Project Coversheet

Full Name	Polislava Kalcheva
Email	polislavakk@gmail.com
Contact Number	+359884761900
Date of Submission	12.08.2025
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

1. Introduction

I have joined the Data Strategy Team at StreamWorks Media, a fast-growing UK-based video streaming platform competing with global players like Netflix and Amazon Prime. With customer acquisition becoming more expensive and competition intensifying, my manager has tasked me with analysing customer churn, users who cancel their subscriptions.

2. Data Cleaning Summary

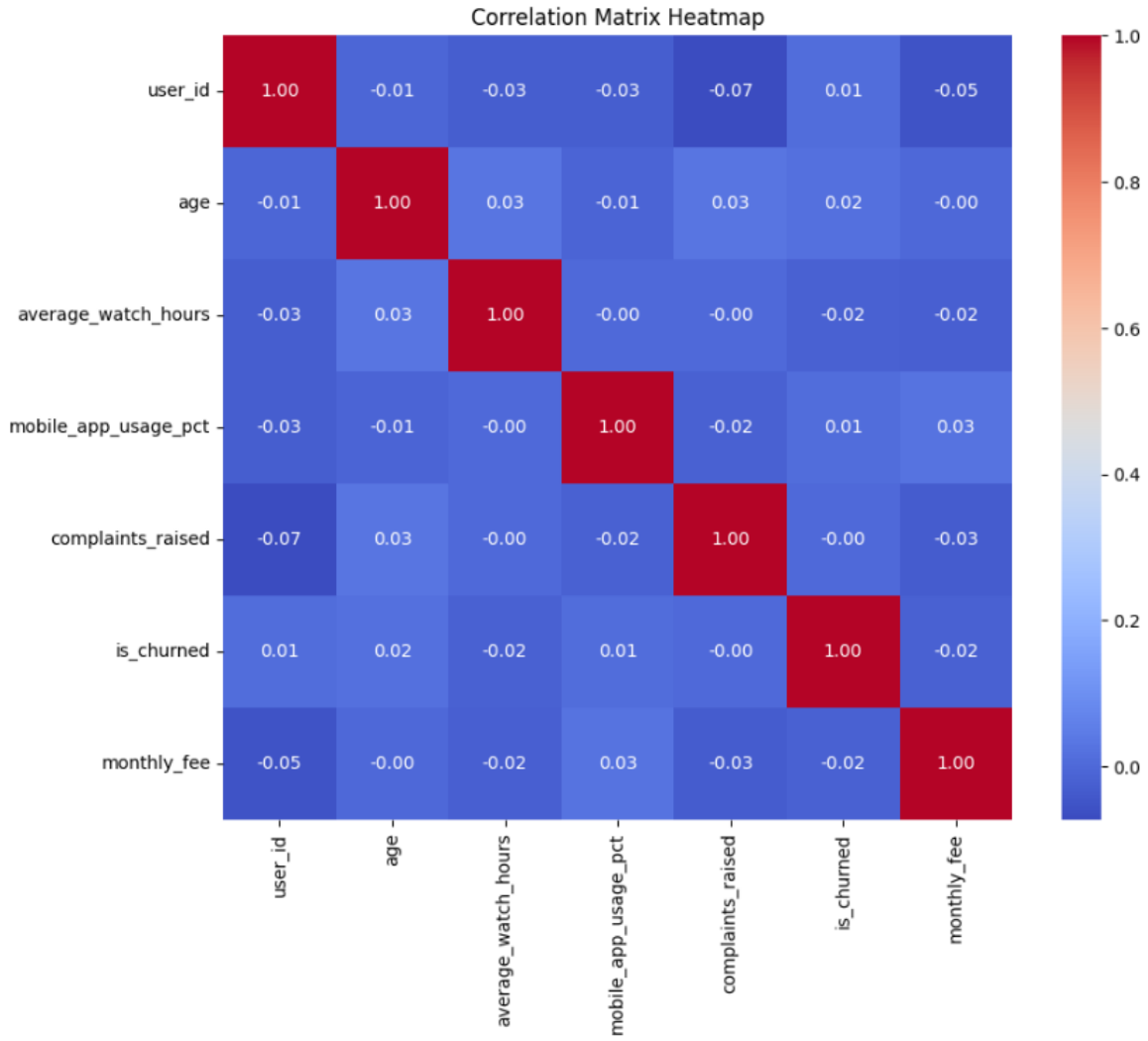
First the empty values were dropped from the numeric columns in the given database called 'streamworks_user_data.csv'. Then a heatmap was created with the help of a correlation matrix. Therefore, the correlation matrix values are the following.

	user_id	age	average_watch_hours	\
user_id	1.000000	-0.007649	-0.028472	
age	-0.007649	1.000000	0.029228	
average_watch_hours	-0.028472	0.029228	1.000000	
mobile_app_usage_pct	-0.026929	-0.013258	-0.000745	
complaints_raised	-0.072418	0.031931	-0.001006	
is_churned	0.007697	0.017001	-0.018079	
monthly_fee	-0.053461	-0.002018	-0.023123	

	mobile_app_usage_pct	complaints_raised	is_churned	\
user_id	-0.026929	-0.072418	0.007697	
age	-0.013258	0.031931	0.017001	
average_watch_hours	-0.000745	-0.001006	-0.018079	
mobile_app_usage_pct	1.000000	-0.019674	0.010189	
complaints_raised	-0.019674	1.000000	-0.001017	
is_churned	0.010189	-0.001017	1.000000	
monthly_fee	0.025936	-0.032747	-0.019127	

	monthly_fee
user_id	-0.053461
age	-0.002018
average_watch_hours	-0.023123
mobile_app_usage_pct	0.025936
complaints_raised	-0.032747
is_churned	-0.019127
monthly_fee	1.000000

And then the heatmap looks like the following image.



Moreover, two columns called ‘signup_date’ and ‘last_active_date’ were converted into date time. See the sample below.

signup_date	last_active_date
2025-04-02	2025-07-13
2023-01-02	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
2025-02-12	2025-07-13
2023-03-01	2025-07-13
2022-10-24	2025-07-13
2023-01-26	2025-07-13

Another modification is that missing values in the table were dropped, because on the visualizations some features will not appear when containing empty values.

3. Feature Engineering Summary

The newly created features are called ‘tenure_days’ and ‘is_loyal’. The first is about the values of the days between ‘signup’ and ‘last_active_date’ features. And the latter is a binary column filtering the tenure days with ‘True’ for more than 180 days and ‘False’ for less or equal to 180 days. Check the table below.

tenure_days	is_loyal
102.0	False
923.0	True
1057.0	True
668.0	True
715.0	True
...	...
595.0	True
151.0	False
865.0	True
993.0	True
899.0	True

Moreover, using dummy variables, categorical features were encoded using the features 'gender', 'country', 'subscription_type' and making them into binary columns also called as one-hot encoding. For example, gender was split into gender_Male, gender_Other (drops gender_Female), country was split into multiple columns such as country_Germany, country_India, country_UK, etc. (dropping the first one alphabetically), and subscription_type becomes: subscription_type_Premium, subscription_type_Standard (drops Basic). Check the sample below.

gender_Male	gender_Other	country_France	country_Germany	country_India	country_UK	country_USA	subscription_type_Premium	subscription_type_Standard
False	True	True	False	False	False	False	False	True
True	False	False	False	True	False	False	False	False
True	False	False	False	False	True	False	True	False
False	True	False	True	False	False	False	True	False
False	False	False	False	True	False	False	False	True
...	...	...	...	...	...	...	...	...
False	False	False	False	False	False	False	False	True
True	False	False	False	False	False	True	False	False
False	True	False	False	False	True	False	True	False
False	False	False	False	False	False	True	False	False
False	True	False	False	False	False	False	False	False

Another created features are ‘watch_per_fee_ratio’ and ‘heavy_mobile_user’. The first is about ‘average_watch_hours’ devided by ‘monthly_fee’ and the latter is about ‘mobile_app_usage_pct’ values higher than 70 (also every value was converted into numeric value). You can see how the new features look like below.

	tenure_days	is_loyal	watch_per_fee_ratio	heavy_mobile_user
0	102.0	False	3.876251	1
1	923.0	True	10.901503	1
2	1057.0	True	2.866333	0
3	668.0	True	0.414582	0
4	715.0	True	3.273273	0

There are transformations made in the newly added ‘tenure_days’ and ‘watch_per_fee_ratio’ features only because they are numeric and ‘is_loyal’ and ‘heavy_mobile_user’ are binary so they do not need transformation. Consequently, on the two numeric features Log transform was applied then Min–max scaling (0–1). You can compare the old and the new versions of the two features below.

```
   tenure_days  log_tenure_days  tenure_days_scaled  watch_per_fee_ratio  \
0        102.0         4.634729            0.092322             3.876251
1        923.0         6.828712            0.842779            10.901503
2       1057.0         6.964136            0.965265             2.866333
3        668.0         6.505784            0.609689             0.414582
4        715.0         6.573680            0.652651             3.273273

   log_watch_per_fee_ratio  watch_per_fee_ratio_scaled
0                 1.584377                    0.288819
1                 2.476665                    0.816814
2                 1.352307                    0.212917
3                 0.346834                    0.028652
4                 1.452380                    0.243501
```

Moreover, it was checked if there were any remaining features left and if there are, one-hot encoding had to be applied but the result showed that all features are good to go.

```
Categorical columns before encoding: ['received_promotions', 'referred_by_friend']
Categorical columns after encoding: []
```

Another modification was made- binning: two new features called ‘tenure_bin’ categorizing the ‘tenure_days’ if they are fewer than 6 months, more than 2 years or between 1 and 2 years entries and ‘watch_ratio_bin’ categorizing if the ratio in ‘watch_per_fee_ration’ is medium high, high, medium low, or low. Check below the sample.

	tenure_days	tenure_bin	watch_per_fee_ratio	watch_ratio_bin
0	102.0	<6m	3.876251	med_high
1	923.0	>2y	10.901503	high
2	1057.0	>2y	2.866333	med_low
3	668.0	1–2y	0.414582	low
4	715.0	1–2y	3.273273	med_low

Another thing is that an interaction between ‘received_promotions’ and ‘low_watch_time’ features was made. It shows that promoted users who still watch very short time – under 20 hours, out of all users, 156 people are in this group. Therefore, promotions did not increase engagement in watching.

```
Low-watch threshold (hours): 20.1
   received_promotions_Yes  average_watch_hours  promo_low_watch
0                    False                 42.6                0
1                    False                 65.3                0
2                    False                 40.1                0
3                     True                  5.8                1
4                    False                 32.7                0
Counts:
 promo_low_watch
0    1181
1     156
```

The last moditication was dropping low-variance features. But the result showed that there are no non-informational (low-variance) columns to drop.

```
Low-variance columns to drop: []
Remaining columns: 31
```

4. Key Findings

Chi-square test was applied to check if churn is related to ‘gender’, ‘received_promotions’, or ‘referred_by_friend’. The first output is about ‘gender’.

```
is_churned  0.0  1.0
row_0
Female      343  119
Male        336   93
Other       349   97
Chi-square (gender vs is_churned): chi2=2.7822, dof=2, p=0.2488
```

The next image shows the ‘received_promotions’, or ‘referred_by_friend’ statistics.

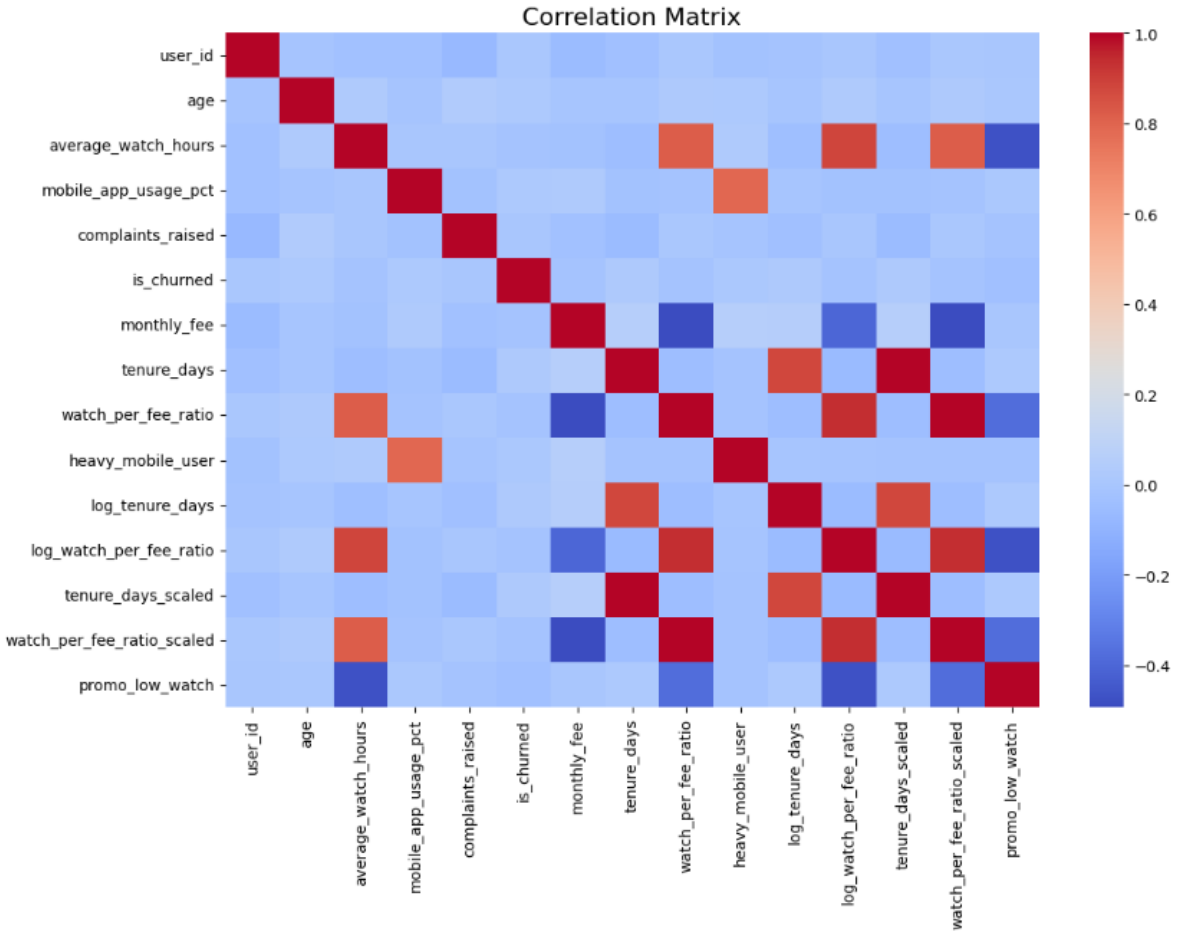
```
received_promotions vs is_churned (using 'received_promotions_Yes')
is_churned                0.0  1.0
received_promotions_Yes
0                         509  175
1                         519  134
chi2=4.5405, dof=1, p=0.0331 → RELATED

referred_by_friend vs is_churned (using 'referred_by_friend_Yes')
is_churned                0.0  1.0
referred_by_friend_Yes
0                         514  162
1                         514  147
chi2=0.4670, dof=1, p=0.4943 → Not related
```

Therefore, Chi-square tests show a significant association between **received promotions** and **churn** (p=0.033), while **gender** (p=0.249) and **referred by friend** (p=0.494) are not significantly related.

Moreover, a t-test was used to check if watch time differs significantly between churned and retained users. The output shows that there is no big significance between the two features. Check below.

```
T-test for average_watch_hours between churned and retained users:
t-statistic = -0.7223, p-value = 0.4705
Result: No significant difference in watch time between churned and retained users.
```

Another finding was in creating a correlation matrix. The red color shows strong positive link, the blue shows strong negative link, and the white- no link. Diagonal is always 1.0 (feature with itself). It helps spot duplicate-like features and see which variables move together. Churn (is_churned) has weak links to most features, meaning no single numeric feature strongly predicts it alone.



Lastly, charts like boxplots, bar plots, histograms were used to visualize key differences between churned and active users.

Average Watch Hours vs Churn- Churned and non-churned users have similar median watch hours. No big visual difference.

Tenure Days vs Churn- Median tenure is slightly lower for churned users, but overlap is large.

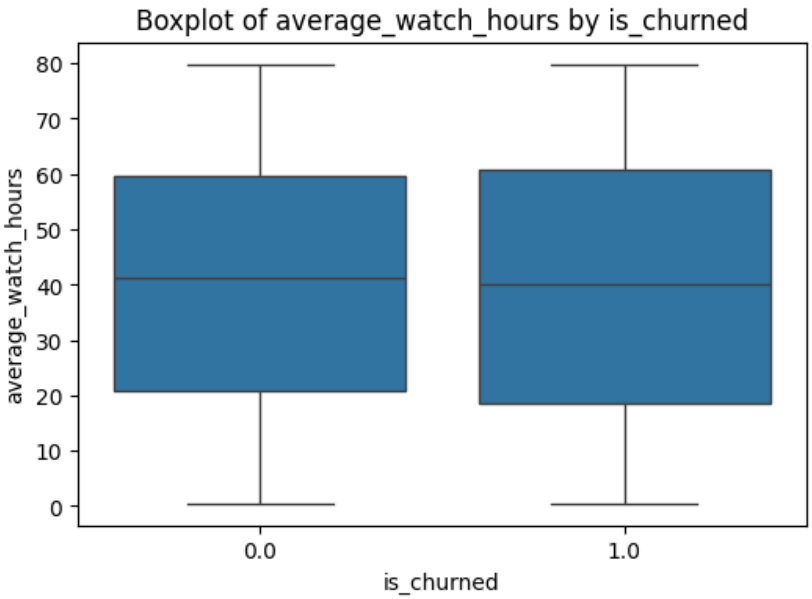
Watch per Fee Ratio vs Churn- Ratios are quite similar; churn doesn't clearly depend on it.

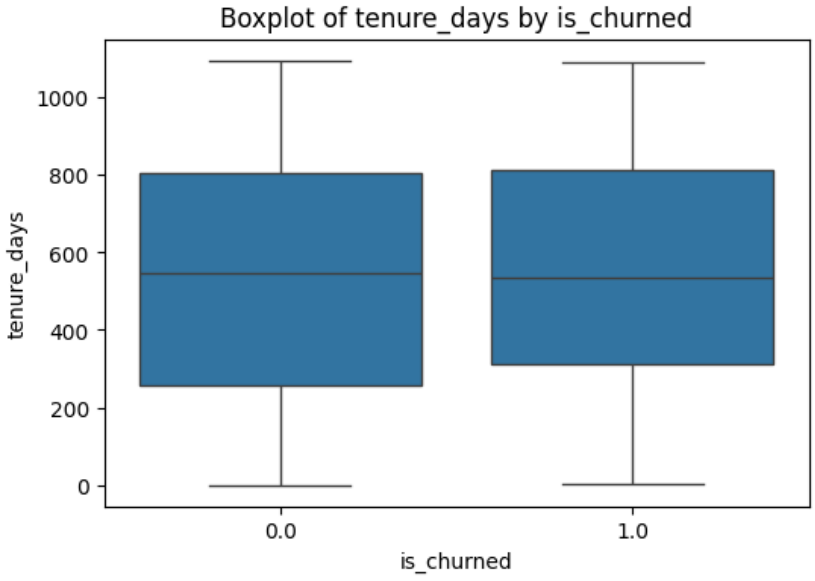
Churn Rate by Loyalty- Surprisingly, loyal users have a slightly higher churn rate than non-loyal ones.

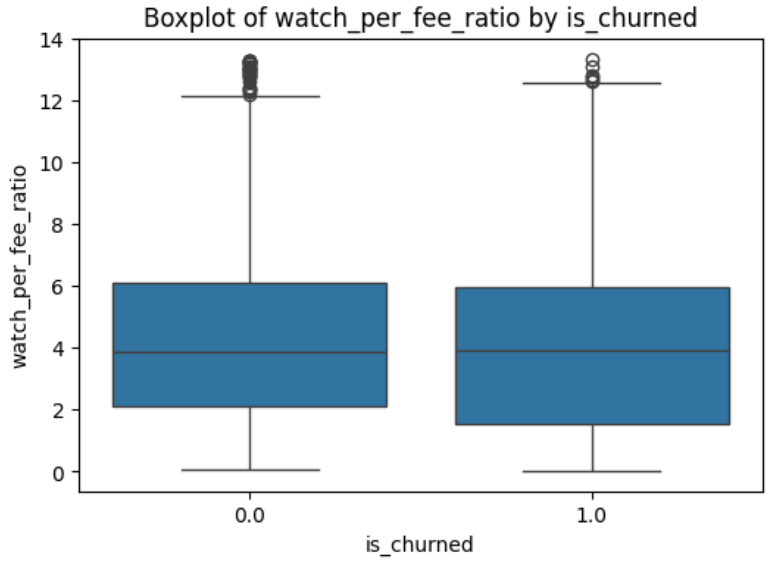
Churn Rate by Heavy Mobile Use- Only a tiny difference; heavy mobile use doesn't strongly impact churn.

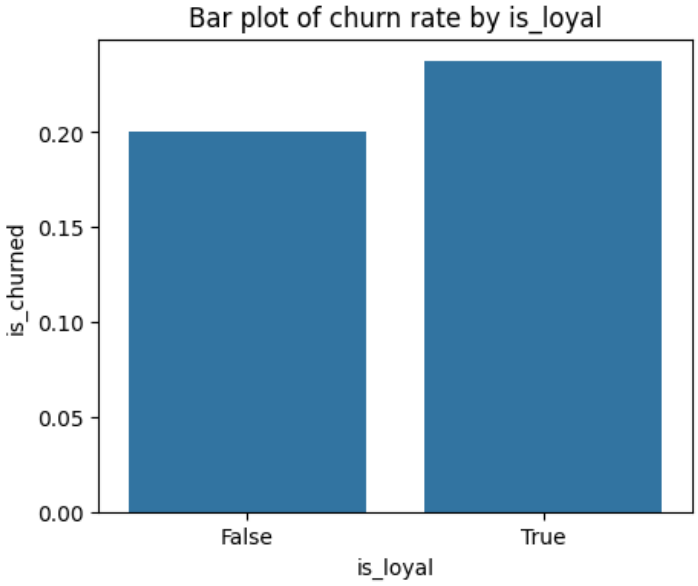
Histograms- For watch hours, tenure, and watch/fee ratio, churned vs non-churned distributions look broadly alike, meaning these features alone don't sharply separate the groups.

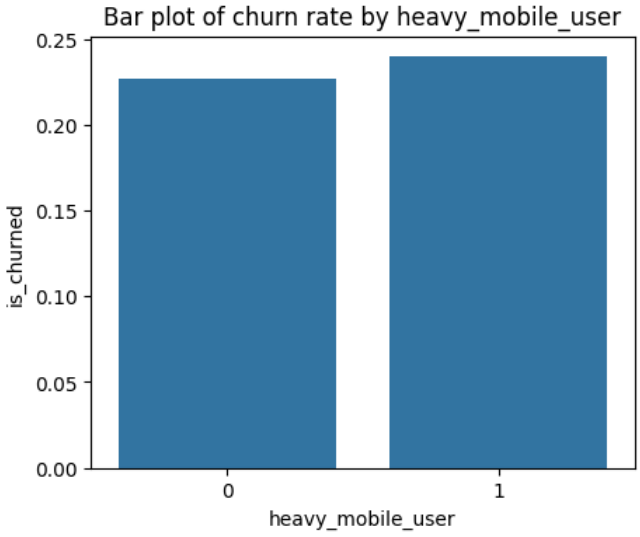
This suggests churn is not strongly driven by any single one of these metrics, it's likely a mix of multiple factors. Check below.

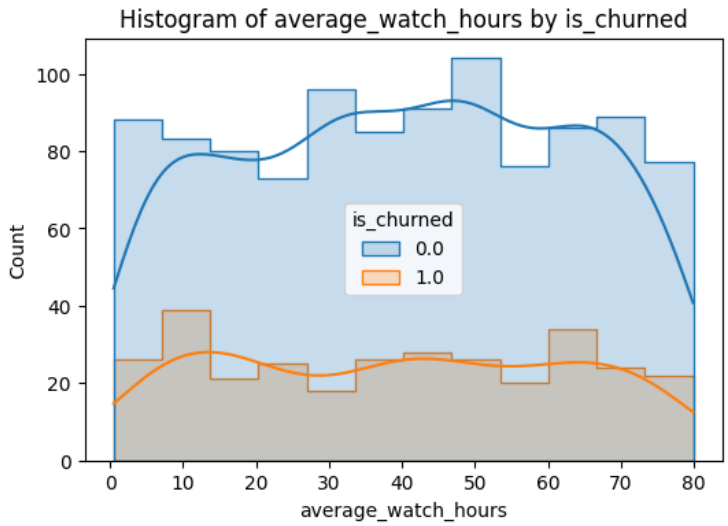


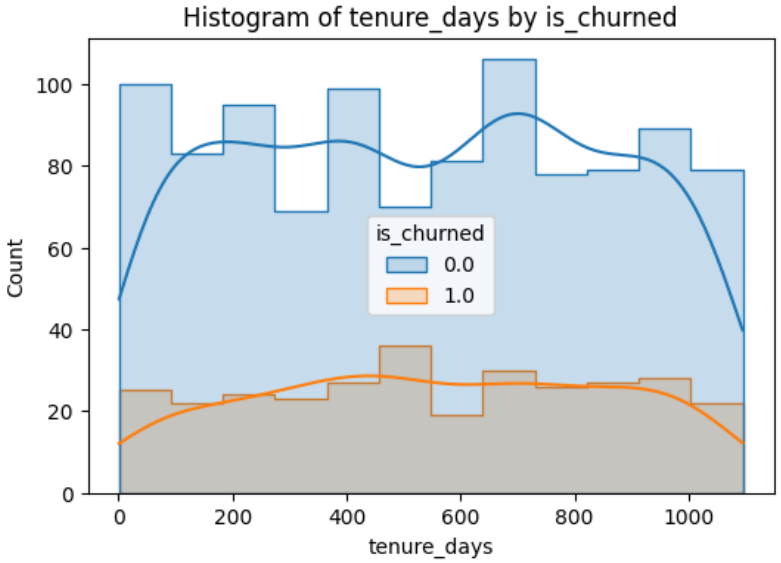


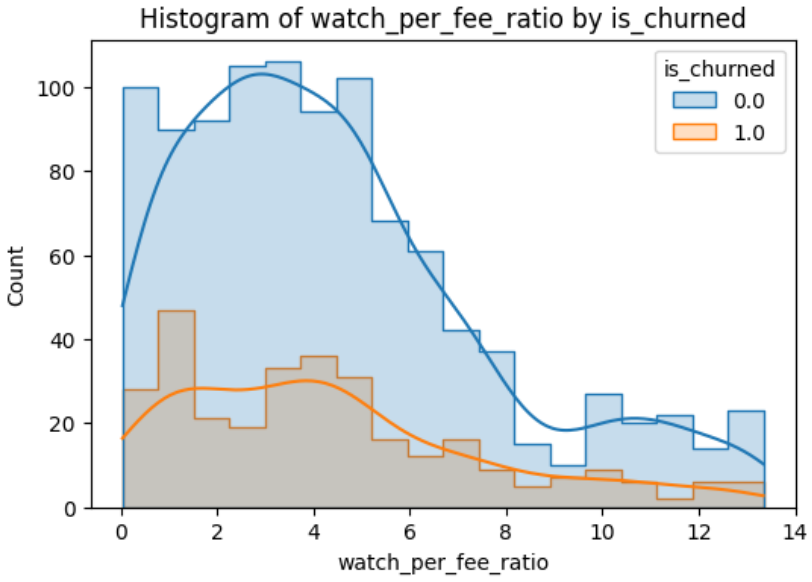

Boxplot of watch_per_fee_ratio by is_churned
watch_per_fee_ratio
14
12
10
8
6
4
2
0
0.0
1.0
is_churned


Bar plot of churn rate by is_loyal
is_churned
0.20
0.15
0.10
0.05
0.00
False
True
is_loyal


Bar plot of churn rate by heavy_mobile_user
is_churned
0.25
0.20
0.15
0.10
0.05
0.00
0
1
heavy_mobile_user


Histogram of average_watch_hours by is_churned
Count
average_watch_hours
is_churned
0.0
1.0


Histogram of tenure_days by is_churned
Count
tenure_days
is_churned
0.0
1.0


Histogram of watch_per_fee_ratio by is_churned
Count
watch_per_fee_ratio
is_churned
0.0
1.0

5. Model Results

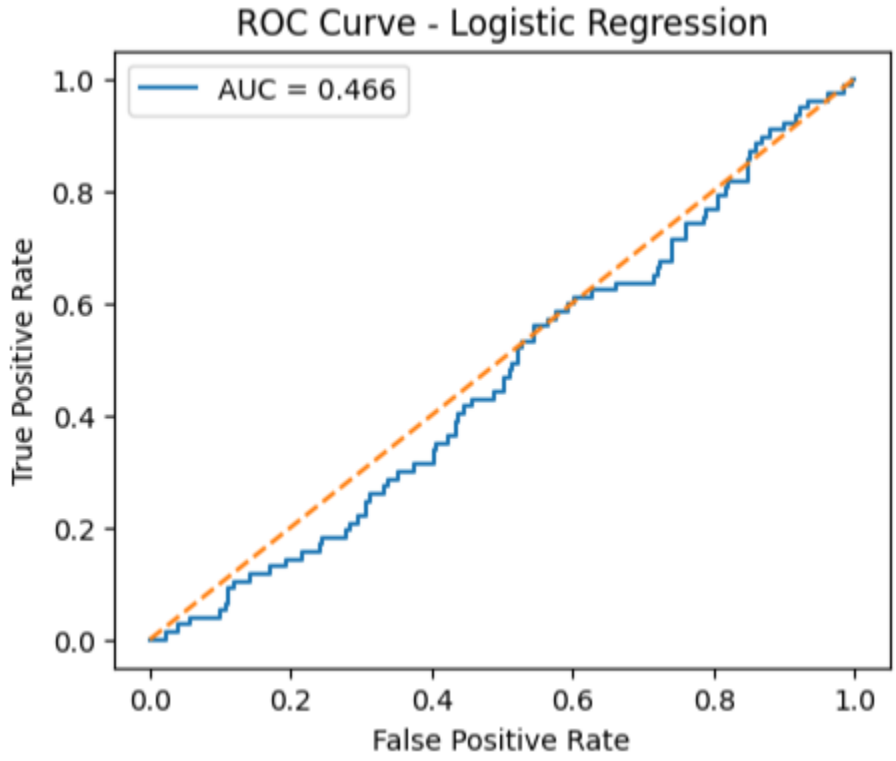
- Logistic Regression (Binary Classification)

  Accuracy of 51.6% barely above random guessing (50%).

  F1-score (positive churn class) of 0.283 is weak at identifying churned customers.

  ROC AUC of 0.466 is worse than random (0.5), meaning the model struggles to distinguish churn vs non-churn. Check below.

```
Confusion matrix (rows=true, cols=pred):
[[141 117]
 [ 45  32]]

Classification report:
              precision    recall  f1-score   support

           0      0.758     0.547     0.635       258
           1      0.215     0.416     0.283        77

    accuracy                          0.516       335
   macro avg      0.486     0.481     0.459       335
weighted avg      0.633     0.516     0.554       335

ROC AUC: 0.466
```



- List of the top 3 predictors of churn and their business interpretation

  Average Watch Hours at 0.254: Customers with higher watch hours are surprisingly more likely to churn. This could indicate binge usage before leaving or using the service heavily just before canceling.

  Log Tenure Days at 0.182: Customers with longer tenure have a slightly higher likelihood of churn, possibly due to contract completion or subscription fatigue.

  Heavy Mobile User at 0.071: Heavy mobile users are somewhat more prone to churn, which could suggest dissatisfaction with mobile experience or using mobile as a temporary engagement channel.

  Check below.

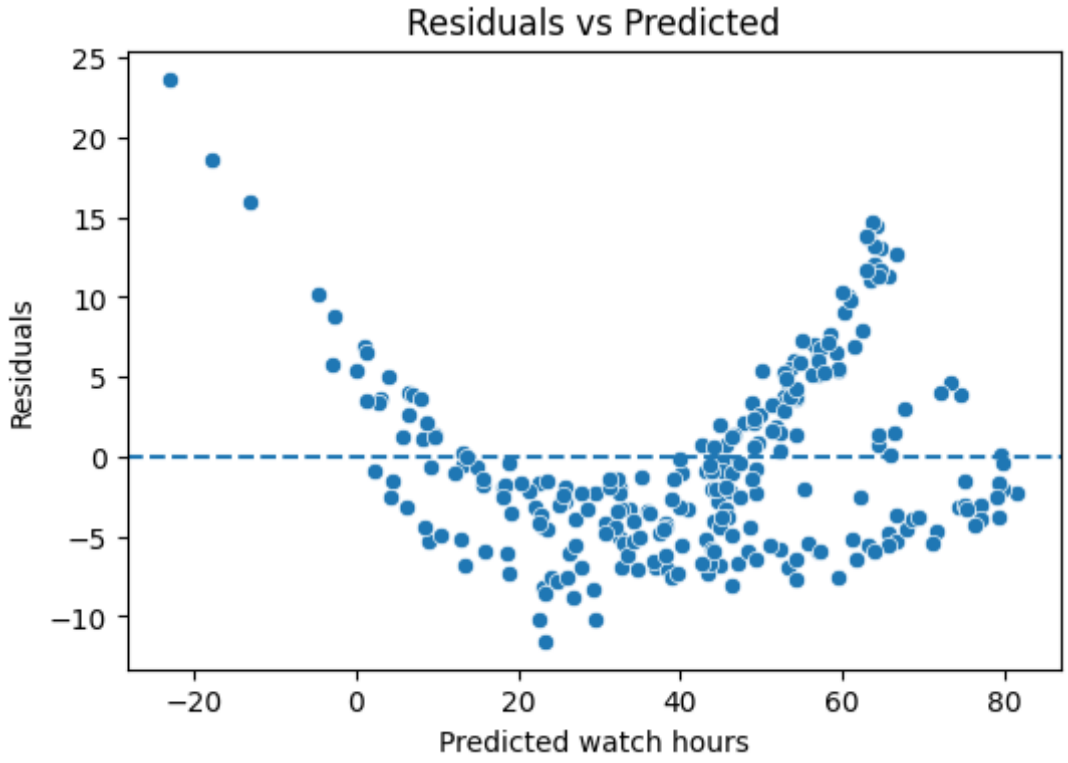
```
Top negative predictors (reduce churn probability):
log_watch_per_fee_ratio      -0.370203
monthly_fee                  -0.282134
promo_low_watch              -0.121773
tenure_days                  -0.094741
tenure_days_scaled           -0.094741
mobile_app_usage_pct         -0.079796
watch_per_fee_ratio          -0.035502
watch_per_fee_ratio_scaled   -0.035502
age                           0.026337
complaints_raised             0.050499
dtype: float64

Top positive predictors (increase churn probability):
tenure_days                  -0.094741
tenure_days_scaled           -0.094741
mobile_app_usage_pct         -0.079796
watch_per_fee_ratio          -0.035502
watch_per_fee_ratio_scaled   -0.035502
age                           0.026337
complaints_raised             0.050499
heavy_mobile_user             0.070706
log_tenure_days               0.181784
average_watch_hours           0.254350
dtype: float64
```

- Report model performance for Linear Regression (R^2 score, RMSE, MAE)

```
R²: 0.9343
RMSE: 5.7762
```

- Include residual plot (predicted vs. actual or residuals vs. fitted values)

  The residuals vs. predicted plot shows a curved pattern rather than a random scatter, suggesting that the model may not fully capture non-linear relationships in the data. Clustering of points above or below zero indicates potential bias at certain predicted watch-hour ranges.



- List of the top 3 predictors of the continuous target variable (e.g. watch time or tenure) and their Business interpretation

  log_watch_per_fee_ratio of 16.85: A higher ratio of watching time relative to the fee strongly increases watch hours, meaning customers who feel they're getting more value for their money tend to watch more.

  monthly_fee of 10.21: Higher subscription fees are linked with more watch hours, possibly because premium subscribers engage more with the service.

watch_per_fee_ratio_scaled of 4.08: Reinforces the first point; when adjusted for scale, better value still drives higher watch times.

Check below.

```
Top positive predictors (increase watch hours):
log_watch_per_fee_ratio       16.851259
monthly_fee                   10.205945
watch_per_fee_ratio_scaled     4.082158
watch_per_fee_ratio            4.082158
heavy_mobile_user              0.851487
promo_low_watch                0.396279
log_tenure_days                0.393932
complaints_raised              0.276389
is_churned                     0.258123
age                            0.030377
dtype: float64

Top negative predictors (decrease watch hours):
watch_per_fee_ratio     4.082158
heavy_mobile_user       0.851487
promo_low_watch         0.396279
log_tenure_days         0.393932
complaints_raised       0.276389
is_churned              0.258123
age                     0.030377
tenure_days_scaled     -0.232921
tenure_days            -0.232921
mobile_app_usage_pct   -0.259889
dtype: float64
```

6. Business Questions Answered

- Do users who receive promotions churn less?: **No.** Chi-square results (p=0.033) show a significant link, but the “promoted + low watch” feature revealed many promoted users still churned, meaning promotions alone did not reduce churn.

```
received_promotions vs is_churned (using 'received_promotions_Yes')
is_churned              0.0  1.0
received_promotions_Yes
0                       509  175
1                       519  134
chi2=4.5405, dof=1, p=0.0331 → RELATED

referred_by_friend vs is_churned (using 'referred_by_friend_Yes')
is_churned              0.0  1.0
referred_by_friend_Yes
0                       514  162
1                       514  147
chi2=0.4670, dof=1, p=0.4943 → Not related
```

- Does watch time impact churn likelihood?: **Yes, but in an unexpected way.** Logistic Regression shows **higher watch hours (average_watch_hours)** are linked with higher churn probability, possibly because of the binge-before-leaving behavior.

```
Top negative predictors (reduce churn probability):
log_watch_per_fee_ratio      -0.370203
monthly_fee                  -0.282134
promo_low_watch              -0.121773
tenure_days                  -0.094741
tenure_days_scaled           -0.094741
mobile_app_usage_pct         -0.079796
watch_per_fee_ratio          -0.035502
watch_per_fee_ratio_scaled   -0.035502
age                           0.026337
complaints_raised             0.050499
dtype: float64

Top positive predictors (increase churn probability):
tenure_days                  -0.094741
tenure_days_scaled           -0.094741
mobile_app_usage_pct         -0.079796
watch_per_fee_ratio          -0.035502
watch_per_fee_ratio_scaled   -0.035502
age                           0.026337
complaints_raised             0.050499
heavy_mobile_user             0.070706
log_tenure_days               0.181784
average_watch_hours           0.254350
dtype: float64
```

- Are mobile dominant users more likely to cancel?: **Slightly yes.** "Heavy mobile user" is a top churn predictor with coefficient of 0.071, suggesting mobile-first users have somewhat higher churn risk.

```
Top negative predictors (reduce churn probability):
log_watch_per_fee_ratio      -0.370203
monthly_fee                  -0.282134
promo_low_watch              -0.121773
tenure_days                  -0.094741
tenure_days_scaled           -0.094741
mobile_app_usage_pct         -0.079796
watch_per_fee_ratio          -0.035502
watch_per_fee_ratio_scaled   -0.035502
age                           0.026337
complaints_raised             0.050499
dtype: float64

Top positive predictors (increase churn probability):
tenure_days                  -0.094741
tenure_days_scaled           -0.094741
mobile_app_usage_pct         -0.079796
watch_per_fee_ratio          -0.035502
watch_per_fee_ratio_scaled   -0.035502
age                           0.026337
complaints_raised             0.050499
heavy_mobile_user             0.070706
log_tenure_days               0.181784
average_watch_hours           0.254350
dtype: float64
```

- What are the top 3 features influencing churn based on your model?:

  Average watch hours: 0.254

  Log tenure days: 0.182

  Heavy mobile user: 0.071

```
Top negative predictors (reduce churn probability):
log_watch_per_fee_ratio      -0.370203
monthly_fee                  -0.282134
promo_low_watch              -0.121773
tenure_days                  -0.094741
tenure_days_scaled           -0.094741
mobile_app_usage_pct         -0.079796
watch_per_fee_ratio          -0.035502
watch_per_fee_ratio_scaled   -0.035502
age                           0.026337
complaints_raised             0.050499
dtype: float64

Top positive predictors (increase churn probability):
tenure_days                  -0.094741
tenure_days_scaled           -0.094741
mobile_app_usage_pct         -0.079796
watch_per_fee_ratio          -0.035502
watch_per_fee_ratio_scaled   -0.035502
age                           0.026337
complaints_raised             0.050499
heavy_mobile_user             0.070706
log_tenure_days               0.181784
average_watch_hours           0.254350
dtype: float64
```

- Which customer segments should the retention team prioritise?:

  High watch-hour users: possible binge-before-leaving

  Long-tenure subscribers: subscription fatigue risk

  Heavy mobile users: possible dissatisfaction with mobile experience

```
Top negative predictors (reduce churn probability):
log_watch_per_fee_ratio      -0.370203
monthly_fee                  -0.282134
promo_low_watch              -0.121773
tenure_days                  -0.094741
tenure_days_scaled           -0.094741
mobile_app_usage_pct         -0.079796
watch_per_fee_ratio          -0.035502
watch_per_fee_ratio_scaled   -0.035502
age                           0.026337
complaints_raised             0.050499
dtype: float64

Top positive predictors (increase churn probability):
tenure_days                  -0.094741
tenure_days_scaled           -0.094741
mobile_app_usage_pct         -0.079796
watch_per_fee_ratio          -0.035502
watch_per_fee_ratio_scaled   -0.035502
age                           0.026337
complaints_raised             0.050499
heavy_mobile_user             0.070706
log_tenure_days               0.181784
average_watch_hours           0.254350
dtype: float64
```

- What factors affect user watch time or tenure?:

  log_watch_per_fee_ratio: customers getting better value watch more.

  monthly_fee: higher-paying users tend to engage more.

  watch_per_fee_ratio_scaled: value perception strongly drives watch time.

```
Top positive predictors (increase watch hours):
log_watch_per_fee_ratio       16.851259
monthly_fee                   10.205945
watch_per_fee_ratio_scaled     4.082158
watch_per_fee_ratio            4.082158
heavy_mobile_user              0.851487
promo_low_watch                0.396279
log_tenure_days                0.393932
complaints_raised              0.276389
is_churned                     0.258123
age                            0.030377
dtype: float64

Top negative predictors (decrease watch hours):
watch_per_fee_ratio     4.082158
heavy_mobile_user       0.851487
promo_low_watch         0.396279
log_tenure_days         0.393932
complaints_raised       0.276389
is_churned              0.258123
age                     0.030377
tenure_days_scaled     -0.232921
tenure_days            -0.232921
mobile_app_usage_pct   -0.259889
dtype: float64
```

7. Recommendations

- **Target high watch-hour churn risks**: Implement early-warning triggers for heavy watchers showing signs of disengagement to address ‘binge-before-leave’ behavior.
- **Enhance mobile experience**: Since heavy mobile usage correlates with higher churn, invest in UI/UX upgrades or exclusive mobile content to boost retention.
- **Refine promotion targeting**: Promotions should be coupled with engagement campaigns for low-watch users rather than sent broadly, as current promotions alone do not reduce churn.

8. Data Issues or Risks

- **Class imbalance**: Churn cases are fewer than non-churn, which may bias models toward predicting non-churn; mitigation via SMOTE or class weights is needed.

- **Potential feature leakage**: Some engineered variables such as tenure and watch hours may be influenced by churn behavior itself, which can inflate model performance if not carefully time-bound.
- **Date inconsistencies**: Missing or irregular signup/last active dates limited deeper time-series analysis.
- **Causality vs. Correlation**: Predictors indicate association with churn but do not confirm they cause it; business interventions should be tested experimentally.