Subscription Churn Analysis

ST422 Week3 Activity 3

Mani Mahal

January 29, 2026

The purpose of this report is to analyse customer subscription data to identify key characteristics associated with customer churn. Specifically, this analysis aims to:

1. **Characterise the sample:** Provide a summary of customer demographics and account details, stratified by churn status (Table 1).
2. **Visualise key drivers:** Explore the relationship between monthly fees, plan types, and regional churn rates.
3. **Ensure reproducibility:** Demonstrate a robust workflow that can be updated with new data versions (v1, v2, v3) with minimal manual intervention.

```
#if (!require("renv", quietly = TRUE)) {
#     install.packages("renv")
#    library(renv)
#}
#renv::init()


if (!require("pacman", quietly = TRUE)) {
     install.packages("pacman")
    library(pacman)
}

pacman::p_load('knitr','rio', 'readr', 'dplyr', 'ggplot2', 'gt', 'gtsummary',
    'kableExtra', 'tidyverse')
knitr::opts_chunk$set(echo = TRUE, message = FALSE, warning = FALSE, dev = "
    png", dpi = 300)
#renv::snapshot()

# This is the input file which can be switched between v1 and v2
input_file <- 'st422_week3_subscription_v1.csv'

# Loading the data
raw_data_path <- file.path('../data/raw', input_file)
df <- import(raw_data_path)

# Quick view of the data, using glimpse means that every column is shown
glimpse(df)

## Rows: 60
## Columns: 10
## $ customer_id          <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14,
    15,~
```

```
## $ signup_date          <IDate> 2024-04-16, 2024-11-17, 2024-05-31,
    2024-12-29, ~
## $ region               <chr> "Midlands", "South", "Midlands", "Midlands", "
    Nort~
## $ plan_type            <chr> "Basic", "Basic", "Standard", "Basic", "
    Standard",~
## $ tenure_months        <int> 10, 21, 19, 4, 26, 8, 31, 21, 35, 31, 19, 24,
    1, 2~
## $ monthly_fee_gbp      <dbl> 15.02, 12.01, 19.15, 13.13, 17.57, 18.10,
    19.03, 1~
## $ support_tickets_90d <int> 3, 2, 0, 2, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 1, 1,
     0,~
## $ last_login_days      <int> 2, 3, 14, 0, 4, 12, 2, 22, 9, 38, 14, 7, 5, 6,
    16,~
## $ nps_score            <dbl> 14, 8, 24, 6, 23, 21, 33, 15, 39, 32, 36, 15,
    30, ~
## $ churned_90d          <int> 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,
     0,~
```

```
# Checking for missing values
df %>%
  summarise(across(everything(), ~ sum(is.na(.x))))
```

```
##   customer_id signup_date region plan_type tenure_months monthly_fee_gbp
## 1           0           0      0         0             0               0
##   support_tickets_90d last_login_days nps_score churned_90d
## 1                   0               0         1           0
```

```
# Checking values for columns with data type character
# Ensuring missing values are not labeled as Unknown, ? etc
df %>%
  select(where(is.character)) %>%
  map(unique)
```

```
## $region
## [1] "Midlands" "South"    "North"    "Scotland" "Wales"
##
## $plan_type
## [1] "Basic"    "Standard" "Premium"
```

```
# Processing the data for table 1

df_clean <- df %>%
  mutate(
    churned_90d = factor(churned_90d, levels = c(0, 1), labels = c("Active", "
        Churned")),
    region = factor(region),
    plan_type = factor(plan_type)
  )
```

```
# Generating the table 1

t1 <- df_clean %>%
  select(churned_90d, tenure_months, monthly_fee_gbp, nps_score, region, plan_
      type) %>%
```

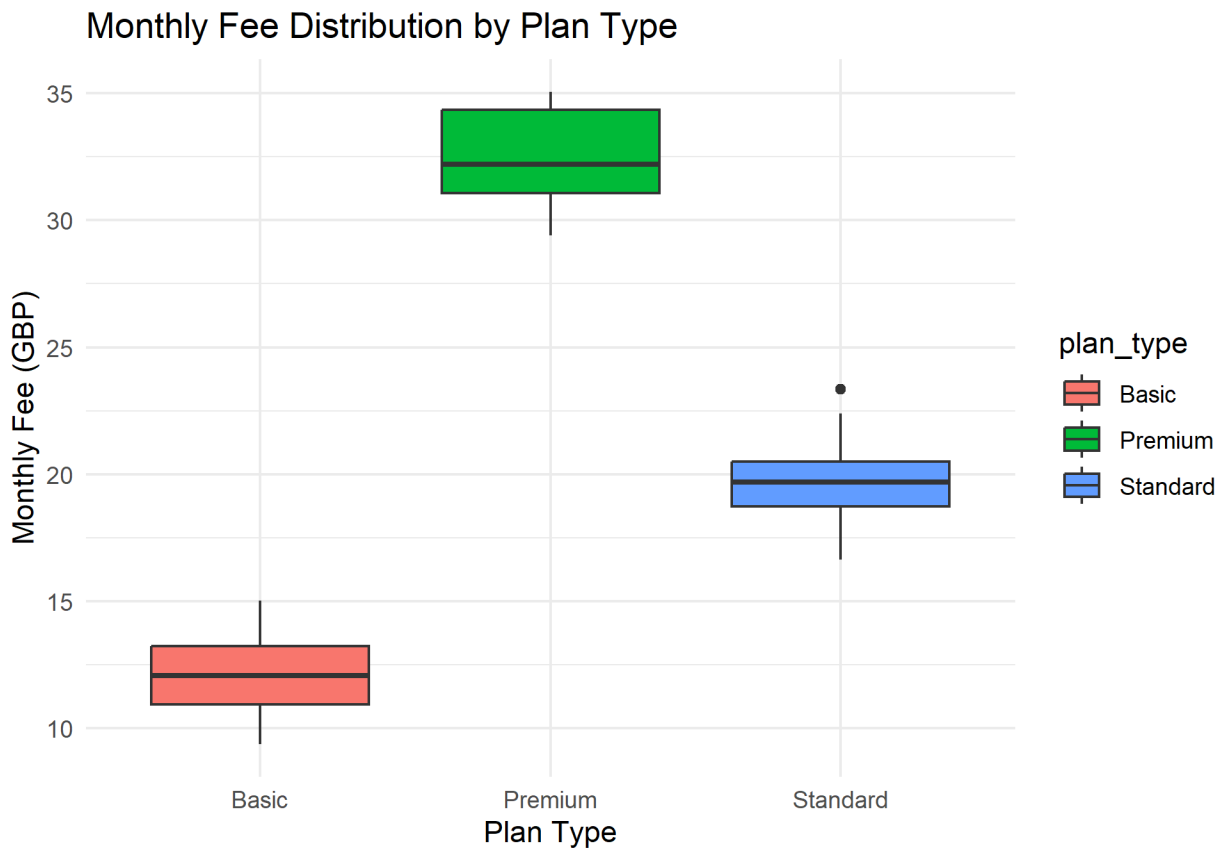
```
  tbl_summary(
    by = churned_90d,                                    # Stratify by Churn
    missing = "ifany",                                   # Explicit missingness
    label = list(
      tenure_months ~ "Tenure (Months)",
      monthly_fee_gbp ~ "Monthly Fee (GBP)",
      nps_score ~ "NPS Score",
      region ~ "Region",
      plan_type ~ "Plan Type"
    ),
    statistic = list(
      all_continuous() ~ "{mean} ({sd})",                # Mean (SD) for
          symmetric
      all_categorical() ~ "{n} ({p}%)"                   # n (%) for
          categorical
    ),
    digits = all_continuous() ~ 1                        # Precision to 1dp
  ) %>%
  add_overall() %>%                                      # Overall column
  add_n() %>%                                            # Add N column
  modify_header(label = "**Variable**") %>%
  modify_caption("**Table 1. Customer Demographics and Account Characteristics
      by Churn Status**") %>%
  as_gt() %>%
  gt::tab_source_note(source_note = "Data represents active subscriptions as
      of Jan 2026. NPS Score missingness indicates customers who did not
      respond to the survey.")


# Outputting the table for the report
t1

# Saving the table 1 as table1.csv in /outputs/tables/

table1_raw <- df_clean %>%
  select(churned_90d, tenure_months, monthly_fee_gbp, nps_score, region, plan_
      type) %>%
  tbl_summary(by = churned_90d) %>%
  as_tibble()

write_csv(table1_raw, "../outputs/tables/table1.csv")

# Plot of Monthly Fee Distribution by Plan Type

p1 <- ggplot(df, aes(x = plan_type, y = monthly_fee_gbp, fill = plan_type)) +
  geom_boxplot() +
  labs(title = "Monthly Fee Distribution by Plan Type",
       y = "Monthly Fee (GBP)", x = "Plan Type") +
  theme_minimal()

p1
```

Table 1: **Table 1. Customer Demographics and Account Characteristics by Churn Status**

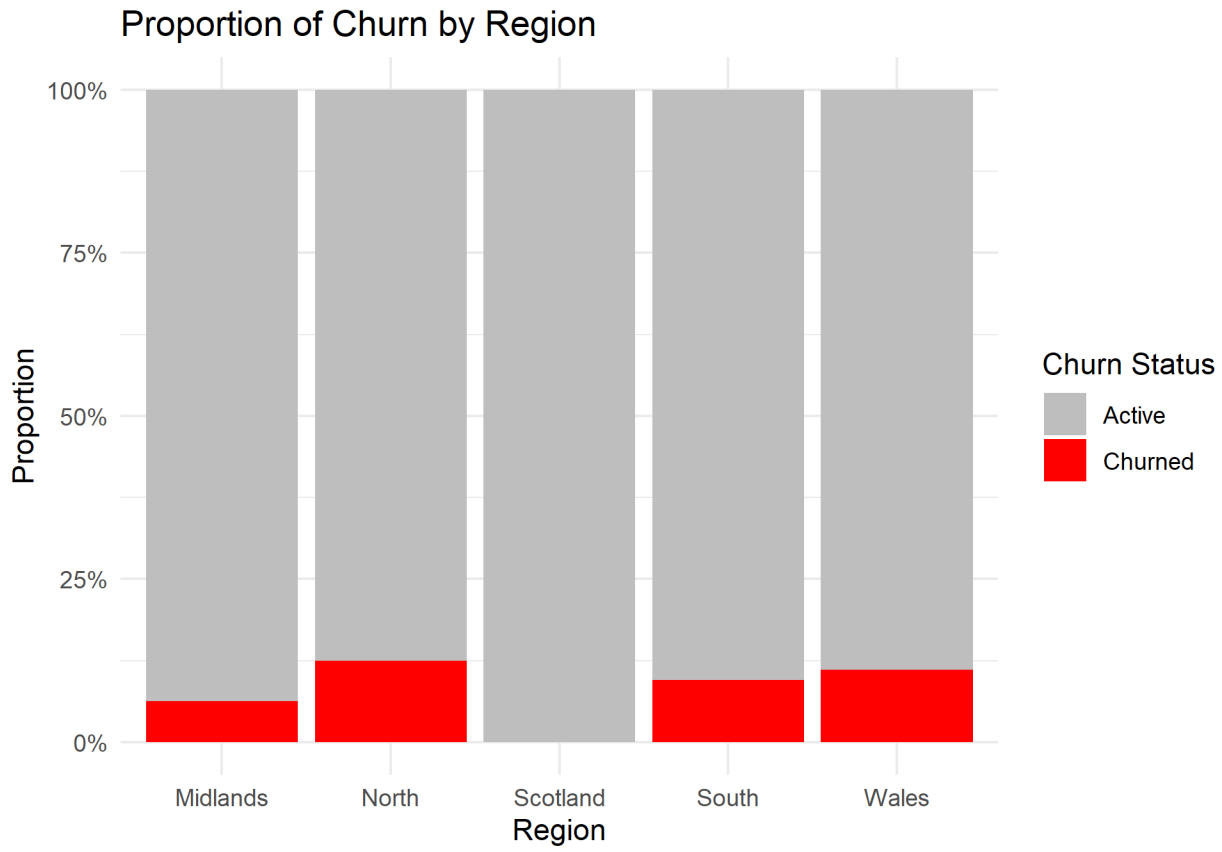
Variable	**N**	**Overall** N = 60[1]	**Active** N = 55[1]	**Churned** N = 5[1]
Tenure (Months)	60	20.2 (10.4)	20.5 (10.3)	16.8 (12.7)
Monthly Fee (GBP)	60	19.1 (7.1)	18.9 (6.8)	21.7 (11.2)
NPS Score	59	18.0 (9.9)	17.9 (10.4)	19.2 (3.0)
Unknown		1	1	0
Region	60			
Midlands		16 (27%)	15 (27%)	1 (20%)
North		8 (13%)	7 (13%)	1 (20%)
Scotland		6 (10%)	6 (11%)	0 (0%)
South		21 (35%)	19 (35%)	2 (40%)
Wales		9 (15%)	8 (15%)	1 (20%)
Plan Type	60			
Basic		22 (37%)	20 (36%)	2 (40%)
Premium		10 (17%)	8 (15%)	2 (40%)
Standard		28 (47%)	27 (49%)	1 (20%)

[1]Mean (SD); n (%)

Data represents active subscriptions as of Jan 2026. NPS Score missingness indicates customers who did not respond to the survey.



```r
# Saving the plot as fig1_fee_dist.png in outputs/figures
ggsave("../outputs/figures/fig1_fee_dist.png", plot = p1, width = 6, height =
    4)

# Plot of Churn Rate by Region

p2 <- ggplot(df, aes(x = region, fill = as.factor(churned_90d))) +
  geom_bar(position = "fill") +
  labs(title = "Proportion of Churn by Region",
       y = "Proportion", x = "Region", fill = "Churn Status") +
  scale_y_continuous(labels = scales::percent) +
  scale_fill_manual(values = c("grey", "red"), labels = c("Active", "Churned")
      ) +
  theme_minimal()

p2
```




```r
# Saving the plot as fig2_churn_region.png in outputs/figures
ggsave("../outputs/figures/fig2_churn_region.png", plot = p2, width = 6,
    height = 4)
```