Telecom Customer Churn Analysis

Quek Hong Rui

2 August 2024

Contents

Introduction

This data analysis project investigates churn for a fictional telecom company based in California. Data on telecom churn was obtained here[1] *(navigate to Telecom Customer Churn)*. This is a data set of **7,043** customers of the telecom company; and contains information on each customer's demographics, area of residence, contracts and subscriptions as well as whether the customer stayed, churned or joined at the end of Q2 2022. A snippet of the data set is shown below.

```
## Rows: 7,043
## Columns: 38
## $ ‘Customer ID‘                       <chr> "0002-ORFBO", "0003-MKNFE", "0004-~
## $ Gender                              <chr> "Female", "Male", "Male", "Male", ~
## $ Age                                 <int> 37, 46, 50, 78, 75, 23, 67, 52, 68~
## $ Married                             <chr> "Yes", "No", "No", "Yes", "Yes", "~
## $ ‘Number of Dependents‘              <int> 0, 0, 0, 0, 0, 3, 0, 0, 0, 1, 0, 2~
## $ City                                <chr> "Frazier Park", "Glendale", "Costa~
## $ ‘Zip Code‘                          <int> 93225, 91206, 92627, 94553, 93010,~
## $ Latitude                            <dbl> 34.82766, 34.16251, 33.64567, 38.0~
## $ Longitude                           <dbl> -118.9991, -118.2039, -117.9226, -~
## $ ‘Number of Referrals‘               <int> 2, 0, 0, 1, 3, 0, 1, 8, 0, 3, 0, 2~
## $ ‘Tenure in Months‘                  <int> 9, 9, 4, 13, 3, 9, 71, 63, 7, 65, ~
## $ Offer                               <chr> "None", "None", "Offer E", "Offer ~
## $ ‘Phone Service‘                     <chr> "Yes", "Yes", "Yes", "Yes", "Yes",~
## $ ‘Avg Monthly Long Distance Charges‘ <dbl> 42.39, 10.69, 33.65, 27.82, 7.38, ~
## $ ‘Multiple Lines‘                    <chr> "No", "Yes", "No", "No", "No", "No~
## $ ‘Internet Service‘                  <chr> "Yes", "Yes", "Yes", "Yes", "Yes",~
## $ ‘Internet Type‘                     <chr> "Cable", "Cable", "Fiber Optic", "~
## $ ‘Avg Monthly GB Download‘           <int> 16, 10, 30, 4, 11, 73, 14, 7, 21, ~
## $ ‘Online Security‘                   <chr> "No", "No", "No", "No", "No", "No"~
## $ ‘Online Backup‘                     <chr> "Yes", "No", "No", "Yes", "No", "N~
## $ ‘Device Protection Plan‘            <chr> "No", "No", "Yes", "Yes", "No", "N~
## $ ‘Premium Tech Support‘              <chr> "Yes", "No", "No", "No", "Yes", "Y~
## $ ‘Streaming TV‘                      <chr> "Yes", "No", "No", "Yes", "Yes", "~
## $ ‘Streaming Movies‘                  <chr> "No", "Yes", "No", "Yes", "No", "Y~
## $ ‘Streaming Music‘                   <chr> "No", "Yes", "No", "No", "No", "Ye~
## $ ‘Unlimited Data‘                    <chr> "Yes", "No", "Yes", "Yes", "Yes", ~
## $ Contract                            <chr> "One Year", "Month-to-Month", "Mon~
## $ ‘Paperless Billing‘                 <chr> "Yes", "No", "Yes", "Yes", "Yes", ~
## $ ‘Payment Method‘                    <chr> "Credit Card", "Credit Card", "Ban~
## $ ‘Monthly Charge‘                    <dbl> 65.60, -4.00, 73.90, 98.00, 83.90,~
## $ ‘Total Charges‘                     <dbl> 593.30, 542.40, 280.85, 1237.85, 2~
```

[1] Contents include a two data sets and a data dictionary `telecom_data_dictionary.csv` to explain the variables in each data set. The data set used in this analysis is `telecom_customer_churn.csv`.

```
## $ ‘Total Refunds‘                   <dbl> 0.00, 38.33, 0.00, 0.00, 0.00, 0.0~
## $ ‘Total Extra Data Charges‘        <int> 0, 10, 0, 0, 0, 0, 0, 20, 0, 0, 0,~
## $ ‘Total Long Distance Charges‘     <dbl> 381.51, 96.21, 134.60, 361.66, 22.~
## $ ‘Total Revenue‘                   <dbl> 974.81, 610.28, 415.45, 1599.51, 2~
## $ ‘Customer Status‘                 <chr> "Stayed", "Stayed", "Churned", "Ch~
## $ ‘Churn Category‘                  <chr> NA, NA, "Competitor", "Dissatisfac~
## $ ‘Churn Reason‘                    <chr> NA, NA, "Competitor had better dev~
```

Through analysis, this project aims to distinguish customers who churn from those who do not churn, so that the telecom company can identify prospective churners. It also aims to create actionable insights on the possible drivers of churn specific to this company. By adopting this two-pronged approach, this project seeks to reduce the telecom company's churn rate and retain its customers in the long run.

This analysis consists of three main sections:

1. **Data Validation and Cleaning** - to check if the data set is ready for analysis and rectify any problematic data values.
2. **Exploratory Data Analysis and Data Visualization** - to analyse data and illustrate customer profiles and drivers of churn.
3. **Final Recommendations** — to close off with business recommendations to minimize churn and improve the company's current business situation.

All steps of analysis are performed in `R`, with the use of `dplyr` and `ggplot2` packages. For presentation purposes, programming code has been excluded from this report. Refer to the `R Markdown` file for the full code.

Data Validation

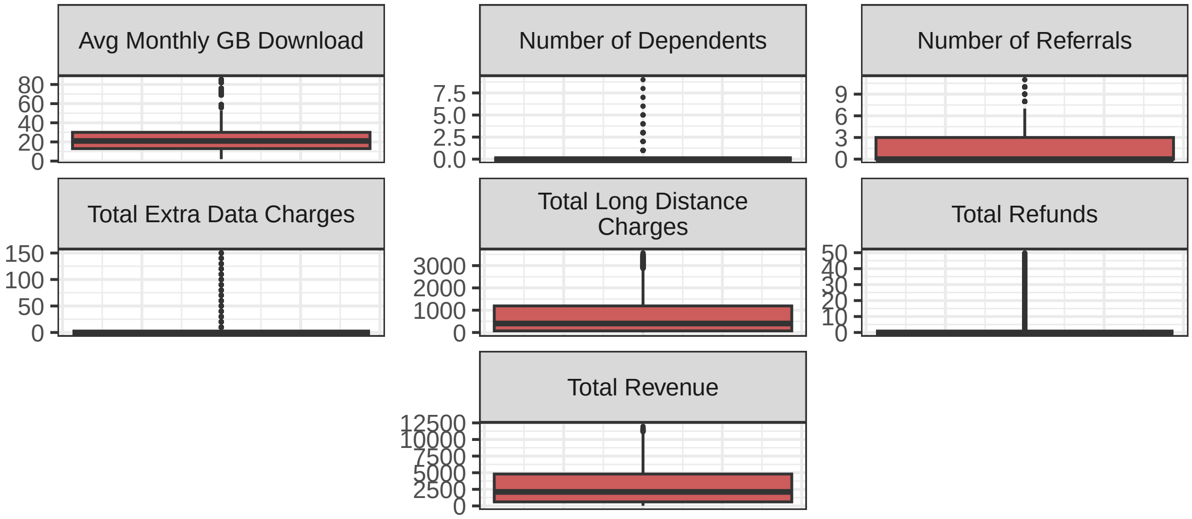
Overview

Some common data problems include anomalies, missing values, incorrect data types, outliers[2], structural errors *eg spelling and case inconsistencies, leading/trailing white spaces.* Data checks will be run in `R` to check if this data set faces any of these issues. We also check the distribution of numerical data in the data set.

Table 1: Summary statistics for numeric columns in data set

Numeric Column	Min.	1st Quartile	Median	Mean	3rd Quartile	Max.
Age	19.00000	32.00000	46.00000	46.5097260	60.00000	80.00000
Number of Dependents	0.00000	0.00000	0.00000	0.4686923	0.00000	9.00000
Zip Code	90001.00000	92101.00000	93518.00000	93486.0705665	95329.00000	96150.00000
Latitude	32.55583	33.99065	36.20546	36.1974548	38.16132	41.96213
Longitude	-124.30137	-121.78809	-119.59529	-119.7566837	-117.96980	-114.19290
Number of Referrals	0.00000	0.00000	0.00000	1.9518671	3.00000	11.00000
Tenure in Months	1.00000	9.00000	29.00000	32.3867670	55.00000	72.00000
Avg Monthly Long Distance Charges	1.01000	13.05000	25.69000	25.4205172	37.68000	49.99000
Avg Monthly GB Download	2.00000	13.00000	21.00000	26.1899583	30.00000	85.00000
Monthly Charge	-10.00000	30.40000	70.05000	63.5961309	89.75000	118.75000
Total Charges	18.80000	400.15000	1394.55000	2280.3812637	3786.60000	8684.80000
Total Refunds	0.00000	0.00000	0.00000	1.9621823	0.00000	49.79000
Total Extra Data Charges	0.00000	0.00000	0.00000	6.8607128	0.00000	150.00000
Total Long Distance Charges	0.00000	70.54500	401.44000	749.0992617	1191.10000	3564.72000
Total Revenue	21.36000	605.61000	2108.64000	3034.3790558	4801.14500	11979.34000

From **Table 1**, there appears to be outliers in some numerical columns. Each box plot below shows the data distribution in each of these columns and confirms the presence of outliers.



Note: Dots represent outliers based on Tukey's definition

Another step of the data validation stage is to verify the integrity of the `Total Revenue`

[2] Tukey's definition — values in the column below **Q1 - 1.5 * IQR** or above **Q3 * 1.5 * IQR**, where **Q1**, **Q3** and **IQR** refers to the 1st quartile, 3rd quartile and interquartile range of the values in the column.

column by checking if `Total Revenue` = `Total Charges` + `Total Extra Data Charges` + `Total Long Distance Charges` - `Total Refunds` for all rows. This formula was provided in the data dictionary. It appears that all values in this column have been computed correctly.

Through these preliminary data checks, data quality was evaluated and the results can be found in **Table A1** in the Appendix. Additionally, all columns in the data set have appropriate data types, so there is no need to perform data type conversion.

Further Checks

Missing Values

Upon further inspection, it is found that the missing values in the data set do not occur randomly.

- The **682** missing values in the `Avg Monthly Long Distance Charges` and `Multiple Lines` columns come from the same observations in the data set, and correspond to the 682 customers whose `Phone Service` value is **No** *ie does not subscribe to home phone services with the telecom company.*
- The **1,526** missing values in the `Internet Type`, `Avg Monthly GB Download`, `Online Security`, `Online Backup`, `Device Protection Plan`, `Premium Tech Support`, `Streaming TV`, `Streaming Movies`, `Streaming Music`, `Unlimited Data` columns also come from the same 1,526 observations where `Internet Service` is **No** *ie does not subscribe to Internet service with the telecom company.*
- The **5,174** missing values in the `Churn Category` and `Churn Reason` columns come from the same 5,174 customers in the data set where `Customer Status` is not **Churned** *ie stayed with or joined the telecom company at the end of Q2 2022.*

In conclusion, the presence of missing values in this data set is at worst a discrepancy between the data dictionary and the data itself and is generally not problematic.

- In the data set, when `Phone Service` is **No**, `Avg Monthly Long Distance Charges` and `Multiple Lines` have missing values as opposed to **0** and **No**, which are the values stated in the data dictionary.
- When `Internet Service` is **No**, the values in the `Internet Type`, `Avg Monthly GB Download`, `Online Security`, `Online Backup`, `Device Protection Plan`, `Premium Tech Support`, `Streaming TV`, `Streaming Movies`, `Streaming Music`, `Unlimited Data` will be missing as opposed to **0**, **No** or **None** as stated in the data dictionary.

- In the data set, the `Churn Category` and `Churn Reason` columns only have missing values when `Customer Status` is not **Churned**. This is valid since the customer will not have any category or reason for churn that is applicable as he or she did not even churn.

Overall, the missing values in this data set appear consistently in a way that is logical and easily understandable. As such, these missing values do not require further cleaning.

Outliers

Conventionally in data science, outliers are problems in data that have to be handled before analysis as they represent anomalous or noisy data. Upon further inspection of the outliers, in this case however, it might be more ideal to leave them untouched.

Table 2: Outliers count and proportion in columns with outliers

Column Name	Number of Outliers	Proportion of Outliers
Number of Dependents	1627	23.1%
Number of Referrals	676	9.6%
Avg Monthly GB Download	649	9.21%
Total Extra Data Charges	728	10.34%
Total Long Distance Charges	196	2.78%
Total Refunds	525	7.45%
Total Revenue	21	0.3%

From **Table 2** above, outliers form a significant proportion of data values in certain columns like `Number of Dependents` and `Total Extra Data Charges`. If outliers in these columns are dropped or replaced, there is a risk of transforming the original data set too much and affecting the results of analysis downstream.

Furthermore, it might be useful to consider these extreme cases in the data set as they might not necessarily be erroneous data. For example, an observation with an outlying `Number of Referrals` value of **11** might represent an extremely satisfied customer who is keen on recommending the telecom company. In this context, the outliers might have some meaning to them. With these considerations in mind, the outliers were left untreated.

Negative Values

Based on **Table 1**, in the `Monthly Charges` column, there are some negative values, as suggested by the column's negative minimum value. Having negative values for total monthly charge is counter-intuitive and could be a sign of incorrect data entry or that the customer received a monthly refund instead. As the `Monthly Charges` column is rather ambiguous, it will be excluded from analysis. This is unlikely a major loss of data as the `Total Charges` also provides information on customer charges and is in fact, a better choice because it contains clean data as shown in **Table A1**.

Summary

Overall, the data values in the data set generally align with the descriptions listed in the data dictionary and is rather clean, with zero to little need for additional data cleaning steps.

In the next section, analysis is carried out on the validated data.

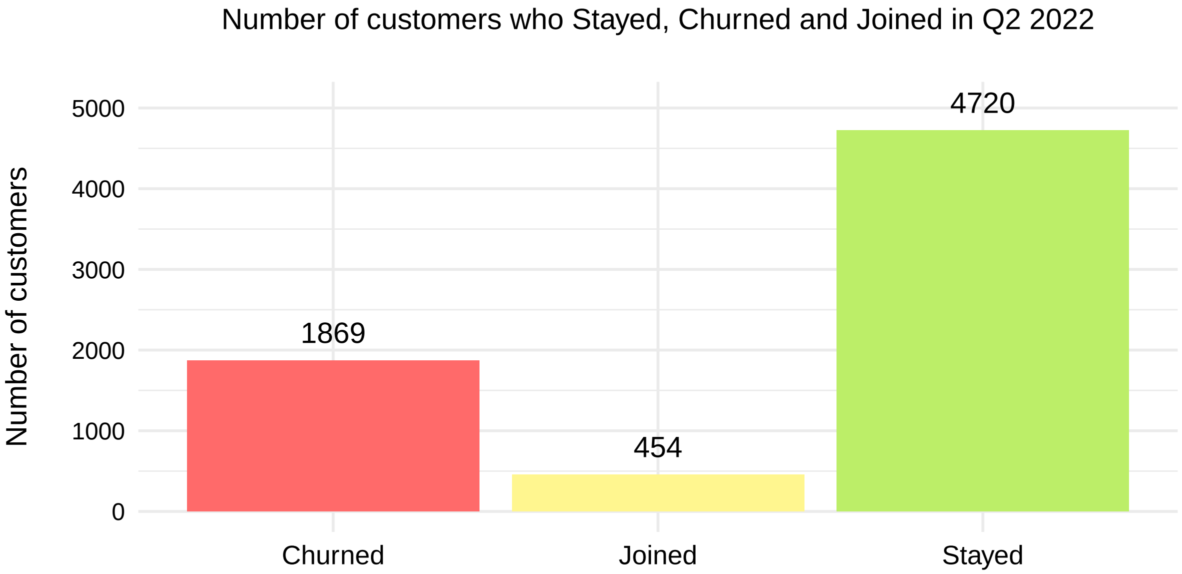
Customer Profiling

Now that the data has been validated and verified ready to be used for analysis, exploratory data analysis will be performed on the data set to segment and profile the customers. *Think*: What characteristics does a Churned customer have? How do they differ from Stayed customers?

Following which, the identified differences will be presented through data visuals, and hypotheses will be proposed to unpack and make sense of these differences.

Overview

Among the 7,043 customers in this data set, there are **1,869** who churned, **454** who just joined and **4,720** who stayed in Q2 2022.

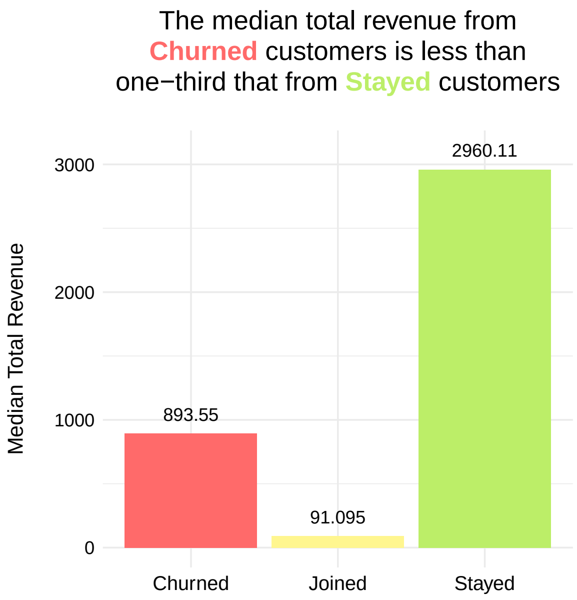


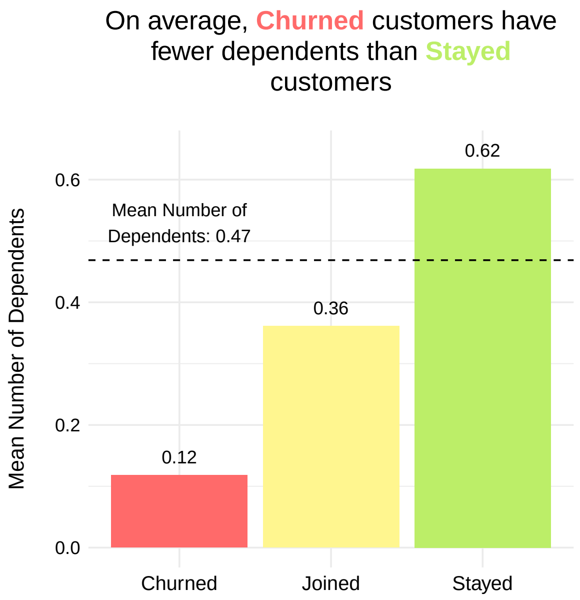
Note: Even though the data set contains customers from the Joined `Customer Status` *category, this analysis will mainly focus on the Churned and Stayed customers since the objective is to help reduce the company's churn rate and retain its Churned customers, as opposed to targeting a more 'neutral' group like the Joined customers. Also, not all variables or columns in the data set will be used for analysis — only appreciable ones are used.*

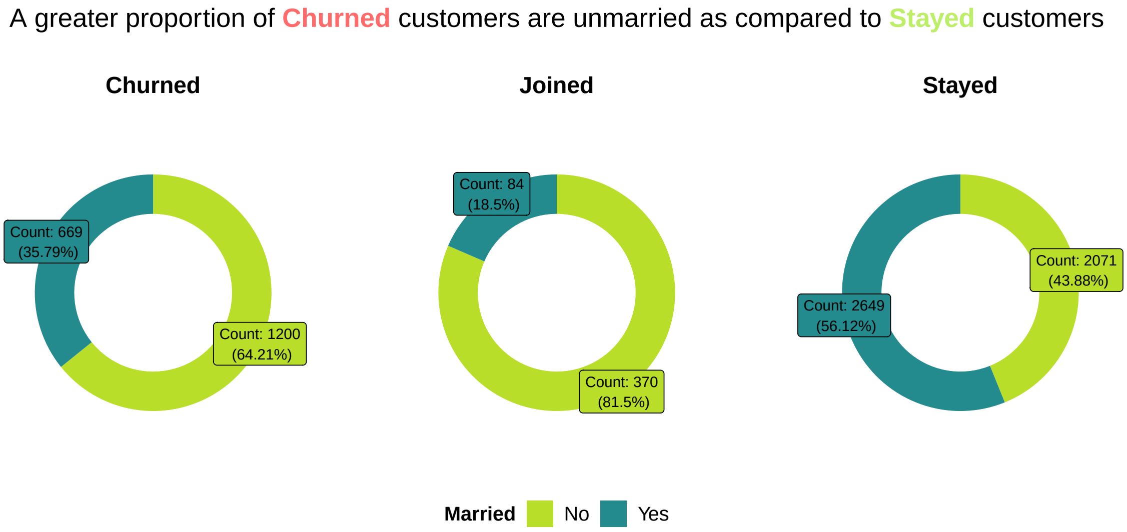
Results and Analysis

Some key observations on the characteristics of the respective `Customer Status` groups:

1. Churned customers typically **spend substantially less on the telecom company's services** as compared to Stayed customers. In addition, they typically have **fewer dependents** and are **unmarried**.


The median total revenue from Churned customers is less than one−third that from Stayed customers


On average, Churned customers have fewer dependents than Stayed customers

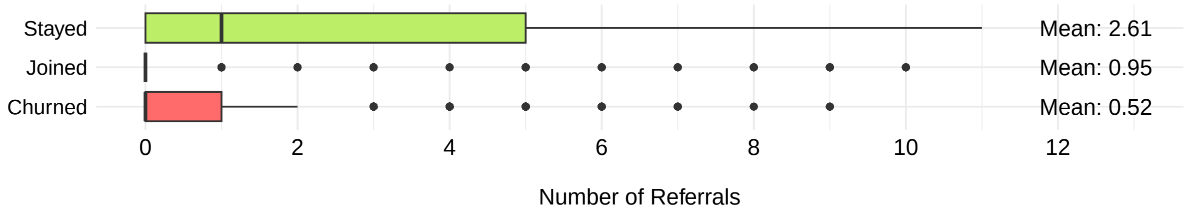

A greater proportion of Churned customers are unmarried as compared to Stayed customers

Lower consumer spending *(which translates to lower revenue for the company)* might be a sign that the customer has little need for or is dissatisfied with the telecom company's services, which might be why they churn. This suggests why Churned customers spend less than Stayed customers.

Customers with fewer dependents and are unmarried are likelier to churn possibly because they have minimal need for telecommunications services beyond the fundamentals. For example, someone who is single or has fewer dependents has fewer close contacts *eg partners and family members* that he/she needs to call or text by phone regularly. This customer might consider cancelling his/her contract plan with the telecom company to switch over to alternative value-for-money plans like prepaid plans.

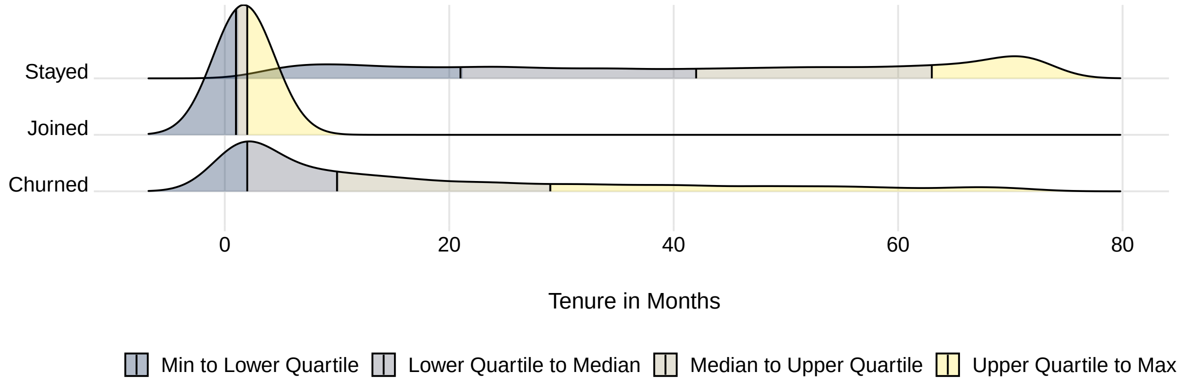
2. Churned customers also tend to **have made fewer referrals of the company** to their family and friends, and **stayed with the company for a shorter period of time** than Stayed customers.

Both mean and median number of referrals of Churned customers is smaller than that of Stayed customers



Note: Dots represent outliers based on Tukey's definition

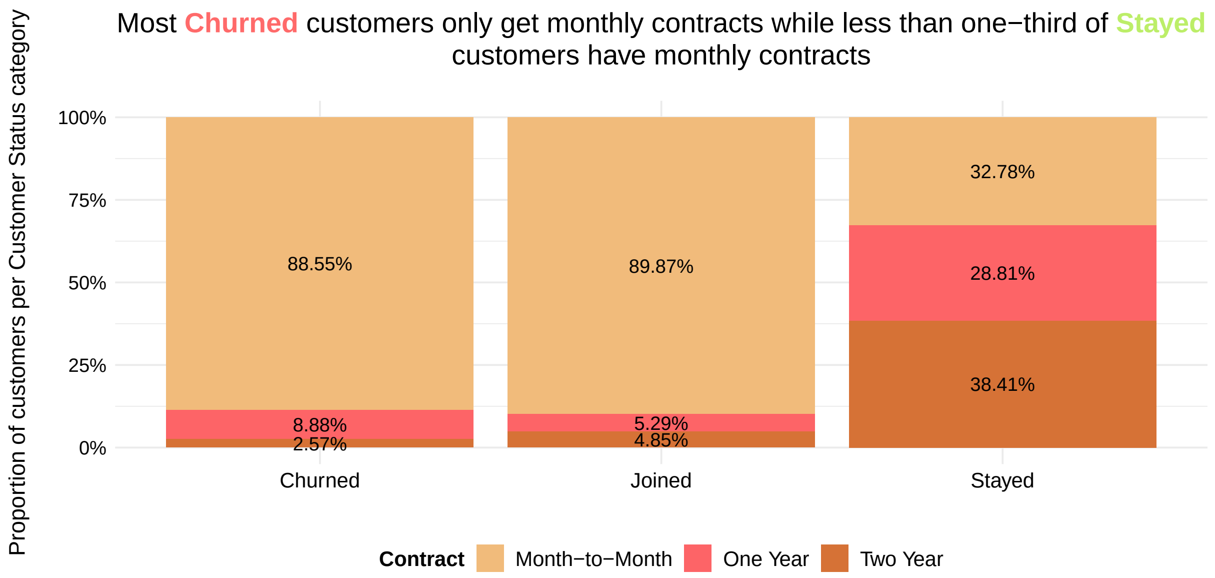
The tenures of Churned customers are concentrated at smaller values than that of Stayed customers, with a lower median too



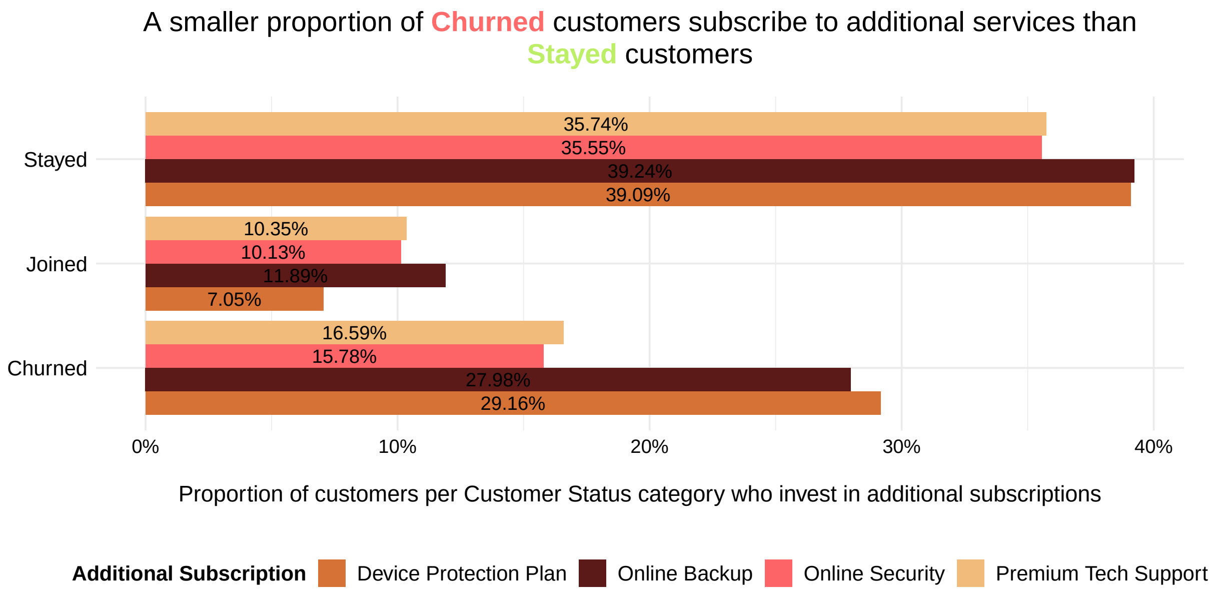
This might be due to some displeasure that Churned customers have towards the telecom company's services, for whatever reason. Consequently, they may give fewer positive reviews or recommendations of the company to the people around them. Also because of their displeasure, they choose to end their subscriptions and churn.

Given that many of them have been with the telecom company for a longer period of time, Stayed customers are more likely to have greater brand loyalty as they grow to be more familiar and comfortable with the company's services over their extended tenure. This might be why they do not end up churning like the Churned customers.

3. As compared to Stayed customers, Churned customers are those who are **less likely to commit to longer-term plans** *eg 1 Year, 2 Year contracts* and are **less likely to spend on supplementary subscriptions** with the company *eg Online Security, Online Backup, Device Protection Plan, Premium Tech Support.*



Most Churned customers only get monthly contracts while less than one−third of Stayed customers have monthly contracts



A smaller proportion of Churned customers subscribe to additional services than Stayed customers

This trend can be explained by the fact that customers who opt for low-commitment Month-to-Month contracts might not even have the intention to stay with the company to begin with. It is therefore not surprising that they churn in the near future. Furthermore, the Month-to-Month contract option makes it easier for customers to prematurely terminate their contracts since their losses would be smaller as compared to when they are on a 1 Year or 2 Year contract. This might suggest why many Churned customers have a monthly contract while many Stayed Customers do not.

As for their lower tendency to spend on additional subscriptions, Churned customers might be customers who have little to no need for such services from the company in the first place. In comparison to Stayed customers, they might be less reliant on the telecom company's services. As a result, they churn.

Summary

Through exploratory data analysis, some distinctions between Churned and Stayed customers have been highlighted. In short, Churned customers tend to be characterized by:

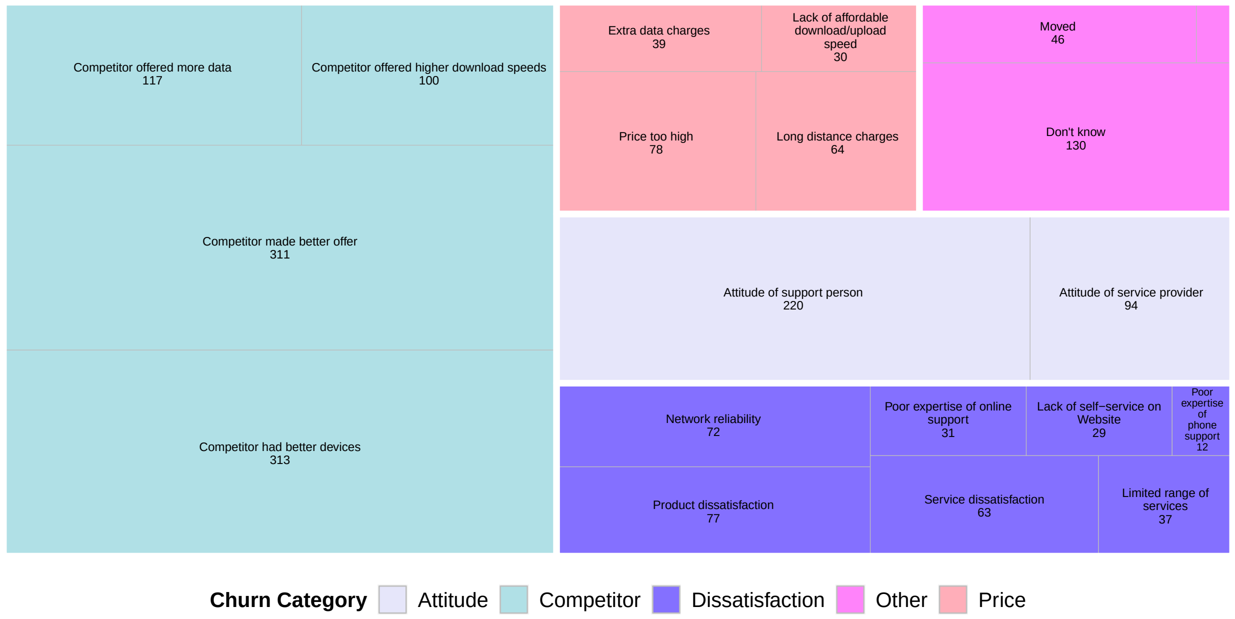
- low revenue customers who are not spending much on the company's services
- customers who do not subscribe to the company's additional paid services
- customers who have few dependents and/or are unmarried
- customers who provide little referrals of the company
- customers who have not been with the company for long
- customers on a monthly contract

With such 'risk factors' in mind, the telecom company can proceed to focus on customers who fit the aforementioned criteria and target them through specific strategies so as to better engage and retain them. Based on the findings of this analysis, these are the customers who might potentially churn in the near future.

Drivers of Churn

Using the `Churn Category` and `Churn Reason` responses provided by customers who churned, this analysis proceeds to find the most popular reason(s) for churning which provides insight into the drivers of churn. The results of analysis are presented in the tree map below.

Frequency of Churn Reason by Churn Category



Note: The unlabelled rectangle in the Other Churn Category represents the Deceased Churn Reason, whose frequency count is 6.

Based on the tree map, the `Churn Category` of **Competitor** exceeds the rest of the categories to account for the most cases of telecom churn recorded in Q2 2022. Particularly, a common gripe was that the telecom company's devices and offers were less attractive than its competitors'. As it contributes the most significantly to churn, the company might want to address this gap to effectively minimize customer churn.

Final Recommendations

To reduce churn most directly and swiftly, the telecom company can consider **revising its contract plans and devices**. Since competitor-related factors are the most common reasons for churn, the company needs to offer more attractive deals to its customers to remain competitive and keep them from switching over to its competitors' services.

One key finding from this analysis is the characteristics of customers who churn. With this knowledge, the company can continue to monitor the behavior of its customers and **engage customers who match the identified characteristics in regular surveys**. These are customers with high potential to churn. By reaching out to these customers via surveys to understand their preferences and demands better, the company can tailor specific strategies to cater to and retain them.

While the characteristics of Joined customers has not been discussed much in this analysis, from the data charts above, we can see that in many ways the characteristics of Joined customers resemble Churned customers. This means there is a possibility that these newly joined customers may churn in the near future too. To encourage them to stay longer, the telecom company can **offer newcomer initiatives and benefits** that grant certain privileges.

Even though the company's objective is to reduce churn, it should not focus only on prospective churners and neglect the customers who have chosen to stay with them. By rolling out **mileage systems** for its customers, the telecom copany can reward customers for their support and further reinforce brand loyalty which contributes to lower churn in the long term.

In short, to minimize churn holistically, it is important for the company to pay attention to the unique characteristics of each group of customers and adopt a mix of strategies to address these characteristics.

Last but not least, the company should **continue with its data collection and analysis efforts** to validate the results obtained in this study. While this study provides the telecom company with useful insight for future action, it at best suggests associations or correlations which do not imply causation. More of such data analyses have to be performed to verify that the results of this study do not come from an isolated customer sample, and is truly reliable in guiding the company's decision-making.

Appendix

Table A1: Remarks on data quality in each column of data set

Column Name	**Remarks**	**Needs Cleaning / Further Checks?**
Customer ID	· 7,043 unique values^a. · 0 missing values. · Appropriate as unique identifier.	No
Gender	· 2 unique categories: Female, Male. · 0 missing values.	No
Age	· 0 missing values. · No outliers.	No
Married	· 2 unique categories: Yes, No. · 0 missing values.	No
Number of Dependents	· 0 missing values. **· Outliers present.**	Yes
City	· 1,106 unique values^b. · 0 missing values.	No
Zip Code	· 0 missing values.	No
Latitude	· 0 missing values.	No
Longitude	· 0 missing values.	No
Number of Referrals	· 0 missing values. **· Outliers present.**	Yes
Tenure in Months	· 0 missing values. · No outliers.	No
Offer	· 6 unique categories: None, Offer A, Offer B, Offer C, Offer D, Offer E. · 0 missing values.	No
Phone Service	· 2 unique categories: Yes, No. · 0 missing values.	No
Avg Monthly Long Distance Charges	**· 682 missing values.** · No outliers.	Yes
Multiple Lines	· 3 unique categories: Yes, No, NA. **· 682 missing values.**	Yes
Internet Service	· 2 unique categories: Yes, No. · 0 missing values.	No
Internet Type	· 4 unique categories: Cable, Fiber Optic, DSL, NA. **· 1,526 missing values.**	Yes

Column Name	**Remarks**	**Needs Cleaning / Further Checks?**
Avg Monthly GB Download	· **1,526 missing values.** · **Outliers present.**	Yes
Online Security	· 3 unique categories: Yes, No, NA. · **1,526 missing values.**	Yes
Online Backup	· 3 unique categories: Yes, No, NA. · **1,526 missing values.**	Yes
Device Protection Plan	· 3 unique categories: Yes, No, NA. · **1,526 missing values.**	Yes
Premium Tech Support	· 3 unique categories: Yes, No, NA. · **1,526 missing values.**	Yes
Streaming TV	· 3 unique categories: Yes, No, NA. · **1,526 missing values.**	Yes
Streaming Movies	· 3 unique categories: Yes, No, NA. · **1,526 missing values.**	Yes
Streaming Music	· 3 unique categories: Yes, No, NA. · **1,526 missing values.**	Yes
Unlimited Data	· 3 unique categories: Yes, No, NA. · **1,526 missing values.**	Yes
Contract	· 3 unique categories: One Year, Month-to-Month, Two Year. · 0 missing values.	No
Paperless Billing	· 2 unique categories: Yes, No. · 0 missing values.	No
Payment Method	· 3 unique categories: Credit Card, Bank Withdrawal, Mailed Check. · 0 missing values	No
Monthly Charge	· 0 missing values. · No outliers. · **Negative values present.**	Yes
Total Charges	· 0 missing values. · No outliers.	No
Total Refunds	· 0 missing values. · **Outliers present.**	Yes
Total Extra Data Charges	· 0 missing values. · **Outliers present.**	Yes
Total Long Distance Charges	· 0 missing values. · **Outliers present.**	Yes

Column Name	Remarks	Needs Cleaning / Further Checks?
Total Revenue	· 0 missing values. · **Outliers present.**	Yes
Customer Status	· 3 unique categories: Stayed, Churned, Joined. · 0 missing values.	No
Churn Category	· 6 unique categories: Competitor, Dissatisfaction, Other, Price, Attitude, NA. · **5,174 missing values.**	Yes
Churn Reason	· 21 unique values[c]. · **5,174 missing values.**	Yes

[a] *7,043 Customer IDs correspond to 7,043 customers, so each Customer ID corresponds to a unique customer, each customer only appears once in the data set — no duplicate rows.*

[b] *High cardinality data, will not be listed one-by-one here.*

[c] *Consists of NA and 20 other churn reasons. Refer to tree map on page 13 for all 20 values.*