Streaming Service Data Exploratory

Toan Nguyen

06 July 2025

Contents

Install packages

```
install.packages("tidyverse")
install.packages("lubridate")
install.packages("ggplot2")
install.packages("scales")
install.packages("skimr")
install.packages("ggthemes")
install.packages("mlr3")
install.packages("mlr3filters")
install.packages("mlr3learners")
install.packages("FSelectorRcpp")
install.packages("mlr3viz")
install.packages("smotefamily")
install.packages("sf")
install.packages("rnaturalearth")
install.packages("rnaturalearthdata")
install.packages("dplyr")
```

Load packages

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(lubridate)
library(ggplot2)
library(scales)
```

```
##
## Attaching package: 'scales'
##
## The following object is masked from 'package:purrr':
##
##     discard
##
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
library(skimr)
library(ggthemes)
library(mlr3)
```

```
##
## Attaching package: 'mlr3'
##
## The following object is masked from 'package:skimr':
##
##     partition
```

```
library(mlr3filters)
library(mlr3learners)
library(FSelectorRcpp)
library(mlr3viz)
library(smotefamily)
library(sf)
```

```
## Linking to GEOS 3.13.0, GDAL 3.10.1, PROJ 9.5.1; sf_use_s2() is TRUE
```

```
library(rnaturalearth)
library(dplyr)
```

1. Introduction

This report explores a customer subscription dataset from StreamWorld, a fictional streaming service. The aim is to uncover patterns related to customer churn and provide insights to support both strategic business decisions and the development of predictive models.

Access to real customer data is often limited due to privacy concerns and confidentiality constraints. To address this, a synthetic dataset was generated using ChatGPT, simulating genuine user behaviors and business conditions. This approach allows for meaningful analysis while avoiding the ethical and legal issues associated with using real-world proprietary data.

The dataset is designed to mimic real-world scenarios but contains some discrepancies and lacks business logic in calculations. We will make reasonable assumptions during analysis to provide practical business insights.

Customer churn was chosen as the central theme of this report because it poses a significant challenge for subscription-based businesses. Retaining existing customers is generally more cost-effective than acquiring new ones, and understanding the drivers of churn is essential for maintaining sustainable growth, maximizing customer lifetime value, and improving profitability.

The analysis is guided by a key research question: What drives a customer to cancel? By addressing this question, the report aims to deliver actionable insights that can inform retention strategies and support the creation of accurate, data-driven churn prediction models.

2. Data Loading & Overview

```
streaming_data <- read_csv("streaming_service_cleaned.csv")
```

```
## Rows: 2375 Columns: 19
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (10): Name, Gender, SubscriptionType, PaymentMethod, Email, Country, Re...
## dbl   (6): CustomerID, Age, MonthlyFee, TotalWatchTime, NumSupportTickets, L...
## lgl   (1): Cancelled
## dttm  (2): JoinDate, LastLoginDate
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
glimpse(streaming_data)
```

```
## Rows: 2,375
## Columns: 19
## $ CustomerID       <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 1~
## $ Name             <chr> "David Jackson", "Emily Martin", "Alex Taylor", "Emi~
## $ Age              <dbl> 19, 53, 75, 58, 27, 58, 36, 22, 23, 70, 19, 41, 80, ~
## $ Gender           <chr> "Male", "Non-binary", "Male", "Female", "Male", "Mal~
## $ JoinDate         <dttm> 2022-06-29, 2021-12-15, 2023-09-11, 2019-06-15, 202~
## $ SubscriptionType <chr> "Standard", "Standard", "Premium", "Basic", "Basic",~
## $ MonthlyFee       <dbl> 16.38, 15.90, 18.67, 10.67, 8.98, 9.72, 15.46, 19.36~
## $ PaymentMethod    <chr> "Paypal", "Credit Card", "Paypal", "Credit Card", "C~
## $ TotalWatchTime   <dbl> 45.55, 7.85, 42.49, 9.29, 14.94, 38.42, 0.36, 7.82, ~
## $ Cancelled        <lgl> FALSE, TRUE, TRUE, FALSE, FALSE, FALSE, TRUE, TRUE, ~
```

```
## $ Email             <chr> "djackson1@yahoo.com", "emartin2@streamworld.com", "~
## $ Country           <chr> "France", "Australia", "USA", "Germany", "Australia"~
## $ Region            <chr> "Region7", "Region10", "Region10", "Central", "East"~
## $ LastLoginDate     <dttm> 2023-08-02, 2025-03-25, 2025-06-26, 2025-06-26, 202~
## $ NumSupportTickets <dbl> 0, 0, 0, 3, 2, 1, 3, 3, 1, 1, 1, 2, 1, 0, 0, 5, 0, 3~
## $ LoyaltyPoints     <dbl> 715, 272, 5, 940, 320, 865, 735, 452, 163, 551, 994,~
## $ ReferralSource    <chr> "Ad", "Website", "SocialMedia", "Friend", "Ad", "Web~
## $ PreferredDevice   <chr> "Mobile", "Mobile", "Mobile", "Mobile", "Tablet", "M~
## $ PaymentFrequency  <chr> "Monthly", "Annual", "Monthly", "Monthly", "Monthly"~
```

```
skim(streaming_data)
```

Table 1: Data summary

Name	streaming_data
Number of rows	2375
Number of columns	19
_____	
Column type frequency:	
character	10
logical	1
numeric	6
POSIXct	2
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Name	0	1	10	16	0	100	0
Gender	0	1	4	17	0	4	0
SubscriptionType	0	1	5	8	0	3	0
PaymentMethod	0	1	6	13	0	4	0
Email	0	1	12	29	0	2306	0
Country	0	1	2	9	0	6	0
Region	0	1	4	8	0	10	0
ReferralSource	0	1	2	13	0	5	0
PreferredDevice	0	1	6	7	0	3	0
PaymentFrequency	0	1	6	9	0	3	0

Variable type: logical

skim_variable	n_missing	complete_rate	mean	count
Cancelled	0	1	0.29	FAL: 1692, TRU: 683

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
CustomerID	0	1	1251.00	723.03	1.00	625.50	1247.00	1877.50	2500.00	□□□□□
Age	0	1	49.33	18.04	18.00	34.00	49.00	64.00	80.00	□□□□□
MonthlyFee	0	1	13.90	3.84	7.08	10.27	14.19	15.98	23.23	□□□□□
TotalWatchTime	0	1	20.92	21.73	0.00	5.70	13.89	28.86	201.44	□□□□□
NumSupportTickets	0	1	1.69	1.41	0.00	1.00	1.00	2.00	6.00	□□□□□
LoyaltyPoints	0	1	501.84	294.41	0.00	239.50	507.00	758.00	1000.00	□□□□□

Variable type: POSIXct

skim_variable	n_missing	complete_rate	min	max	median	n_unique
JoinDate	0	1	2018-01-01	2025-06-24	2021-11-15	1575
LastLoginDate	0	1	2018-03-28	2025-06-26	2025-06-26	910

The dataset, titled "streaming_data," comprises 2375 rows and 19 columns, providing a comprehensive overview of customer information for a streaming service. It includes 10 character-type variables (e.g., Name, Gender, SubscriptionType), 6 numeric variables (e.g., Age, MonthlyFee, TotalWatchTime), 2 POSIXct variables (JoinDate, LastLoginDate) and 1 response variable "Cancelled", with no missing values identified. A detailed skim of the data reveals complete records across all variables, with numeric columns exhibiting varied ranges, such as TotalWatchTime (0 to 201.44 hours) and LoyaltyPoints (0 to 10000), indicating diverse customer engagement and loyalty metrics as of June 29, 2025.

3. Feature Engineering

Datasets often include a large number of features, not all of which contribute meaningfully to the prediction task. Including irrelevant or redundant features can reduce model performance, increase computational cost, and make the results harder to interpret.

In this project, our goal is to predict customer churn, so we first perform feature engineering to create useful new variables, and then we focus on feature selection to identify the most relevant ones for the churn outcome.

Create new variables which are not in the dataset

Based on the dataset, certain variables can be derived from the available data.

Tenure

Description: This represents the number of days a customer has been with the service. It is calculated as the time between the date they joined (JoinDate) and their most recent login (LastLoginDate).

```
streaming_data <- streaming_data %>%
  mutate(TenureDays = as.numeric(difftime(LastLoginDate, JoinDate, units = "days")),
         TenureMonths = round(TenureDays / 30, 1))
```

Revenue Earned for each customer

Description: Revenue based on MonthlyFee and PaymentFrequency.

```r
streaming_data <- streaming_data %>%
  mutate(
    Revenue = case_when(
      PaymentFrequency == "Monthly" ~ MonthlyFee,
      PaymentFrequency == "Quarterly" ~ MonthlyFee * 3,
      PaymentFrequency == "Annual" ~ MonthlyFee * 12,
      TRUE ~ NA_real_
    )
  )
```

Rank features and select the most relevant variables

We will apply a filter method using the mlr3 library in R. Specifically, we use information gain to rank features based on their importance, helping us select the most relevant ones for our classification task. Since there are 1692 "not cancelled" and 683 "cancelled" records, the dataset has a class imbalance of about 71% vs 29%, which is noticeable but not extreme. This level can still affect filter-based feature selection methods like information_gain, especially if the features don't strongly differentiate the classes. We will try applying a class balancing technique like SMOTE. before using information gain filter.

```r
# Create a copy of the original dataset for transformations
transformed_data <- streaming_data

# Transform date fields
transformed_data$JoinDate <- NULL
transformed_data$LastLoginDate <- NULL

# Drop identifier columns
excluded_cols <- c("Email", "Name", "CustomerID")
transformed_data <- transformed_data[, !(names(transformed_data) %in% excluded_cols)]

# Label encode categorical variables
categorical_vars <- c(
  "Region",
  "Country",
  "PreferredDevice",
  "Gender",
  "ReferralSource",
  "SubscriptionType",
  "PaymentFrequency",
  "PaymentMethod"
)

for (col in categorical_vars) {
  if (col %in% names(transformed_data)) {
    transformed_data[[col]] <- as.integer(as.factor(transformed_data[[col]]))
  }
}

# Convert target to numeric (0 = not cancelled, 1 = cancelled)
transformed_data$Cancelled <- as.numeric(as.factor(transformed_data$Cancelled)) - 1

# Separate features and target
X_transformed <- transformed_data[, setdiff(names(transformed_data), "Cancelled")]
```

```r
Y_transformed <- transformed_data$Cancelled

# Calculate how many duplicates are needed to balance the dataset
minority_size <- sum(Y_transformed == 1)
majority_size <- sum(Y_transformed == 0)
dup_size <- floor((majority_size - minority_size) / minority_size)

# Apply SMOTE
smote_result <- SMOTE(X_transformed, Y_transformed, K = 5, dup_size = dup_size)

# Recombine data
balanced_data <- smote_result$data
balanced_data$class <- as.factor(balanced_data$class)
colnames(balanced_data)[ncol(balanced_data)] <- "Cancelled"

# Create classification task
streaming_task <- as_task_classif(balanced_data, target = "Cancelled")

# Apply Information Gain filter
filter_importance <- flt("information_gain")
streaming_feature_importance <- filter_importance$calculate(streaming_task)

# Display results
importance_table <- as.data.table(streaming_feature_importance)
print(importance_table[order(-score)])
```

```
##               feature       score
##                <char>       <num>
##  1:           Country 0.174642496
##  2:    ReferralSource 0.163806361
##  3:  SubscriptionType 0.130263823
##  4:            Gender 0.114970335
##  5:     PaymentMethod 0.109383131
##  6: NumSupportTickets 0.096253787
##  7:   PreferredDevice 0.079732075
##  8:  PaymentFrequency 0.058307716
##  9:            Region 0.035809914
## 10:               Age 0.004365718
## 11:      TenureMonths 0.000000000
## 12:        MonthlyFee 0.000000000
## 13:    TotalWatchTime 0.000000000
## 14:        TenureDays 0.000000000
## 15:           Revenue 0.000000000
## 16:     LoyaltyPoints 0.000000000
```

```r
# Plot feature importance
autoplot(streaming_feature_importance)
```

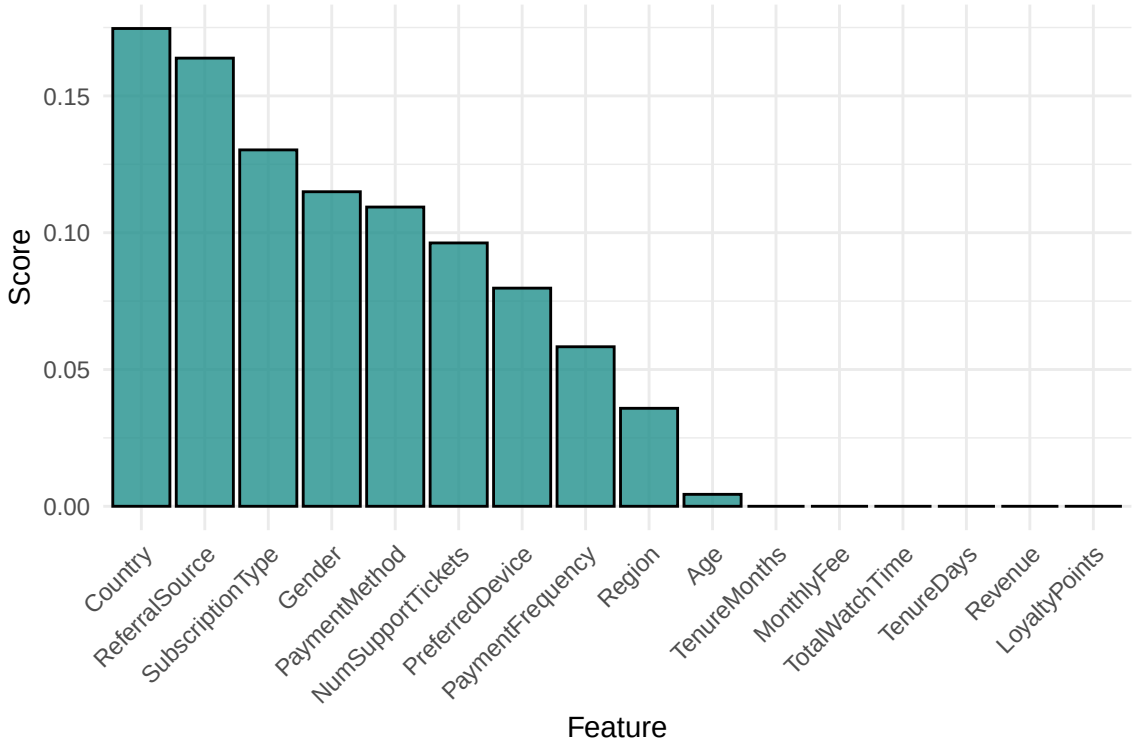

Figure 1: Streaming Feature Importance

Based on the feature importance scores in the provided chart, the features with a score of zero such as MonthlyFee, TenureMonths, TenureDays, LoyaltyPoints, Revenue and TotalWatchTime should be excluded from the analysis. The remaining features, including Region, Country, PreferredDevice, Gender, SubscriptionType, PaymentFrequency, PaymentMethod, Age, NumSupportTickets and ReferalSource, exhibit non-zero scores and should be retained for further investigation.

4. What Drives a Customer to Cancel?

We will now explore how the features influence the likelihood of churn.

4.1 Churn by Referal Source

```
# Ensure Cancelled is a factor
streaming_data$Cancelled <- as.factor(streaming_data$Cancelled)

# Now plotting the churn rate by Subscription Type
ggplot(streaming_data, aes(x = ReferralSource, fill = Cancelled)) +
  geom_bar(position = "fill") +
  labs(title = "Churn Rate by Referral Source", y = "Proportion of Customers") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```

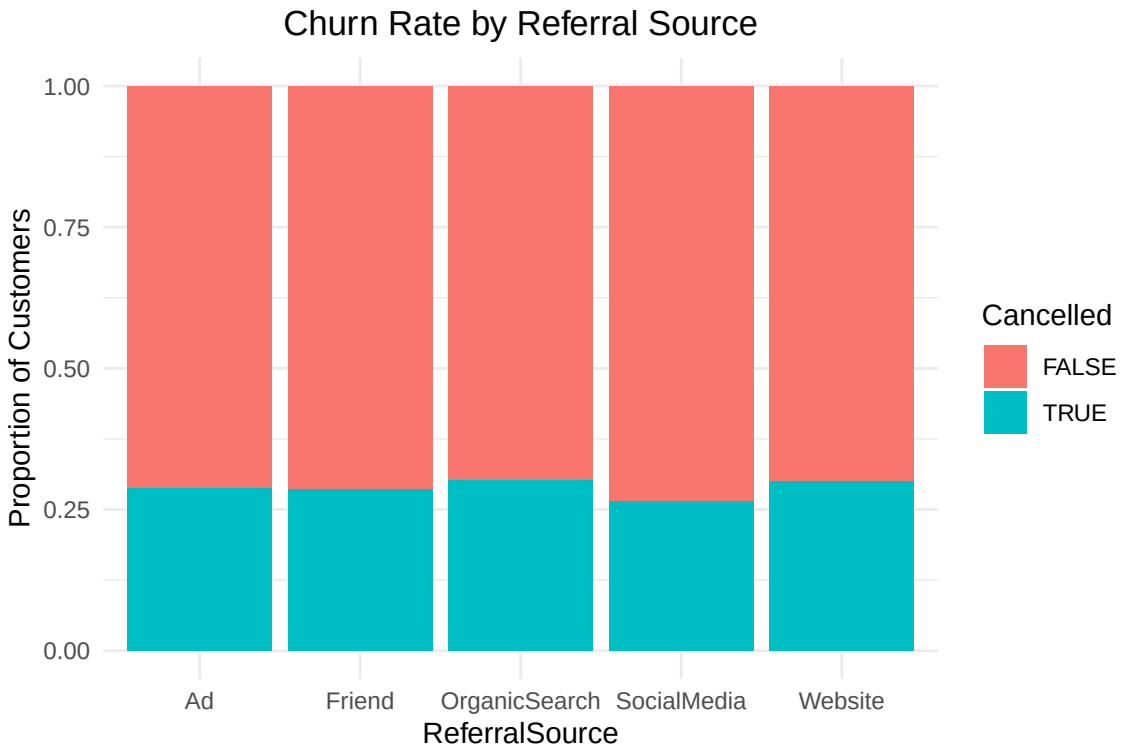

Figure 2: Churn Rate by Referral Source

Customers acquired through "OrganicSearch" and "Website" exhibit the highest cancellation rates (approximately 0.30), while those from "Friend" and Ad show a moderate rate (around 0.27), and "SocialMedia" display the lowest rates (about 0.25). This suggests that organic search traffic and direct website visits may contribute to higher cancellations, potentially due to unmet expectations, whereas friend referrals, social media, and paid advertisements appear to enhance retention.

4.2 Churn by Preferred Device

```
# Now plotting the churn rate by Subscription Type
ggplot(streaming_data, aes(x = PreferredDevice, fill = Cancelled)) +
  geom_bar(position = "fill") +
  labs(title = "Churn Rate by Preferred Device", y = "Proportion of Customers") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```

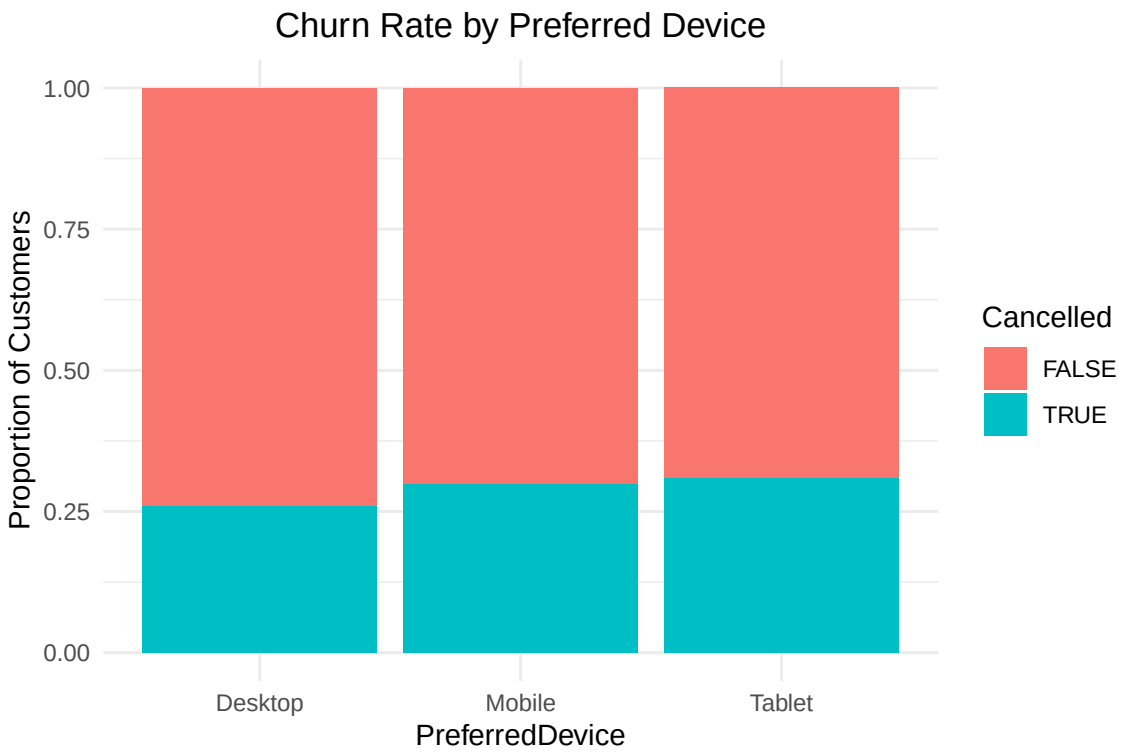

Figure 3: Churn Rate by Preferred Device

Customers utilizing “Tablet” and “Mobile” demonstrate the highest cancellation rates, with approximately 0.30 of customers in each group having terminated their subscriptions. Conversely, customers preferring “Desktop” exhibit a significantly lower cancellation rate, approximately 0.25. This indicates that a preference for tablet or mobile devices may be associated with higher churn rates, potentially due to unfriendly UI/UX designs compared to the desktop version, which may lead to user dissatisfaction or technical difficulties. In contrast, the desktop version appears to facilitate greater retention levels, likely owing to a more user-friendly interface and experience.

4.3 Churn by Subscription Type

```
# Now plotting the churn rate by Subscription Type
ggplot(streaming_data, aes(x = SubscriptionType, fill = Cancelled)) +
  geom_bar(position = "fill") +
  labs(title = "Churn Rate by Subscription Type", y = "Proportion of Customers") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```

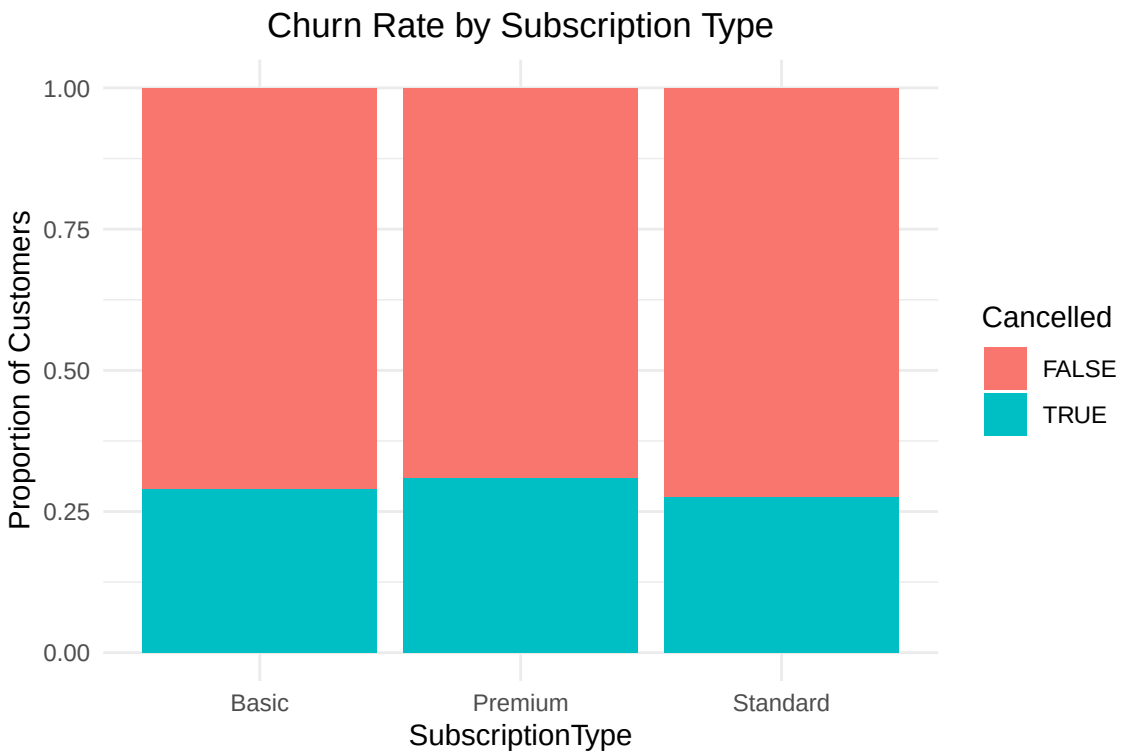

Figure 4: Churn Rate by Subscription Type

Customers subscribed to the "Premium" plans exhibit the highest cancellation rates, with approximately 0.30 of customers in each category having canceled. In contrast, customers with the "Standard" and "Basic" plans show a lower cancellation rate, approximately 0.25. This implies that "Premium" subscription types might be linked to higher churn, possibly due to perceived value or dissatisfaction with the service, while the "Standard" and "Basic" plans seem to encourage better retention due to their affordable service cost.

4.4 Churn by Payment Frequency

```
ggplot(streaming_data, aes(x = PaymentFrequency, fill = Cancelled)) +
  geom_bar(position = "fill") +
  labs(title = "Churn by Payment Frequency", y = "Proportion of Customers") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```

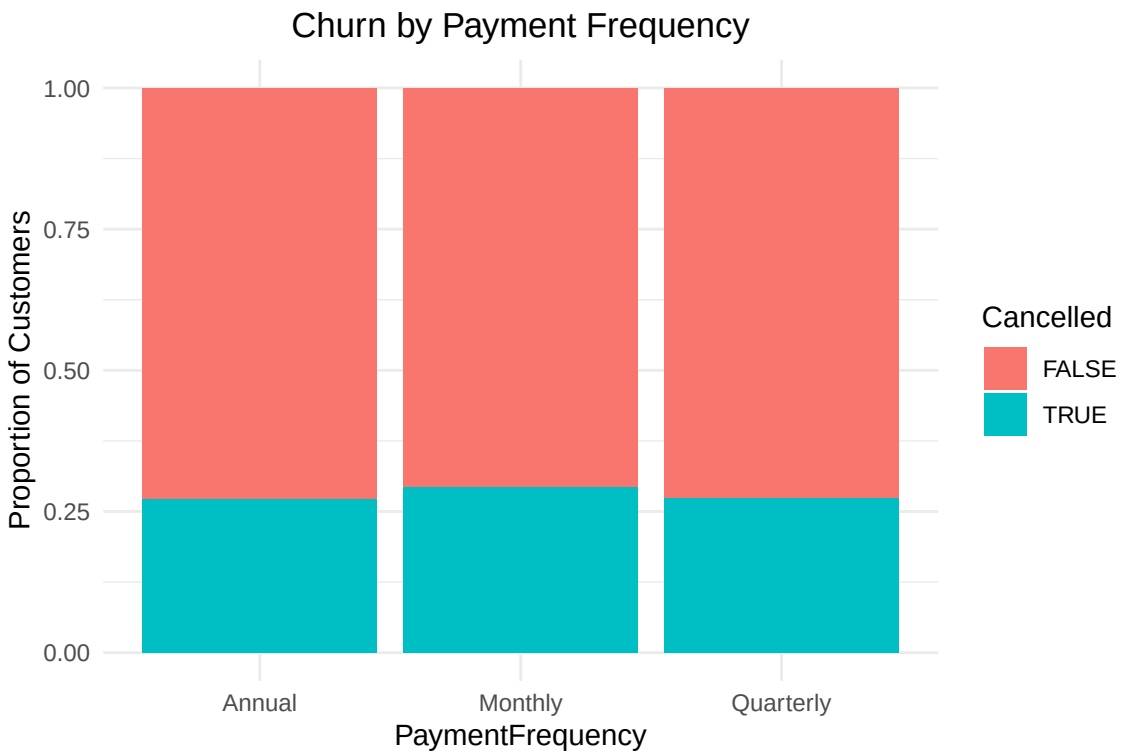

Figure 5: Churn Rate by Payment Frequency

Customers with an "Monthly" payment frequency exhibit the highest cancellation rate, with approximately 0.30 having canceled. In contrast, customers with a "Quarterly" or "Annual" payment frequency show a lower cancellation rate, approximately 0.25. This suggests that monthly payment schedules may be associated with higher churn, potentially due to financial commitment concerns or dissatisfaction, whereas a quarterly payment frequency appears to support greater retention due to its reasonable payment fee.

4.5 Churn by Payment Method

```
ggplot(streaming_data, aes(x = PaymentMethod, fill = Cancelled)) +
  geom_bar(position = "fill") +
  labs(title = "Churn by Payment Method", y = "Proportion of Customers") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```

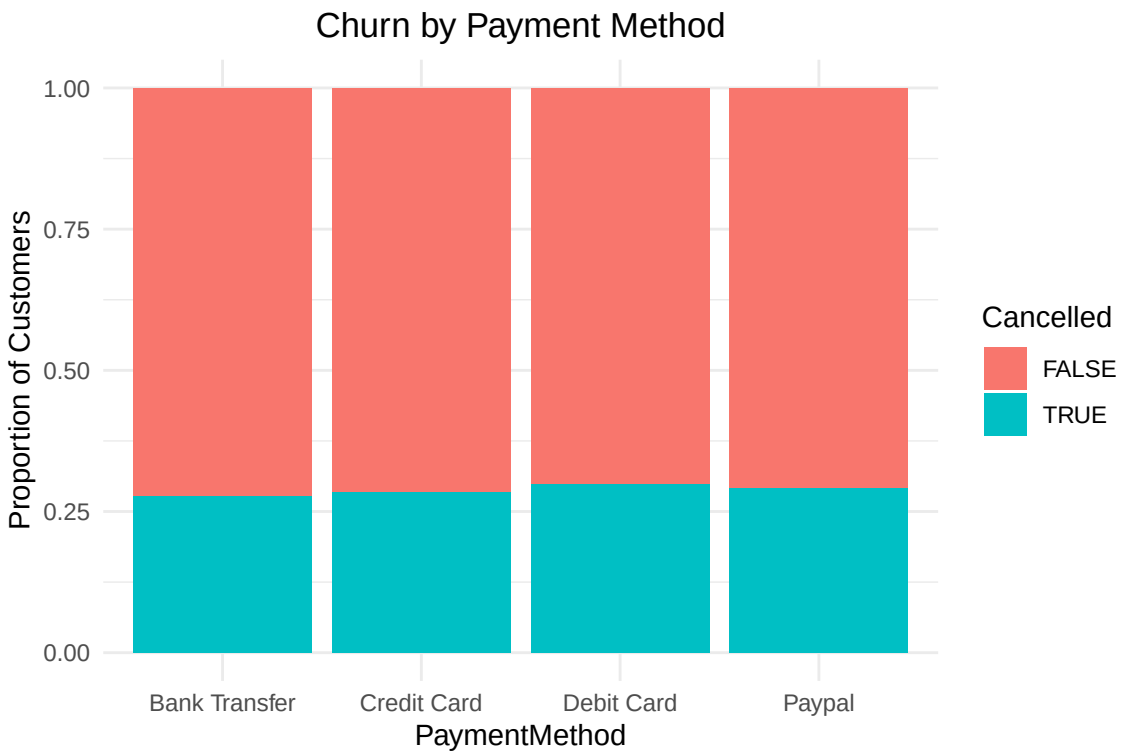

Figure 6: Churn Rate by Payment Method

Customers using "Debit Card" have the highest cancellation rates, with around 0.30 of customers in each group canceling. While "Paypal", "Bank Transfer" and "Credit Card" customers have a lower cancellation rate, around 0.25. This suggests that payment methods like debit cards might be associated with higher churn because of factors like payment failure risks, perceived security, and transaction complexities. On the other hand, flexibility, security features, and customer protection mechanisms pf "Bank Transfer", "Credit Card" and "Paypal" could contribute to a more positive user experience, encouraging customers to remain subscribed.

4.6 Churn Rate by Age Group

```
ggplot(streaming_data, aes(x = Cancelled, y = Age, fill = Cancelled)) +
  geom_boxplot() +
  labs(title = "Age Distribution by Churn Status",
       x = "Churn Status", y = "Age") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```

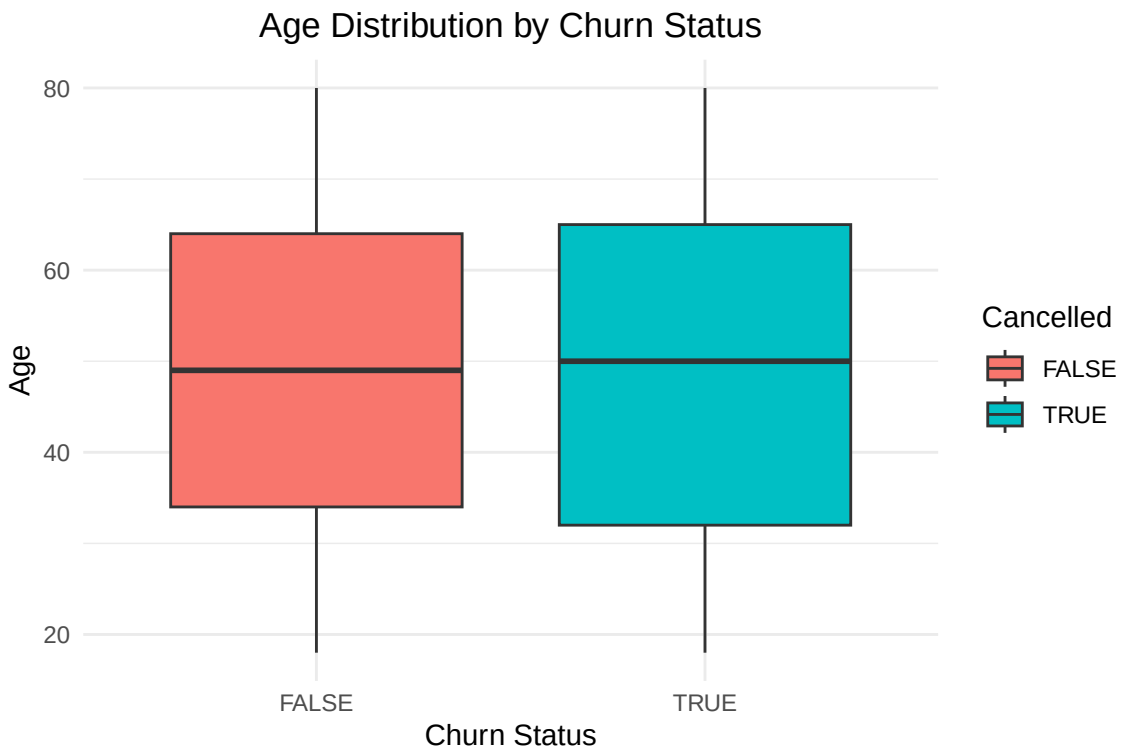

Figure 7: Age Distribution by Churn Status

For customers who have not canceled (FALSE), the age range is approximately 30 to 66, with a median age around 50. For customers who have canceled (TRUE), the age range is similarly approximately 35 to 65, with a median age also around 50. This suggests that age distribution does not significantly differ between retained and canceled customers, indicating that age may not be a primary factor influencing churn in this dataset.

4.7 Churn by Country

```
# Calculate churn rate per country
country_churn_data <- streaming_data %>%
  group_by(Country) %>%
  summarise(churn_rate = mean(as.numeric(Cancelled)))
```

```
world_map <- ne_download(scale = 110, type = "countries", category = "cultural",
 ↪  returnclass = "sf")
```

```
# Merge country churn data with world map
map_data <- world_map %>%
  left_join(country_churn_data, by = c("ADMIN" = "Country"))
```

```
ggplot(map_data) +
  geom_sf(aes(fill = churn_rate), color = "white") +  # Fill countries by churn_rate
  scale_fill_viridis_c(option = "plasma") +           # Color scale for churn rate
  labs(title = "Churn Rate by Country", fill = "Churn Rate") +
```

```
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))        # Center title
```

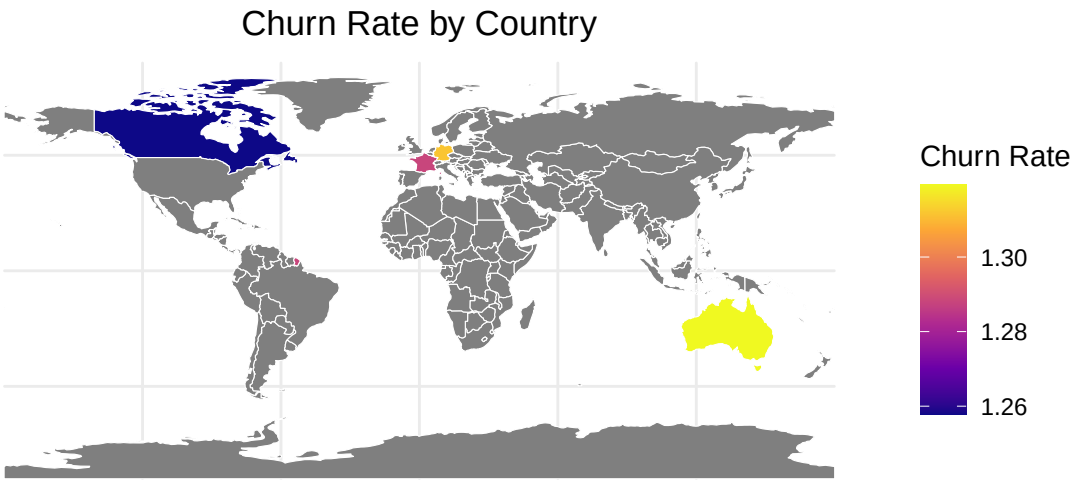

Figure 8: Churn Rate by Country

The chart "Churn Rate by Country" shows how often customers cancel their subscriptions in different countries, with values between 1.26 and 1.30. Darker colors, like those over Canada and parts of Western Europe, represent lower churn rates (around 1.26), while bright yellow over Australia shows the highest churn rate (around 1.30).

This difference might be caused by several reasons. In Australia, more people might cancel because there are lots of other options, prices might be high, or the service may not meet their needs. Australia's distance from other countries and reliance on internet services could also make it harder to deliver good service.

On the other hand, Canada and Western Europe may have fewer cancellations because customers are more loyal, get better support, or the service fits their needs better. These regions might also benefit from companies having more experience and better marketing in those areas.

4.8 Churn by Region

```
ggplot(streaming_data, aes(x = Region, fill = Cancelled)) +
  geom_bar(position = "fill") +
  labs(title = "Proportion of Churn by Region",
       y = "Proportion of Customers") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 45, hjust = 1))
```

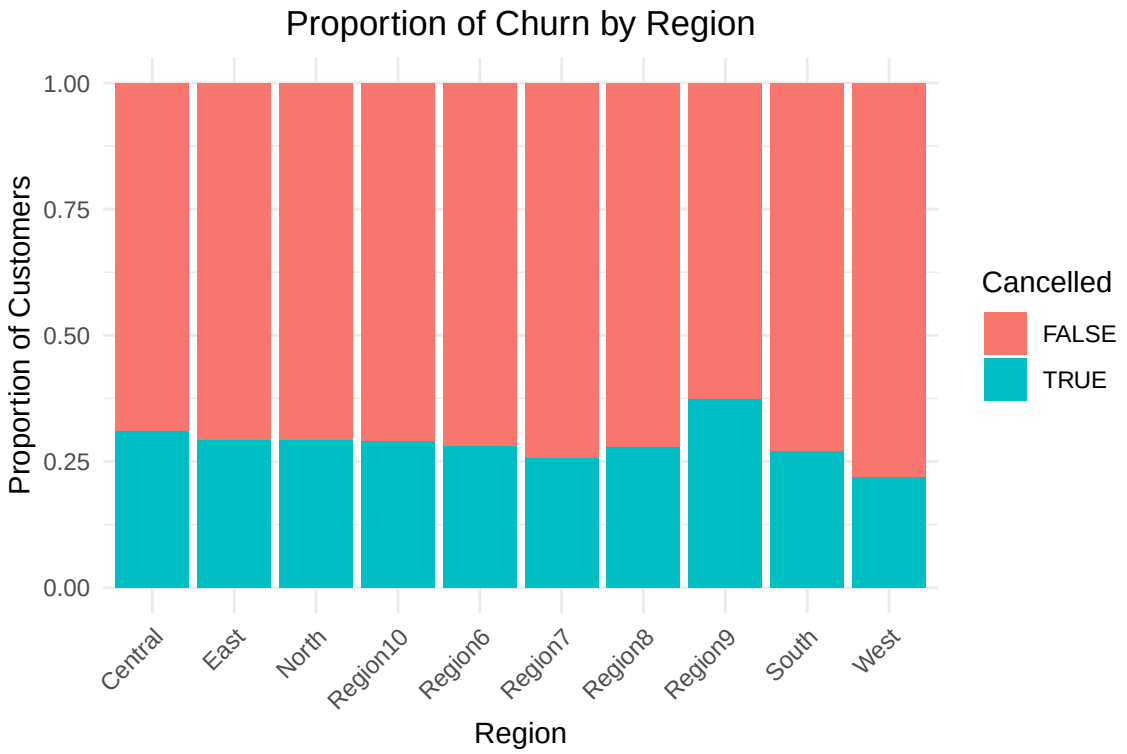

Figure 9: Churn Rate by Region

"Region9" exhibit the highest cancellation rates, with approximately 0.75 of customers having canceled. In contrast, remaining regions show lower cancellation rates, with approximately 0.25 to 0.30 of customers having canceled. This indicates that regions with higher churn like "Region9" may face challenges such as inadequate customer support or market saturation, whereas other regions may benefit from enhanced service delivery or greater customer satisfaction, contributing to improved retention.

4.9 Churn by Gender

```
# Separate datasets for churned and non-churned customers
gender_churn_true <- streaming_data %>%
  filter(Cancelled == TRUE) %>%
  count(Gender) %>%
  mutate(Prop = n / sum(n),
         PercentLabel = paste0(Gender, " (", round(Prop * 100, 1), "%)"))

gender_churn_false <- streaming_data %>%
  filter(Cancelled == FALSE) %>%
  count(Gender) %>%
  mutate(Prop = n / sum(n),
         PercentLabel = paste0(Gender, " (", round(Prop * 100, 1), "%)"))

# Create doughnut charts for churned and non-churned customers

ggplot(gender_churn_true, aes(x = 2, y = Prop, fill = PercentLabel)) +
  geom_bar(stat = "identity",
```

```
          width = 1,
          color = "white") +
  coord_polar(theta = "y") +
  xlim(0.5, 2.5) +
  theme_void() +
  labs(title = "Gender Distribution (Churned Customers)", fill = "Gender") +
  theme(plot.title = element_text(hjust = 0.5))
```

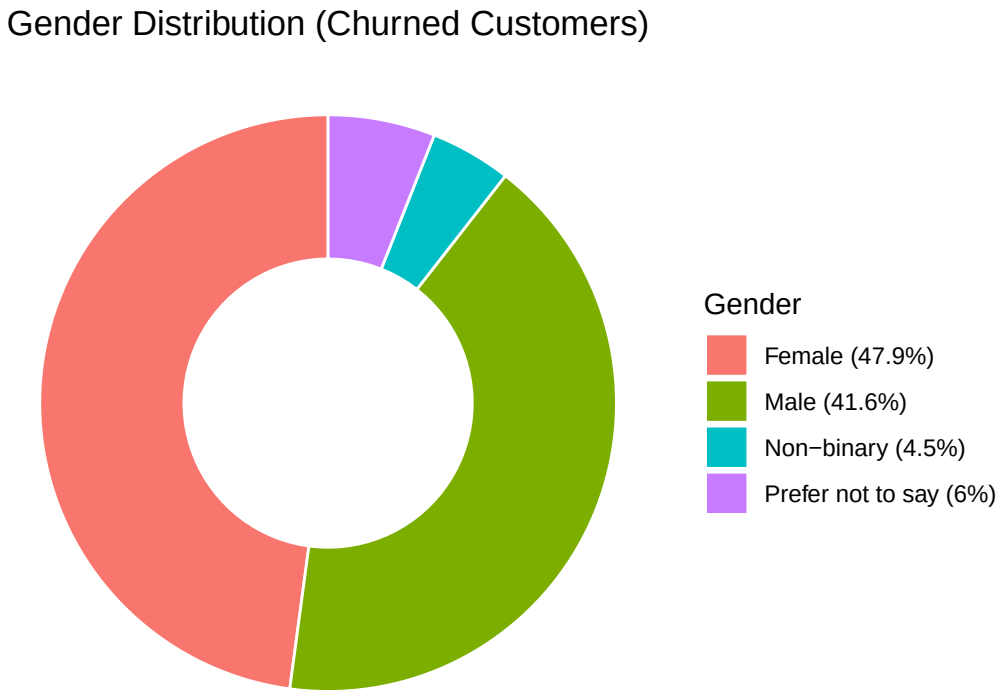

Figure 10: Churn Rate by Gender Distribution

```
ggplot(gender_churn_false, aes(x = 2, y = Prop, fill = PercentLabel)) +
  geom_bar(stat = "identity",
          width = 1,
          color = "white") +
  coord_polar(theta = "y") +
  xlim(0.5, 2.5) +
  theme_void() +
  labs(title = "Gender Distribution (Retained Customers)", fill = "Gender") +
  theme(plot.title = element_text(hjust = 0.5))
```

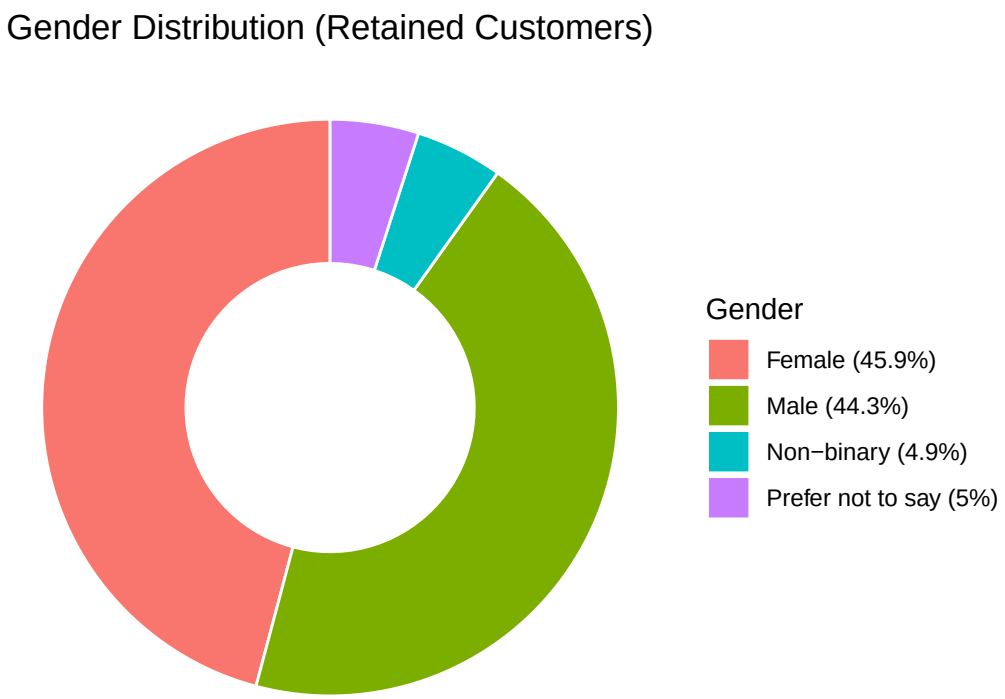

Figure 11: Churn Rate by Gender Distribution

The "Churned Customer" doughnut chart shows that females (47.9%) have a slightly higher churn rate than males (41.6%). This could be due to factors like service experience, product relevance, or marketing alignment not fully meeting female customers' needs. Women might be more sensitive to pricing and value, and differences in engagement patterns could also contribute to higher churn.

In the "Retained Customer" chart, the distribution between females (45.9%) and males (44.3%) is quite balanced, suggesting that gender doesn't strongly impact retention. However, the higher churn rate among females indicates that gender may influence attrition more than retention, pointing to areas like customer experience or product offerings that could be improved to better cater to female customers.

4.10 Churn by Number of Support Tickets

```
ggplot(streaming_data, aes(x = Cancelled, y = NumSupportTickets, color = Cancelled)) +
  geom_jitter(width = 0.2, alpha = 0.5) +
  labs(
    title = "Individual Support Tickets by Churn Status",
    x = "Churned?",
    y = "Number of Support Tickets"
  ) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```

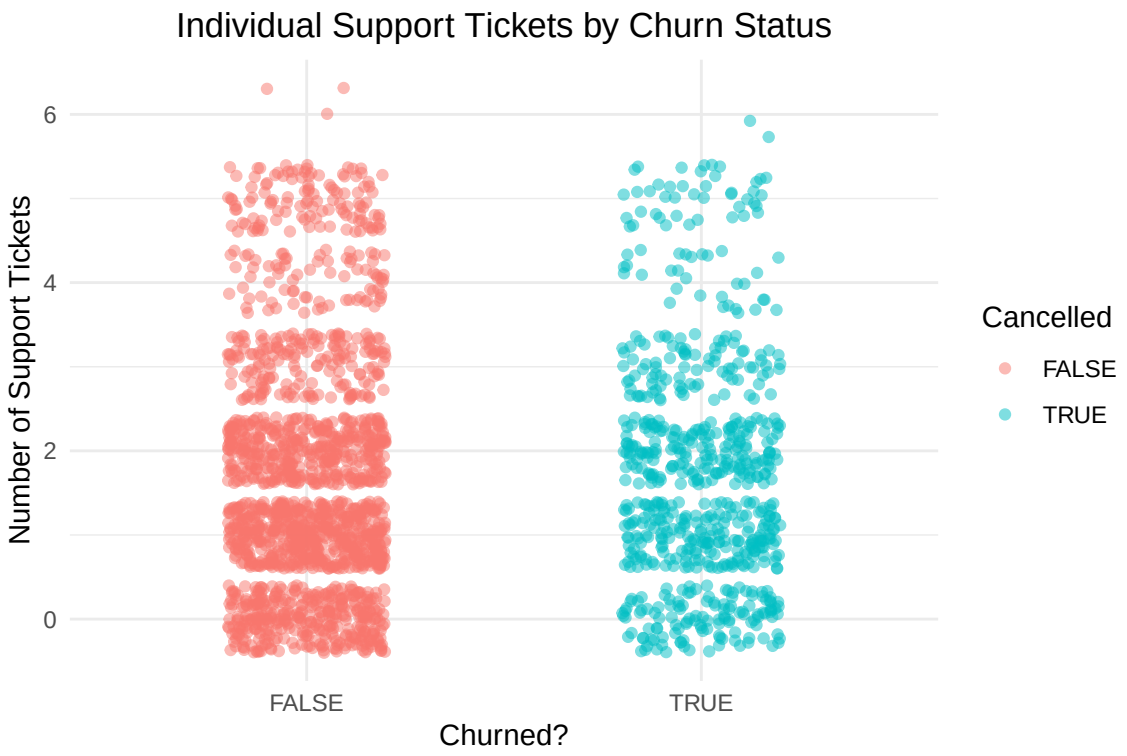

Figure 12: Individual Support Tickets by Churn Status

The chart shows that the number of support tickets is similar between churned and retained customers, mostly ranging from 0 to 5. This suggests that the volume of support requests alone does not drive churn. Many customers who submitted multiple tickets did not cancel, which may indicate that the support team effectively resolved their issues and helped retain them. Therefore, good customer support may have played a key role in encouraging customers to stay.

Conclusion and Recommendations

The analysis reveals that customer churn is primarily driven by a mismatch between customer expectations and the actual experience delivered across different touchpoints. Customers acquired through organic search and direct website visits show the highest cancellation rates, suggesting that these acquisition channels may not be effectively setting accurate expectations about the service. In contrast, those who join via referrals or social media are more likely to stay, possibly because they receive more realistic insights from trusted sources or engaging content, leading to better alignment with the product experience.

Device preference also plays a significant role. Users who access the service through mobile phones or tablets cancel at higher rates, which implies that the user interface or performance on these platforms may be less satisfactory. A likely explanation is that the mobile and tablet versions of the platform do not provide the same seamless or intuitive experience as the desktop version, which is associated with significantly lower churn. This gap in usability could frustrate users and drive them away, particularly when technical issues or navigation difficulties occur on smaller devices.

Subscription type is another influential factor. Customers on premium plans cancel more frequently, indicating that the perceived value of premium features does not consistently justify the higher cost. This may reflect either an over-promise in the premium tier or an under-delivery in terms of exclusive benefits. Conversely, standard and basic plans offer a more predictable and cost-effective service, which appears to better meet customer expectations and encourages longer retention.

Payment frequency also correlates with churn. Monthly subscribers are more likely to cancel, which could be due to the perceived financial burden of frequent payments or greater opportunities to reconsider their commitment. Customers paying quarterly or annually demonstrate higher retention, suggesting that longer-term commitments may foster a more sustained relationship with the service, possibly because these users are more invested or perceive greater value over time.

The choice of payment method contributes further to this pattern. Debit card users exhibit higher churn rates, which may be attributed to payment failure risks, lower perceived security, or limited flexibility. In contrast, payment methods such as PayPal, credit cards, and bank transfers are linked to better retention, potentially due to their convenience, trustworthiness, and consumer protections.

Interestingly, age does not appear to be a significant driver of churn, as the age distribution between canceled and retained customers is largely similar. Likewise, the number of support tickets submitted does not correspond to higher churn. In fact, many customers who made support requests remained subscribed, indicating that the support team is playing a key role in resolving issues effectively and enhancing satisfaction.

Regional and demographic patterns also reveal important context. For instance, certain regions like Region9 and countries such as Australia have notably higher churn rates. These differences may reflect local market conditions, such as competitive pressure, pricing sensitivity, or difficulties in service delivery due to geographic constraints. Moreover, although gender alone does not strongly determine retention, female customers exhibit slightly higher churn, hinting at possible gaps in how the service meets their specific needs or preferences.

In conclusion, customer cancellation is largely influenced by unmet expectations, particularly in terms of user experience, pricing justification, and perceived service value. Addressing these drivers by improving mobile usability, enhancing the appeal and clarity of premium plans, encouraging longer payment commitments, and refining acquisition strategies to better align expectations will be essential in reducing churn. Furthermore, continuing to invest in customer support and understanding regional differences will help tailor services that retain more customers over time.