



APRIL 2, 2025


# CHURN ANALYSIS

IBM TELCO DATASET

SHINDELMAN, ROBIN

CS 406-PROJECTS

Oregon State University



## Table of Contents

|  |          |
|--|----------|
| <b>WHAT IS CHURN?</b>                  | <b>2</b> |
| <b>PROBLEM BACKGROUND</b>              | <b>2</b> |
| QUESTION STATEMENT                     | 2        |
| DATA                                   | 2        |
| <b>WAREHOUSING</b>                     | <b>3</b> |
| <b>EXPLORATORY DATA ANALYSIS – EDA</b> | <b>4</b> |
| DEMOGRAPHICS                           | 4        |
| SERVICES                               | 5        |
| <b>INFERENCE</b>                       | <b>6</b> |
| CONDITIONS FOR INFERENCE               | 6        |
| HYPOTHESES                             | 6        |
| TESTING                                | 6        |
| INFERENCE CONCLUSIONS                  | 7        |
| <b>RECOMMENDATIONS</b>                 | <b>7</b> |
| <b>REFERENCES</b>                      | <b>8</b> |

# What is Churn?

A common business problem is the question of customer retention. How do we, as an organization, keep customers from leaving to find another product or service? Viable answers to this issue can be difficult to come by, require access to large amounts of customer data, and an understanding of the market or domain.



*Image Source: <https://freetools.textmagic.com/churn-rate-calculator>*

Given these conditions are met, a data professional can mine this information for insights which could then be transformed into valuable drivers of business strategy. While new customers can always be found via marketing campaigns and promotions, these strategies are expensive and time consuming. Current customers cost far less to maintain and can be depended on to continue bringing revenue to your business. Therefore, a valuable use of company resources is to conduct churn analysis.

## Problem Background

### Question Statement

The Telco Customer Churn dataset represents customer data from a fictional telecommunications company. In Q3, Telco provided home and internet services to over 7000 customers in California. Information is provided on each customer's location, demographic, and services received. Most importantly, whether or not the customer signed up, left, or continued using Telco during the quarter is included in the data [1].

Our goal with this dataset is to first identify any interesting customer trends related to churn rate. Following that, we'll respond to this insight with recommendations for strategies to increase retention in key areas.

### Data

Detailed information from IBM on the Telco Customer Churn dataset can be found [here](#). This dataset is ideal for a student conducting a formalized churn analysis. A zip file can be downloaded containing five spreadsheets which act as tables. Since SQL is an important aspect of this demonstration, these spreadsheets were uploaded to Google Cloud Platform's BigQuery as tables in a single relational database.

## Warehousing

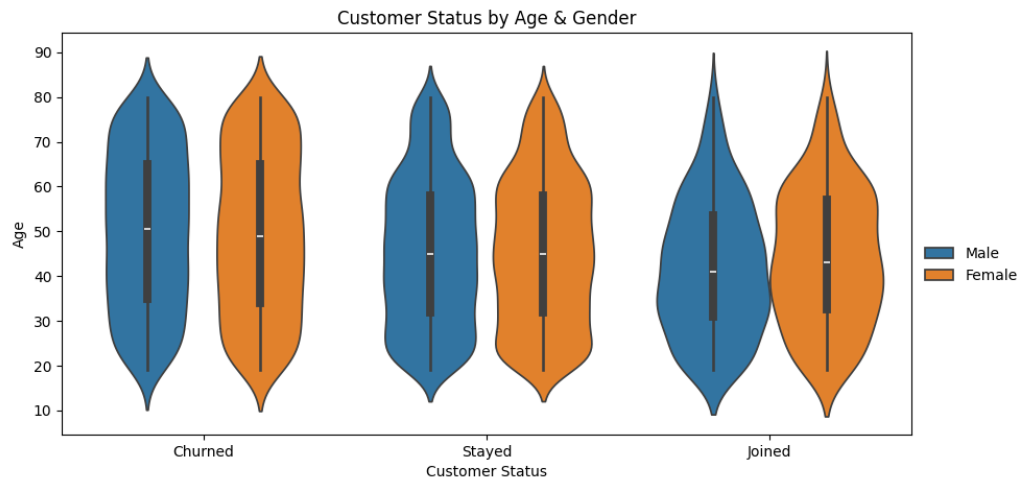
Data warehousing is an essential component in the analysis of enterprise data. In the typical process, data will undergo a process of Extract-Transform-Load (ETL) or Extract-Load-Transform (ELT) before being usable by analysts, engineers, or leadership. In this case, the data was given a very abbreviated treatment of ELT. The data was first “extracted” from IBM’s Cognos Analytics platform as a series of disjoint Excel spreadsheets. Each sheet contained a variety of information to build a complete picture of each customer profile and their churn status.

These spreadsheets then needed to be converted over to CSV files in order to be “loaded” to the chosen data warehousing solution, Google BigQuery. While there are a vast plethora of very strong choices for housing the churn data, Google Cloud Platform offers a robust integration with a variety of services such as Looker, which will be used during the reporting phase of this analysis. BigQuery is a fully administrated relational database system, allowing users to focus on the data itself rather than engineering the SQL server and worrying about things like scaling resources during peak access times. Of the 5 spreadsheets downloaded and converted to CSVs, each was uploaded to BigQuery as a separate table, connected through foreign and primary keys.

Finally, the data was “transformed” through a series of cleansing operations done with SQL queries on BigQuery. Missing values were dropped and a couple of columns needed to have their data type altered, such as zip codes being changed from integers to strings. For the most part, the data provided by IBM was extremely clean and free from errors. During the EDA step in the next stage of this process, some statistical techniques will be used to discover any significant outliers in the data.

# Exploratory Data Analysis – EDA

## Demographics



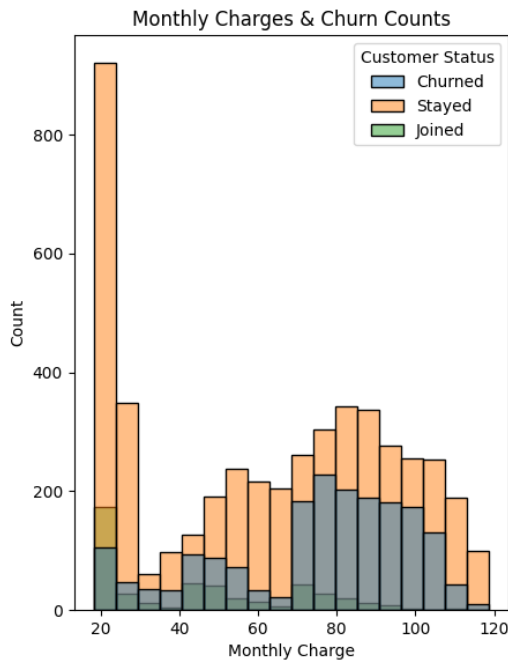
1. Violin plot describing the distribution of customer age and gender in relation to their churn status.

The above plot describes the customer distributions by age and gender. The violin plot reflects the relative number of people in each demographic. Wider points in the chart indicate a larger population, narrower portions indicate fewer people. Fairly quickly, it should be noted that those customers who churned have a very even distribution throughout the age groups. It would appear that age and gender have very little effect on the probability of a customer *leaving* Telco's services.

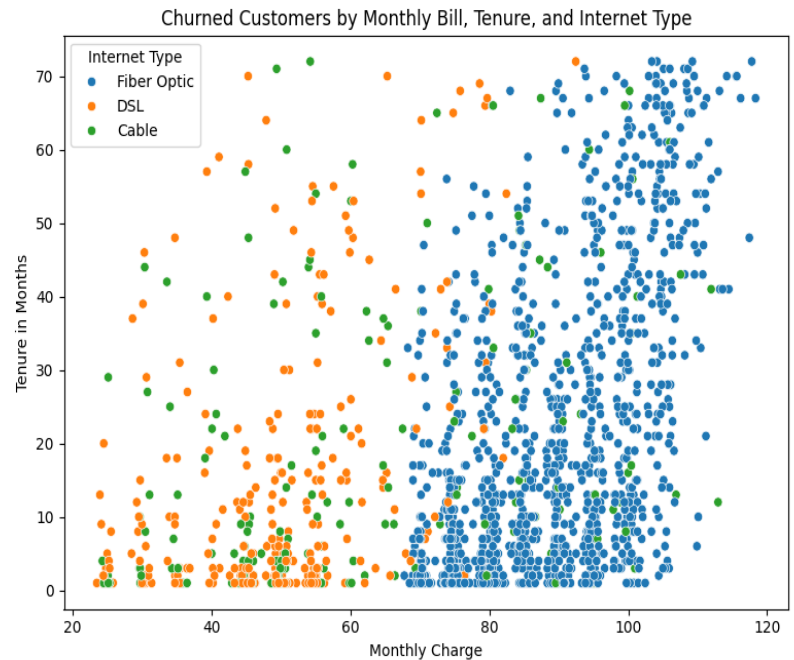
However, Q3 did seem to see somewhat more male customers join Telco in their 30s or 40s than other age groups. This is indicated by the pear shaped blue violin above the Joined category. Interestingly, the inverse seemed to have happened for those who stayed. Both men and women 60 and above make up just a small portion of customers who neither churned nor joined Telco. This may be reflected in the slight bulging of the violins in the Churned category for the same age group.

Simple exploratory analysis of this particular set of variables did not reveal any strikingly significant patterns in the data related to customer churn. Though, the mean bar of male customers who churned does appear to be slightly higher than the females. However, to ascertain whether significantly more male customers churned than females will require [Inference](#).

## Services



3. Stacked histogram showing customers' monthly bill and their churn status.



2. Scatterplot depicting the relationship between tenure, billing, and internet type for solely churned customers.

Both plots above explore the services rendered to each customer. On the left, a stacked histogram shows the distribution of those who Churned, Stayed, or Joined and their monthly bill from \$20 to \$120. From this chart, a possibly significant trend can be seen in the \$70 to \$105 monthly bill range. It would appear that the greatest number of customers who left were paying these higher rates. Interestingly, a majority of customers in Q3 were only paying around \$20 per month for their services. Of the customers in that bracket, only a small portion left Telco. The relationship between monthly charge and churn rate should certainly be tested for significance with statistical inference.

The scatterplot on the right is derived from a query consisting only of those customers who churned in Telco's Q3. There are a couple of insights that jump out right away from this plot. First, there is a very clear delineation between price brackets and the type of internet service rendered to that customer. Second, as tenure grows longer, there does seem to be fewer customers churning. This is relative to both DSL and Fiber Optic plans. So, while there are more customers on Fiber churning, it's possible that the proportion of Fiber customers churning to DSL customers churning is not significant. This relationship between two proportions is also worth exploring in more detail with a statistical analysis.

# Inference

The primary question of inference to explore in this case is whether or not there is a significant difference in churn rate between those customers with DSL compared to those with Fiber Optic. Particularly, are did more customers in Q3 who paid for Fiber Optic churn than those subscribed to DSL? Here, the use of a two-proportion test will be employed to judge the validity of the hypothesis.

## Conditions For Inference

Before anything else, the initial conditions for inference must be satisfied. For a two-proportion test with a categorical variable of interest, **internet type**, there will also be a discrete random variable, **number of customers**, which falls into a binomial distribution. This is assumption of distribution is true for the following reasons:

1. The response variable, **churn status**, has only two possible outcomes. The customer will either churn, or they will not regardless of their chosen **internet type**.
2. There is a fixed **number of customers** in this test. There are 3035 customers in Q3 subscribed to Fiber Optic internet and 1652 customers using DSL. A total of  $n = 4687$ .
3. It is reasonable to assume that all observations in the sample are independent of one another, as they are distinct households with their own subscriptions.
4. Since each of the observations are independent, it should be true that the probability of one customer churning is equal across all customers within each group.

## Hypotheses

For this case, the null hypothesis can be stated as such: The proportion of all DSL internet customers in Telco's Q3 who churned is the same as the proportion of all Fiber Optic internet customers in Telco's Q3 who churned.

Let  $p_{dsl}$  represent the proportion of all customers who used DSL for internet in Telco's Q3, let  $p_{fo}$  represent those who used Fiber Optic in Q3.

$$H_0: p_{dsl} - p_{fo} = 0$$

$$H_A: p_{dsl} - p_{fo} < 0$$

## Testing

To begin the manual two-proportion test using the normal approximation method, a summary table of proportions is produced:

|                   | DSL                         | Fiber Optic                | Total             |
|-------------------|-----------------------------|----------------------------|-------------------|
| Sample size       | $n_{dsl} = 1652$            | $n_{fo} = 3035$            | $n = 4687$        |
| Number churned    | $x_{dsl} = 307$             | $x_{fo} = 1236$            | $x = 1543$        |
| Sample proportion | $\widehat{p}_{dsl} = 0.186$ | $\widehat{p}_{fo} = 0.407$ | $\hat{p} = 0.329$ |

The Standard Error of the distribution of sample proportions is performed first.

$$SE_{\widehat{p}_{dsl} - \widehat{p}_{fo}} = \sqrt{0.329(1 - 0.329) \left( \frac{1}{1652} + \frac{1}{3035} \right)} = 0.014$$

Thus, the z-statistic can be calculated as,

$$z = \frac{(0.186 - 0.407) - 0}{0.014} = -15.79$$

Such a low z-statistic results in a very small p-value approaching 0.

## Inference Conclusions

There is strong evidence to indicate that the proportion of Telco customers who are subscribed to Fiber Optic and churned by the end of Q3 is larger than the proportion of Telco DSL customers who churned.

## Recommendations

There appears to be strong evidence for Telco to consider investigating the segment of their customers who subscribe to Fiber Optic service. A number of reasons could be responsible for the disproportionately high churn rate of customers in this bracket as compared to those simply on DSL.

Chief among those reasons could be the difference in monthly price between DSL and Fiber. As shown in Figure 3, the starting charge for most DSL customers hovers around \$25/month and continues up to about \$60/month. Fiber Optic on the other hand, starts at \$60/month and reaches nearly \$120/month at the highest point.

The most drastic response to this information would be to reduce the price of Fiber Optic to be closer to that of DSL. However, this is most likely not an ideal tactic for Telco to take as Fiber infrastructure may incur more cost to the company. Instead, Telco could create introductory promotions for customers who are interested in switching over to Fiber. These promotions could be offered at a lower price to reduce churn for 2-4 months, then return to their normal pricing at the end of the promotion. This strategy could help customers build momentum with the company and possibly reduce the rate at which they leave.



## References

- [1] S. Macko, “Telco Customer Churn,” IBM Business Analytics Blog, Jul. 11, 2019. [Online]. Available: <https://community.ibm.com/community/user/businessanalytics/blogs/steven-macko/2019/07/11/telco-customer-churn-1113>