KoekoEdu Feature Impact Analysis and Predictive Modelling

Dear [Your Name],

I hope this message finds you well. We have an exciting and challenging project ahead of us. As you know, our primary goal is to evaluate the impact of the new features we introduced in the latter half of 2021.

These features were designed with the dual aim of increasing user engagement and enhancing profitability by the second quarter of 2022. We have high hopes that these updates have made a significant difference, and it's now up to us to put these assumptions to the test through rigorous data analysis and statistical evaluation.

Here's a detailed breakdown of the tasks and expectations for this project:

DATA EXTRACTION

1. **SECTION ONE**

Start by utilizing the **student_purchases** table from the database to create a result set with the following columns:

- purchase_id
- student_id
- plan_id
- date_start
- date_end
- date_refunded

In this section, the **date_purchased** column is renamed **date_start** and adjusted for consistency with the subsequent **date_end** column.

To calculate the end date of a subscription (**date_end**), add one month, three months, or 12 months to the start date of a subscription for a Monthly (represented as **0** in the **plan_id** column), Quarterly (**1**), or an Annual (**2**) purchase, respectively. The only exception is the lifetime subscription (denoted by **3**), which has no end date (i.e. NULL).

2. **SECTION TWO**

Using the query from section one as a sub-query, create a new one retrieving the following columns:

- purchase_id
- student_id
- plan_id
- date_start
- date_end

Re-calculate the **date_end** column so that if an order was refunded—indicated by a non-**NULL** value in the **date_refunded** field—the student's subscription terminates at the refund date.

3. **SECTION THREE**

Using the query you designed in section two, create a new SQL query that, when executed, stores in the database, a view called purchases_info which we'll use in subsequent parts. The view should include the following columns:

- purchase_id
- student_id
- plan_id
- date_start
- date_end
- paid_q2_2021
- paid_q2_2022

The **paid_q2_2021** and **paid_q2_2022** columns contain binary values indicating whether a student had an active subscription during the respective year's second quarter (April 1 to June 30, inclusive). A **0** in the column indicates a free-plan student in Q2, while a **1** represents an active subscription in that period.

NB: if you did everything right, you should have a view with 18207 rows and the first purchase record doesn't fall under any of the second quarters.

4. SECTION FOUR

We're now interested in analysing the engagement of our users in terms of the total minutes watched during Q2 2021 and Q2 2022 separately. Additionally, we want to identify which users were paid subscribers during each of these periods.

Your task is to write an SQL query that returns the following columns:

- **student_id** – a list of student IDs
- **minutes_watched** – the sum of minutes students have watched in both periods—return a separate table for each period

The information about the minutes watched by each student is available in the **student_video_watched** table. Remember you're doing this for Q2 2021 & Q2 2022 separately.

NB: If you did everything right, the result for Q2 2021 query will have 7639 rows and that of Q2 2022 will have 8841 rows.

5. SECTION FIVE

Now, use the queries you designed in section four to create a result set with the following columns:

- student_id
- minutes_watched
- paid_in_q2

The last column indicates whether a student had an active subscription in Q2 (represented by 1) or not (represented by 0).

Retrieve the following four datasets and store them in the corresponding CSV files:

- Students engaged in Q2 2021 who haven't had a paid subscription in Q2 2021

- Students engaged in Q2 2022 who haven't had a paid subscription in Q2 2022
- Students engaged in Q2 2021 who have been paid subscribers in Q2 2021
- Students engaged in Q2 2022 who have been paid subscribers in Q2 2022

NB: If you did everything right, you'd retrieve the following number of rows for each case, respectively: 5334, 6055, 2305, 2786

6. SECTION SIX

We are now going to consider only the students who've been issued a certificate. Create an SQL query to extract the following information for each such student:

- The student ID
- The total minutes watched
- The total number of certificates issued

Assign the corresponding value for students with no minutes recorded as 0. Save the resulting table as minutes_and_certificates.csv for later use.

NB: If you did everything right, your csv file should have 658 rows.

DATA PREPROCESSING

1. SECTION ONE

Using a kdeplot, plot the distribution of the minutes_watched variable of each of the four datasets (from section five of data extraction) and examine its shape. Are the distributions skewed? If yes, how? What does this tell us about the distribution of minutes watched?

2. SECTION TWO

Remove outliers by retaining only values below the 99th percentile for each dataset. Once you've retrieved the final datasets, save the cleaned dataset as four separate CSV files on your computer:

- minutes_watched_2021_paid_0_no_outliers.csv
- minutes_watched_2022_paid_0_no_outliers.csv
- minutes_watched_2021_paid_1_no_outliers.csv
- minutes_watched_2022_paid_1_no_outliers.csv

DESCRIPTIVE STATISTICS

Calculate the mean and median minutes watched by students in the four groups (2021 free-plan students, 2022 free-plan students, 2021 paying students, 2022 paying students). How does the median compare with the mean in each group?

Referring to the distribution plots you created in the first section of data preprocessing, does this result meet your expectations?

CONFIDENCE INTERVALS

For each of the four groups, find the minute interval for which you are 95% confident a random person will fall in that interval. Assume a normal distribution.

What conclusions can you draw about students' engagement in Q2 2021 and Q2 2022 for both free-plan and paying students?

Create a confidence interval bar chart to support your arguments better. Bars are the sample mean and error bars should be the lower and upper bounds of their confidence intervals.

HYPOTHESIS TESTING

You want to reach a data-driven decision on whether the new features (courses, career tracks, and exams) contribute to the increased number of minutes watched on the platform for free-plan and paying students—i.e., increased student engagement in their study process. You use hypothesis testing on both groups (free-plan and paying) for 2021 and 2022.

Let your null and alternative hypotheses (respectively) be:

- The engagement (minutes watched) in Q2 2021 is higher than or equal to the one in Q2 2022. We test free-plan and paying students separately.
- The engagement (minutes watched) in Q2 2021 is lower than the one in Q2 2022. We test free-plan and paying students separately.

Additionally, make the following assumptions:

- Assume a normal distribution.
- For free-plan students, perform a two-sample t-test assuming equal variances.
- For paying students, perform a two-sample t-test assuming unequal variances.

Optional: Perform a two-sample f-test for variances to support the assumptions.

What conclusion can you draw from this test? Comment on the results of committing a Type I or a Type II error in this study. Which one would result in higher costs to the company?

CORRELATION ANALYSIS

Remember the data we extracted in section six of data extraction, find the correlation coefficient between the minutes watched and the certificates issued. Interpret the results. Create a scatter plot to support your arguments better.

DEPENDENCIES AND PROBABILITIES

We are now going to analyse the engaged students on the platform. Return to the database in My SQL and consider all students who've watched a lecture in Q2 2021 and those who've watched a lecture in Q2 2022 as two sets. Let the universal set be all students who've watched a lecture on the platform—the union of the two sets defined above. Don't omit any outliers we've removed during this project.

- Determine if watching a lecture in Q2 2021 and Q2 2022 are dependent or independent events. Explain your result.
- What is the probability that a student has watched a lecture in Q2 2021, given that they've watched a lecture in Q2 2022?

MACHINE LEARNING MODELING AND DEPLOYMENT

Perform a linear regression using the minutes_watched column as a predictor and certificates_issued as a target. Having done that, answer the following:

- What is the linear equation that explains the behaviour of the relationship?
- What is the R-squared value of the regression? How would you interpret it?
- What is the predicted number of certificates taken by a student who has watched 1200 minutes of content? (Round your result up to the nearest integer.)

Note: Use 20% of your data as a test set. Use the number 365 as a random state.

After training the model with 80% of the data and evaluating with 20%, deploy it as a web application. Ensure the model is production-ready and can provide predictions effectively.

Please ensure that all tasks are carried out with attention to detail and accuracy. Your insights and findings will be crucial in shaping the future direction of our platform. If you have any questions or need additional clarifications, don't hesitate to reach out.

Looking forward to your valuable contributions!

Best regards,

John Doe,
Team Lead, Analytics
KoekoEdu