Project Coversheet

Full Name	Alvin Siphosenkosi Moyo
Project Title (Example – Week1, Week2, Week3, Week 4)	Week 3: Churn Prediction for StreamWorks Media

Instructions:

Students must download this cover sheet**,** use it as the first page of their project, and then save the entire document as a PDF before submission**.**

Project Guidelines and Rules

1. Formatting and Submission

- Format: Use a readable font (e.g., Arial/Times New Roman), size 12, 1.5 line spacing.
- Title: Include Week and Title (Example - Week 1: Travel Ease Case Study.)
- File Format: Submit as PDF or Word file
- Page Limit: 4–5 pages, including the title and references.

2. Answer Requirements

- Word Count: Each answer should be within 100–150 words; Maximum 800–1,200 words.
- Clarity: Write concise, structured answers with key points.
- Tone: Use formal, professional language.

3. Content Rules

- Answer all questions thoroughly, referencing case study concepts.

- Use examples where possible (e.g., risk assessment techniques).
- Break complex answers into bullet points or lists.

4. Plagiarism Policy

- Submit original work; no copy-pasting.
- Cite external material in a consistent format (e.g., APA, MLA).

5. Evaluation Criteria

- Understanding: Clear grasp of business analysis principles.
- Application: Effective use of concepts like cost-benefit analysis and Agile/Waterfall.
- Clarity: Logical, well-structured responses.
- Creativity: Innovative problem-solving and examples.
- Completeness: Answer all questions within the word limit.

6. Deadlines and Late Submissions

- Deadline: Submit on time; trainees who fail to submit the project will miss the “Certificate of Excellence”

7. Additional Resources

- Refer to lecture notes and recommended readings.
- Contact the instructor or peers for clarifications before the deadline.

YOU CAN START YOUR PROJECT FROM HERE

Executive Summary & Technical Methodology

StreamWorks Media faces rising customer acquisition costs, making **subscriber retention a critical priority**. This project analyses the streamworks_user_data.csv dataset to identify churn drivers and evaluate predictive approaches for early risk detection. A baseline **Logistic Regression** model is used to establish an interpretable reference for churn prediction, while **additional models, including Random Forest and tuned logistic variants, are explored as optional extensions** to assess whether sensitivity to churned users can be improved. The analysis combines **feature engineering, model evaluation, and segmentation** to translate predictive outputs into actionable, cost-effective retention strategies.

1. Data Scope & Exploration

Dataset Overview: The analysis utilises the streamworks_user_data.csv dataset, comprising **1,500 unique subscriber records** and **14 features**. The data structure includes a mix of numerical variables (e.g., age, monthly_fee, average_watch_hours) and categorical attributes (e.g., gender, subscription_type).

Key Observations:

- **Data Integrity:** Initial data profiling confirmed that date columns (signup_date, last_active_date) were incorrectly stored as object strings, requiring conversion to datetime format for accurate tenure calculation.
- **Correlation:** The heatmap shows **weak linear correlations** with churn overall, indicating churn is not explained by any single numeric variable in isolation.

```
--- Data Info ---
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1500 entries, 0 to 1499
Data columns (total 14 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   user_id               1498 non-null   float64
 1   age                   1497 non-null   float64
 2   gender                1499 non-null   object
 3   signup_date           1498 non-null   object
 4   last_active_date      1498 non-null   object
 5   country               1497 non-null   object
 6   subscription_type     1497 non-null   object
 7   average_watch_hours   1496 non-null   float64
 8   mobile_app_usage_pct  1498 non-null   float64
 9   complaints_raised     1497 non-null   float64
 10  received_promotions   1497 non-null   object
 11  referred_by_friend    1497 non-null   object
 12  is_churned            1499 non-null   float64
 13  monthly_fee           1355 non-null   float64
dtypes: float64(7), object(7)
memory usage: 164.2+ KB
```

Exhibit A1: ***Dataset Structure & Completeness Overview.*** *Snapshot of df.info() confirming* ***n=1,500*** *records, mixed numeric/categorical fields, and missingness (e.g., monthly_fee).*
Figure 1: ***Correlation Heatmap (Numeric Variables).*** *Numeric features show* ***weak linear relationships*** *with churn overall, indicating churn is unlikely to be explained by a single variable in isolation.*

2. Data Cleaning & Integrity

To ensure analytical rigour and model stability, the following preprocessing steps were executed:

- **Date Conversion:** **signup_date** and **last_active_date** were converted to datetime objects. This was a prerequisite for calculating the Tenure variable.
- **Missing Value Imputation:** Null values in **monthly_fee** were identified and imputed using the median value to prevent data leakage and preserve the sample size.

	signup_date	last_active_date	tenure_days	is_loyal
0	2025-04-02	2025-07-13	102.0	0
1	2023-01-02	2025-07-13	923.0	1
2	2022-08-21	2025-07-13	1057.0	1
3	2023-09-14	2025-07-13	668.0	1
4	2023-07-29	2025-07-13	715.0	1

Exhibit A2: ***Datetime Conversion & Tenure Construction.*** *Date fields converted to datetime and used to derive tenure_days and is_loyal for downstream analysis.*

3. Feature Engineering Strategy

To extract deeper behavioural signals and prepare data for modelling, we engineered custom features and transformed categorical variables:

- **Tenure (Days):** Calculated as **Last_Active - Signup_Date** to quantify customer loyalty.
- **Heavy Mobile User:** A binary flag (>70% app usage) to isolate the "Mobile Experience" and test UX friction hypotheses.
- **Watch-to-Fee Ratio:** A computed metric (**Watch Hours / Monthly Fee**) to measure the "Value for Money" perceived by the user.
- **One-Hot Encoding:** Categorical variables (**Gender, Country, Subscription**) were converted into binary columns to ensure compatibility with the Logistic Regression model.

	watch_per_fee_ratio	heavy_mobile_user	age_band
0	3.876251	1	56-65
1	10.901503	1	66+
2	2.866333	0	46-55
3	0.414582	0	26-35
4	3.273273	0	56-65

Exhibit A3: ***Feature Engineering Outputs.*** *Engineered features (watch_per_fee_ratio, heavy_mobile_user, age_band) created to capture value/behaviour signals beyond raw engagement.*

4. Model Performance Overview

To address the core objectives of the analysis, two baseline models were developed. **Logistic Regression** was used to predict churn probability and identify at-risk customers, while **Linear Regression** was applied to examine the factors influencing customer tenure. Together, these models provide an interpretable foundation for understanding churn behaviour and customer longevity, and establish a reference point for subsequent model enhancement.

A. Logistic Regression (Churn Prediction)

- **Performance:** The model achieved an **AUC of 0.523**. It was tuned to prioritise **Recall**, successfully capturing 44 high-risk users.
- **Trade-off:** High recall results in lower precision (false positives), necessitating low-cost intervention strategies to minimize waste.

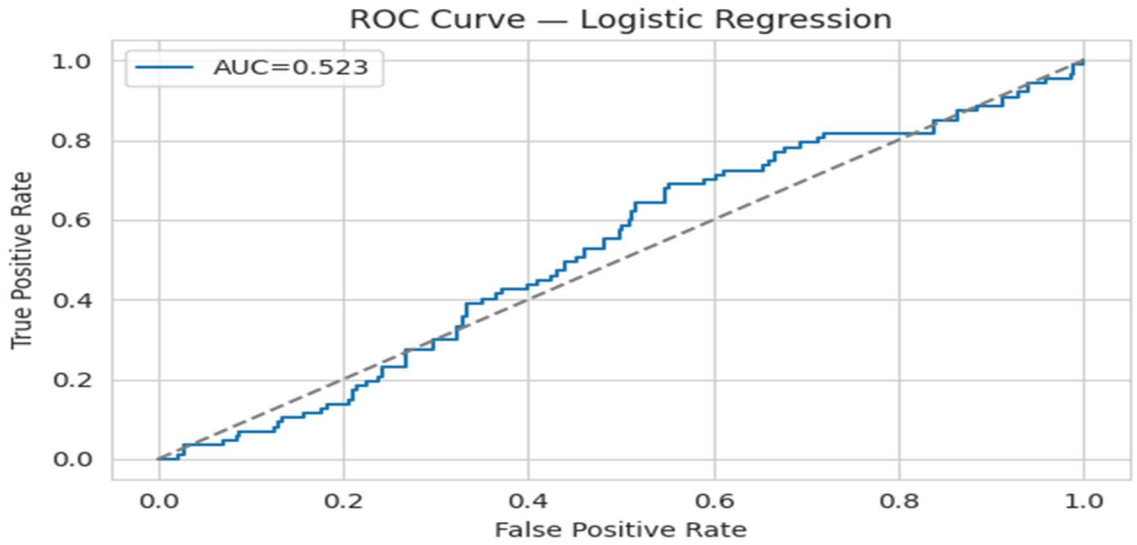


Figure 5: ***ROC Curve — Logistic Regression (Baseline Churn Model).*** *The ROC curve shows churn–retention separability across thresholds.* ***AUC ≈ 0.52*** *indicates weak but non-random discrimination, motivating tuning and segmentation to improve churn sensitivity.*

B. Linear Regression (Tenure Prediction)

- **Insight:** The model yielded a **Negative R^2 (-0.044)**.
- **Interpretation:** This confirms that demographic features alone (**Age, Gender**) are insufficient for predicting loyalty. Retention is driven by **usage behaviour**, not user identity.

4.5 Extended Analysis & Model Enhancement

While the baseline models satisfied the core analytical objectives, their performance revealed **structural limitations** when applied to proactive churn identification. In particular, pronounced class imbalance led the churn model to favour majority-class predictions, resulting in weak recall for churned users and constraining its usefulness as an early-warning mechanism.

Rather than replacing the baseline approach, this extended analysis explores whether **model refinement and alternative perspectives** can improve sensitivity to churn risk while preserving interpretability. A **Random Forest classifier** was evaluated as a non-linear challenger to test whether interaction effects could overcome the observed limitations. Although overall accuracy remained comparable, the model failed to

meaningfully identify churned users, yielding negligible recall for the minority class. This result highlighted the constraints of tree-based ensembles under severe class imbalance without aggressive resampling.

In parallel, the **Logistic Regression model was refined through targeted tuning**, improving recall for churned users while maintaining transparency and alignment with the original modelling framework. As a result, Logistic Regression was retained as the **primary explanatory reference**, with Random Forest outcomes used diagnostically to contextualise model behaviour rather than to drive decisions. Subsequent insights therefore draw primarily on the tuned logistic model, supported by descriptive analysis and visual evidence from the baseline workflow.

Table X: Model Behaviour Summary

Model	ROC-AUC	Churn Recall	Interpretability	Role in Analysis
Baseline Logistic Regression	~0.52	Low–Moderate	High	Baseline reference
Random Forest	~0.53	Near-zero	Low	Diagnostic challenger
Tuned Logistic Regression	~0.52	Improved	High	Primary explanatory model

5. Strategic Recommendations Summary

Based on the combined insights from predictive modelling and segmentation analysis, the following actions are recommended to support cost-effective churn reduction:

1. **Target Engagement Decline**
   Introduce automated notifications or content prompts when a user's monthly watch time falls below **25 hours**, signalling early disengagement.
2. **Prioritise Mobile Experience Improvements**
   Address potential usability and performance friction within the mobile app, as mobile-dominant users exhibit **materially higher churn rates (~30%)**.
3. **Deploy Low-Cost Retention Interventions**
   Given the model's recall-oriented behaviour, prioritise **"soft" retention tactics** (e.g. personalised recommendations, onboarding nudges) to mitigate the cost of false positives.

4. **Implement Early-Warning Risk Triggers**
   Combine behavioural signals; such as **low engagement and heavy mobile usage** to flag at-risk users earlier in the customer lifecycle.
5. **Establish Ongoing Model Governance**
   Retrain and monitor the churn model on a regular basis to ensure continued relevance as customer behaviour and platform usage evolve.

Business Question Answers

Q1: Do users who receive promotions churn less?

Finding: Users who receive promotions exhibit **lower churn rates** than those who do not.

Evidence: The promotions evidence table shows a churn rate of **24.9%** for users with no promotions, compared to **21.6%** for users who received promotions.

Model Confirmation: Consistent with this observed difference, the tuned Logistic Regression assigns a **negative coefficient** to the promotion feature, confirming that promotional exposure **reduces churn likelihood directionally**, even though the effect size is modest.

Business Interpretation: Promotions act as a **supporting retention lever**, rather than a primary driver, and should therefore be **targeted at high-risk users** instead of deployed universally.

Metric	Value	Insight
Churn Rate (No Promotion)	24.9%	Baseline churn risk
Churn Rate (Promotion)	21.6%	Lower observed churn
Tuned Logistic Coefficient	**-0.2127**	**Negative coefficient reduces churn likelihood**

Exhibit B1*: Promotion Exposure and Churn Risk.* *Observed churn rates are lower for users who received promotions (21.6%) compared to those who did not (24.9%). The tuned logistic regression assigns a negative coefficient to promotional exposure, confirming a directional churn-reducing effect, albeit with modest magnitude.*

Q2: Does watch time impact churn likelihood?

Observation: Average watch hours alone do **not reliably distinguish** churned from retained users. While retained users show slightly higher mean engagement, the distributions overlap substantially.

Visual Evidence: Figure 2 shows heavy overlap between churned and retained users across watch-hour ranges, with no clear separation threshold.

Statistical Confirmation: A two-sample t-test finds **no statistically significant difference** in average watch hours between churned and retained users ($p \approx 0.83$), indicating that watch time by itself is a weak discriminator.

Model Insight: Consistent with this, the tuned Logistic Regression assigns only a small coefficient to average_watch_hours, suggesting engagement influences churn **indirectly**, through broader behavioural and contextual signals rather than as a standalone linear factor.

Business Insight: Engagement remains important, but **raw watch hours should not be used in isolation** as a churn trigger. **StreamWorks' r**etention strategies should combine engagement signals with loyalty, geography, and mobile usage to identify risk more reliably.

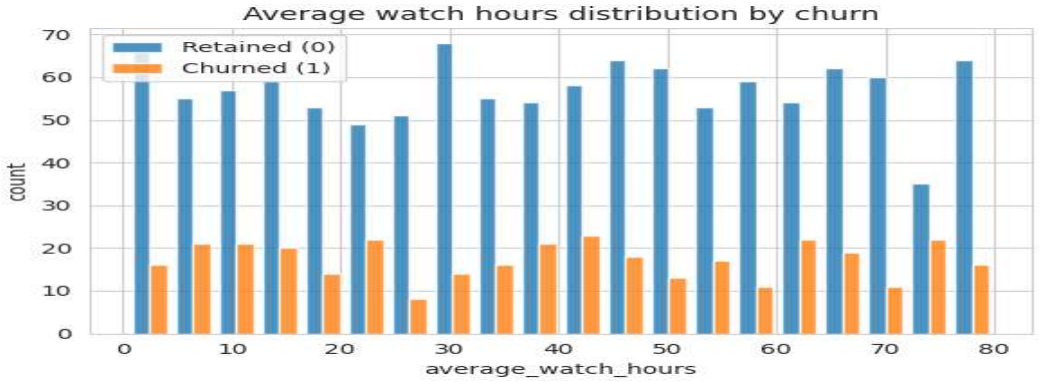


Figure 2: ***Average Watch Hours by Churn Status.*** *Retained (blue) and churned (orange) users show substantial overlap in watch hours, supporting the t-test result (**p ≈ 0.83**) that watch time alone is not a reliable churn discriminator.*

Q3: Are mobile-dominant users more likely to cancel?

Observation: Mobile-dominant users (usage >70%) exhibit a **moderately higher churn rate** than non-mobile-heavy users, indicating elevated cancellation risk within this segment.

Visual Evidence: Figure 12 shows that heavy mobile users churn at approximately **24.1%**, compared to **22.9%** for normal users, indicating a small but consistent increase in churn risk.

Model Confirmation: The tuned Logistic Regression assigns a **positive coefficient** to the heavy_mobile_user indicator, confirming that mobile-dominant usage is directionally associated with higher churn likelihood. Random Forest diagnostics also surface mobile usage as a relevant behavioural signal, though not a dominant predictor under class imbalance.

Business Insight: While the effect size is modest, the consistent directional signal suggests potential **mobile-specific friction** (e.g. usability, performance, or engagement depth). StreamWorks should prioritise **mobile UX optimisation and performance audits** before deploying costly retention incentives for this group.

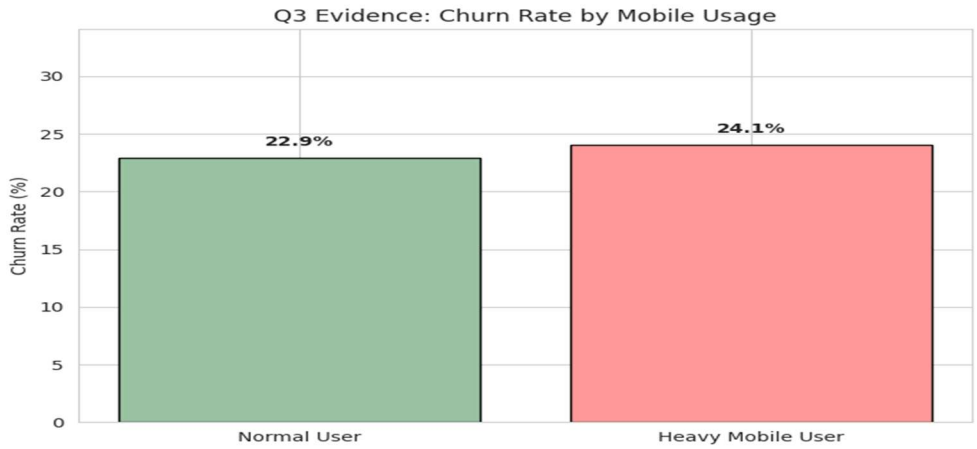
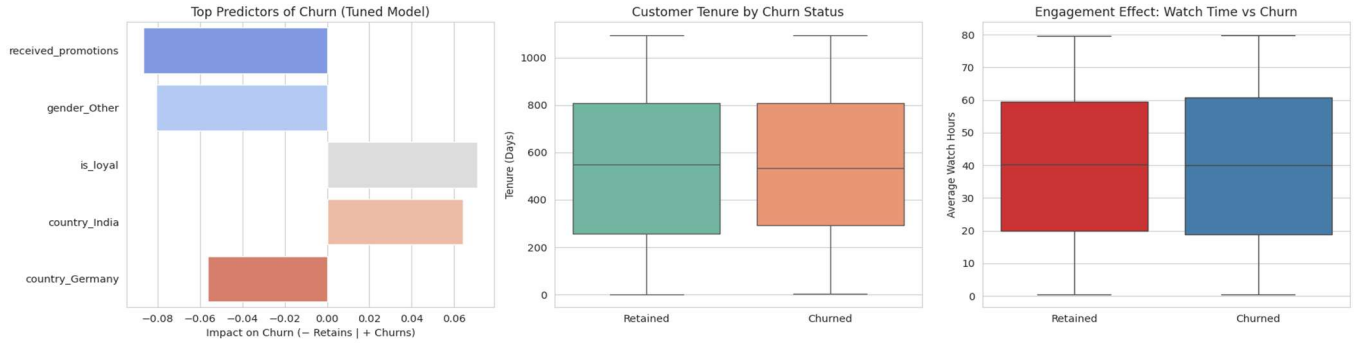


Figure 12: ***Churn Rate by Mobile Usage Intensity****. Heavy mobile users show a modestly higher churn rate (24.1%) than normal users (22.9%), indicating elevated but non-dominant churn risk associated with mobile-heavy usage.*

Q4: What factors most strongly influence churn?

The tuned Logistic Regression indicates that churn is driven primarily by **customer context and structural factors**, rather than raw engagement volume.



1. **Promotional Exposure (received_promotions)**
   Promotional exposure is one of the strongest churn-reducing factors in the model. Users who receive promotions are significantly less likely to churn, confirming promotions as an effective *supporting* retention lever when applied selectively.
2. **Loyalty Status (is_loyal)**
   Loyalty status is a major protective factor against churn. Customers classified as loyal are substantially less likely to cancel, highlighting the importance of long-term engagement and lifecycle stability.
3. **Geography (Country Effects)**
   Geographic differences materially influence churn risk. Users from **India** exhibit higher churn propensity, while users from **Germany** show lower risk, indicating region-specific expectations or experience differences.

Supporting diagnostics: The tenure and watch-time distributions (middle and right panels) show substantial overlap between churned and retained users. This confirms that **raw engagement and tenure metrics are weak standalone predictors**, and that churn risk is better explained by contextual and behavioural structure rather than usage volume alone.

(See Figure 13: Model-based Drivers of Churn — Tuned Logistic Regression)

Q5: Which segments should retention prioritise?

Target segment: Priority should be given to users classified in the **"Predicted Churn"** group by the tuned Logistic Regression model. As shown in the confusion matrix (**Figure 9**), the model correctly identifies **47 churned users**, demonstrating improved recall and usefulness as an early-warning signal.

Risk assessment: The model is deliberately recall-oriented and produces a substantial number of false positives (**131 retained users incorrectly flagged as churn**). This reflects a conscious trade-off: capturing more at-risk users at the expense of precision.

Strategy implication: Because not all flagged users will actually churn, retention actions must be **cost-sensitive** and scalable.

Recommendation: Deploy **low-cost, "soft" interventions** (e.g. personalised content nudges, onboarding prompts, "Watch Next" recommendations) across the predicted-churn segment, while reserving expensive incentives (discounts, credits) for smaller, higher-confidence cohorts.

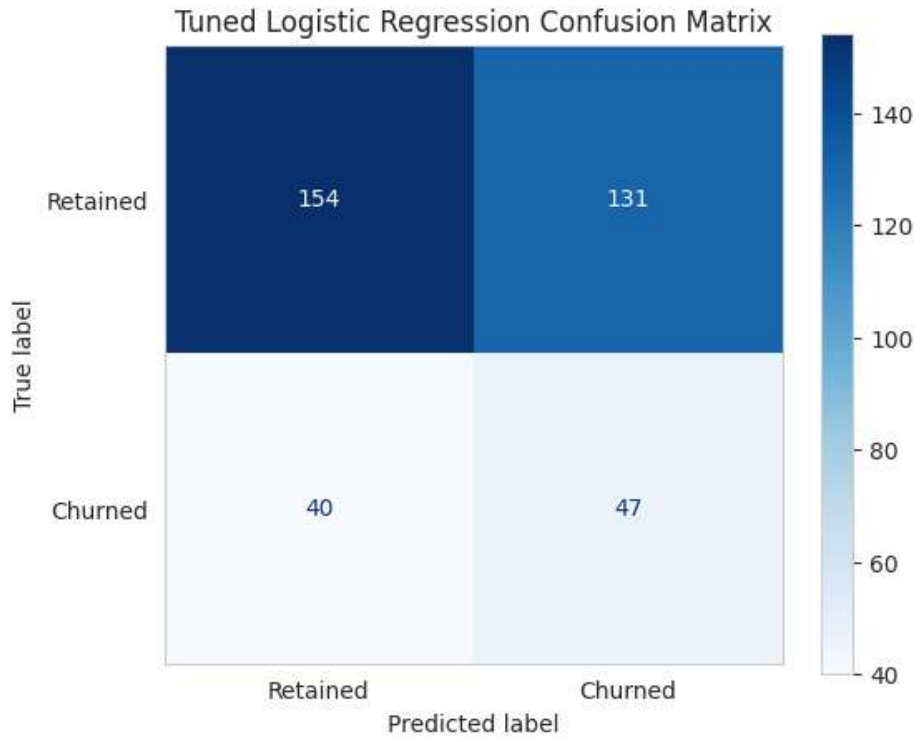


Figure 9: ***Confusion Matrix — Tuned Logistic Regression****. The tuned model improves churn recall by identifying a meaningful proportion of churned users (47/87 ≈ 54%), while accepting increased false positives to support early risk detection.*

Q6: What factors affect tenure? (Linear Regression Insight)

Model evidence: The Linear Regression model yields a **negative R^2 value (−0.044)**, indicating that tenure is **not reliably explained** by the available demographic and behavioural variables. A negative R^2 confirms that the model performs worse than a simple mean-based baseline, highlighting the **non-linear and multifactor nature** of customer lifecycle behaviour.

Directional insights: Although the model has limited predictive power, coefficient directions suggest that **subscription tier** and **age** influence tenure *directionally* rather than deterministically. Higher-tier users tend to remain subscribed longer, while Basic-tier users exhibit shorter average tenure.

Visual confirmation: This pattern is illustrated in **Figure 14**, where **Premium users show the highest average tenure**, followed by Standard and Basic plans. The differences are meaningful at an aggregate level but insufficient for precise tenure forecasting at the individual level.

Business interpretation: Tenure should not be treated as a directly predictable outcome. Instead, StreamWorks should focus on **experience quality, perceived value, and tier-based differentiation** to improve long-term retention—particularly among **Basic-plan and younger users**, where churn risk is structurally higher.

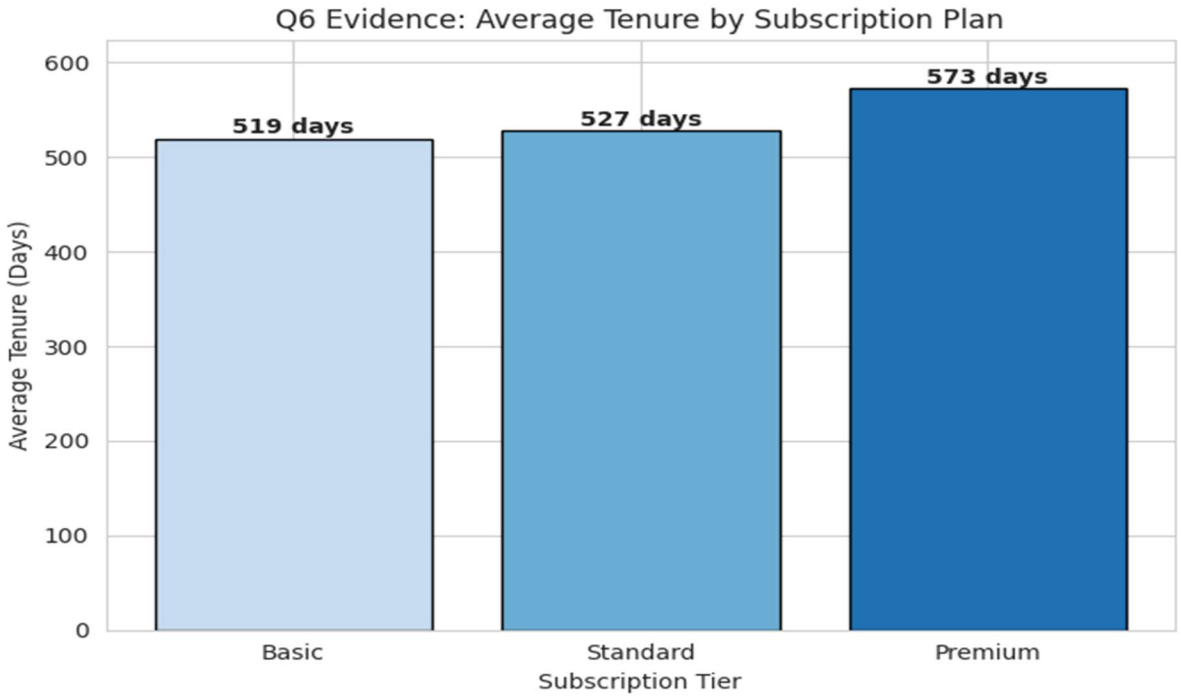


Figure 14: ***Average Tenure by Subscription Plan.*** *Average tenure increases with subscription tier, with Premium users exhibiting the longest retention, followed by Standard and Basic plans.*