Data Cleaning

Adjust Column Width: To have a comprehensive view on the data
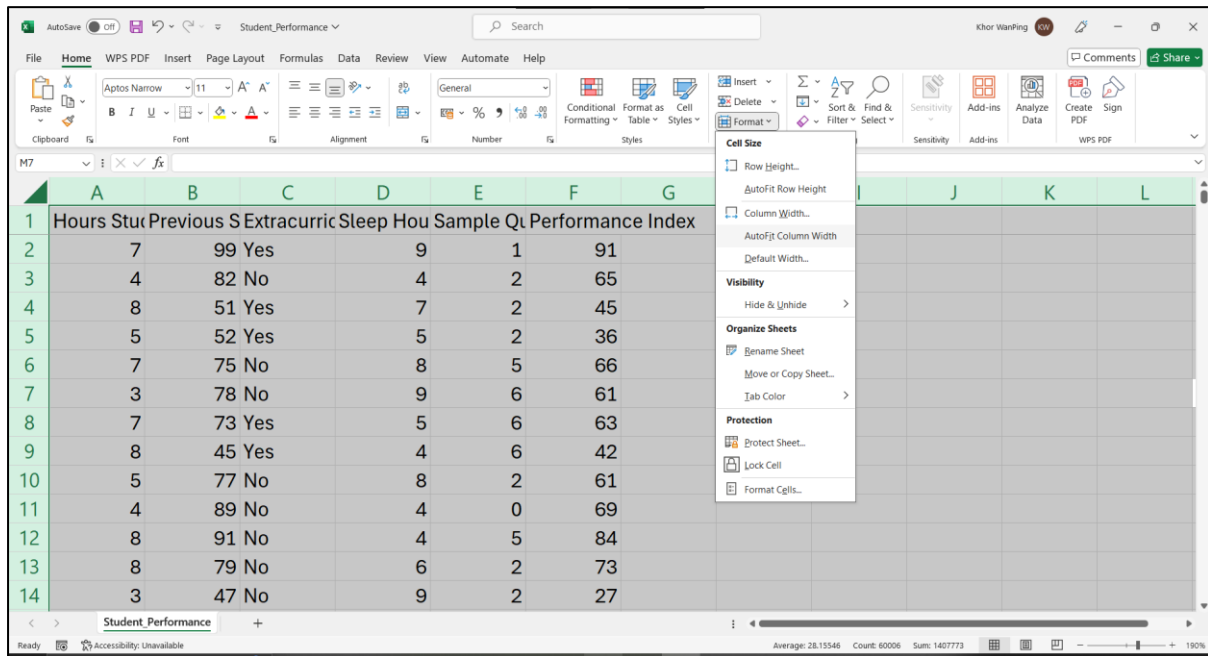

Figure 3.1.1 Adjust column width
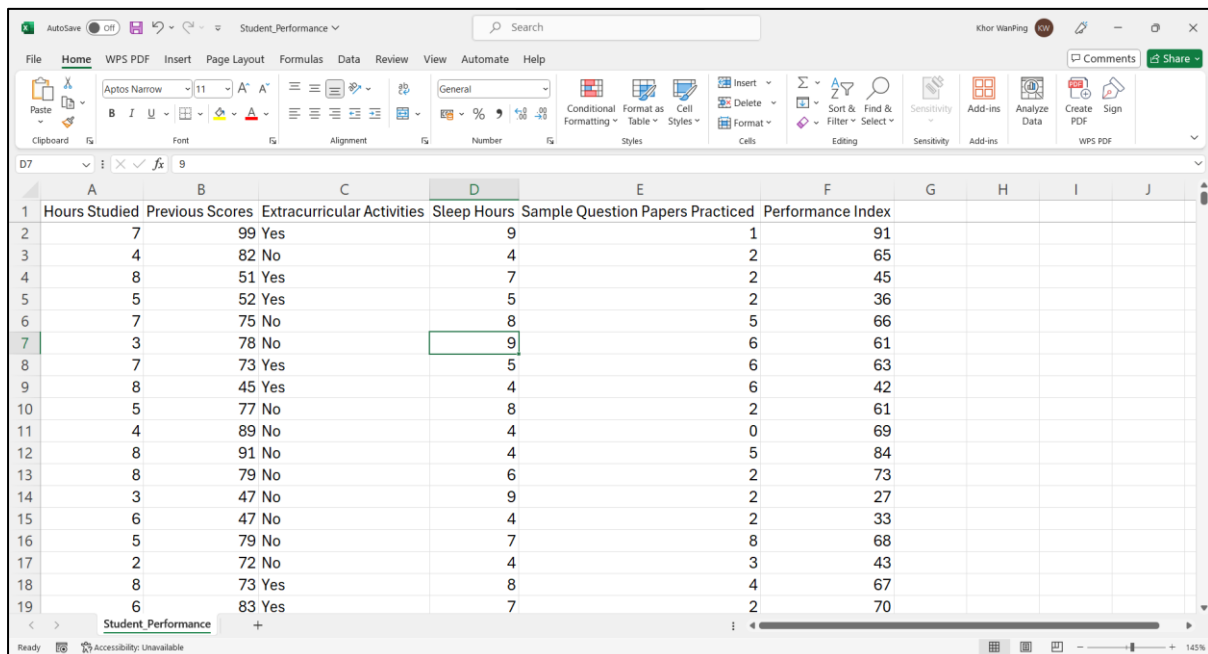

Figure 3.1.2 Adjust column width

Spelling Check: To ensure that the data analysis can be done correctly
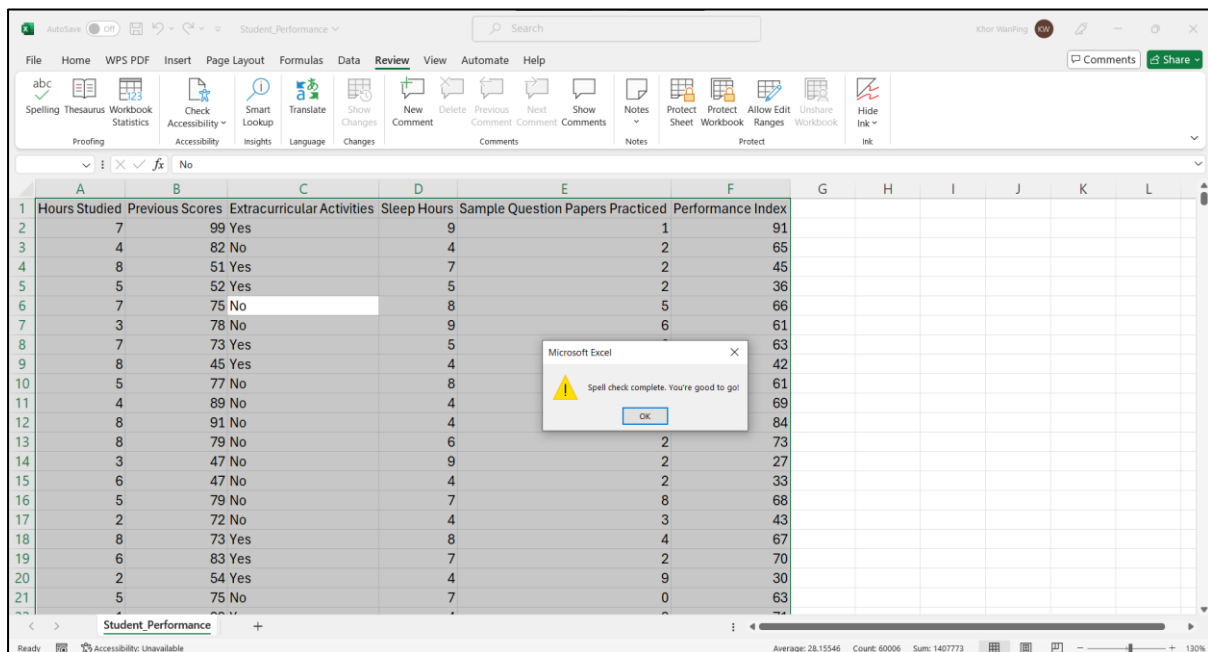

Figure 3.2.1 Spell checking

Sorting: Sort to ascending based on the 'Study Hours' to make it easier to analyze patterns and correlations between study time and other variables

Figure 3.3.1 Sorting with 'Hours Studies'

Figure 3.3.2 Sorting with 'Hours Studies'

Create Table: Converting the dataset into table format allows for easier visualization, comparison, and analysis of the data

Figure 3.4.1 Insert as a table format

Figure 3.4.2 Insert as a table format

Data Process

Data Attributes

Attributes	Description	Data Types
Hours Study	Collect the number of hours students are spending on study each day	Ratio | Integer
Previous Scores	List scores of students obtains in previous test / exam	Ratio | Integer
Extracurricular Activities	Whether the student is participating in any extracurricular activities	Categorical (Nominal) | Boolean
Numeric representation for Extracurricular Activities	Show '1' if the student is taking part in extracurricular activities; '0' vice versa	Interval | Integer
Sleep Hours	The number of hours student sleep in a day	Interval | Integer
Sample Question Papers Practiced	Number of practices done by the student	Interval | Integer
Performance Index	Student's overall performance in current test / exam	Interval | Integer
High Performer	Whether a student is a high performer based on their performance index	Interval | Integer

Table 4.1 Data attributes included

Data Processing

The data are pre-processed before being used for data analysis. For better analyze, some columns were added into this dataset. The first one is 'Numeric representation for Extracurricular Activities', it is used to represent the 'Extracurricular Activities' in numeric format ('1' and '0'). By doing this, it will reduce the complexity and increase the evenness of the dataset.

Extracurricular Activities	Numeric representation for Extracurricular Activities
Yes	1
No	0
Yes	1
Yes	1
No	0

Table 4.2 Sample of 'Numeric representation for Extracurricular Activities

In addition, 'High Performer' show '1' when the student's 'Performance Index' is greater than 81; '0' when less or equal to 81, which is high performance threshold. The threshold is calculated by using ` =PERCENTILE.EXC('Performance Index', 0.9)`, to calculate the top 10% of performance index.

Performance Index	High Performer
91	1
65	0
45	0
36	0

Table 4.3 Sample of 'High Performer'

In conclusion, steps above are executed to present a neatly, completely and easier to understand dataset to make it better to use and analyze after.

Screenshot of dataset

Hours Studied	Previous Scores	Extracurricular Activities	Numeric representation for Extracurricular Activities
7	99	Yes	1
4	82	No	0
8	51	Yes	1
5	52	Yes	1
7	75	No	0
3	78	No	0
7	73	Yes	1
8	45	Yes	1
5	77	No	0

Figure 4.1 Screenshot of dataset

Sleep Hours	Sample Question Papers Practiced	Performance Index	High Performer
9	1	91	1
4	2	65	0
7	2	45	0
5	2	36	0
8	5	66	0
9	6	61	0
5	6	63	0
4	6	42	0
8	2	61	0

Figure 4.2 Screenshot of dataset (cont.)

Data Analysis and Result Discussion

Descriptive Statistics

Result:

	Measure of central tendency			Measures of variability							
	Mean	Medina	Mode	Range		Variance	Standard Deviation	Quartile			
				Min	Max			Q1	Q3	IQR	
Hours Studies	4.99	5	1	1	9	6.70	2.59	3	7	4	
Previous Scores	69.4	69	54	40	99	300.78	17.34	54	85	31	
Sleep Hours	6.53	7	8	4	9	2.88	1.70	5	8	3	
Sample Question Paper Practiced	4.58	5	6	0	9	8.22	2.87	2	7	5	
Performance Index	55.2	55	67	10	100	369.12	19.2	40	71	31	

Table 5.1 Statistic table of dataset

Discussion:

This insight offers a deep understanding into all aspects of student life, including study hours, academic performance etc. By analyzing the measures of central tendency and variability, we can fully understand the student behaviour and performance.

In terms of study hours, student average study for 5 hours in a day, and the median was also 5 hours, indicating that the distribution of study hours is relatively symmetrical. However, the mode represents that most students only study for 1 hour a day. There was a significant

variance in the study habits as shown by the 8-hour study time range and 2.59-hours standard deviation.

The attribute with the prior scores had a mean of 69.4 and a median of 69. The mode, which is 54, is noteworthy because it is lower than the mean and median. There was a substantial variation in academic achievement, as seen by the range of 59 and the standard deviation of 17.34. The mean and median of the students' performance index were 55.2 and 55, respectively. While the mode of this attribute was 67, which was higher than the mean and median, show a right-skewed distribution. The range of it is 90 with a standard deviation of 19.2, which shows a significant difference in student performance level.

In terms of sleep hours, students sleep about 6.53 hours per day, and most of them sleep for 8 hours. Sleep duration ranged from 4 to 9 hours with a standard deviation of 1.70 hours. Some students may be sleep deprived while others are better rested.

Students practiced an average of 4.58 sample papers, with a median of 5. The mode is 6, indicating that many students practiced a slightly higher number of papers. The range and standard deviation showed significant differences in how students practiced the sample papers, with some students not practicing and some practicing as many as 9 papers.

Overall, the data revealed significant differences in students' study time, previous grades, sleep time, practice sample papers, and performance indexes. These results reflect the diversity of student behavior and performance:

The study time shows that the students' study habits are diverse; Previous grades and performance indices indicate large differences in academic performance and performance levels; Sleep time showed that most students had enough sleep, but there were some variations. The number of sample practice papers shows that students' preparation strategies are diverse.

This data can help educators understand a wide range of differences in student behavior and performance, which can identify areas that need support. For example, disparities in grades can be reduced through targeted academic support, and sleep problems can be addressed through wellness programs.

Sampling and estimation

Result:

High-Performance Threshold	81
The number of high-performers	954
Alpha	0.05
Sample proportion	0.1
Sample size	10000
Confidence Interval	95%
z-value	1.96
standard error	0.003
Margin of error	0.006
Upper Limit	0.101
Lower Limit	0.090

Figure 5.7.1 Estimation of the proportion of high-performance students

Discussion:

Based on the results, about 10% of the students obtain a high-performance index (score greater than 81). The genuine percentage of top performers in the population is assessed to be between 9% and 10.1% with a 95% confidence level. This result suggests that approximately one in ten students is a high performer. A standard error of 0.003 indicates that the margin of error between the sample proportion and the overall proportion is small. The margin of error is 0.006, so we have a high level of confidence in the forecast for this.

Getting 8 to 10 hours of sleep every night has a major impact on academic achievement, according to the National Sleep Foundation. High performers might have improved sleeping and study techniques. Furthermore, B. Lovato and M. Lack's (2010) study discovered a strong relationship between enhanced cognitive function and academic achievement and good sleep.

Result:

Sample Mean	55.22
Sample Standard Deviation	19.21
Sample size	10000
Alpha	0.05
t-value	1.96
Error	0.377
Upper Limit	55.601
Lower Limit	54.848

Figure 5.7.2 Prediction intervals

Discussion:

A new student's performance index has a prediction interval of [54.848, 55.601], which means that we have a 95% confidence interval in which their performance will fall. The sample means deviate slightly from the population means. With a 0.377 margin of error, the estimate appears to be rather accurate. Credé, Roch, and Kieszczynka's (2010) study found a strong correlation between academic achievement and attendance in class. Increased attendance rates have the potential to improve students' overall performance by boosting their grasp of the course material. In addition, T. J. Cleary and B. Platten (2013) discovered that self-regulation techniques and goal setting greatly enhance academic achievement.
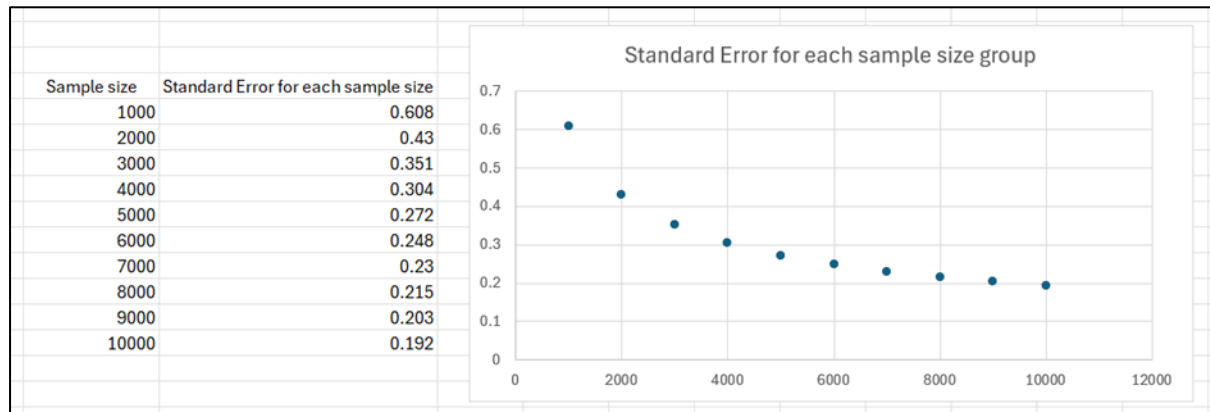
Result:

Sample size	Standard Error for each sample size
1000	0.608
2000	0.43
3000	0.351
4000	0.304
5000	0.272
6000	0.248
7000	0.23
8000	0.215
9000	0.203
10000	0.192

Figure 5.7.3 Sampling error

Discussion:

The graph indicates that the standard error falls with increasing sample size. This indicates that higher sample numbers increase the estimate's accuracy. The Central Limit Theorem states that as sample size increases, the sample means approaches the population mean and the standard error decreases, improving the accuracy and dependability of our estimations. Sample size has a major impact on statistical analysis, particularly when estimating population parameters, as Sirin (2005) pointed out. Greater findings confidence and effective error reduction are achieved with larger sample numbers.

Discussion:

Through these three analyses, we can effectively help educational institutions, educators and administrators in making decisions. First of all, the application of high-performance proportion can be used to develop targeted support programs. By identifying and assisting students who are about to achieve at a high level, more resources and tutoring can be made available to raise academic standards more broadly. Additionally, Prediction intervals give educators and administrators a realistic expectation for the academic achievement of incoming pupils. To guarantee that the majority of students fulfil the required academic requirements, this can be used to inform course design, modifications to teaching strategies, and resource allocation.

In conclusion, educational administrators can use the data from these findings to gain a better understanding of and ability to enhance student academic performance. Through a comprehensive understanding of the variables that impact high performance, like sleep patterns, study habits, and socioeconomic status, educational institutions may design and execute focused interventions that improve overall student outcomes. A wider framework for understanding these results is provided by comparisons with other studies and real-world data, emphasizing the significance of helping high-achieving students to uphold and raise academic standards.