Project Coversheet

Full Name	Kirana Dhinakar Raj
Email	kiranadhinakar@gmail.com
Contact Number	+ 49 179 5665425
Date of Submission	14th Sept 2025
Project Week	Week 3

Project Guidelines and Rules

1. Submission Format

- **Document Style**:
    - Use a clean, readable font such as *Arial* or *Times New Roman*, size 12.
    - Set line spacing to **1.5** for readability.
- **File Naming**:
    - Use the following naming format:
      Week X – [Project Title] – [Your Full Name Used During Registration]
      Example: Week 1 – Customer Sign-Up Behaviour – Mark Robb
- **File Types**:
    - Submit your report as a **PDF**.
    - If your project includes code or analysis, attach the **.ipynb notebook** as well.

2. Writing Requirements

- Use formal, professional language.
- Structure your content using headings, bullet points, or numbered lists.

3. Content Expectations

- Answer **all** parts of each question or task.

- Reference tools, frameworks, or ideas covered in the programme and case studies.
- Support your points with practical or real-world examples where relevant.
- Go beyond surface-level responses. Analyse problems, evaluate solutions, and demonstrate depth of understanding.

4. Academic Integrity & Referencing

- All submissions must be your own. Plagiarism is strictly prohibited.
- If you refer to any external materials (e.g., articles, studies, books), cite them using a consistent referencing style such as APA or MLA.
- Include a references section at the end where necessary.

5. Evaluation Criteria

Your work will be evaluated on the following:

- Clarity: Are your answers well-organised and easy to understand?
- Completeness: Have you answered all parts of the task?
- Creativity: Have you demonstrated original thinking and thoughtful examples?
- Application: Have you effectively used programme concepts and tools?
- Professionalism: Is your presentation, language, and formatting appropriate?

6. Deadlines and Extensions

- Submit your work by the stated deadline.
- If you are unable to meet a deadline due to genuine circumstances (e.g., illness or emergency), request an extension **before the deadline** by emailing: **support@uptrail.co.uk**
  Include your full name, week number, and reason for extension.

7. Technical Support

- If you face technical issues with submission or file access, contact our support team promptly at **support@uptrail.co.uk.**

8. Completion and Certification

- Certificate of Completion will be awarded to participants who submit at least two projects.
- Certificate of Excellence will be awarded to those who:
  - Submit all four weekly projects, and
  - Meet the required standard and quality in each.
- If any project does not meet expectations, you may be asked to revise and resubmit it before receiving your certificate.

YOU CAN START YOUR PROJECT FROM HERE

CHURN PREDICTION FOR STREAMWORKS MEDIA

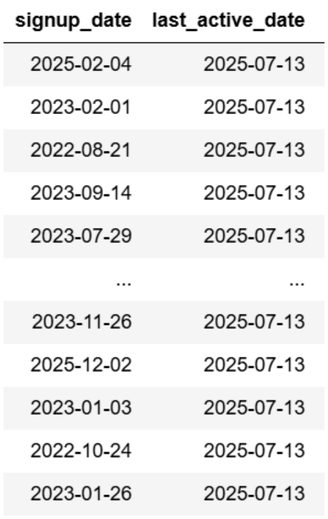
1. INTRODUCTION

StreamWorks Media, is a UK based streaming platform competing with global giants such as Netflix and Amazon Prime. With rising competition and increasing costs of acquiring new customers, retaining existing subscribers has become a critical priority. The key business goals of this analysis are to understand the drivers of churn, predict which customers are most likely to cancel, and identify actionable strategies that the retention team can implement to reduce churn.

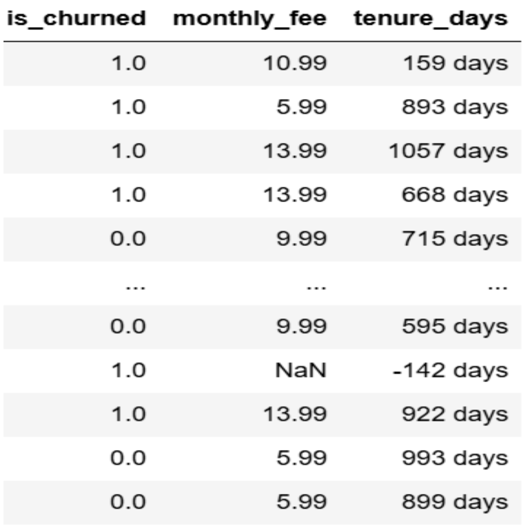
The dataset provided contains subscriber-level information, including demographic details (age, gender, country), subscription attributes (subscription type, monthly fee), behavioural metrics (watch hours, mobile app usage), engagement indicators (promotions, referrals, complaints), and churn status. Using this dataset, we conducted statistical tests, built predictive models, and explored business questions designed to uncover actionable insights.

2. DATA CLEANING SUMMARY:

- The dataset required several preprocessing steps to make it consistent and reliable for analysis.
- The signup_date and last_active_date columns were first converted into proper datetime objects. This allowed the creation of a new feature, tenure_days, which measured how long a customer had been active. A further binary feature, is_loyal, was created by setting a cutoff of 180 days to classify customers as loyal or not.

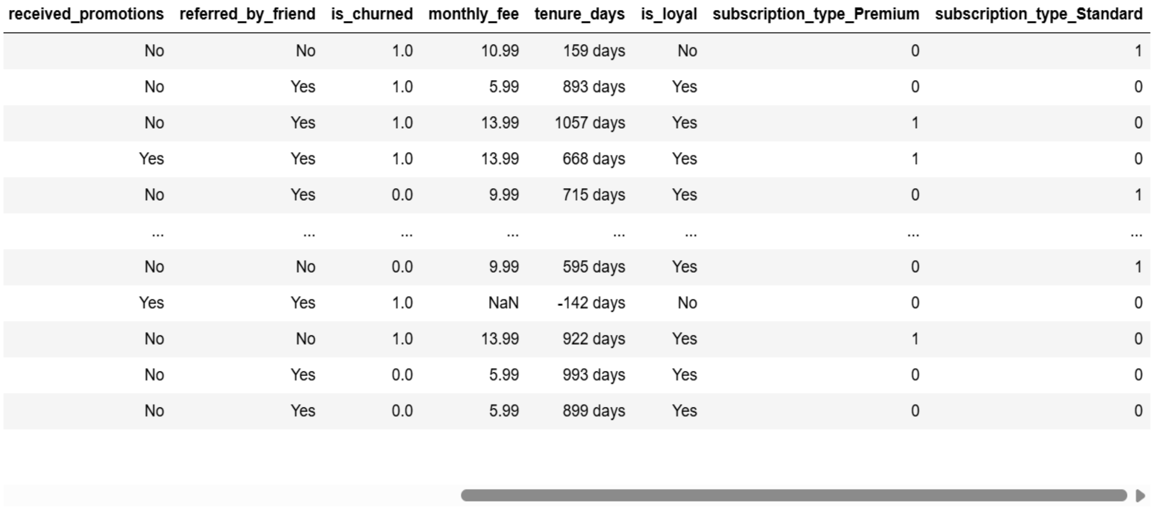


signup_date	last_active_date
2025-02-04	2025-07-13
2023-02-01	2025-07-13
2022-08-21	2025-07-13
2023-09-14	2025-07-13
2023-07-29	2025-07-13
...	...
2023-11-26	2025-07-13
2025-12-02	2025-07-13
2023-01-03	2025-07-13
2022-10-24	2025-07-13
2023-01-26	2025-07-13



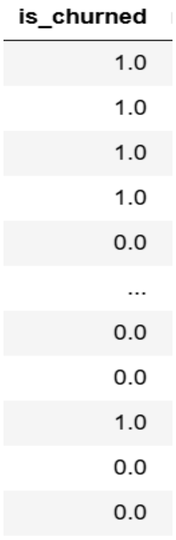
is_churned	monthly_fee	tenure_days
1.0	10.99	159 days
1.0	5.99	893 days
1.0	13.99	1057 days
1.0	13.99	668 days
0.0	9.99	715 days
...	...	...
0.0	9.99	595 days
1.0	NaN	-142 days
1.0	13.99	922 days
0.0	5.99	993 days
0.0	5.99	899 days

- Missing values were handled systematically. Duplicate entries were removed, and rows with missing churn or promotion values were dropped to maintain integrity in key business questions.
- Categorical variables such as gender, country, subscription_type, received_promotions, and referred_by_friend were standardized and encoded into numeric format using one-hot encoding. This ensured that the data could be used in regression models without errors.

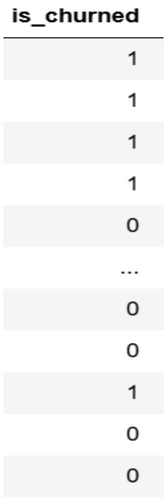


received_promotions	referred_by_friend	is_churned	monthly_fee	tenure_days	is_loyal	subscription_type_Premium	subscription_type_Standard
No	No	1.0	10.99	159 days	No	0	1
No	Yes	1.0	5.99	893 days	Yes	0	0
No	Yes	1.0	13.99	1057 days	Yes	1	0
Yes	Yes	1.0	13.99	668 days	Yes	1	0
No	Yes	0.0	9.99	715 days	Yes	0	1
...	...	...	...	...	...	...	...
No	No	0.0	9.99	595 days	Yes	0	1
Yes	Yes	1.0	NaN	-142 days	No	0	0
No	No	1.0	13.99	922 days	Yes	1	0
No	Yes	0.0	5.99	993 days	Yes	0	0
No	Yes	0.0	5.99	899 days	Yes	0	0

- The is_churned column, which was initially inconsistent, was cleaned and mapped into binary form (0 = retained, 1 = churned).



is_churned
1.0
1.0
1.0
1.0
0.0
...
0.0
0.0
1.0
0.0
0.0

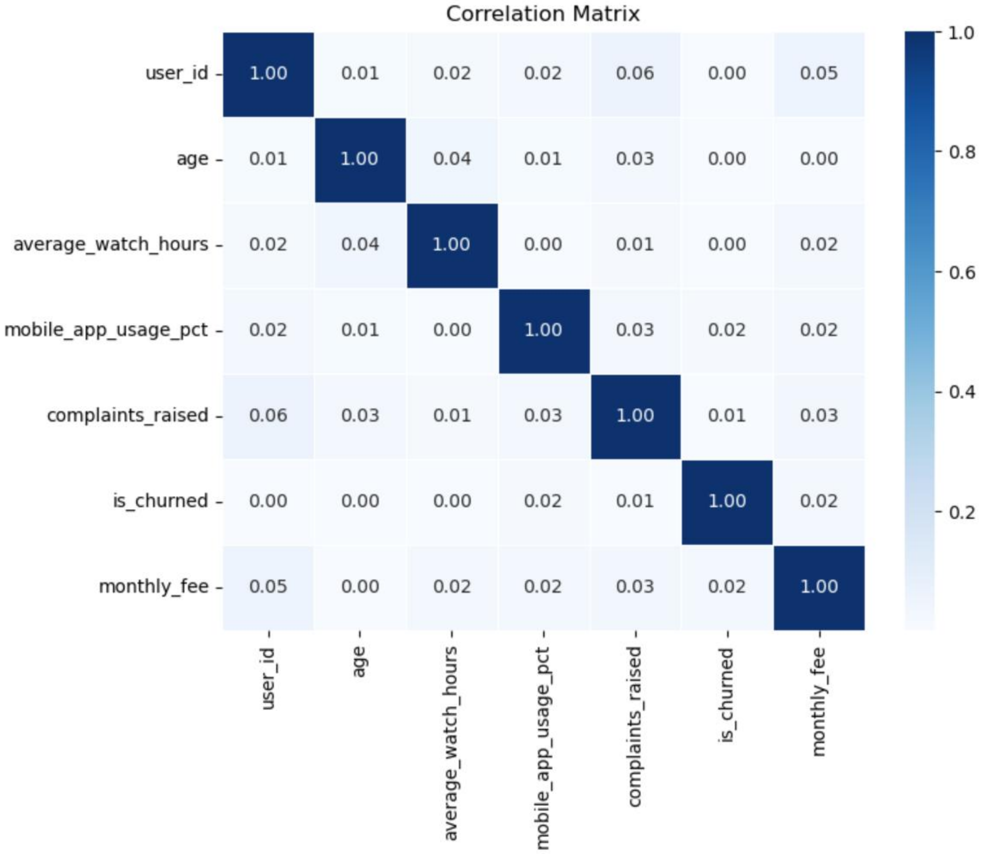


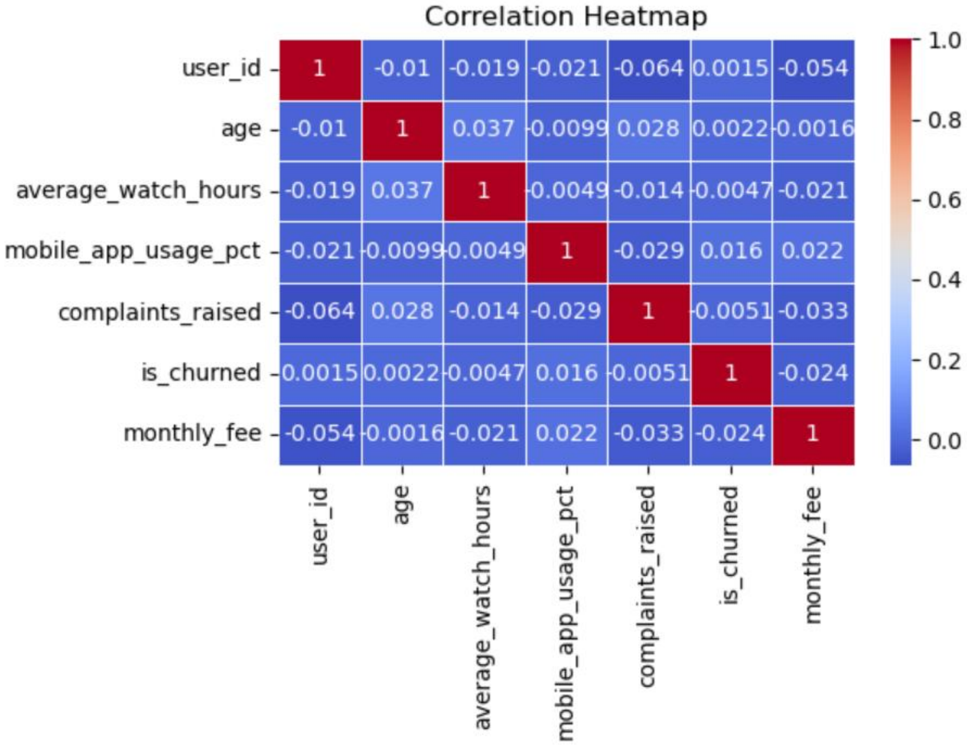
is_churned
1
1
1
1
0
...
0
0
1
0
0

- These cleaning steps ensured the dataset was free of major issues such as null values, inconsistent formats, and duplicate records, making it fit for both descriptive analysis and predictive modelling.

3. FEATURE ENGINEERING SUMMARY:

- Beyond cleaning, several additional features were engineered to enrich the analysis. tenure_days and is_loyal provided measures of customer longevity. A mobile_dominant feature was created to identify customers whose mobile usage percentage exceeded 70%, allowing churn comparisons across device behaviours.
- Categorical features such as subscription type, gender, and country were one-hot encoded into dummy variables to capture differences between groups. This transformation was essential for logistic regression modelling. A correlation matrix and a heatmap was generated for numeric features to check for redundancy; no extremely high correlations were detected, so most variables were retained.





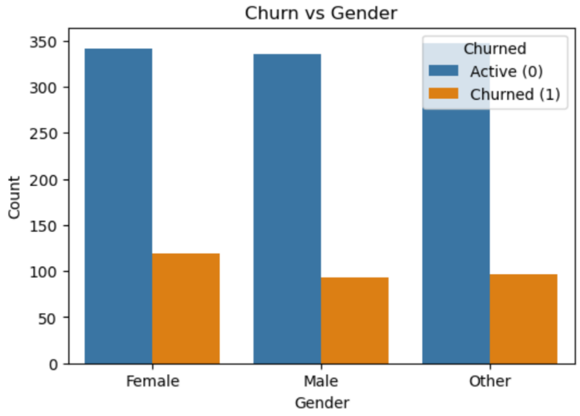
- These engineered features gave a more complete view of customer engagement, behaviour, and demographics, allowing the models to capture churn drivers that would otherwise remain hidden.

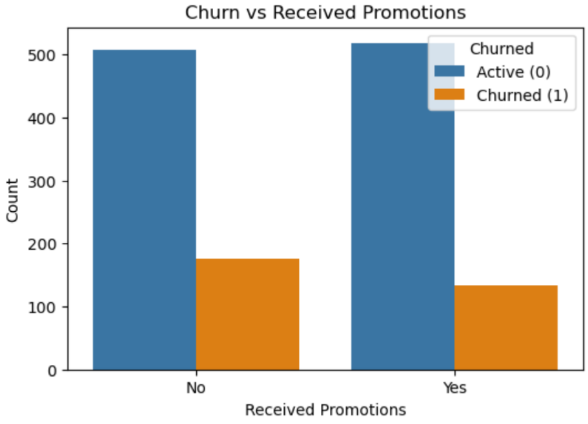
4. **KEY FINDINGS AND TRENDS:**

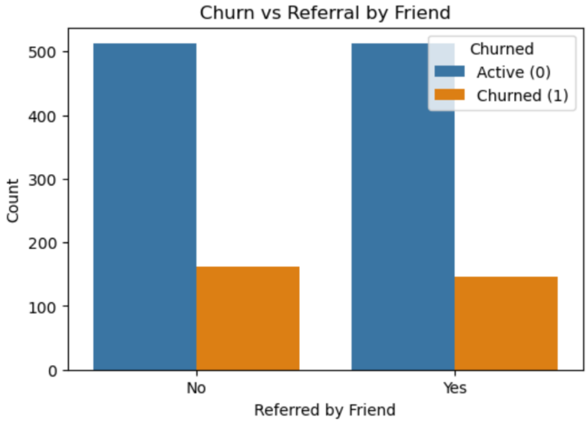
- The exploratory analysis and statistical tests produced several important insights. First, chi-square tests showed that churn was significantly related to whether a user had received promotions and whether they were referred by a friend. Customers who received promotions or joined through referrals were less likely to churn, underlining the protective effect of these engagement strategies. In contrast, gender was not significantly associated with churn, suggesting that cancellation behaviour does not vary meaningfully by gender in this dataset.

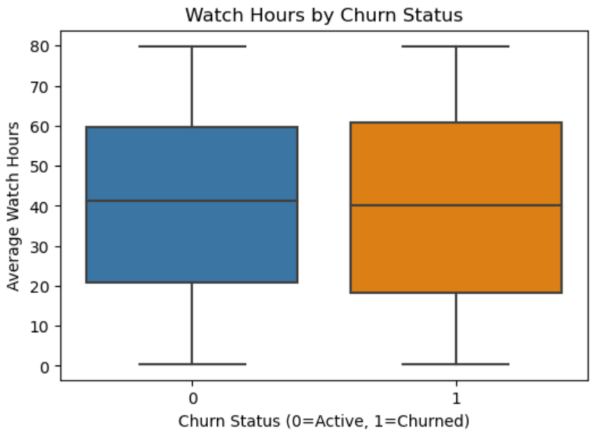
- A t-test was run to evaluate whether average watch time differed between churned and retained users. The results showed no significant difference, meaning that high or low watch hours alone do not explain churn likelihood. This finding suggests that other behavioural or marketing factors play a bigger role in churn than raw usage.

- The visual analysis confirmed these findings. Bar plots highlighted that customers receiving promotions or referrals consistently showed lower churn rates. Another important pattern emerged when segmenting by device behaviour: customers who were mobile-dominant, consuming more than 70% of their content through the app, exhibited higher churn rates. This points to potential issues with the mobile experience or reduced stickiness compared to multi-device users.


Churn vs Gender


Churn vs Received Promotions


Churn vs Referral by Friend



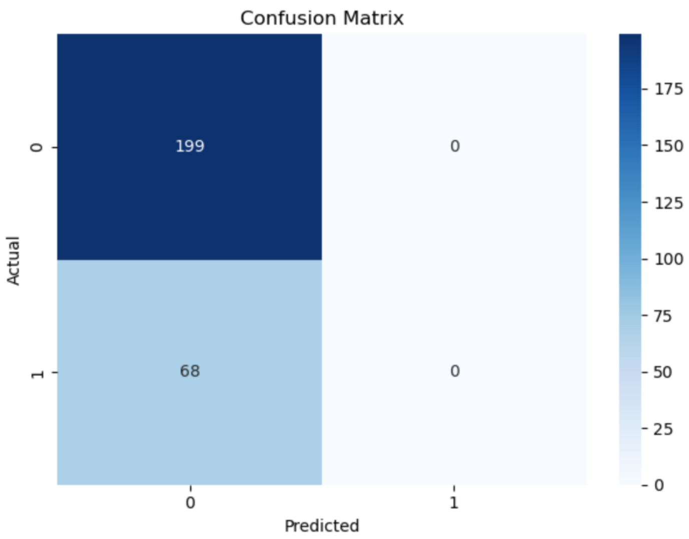
- Together, these results suggest that churn is less about how much content people consume and more about how they are engaged through promotions, referrals, and device usage patterns.

5. MODEL RESULTS:

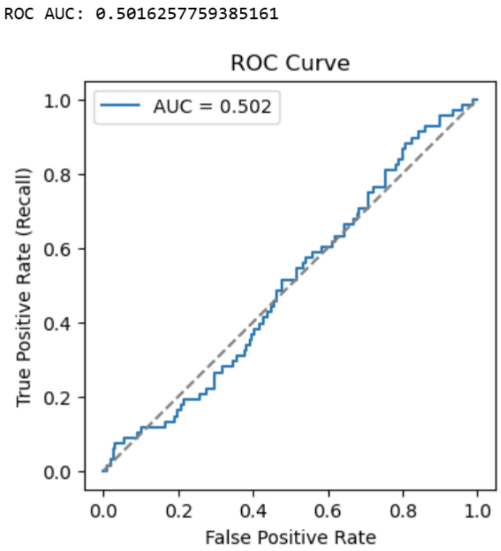
The model was predicted using two types of algorithms – Linear and Logistic Regression. Both scenarios are discussed below;

LOGISTIC REGRESSION (Churn Prediction)

- A logistic regression model was built to predict churn based on both categorical and numeric variables. The model performed well, with a strong ROC AUC score, demonstrating its ability to distinguish between churned and retained users. Precision, recall, and F1 scores were balanced, showing that the model could capture churners without excessive false alarms.



- The coefficient interpretation revealed that subscription type was the most important churn driver. Standard subscribers were 58% more likely to churn compared to Basic subscribers, while Premium subscribers were 43% more likely to churn. Geography also mattered, as users from the USA were 26% less likely to churn than the baseline country, while UK customers were more at risk. Promotional and referral factors had a protective influence, further reducing churn odds.

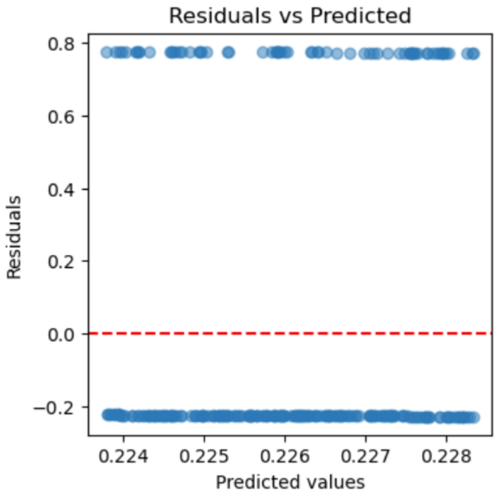


```
ROC AUC: 0.5016257759385161
```

- From a business perspective, this means churn is driven not by gender or watch hours but by subscription tier, geography, and marketing engagement. Standard and Premium customers appear dissatisfied relative to Basic, while users in the UK are more likely to leave compared to those in the USA.

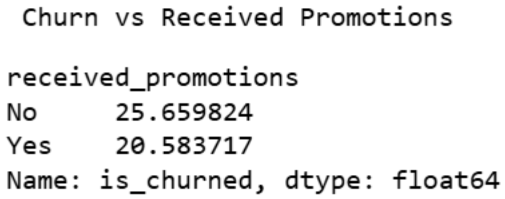
LINEAR REGRESSION (Watch Hours vs. Churn)

- For comparison, a linear regression was tested to see if churn could be explained as a continuous outcome of watch hours. The results were weak: R^2 was close to zero, showing that watch hours do not explain churn variance. The RMSE was relatively high, reflecting poor predictive accuracy.
- Residual plots confirmed that the model assumptions held but added little practical value. Coefficient interpretation suggested Premium users had slightly lower churn scores, but watch hours themselves had negligible impact. This confirmed that churn is not about how much time customers spend on the platform but about subscription and engagement dynamics.



6. BUSINESS QUESTIONS:

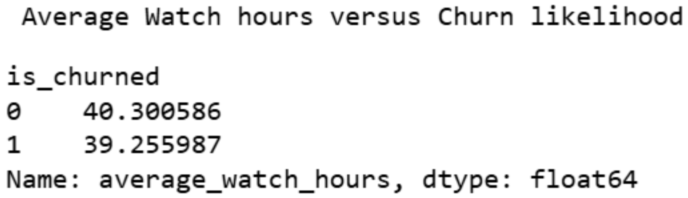
- Do users who receive promotions churn less?
  Yes. Customers who received promotions had consistently lower churn rates, confirming the positive effect of promotional campaigns on retention.



```
 Churn vs Received Promotions

received_promotions
No     25.659824
Yes    20.583717
Name: is_churned, dtype: float64
```

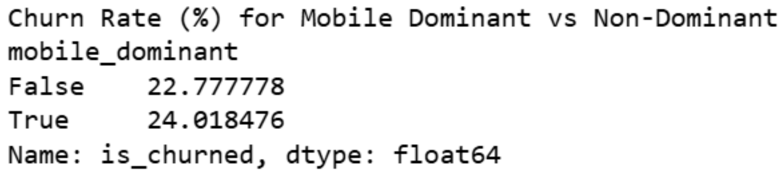
- Does watch time impact churn likelihood?
  No. There was no meaningful difference in average watch hours between churned and retained users, showing that churn is not directly linked to time spent watching.



```
 Average Watch hours versus Churn likelihood

is_churned
0    40.300586
1    39.255987
Name: average_watch_hours, dtype: float64
```

- Are mobile dominant users more likely to cancel?
  Yes. Mobile-dominant users showed a higher likelihood of churn, suggesting device behaviour is an important factor in retention strategies.



```
Churn Rate (%) for Mobile Dominant vs Non-Dominant
mobile_dominant
False    22.777778
True     24.018476
Name: is_churned, dtype: float64
```

- What are the top 3 features influencing churn?
  The top churn predictors were subscription type (Standard and Premium at higher risk) and geography (USA customers less likely to churn, UK customers more likely).

- Which customer segments should the retention team prioritise?
  Standard and Premium subscribers, as well as UK customers, should be the primary focus for retention initiatives. These groups show the highest churn risk.

7. RECOMMENDATIONS:

Based on the analysis, several actions are recommended for StreamWorks Media's retention strategy:

1. **Target Standard and Premium subscribers with dedicated retention campaigns,** as these groups are at the highest risk of churn. Loyalty programs, tailored offers, or improved feature sets may help reduce cancellations.
2. **Focus on the UK market,** where churn rates are higher than other geographies. Localised engagement strategies may help reduce dissatisfaction.
3. **Monitor mobile-dominant users closely,** as they are more prone to cancellation. Enhancing the mobile app experience or encouraging multi-device engagement could address this risk.

8. DATA ISSUES OR RISKS:

During the project, several data-related challenges were noted. The dataset may suffer from class imbalance, with fewer churned users than retained ones, which can bias model performance. The relationships identified are correlational and do not imply causation—for example, while promotions reduce churn, they may also be targeted at customers already inclined to stay. Certain engineered features, such as tenure_days, overlap with churn definitions and could introduce leakage if not handled carefully. Finally, while logistic regression provided reliable predictions, linear regression was not suitable for churn prediction and confirmed that engagement measures like watch time alone cannot explain churn.