FPCP3

Robin Tran

Introduction

This project explores customer behavior using a dataset with demographic, transactional, and engagement features. There are two main sections. In the first section, we implemented linear regression models to predict each customer's average transaction value, comparing different model specifications using 10-fold cross-validation. In the second section, we used logistic regression to classify whether a customer is at high churn risk based on selected predictors. The goal is to understand what factors are associated with customer spending and retention, and to evaluate model performance using appropriate validation techniques.

Variables of Interest

I will be considering 7 numerical variables and 12 categorical variables. They are listed below along.

Table 1: Numerical Variables

Variable	Description
age	The age of the user, measured in years.
days_since_last_login	The number of days since the user last logged into the website, measured in days.
avg_time_spent	The average amount of time the user spends per visit on the website, measured in minutes.
avg_transaction_value	The average value of transactions made by the user, measured in monetary units (currency).
avg_frequency_login_days	The average number of days between the user's consecutive logins, measured in days.
points_in_wallet	The amount of loyalty or reward points currently available in the user's wallet, measured in points.
churn_risk_score	The churn risk score assigned to the user, ranging from 1 to 5, indicating the likelihood of the user leaving the service (higher scores indicate higher risk).

Table 2: Categorical Variables

Variable	Categories
gender	F, M, Unknown
region_category	Village, City, Town
membership_category	Platinum Membership, Premium Membership, No Membership, Gold Membership, Silver Membership, Basic Membership
joined_through_referral	Yes, No
preferred_offer_types	Gift Vouchers/Coupons, Credit/Debit Card Offers, Without Offers
medium_of_operation	Desktop, Smartphone, Both
internet_option	Wi-Fi, Mobile_Data, Fiber_Optic
used_special_discount	Yes, No
offer_application_preference	Yes, No
past_complaint	Yes, No
complaint_status	Not Applicable, Solved, Solved in Follow-up, Unsolved, No Information Available
feedback	Products always in Stock, Quality Customer Care, Poor Website, No reason specified, Poor Product Quality, Poor Customer Service, Too many ads, User Friendly Website, Reasonable Price

Observational Unit

Each row in the data set represents one individual customer who has engaged with the platform and has at least one recorded purchase.

We also verified the categorical variables' types.

Exploratory Data Analysis

Summaries

We included statistical summary for our numerical variables as below.

We also included statistical summaries for some of our categorical variables as below. The summaries are quite insightful. We observe balanced classes in `gender`, but there might be some imbalances in `membership_category` and `region_category`.

Visualizations

To better understand the relationship between user behavior and churn risk, we included a variety of plots.

Table 3: Summary Statistics for Numerical Variables

Variable	mean	median	sd	IQR	min	max	n
age	37.08	37.00	15.91	28.00	10.00	64.00	24853
days_since_last_login	-42.61	12.00	230.18	8.00	-999.00	26.00	24853
avg_time_spent	244.14	162.37	398.10	295.65	-2281.24	3040.41	24853
avg_transaction_value	29321.54	27534.68	19499.51	26605.22	800.46	99914.05	24853
avg_frequency_login_days	15.97	16.00	9.23	14.00	-43.65	73.06	24853
points_in_wallet	688.30	698.66	195.79	148.52	-760.66	2069.07	24853
churn_risk_score	3.61	4.00	1.18	2.00	1.00	5.00	24853

Table 4: Distribution of Gender

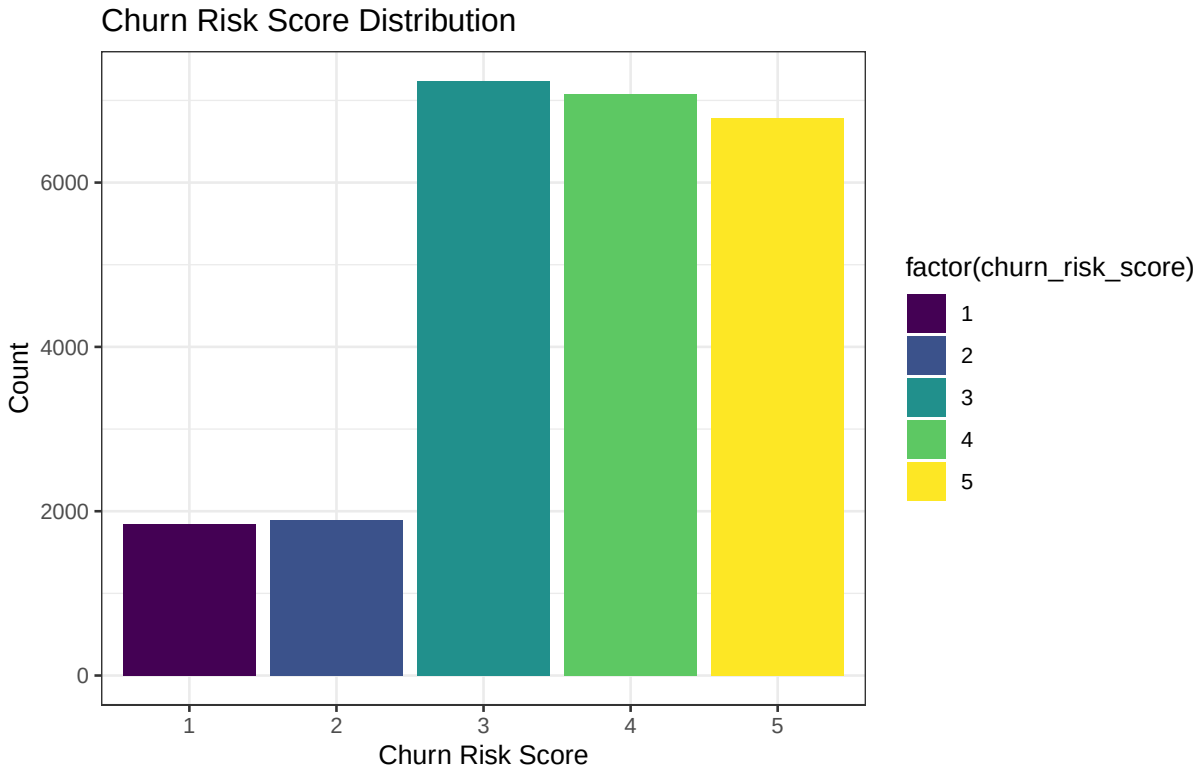
gender	n	percentage
F	12463	50.15
M	12351	49.70
Unknown	39	0.16

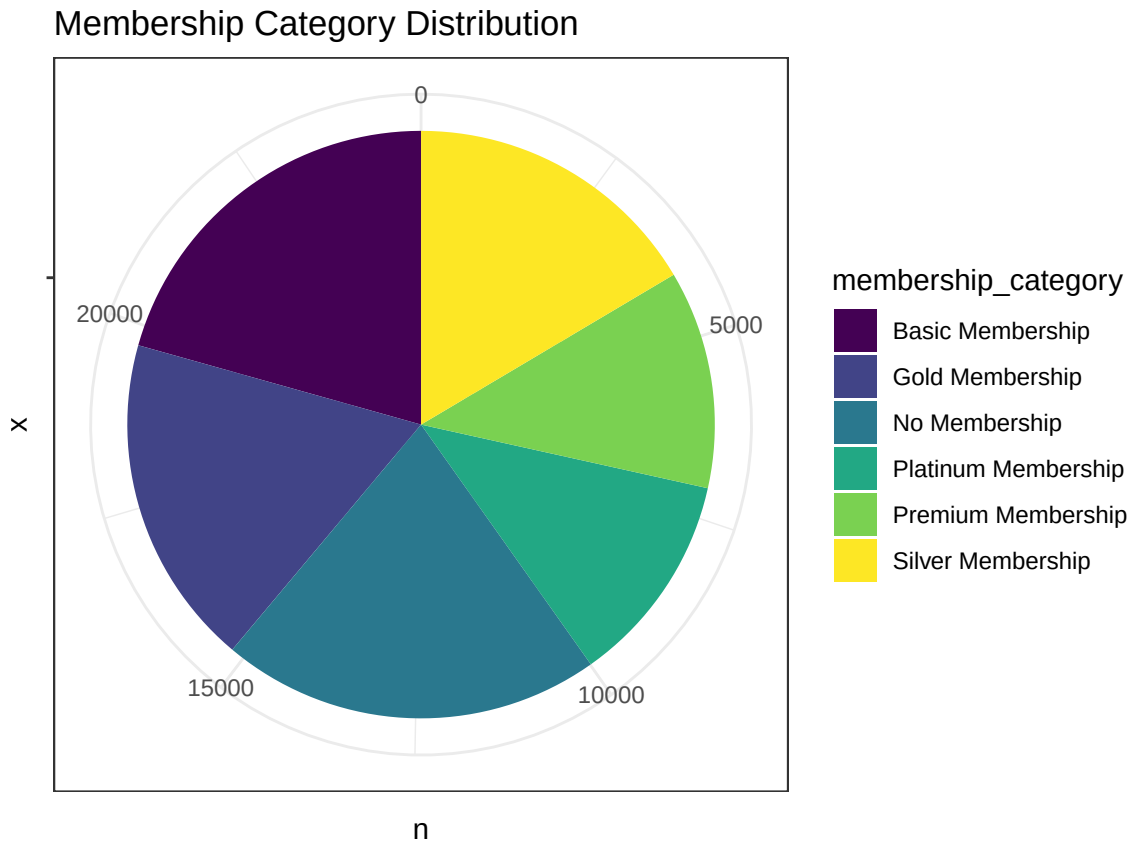
Table 5: Distribution of Membership

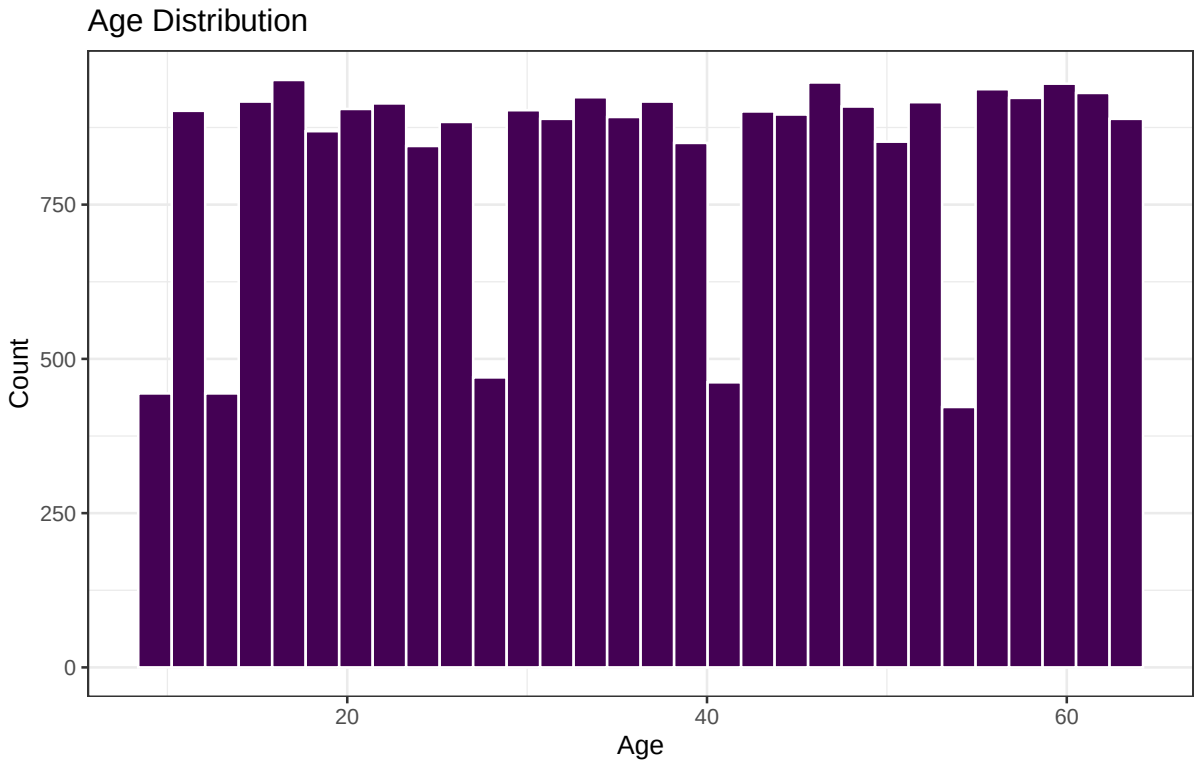
membership_category	n	percentage
Basic Membership	5131	20.65
Gold Membership	4535	18.25
No Membership	5198	20.91
Platinum Membership	2912	11.72
Premium Membership	2982	12.00
Silver Membership	4095	16.48

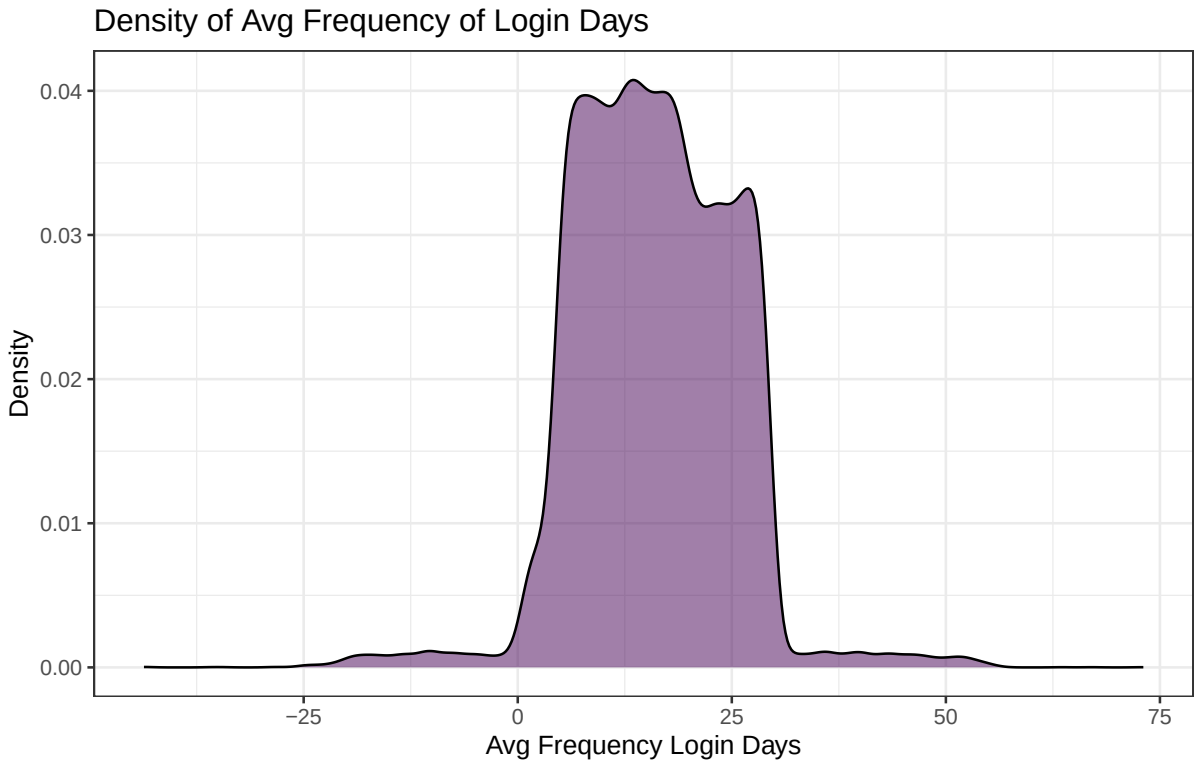
Table 6: Distribution of Region

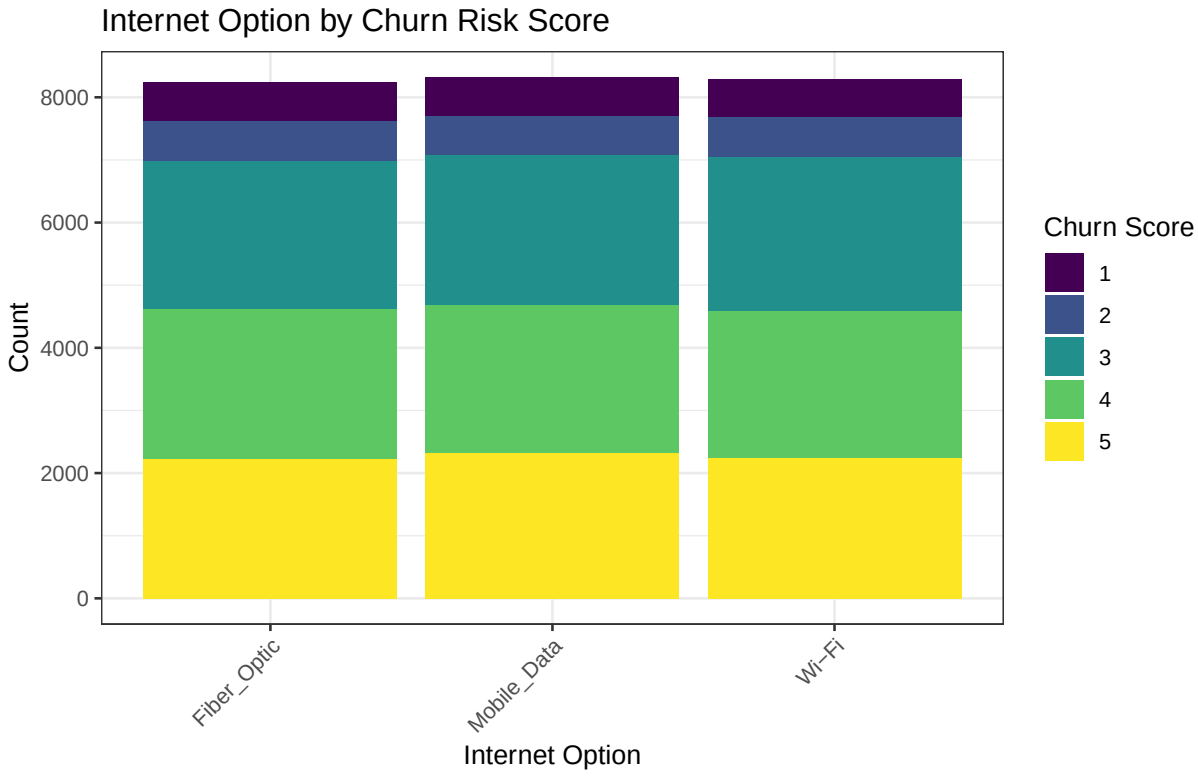
region_category	n	percentage
City	10020	40.32
Town	11090	44.62
Village	3743	15.06

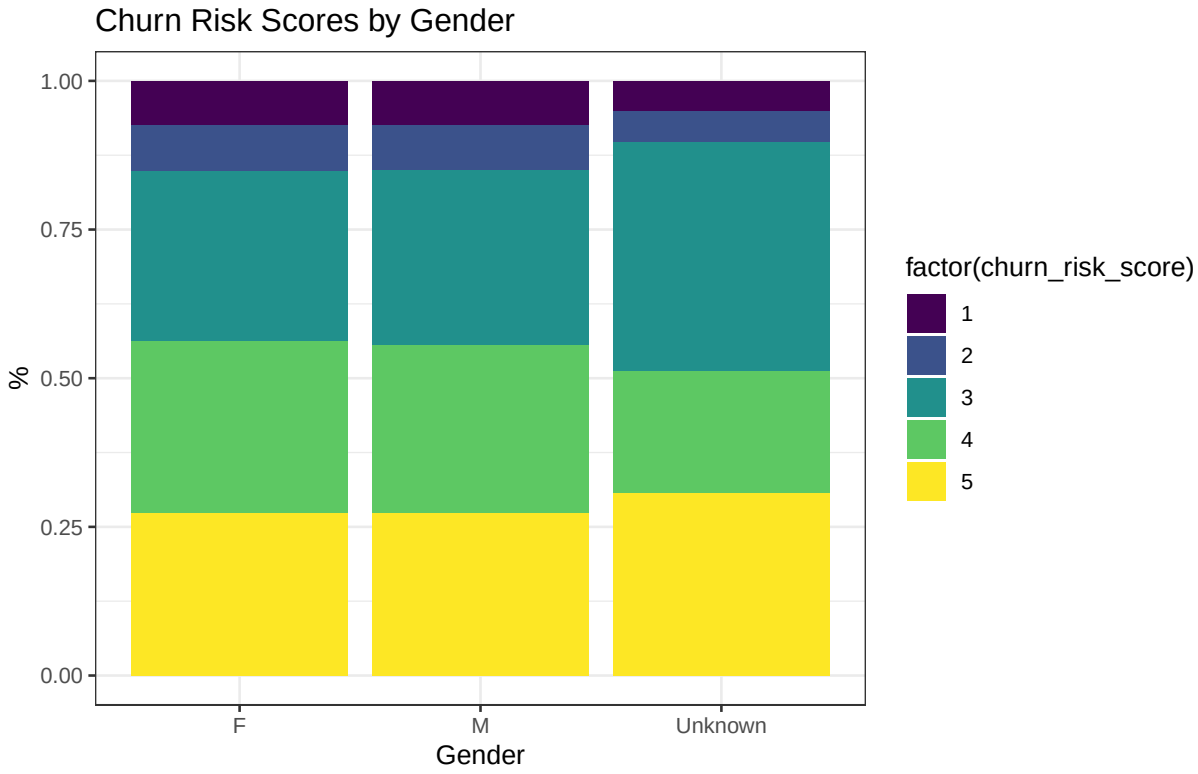

Churn Risk Score Distribution
Count
6000
4000
2000
0
1
2
3
4
5
Churn Risk Score
factor(churn_risk_score)
1
2
3
4
5

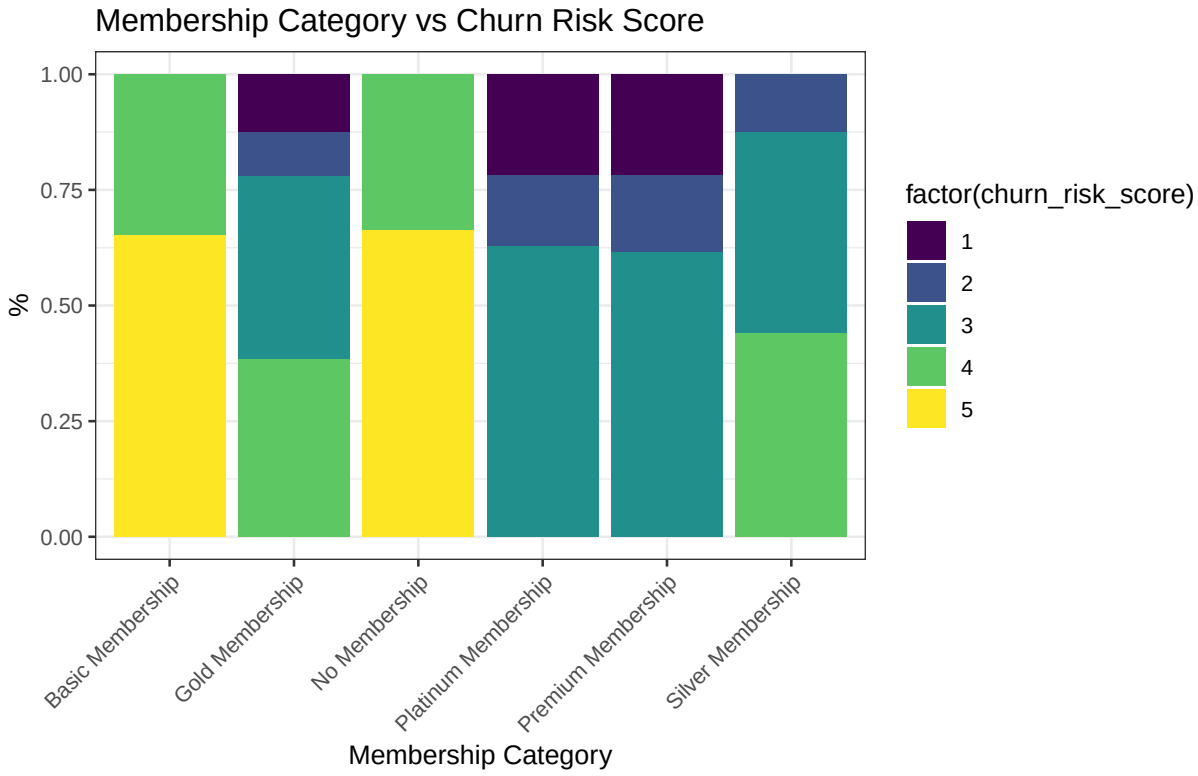

Membership Category Distribution
0
5000
10000
15000
20000
x
n
membership_category
Basic Membership
Gold Membership
No Membership
Platinum Membership
Premium Membership
Silver Membership

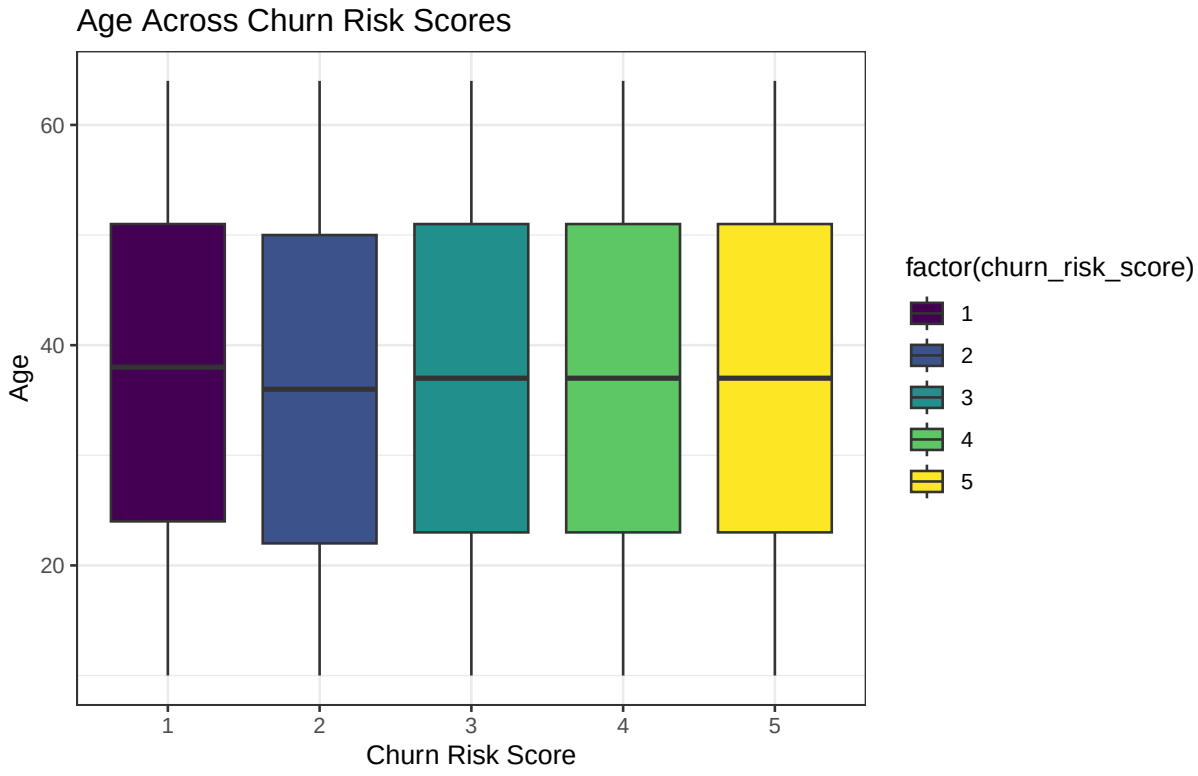

Age Distribution
Count
750
500
250
0
20
40
60
Age

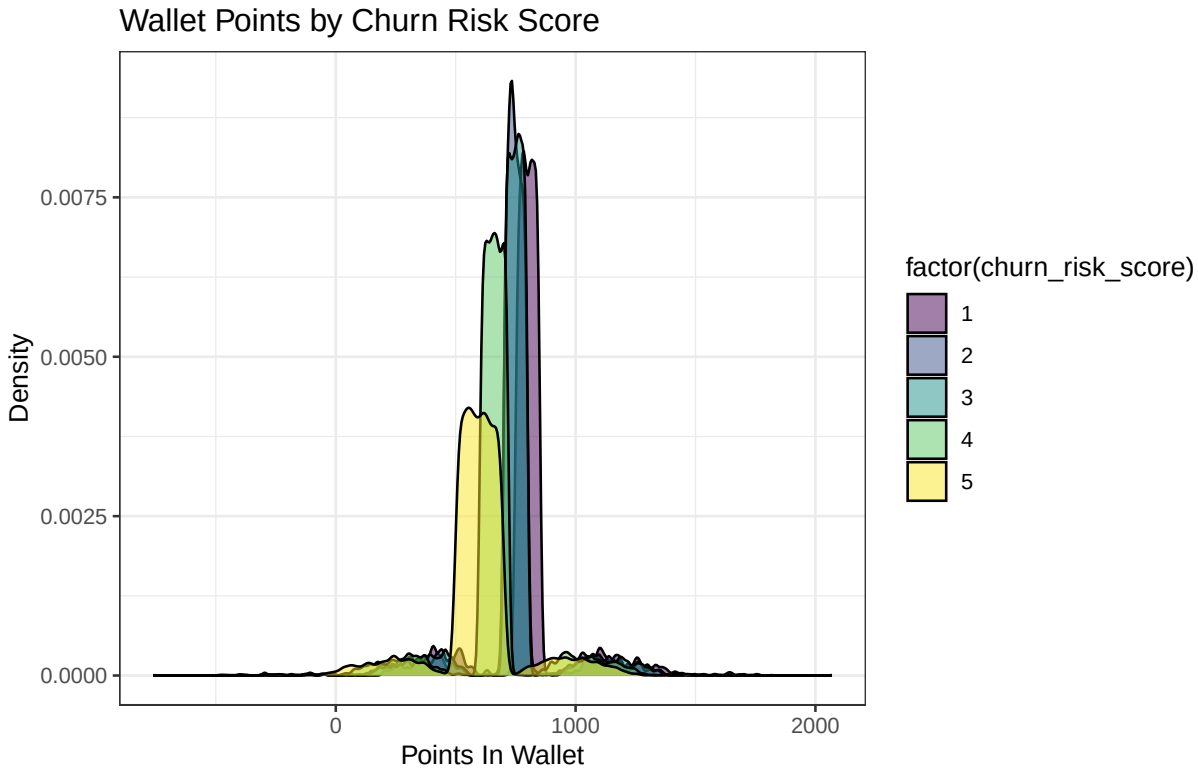

Density of Avg Frequency of Login Days
Density
0.04
0.03
0.02
0.01
0.00
−25
0
25
50
75
Avg Frequency Login Days

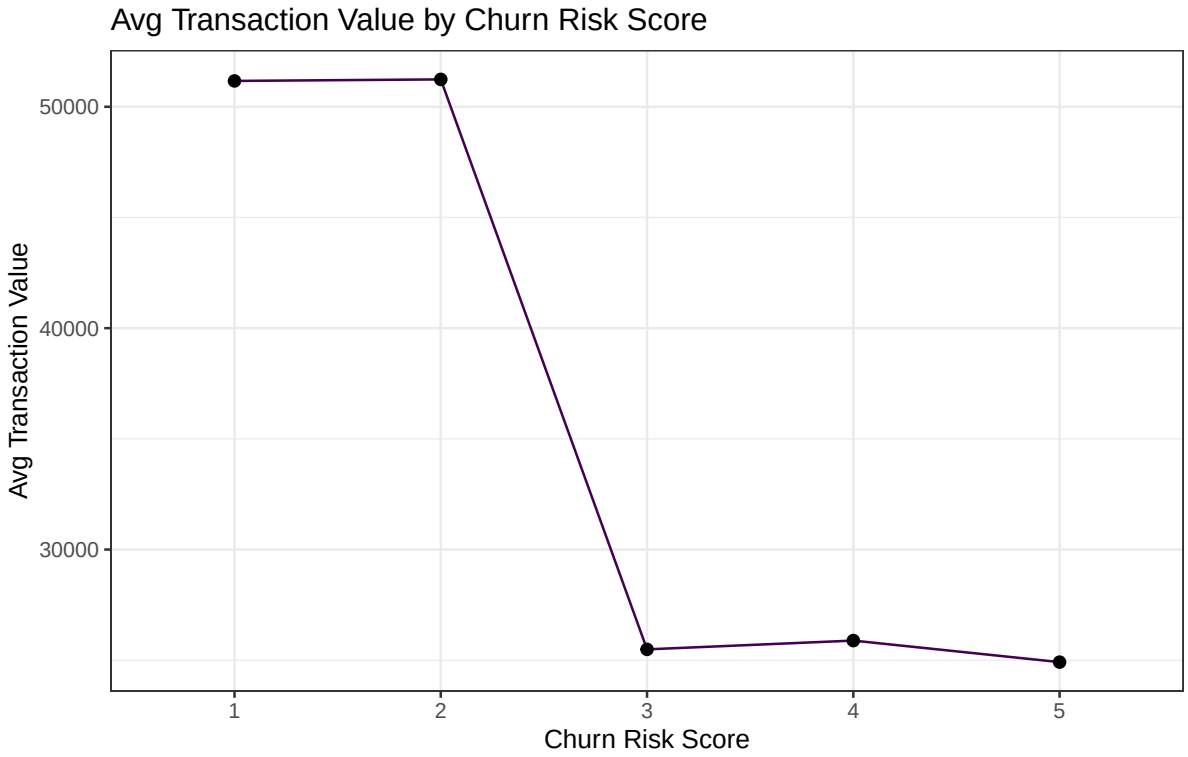

Internet Option by Churn Risk Score
Count
8000
6000
4000
2000
0
Fiber_Optic
Mobile_Data
Wi-Fi
Internet Option
Churn Score
1
2
3
4
5

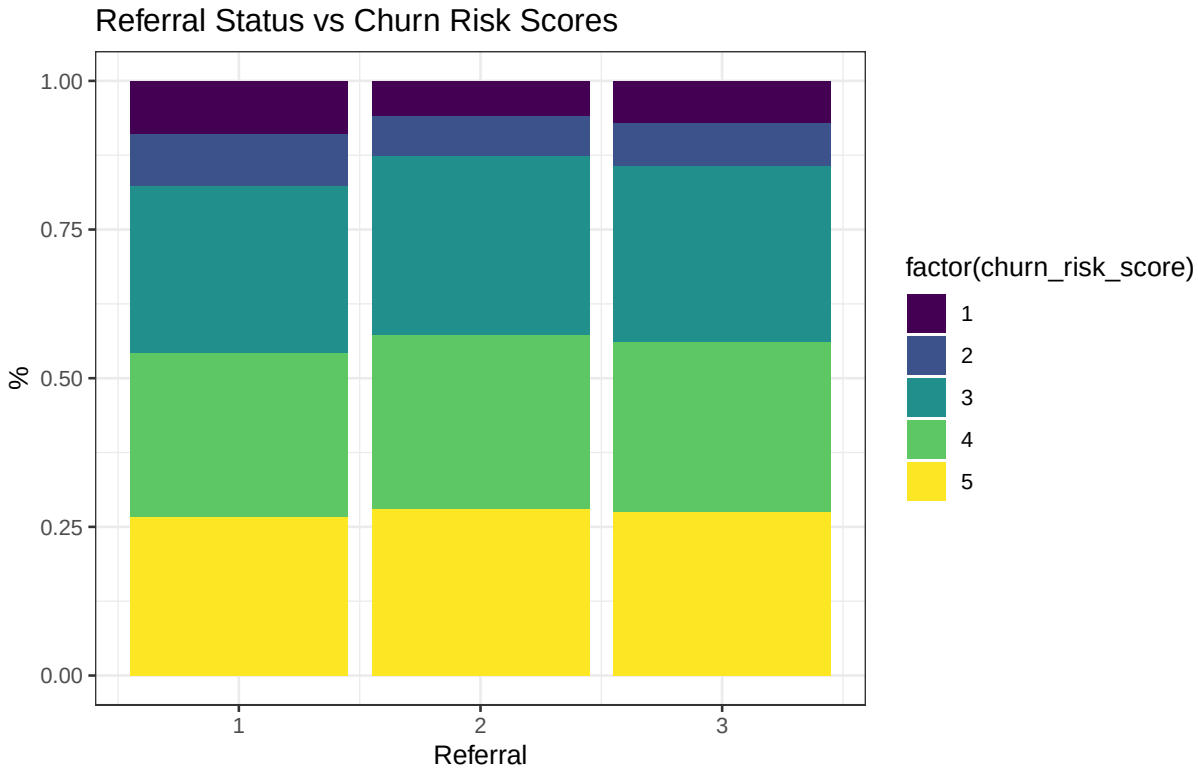

Churn Risk Scores by Gender
1.00
0.75
0.50
0.25
0.00
%
F
M
Unknown
Gender
factor(churn_risk_score)
1
2
3
4
5


Membership Category vs Churn Risk Score
1.00
0.75
0.50
0.25
0.00
%
Basic Membership
Gold Membership
No Membership
Platinum Membership
Premium Membership
Silver Membership
Membership Category
factor(churn_risk_score)
1
2
3
4
5


Age Across Churn Risk Scores
Age
60
40
20
1
2
3
4
5
Churn Risk Score
factor(churn_risk_score)
1
2
3
4
5


Wallet Points by Churn Risk Score
Density
0.0075
0.0050
0.0025
0.0000
0
1000
2000
Points In Wallet
factor(churn_risk_score)
1
2
3
4
5


Avg Transaction Value by Churn Risk Score
Avg Transaction Value
50000
40000
30000
1
2
3
4
5
Churn Risk Score


Referral Status vs Churn Risk Scores
1.00
0.75
0.50
0.25
0.00
%
1
2
3
Referral
factor(churn_risk_score)
1
2
3
4
5