Optimizing SQL Learning: Identifying Prime Study Times Using Time-Data Analysis

Sophia Yang*, Colin Li*, Abdussalam Alawini
Department of Computer Science, University of Illinois Urbana-Champaign
{sophiay2, colinli3, alawini}@illinois.edu

***Abstract*—This research full paper explores the optimal times for studying and learning Structured Query Language (SQL), a critical skill in managing relational databases across various domains, to improve learning, problem-solving, and academic performance. By identifying prime study times, productivity and retention can be enhanced, particularly by scheduling breaks when cognitive function wanes. This study analyzes over 129,000 SQL submissions from students in a Fall 2022 Database Systems course at the University of Illinois Urbana-Champaign, examining correlations between time of day and answer accuracy (correct, syntax error, or semantic error). Time series analysis was done to analyze the data collected over a set of intervals, and null-hypothesis significance testing was utilized to calculate the p-value, determining the statistical significance of the collected data for drawing confident conclusions.**

***Index Terms*—SQL; database education; online assessment; syntax; semantics; error**

I. Introduction

The Structural Query Language (SQL) is the de facto standard language for managing and querying relational databases [1]. Hence, proficiency in SQL emerges as a requisite skill for individuals engaged in database-related tasks, spanning researchers, developers, and similar professionals [2, 3]. Featuring an English-like syntax, SQL offers a user-friendly entry point for novices, eliminating the need for extensive prior programming knowledge and potentially fostering inclusivity in computing fields for non-Computer Science students. Despite these advantages, persistent learning challenges in acquiring SQL skills prompt the need for further investigation [2]. Specifically, our research aims to explore optimal times for students to engage with SQL homework assignments by analyzing the types of errors encountered during different hours of the day.

Our study seeks to build upon previous research that indicates an optimal start time for learning between 11 a.m. and 1 p.m. [4]. Additionally, the impact of school start times on academic achievement has been underscored by other scholars, with delays of just 50 minutes showing a significant positive effect [5]. Improved sleep quality, characterized by longer duration and greater consistency, is linked with enhanced academic performance, accounting for approximately one-quarter of the observed variability in academic achievements [6]. Hence, our study aims to understand the best times for learning SQL by examining when students encounter errors (and which types) while solving SQL questions, similar to previous research on optimal learning times. This insight could help improve course scheduling for better learning outcomes. To do this, we analyze the SQL homework assignment, considering different problems intending to assess various SQL concepts. Our analysis of the SQL homework assignment includes both an individual problem-wise and an across-all-problems examination. Through this dual perspective, we assess student performance across various SQL concepts and investigate potential disparities in time-dependent trends.

For this analysis, we collected submissions to homework assignment questions from students enrolled in the Database Systems course during the Fall 2022 semester at the University of Illinois Urbana-Champaign, with a total enrollment of 730 students. The course is open to both undergraduate and graduate students, who may opt to take it for either learning purposes or as a technical elective within the Computer Science curriculum required for graduation. Enrollment in this course necessitates prior completion of a data structures course, typically undertaken in the second year of undergraduate study, ensuring a foundational programming background, given the integration of programming assignments within the curriculum. The dataset includes responses to the SQL homework assignment, which consists of 15 programming questions one a single assignment, all subject to auto-grading with immediate feedback. The immediate feedback indicates to the students whether their answers are correct or incorrect, with the auto-grader providing additional feedback in cases of incorrect responses, indicating whether errors are syntactic (the result of compilation errors) or semantic (pertaining to differences between student-generated output and the required data output).

All SQL-related homework assignments were released to students via PrairieLearn[7], an online learning management system equipped with an embedded autograder. This system enables comprehensive data collection, including recording every submission, timestamping each submission, and identifying the correctness of responses - distinguishing between syntax errors, semantic errors, and correct answers. With this dataset, we look at the distribution of syntax errors, semantic errors, and correct responses across different times of the day. Our study aims to address two primary research questions: 1) *Does a significant correlation exist between the ratio of correct to incorrect answers and the time of day?* and 2) *Are these patterns consistent across both fundamental and advanced SQL problem concepts?* The findings hold potential

*These authors contributed equally to this work implications for instructors in optimizing course logistics and scheduling lecture discussions to capitalize on students' concentration and retention levels.

II. Related Works

While existing research has extensively looked at the errors students encounter in procedural programming languages such as Java [8, 9, 10], C++ [9, 11, 12, 13], and Python [8, 9, 13], relatively less attention has been directed towards errors in database query languages or declarative languages like SQL [14, 15, 16, 17, 18]. The limited research in this domain primarily focuses on the SQL problems students are commonly challenged by and their errors. For instance, Taipalus et al. [16], in their analysis of over 33,000 SQL queries submitted by students, observed a range of syntax, semantic, and logic errors. While some of these syntax errors align with those documented in prior literature, such as undefined parameters and data type mismatches [14], Taipalus et al. [16] also identified novel syntax errors like "IS where not applicable," "duplicate clause," and "confusing table names with column names." Additionally, Taipalus and Perälä [17] investigated persistent error types in students' SQL queries and the associated SQL query concepts leading to such errors, while others explored the challenging SQL query types for students [19, 20, 21]. Furthermore, researchers have delved into methodologies for visualizing and detecting students' learning obstacles and approaches [22, 23, 24, 25], as well as visualizing SQL queries to enhance comprehension [26]. On the other hand, our research focuses on determining if a significant correlation exists between the ratio of correct and incorrect answers and the time of day, and whether these patterns persist across both fundamental and advanced SQL problem concepts. Rather than studying the types of errors and SQL concepts that students struggle on [14, 15, 16, 17, 19, 20, 21], we look at the time of day when students struggle with SQL problem-solving and use errors as a measure of student performance. We aim to explore the implications of our findings for optimizing course logistics and scheduling strategies to facilitate more effective learning of SQL by students.

Existing research extensively examines the influence of superior sleep quality and later start times on academic performance within high school and college contexts [4, 5, 6, 27, 28]. However, there are much less studies investigating whether specific times of day correlate with a higher frequency of correctly submitted answers in particular subjects [5, 29, 30]. Moreover, the limited research in this area predominantly focuses on broader metrics of academic achievement, such as GPA or standardized test scores. For instance, Pope [30] analyzed 1.8 million student-year observations in the Los Angeles area, revealing that scheduling Math classes earlier in the school day, as opposed to later, led to a 0.072 increase in Math GPA. Similarly, scheduling English classes in the morning rather than the afternoon resulted in a significant GPA increase of 0.032, equivalent to enhancing teaching quality by a quarter [30]. Carrell et al. [27] and Hinrichs [28] utilized diverse metrics such as SAT scores and standardized test results to explore the impact of school start times on academic performance. Notably, Hinrichs [28] examined the effects of school schedule reforms on standardized test scores, investigating whether the timing of the school day influenced testing performance [28]. While these studies employ broad measures of academic performance, our research aims to provide insights specifically into the performance of students learning SQL. Building on the findings of Pope [30], which noted the significance of the time of day in learning subjects like Math and English, we hypothesize that similar insights regarding SQL learning could inform more efficient learning strategies and course scheduling practices.

Previous research on the impact of school start times and scheduling on learning has typically been constrained by specific time intervals, dictated by the structure of their studies [29, 30, 31]. For instance, Pope [30] investigated the effects of scheduling subjects at various times within the school day, constraining data collection by the typical school hour window from 8:00 a.m. to 3:10 p.m. [30]. Due to these time constraints, this model of data collection inherently precludes the exploration of significant results in the evening or early morning. Similarly, Gaggero and Tommasi [29] focused on three exam slots: morning at 9 a.m., post-lunch at 1:30 p.m., and afternoon at 4:30 p.m., revealing significant positive performance differences at the 1:30 p.m. time slot [29]. Likewise, Facer-Childs et al. [31] examined cognitive performance at three specific times of day: 8 a.m., 2 p.m., and 8 p.m., limiting the diversity of time recordings and thereby restricting the breadth of their findings [31]. To address this limitation, we utilized PrairieLearn[7], online learning management system student submissions to gather continuous quantitative data. This platform automatically records student responses to homework questions with timestamps, enabling data collection across all 24 hours of the day. This approach facilitates a more comprehensive examination of student behavior beyond traditional school hours or restricted time intervals.

Prior research has also employed various methods to measure quantitative data, including wrist actigraphy and wearable activity trackers, in addition to online learning management platforms [6, 31]. For instance, Facer-Childs et al. [31] utilized actigraphy to classify individuals into early (ECT) or late chronotypes (LCT), commonly known as early birds or night owls. This classification informed their assessment of subjects' sleepiness using the Karolinska Sleepiness Scale [31]. Their findings revealed that ECTs outperformed LCTs across all tasks in the morning and exhibited significantly lower levels of sleepiness at 8 a.m. [31]. Similarly, Okano et al. [6] employed wearable activity trackers to gauge sleep duration and consistency when examining the relationship between sleep quality and academic performance. Their study identified a significant negative correlation between bedtime and overall scores in an introductory college chemistry class, with students sleeping earlier generally achieving higher scores [6]. Moreover, research has demonstrated that later school start times, allowing students to sleep longer in the morning, contribute to improved academic performance [5, 27]. Specifically, Luong

[5] found that students taking courses at 8:15 a.m. performed better than those at earlier time slots (such as 6:30 a.m. and 7:20 a.m.), with peak performance observed at 8:15 a.m. and 1:20 p.m. Additionally, studies have indicated that students attending their first-period class at later hours tend to perform better academically compared to those with earlier start times (0.140 standard deviations lower for students who start at 7:00 a.m.) [27]. Building upon this literature, our study aims to investigate whether late-night or early-morning hours, typically associated with sleep times, correlate with decreased academic performance during SQL learning. By leveraging our findings, we aim to offer recommendations to database instructors regarding optimal times for SQL study sessions, thereby enhancing their ability to support students effectively.

III. Methodology & Data Collection

Our dataset is collected from the Database Systems course offered to upper-level undergraduate and graduate students at the University of Illinois Urbana-Champaign. Data collection took place during the Fall 2022 semester, with a student enrollment of 730 individuals. The course followed an in-person format, implementing a flipped-classroom approach. Prior to class meetings, students were assigned pre-recorded lecture videos accompanied by quick knowledge check quizzes for review. During class sessions, the instructor reviewed the quiz solutions, solved practice problems, and addressed any student questions or misconceptions. The remaining class time was dedicated to collaborative group exercises aimed at reinforcing the concepts covered in the pre-recorded lectures.

In conjunction with collaborative assignments, students were assigned an SQL homework assignment containing 15 programming questions designed to assess abstract data operation concepts including selection, projection, grouping, aggregation, and joining. These concepts represent fundamental operations pivotal in data manipulation, transformation, and analysis across diverse database systems: selection extracts rows based on specified criteria, projection selects columns to achieve reduced dimensionality, grouping categorizes data for aggregations and summaries, aggregation combines rows for operations like summation and averaging, and joining merges data from multiple sources based on shared columns or keys. The assignment featured questions of increasing difficulty and complexity, evident in the number of concepts tested per question and the intricacy of criteria required to retrieve the targeted dataset. With a two-week deadline, students' submissions were graded based on accuracy using the embedded auto-grader on PrairieLearn[7]. Students could still earn partial credit for incorrect answers on their final attempt through manual grading by the instructors.

A. Description of SQL Problems

The autograder on PrairieLearn[7] assesses credit for SQL problems by comparing the data outputs produced by the student's SQL query with those generated by the instructor's solution query to verify correctness. In the event of data discrepancies, students can review their data outputs alongside the solution query's outputs to rectify semantic errors (resulting from logical errors in the query) or refer to syntax error codes to address syntactical issues (related to unsuccessful compilation). Students are allowed unlimited attempts to answer questions in any sequence until the assignment deadline elapses. An example of an SQL problem and its instructor solution is shown in Figures 1 and 2.

Write an SQL query that returns the ProductName of each product made by the brand 'Samsung' and the number of customers who purchased that product. Only count customers who have purchased more than 1 Samsung product. Order the results in descending order of the number of customers and in descending order of ProductName.

FIG. 1: SQL Homework Problem Statement Example

```
SELECT Pr1.ProductName, COUNT(C1.CustomerId) as numCustomers
FROM Products Pr1 NATURAL JOIN Purchases Pu1
  NATURAL JOIN Customers C1
WHERE Pr1.BrandName = 'Samsung'
  AND C1.CustomerId IN (
    SELECT C2.CustomerId
    FROM Customers C2 NATURAL JOIN Purchases Pu2
      NATURAL JOIN Products Pr2
    WHERE Pr2.BrandName = 'Samsung'
    GROUP BY C2.CustomerId
    HAVING COUNT(C2.CustomerId) > 1
  )
GROUP BY Pr1.ProductName
ORDER BY numCustomers DESC, Pr1.ProductName DESC;
```

FIG. 2: SQL Homework Solution Example

The resultant dataset comprises of 129,408 log files, collected following the Institutional Review Board (IRB) specified data safety protocols of our university to safeguard student privacy. To ensure anonymity, all identifiers were removed from the submission files, with randomized numbers allocated to represent individual students.

B. Data Cleansing & Processing

The dataset log files extracted from PrairieLearn[7] contain the timestamp information for each submission, as well as whether it was a correct submission, incorrect submission due to a syntax error (and the error code), or incorrect submission due to a semantic error.

We segmented submission times into hourly or 15-minute intervals and compiled dataframes containing homework question numbers, submission times, and correctness statuses to analyze submission patterns. Subsequently, we tabulated counts of correct, incorrect semantic, and incorrect syntax submissions for each time interval, calculating the ratios between correct, incorrect syntax, and incorrect semantic submissions. These ratios were computed for both individual problem-wise data and across all problems to discern trends across varying difficulty levels defined by the tested abstract data operation concepts for the SQL query. Undergraduate and graduate student submissions were not separately categorized due to following data privacy protocols. Null-hypothesis testing was then performed on the percentages of correct, incorrect syntax, and incorrect semantic submissions for each hour of the day and each question. A subset of our results are shown in Table I.

Additionally, we performed regression analyses with different types of regressions (linear, quadratic, and sinusoidal) to determine the relationship between submission times and correctness status, examining whether submission patterns exhibit any discernible temporal trends or cyclic behavior. Regression analyses were done for each problem and across all problems using 15-minute intervals. A small subset of our results are shown in Figures 4 and 5 to abide by page limits.

The dataset exclusively consists of homework assignment submissions and excludes exam and collaborative assignment submissions. This decision was intentional, as homework submissions have flexible deadlines, allowing for submissions at any hour, facilitating comprehensive time-series analysis throughout the day. Additionally, homework assignments are completed independently, in contrast to collaborative assignments, which may involve assistance from classmates.

IV. Results

In this section, we present our findings and insights in the order of our research questions. To answer our first research question: *Does a significant correlation exist between the ratio of correct to incorrect answers and the time of day?*, we showcase the null hypothesis testing results that find hours of the day with statistically significant percentages of correct, incorrect due to syntax, or incorrect due to semantic responses. Next, to answer our second research question: *Are these patterns consistent across both fundamental and advanced SQL problem concepts?*, we provide our analyses based on the null hypothesis testing results across problems that test on fundamental and/or more advanced SQL concepts. Furthermore, we also look at the trends for both the number and percentage of correct, incorrect due to syntax, and incorrect due to semantic submissions over 15-minute increments using various regression models. These regression analyses allow us to not only identify temporal trends in students' performance but also to quantify the relationship between submission times and correctness status; we then make recommendations to provide insights for instructional practices and curriculum design for students learning SQL.

A. Analysis of Correlation between Correctness Ratio and Time of Day

To explore whether a significant correlation exists between the ratio of correct and incorrect answers and the time of day, we first looked at the distribution of correct, syntax error, and semantic error submissions across all problems, shown in Figure 3.

Syntax errors predominated in most time bins, followed by semantic errors, with correct responses being the least frequent. This logically follows, since students must resolve any compilation errors with their query (syntax errors) before they can reach semantic (logical) errors; students also only require one correct submission to receive full credit on the

Hour of Day	p-value	Standard Score	Correct Incorrect Syntax Incorrect Semantic
	Across All Problems		
4	0.01665067	-2.394332	Correct
5	0.02527502	2.237173	Incorrect Semantic
6	0.00440685	2.847468	Incorrect Semantic
6	0.00469168	-2.827474	Incorrect Syntax
7	0.00219865	3.061999	Correct
	Question 1		
6	0.02640057	-2.220268	Incorrect Semantic
6	0.00035532	3.571225	Incorrect Syntax
7	0.00004276	4.092044	Correct
7	0.00047256	-3.495848	Incorrect Semantic
	Question 2		
3	0.03041199	2.164683	Incorrect Semantic
3	0.02458298	-2.247894	Incorrect Syntax
6	0.00004818	4.064262	Correct
6	0.00680566	-2.706207	Incorrect Semantic
6	0.02458298	-2.247894	Incorrect Syntax
7	0.00680566	-2.706207	Incorrect Semantic
7	0.00572570	2.763090	Incorrect Syntax
	Question 3		
6	0.00237588	3.038717	Correct
6	0.00464287	-2.830821	Incorrect Semantic
6	0.00957265	-2.590896	Incorrect Syntax
7	0.00237588	3.038717	Correct
7	0.00464287	-2.830821	Incorrect Semantic
7	0.00957265	-2.590896	Incorrect Syntax
	Question 4		
3	0.04426422	-2.011580	Incorrect Semantic
3	0.02994226	2.170853	Incorrect Syntax
5	0.00135192	3.204724	Incorrect Semantic
5	0.00279546	-2.989378	Incorrect Syntax
7	0.01498883	-2.432649	Correct
7	0.04293162	-2.024375	Incorrect Semantic
7	0.01409519	2.454828	Incorrect Syntax
23	0.04231964	2.030363	Correct
	Question 5		
3	0.02976962	-2.173142	Incorrect Syntax
5	0.00553655	2.774036	Incorrect Semantic
5	0.01388217	-2.460298	Incorrect Syntax
7	0.00362650	2.908945	Correct
	Question 6		
5	0.01185040	1.185040	Incorrect Semantic
6	0.00003033	0.003033	Correct
6	0.00017609	0.017609	Incorrect Syntax
7	0.03142574	3.142574	Incorrect Semantic
	Question 7		
5	0.00977399	2.583724	Incorrect Semantic
5	0.00076890	-3.363755	Incorrect Syntax
6	0.00007920	3.946804	Correct
6	0.02276271	-2.277394	Incorrect Semantic
6	0.04280909	-2.025568	Incorrect Syntax
	Question 8		
3	0.01650061	2.397650	Incorrect Semantic
3	0.03958269	-2.058078	Incorrect Syntax
5	0.00733876	-2.681068	Incorrect Semantic
5	0.00963917	2.588512	Incorrect Syntax
6	0.03136570	2.152402	Correct
7	0.01207560	2.509927	Correct
7	0.01349255	-2.470497	Incorrect Syntax
	Question 9		
6	0.01753187	2.375359	Incorrect Semantic
6	0.01359850	-2.467698	Incorrect Syntax
7	0.00172396	-3.134077	Correct
7	0.00060471	3.429492	Incorrect Semantic
7	0.00555279	-2.773083	Incorrect Syntax
8	0.01194479	2.513771	Correct
	Question 10		
6	0.01534092	2.424229	Incorrect Semantic
7	0.00002151	4.248639	Correct
7	0.00069672	-3.390867	Incorrect Syntax
8	0.03099922	-2.157083	Incorrect Semantic
8	0.02622567	2.222854	Incorrect Syntax
	Question 11		
5	0.04976885	-1.961945	Correct
5	0.01946679	2.336469	Incorrect Syntax
8	0.00689381	2.701931	Incorrect Semantic
8	0.04124576	-2.041051	Incorrect Syntax
9	0.04815377	1.976009	Correct
	Question 12		
4	0.03971778	-2.056672	Correct
6	0.00253409	3.019241	Incorrect Semantic
6	0.00083225	-3.341839	Incorrect Syntax
8	0.02278984	-2.276940	Incorrect Semantic
8	0.00948109	2.594201	Incorrect Syntax
	Question 13		
2	0.04907261	1.967960	Correct
4	0.01218141	-2.506845	Correct
4	0.00111895	3.258773	Incorrect Semantic
4	0.01244209	-2.499351	Incorrect Syntax
6	0.01930651	2.339559	Incorrect Syntax
	Question 14		
3	0.00834395	2.637825	Correct
5	0.04261746	-2.027440	Incorrect Semantic
5	0.03375353	2.123004	Incorrect Syntax
6	0.03989101	2.054876	Incorrect Syntax
7	0.04338313	2.020002	Correct
	Question 15		
5	0.00048186	3.490642	Incorrect Semantic
5	0.00237634	-3.038659	Incorrect Syntax
6	0.03475754	2.111172	Correct
6	0.01812656	2.363023	Incorrect Semantic
6	0.00375270	-2.898234	Incorrect Syntax
7	0.00052454	3.467904	Correct

TABLE I: Significant hours by problem and correctness status

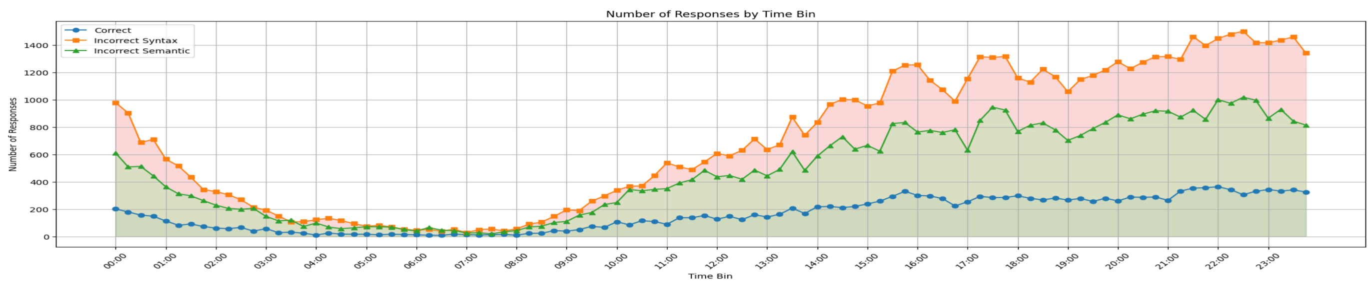

FIG. 3: This figure shows the total number of correct, syntax error, and semantic error submissions made by students across all SQL homework problems. Note that the time bins are defined every 15 minutes instead of every hour.

autograder. We observed that the highest volume of responses occurred later in the day, peaking between 9:30 p.m. and 11:30 p.m., with 10 p.m. recording the highest total responses at 10,979. Conversely, the early morning hours from 3:30 a.m. to 8:00 a.m. exhibited the lowest response rates, with 7 a.m. registering only 323 responses.

Although line graphs (similar to Figure 3) for individual problems are not depicted due to space constraints, these trends persisted uniformly across all problems, irrespective of the tested concepts. For instance, problem #12 recorded 957 submissions at 10 p.m. and 23 submissions at 7 a.m. As problem difficulty increased, no discernible effect was noted on the submission time distribution.

To address the temporal aspect of our first research question, we present the outcomes of null hypothesis testing on the percentages of correct, syntax error, and semantic error submissions across each hour of the day and for each question. Table I only includes the statistically significant hours, considering space constraints, while also providing the standard score to indicate the directionality of the correlation.

Across all problems, significant values were observed at 4 a.m., 5 a.m., 6 a.m., and 7 a.m. Specifically, at 4 a.m., the percentage of correct responses was notably lower than the average, while at 7 a.m., it was significantly higher. Furthermore, both 5 a.m. and 6 a.m. exhibited a higher percentage of semantic error submissions compared to the average, while at 6 a.m., the percentage of syntax error submissions was notably lower than average. This observation implies a potential correlation between cognitive performance and diurnal rhythm, indicating that students demonstrate increased proficiency in SQL query syntax and problem-solving skills during early morning hours, such as 6 a.m. to 7 a.m. Conversely, diminished problem-solving abilities are associated with late-night periods, particularly at 4 a.m., when students tend to struggle with SQL problem-solving tasks. In the next subsection, we look at the correlation between the correctness ratio and time of day across individual problems based on their SQL problem concepts.

B. Consistency of Patterns Across Fundamental and Advanced SQL Problem Concepts

To contextualize the SQL problem concepts, we will delineate the specific skills assessed by each question. As part of the homework assignment's framework, all SQL problems inherently involve selection and projection abstract data operations. For instance, Questions 1 and 2 entail basic queries with single-table selection and projection. Question 3 expands on this by incorporating a multiple table join alongside selection and projection. Meanwhile, Question 4 introduces a more intricate problem with multiple table joins and a basic subquery. Question 5 builds upon this complexity with a slightly more intricate subquery structure. Question 6 introduces basic grouping and aggregation within a multiple table join context. In contrast, Question 7 deviates from the norm by testing inserts and updates. Building on the skills required for Question 6, Question 8 introduces constraints within the HAVING clause. Question 9 advances further with a subquery involving aggregations and the HAVING clause; Questions 9 and beyond are typically the more complex ones that test various advanced SQL problem concepts. Similarly, Question 10 introduces set operations, necessitating two subqueries akin to the complexity level of Question 6. Questions 11 and 12 escalate the difficulty by incorporating multiple subqueries alongside aggregations, constraints, and multiple table joins. Question 13 elevates the complexity by introducing correlated subqueries instead of regular subqueries, building upon the foundation of Questions 11 and 12. Additionally, Question 14 assesses trigger usage, while Question 15 evaluates stored procedure usage, both representing non-traditional SQL queries.

Although no discernible effect was noted on the submission time distribution as the problem difficulty increased, discrepancies emerged in the prevalence of error types across different problems. We generally observed that as problem complexity heightened (question number increased), the proportion of syntax errors increased while the proportions of semantic errors and correct responses decreased.

Furthermore, the trends observed across all problems remained consistent until Question 8, along with the non-traditional SQL queries (Questions 7, 14, and 15). Specifically, Table I highlights statistically significant increases in correct submissions generally occurring at 6 a.m. and/or 7 a.m. However, from Question 9 through Question 13, we noticed shifts in these patterns. While significant negative correlations persisted between correct responses and 4 a.m. and 5 a.m., in alignment with our findings across all problems,

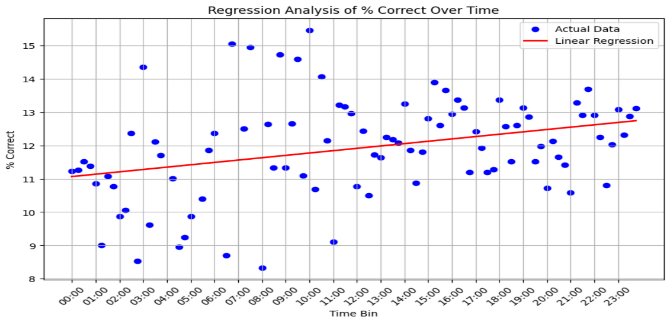

FIG. 4: This figure shows the linear regression of the percent of correct responses made by students across all problems aggregated together. Outliers (1.5 times the Interquartile Range [IQR] above the 3rd quartile or 1.5 times the IQR below the 1st quartile) were excluded from the analysis. Regression Equation for Percent of Correct Responses: $y = 11.07 + 0.02x$. Coefficient of Correlation: 0.331

Question 9 exhibited a significant negative correlation between correct responses and 7 a.m. The time frames wherein students excel transitioned from 7 a.m. for more basic SQL problems to 8 a.m. (Question 9) and 9 a.m. (Question 11) later in the morning for more advanced SQL problems. Interestingly, students exhibited positive correlations earlier in the late-night, as evidenced by significant trends at 11 p.m. for Question 4 and 2 a.m. for Question 13.

C. Regression Analysis of Submission Trends

When analyzing the relationship between the *percentage* of correct responses and the time of day, our findings suggest a moderate positive correlation, best fitted by a linear regression model (Figure 4). The correlation coefficient of 0.331 indicates that, on average, there is a modest increase in the percentage of correct responses from 12:00 a.m. to 11:45 p.m. However, since the strength of this relationship is only moderate, other factors beyond the time of day also likely influence student performance. Based on Figure 4, students tend to perform less effectively during the early hours of the morning and more effectively during the late hours of the evening.

We also looked at the absolute *number* of correct, syntax error, and semantic error submissions with the time of day using various types of regression models, including linear, quadratic, and sinusoidal. We found that regardless of the question number or submission outcome (correct, syntax error, or semantic error), the submissions best fit a sinusoidal regression. We discuss the implications of this trend in our next subsection. For the sake of space constraints, we have only included the sinusoidal regressions over all the problems and for question 12, as shown in Figures 5 and 6. Regardless of the submission outcome, we observed a strong fit on the sinusoidal regression, with coefficients of determination of 0.942, 0.953, and 0.950, respectively over all the problems. This pattern persists across different individual questions and problem complexities, exemplified by the sinusoidal regression model applied to problem #12, as shown in Figure 6. Despite variations in amplitude, the submissions consistently align with the sinusoidal regression, as evidenced by coefficients of determination of 0.766, 0.866, and 0.787 for correct, syntax error, and semantic error submissions, respectively. This observation extends to problem #4, indicating a fairly strong fit with coefficients of determination of 0.738, 0.827, and 0.819 for correct, syntax error, and semantic error submissions, respectively.

D. Discussion & Interpretations

In interpreting the findings from the preceding three subsections, we consider various factors related to cognitive performance, sleep patterns, and circadian rhythms.

In analyzing whether a significant correlation exists between the ratio of correct to incorrect answers and the time of day, the significant decrease in correct responses observed at 4 a.m. may be attributed to fatigue, reduced attention, or diminished cognitive functioning during late-night hours, potentially hindering problem-solving abilities. Conversely, the notable increase in correct responses at 7 a.m. could be associated with heightened alertness, improved cognitive functioning upon waking, or the fresh start of a new day, leading to enhanced motivation. In addition, the lower percentage of syntax error submissions at 6 a.m. suggests that students may demonstrate a better understanding of SQL syntax during this time, possibly due to increased cognitive functioning or focus.

In analyzing the consistency of patterns across fundamental and advanced SQL problem concepts, the observed increase in syntax errors and decline in correct responses with heightened problem complexity suggests students' difficulties in mastering advanced SQL concepts or combining multiple SQL concepts accurately. Factors such as inadequate understanding of foundational SQL concepts or challenges in translating complex problem requirements into multiple basic SQL queries (or subqueries) could contribute to this trend. Previous research has also highlighted syntax errors as significant obstacles in SQL learning [17, 15]. Shifts in submission patterns from basic to advanced SQL problems, especially in peak performance timing, suggest that students may adapt their study habits and time management strategies in response to varying levels of problem complexity. The observed early late-night productivity peaks (for example Questions 4 and 13) may reflect heightened focus due to reduced distractions or increased motivation as assignment deadlines approach. In particular, students' productivity seems to shift towards slightly later hours, such as 8 a.m. or 9 a.m. (Questions 9 and 11) for more advanced problems, from the original 7 a.m. (Questions 1-8, 14, and 15) for less advanced problems.

In our regression analysis of submission trends, the observed moderate positive correlation between the percentage of correct responses and time of day (Figure 4) may be caused by college students going to bed relatively late, around 2 a.m. on average, and waking up around 9 a.m. [6]. Therefore, as the day progresses, students may experience increasing alertness and cognitive functioning, potentially contributing

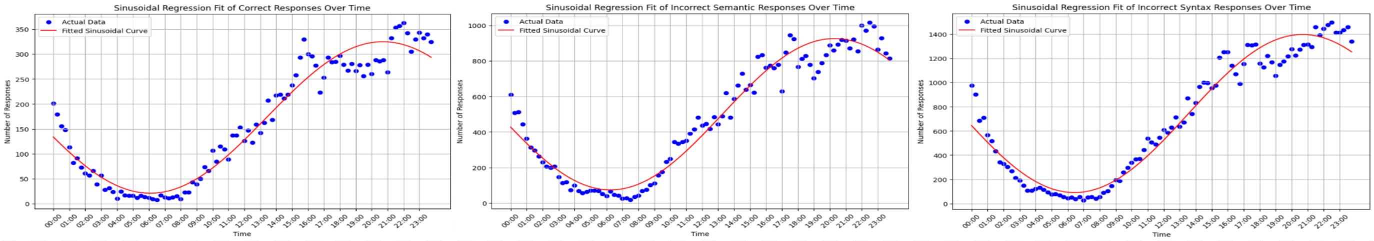

FIG. 5: This figure shows the sinusoidal regressions over the number of correct, incorrect syntax, and incorrect semantic responses made by students across all the problems aggregated together. Regression Equation for Correct Responses: $y = 151.972sin(0.054(x - 3.402)) + 172.951$. Regression Equation for Incorrect Syntax Responses: $y = 653.840sin(0.055(x - 3.294)) + 744.320$. Regression Equation for Incorrect Semantic Responses: $y = 426.419sin(0.056(x - 3.311)) + 500.008$. Coefficient of Determination for Correct Responses: 0.942. Coefficient of Determination for Incorrect Syntax Responses: 0.953. Coefficient of Determination for Incorrect Semantic Responses: 0.950

to improved problem-solving abilities and higher percentages of correct responses in their submissions. While there's a modest increase in correct responses from early morning to late evening, the moderate strength of this correlation implies that factors beyond time likely play significant roles in student performance.

In addition, the high coefficients of determination for the sinusoidal regression models applied to the absolute number of correct, syntax error, and semantic error submissions suggest that students exhibit cyclical behavior, alternating between periods of high engagement and low engagement, in line with the times mentioned by Okano et al. [6] regarding college student bedtimes. Notably, the crest and trough of the sinusoidal curves consistently occur at similar times, suggesting a consistent circadian rhythm among students learning SQL, independent of the concept complexity they encounter.

E. Recommendations for Instructors

Based on our findings, we recommend instructors schedule course material releases during periods when students exhibit higher-than-average performance, such as after 7 a.m., and close submissions during periods of lower performance, like between 4 a.m. and 7 a.m. This strategy can help optimize student engagement and minimize errors, particularly semantic errors, during the relaxed submission deadline of two weeks.

Furthermore, we noticed a positive correlation between the time of day and the percentage of correct responses, with students exhibiting peak activity during later hours of the afternoon and evening. To capitalize on this trend, we recommend instructors try to avoid scheduling lectures early in the morning and instead, opt for afternoon time slots. This recommendation aligns with prior studies, which found that students have improved academic performance when attending classes or exams in the early afternoon compared to the early morning [5, 29].

Although our analysis revealed several significant periods during which student performance deviated notably from the average, there were numerous instances where conclusive evidence was lacking. This variability could stem from individual variations in peak productivity periods experienced by each student throughout the day. Thus, we propose that instructors implement relaxed (and therefore flexible) assignment deadlines to accommodate the diverse study schedules of students. Allowing students to select submission times within a specified window encourages them to align their work with their peak cognitive functioning periods, thereby enhancing individual performance. Students of the database course where we collected data were able to submit at any time within two weeks, as the assignment was released on an online learning management system - PrairieLearn[7].

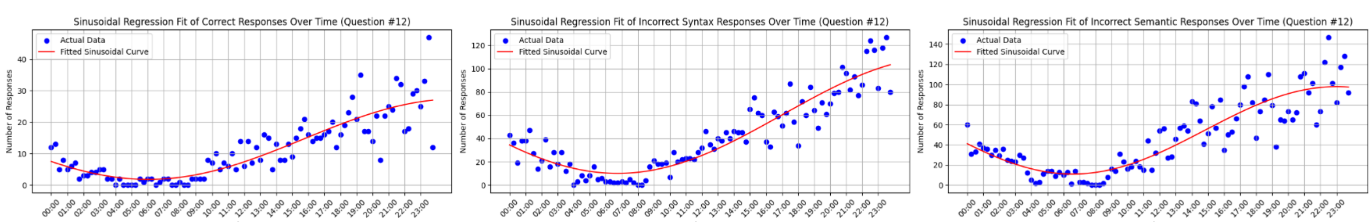

FIG. 6: This figure shows the sinusoidal regressions over the number of correct, incorrect syntax, and incorrect semantic responses made by students answering question 12. Regression Equation for Correct Responses: $y = 12.807sin(0.041(x - 3.731)) + 14.611$. Regression Equation for Incorrect Syntax Responses: $y = 50.169sin(0.038(x - 3.679)) + 60.174$. Regression Equation for Incorrect Semantic Responses: $y = 43.446sin(0.048(x - 3.453)) + 54.396$. Coefficient of Determination for Correct Responses: 0.766. Coefficient of Determination for Incorrect Syntax Responses: 0.866. Coefficient of Determination for Incorrect Semantic Responses: 0.787

V. Limitations and Future Work

Since our study is based on data collected from the University of Illinois Urbana-Champaign with a top-ranked Computer Science department, the students and their data may impact the applicability of our findings to broader contexts. To enhance the generalizability of our results, data from other universities or institutions should be collected and analyzed. Furthermore, collecting and analyzing data from multiple semesters would allow for an analysis of the consistency of the observed trends over time.

Another limitation stems from our approach to computing the linear correlation coefficient for the percentage of correct responses with the time of day. In this analysis, we excluded outliers exceeding 1.5 times the interquartile range (IQR) above the third quartile and below the first quartile. However, removing outliers could mean disregarding student responses during these periods. While these outlier submissions may not fully represent the population, they could contain relevant information capable of yielding insights into the data.

The identified outliers likely stem from limited sample sizes during specific time intervals, notably between 3 a.m. and 8 a.m., as shown in Figure 3. These patterns are inherent to students' circadian rhythms, aligning with wakefulness and productivity during daytime hours and rest during nighttime. As a result, data collected during periods of reduced student activity may be disproportionately influenced by a few individuals with unique sleep schedules or situations, deviating from the broader student population. To address this limitation and accommodate individual sleep-wake patterns in future studies, students may be grouped based on whether they had an early or late chronotype, adopting a similar approach by Facer-Childs et al. [31]. However, logistical constraints such as the large student population in our research made it impractical to record individual chronotypes for all students.

Future research should also address the underlying factors contributing to the variations in student performance across different times of the day. A qualitative study could shed light on reasons why students exhibit enhanced productivity during specific time intervals when learning SQL. By integrating insights from both quantitative and qualitative research work, we can provide instructors with more context behind our recommendations for optimizing teaching strategies and course content delivery methods.

VI. Conclusion

In this research work, we presented our findings from 129,408 student submissions on an SQL homework assignment, collected from the Database Systems course at the University of Illinois Urbana-Champaign. For our research questions - 1) *Does a significant correlation exist between the ratio of correct to incorrect answers and the time of day?* and 2) *Are these patterns consistent across both fundamental and advanced SQL problem concepts?* - we found a significant decrease in correct responses at 4 a.m. and a peak in correct responses at 7 a.m., aligning with student's diurnal rhythms. In addition, variations in syntax error rates and submission patterns across problem complexities suggest that students face challenges in mastering advanced SQL concepts. Students demonstrate adaptability in study habits, with their productivity shifting towards later hours for more complex problems. Regression analysis supports the influence of circadian rhythms, with student engagement showing cyclical patterns aligned with typical college student sleep schedules. These insights highlight the importance of considering time-related factors in instructional design and support the need for further research to explain underlying mechanisms driving student performance fluctuations. Based on our findings, we suggested instructors strategically time course material releases and submissions to coincide with periods of higher student performance, such as post-7 a.m. for releases, and closing assignments between 4-7 a.m. when student performance drops. Additionally, in line with prior research, scheduling classes during later afternoon hours may help increase student engagement.

References

[1] A. DiFranza, "5 reasons sql is the need-to-know skill for data analysts," 2020.

[2] A. Mitrovic, "Learning sql with a computerized tutor," in *Proceedings of the Twenty-Ninth SIGCSE Technical Symposium on Computer Science Education*, ser. SIGCSE '98. New York, NY, USA: ACM, 1998, p. 307–311.

[3] J. Vigo, "5 programming languages database administrators should learn," November 2020.

[4] M. D. Evans, P. Kelley, and J. Kelley, "Identifying the best times for cognitive functioning using new methods: matching university times to undergraduate chronotypes," *Frontiers in human neuroscience*, vol. 11, p. 239492, 2017.

[5] P. T. Luong, "Time-of-day effect on academic achievement of adolescents," 2017.

[6] K. Okano, J. R. Kaczmarzyk, N. Dave, J. D. Gabrieli, and J. C. Grossman, "Sleep quality, duration, and consistency are associated with better academic performance in college students," *NPJ science of learning*, vol. 4, no. 1, p. 16, 2019.

[7] M. West, G. L. Herman, and C. B. Zilles, "Prairielearn: Mastery-based online problem solving with adaptive scoring and recommendations driven by machine learning," 2015.

[8] L. Mannila, M. Peltomäki, and T. Salakoski, "What about a simple language? analyzing the difficulties in learning to program," *Computer science education*, vol. 16, no. 3, pp. 211–227, 2006.

[9] E. Lahtinen, K. Ala-Mutka, and H.-M. Järvinen, "A study of the difficulties of novice programmers," *Acm sigcse bulletin*, vol. 37, no. 3, pp. 14–18, 2005.

[10] A. E. Fleury, "Programming in java: Student-constructed rules," in *Proceedings of the thirty-first SIGCSE technical symposium on Computer science education*, 2000, pp. 197–201.

[11] H.-C. Woon and Y.-T. Bau, "Difficulties in learning c++ and gui programming with qt platform: View of students," in *Proceedings of the 2017 International Conference on E-commerce, E-Business and E-Government*. New York, NY, USA: ACM, 2017, pp. 15–19.

[12] J. Bergin, "Java as a better c++," *ACM SIGPLAN Notices*, vol. 31, no. 11, pp. 21–27, 1996.

[13] N. Alzahrani, F. Vahid, A. Edgcomb, K. Nguyen, and R. Lysecky, "Python versus c++ an analysis of student struggle on small coding exercises in introductory programming courses," in *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*. New York, NY, USA: ACM, 2018, pp. 86–91.

[14] A. Ahadi, V. Behbood, A. Vihavainen, J. Prior, and R. Lister, "Students' syntactic mistakes in writing seven different types of sql queries and its application to predicting students' success," in *Proceedings of the 47th ACM Technical Symposium on Computing Science Education*. New York, NY, USA: ACM, 2016, pp. 401–406.

[15] S. Poulsen, L. Butler, A. Alawini, and G. L. Herman, "Insights from student solutions to sql homework problems," in *Proceedings of the 2020 ACM Conference on Innovation and Technology in Computer Science Education*. New York, NY, USA: ACM, 2020, pp. 404–410.

[16] T. Taipalus, M. Siponen, and T. Vartiainen, "Errors and complications in sql query formulation," *ACM Transactions on Computing Education (TOCE)*, vol. 18, no. 3, pp. 1–29, 2018.

[17] T. Taipalus and P. Perälä, "What to expect and what to focus on in sql query teaching," in *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*, ser. SIGCSE '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 198–203. [Online]. Available: https://doi.org/10.1145/3287324.3287359

[18] D. Miedema, E. Aivaloglou, and G. Fletcher, "Identifying sql misconceptions of novices: Findings from a think-aloud study," in *Proceedings of the 17th ACM Conference on International Computing Education Research*, ser. ICER 2021. New York, NY, USA: Association for Computing Machinery, 2021, p. 355–367. [Online]. Available: https://doi.org/10.1145/3446871.3469759

[19] A. Ahadi, J. Prior, V. Behbood, and R. Lister, "A quantitative study of the relative difficulty for novices of writing seven different types of sql queries," in *Proceedings of the 2015 ACM Conference on Innovation and Technology in Computer Science Education*, ser. ITiCSE '15. New York, NY, USA: ACM, 2015, p. 201–206.

[20] A. Migler and A. Dekhtyar, "Mapping the sql learning process in introductory database courses," in *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, ser. SIGCSE '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 619–625. [Online]. Available: https://doi.org/10.1145/3328778.3366869

[21] R. Alkhabaz, Z. Li, S. Yang, and A. Alawini, "Student's learning challenges with relational, document, and graph query languages," in *Proceedings of the 2nd International Workshop on Data Systems Education: Bridging Education Practice with Education Research*, ser. DataEd '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 30–36. [Online]. Available: https://doi.org/10.1145/3596673.3596976

[22] S. Yang, Z. Wei, G. L. Herman, and A. Alawini, "Analyzing patterns in student sql solutions via levenshtein edit distance," in *Proceedings of the Eighth ACM Conference on Learning @ Scale*, ser. L@S '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 323–326. [Online]. Available: https://doi.org/10.1145/3430895.3460979

[23] S. Yang, G. L. Herman, and A. Alawini, "Analyzing student sql solutions via hierarchical clustering and sequence alignment scores," in *1st International Workshop on Data Systems Education*, ser. DataEd '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 10–15. [Online]. Available: https://doi.org/10.1145/3531072.3535319

[24] ——, "Mining sql problem solving patterns using advanced sequence processing algorithms," in *Proceedings of the 2nd International Workshop on Data Systems Education: Bridging Education Practice with Education Research*, ser. DataEd '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 37–43. [Online]. Available: https://doi.org/10.1145/3596673.3596973

[25] ——, "Comparison of student learning outcomes among sql problem-solving patterns," in *2023 IEEE Frontiers in Education Conference (FIE)*, 2023, pp. 1–9.

[26] J. Danaparamita and W. Gatterbauer, "Queryviz: Helping users understand sql queries and their patterns," in *Proceedings of the 14th International Conference on Extending Database Technology*, ser. EDBT/ICDT '11. New York, NY, USA: Association for Computing Machinery, 2011, p. 558–561. [Online]. Available: https://doi.org/10.1145/1951365.1951440

[27] S. E. Carrell, T. Maghakian, and J. E. West, "A's from zzzz's? the causal effect of school start time on the academic achievement of adolescents," *American Economic Journal: Economic Policy*, vol. 3, no. 3, pp. 62–81, 2011.

[28] P. Hinrichs, "When the bell tolls: The effects of school starting times on academic achievement," *Education Finance and Policy*, vol. 6, no. 4, pp. 486–507, 2011.

[29] A. Gaggero and D. Tommasi, "Time of day, cognitive tasks and efficiency gains," 2020.

[30] N. G. Pope, "How the time of day affects productivity: Evidence from school schedules," *Review of Economics and Statistics*, vol. 98, no. 1, pp. 1–11, 2016.

[31] E. R. Facer-Childs, S. Boiling, and G. M. Balanos, "The effects of time of day and chronotype on cognitive and physical performance in healthy volunteers," *Sports medicine-open*, vol. 4, pp. 1–12, 2018.