Data Preparation with SQL

`Data_scientist_project.sql` - data file

`View.sql`

To prepare the data for analysis, I first standardized the subscription dates by renaming the `date_purchased` column to `date_start`, ensuring consistency with the existing `date_end` column. I then recalculated the `date_end` values to accurately reflect subscription duration, adjusting them for any refunds: if a refund was issued (indicated by a non-NULL value in the `date_refunded` column), the subscription was considered to have ended on the refund date. Next, I created a SQL view named `purchases_info` within the `data_scientist_project` schema. This view includes two binary columns, `paid_q2_2021` and `paid_q2_2022`, which indicate whether a student held an active subscription during the second quarter (April 1 to June 30) of 2021 or 2022. A value of 1 denotes an active paid subscription, while 0 indicates a free-plan student for that period. This was accomplished using subqueries to check for subscription activity within the specified date ranges.

`Periods.sql`

To further analyze engagement, I filtered the data to identify students who were active in Q2 2021 and Q2 2022 but did not hold a paid subscription during those periods. Using the `paid_in_q2` column—calculated separately for each year (2021 and 2022)—I isolated students with `paid_in_q2 = 0`, representing free-plan users. The resulting datasets were exported as `minutes_watched_2021_paid_0.csv`, `minutes_watched_2022_paid_0.csv`,`minutes_watched_2021_paid_1.csv` and `minutes_watched_2022_paid_1.csv`, which contain students who watched content during the respective quarters without having an active paid subscription and the ones that did have an active one. This allowed for a targeted analysis of engagement behavior among non-paying users across both time frames.

`Certificates.sql:`

To explore the relationship between engagement and achievement, I focused on students who had been issued at least one certificate. Using SQL, I created a query to extract each student’s `student_id`, the total `minutes_watched`, and the total number of certificates they received. For students with no recorded viewing time, a value of 0 was assigned to ensure completeness of the dataset. This was accomplished through table joins and subqueries. The resulting dataset was saved as `minutes_and_certificates.csv` for further analysis.

Data Preprocessing with Python

`Outliers.ipynb`

To better understand user engagement patterns, I visualized the distribution of the `minutes_watched` variable across four datasets, each representing students with or without a paid subscription in Q2 2021 and Q2 2022. The plots revealed skewed distributions with a small number of students watching significantly more content than the rest. To reduce the impact of extreme values and focus on typical behavior, I removed outliers by filtering out data points above the 99th percentile in each dataset. The cleaned datasets were then saved as four separate CSV files:
`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`.

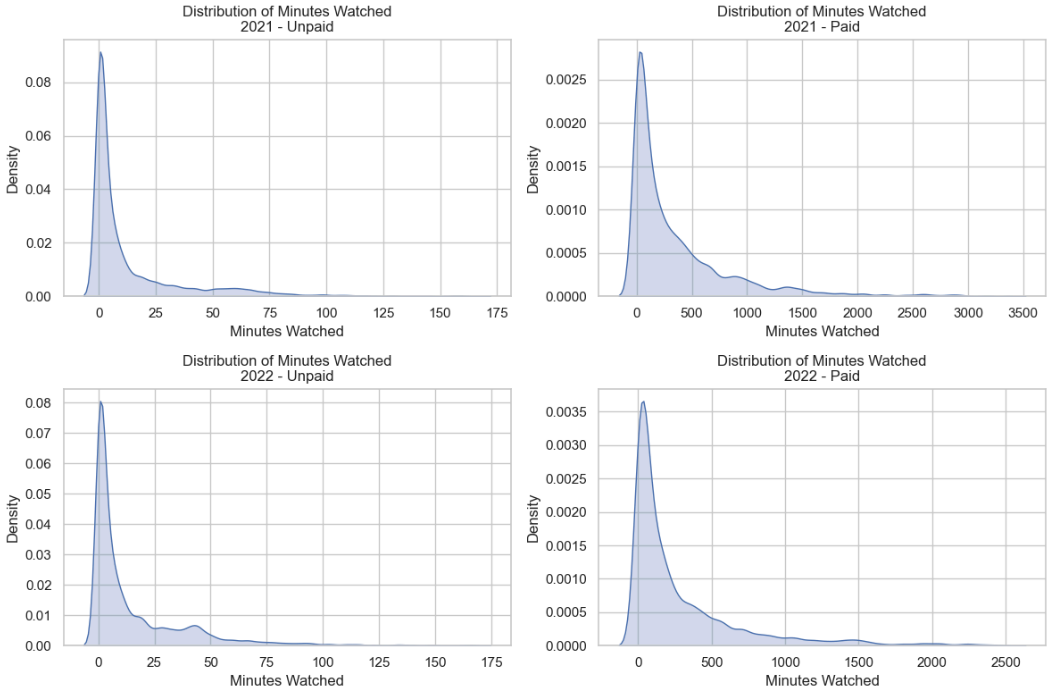

Figure 1. Distributions of minutes watched for paid and unpaid users in Q2 of 2021 and 2022

Data Analysis with Excel

`Hypothesis_Testing.xlsx`

I calculated the **mean and median minutes watched** for each of the four student groups—paid and free-plan users in Q2 2021 and Q2 2022. Next, I computed **95% confidence intervals** for the minutes watched in each group, assuming a normal distribution, to estimate the range within which a randomly selected student's engagement is likely to fall.

I conducted **hypothesis testing** separately for free-plan and paying students. The null hypothesis stated that engagement in Q2 2021 was greater than or equal to that in Q2 2022 ($\mu_1 \geq \mu_2$), while the alternative hypothesis posited that engagement in Q2 2021 was lower ($\mu_1 < \mu_2$).

For **free-plan users**, a **two-sample t-test assuming equal variances** yielded a t-statistic of –3.95, which is less than the critical value of –1.645. As a result, we **rejected** the null hypothesis and concluded that engagement among free-plan students significantly increased in Q2 2022 ($\alpha = 0.05$).

For **paying users**, a **two-sample t-test assuming unequal variances** resulted in a t-statistic of 5.15, which does not meet the rejection criterion. Thus, **we failed to reject** the null hypothesis, suggesting no statistically significant increase in engagement. In fact, this result indicates that engagement in Q2 2021 may have been equal to or higher than in Q2 2022 for paid users. To support the variance assumptions used in the t-tests, I also performed **F-tests for equality of variances** on both groups.

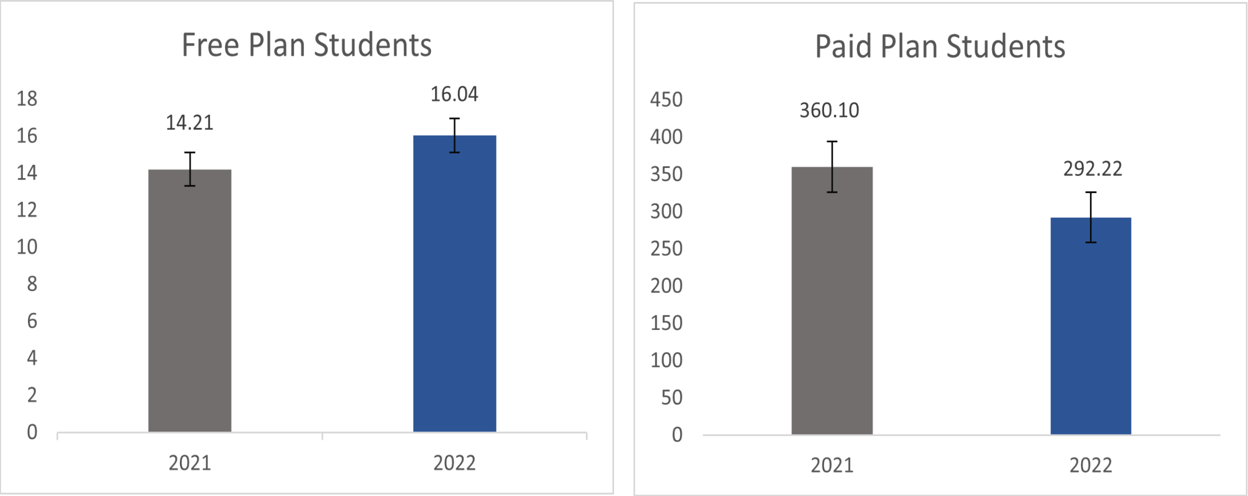

Figure 2. Bar charts showing the mean minutes watched by free-plan students (left) and paid-plan students (right) in Q2 of 2021 and 2022

`Correlation_Coefficients.xlsx`

To examine the relationship between student engagement and achievement, I calculated the **correlation coefficient** between the total **minutes watched** and the **number of certificates issued**. The result was **0.512**, indicating a **moderate positive linear relationship**. As we can see from the scatter plot below, there is a moderate positive linear relationship between minutes watched and the number of certificates issued. This suggests that students who watch more content are generally more likely to receive certificates, and those who watch less tend to receive fewer. The scatter plot further supports this trend, although the data points show some variability, indicating that while engagement is associated with achievement, the relationship is not perfectly linear. Other factors, such as course difficulty or certificate criteria, may also influence the outcome.

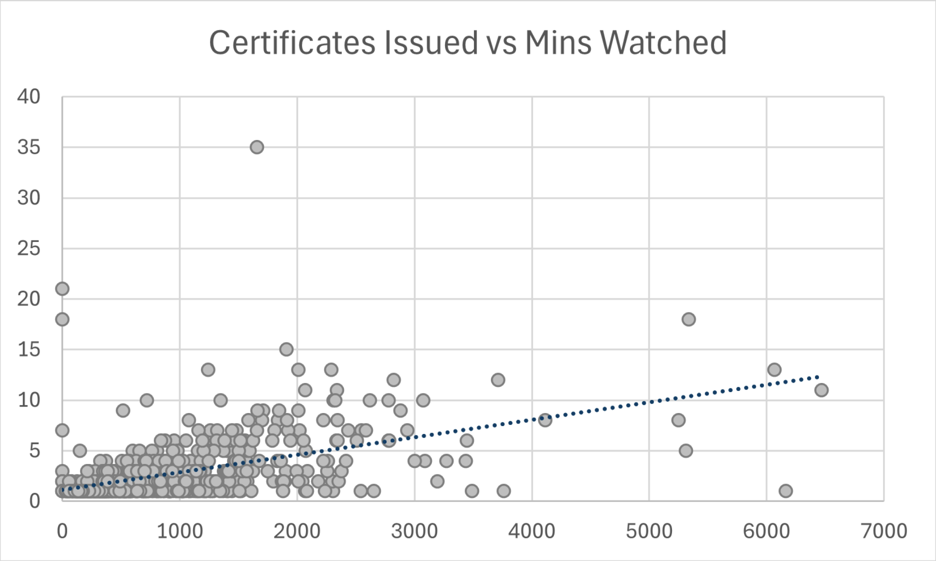

Figure 3. **X-axis:** Number of certificates issued to the student. **Y-axis:** Minutes watched by the student.

Data Prediction with Python

`LinearRegression.ipynb`

This section performs a **linear regression** using the `minutes_watched` column as the input variable (feature) and `certificates_issued` as the target variable. The goal is to explore the relationship between content consumption and certificate issuance.

Linear Equation

The fitted regression line is:

$$\hat{y} = 1.5074x + 2.5798$$

where:

- $\hat{y}$ - is the predicted number of certificates issued

- x - is the number of minutes watched

R-squared Value

$$R^2 = 0.0968$$

Only about 9.7% of the variation in certificates issued is explained by the variation in minutes watched. This suggests a **weak linear relationship** between the two variables.

- **Mean Squared Error (MSE)** = 13.6566

- **Root Mean Squared Error (RMSE)** = 3.6955

The average error between the predicted and actual values. RMSE is about equal to ~3.7 which means the model is, on average, off by about 3–4 certificates.

Prediction for 1200 Minutes Watched

Using the equation:

$$\hat{y} = 1.5074 \times 1200 + 2.5798 = 1811.4598 \approx 1812$$