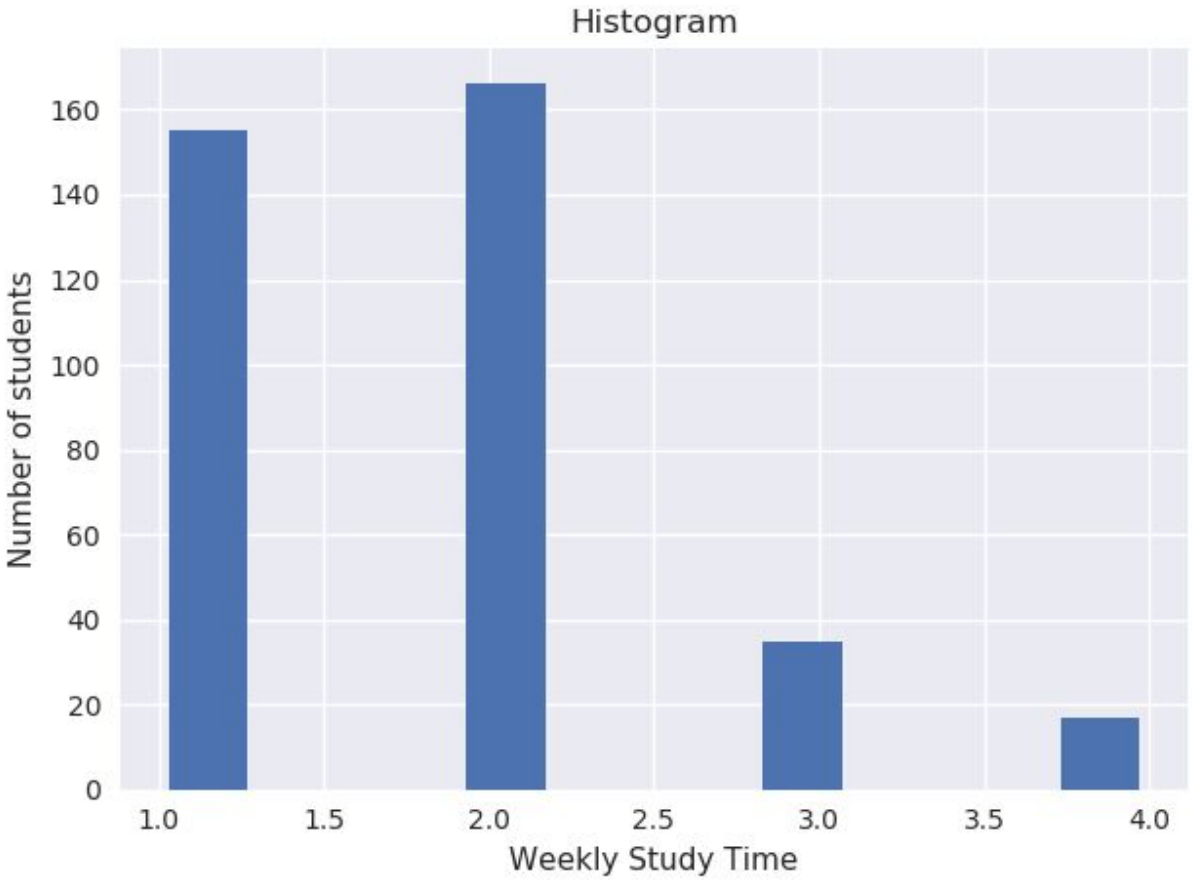
For students with final grade G3<=12:

Histograms for the attributes: a) Fedu, b) Mjob, c) studytime, d) failures, and e) goout

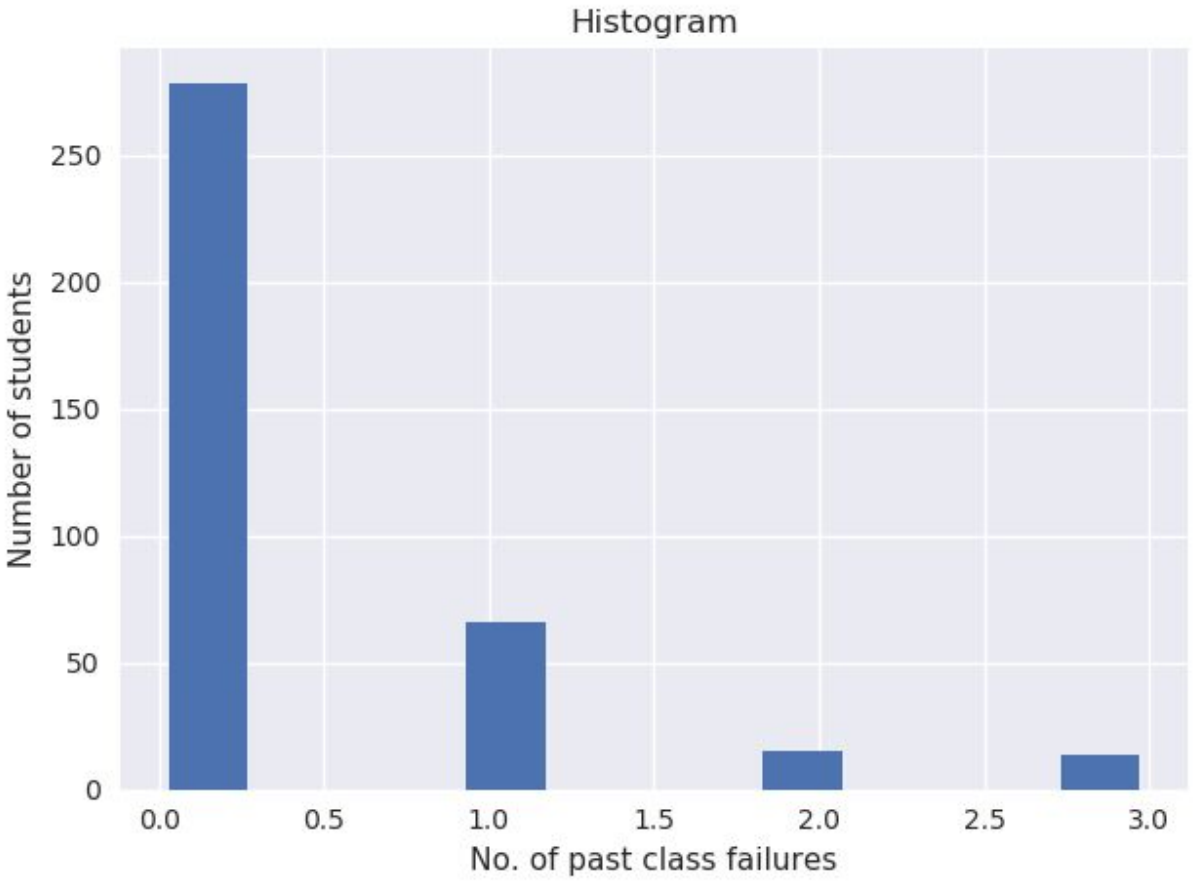
Fathers of students having lower grades are also less educated. With general education being either upto 4th grade or between 5th to 9th grade mostly.



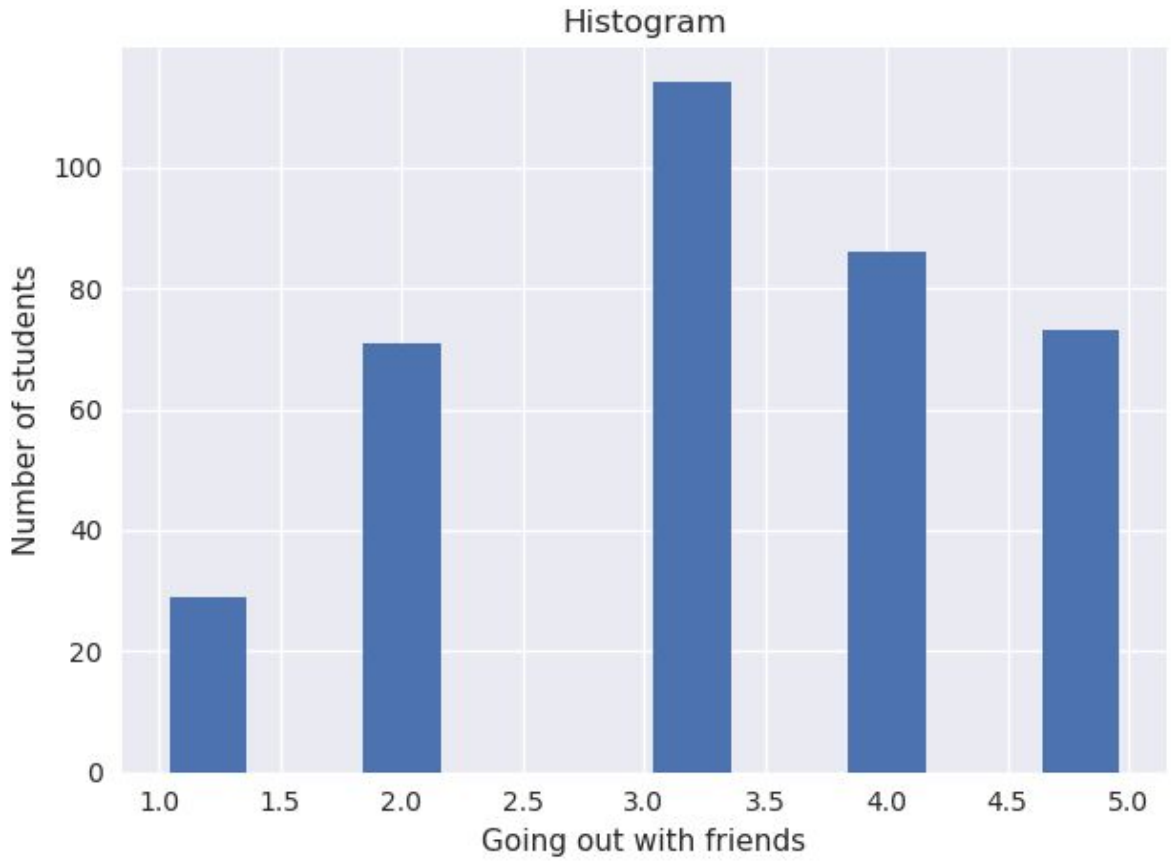
This is identical to the general view. So, mother of students with good grade also have “other” jobs.



Students with low grade study for less number of hours (As most are either in bin 1 or 2).

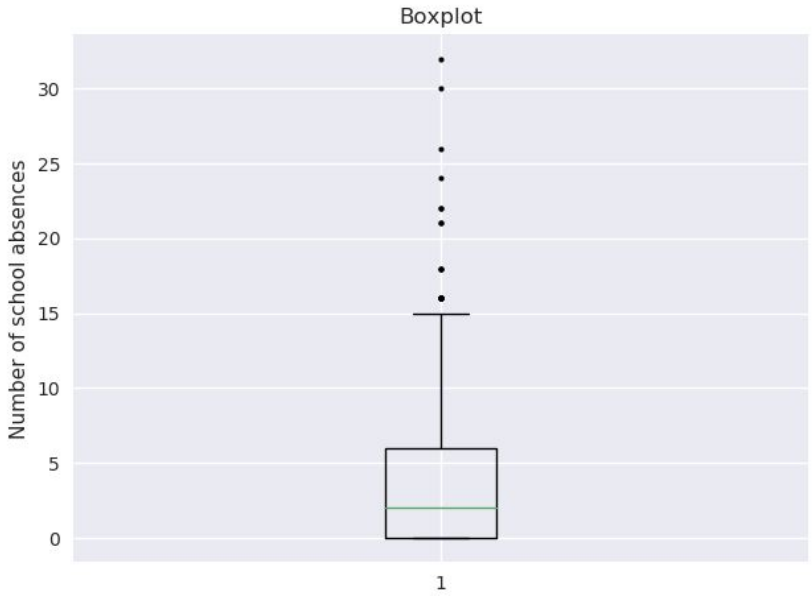


A significant number of students having low grades have failed in the past classes.

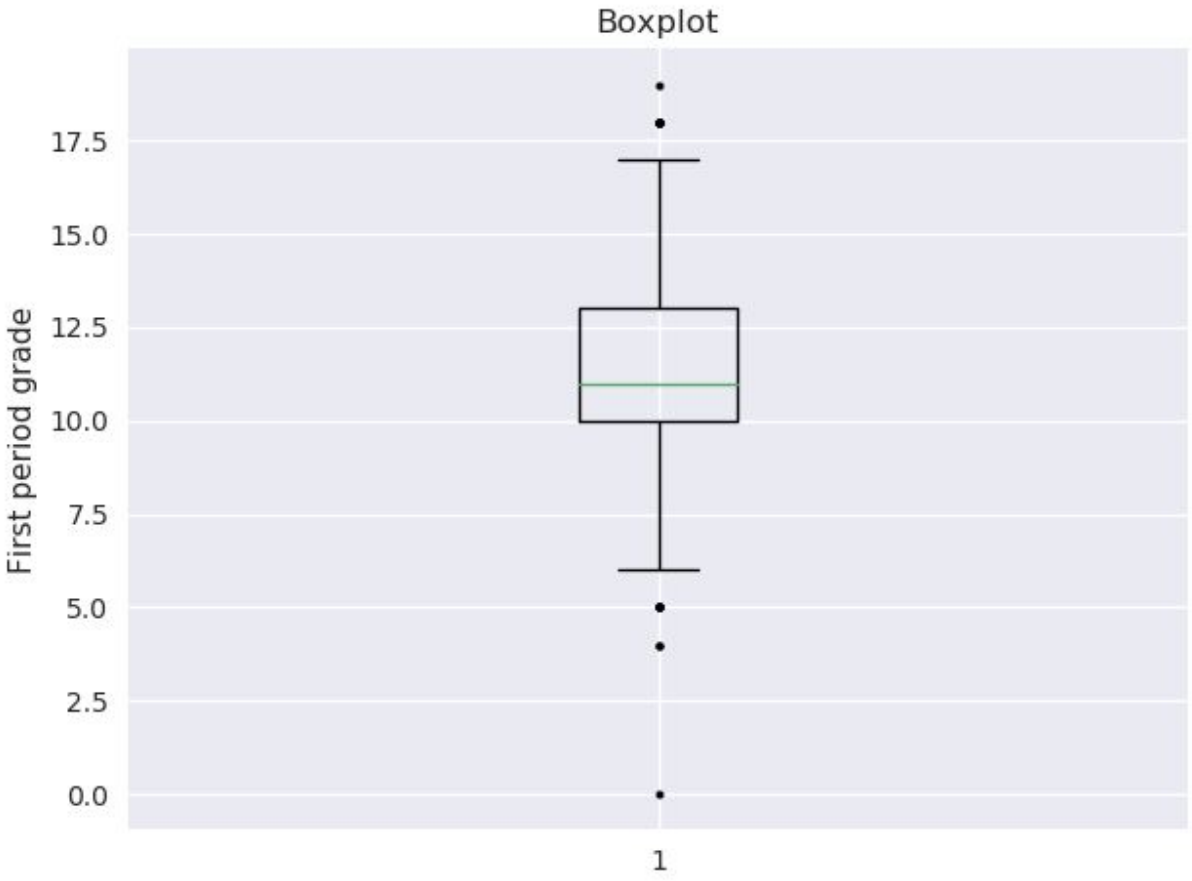


Students with low grades do go out a lot with their friends thus having an impact on their studies.

Task-5

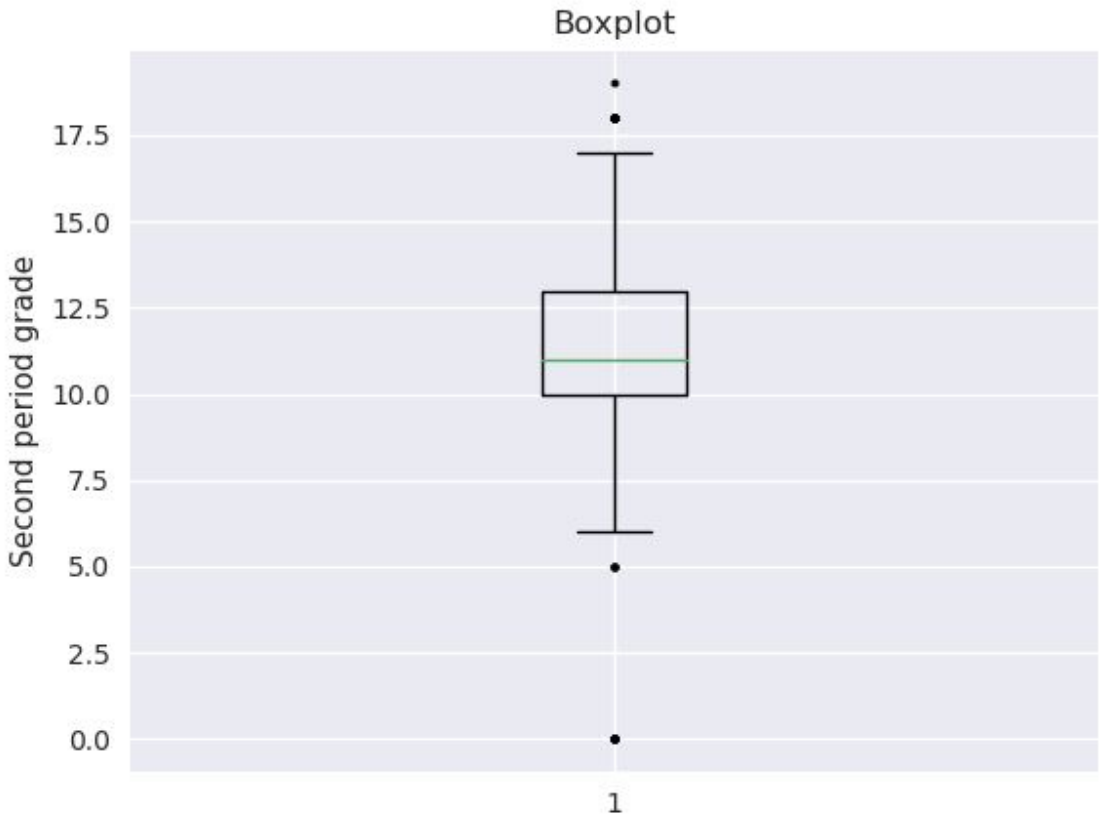


The box plot shows that the median is around 2. Quartile 3 or 75% of students have less than 6 absents. There are some outliers which shows some students who have took 20-30 absents.



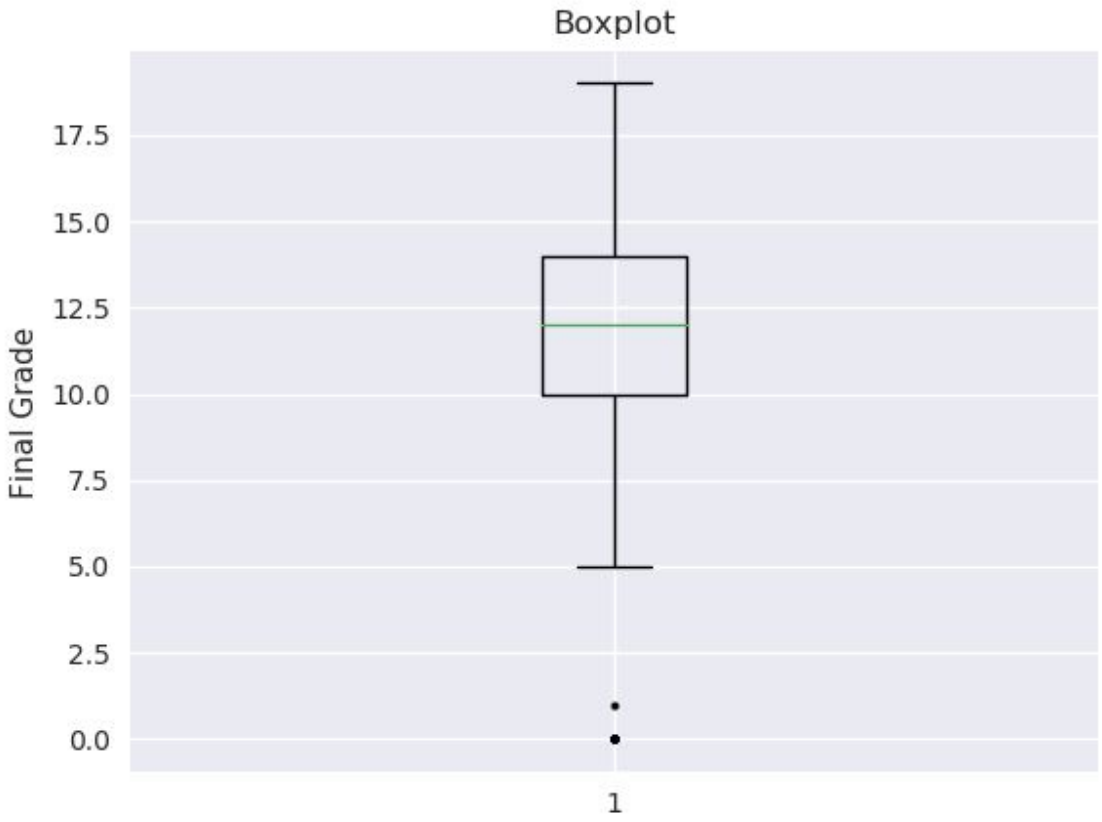
Grade is about 10-13 for 50% of students. IQR = 3

There are again some outliers which shows some poor students and some extraordinary students. (In Studies- First Period)



Grade is about 10-13 for 50% of students. IQR = 3

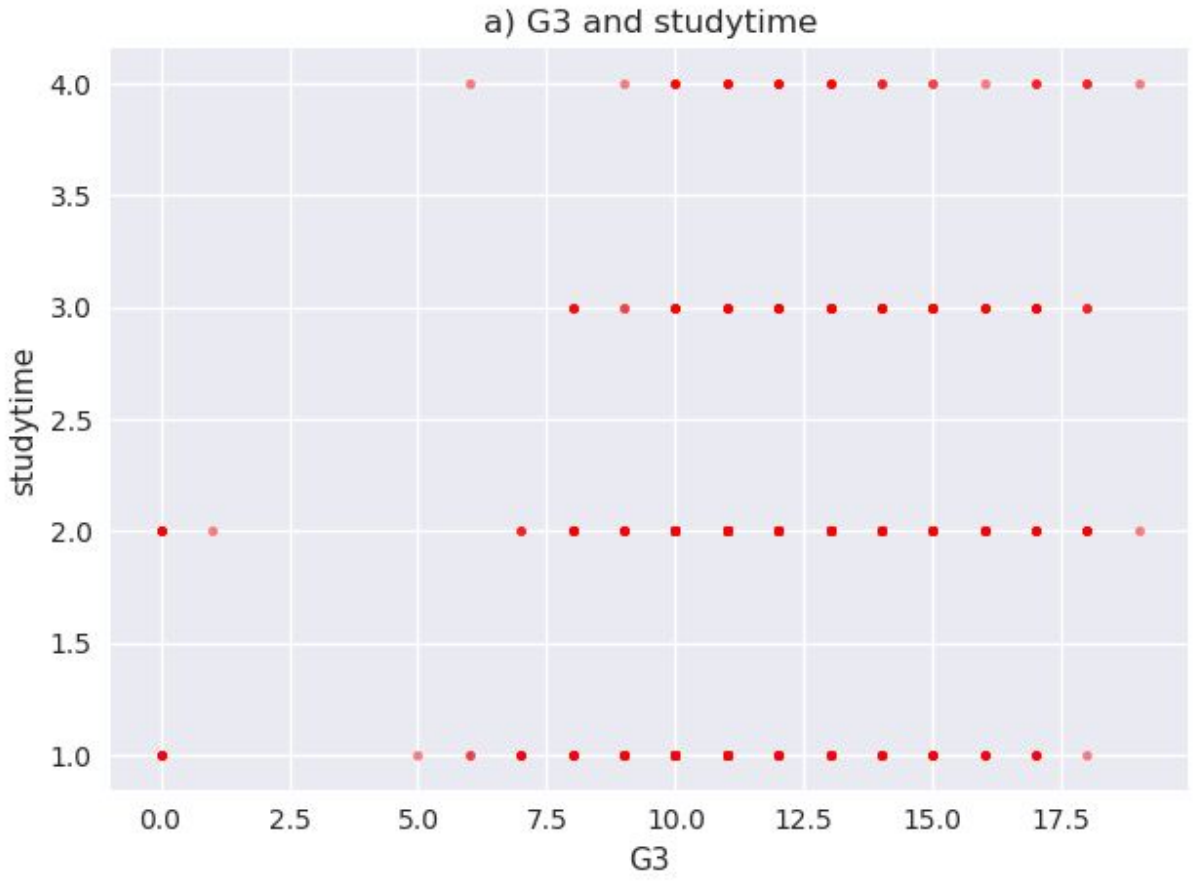
There are again some outliers which shows some poor students and some extraordinary students. (In Studies- Second Period)



Grade is about 10-14 for 50% of students. IQR = 4. Max range =19, Min = 5.

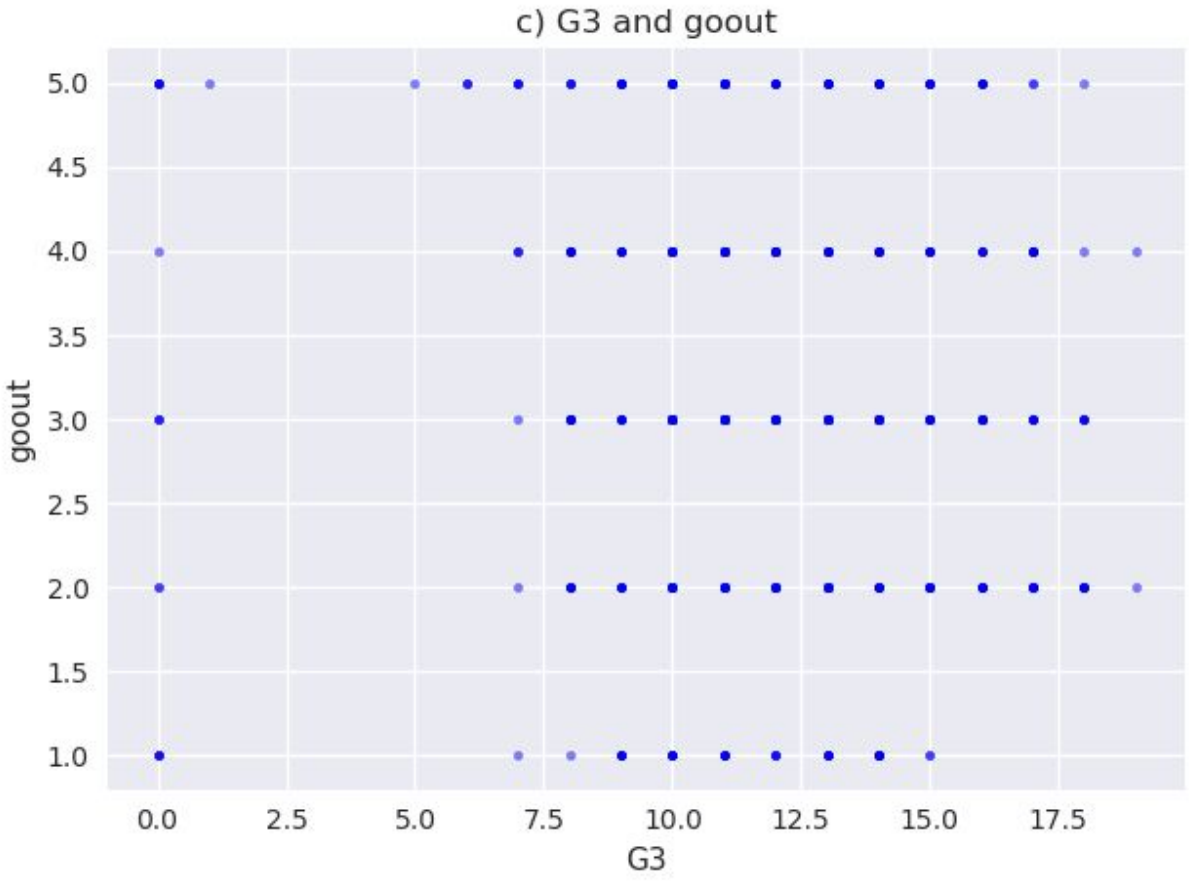
There are again some outliers which shows some poor students and some extraordinary students. (In Studies- Final grade)

Task-6

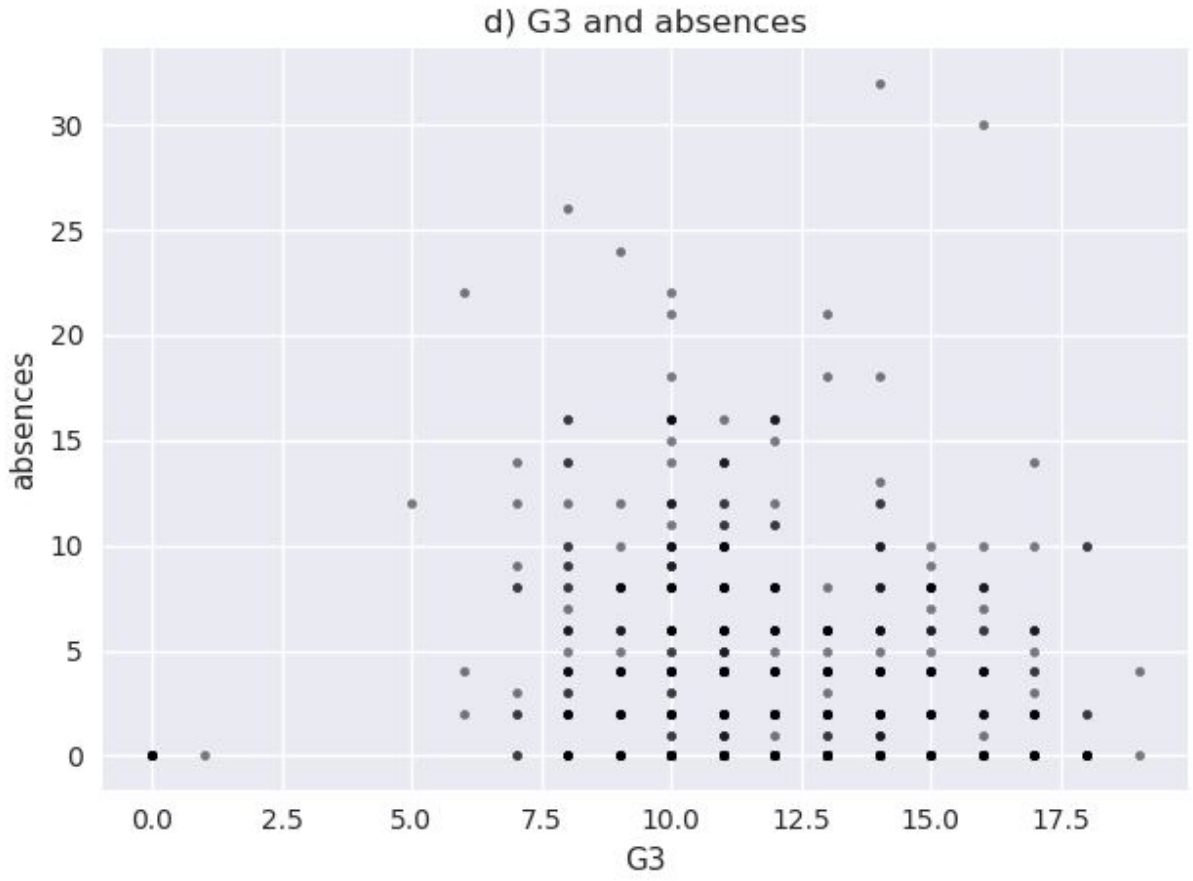


This shows that students with high study hours generally score relatively better final grades than students with low study hours.

Students with low failures tend to have higher grades.



Students who go out very often (value-5) tend to have lower grades compared to others.



Students with low absents perform better in class, thus have good grades.

Task-7

```
heil_keshav@heil-wallace:~/Desktop/Last/CSL 524/2015csb1016_A1$ python Script.py

Triangular Dissimilarity Matrix:
0.000000
0.287743 0.000000
0.467489 0.331564 0.000000
0.561067 0.484156 0.469071 0.000000
0.513503 0.484042 0.475306 0.346570 0.000000
0.518214 0.484701 0.474868 0.469071 0.448237 0.000000
0.518348 0.466686 0.482096 0.370960 0.180278 0.481577 0.000000
0.476494 0.514539 0.531681 0.407010 0.177613 0.482374 0.165971 0.000000
0.567173 0.486617 0.466279 0.338543 0.341321 0.341930 0.366060 0.403170 0.000000
0.541927 0.533663 0.513241 0.380278 0.114018 0.461428 0.285482 0.239888 0.376165 0.000000
```

Here, d(i, j) shows dissimilarity between i th row and j th row of dataset considering all attributes.

If they are completely similar, then d is 0. If they are completely different d is 1.

For row 10, it is most similar to row 5 with value of d being 0.114018

1	4	at_home	course	2	0	4	4	0	11	11
2	1	at_home	course	2	0	3	2	9	11	11
3	1	at_home	other	2	0	2	6	12	13	12
4	2	health	home	3	0	2	0	14	14	14
5	3	other	home	2	0	2	0	11	13	13
6	3	services	reputation	2	0	2	6	12	12	13
7	2	other	home	2	0	4	0	13	12	13
8	4	other	home	2	0	4	2	10	13	13
9	2	services	home	2	0	2	0	15	16	17
10	4	other	home	2	0	1	0	12	12	13

This can be cross checked using the dataset. Row 10 has most attribute values similar to Row 5 then any other row.

To know about the dissimilarity between any two rows i and j, just check their d value from matrix (Dij or Dji).
Closer to 1 being completely dissimilar.