Data Science Course Project

Phase 1: Proposal

Team 2

Team members:-

Name	Sec	BN
Esraa Gamal	1	15
Esraa Amr	1	16
Ayman Mohamed	1	19
Rawaa Ahmed	1	32

Preprocessing and Cleansing

- We turned a lot of categorical data into numeric
- Added new features which are a combination of others or binarized ones
- Dealt with null values by imputing mean or mode
- Resampled some unbalanced classes

Descriptive Questions:

1) **What is the Percentage of people in the population getting bullied?**

   Using statistical methods.

   We found that

Bullied inside school (%)	Bullied Outside school (%)	Cyberbullied (%)
20.857368	22.202319	22.405731

However, those percentages are not disjoint, some students may experience bullying inside school, outside school and cyberbullied at the same time, so it is better to get the Percentage of bullied students as a joint matrix…

	Cyber_bullied_in_past_12_months	No	Yes
Bullied_on_school_property_in_past_12_months	Bullied_not_on_school_property_in_past_12_months		
No	No	0.593630	0.083490
	Yes	0.069069	0.045237
Yes	No	0.072803	0.028053
	Yes	0.040440	0.067278

From this cross-table we can see that 59.3% didn't report any form of bullying.
Meaning that the remaining 40.7% suffer from at least one form of bullying which is a large percentage.

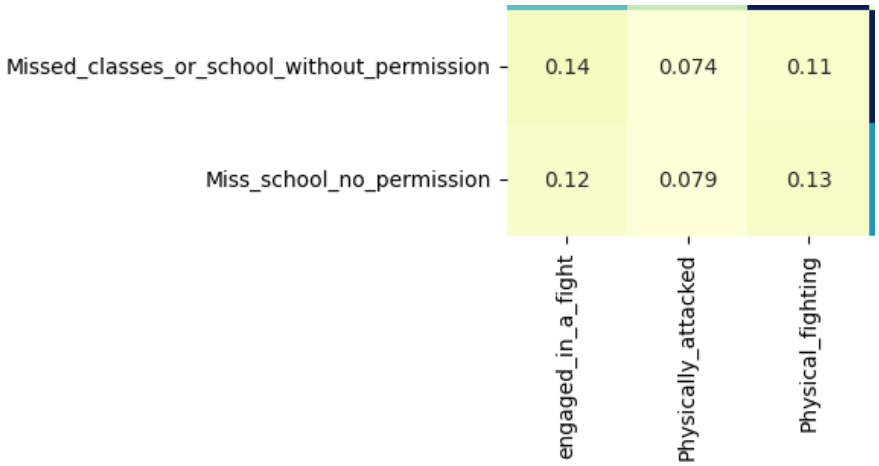
Students bullied inside school AND outside school = 10.7717
Meaning that HALF of the students bullied inside school, experience bullying outside school.
Bullied both is school & outside school is a large percentage of all students.

Very small number of students who don't get bullied in school experience bullying outside school, this concludes that a student is more likely to experience different forms of bullying if he was bullied in school.

This also means that the psychological reasons for bullying a kid is in the kid himself (his personality/appearance/talking skills) not due to the school's other students.

Exploratory Questions:

1) **What is the relationship between being physically attacked and missing school ?**



1) **Expectations:** the more a student gets physically attacked, the more days he misses from school. Data should help.
2) **Collect Data:** Data is numeric, the output correlation is very low.
3) **Matching Expectation with results:** Mismatch, we cannot conclude anything yet.

→ We do not care about number of missed classes, we only care whether the student misses classes or not, we binarized the column 'missed classes'

1) **Expectations:** the more a student gets physically attacked, probability of missing school increases
2) **Collect Data:** physically attacked is still a numeric column, get correlation bet physically attacked and missed school.
3) **Matching Expectation with results:** Mismatch, low correlation, we cannot conclude anything yet.

→ there is another numeric column called 'physical fighting' which is strongly correlated to physically attacked, maybe we should merge both columns with each other and binarize them, we still don't care about number of physical fights, rather, we care about whether they happen or not, the new binary column will be named 'engaged in a fight'

1) **Expectations:** if a student engages in a fight, his probability of missing school increases
2) **Collect Data:** Binary engagement in fights correlates to missing school.
3) **Matching Expectation with results:** Match! 14% correlation is a good percentage.

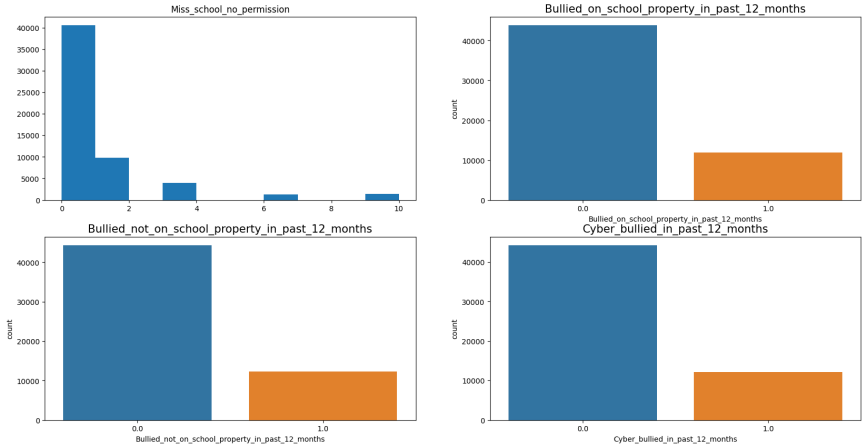
Interpretation and Communication:
We should try to minimize fights as little as possible, so kids don't miss school and other important days.
The variable "engaged_in_a_fight" is a binary variable if the student undergoes any form of fighting or attacking,

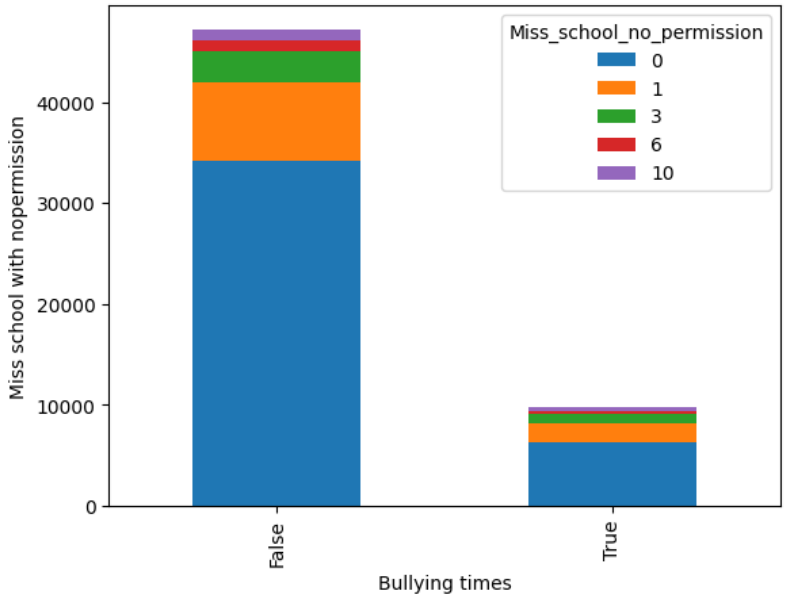
There is a weak correlation between engaging in a fight and missing school which was against our expectations.

Notice that physical fights correlate more to missing school than attacks. Because fights are more violent than just plain attacks.

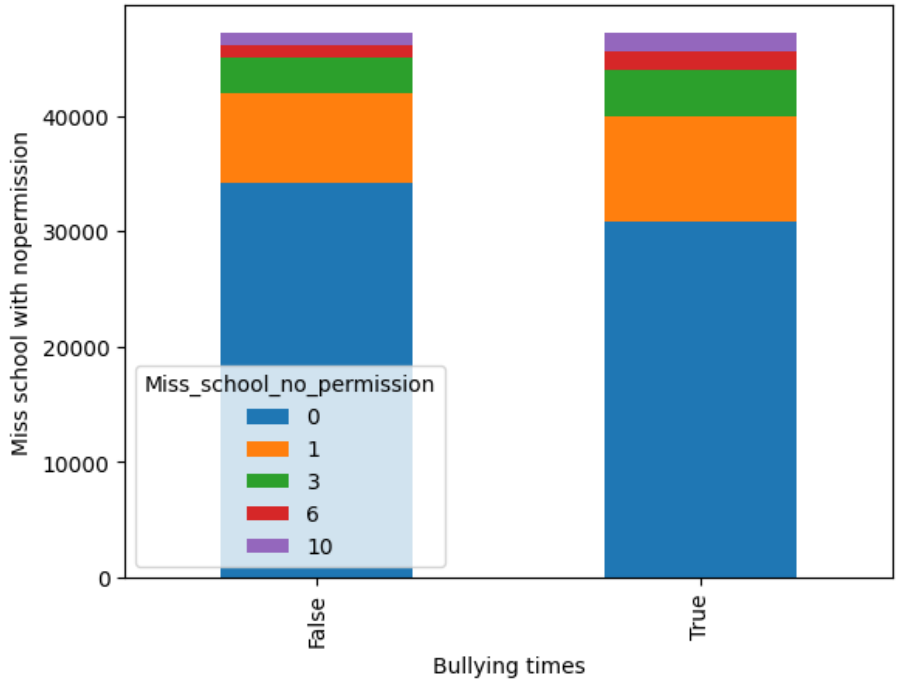
2) **How many days does a student being bullied usually skip on average?**

We combined between all variables of being bullied (bullied in school on past 12 months, bullied not in school in past 12 months, cyber bullied) if anyone is yes so the student suffer from bullied
And the result that there are a few student suffer from any sort of bullying





And when we balanced data we found the distribution like that



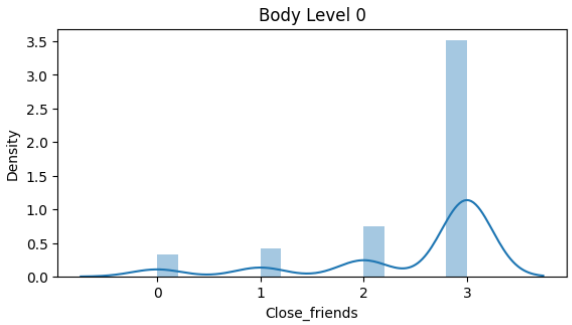
The average day missing school without permission

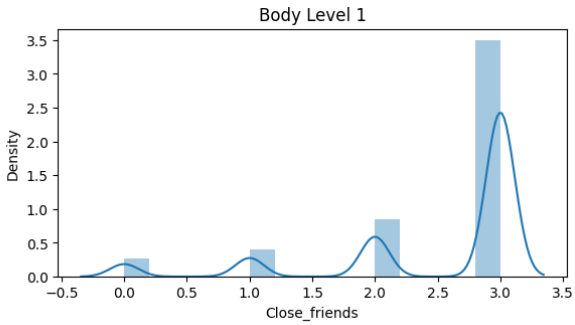
	Bullying	Miss_school_no_permission
0	Don't suffer from bullying	1.0
1	suffer from bullying	2.0

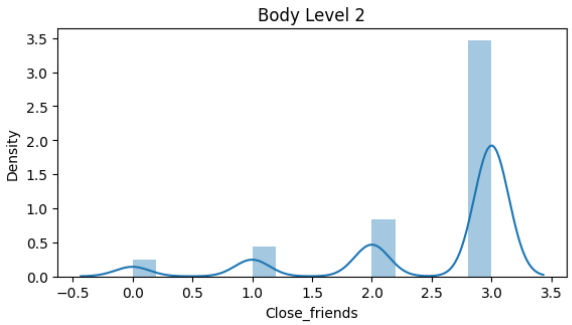
3) Is body level related to the number of close friends?

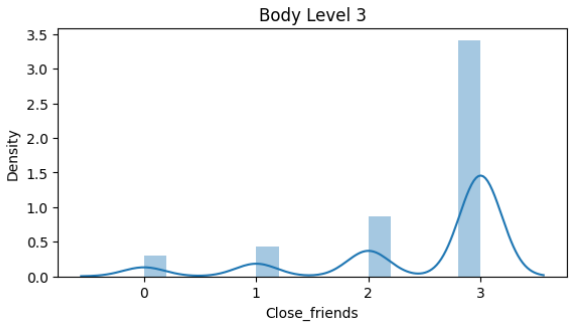
Our expectation is that they are indeed related and that being overweight or underweight would increase the chances of isolation and consequently lacking friends out of low self esteem or being outcasted or mocked.

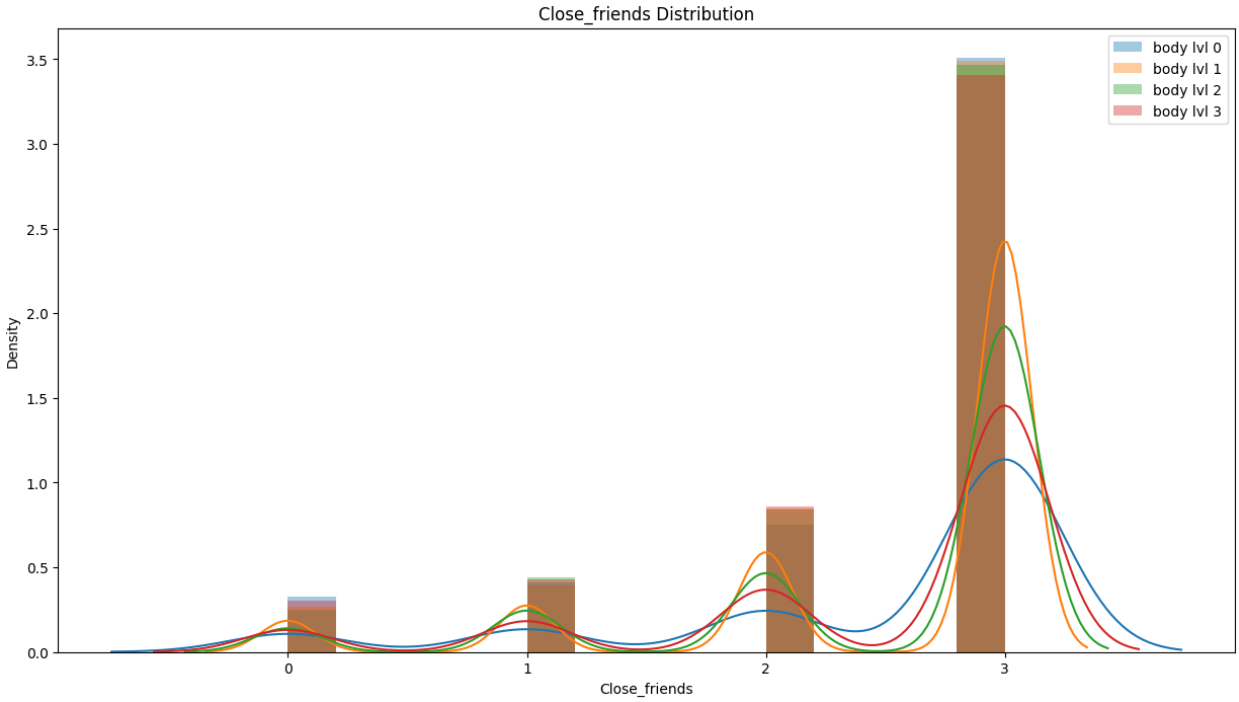
So we tried to visualize the distribution of close friends for each body level and the results as we can see is that they are very similar, probably the only difference is the number of students in each class because the data of body level is unbalanced.



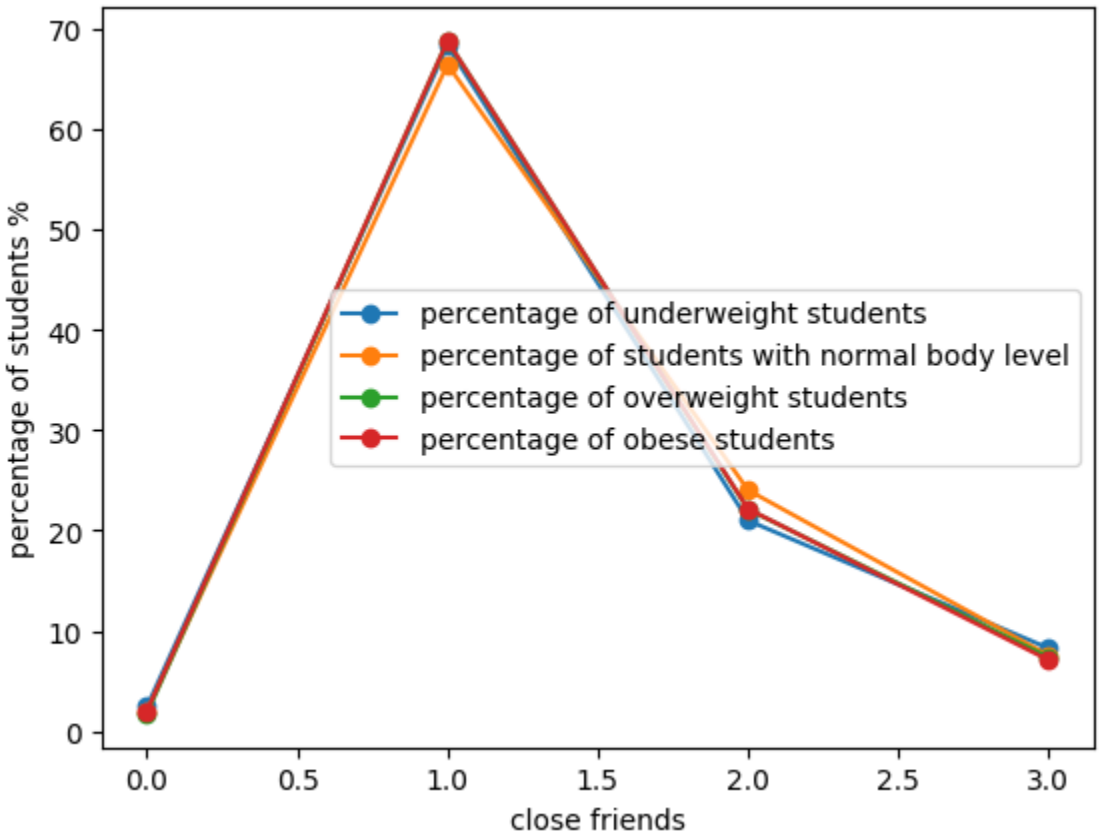








After that we went to check the percentage for a more accurate outcome and as we can see they are almost overlapping!



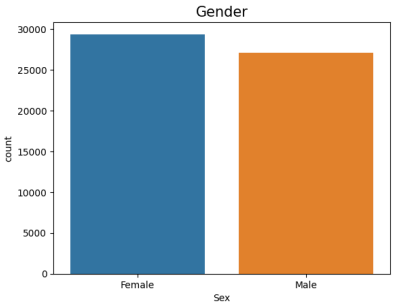
Finally we tried using the Spearman and Pearson correlation test with the hypothesis that those two variables are related and dependent and that hypothesis was rejected.

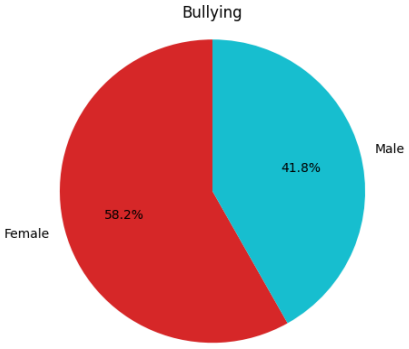
```
Pearson
        stat=-0.009, p=0.092
        Probably independent
Spearman
        stat=-0.009, p=0.106
        Probably independent
```

In conclusion we can say that having any type of body level has absolutely no effect on making friends like anybody else which was definitely a pleasant surprise.

4) Does bullying differ based on gender?

We first choose the student suffer from bullying and classify and we find female suffer from bullying more than male
→ for any type of bullying





And all types of bullying, the female suffering from bullying.

Inferential Questions:

Q: Can we Generalize bullying in school to other forms of bullying? (bullying outside school and cyberbullying)

1. **Stating the question:**
    1) Expectations: question is answerable, with ‘yes’.
    2) Collecting Data:
    There were 3 columns, bullied inside school, bullied outside school, cyberbullied.

3) Comparing Data with expectations: Match, the smaller population was students getting bullied inside school only, while the larger population is students who get bullied anywhere by any form.

2. EDA:

1) Expectations: Data is full and clean.
2) Collecting Data: those 3 columns carried ‘yes’ and ‘no’ and a lot on null entries.
3) Comparing Data with Expectations: mismatch, null entries can be MNAR.

   → impute missing Data using zero entries, null means zero, binarize the 3 columns, We made up a new column called ‘bullied’ which represents ANY form of bullying whether it was in school or outside or cyberbullied, this is our larger population, This column carries binary values.

1) Expectations: Data is full and clean.
2) Collecting Data: 2 columns, 1 for the smaller population and the other for the bigger population.
3) Comparing Data with Expectations: Match, we can now train a model on the smaller population.

3. Building Model:

We trained a SVM model on the dataset with the column ‘bullied in school’ as a label, we hid the rest of the labels as if we were training our model on a smaller population.
Those were the results of our SVM model when testing on the larger population:

```
########################################### USING LINEAR KERNEL ##########################
Accuracy: 0.7945598128107634
Confusion Matrix:
 [[13583     0]
 [ 3512     0]]
classification_report:
              precision    recall  f1-score   support

       False       0.79      1.00      0.89     13583
        True       0.00      0.00      0.00      3512

    accuracy                           0.79     17095
   macro avg       0.40      0.50      0.44     17095
weighted avg       0.63      0.79      0.70     17095
```

4. Interpretation:

1) Expectation: F1 scores above 70% when using our model on the bigger population
2) Collect Data: Classification report and focus on F1 score.
3) Comparing Data with Expectations: Match, we can generalize our model.

5. Communication:

1) Expectation: Bullying is not experienced in school only due to personal traits.
2) Collect Data: our model results and its ability to predict bullying anywhere.
3) Comparing Data with Expectations: Match, Bullying is experienced anywhere because of visible personal traits, a lot of other hidden factors contribute to bullying but we cannot yet measure.

Q:How many close friends does a student have on average?

1. Stating the question:

1- Expectations:

Question could be answered by the data.

2- Collect data:

The close friends column is categorical.

3- Comparing data and expectations:

Mismatch. Can't get the average for categorical data.

Instead: What is the mode number of close friends of a student?

2. EDA:

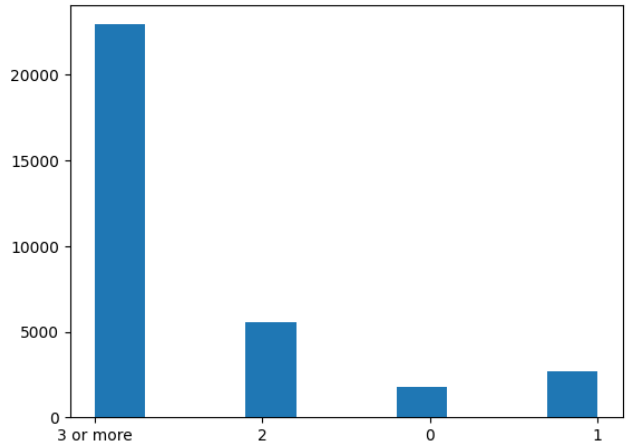
1- Expectations:

I can use the mode function to get the mode easily. Then plot the frequency of all variables.

2- Collect data:

Call the function on the column and build the histogram or pie chart.

```
0    3 or more
Name: Close_friends, dtype: object
```



3- Comparing data and expectations:

Match. I could easily get the mode and see the frequency of all variables.

3. Building a model: -

4. Interpreting results:

1- Expectations:

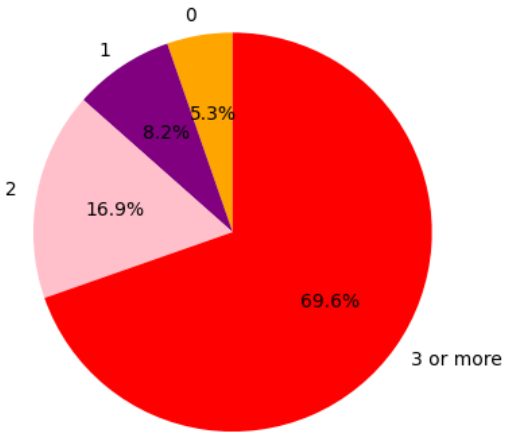
Research says: almost half (49%) report having 3 or fewer close friends. over one-third (36%) report having between 4 and 9 close friends. 13% say they have 10 or more close friends. 12% say they have no close friends.

2- Collect data:

69% of students have 3 or more friends.

3- Comparing data and expectations:

Match.The research and results are close.



5. Communication:

1- Expectations:

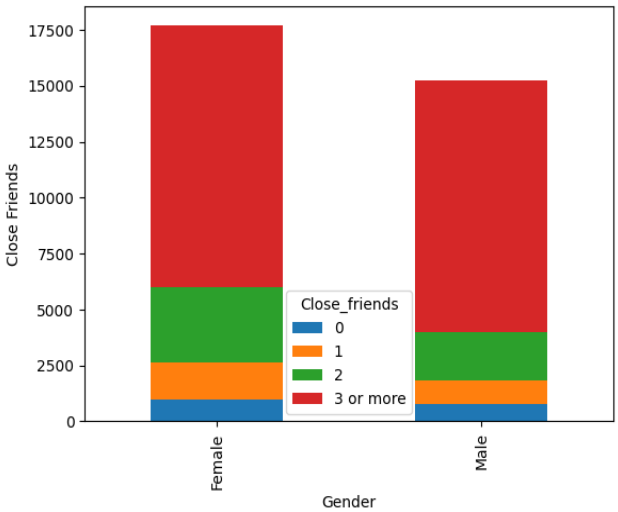
The report would satisfy the boss.

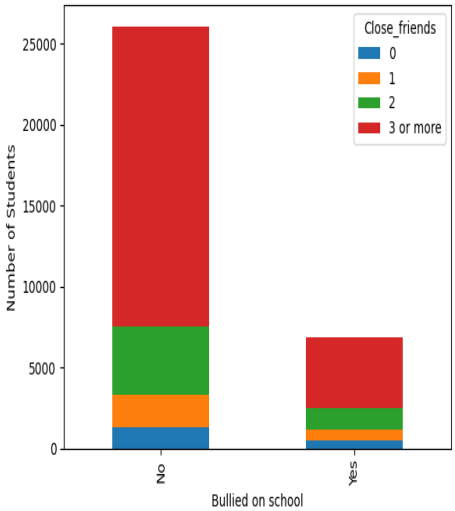
2- Collect data:

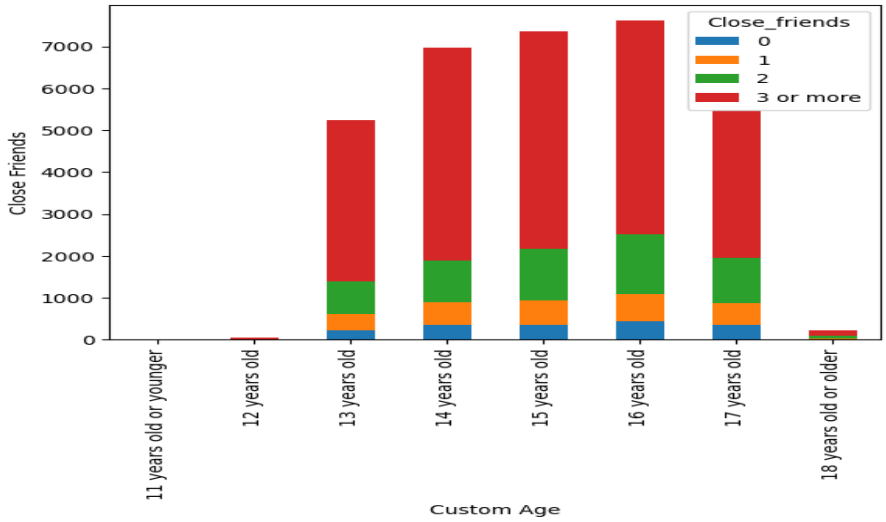
Boss asks if this is true for bullied students.

3- Comparing data and expectations:

Draw some stacked bars between the numbe bullied and also add more graphs for age and Gender.







Q2: What's the relationship between body level and being bullied?

2. Stating the question:

1- Expectations:

There is a body level column in the data.

2- Collect data:

Not found. Found were_obese, were_overweight and were_underweight.

3- Comparing data and expectations:

Mismatch. Need to build it myself by combining existing columns.

2. EDA:

1- Expectations:

I can draw stacked bars, heat maps and make some hypothesis tests to support the hypothesis and expect them to agree on the result.

2- Collect data:

All tests and graphs agree on the hypothesis.

3- Comparing data and expectations:

Match.

3. Building a model: -

4. Interpreting results:

1- Expectations:

Existing studies have yielded inconsistent results on gender differences in the effects of BMI on school bullying. Some studies have shown that both being underweight and overweight are the predictors of victimization of bullying for both boys and girls

2- Collect data:

Pearsonr Test:
stat=0.018842045462178277, p=0.0006267077924602461
Being bullied probably depends on body level
[Tableau Dashboard]

3- Comparing data and expectations:

Match.The research and results are close.

5. Communication:

1- Expectations:

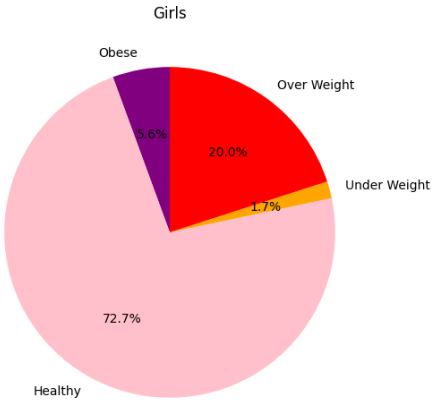
The report would satisfy the boss.

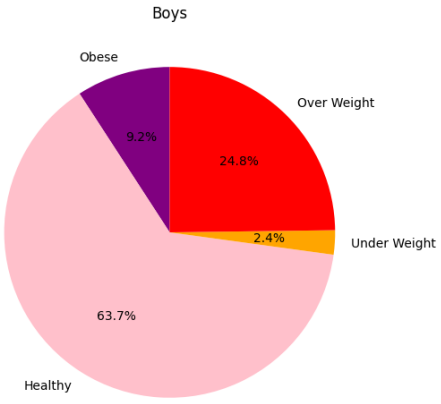
2- Collect data:

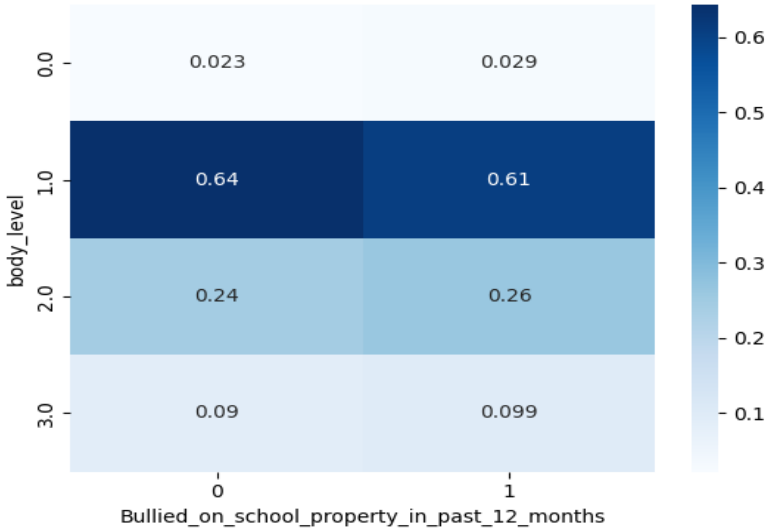
Boss asks if the hypothesis holds on both genders and if body_level affects physical fighting.

3- Comparing data and expectations:

Make more tests to prove or disprove these cases.

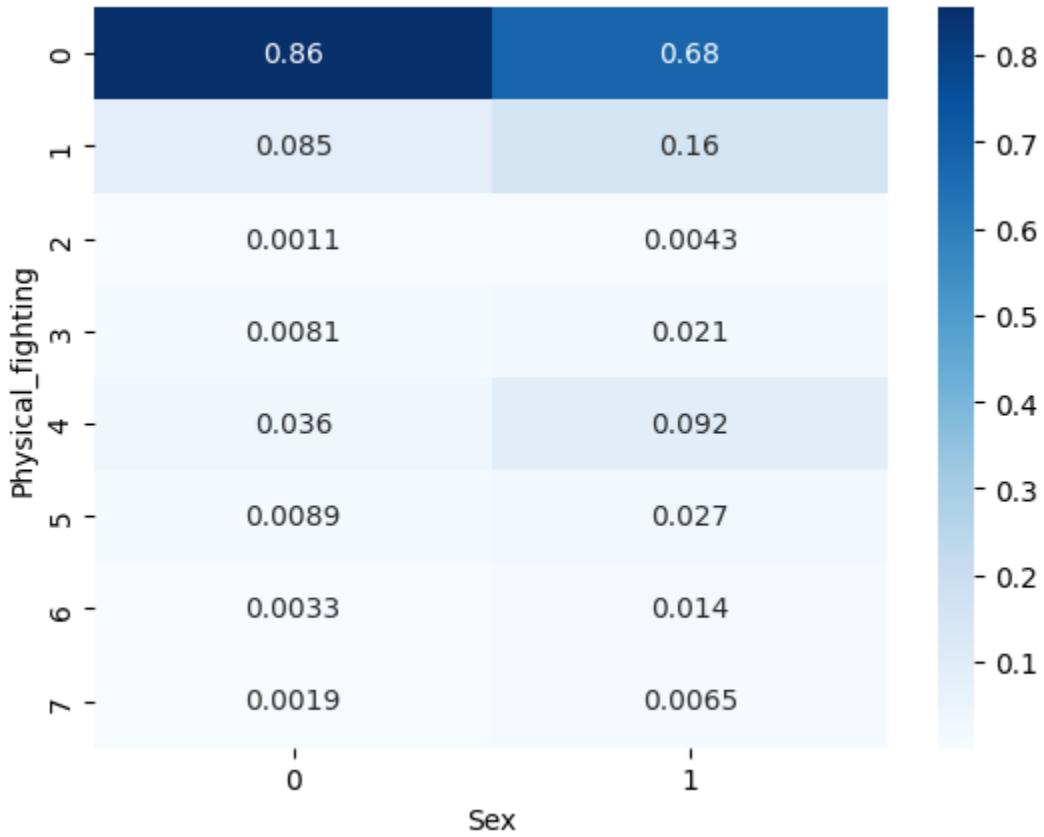






Boys

Girls




```
Pearsonr Test:
stat=0.19275106778658585, p=3.795498180299108e-273
```

Being bullied probably depends on body level

Pearsonr Test:

$$r = \frac{\sum(x - m_x)(y - m_y)}{\sqrt{\sum(x - m_x)^2 \sum(y - m_y)^2}}$$

$$f(r) = \frac{(1 - r^2)^{n/2-2}}{\mathrm{B}(\frac{1}{2}, \frac{n}{2} - 1)}$$

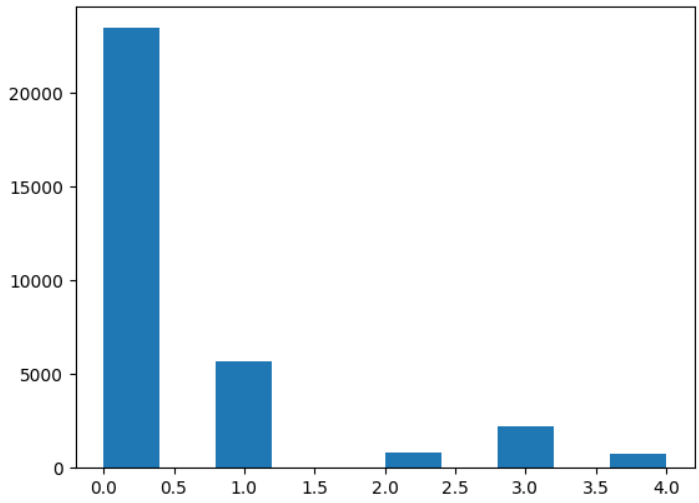
```
dist = scipy.stats.beta(n/2 - 1, n/2 - 1, loc=-1, scale=2)
```

```
p = 2*dist.cdf(-abs(r))
```

Predictive Questions:

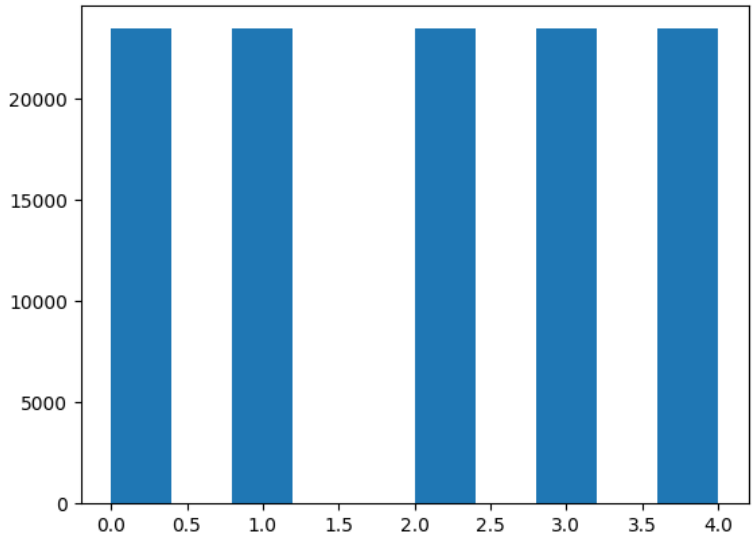
Q3: Can we predict that a student misses school with no permission given the number of close friends, parents understanding problems, feeling lonely and other students kind and helpful?

1. **Stating the question:**
   1- Expectations:
      Question is not answered.
   2- Collect data:
      Browse the web.
   3- Comparing data and expectations:
      Match. None has answered this question.
2. **EDA:**
   1- Expectations:
      Classes are balanced.
   2- Collect data:
      The classes are not balanced.



3- Comparing data and expectations:

Mismatch. Perform over sampling to balance classes.



3. Building a model:

1- Expectations:

Selected features are highly correlated with the label.

2- Collect data: Correlations are very small.

3- Comparing data and expectations:

Mismatch. Try other features.

4. Interpreting results:

1- Expectations:

A student having close friends and doesn't feel lonely won't like to miss school with no permission.

2- Collect data:

```
Accuracy = 31%
```

3- Comparing data and expectations:

Mismatch. The features can't predict the label.

5. Communication:

1- Expectations:

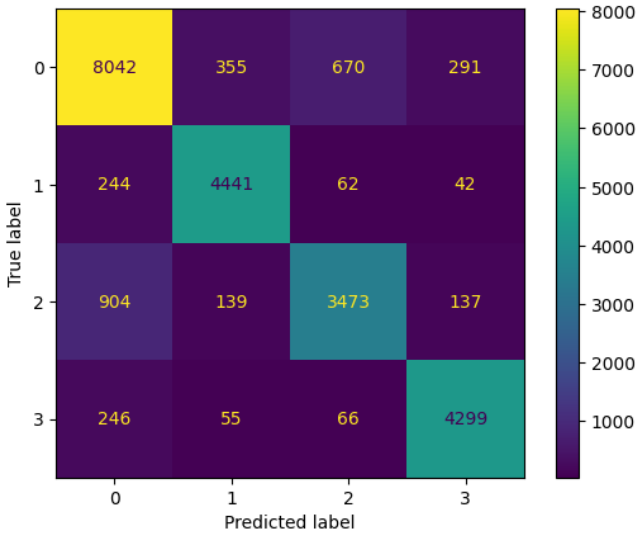
The report would satisfy the boss.

2- Collect data:

Boss asks if we can predict the label given all features.

3- Comparing data and expectations:

Use all other features to build the model and get accuracy 86%.

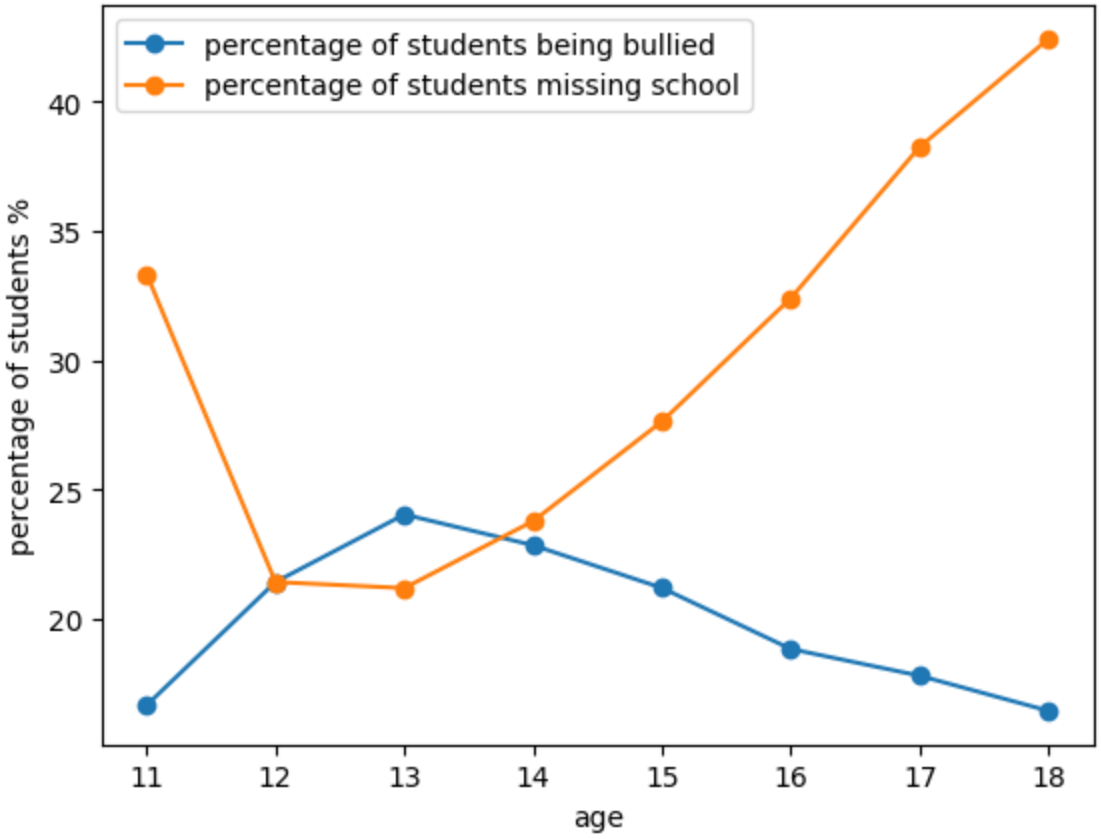


Are bullying and missing school going to increase in the future?

We considered the age here to be representing time and then tried to find how bullying and missing school are behaving based on it to try picking a pattern and predicting the future.

Our assumption here is that if bullying increases over time that would lead to an increase in skipping school as well.

We checked the percentage of bullying in each grade and the percentage of skipping school in it.

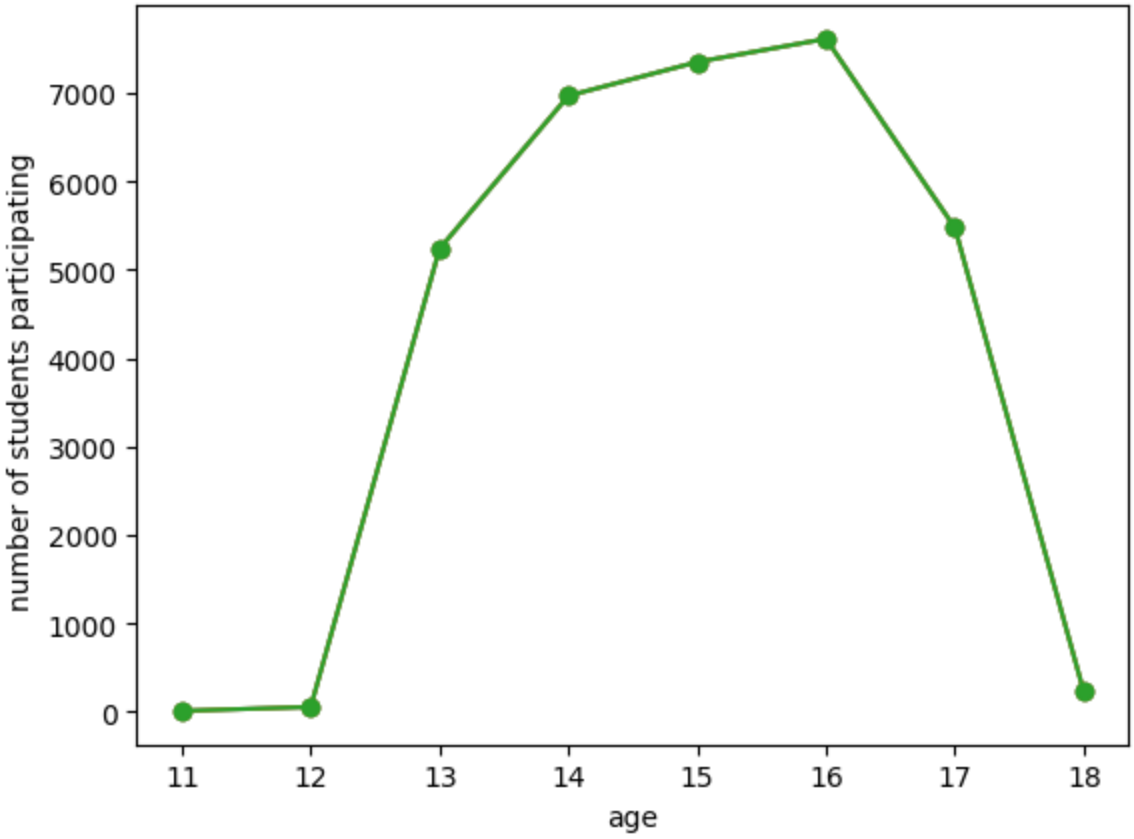


The results as we can see were quite interesting since bullying started very low and started increasing with each grade till we got to a peak and it started declining since then.

What was not expected though was that we could see a complete opposite to that in missing school.

We can attribute that behavior to the fact that more missing school alone minimizes the chances of bullying since the bullying and the bullied can be among the absent students.
And probably with age children mature and realize the consequences of their actions and also start having more responsibilities they care more about other things. And school and parents probably become more lenient about skipping school can also be a possible factor.

Last point we tried to see the number of people from each grade in our data to see how reliable their statements can be and this is what we found:



As we can see the majority of responders were of the age between 13-17 years old and based on that we can put more weight on that particular time and so we see the bullying decreasing and the missing school increasing.

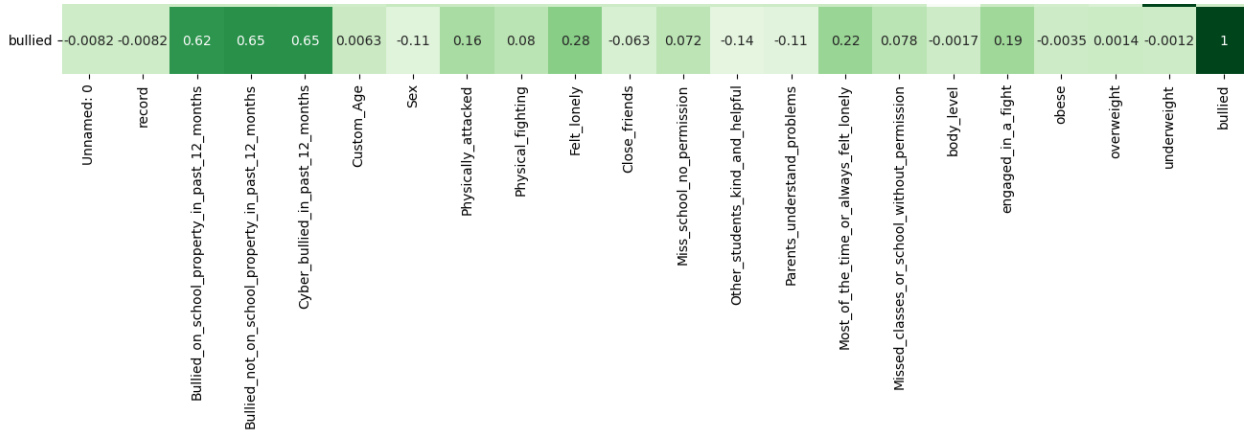
From the results we can expect that bullying declines with growing up and missing school becomes more common.

Mechanistic Questions:

Why does bullying happen?

Based on our data we expected to pinpoint the main reason for bullying aside from the individual relations between variables so we mixed all types of bullying into a new feature as 1 or true if any type of bullying happened even once.

We then checked the correlation between that feature and all other variables.



What we notice is that there is no dominant feature that we can pick as the main culprit behind bullying and even the most related features are mostly consequences to bullying rather than causes for it.
Another attempt was made to check the dependability between the bullying and the other features by using Pearson and Spearman between them.
Our hypothesis is that there is a relation between them.

```
Close friends:
      Pearson
            stat=-0.063, p=0.000
            Probably dependent
      Spearman
            stat=-0.069, p=0.000
            Probably dependent

Body level:
      Pearson
            stat=-0.002, p=0.763
            Probably independent
      Spearman
            stat=-0.001, p=0.862
            Probably independent

Other students kind and helpful:
      Pearson
            stat=-0.136, p=0.000
            Probably dependent
      Spearman
            stat=-0.069, p=0.000
            Probably dependent

Custom Age:
      Pearson
            stat=0.006, p=0.252
```

```
            Probably independent
      Spearman
            stat=0.006, p=0.247
            Probably independent

Sex:
      Pearson
            stat=-0.108, p=0.000
            Probably dependent
      Spearman
            stat=-0.108, p=0.000
            Probably dependent

Parents understand problems:
      Pearson
            stat=-0.112, p=0.000
            Probably dependent
      Spearman
            stat=-0.112, p=0.000
            Probably dependent
```

So bullying may still be dependent on some of the given features even with the not so high correlation.

We can conclude that all the features combined are having their own share of relevance but we may need more data to get a more concrete definite answer.

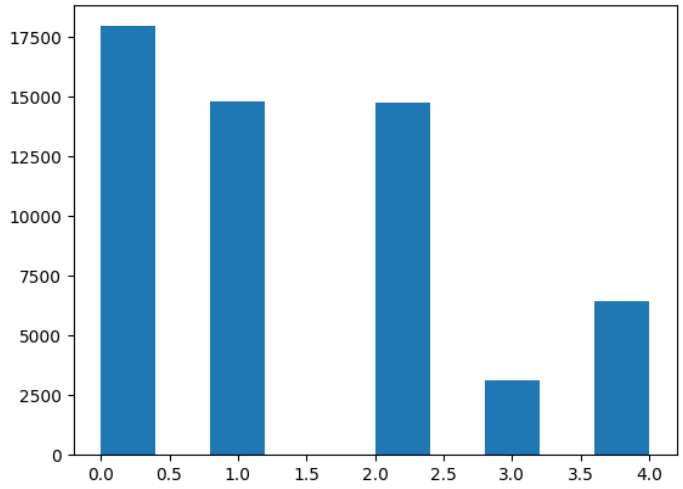
Causal Questions:

1. **Does increasing close friends lower the probability of being lonely?**

   Expectation will close friend have a negative effect of feeling lonely
   And the random forest module accuracy

	precision	recall	f1-score	support
0	0.35	0.79	0.48	2012
1	0.00	0.00	0.00	1725
2	0.31	0.35	0.33	1787
3	0.00	0.00	0.00	320
4	0.00	0.00	0.00	744
accuracy			0.34	6588
macro avg	0.13	0.23	0.16	6588
weighted avg	0.19	0.34	0.24	6588

Distribution of felt lonely



Hypothesis test person is

```
stat=0.061, p=0.000
Close_friends dependent on Felt_lonely
```