SEG4300 - Assignment 2 Report - Anthony Nasr - 300170809

Dataset overview:
The dataset analyzed is a Kaggle churn dataset, sourced from Hugging Face: https://huggingface.co/datasets/kodylow/kaggle_churn. I chose this dataset due to my previous experience attempting to build a churn prediction model for a local business, where I faced challenges due to insufficient quality data. This dataset, containing approximately 346,000 rows, provides a more substantial basis for exploration. Each row represents a user with various features detailing their engagement on the Kaggle platform.

Sampling:
For the sake of efficiency, the dataset was loaded and streamed in. I decided to randomly sample 10,000 total rows using a simple shuffle function with a random seed of 42.

EDA summary:
The dataset primarily consists of integer and float features, with categorical variables (e.g., gender, city, registration method) mapped to numerical values. A heatmap of null values confirmed no missing data. However, class imbalance is evident, as non-churned users significantly outnumber churned users.

Feature distributions appear varied, except for the bd (age) column, which is heavily skewed toward 0, likely due to optional age entry at sign-up. A correlation heatmap revealed logical relationships. For example, higher payments and longer plans correlate positively with retention, while auto-renewal strongly negatively correlates with churn, indicating that users enabling auto-renew are less likely to churn.

GX:
A new GX suite was built, connecting a Pandas batch for validation. Several expectations were defined, primarily binary checks to ensure features such as is_churn, gender, is_auto_renew, and is_cancel adhere to expected integer values. Most validations passed, except for bd (age), where a significant number of users had invalid ages (<13 or >100), mostly due to missing data being stored as 0.

Insights:
Because the age column is skewed, it would be interesting to treat it via imputing values as a median or it as a feature all together in order not to affect the churn prediction. However, the data is mostly clean, having no other notable issues and virtually no null values. It also includes some great features such as daily playtime which can help the model capture user usage patterns in its churn prediction.