Group based Exploratory Data Analysis

Saren Chatham, Fran Jimenez, Shamsa Khoja, Willow Noltemeyer, Arnold Songa
18 January 2026
MSBA 635

Case Context and Problem Framing

What is the core business problem to solve? Explain the relevance of this problem to the subscription company.

From the telco standpoint, the core business problem is predicting and preventing customer churn. This occurs when customers discontinue their service, directly reducing recurring revenue and increasing acquisition costs.

We want to identify customer churn patterns to identify customers at risk of churning before it happens, allowing the business to intervene. The objective is not only to understand why customers leave, but to identify early warning signals that indicate which customers are most likely to churn in the future.

This is a highly relevant problem because it affects business revenue and overall business profitability. The higher the customer lifecycle value, the more the company's topline revenue. It can also mean that the business does not need to continuously spend money on acquiring new users, which can be very expensive. Therefore, it is vital to understand customer churn behavior and implement a relevant retention strategy accordingly.

Identify whether this is a supervised or unsupervised learning problem. If supervised, specify whether it is a classification or regression problem and justify your reasoning.

Because our data has a response variable (customer churn, labeled "Left Flag") and our goal is to predict the value of that response variable (whether or not a customer will churn), this is a supervised learning problem. Specifically, it is a classification problem because our response variable is categorical (Yes/No) and not numeric.

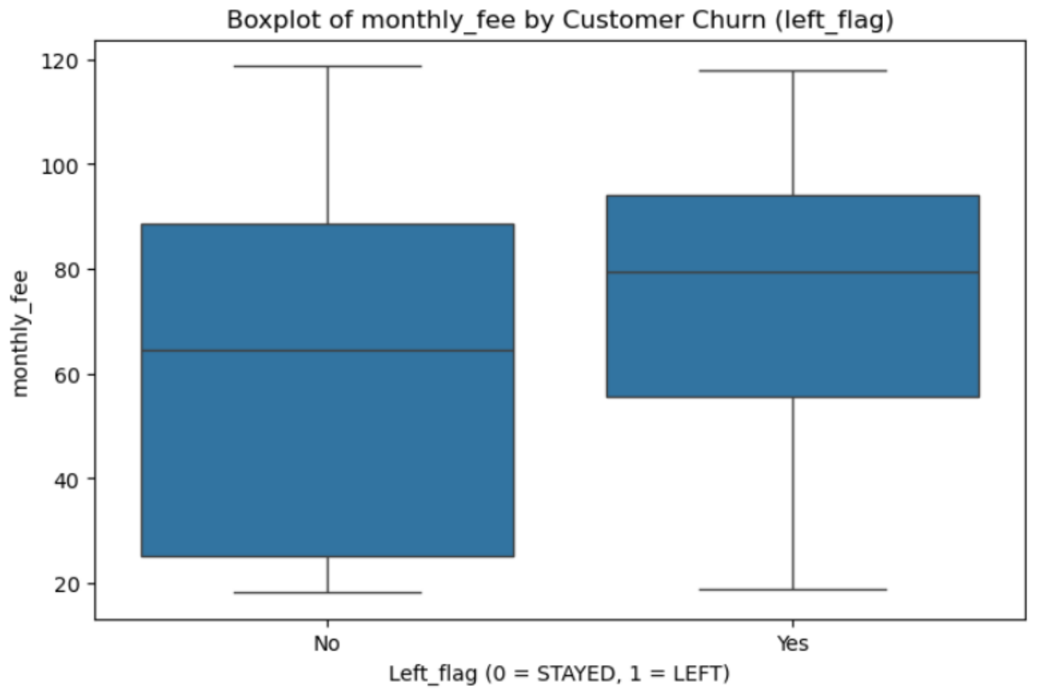
The dataset can be used to train a predictive model that learns correlations between customer variables (tenure, service type, billing behavior, support features, etc.) and churn outcomes because previous churn results are known. Our goal is to predict the likelihood that a current customer will leave in the future so that the company can take action to keep them.

Data Exploration (EDA)

General Overview of the Dataset

What is the response variable? Provide one appropriate visualization or table to show the distribution of the response variable.

The response variable that the model is trying to predict is churn or not churn (left_flag). The dataset indicates that most customers remain, but there is still a significant segment that churns and understanding why they leave is key to reducing that number.



The response variable is “left_flag”, which indicates whether the customer left the company during the given period. (“Yes” indicates churn)

left_flag	count	percent
No	4140	73.5
Yes	1496	26.5

State the number of observations and predictor variables in the dataset.

Number of observations: 5,636 customers

Predictor variables in the dataset: 35 predictor variables (1 response variable: “left_flag”)

Classification of variables:

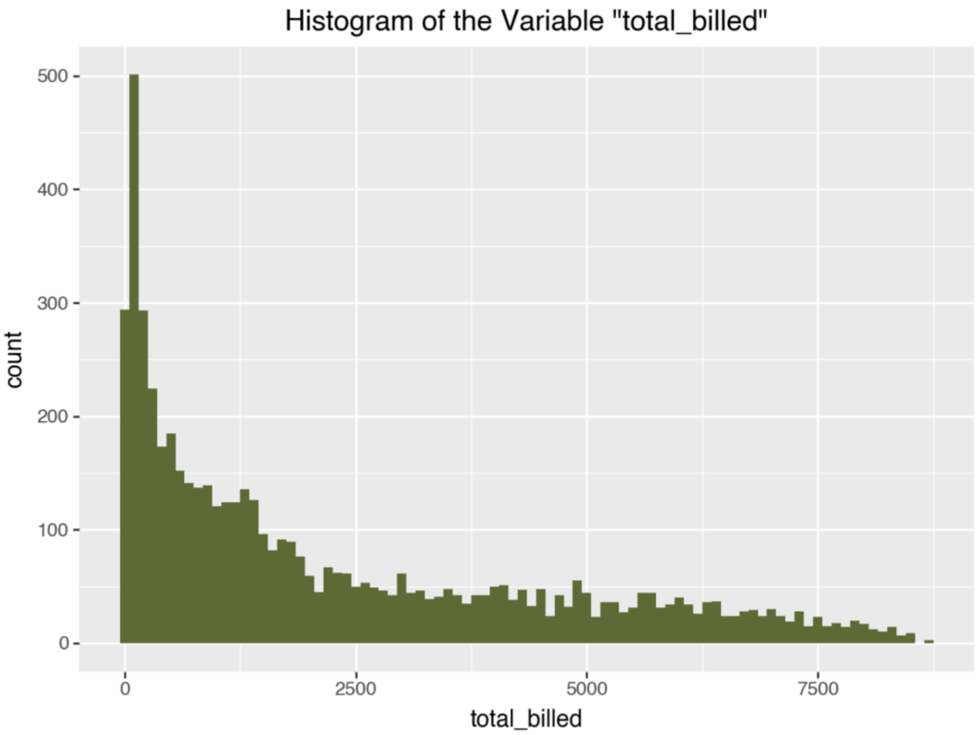
Variable Name	Type
has_dependents	Categorical - Binary
home_phone	Categorical - Binary
left flag	Categorical - Binary
e_bill_opt_in	Categorical - Binary
gender	Categorical - Binary
is married	Categorical - Binary
referred friend	Categorical - Binary
premium_support	Categorical - Binary
stream music	Categorical - Binary
unlimited_data_opt	Categorical - Binary
contract term	Categorical - Nominal
pay_method	Categorical - Nominal
acct ref	Categorical - Nominal
cust_ref	Categorical - Nominal
multi line	Categorical - Nominal
internet_plan	Categorical - Nominal
add_on_security	Categorical - Nominal
add_on_backup	Categorical - Nominal
add_on_protection	Categorical - Nominal
tech_support_std	Categorical - Nominal
stream tv	Categorical - Nominal
stream_movies	Categorical - Nominal
recent offer	Categorical - Nominal
internet tech	Categorical - Nominal
fiscal_qtr	Date
tenure_mo	Numeric
monthly_fee	Numeric
total billed	Numeric
age years	Numeric
dependents_count	Numeric

referrals_count	Numeric
avg_long_dist_fee	Numeric
avg_gb_download	Numeric
refunds total	Numeric
extra data fees total	Numeric
long dist_fees_total	Numeric

Data Quality Assessment

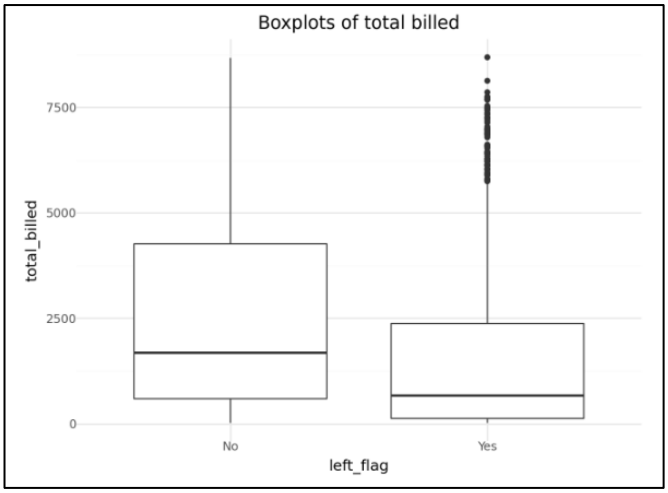
Three variables had missing values: “recent_offer” had 3106 NA values, “internet_tech” had 1212 NA values, and “total_billed” had 8 NA values. We filled the NA values from the “recent_offer” column with “No Offer” and filled the NA values from the column “internet_tech” with “Unknown”.

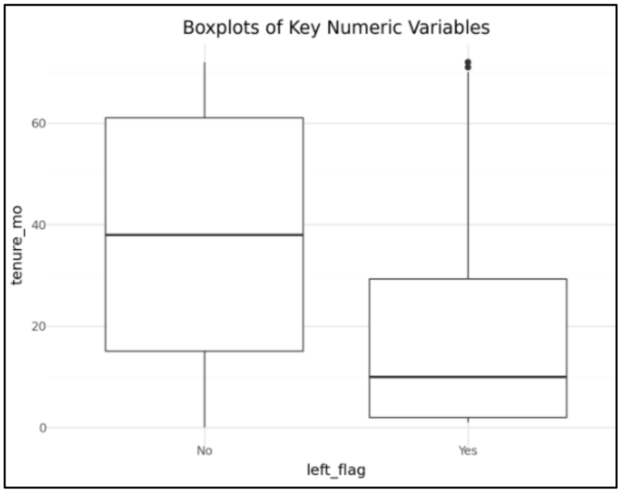
We chose to replace NA values for the column “total_billed” with the median of the dataset. We chose to use the median over the mean because the distribution of the variable is heavily left-skewed, which biases the mean towards lower “total_billed” values.

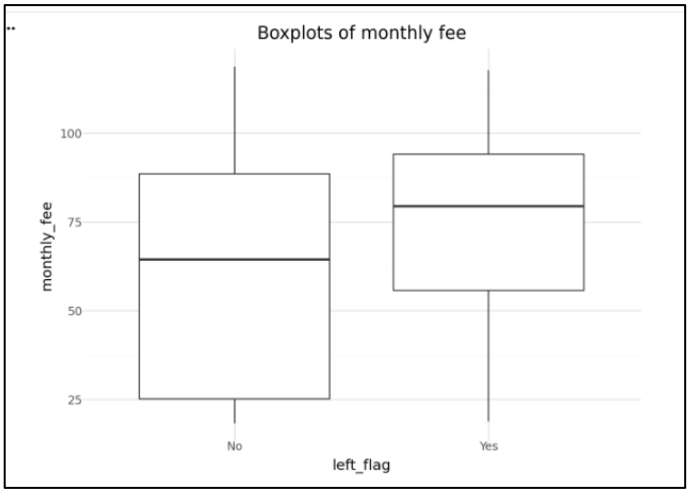


There were no duplicates in the data

We created box plots for “monthly_fee”, “total_billed”, and “tenure_mo”. There were no outliers in “monthly_fee”, a couple of outliers in “tenure_mo” and 20+ in “total_billed”. Because there are very few outliers when compared to the total number of records (5,636), we determined that the effect of the outliers on the dataset was insignificant and did not remove them.

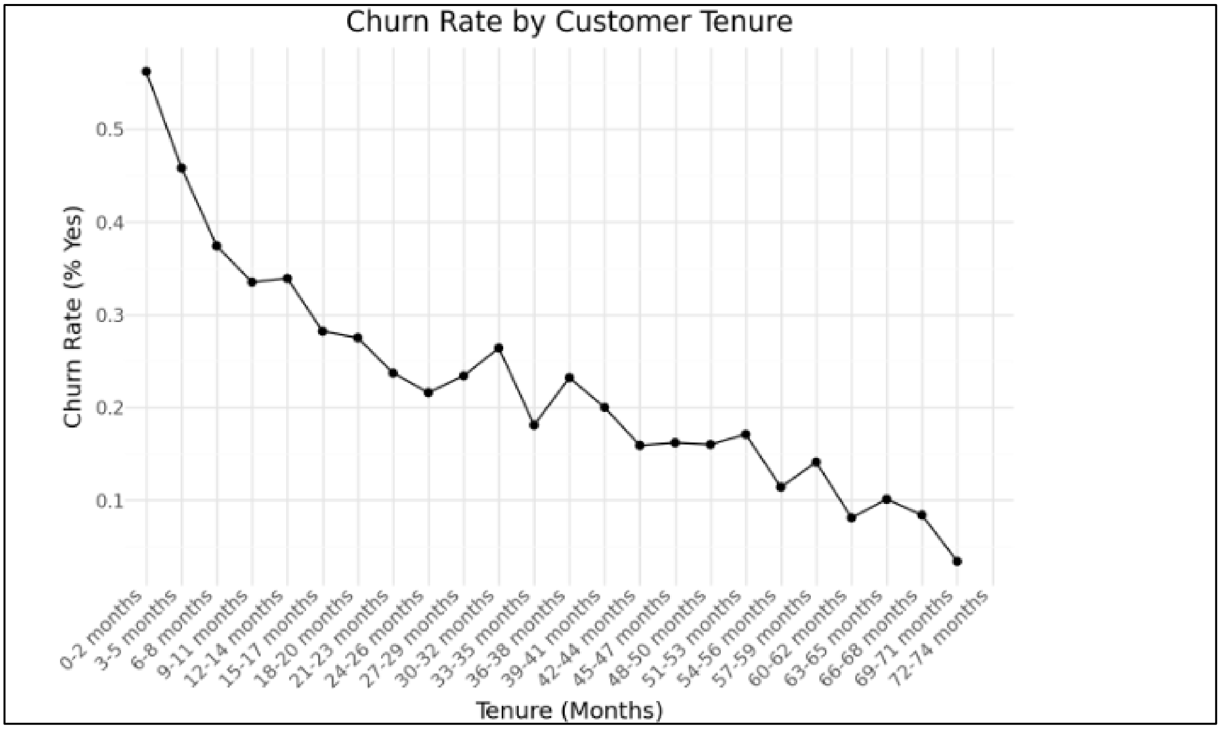






Individual Assignment:

Shamsa: Plot 1

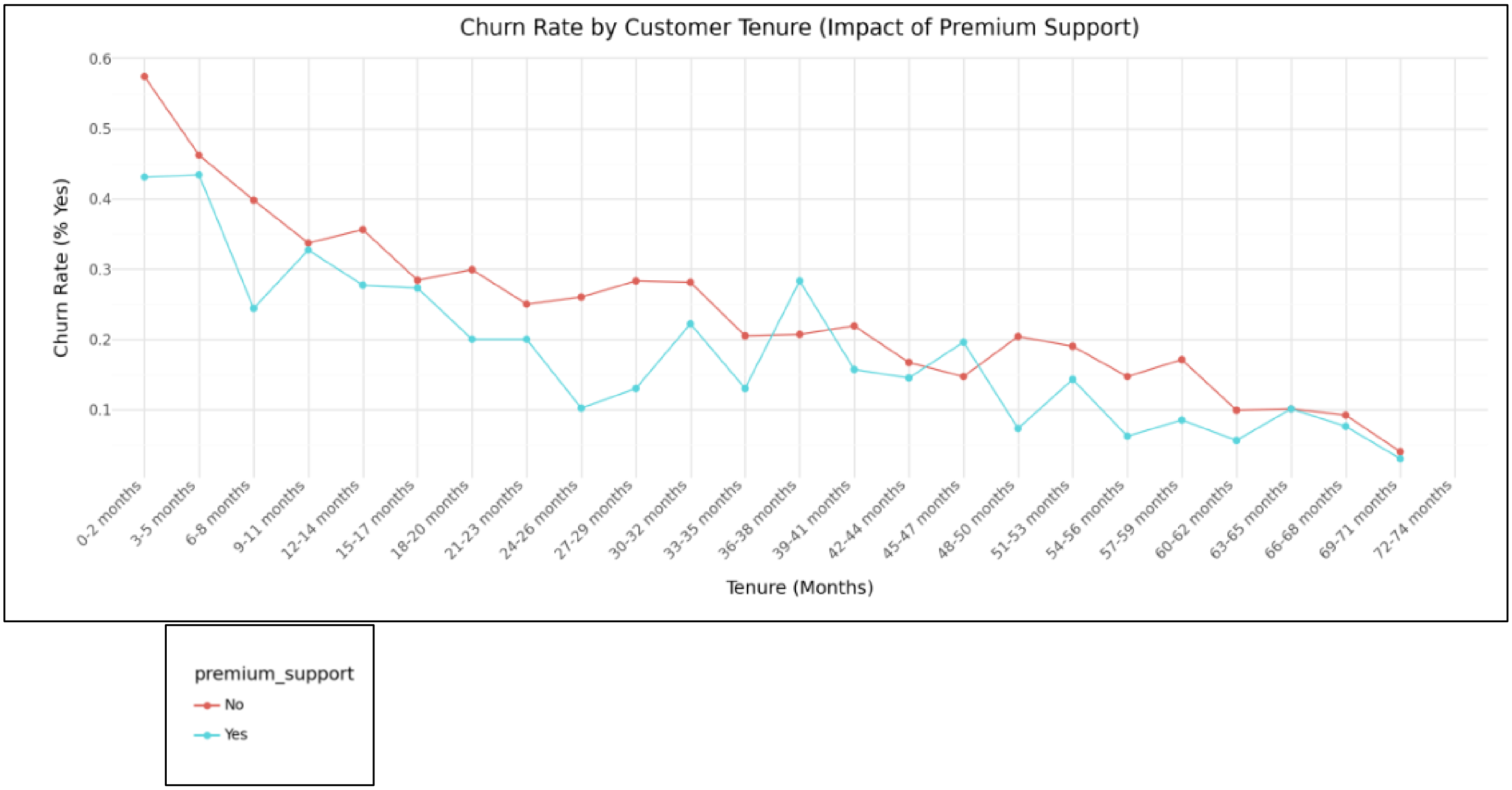


Purpose: This chart shows how customer churn rates change over time as monthly tenure increases.

So what? Churn is an early lifecycle problem. It drops from about 56% in the first 3 months to about 37% after 6 months. So, this shows that new customers are far more likely to leave in first 6-8 months and after which churn stabilizes past year one at much lower levels, suggesting customers who stay past early tenure are increasingly loyal.

Business takeaway: The company should focus on retention efforts on the first few months of the customer lifecycle, where churn risk is highest. A better onboarding experience and customer support in early lifecycle can help in reducing churn.

Shamsa: Plot 2

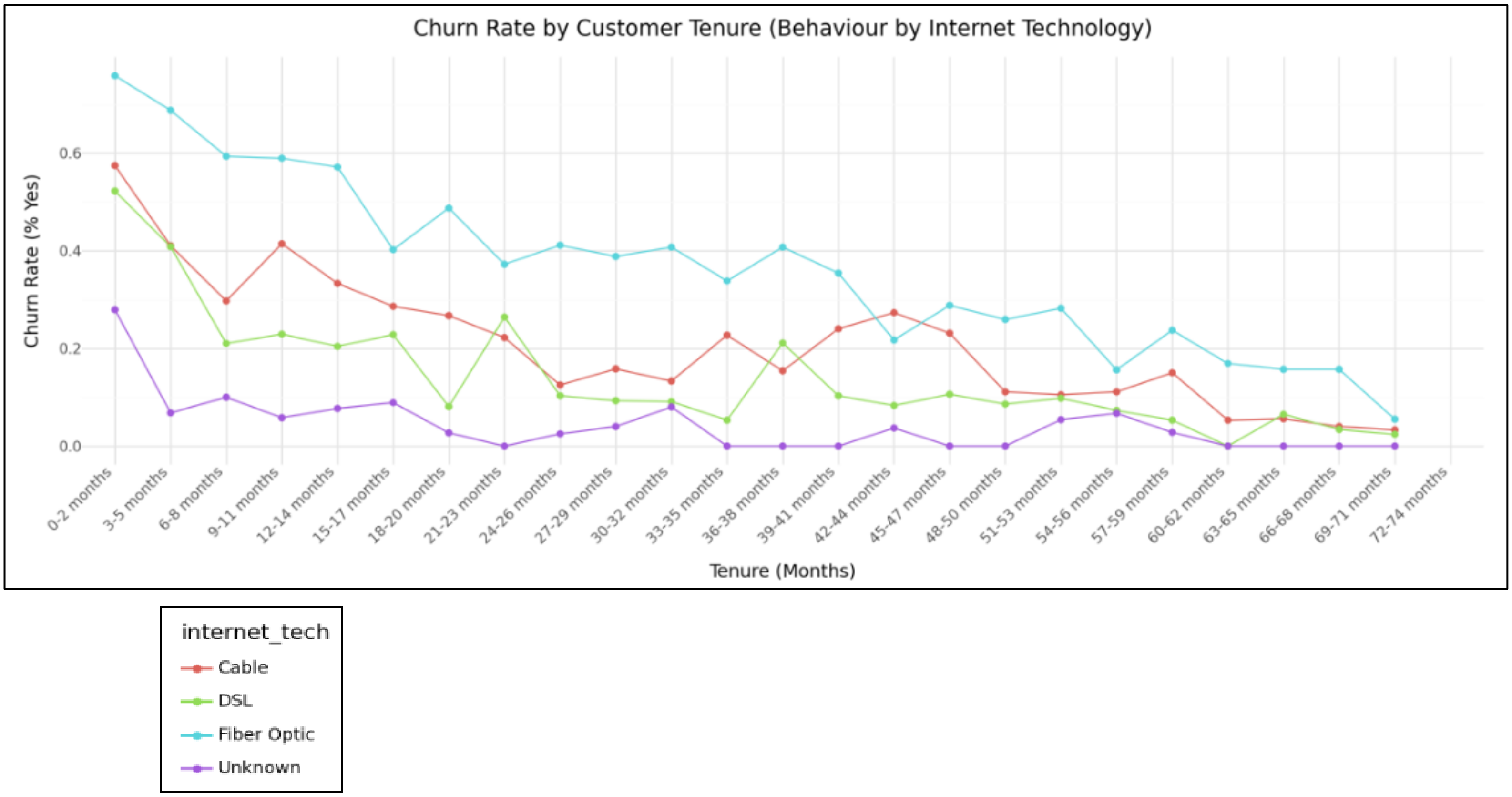


Purpose: This chart shows how churn rate changes across customer tenure in months, comparing customers with and without premium support.

So what? Churn is highest in the early months and stabilizes past year one, but customers without premium support consistently show higher churn across lifecycle. The impact of premium services on churn rate is highest in the first 12-24 months.

Business takeaway: Business should offer premium support during onboarding customers at early lifecycle, perhaps as a free trial or an onboarding offer. This can reduce churn and improve retention rate significantly.

Shamsa: Plot 3



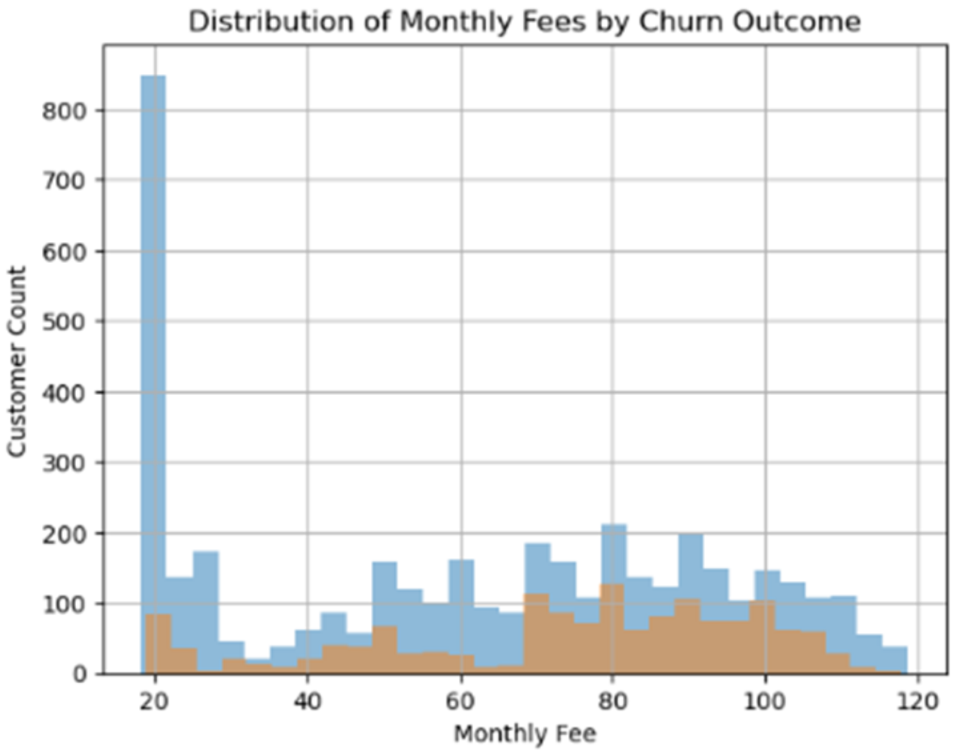
Purpose: This chart shows how churn rates change across customer tenure, segmented by internet technology type.

So what? Churn is highest in the early tenure periods for all technologies but is consistently highest for Fiber Optic customers, followed by Cable and DSL users. Comparing churn rates at 6-8 months, Fiber Optic users are at 60% churn rate while DSL users are at 20%, and cable users are at 30%. Customers using Fiber Optic seem to be at the highest churn risk;perhaps they have high service expectations from the telco.

Business takeaway: Retention efforts should be focused on early lifecycle with more aggressive efforts on Fiber Optic customers. Connecting my plot –2, a business recommendation could be to bundle premium support offering for at least Year1 for all Fiber Optic customers.

Fran: Plot 1

Distribution of monthly fees by churn outcome (Distribution of a single variable).



Churners are NOT concentrated only at the highest prices. The lowest price tier (<$25) shows disproportionately low churn despite contributing very little to revenue. The mid-range monthly fee charge ($60 to $90) has a bigger overlap.

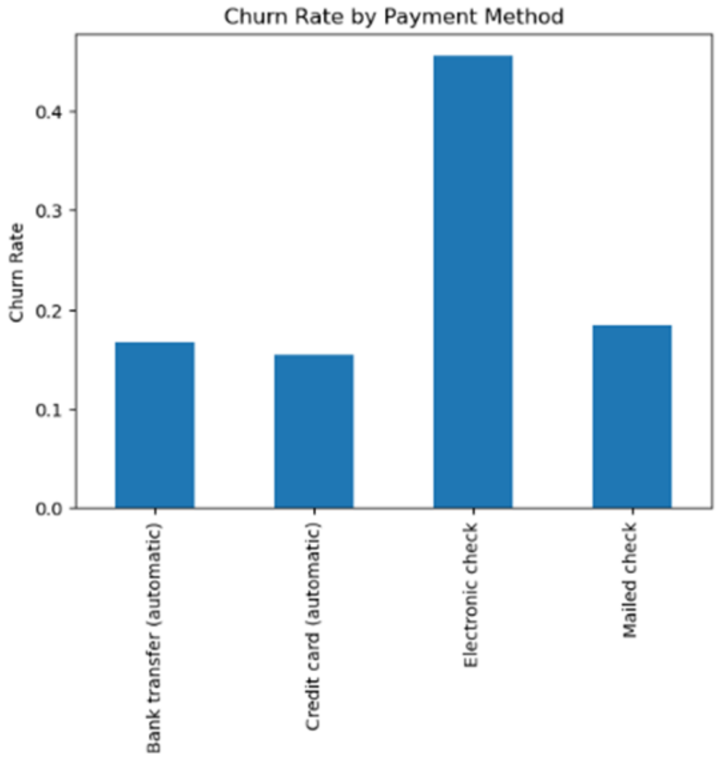
Purpose: This distribution plot's goal is to demonstrate how pricing exposure varies between churners and non-churners rather than just averages. It is appropriate to show the level of effect and hidden risk bands.

So what? Price alone is not the triggering factor once a customers cross into mid-range pricing. As customers who pay comparable amounts show a very diverse behavior (absolute prices are less important than experience, this includes overall service).

Business takeaway: For mid -price customers showing early danger signals, we can use initiate experience-based interventions. This can include service reviews, plan optimization and similar.

Fran: Plot 2

Churn vs One Predictor (Churn rate by Payment Method)



Customers that use electronic checks churn at a rate of about 45%, which is more than double the norm. Churn is significantly lower for automatic payments such as credit card and bank transfer.

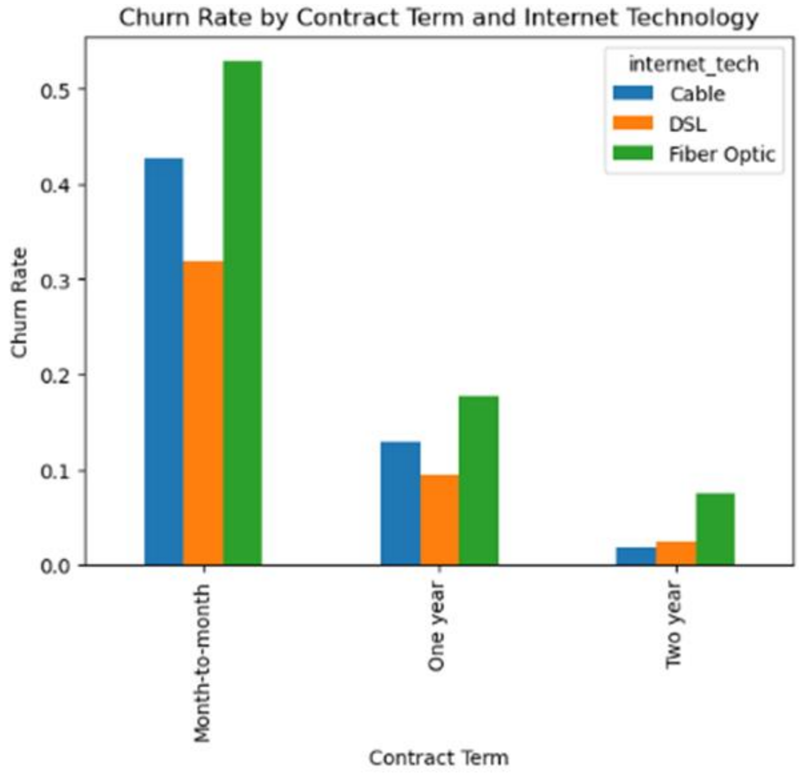
Purpose: Payment method is a behavioral variable, meaning that we can capture customer effort, inefficiency, and trust.

So what? We can say that this is not about payment preference; it is more about missed payment, higher cognitive effort, lower engagement, and trust. Not a huge or extremely high risk. With mail checks, the customers expect a delay. With electronic checks, customers expect instant confirmation, but ACH processing still takes days. Usually, electric checks fail more often due to typing errors in routing/account numbers, bank verification mismatches, insufficient fund delays, and longer clearing times.

Business takeaway: We can develop a focused “Auto-Pay Migration” initiative. This basically entails providing non-discount incentives, such as a free support month, bonus data, and other services.

Fran: Plot 3

Interaction Effects (Response + 2 Predictor) (Churn Rate by Contract Term x Internet Technology)



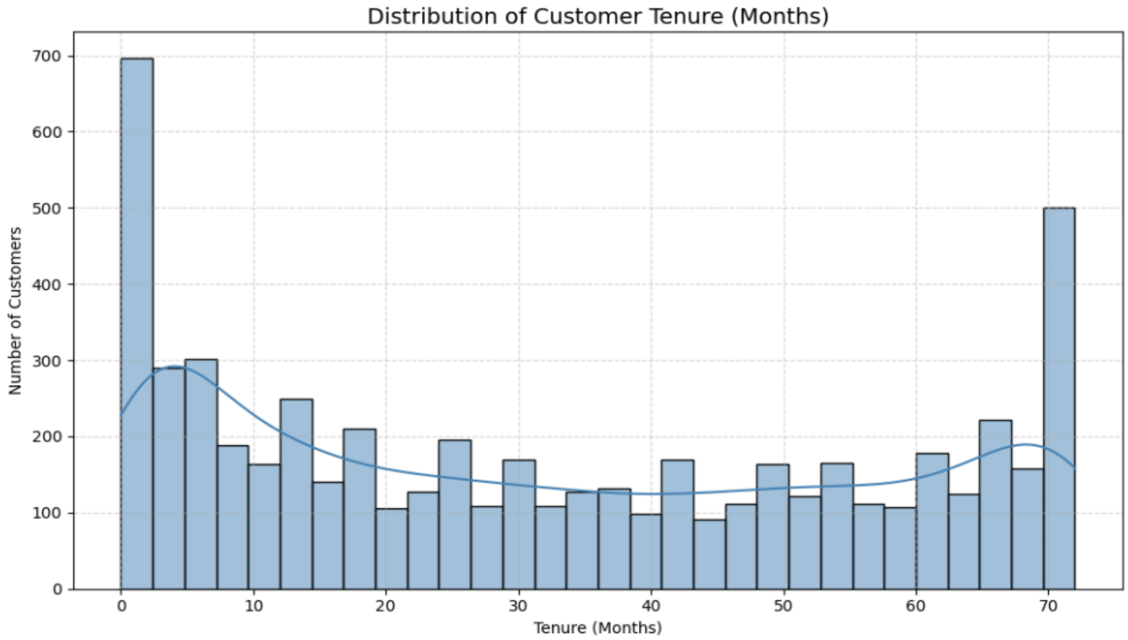
We can see different patterns such as put this in simple terms: Month-to-month plus fiber optic plans leave at very high rates, more than half of them churn (>50%). Even with one-year leases, fiber is still riskier than cable or DSL. Most technical disparities vanish on two-year contracts.

Purpose: This visualization reveals the weaknesses of conventional thinking. The length of a contract on its own is clear, but this shows what happens when technology is added.

So what? The visual shows us that month-to-month users of fiber are a small but critical risk segment. We should not treat fiber like other internet products, because fiber is not the problem itself; fiber without commitment is. Long-term contracts essentially anchor short-term fiber consumers, who frequently switch.

Business takeaway: This market is small, highly preventable, and costly to lose. Expectations, not price, are the reason why fiber consumers leave. Contracts reduce turnover but do not address discontent. We can improve this by adding proactive speed checks and service confirmation, dedicated fiber retention playbooks, and first 90-day onboarding.

Arnold: Plot 1

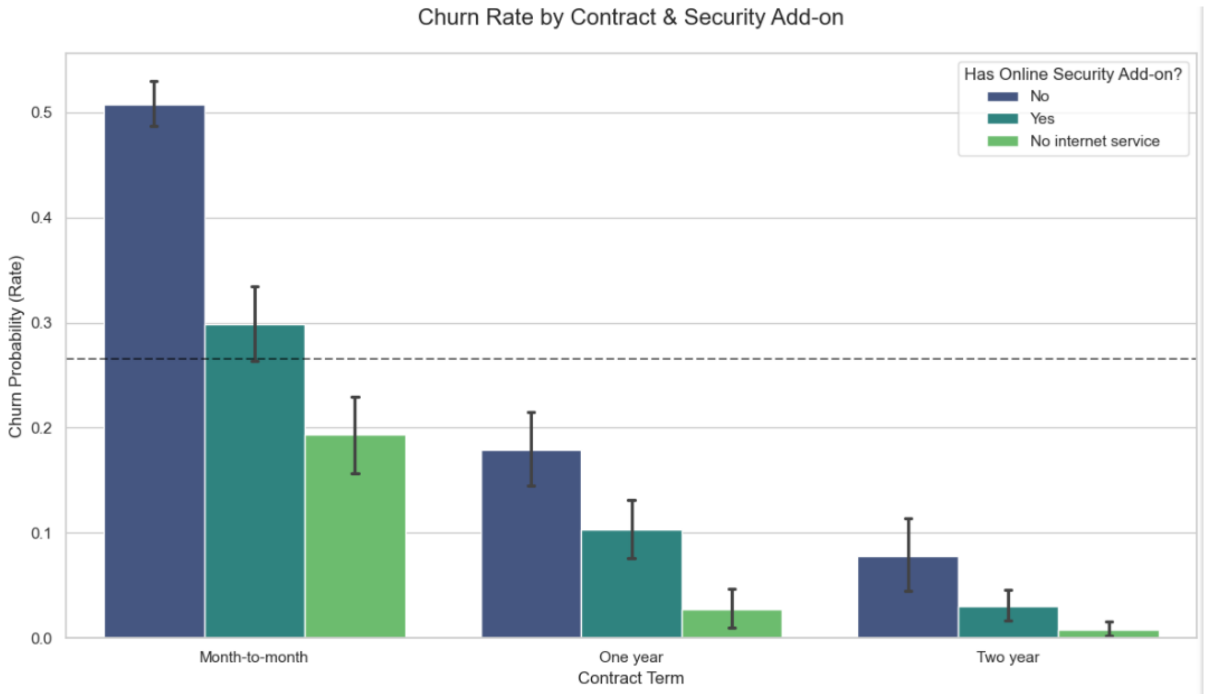


The peaks indicate where most customers fall i.e new vs long-term customers
A histogram with Kernel Density Estimate (KDE) curve is useful in understanding the distribution of a single numeric variable, in this case scenario, it depicts how long the customers have been with the company.

In any subscription business, tenure is one of the key predictors of churn. The distribution is vital because a high number of low-tenure customers usually signify early churn risk, a declining curve suggests that fewer customers survive long enough to become loyal and finally the long-tenure tail depicts a segment of highly loyal customers who are less likely to churn.

The business should identify and resolve early pain points i.e first-month complaints and billing confusion. The business needs to strengthen early life retention by offering loyalty incentives

Arnold: Plot 2

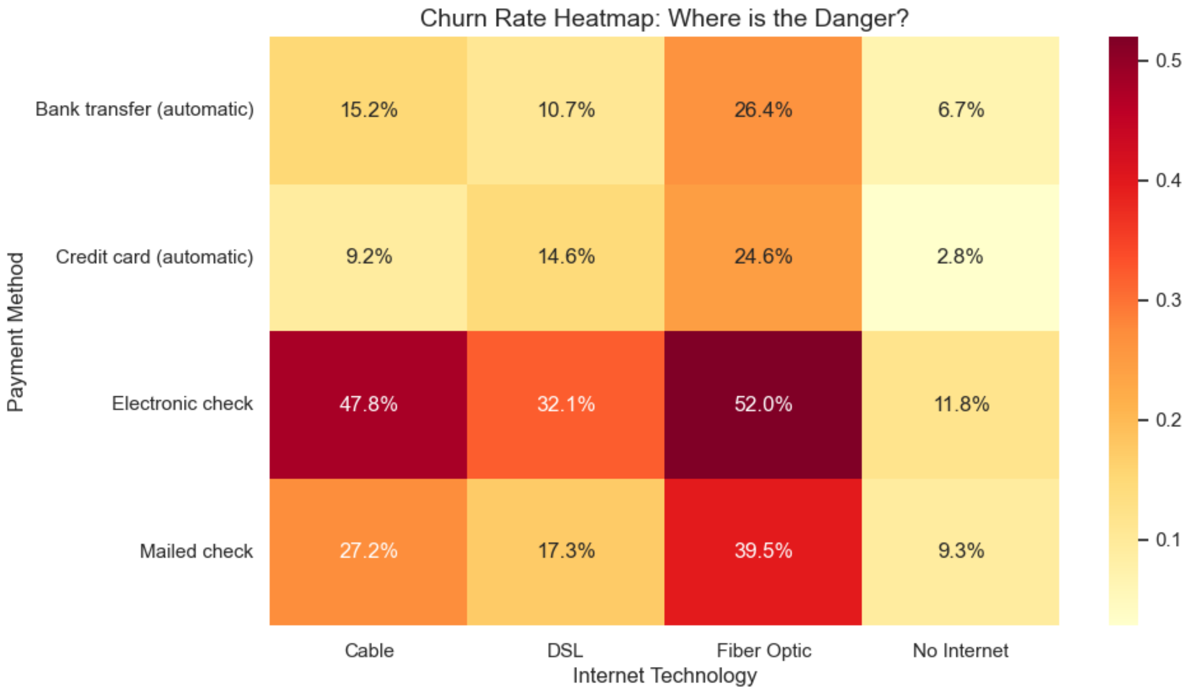


The clustered bar chart is ideal for depicting how two categorical variables, contract type and online security add-on, both interact to influence churn rate.

This is important as it pinpoints which customer segments are most vulnerable and which are most stable: i.e Month-to-month with no security is a high-risk churn segment

This chart shows that churn is highest for short-term, unsecured customers and the business should focus on upgrading contracts, bundling security, and protecting loyal segments to improve retention. The business should focus on providing loyalty rewards for 1-year and 2-year customers

Arnold: Plot 3



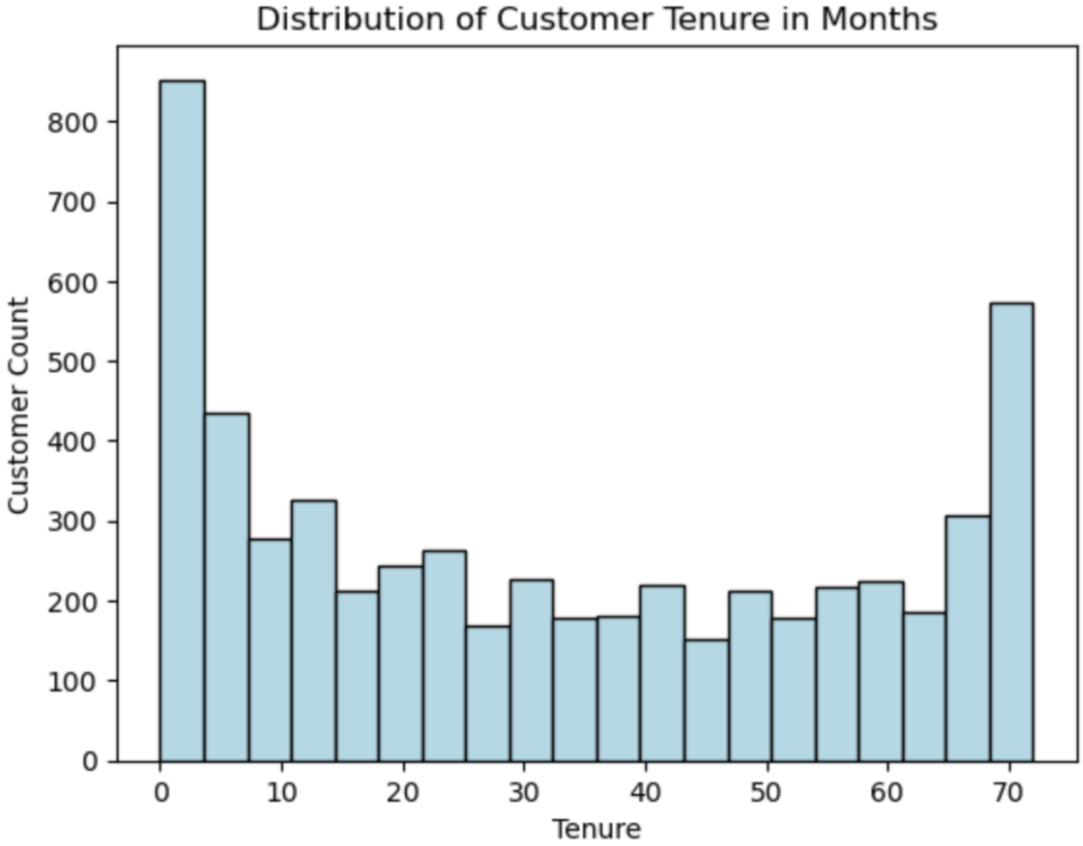
A Heatmap is a distinct way to show the relationship between the three variables, payment method, internet technology and churn

It creates an immediate "danger zone" visual. The darker/redder the square, the higher the risk. It shows exactly which combination of tech and payment is the worst.

It clearly isolates the "Red Zone": Fiber Optic customers using Electronic Checks (52% churn). It proves that the payment method can neutralize some of the risks associated with high-churn technologies like Fiber. Automatic payment methods like credit card and bank transfers records lower churn rates as it has the component of "locking the customers in"

The business should implement a mandatory Auto-Pay policy for new Fiber Optic installations. By forcing the technology to be paired with a "Green Zone" payment method, the company could theoretically cut the Fiber Optic churn rate in half.

Saren: Plot 1

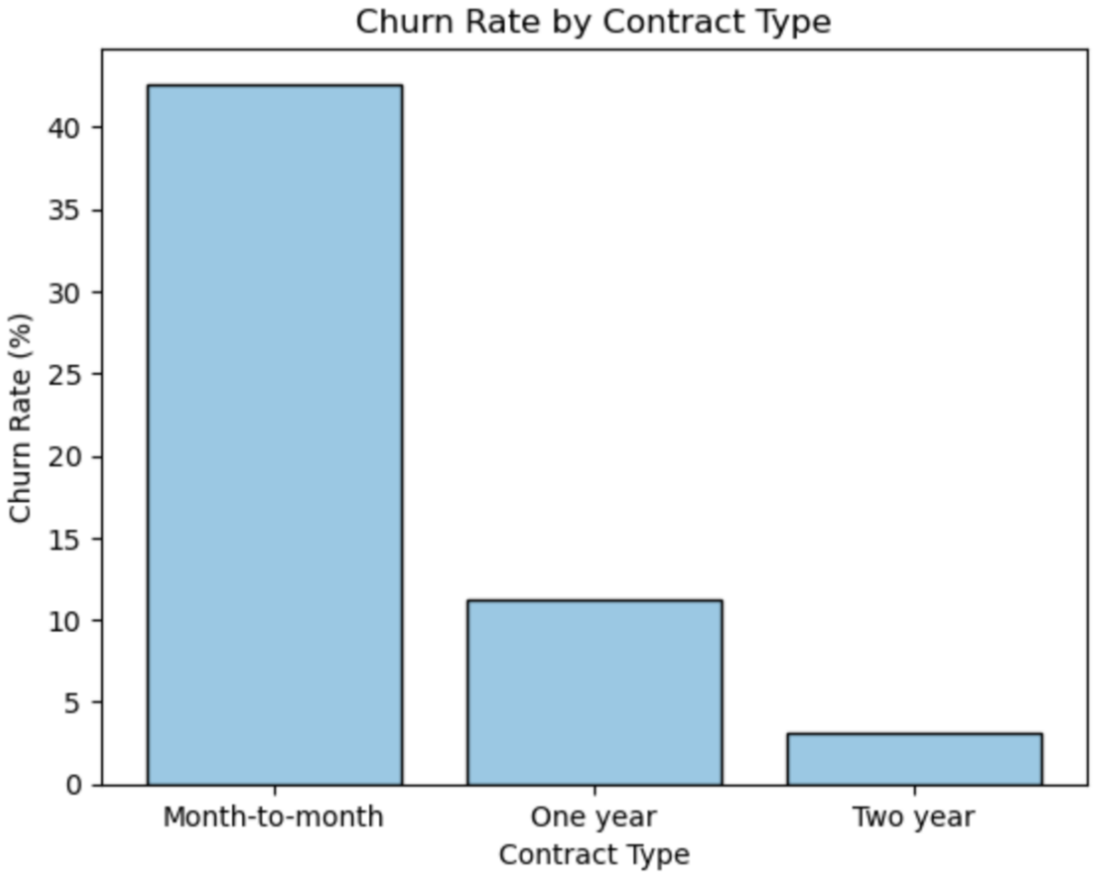


Purpose: This chart shows how long customers stay with the company. The U-shape indicates that majority of customers are new to the company (0-5 months), or they are very loyal (70+ months) while the rest fall somewhere in between.

So what? This chart shows that the highest risk of churning is in the first five months of obtaining a new customer. If a customer makes it past their first year, they are much more likely to stay long-term. Because so many customers leave before their first year, the company is losing customers before they become profitable.

Business takeaway: The company should introduce a retention program aimed at customers in the first three months of their membership. Focusing on onboarding and making sure that customers are prepared for their first bill will increase customer satisfaction and retention.

Saren: Plot 2

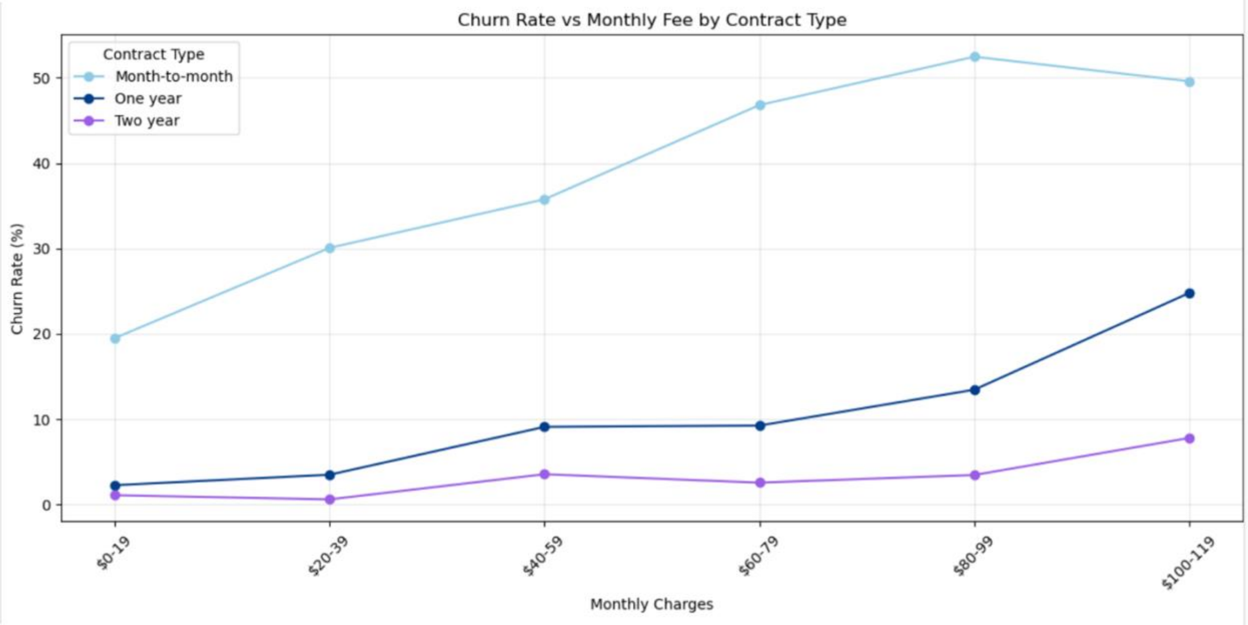


Purpose: This bar chart compares the percentage of customers canceling their service across different contract lengths. It’s unsurprising that month-to-month customers churn at a much higher rate than one-year and two-year contracts.

So what? This plot confirms that the flexibility offered by month-to-month contracts, which may initially entice people to sign up for phone services, are ineffective at retaining customers because they can leave very easily if they are dissatisfied with the service in any way or are tempted with a better offer from another company.

Business takeaway: The business should incentivize month-to-month customers to transition to longer-term plans with offers such as discounts or free upgrades. Getting customers to sign up for a long-term contract is the most effective way to prevent them from switching to a different phone service with another company.

Saren: Plot 3

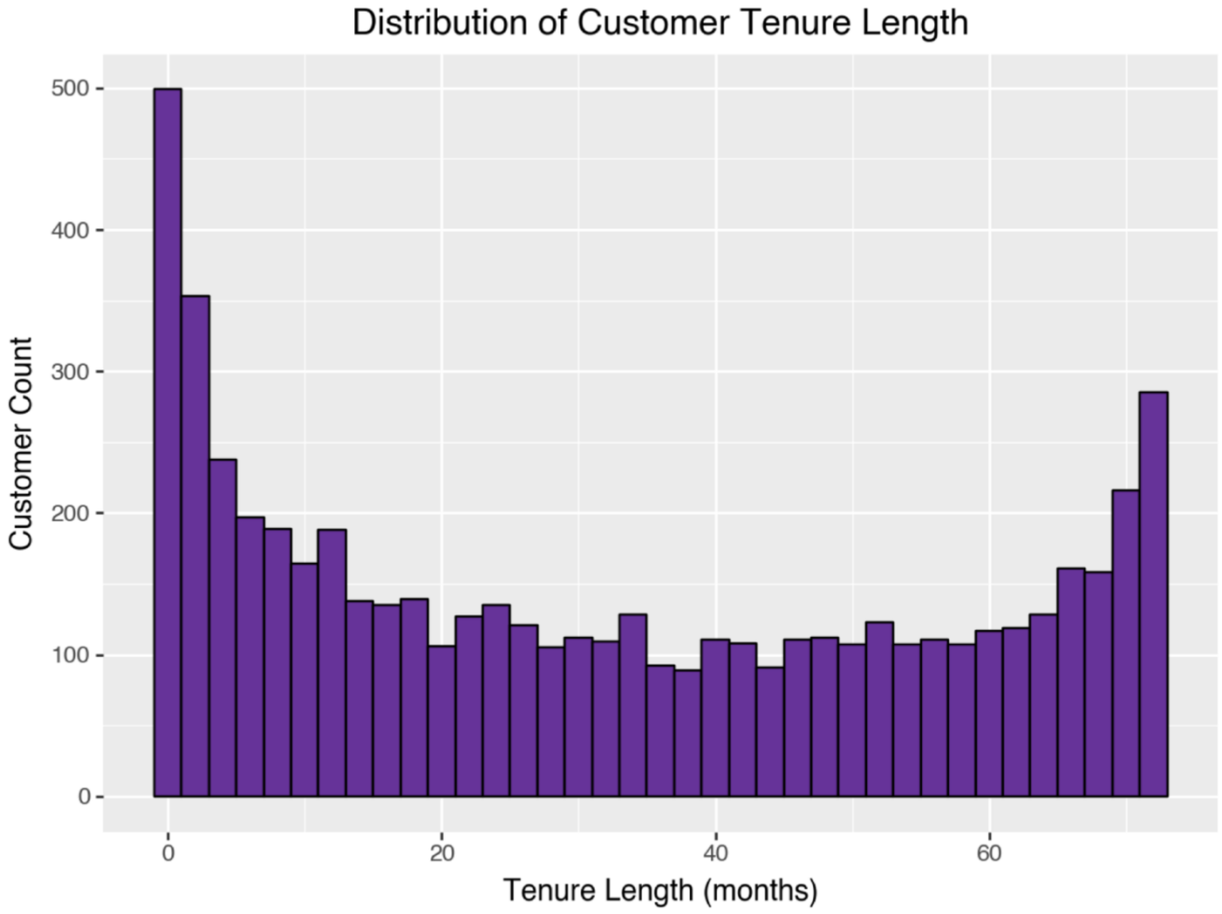


Purpose: This line chart looks at how customers' monthly fees affect churn rates across different contract types. The pattern shows that as the monthly bill goes up so does the churn rate, but this trend is much more stark for month-to-month customers.

So what? High monthly fees, particularly over $60, are a primary trigger for customers to cancel their service, particularly if they aren't in a long-term contract. It shows that the company's most expensive plans are also the least likely to retain customers.

Business takeaway: The company should review the value proposition of their most expensive plans. Customers paying more than $60 a month are at risk of churning. The business should offer these high-value customers additional discounts for loyalty or include additional services to justify the high price point of their plans.

Willow: Plot 1

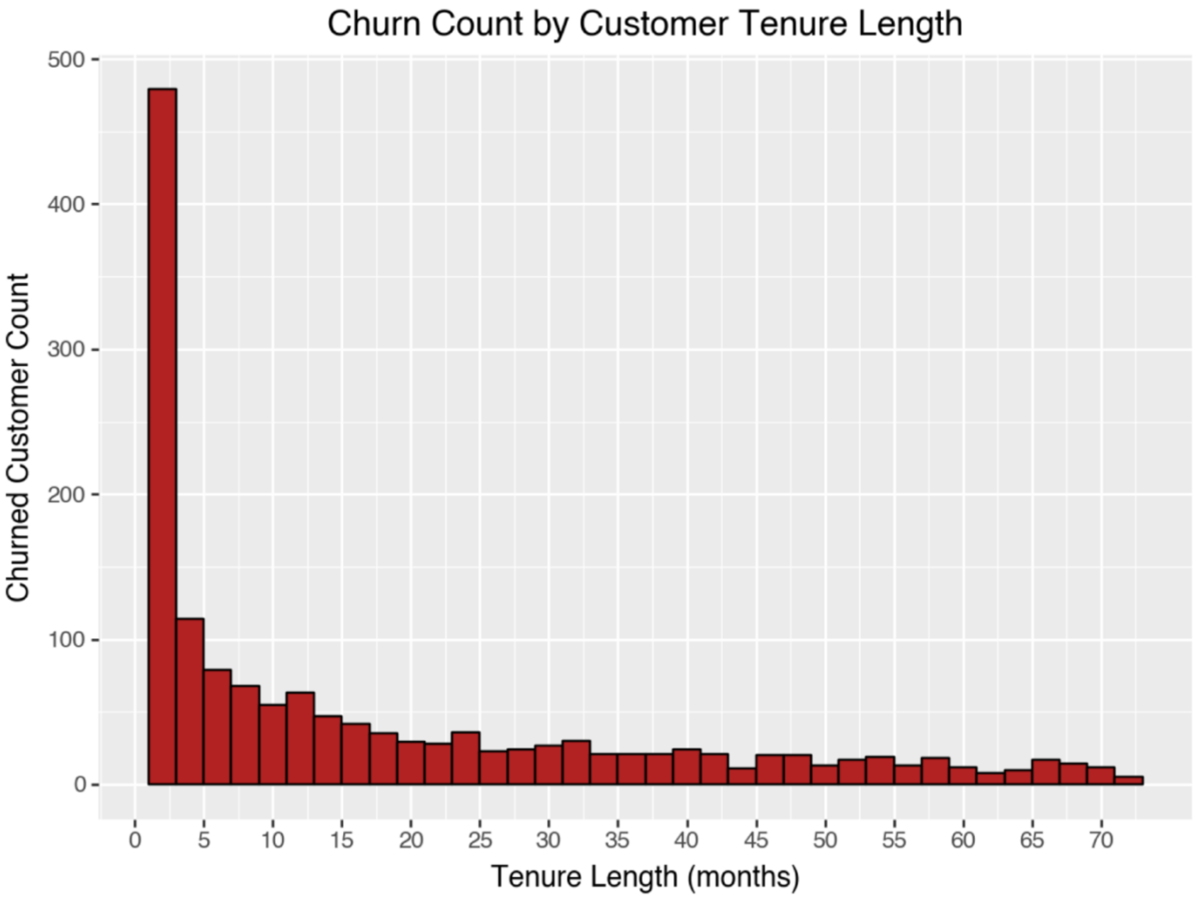


Purpose: The chart's pattern shows that the highest number of customers have a short tenure, and a decreasing number of customers stay for a longer tenure. However, this trend reverses once the tenure length reaches approximately 40 months.

So what? This matters for customer retention because it shows that the highest proportion of customers have a tenure length of 1 month, implying that a high number of customers churn after 1 month of service. The smallest proportion of customers have a tenure in the range of approximately 20-40 months, implying that customers leave around that point. However, after about 40 months of service customers more customers stay.

Business takeaway: Given these findings, the business should introduce incentives for customers with longer tenure, specifically those whose tenure falls in the range of approximately 20-40 months, to prevent customer churn for those whose tenure falls in that range.

Willow: Plot 2

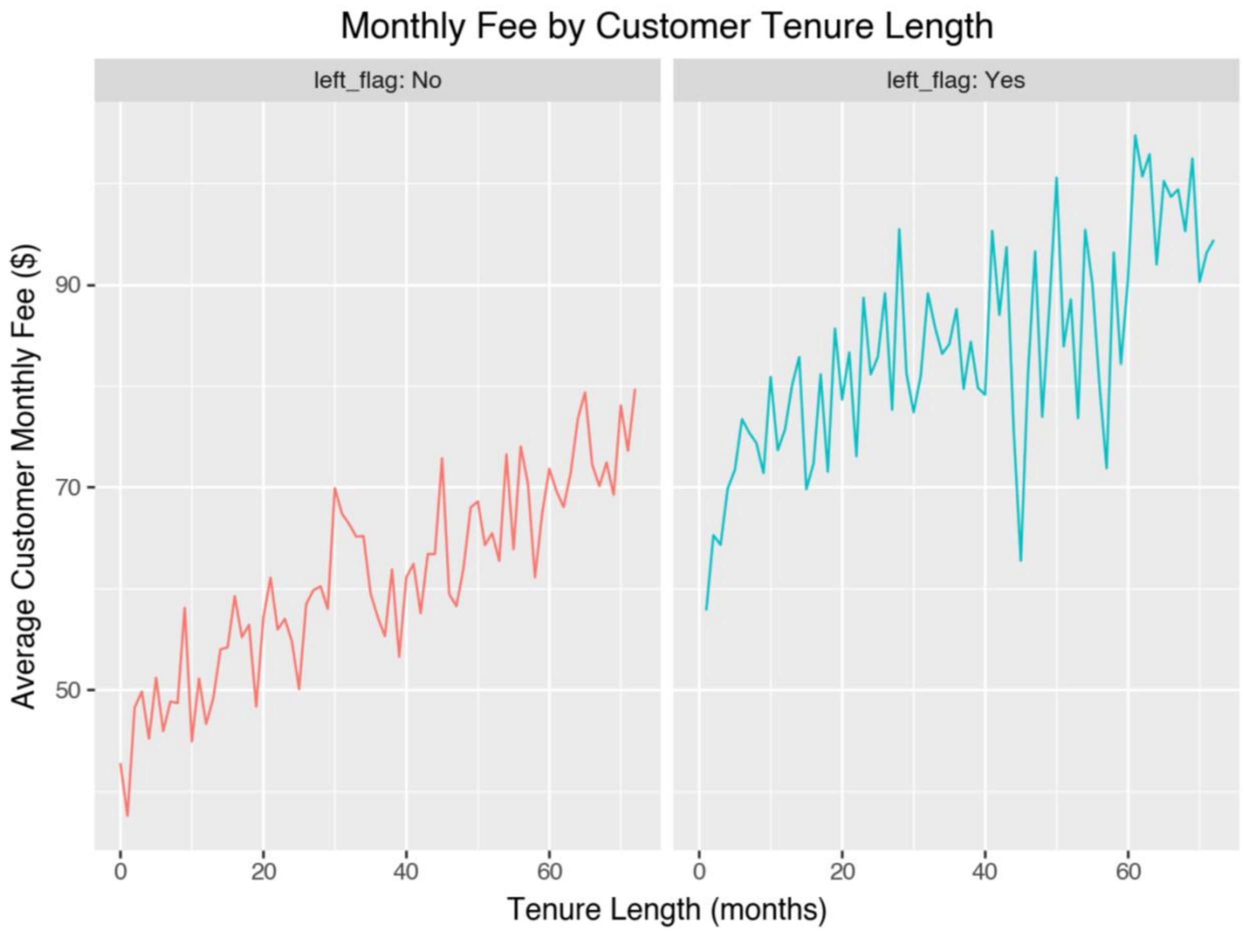


Purpose: This chart shows that the highest number of churns occur within the first month of a customer's tenure, and the number of churns per month decreases sharply after that.

So what? This is crucial for customer retention because it shows that customer retention is most important in the first month of service. After 1 month of service, customers are much less likely to churn.

Business takeaway: Because a customer is the most likely to churn after the first month of service, the business needs to discern what factors are causing customers to churn specifically after 1 month and address them.

Willow: Plot 3



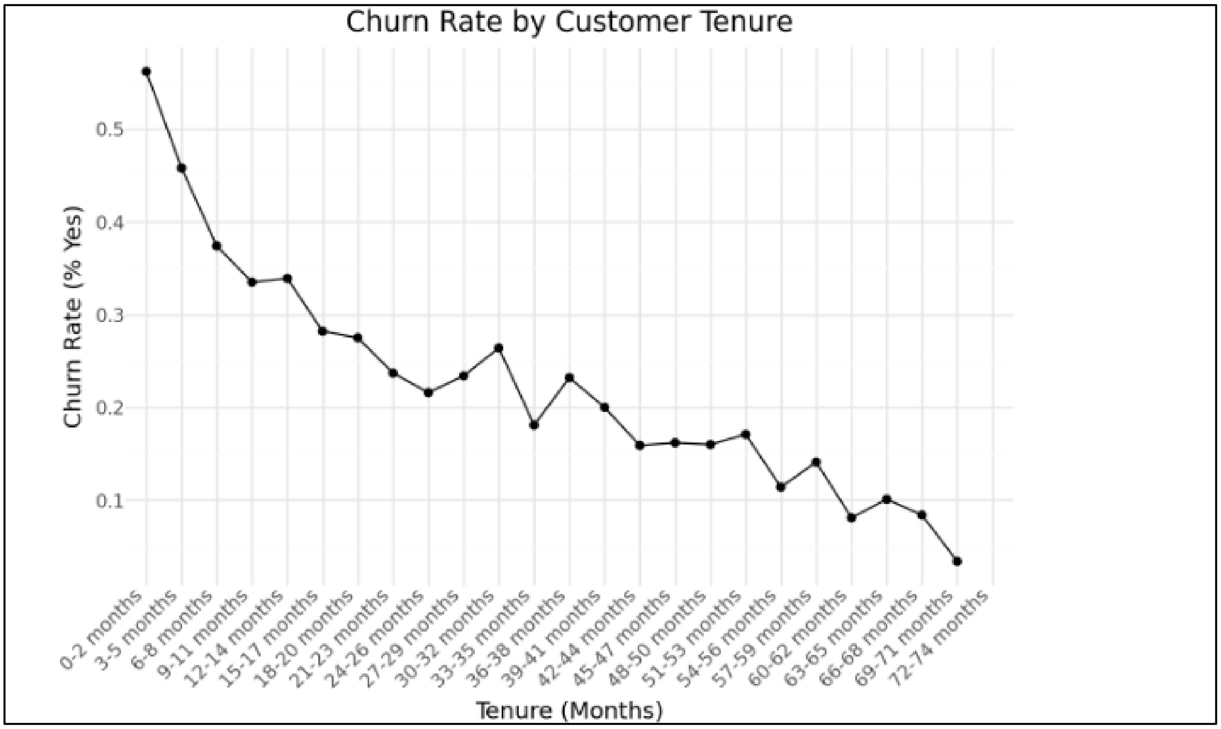
Purpose: The chart shows that monthly fees increase with a longer tenure, and customers who churn tend to have higher monthly fees.

So what? This matters for customer retention because it shows that all customers have higher fees over time, and the customers with higher average fees are more likely to churn.

Business takeaway: The business should minimize monthly fees for customers with higher tenures. Although the reason for increased monthly fees is likely related to the addition of other services, the total monthly cost may still be a significant reason for customer churn.

Recommendations for Actionable Insights

Plot 1:

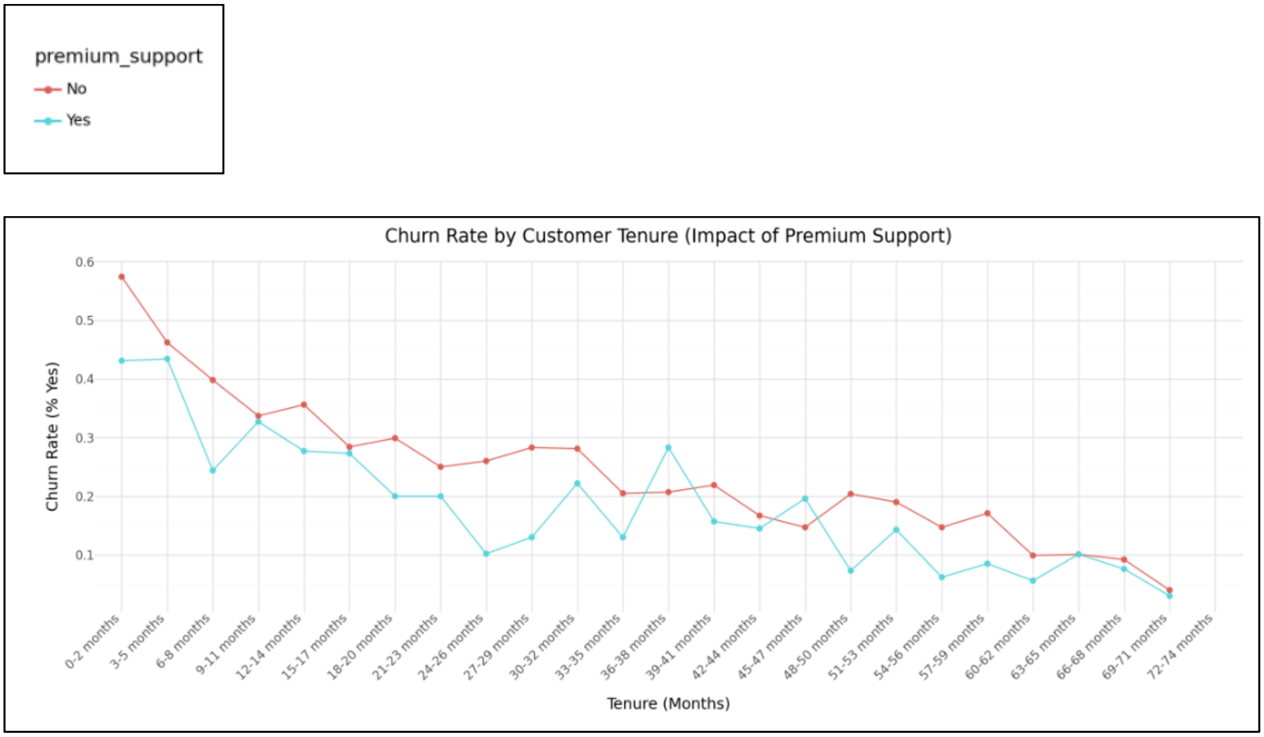


Purpose: This chart shows how customer churn rates change over time as monthly tenure increases.

So what? Churn is an early lifecycle problem. It drops from about 56% in the first 3 months to about 37% after 6 months. So, this shows that new customers are far more likely to leave in first 6-8 months and after which churn stabilizes past year one at much lower levels, suggesting customers who stay past early tenure are increasingly loyal.

Business takeaway: The company should focus on retention efforts on the first few months of the customer lifecycle, where churn risk is highest. A better onboarding experience and customer support in early lifecycle can help in reducing churn.

Plot 2:

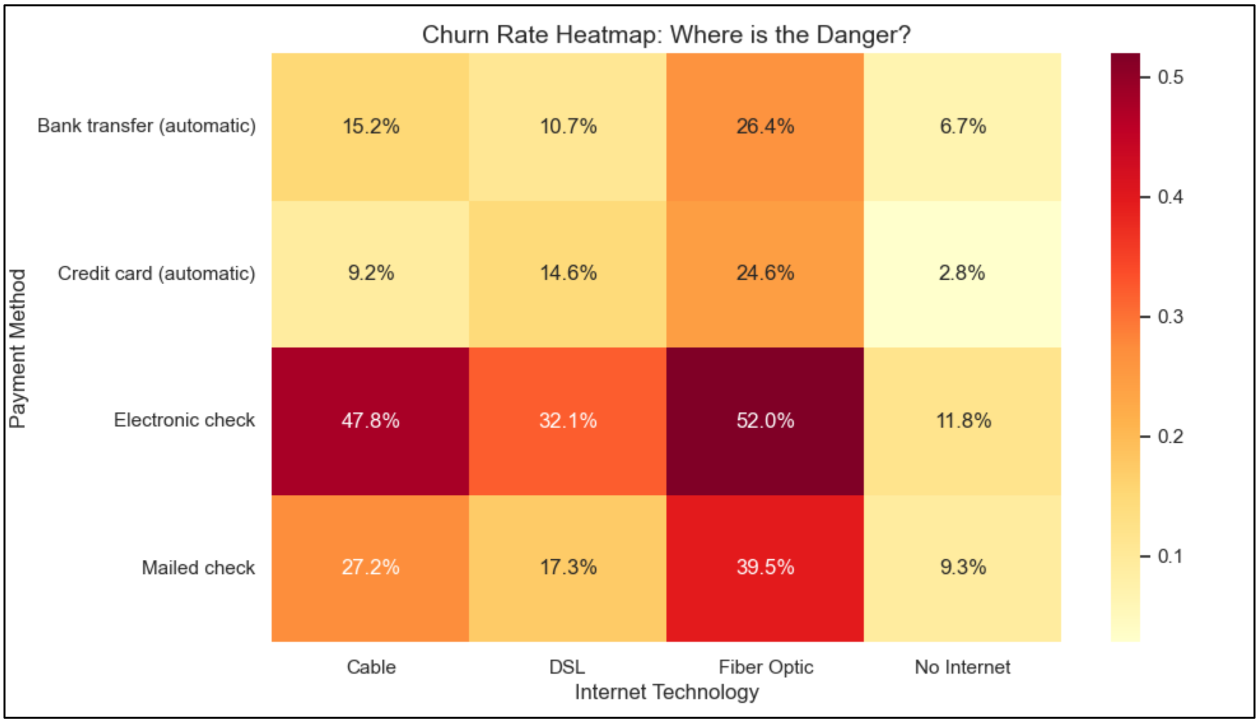


Purpose: This chart shows how churn rate changes across customer tenure in months, comparing customers with and without premium support.

So what? Churn is highest in the early months and stabilizes past year one, but customers without premium support consistently show higher churn across lifecycle. The impact of premium services on churn rate is highest in the first 12-24 months.

Business takeaway: Business should offer premium support during onboarding customers at early lifecycle, perhaps as a free trial or an onboarding offer. This can reduce churn and improve retention rate significantly.

Plot 3:

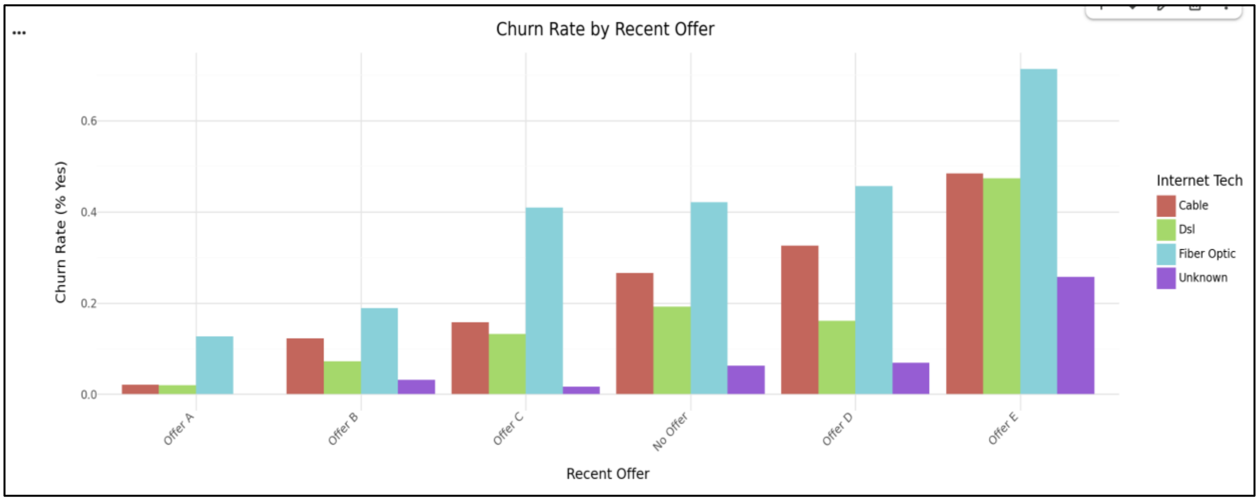


Purpose: This heatmap shows how customer churn rates differ across combinations of payment methods and internet technology. It highlights where churn risk is highest when payment behavior overlaps with internet technology type of the customer.

So what? Customers paying via electronic check show the highest churn across all internet technologies, with the most severe churn among Fiber Optic users. Automatic payment methods are associated with substantially lower churn regardless of technology.

Business takeaway: The highest churn risk segment is the Fiber Optic customers using electronic checks. The business should prioritize migrating these customers to automatic payment methods to reduce churn and retain this customer segment.

Plot 4:



Purpose: The chart evaluates the effectiveness of different offers on churn rate segmented by the user's internet technology type.

So what? Considering no-offer as our control group, we see offer E attracts the worst type of customers across all internet technologies while offers A and B work well in reducing churn rate across all users. Offer D is only effective for DSL users, and Offer C has no impact on Fiber Optic users. It also suggests that business needs to target offers carefully and keep the internet type segmentation in mind.

Business takeaway: Business should consider discontinuing offer E as it shows worse performance than the no-offer control group. Offer D should also be re-evaluated for all segments except DSL users where it shows some improvement in churn rate and Offer C should be re-evaluated for Fiber Optic users.