|  |  |
| --- | --- |
| Generated image  Group Project Team 7 | **Topic**  How Student Sleep Patterns and Lifestyle Affect Academic Performance |

Conestoga College

School of Applied Computer Science & Information Technology

SENG8081 - Case Studies Big Data

**PROF: - DAVID MARSH**

How Student Sleep Patterns and Lifestyle Affect Academic Performance

**Group members: Andrews Owusu**

**Kulwinder Singh**

**Jai Vadhani**

**Shivani Shivani**

**Abstract**

Sleep is a fundamental component of student well-being and academic success, yet it is often overlooked in higher education environments. This project investigates how sleep patterns, along with lifestyle habits such as screen time, physical activity, study hours, and caffeine consumption affect the academic performance of university students. Using three publicly available datasets (from Kaggle, Mendeley, and RPubs), the analysis aims to uncover meaningful relationships between sleep quality and some lifestyle factors; and how these variables correlate with performance metrics like Grade Point Average (GPA).

The data for the project is organized and managed in a SQL Server relational database to support structured querying, integration, and efficient filtering. Exploratory data analysis, statistical methods, and data visualization tools like Python, Tableau, and R are applied to identify trends and correlations. Key research questions include whether higher screen time contributes to poor sleep, how physical activity influences sleep length, and whether students with low sleep quality perform worse academically. The findings aim to support students, educators, and wellness professionals in recognizing the importance of healthy sleep as a driver of academic and personal success.

**Contents**

[Introduction 4](#_Toc204894662)

[System Diagram 5](#_Toc204894663)

[Data Research and Integration 6](#_Toc204894664)

[Data Cleaning 7](#_Toc204894665)

[Data Storage and Maintenance with SQL Server 8](#_Toc204894666)

[Data Maintenance 9](#_Toc204894667)

[Data Quality 9](#_Toc204894668)

[Data Analysis & Visualization: 11](#_Toc204894669)

[Extension 12](#_Toc204894670)

[Project Timeline 13](#_Toc204894671)

[References 14](#_Toc204894672)

# Introduction

Sleep is a critical yet often overlooked factor in the academic journey of university students. Amid busy class schedules, late-night studying, and high levels of digital engagement, many students experience irregular or insufficient sleep. While it is widely accepted that proper sleep contributes to physical and mental well-being, its role in academic performance deserves closer examination, particularly in the context of student life where stress, lifestyle habits, and environmental pressures constantly interact.

This project explores how sleep patterns relate to various lifestyle choices, including screen time, study duration, physical activity, and caffeine intake. Our goal is to analyze whether these behaviors have a measurable impact on both sleep quality and academic indicators such as grade point average (GPA), stress levels, and overall student performance. By using real-world data, and also synthetic data generated based on real-world scenarios, we aim to uncover patterns and correlations that may explain how quality sleep may contribute to academic excellence.

To achieve this, we have gathered and analyzed data from three publicly available sources: Kaggle’s Student Sleep Patterns dataset, Mendeley’s Insomnia and College Performance dataset, and a structured dataset from RPubs. These datasets have been integrated into a SQL Server relational database for efficient organization and querying. Through statistical analysis and data visualization, we aim to draw insights that not only enhance our understanding of student sleep behavior but also inform future strategies for promoting academic success through improved wellness practices.

# System Diagram

A representation of the system is shown in the diagram below. It represents the end-to-end flow of the system from extraction of data from original sources, transformation/cleaning of the data and loading it into the SQL database for permanent storage, and the generation of insights from the stored data.

A diagram of a computer program

AI-generated content may be incorrect.

# Data Research and Integration

In order to determine appropriate datasets, we used search terms such as "college sleep behavior", "student sleeping patterns", and "sleep and academic performance" on Google Dataset Search, Kaggle, GitHub, and RPubs. This helped us to locate datasets that aligned with our project expectations. Having evaluated a few sources, we chose three sets of data that are close to our study objectives:

* Student Sleep Patterns (Kaggle, by Arsalan Jamal):
* Insomnia of students and college performance (Mendeley):
* RPubs Sleep Study Dataset:

The dataset contains information on students’ sleep duration, quality of sleep, screen time, physical activity, and overall health, among other factors. Each row represents a student, with columns which include their age, gender, university year, sleep duration and timing (weekday and weekend start and end times), study hours, screen time, caffeine intake, physical activity levels, and self-reported sleep quality. It includes a manageable number of records and is available in a clean CSV format. It has consistent data types, making it easy to load and process using Python libraries like pandas or visualization tools such as Tableau or Power BI.

* **Student Sleep Patterns Dataset (Kaggle, by Arsalan Jamal):**

The data has a good structure and the variables present in the data are the hours of sleep, quality of sleep, the time spent using screens, stress, study time, and physical activity. It is best suited to conduct visual analysis to determine some patterns of student sleeping behavior.

(JAMAL, 2024)

* **Insomnia of students and college performance Dataset (Mendeley):**

This is a collection of the survey data of 985 students that covers information regarding the frequency of insomnia, their GPA, stress, and characteristics of their lifestyle. It offers great insights concerning the effect of sleep disorders on academic performance.

(Abdullah, 2025)

* **RPubs Sleep Study Dataset:**

This dataset contains structured variables of sleep in a clean form, which is appropriate in regression, correlation analyses, and visualizations.

(Seng, 2024)

# Data Cleaning

The three sets of data that we obtained were in .csv format and using Python and the pyodbc package, we loaded these datasets into the SQL Server database.

Dataset Used :

* Student\_Sleep\_Patterns.csv
* Cleaned\_Student\_Insomnia\_Dataset.csv
* Cleaned\_SleepStudy.csv

We performed data cleaning activities to ensure that all the datasets were good to perform the analysis easily. Our data cleaning process was performed with Python, using the Pandas Library.

Generally, we followed these steps to clean our datasets:

* Deleted all the missing and duplicated records to maintain the consistency of the data.
* Column names were renamed to lowercase and those that had spaces in the column names were replaced with underscores to be SQL and R-friendly.
* Round continuous numeric variables in sleep\_hours, screen\_time and study\_time to 1 decimal point to enhance easy reading and understandability.
* Columns having long survey text were replaced by a correspondingly short variable name.
* Removed the Timestamp column that was not useful with respect to analysis.
* Converted survey data into an ordered set of categorical data to preserve logical order of the responses (for example: “Never”, “Rarely”, “Often”, “Always”).
* Cleaned and got rid of duplications, storing the new clean data as a new CSV, and also inserting the cleaned data into SQL Server.

# Data Storage and Maintenance with SQL Server

In the given project, SQL Server was selected as the main data storage tool because of its relational organization and ease of analysis with the help of data analysis applications, such as RStudio and Python. This was aimed at making sure that the cleaned and normalized datasets were safely stored, readily accessed and also effectively arranged so as to be able to perform analytical operations upon which one can filter, merge and query.

In the preparatory phase of data cleaning, the python and RStudio were used to process and refine those files and write them back as cleaned .csv files. Such CSV files were loaded into SQL Server tables to store them longer and have handy querying.

**Tools: -**

* SQL Server- This is employed in unstructured storage of data and performing relational queries.
* pyodbc a Python module that helped connect to SQL Server, which was used to insert and retrieve data.
* RStudio (DBI + odbc) - It gave data access within R in the form of SQL to inspect and analyze visual expressions in R.

**The advantages that the SQL Server holds in this project are:**

* Relational Joins: Joins are extremely easy to create with common for more multi-dimensional queries.
* Data Integrity: SQL constraints and schema definitions was used to ensure that records had standard format and structures.
* Centralized Updates: It was possible to query updated tables by all team members without creating any data silos or version conflicts.
* Analysis-Ready Format: All the tables were queried in RStudio to obtain the descriptive statistics, visualizations, and correlation tests.
* Back-up and Security: SQL Server has in-built back up and user access facilities to ensure that the data was secure and that an option of back-up available.

# Data Maintenance

**Why SQL Server Over Other Tools?**

Whereas small footprint-based tools such as SQLite or cloud-based solutions such as Google BigQuery are flexible, we chose to use SQL Server because of its:

* The ease of deployment use of RStudio and Python development in the local environment
* Powerful data querying functions in connection to relationships.
* Strong support of structured data.

# Data Quality

This was also one of the important aspects of this project because the quality, reliability and consistency of the datasets played significant role in determining the accuracy of our analysis. In order to guarantee the quality of the process, we used diverse approaches to data cleaning based on the structure and the form of each dataset.

**Accuracy**

The accuracy was determined by having a close look at all variables by grouping and counting. This would enable us to identify any unexpected entries or mislabels of the categorical columns. Regular expressions were used in such a way that in the case of irregular answers such as Rarely (1-2x/week), but instead were interpreted as Rarely. This made the values easier to categorize and summarize.

**Completeness**

To enhance completeness, it was observed that there were some missing values and the missing values were filled in with the mode of the particular column. The reason such method was applied is that the data used were categorical survey data, and in this case, the mode (the most often provided answer) is a reasonable and non-biased substitute. Such columns as the screen time, stress level, caffeine intake were cleaned with the help of this strategy.

**Consistency**

The second step to achieve consistency between the datasets was renaming of columns to include one standard naming pattern, e.g., lowercase letters and the symbols of the underscore. The categorical values (e.g., all responses on the sleep quality) to be compared by the same methodology were also standardized (e.g., all responses on sleep quality were made in a common ordinal form), which played a substantial role in merging the data and performing it in SQL and R through the grouped analysis.

**Reliability**

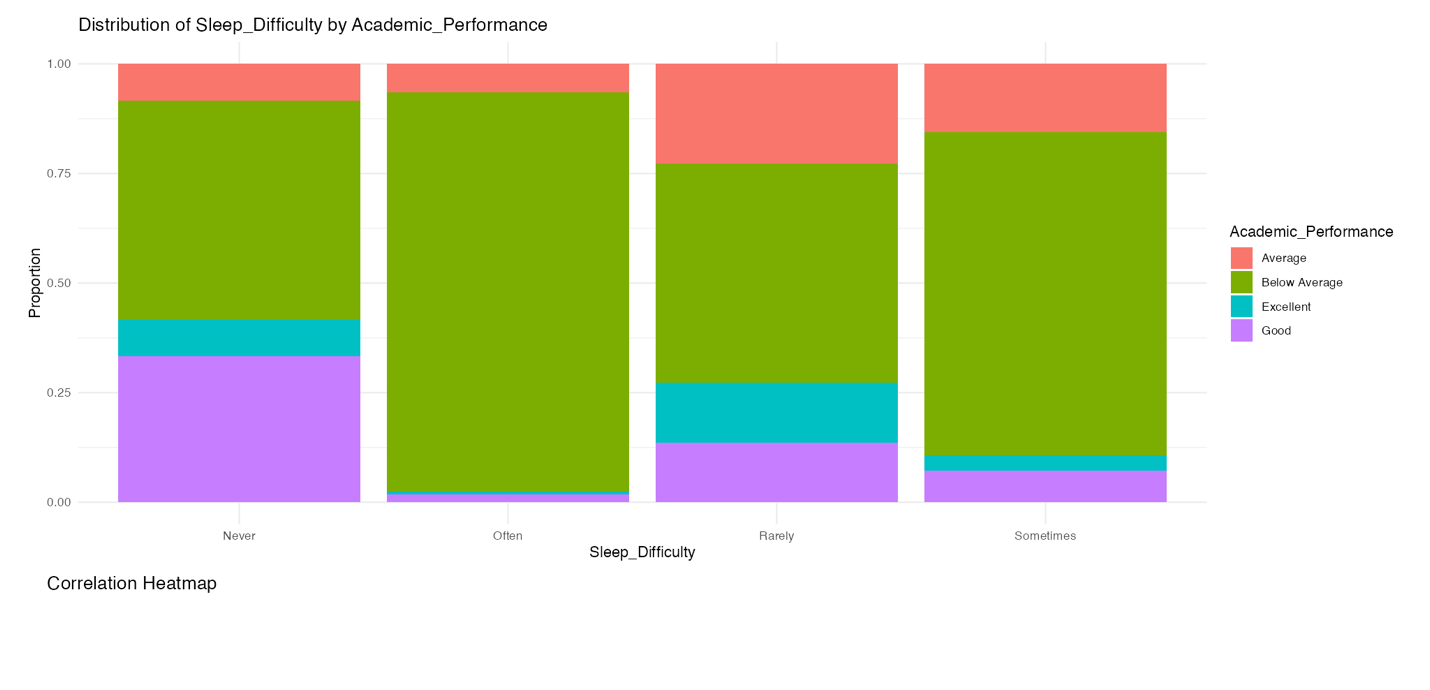
We increased reliability by cleaning up similar records and transforming variables to realistic data type. For example, putting GPA as a numeric value and sleep quality as a ordered factors. This facilitated appropriate statistical modeling and grouping of visuals.

**Validation**

Finally, we applied SQL commands to verify whether the final cleaned tables were accepted to check the frequency of distribution and detect any outliers or erroneous data. Such validation tests helped to make the data logically correct and prepared to be used in RStudio to run logical analysis.

# Data Analysis & Visualization:

Distribution of Sleep by Academic Performance (sample visualization)



## Findings:

From our analysis, we found that **sleep quality plays a more significant role in academic performance than simply the number of hours slept**. In other words, students who reported better quality sleep tended to perform better academically, regardless of how long they slept. Conversely, we observed that **longer sleep duration alone does not have a strong positive correlation with academic performance s**uggesting that just sleeping more is not necessarily beneficial unless the sleep is restful and uninterrupted.

Additionally, the data indicates that **regular, or increased alcohol consumption is associated with lower academic performance**, highlighting a potential negative impact of alcohol use on students’ ability to achieve higher grades.

# Extension

This project has limitations and for the future, it could be improved by:

* Full Automation: currently the process is not fully automated. The Python script is run to perform the data cleaning logic and load the data. Then RStudio is connected to get the data for analysis. A full end-to-end automation with modern tools will improve the workflow significantly.
* IoT Integration- Smartwatch or a fitness band to record the sleep and fitness data in real-time.
* ML Models - Decision trees or logistic regression are employed to determine how the GPA can be predicted on the basis of lifestyle patterns.
* Student Dashboard - Develop a web-based application that presents insights to help students develop better behavior.
* Real-Time Sync Even the data you get daily through wearable apps or mobile sleep trackers can be imported automatically into the SQL Server.

# Project Timeline

|  |  |  |
| --- | --- | --- |
| **Date** | **Deliverable** | **Responsible** |
| MAY 28 | Resource finding dataset | ALL |
| JUNE 04 | Cleaning & Integration in Python/SQL | Jai, Kulwinder |
| JUNE 12 | Store data in the database | Shivani |
| JUNE 15 | Midterm report | ALL |
| JUNE 15 | Further cleaning of data | Jai, Kulwinder,Shivani |
| JUNE 22 | System Diagram | Andrews |
| JULY 06 | Analysis | Shivani,Andrews |
| JULY 20 | Visualization | Andrews, Shivani |
| JULY 29 | PPT | Kulwinder,Jai |
| JULY 30 | Final Report | Jai,Kulwinder |

# References

Abdullah, A. (2025, May 16). *Student Insomnia and Educational Outcomes Dataset*. From Mendeley Data: https://data.mendeley.com/datasets/5mvrx4v62z/3

JAMAL, A. (2024, Sep 04). *Student Sleep Patterns*. From Kaggle: https://www.kaggle.com/datasets/arsalanjamal002/student-sleep-patterns/data

Microsoft . (2025, Nov 17). *SQL Server technical documentation*. From Microsoft Ignite: https://learn.microsoft.com/en-us/sql/sql-server/?view=sql-server-ver17

pandas . (2025, July 07). *pandas documentation*. From pandas : https://pandas.pydata.org/docs/

Seng, H. (2024, 12 02). *STAT 353 Project 2: Exploring College Data*. From Rpubs: https://rpubs.com/cp2185kc/1253151

Solutions. (2025, 7 25). *Best Practices in Working with Databases*. From Solutions: https://solutions.posit.co/connections/db/a

freeCodeCamp.org. (2021, September 22). *R Shiny for Data Science Tutorial – Build Interactive Data-Driven Web Apps* [Video]. YouTube. <https://www.youtube.com/watch?v=9uFQECk30kA>