**Data analysis:**

**Step 1**

1. **Which courses are the most difficult/easy?**

Take the median of each course and check out which one is the highest and which is the lowest. The median is more resilient to data outliers than the mean. It will tell us where the average student will fall. Less affected by high scorers and low scorers.

How to:

1. Sort the array or collection of data in ascending order.
2. Find the middle value. If the number of data points is odd, the median is the middle value. If the number of data points is even, the median is the average of the two middle values.

We can also check the average grade in a course to get a better understanding of the data since it is sensitive to all data points.

How to:

1. Sum up all the numbers and then divide the sum by the total count of numbers.

Maybe use spread to check for outliers.

We can also check the percentage of people above passing grade. The percentage above passing can help assess the overall difficulty of a course. If a high percentage of students are above passing, it suggests that the course may not be as challenging regardless of the average grade.

How to:

1. Determine the passing grade or threshold.
2. Count the number of values in the array that are above the passing grade.
3. Calculate the percentage of values above the passing grade relative to the total number of values in the array.
4. **Which students graduated cum laude?**

Check which students have a gpa over 8.0.

How to:

* Same as the method before but with a higher threshold

1. **Are there courses that seem similar or related?**

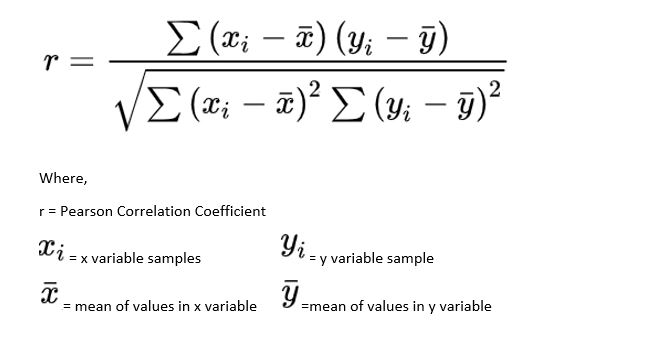
Check the similarity by comparing them head-to-head in terms of grades that were gotten there direct proportional or inverse proportional.

Create a correlation matrix manually by comparing the grades of different courses for the same students. Calculate correlations using a method like the Pearson correlation coefficient. High positive correlations suggest related courses, while high negative correlations suggest inversely related courses.

As example we can have something like math where you have a 10 and that is directly proportional with programming where you also have a 10 that means that they re related or the same case where you have a 10 at math and a 5 at gym class

How to :

1. Calculate Means for X and Y
2. Calculate Sum of Products and Sum of Squares
3. Calculate deviation
4. Calculate Pearsons correlation:



Why:

* Simple and intuitive, Linear Relationships, Standardized, Easy to Interpret.

**Question 4: Is there any suspicion of fraud between students?**

Is there any correlation between student ids and the gpa they have

as a means to check for fraud

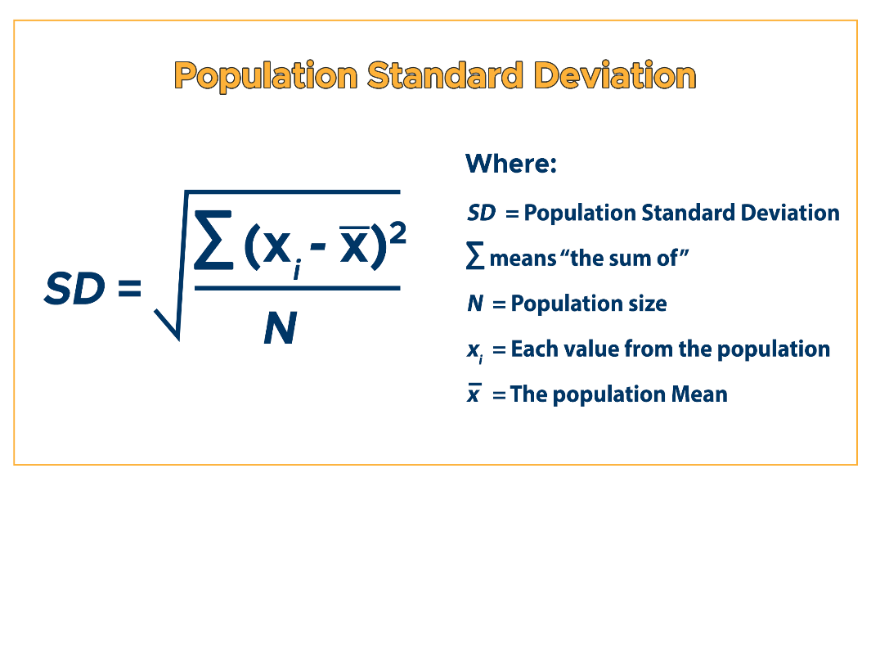
example:

we could have student id #1936 which has a 9.2 gpa and student id #1937 who has 9.3 which is a questionable coincidence. With that we can count the number of times this coincidence happens and if it's way too often then investigation under suspicion of fraud should be undergone.

**Other possible questions:**

**Is there a significant variation in grades within a student?**

* Calculate the standard deviation of grades for each subject to measure variability.



**Are there any subjects where grades tend to cluster around specific values (e.g., many students getting similar grades)?**

* Explore if there are grade clusters within subjects.

**Is there a difference in the variability of grades between subjects?**

* Compare the standard deviations of grades across different subjects.

**Step 2**

1. **Is there an order to courses in which they are offered to students for example?**

Check between students if they share the same NG on the same courses but do it sequentially from the first one. You can use an array list for each course. If NG is present in the same course for all students, then one can assume that it is due to the course and not due to the individual.

How to:

1. Which courses have more NG and rank them in that order.
2. Interpret the results – if students stare the same ng on the courses
3. If it's due to the course and not the student, we can interpret that the courses are sequential
4. **Which students are going to graduate soon?**

Check which names appear the least on the array lists made beforehand

and have them written out.

?? We can also interpret this based on which students don’t have any NG courses and their average score is above the passing grade

How to:

1. If (not NG) && (average score > 5.0)
2. **How many students are eligible to graduate this year?**

Find out the reason for graduating last year which people had what gpa and the lowest one they had. Then compare it with the current grades present as of this moment with a 0.5 correction rate: a student with a gpa of 4.5 and the minimum gpa is 5 then it should not be problematic as we expect him to pass.

Check whether they can graduate if they have a really low grade from a certain course, using the graduates as a reference point.

How to:

1. Average and lowest GPA from the graduated students ( this way we know what a passing score for the school)
2. From the graduated students check what is the lowest score and how many subjects they have such low scores (Maybe you are allowed to fail a course to graduate )
3. Interpret the results from the finding.
4. Check which students have more than the lowest GPA recorded and apply the correction rate.
5. **Is it possible to predict passing percentages for the open courses?**

Yes using the other grades that the particular students have at their disposal. If they have all around 10 then we expect a student to pass every other course. If a student's gpa is for example over 7 we expect them to pass.

* Check the average of the graduates to get an estimate to compare to this year
* Check if there is a correlation with the available data between other subjects

**Possible idea:**

Set a difficulty rating on the courses: Using the student grades one can see where the general population struggles more. Doing an average of all the grades in each course could prove useful for this. Using box plots which present outliers, the mode, the mean and the median.