BIRMINGHAM CITY University

Customer Churn Analysis

Module Coordinate & Code:

Prof . AbdulRahman Alsewari

CMP 7205

Student Name & SID:

Yash Mehta

23145127

Submission Date:

2ndth September 2024

Table of Contents

Executive Summary 2

1. Introduction 2

1.1 Problem Domain 2

1.2 Statistical Questions 2

2. Methodology 3

2.1 Exploratory Data Analysis 3

2.2 Correlation Analysis 3

2.3 Hypothesis Testing 3

2.4 Regression Analysis 3

3. Dataset 7

3.1 Dataset Description 3

3.2 Data Pre-processing 4

4. Results and Discussion 5

4.1 Exploratory Data Analysis 5

4.2 Correlation Analysis 7

4.3 Hypothesis Testing 8

4.4 Regression Analysis 9

5. Conclusion 10

References 11

Appendix:Code 14

Executive Summary

Analyzing customer churn is important to business because it directly addresses the critical problem domain that impacts their sustainability and growth. Looking into the dataset, the transformation of the 'Churn' attribute from binary "Yes" and "No" into numerical 0 and 1 helped drive consistent methodologies across analyses. After the summary of the "tenure" attributes, it indicates that there is diversity-a mix of both long-term and new customers. The next Because the analysis of correlation revealed that month-to-month contracts have a strong negative correlation, approximately -0.40, with churn. According to the matrix plot, the correlation of contract type and churn is highly negative; short contract tenure is associated with high churn rates. Also interesting were customer tenure and monthly charges. As indicated by a boxplot analysis, long-term customers pay more fees and thus give more revenue importance to customer loyalty. The analysis of Internet services showed that the Fiber optic users are more likely to churn, and thus service satisfaction is of utter importance. Besides, hypothesis testing confirmed that customer tenure plays an important role in that shorter-tenured customers are more likely to churn, and nicely supported it with the visually appealing plot of distribution of churned and non-churned customer's tenure. Besides, the linear regression model provided insight into the relationship between tenure, monthly charges, and total charges. Monthly Charges and Tenure coefficients are 35.88 and 65.41 correspondingly, this means that on average for each additional unit increase in Monthly Charges or tenure, Total Charges is expected to rise by the corresponding coefficient assuming the other variable will be held. constant. The trend of higher monthly fees contributing to longer customer tenure and higher overall charges was graphically summarized using a scatter plot with the regression line.

1. Introduction

1.1. Problem Domain

Customer churn, which can be simply described as customers changing service providers, is a serious concern for almost all industries — be it IT, social network, and telecommunication. In particular, when discussing customer churn in the telecommunication business environment, it is chronicled to occur regularly. Customers will switch to any service provider for better service or price point (Vafeiadis et al., 2015).

This report will statistically analyze factors that correlate or influence on the churn rate (Amin et al., 2019). Evaluating, what factors do contribute to, the churn rate will provide insight, for telecommunication business service providers in establishing what are the fundamental and contributing factors they need to evaluate and manage in their markets to limit their organization's churn rate.

It is recognized that the pursuit of new customers is no longer the focal point of companies' objectives; instead, the significant approach is to plan properly for satisfying the current customer base and attracting them more to remain with them. Because of the amount of cost the company is going to incur to attract new customers, numerous scholars have expressed this point of view, including (Day, 1999). Likewise, whenever a scholar has only discussed that retained customers enables organizations to increase benefits, it can also lead to additional profit for companies as they attempt to reduce churn, even if the profit will only improve a little.

Maximizing firm's profitability today is reliant on the research of customer churn variability particularly within service industries. "Our customers' relationships is a backbone of an organization generally, and service firms in particular"(Williamson, 1966).

1.2. Statistical Questions

To obtain insights from the dataset containing Customer Churn data, there are several statistical problems to consider. This report must provide answers to the following queries:

1. Which contract type shows a lower likelihood of customer churn?
2. Is there a notable difference in monthly charges among customers with varying lengths of tenure?
3. What type of internet service has the greatest impact on the churn rate?
4. Between contract type and monthly charges, which factor has a stronger influence on the churn rate?

5. Are customers who are willing to commit to longer-term contracts less likely to churn?
6. Do customers with longer tenure exhibit a lower churn rate compared to those with shorter tenure?
7. Is there a linear relationship between monthly service charges and the total charges incurred by customers?

2. Methodology

To achieve the objectives of this study, the dataset is analyzed using the following statistical methods to extract pertinent information.

2.1. Exploratory Data Analysis

Exploratory Data Analysis refers to an approach in data analysis that provides a summary or overview of the major characteristics of the dataset, often employing visual methods. EDA helps analysts discover patterns, identify anomalies, test hypotheses, and check assumptions through various data visualization techniques. EDA is important for understanding the data's underlying structure and the data should be understood before more complex statistical modeling (Morgenthaler, 2009).

The figures in the results section demonstrate factors that contribute to customer churn.

2.2. Correlation Analysis

In combination, the correlation matrix will be of great help in rigorously assessing and contrasting the impact of month to month charges and contract type on churn, allowing the evaluation of which of those factors may weigh more heavily on customer churn decisions (Senthilnathan, 2019). We will also take the time to develop an in-depth understanding of the relationship between contract type and churn, which will produce detailed, subtle insights based on qualitative evidence leading to conclusions as to how to improve processes of customer retention in a more effective manner for their individual contracts. This thorough analysis will provide clarity for understanding the significant determinants underlying customer loyalty and behavior, to support the coordination of actually targeted interruptions (Makowski et al., 2020).

2.3. Hypothesis Testing

The hypothesis test described here is a basic statistical method used to determine whether an opinion or claim regarding a population parameter is reasonable (Klein et al., 2003). We report our p-value calculation, which is an important statistic used to determine how much evidence is against the null hypothesis. With the calculated p-value we carefully evaluate, based on the p-value, whether or not to reject the null hypothesis. We provide a clear null hypothesis and use a t-test to evaluate whether we would accept or reject the null hypothesis based on our analysis. This will yield valuable findings (Cover, 2016).

2.4. Regression Analysis

There are numerous regression methodologies available for analyzing customer churn datasets, and linear regression will be used for the simple example provided here. The investigation will be done to see if any kind of relationship exists in terms of monthly service charges and the total service charges that the customers are perceived to be paying (Kavitha et al., 2016).

3. Dataset

3.1. Dataset Description

The dataset I employed for my analysis is a dataset from Kaggle. Regarding its dimensions, there are in the dataset entirely 25 columns and 7044 rows before any pre-processing occurs of course. Essentially this dataset includes customer data for a telecommunications service provider, consisting of various attributes for a number of demographic, service-related type, and also for a few billing attributes. Significantly, the dataset has information about Customer Age, Gender, Senior Citizen, Partner, tenure, phone service, internet service, type of contract, Monthly Charges, Total Charges and one of the most important attribute is the Churn. Therefore the dataset's Churn variable can serve as the response variable, thus the dataset can fit predictive modelling tasks in particular involving to customer churn prediction model. Moreover this dataset can aid in exploratory data analysis(eda) using few of these attributes which will give you some idea of underlying factors responsible for customer churn and can pursue for the geospatial frameworks for our marketing strategic decisions in order to maintain customer retention within this competitive landscape who may consider the service providers of their choice to establish for the a geospatial services. Overall dataset is partially clean but proper preprocessing techniques should be relatively simple for me to carry out and for have the glimpse of the customer data I will discuss more on that aspect in Pre-processing either as a standalone section or a part of full report.

3.2. Data Pre-processing

```
> glimpse(churn_clean)
Rows: 7,032
Columns: 21
$ customerID       <chr> "7590-VHVEG", "5575-GNVDE", "3668-QPYBK", "7795-CFOCW", "9237-HQITU", "9305-CDS…
$ gender           <chr> "Female", "Male", "Male", "Male", "Female", "Female", "Male", "Female", "Female…
$ SeniorCitizen    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0…
$ Partner          <chr> "Yes", "No", "No", "No", "No", "No", "No", "No", "Yes", "No", "Yes", "No", "Yes…
$ Dependents       <chr> "No", "No", "No", "No", "No", "No", "Yes", "No", "No", "Yes", "Yes", "No", "No"…
$ tenure           <int> 1, 34, 2, 45, 2, 8, 22, 10, 28, 62, 13, 16, 58, 49, 25, 69, 52, 71, 10, 21, 1, …
$ PhoneService     <chr> "No", "Yes", "Yes", "No", "Yes", "Yes", "Yes", "No", "Yes", "Yes", "Yes", "Yes"…
$ MultipleLines    <chr> "No phone service", "No", "No", "No phone service", "No", "Yes", "Yes", "No pho…
$ InternetService  <dbl> 1, 1, 1, 1, NA, NA, NA, 1, NA, 1, 1, 0, NA, NA, NA, NA, 0, NA, 1, NA, 1, 0, 0, …
$ OnlineSecurity   <chr> "No", "Yes", "Yes", "Yes", "No", "No", "No", "Yes", "No", "Yes", "Yes", "No int…
$ OnlineBackup     <chr> "Yes", "No", "Yes", "No", "No", "No", "Yes", "No", "No", "Yes", "No", "No inter…
$ DeviceProtection <chr> "No", "Yes", "No", "Yes", "No", "Yes", "No", "No", "Yes", "No", "No", "No inter…
$ TechSupport      <chr> "No", "No", "No", "Yes", "No", "No", "No", "No", "Yes", "No", "No", "No interne…
$ StreamingTV      <chr> "No", "No", "No", "No", "No", "Yes", "Yes", "No", "Yes", "No", "No", "No intern…
$ StreamingMovies  <chr> "No", "No", "No", "No", "No", "Yes", "No", "No", "Yes", "No", "No", "No interne…
$ Contract         <chr> "Month-to-month", "One year", "Month-to-month", "One year", "Month-to-month", "…
$ PaperlessBilling <chr> "Yes", "No", "Yes", "No", "Yes", "Yes", "Yes", "No", "Yes", "No", "Yes", "No", …
$ PaymentMethod    <chr> "Electronic check", "Mailed check", "Mailed check", "Bank transfer (automatic)"…
$ MonthlyCharges   <dbl> 29.85, 56.95, 53.85, 42.30, 70.70, 99.65, 89.10, 29.75, 104.80, 56.15, 49.95, 1…
$ TotalCharges     <dbl> 29.85, 1889.50, 108.15, 1840.75, 151.65, 820.50, 1949.40, 301.90, 3046.05, 3487…
$ Churn            <chr> "No", "No", "Yes", "No", "Yes", "Yes", "No", "No", "Yes", "No", "No", "No", "No…
```

Figure 1. Dataset Overview

Above Figure 1 depicts the dataset, and a few discussions are presented. To start, the original data was unstructured and the "TotalCharges" column is found to have 11 missing values. In figure 2 we see that, relatively, the missing value column does not contribute significantly to the datatset and therefore, we remove these row.

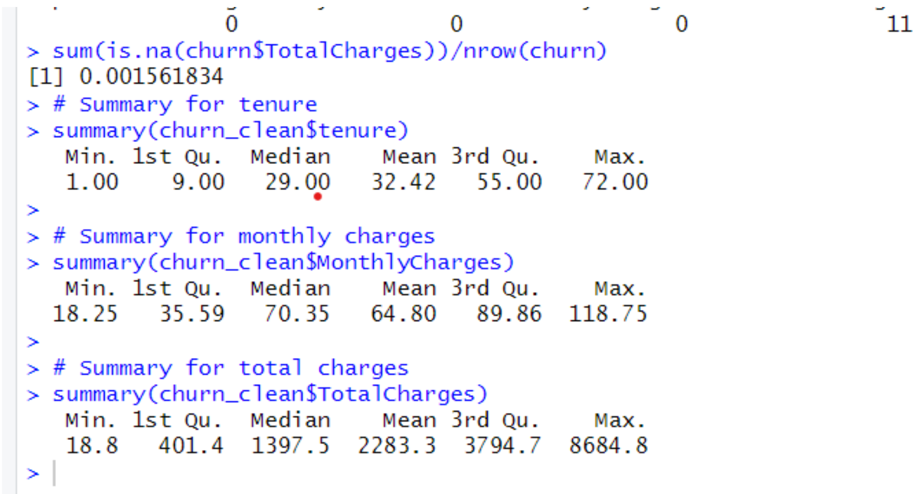
```
> sapply(churn, function(x) sum(is.na(x)))
      customerID           gender    SeniorCitizen          Partner       Dependents           tenure     PhoneService    MultipleLines
               0                0                0                0                0                0                0                0
 InternetService   OnlineSecurity     OnlineBackup DeviceProtection      TechSupport      StreamingTV  StreamingMovies         Contract
               0                0                0                0                0                0                0                0
PaperlessBilling    PaymentMethod   MonthlyCharges     TotalCharges            Churn
               0                0                0               11                0
> sum(is.na(churn$TotalCharges))/nrow(churn)
[1] 0.001561834
>
```

Figure 2. Missing Value Proportion

After the process given that the categorical character variable of the "Churn" attribute could not be used for correlation analysis, we went through the conversion from "Yes", "No" to "1", "0". The same process was followed with the "Contract" attribute.

4. Results and Discussion

4.1. Exploratory Data Analysis



```
                  0                 0                 0                 11
> sum(is.na(churn$TotalCharges))/nrow(churn)
[1] 0.001561834
> # Summary for tenure
> summary(churn_clean$tenure)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1.00    9.00   29.00   32.42   55.00   72.00 
> 
> # Summary for monthly charges
> summary(churn_clean$MonthlyCharges)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  18.25   35.59   70.35   64.80   89.86  118.75 
> 
> # Summary for total charges
> summary(churn_clean$TotalCharges)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   18.8   401.4  1397.5  2283.3  3794.7  8684.8 
> 
```

Figure 3. Summary

The above image provides a summary of critical variables within a customer churn dataset, specifically focusing on tenure, monthly charges, and total charges. Customer tenure ranges from 1 to 72 months, with a median of 29 months, indicating that most customers have maintained service for a substantial period. Monthly charges exhibit considerable variability, with a median of $70.35 and a maximum of $118.75. Similarly, total charges show significant variation, ranging from $18.8 to $8684.8, with a median of $1397.5. Notably, the percentage of missing values in the Total Charges column is extremely low, at just 0.16%.

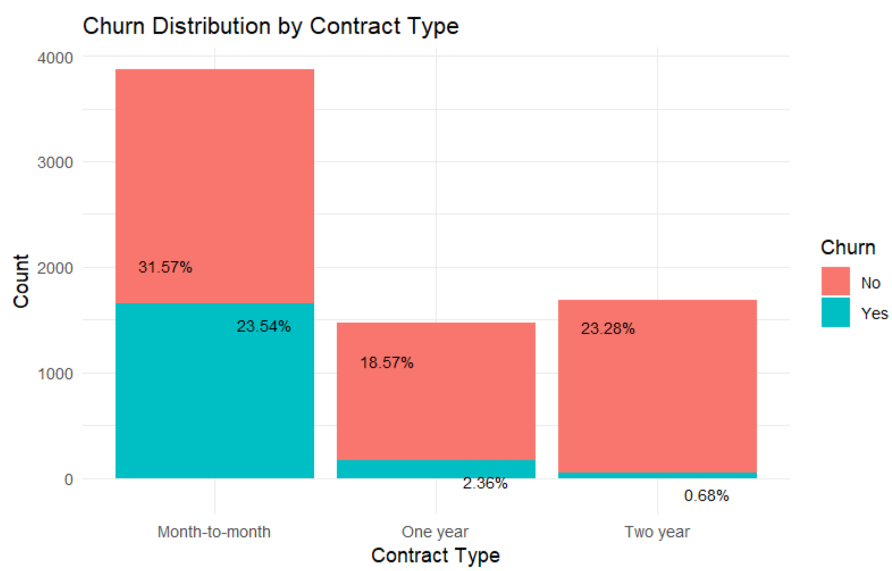


Figure 4. Customer Churn By Proportion Type

The data represented on the chart indicates that customers with a month-to-month contract experience the highest churn. It shows a noticeable decline in churn for customers with a one-year contract, an even larger decline for customers with a two-year contract, suggesting increased customer retention as a result of longer contracts. Furthermore, two-year contracts exhibit the lowest churn rate within the analysis.

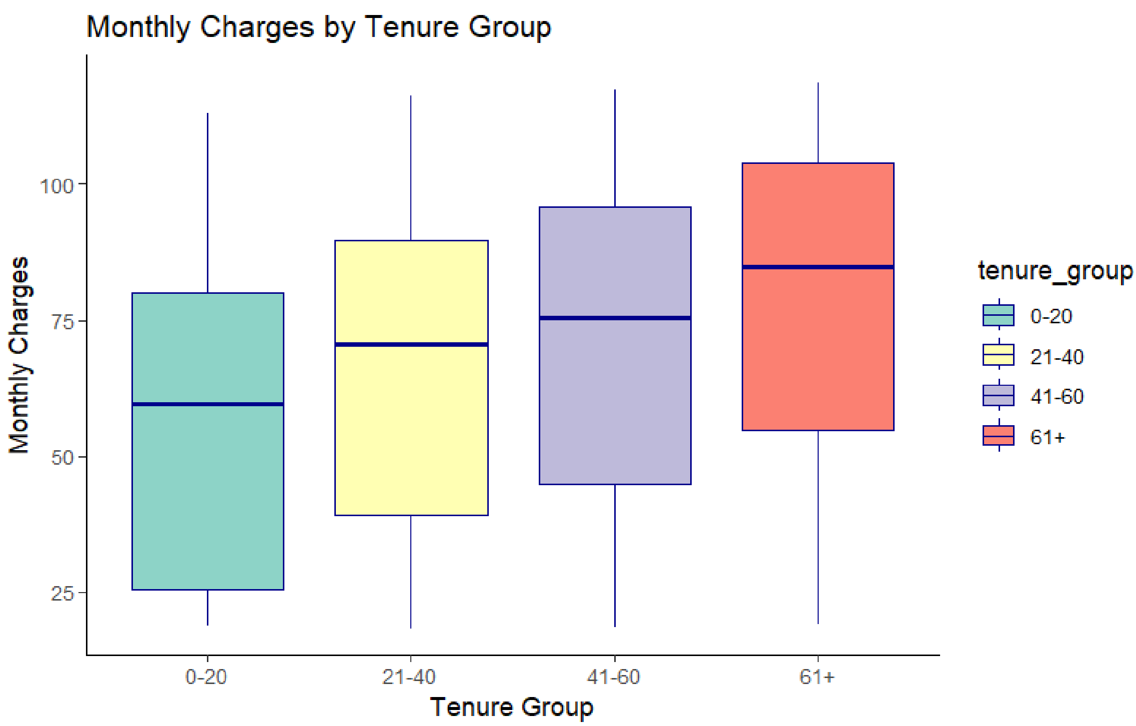


Figure 5. Monthly Charges By Tenure Group

This boxplot represents the distribution of monthly charges by tenure groups, 0-20, 21-40, 41-60, and 61+ months. There is a clear trend that customers in the longest tenure group (61+ months) have a higher corresponding monthly charge (median slightly above $75). Additionally, the variability of charges in this group is greatest, suggesting a wider range of options. Customers from the short tenure group (0-20 months) generally pay lower monthly charges, with less variability. The median charge appears to correlate; the longer the tenure, the higher the typical monthly charge.

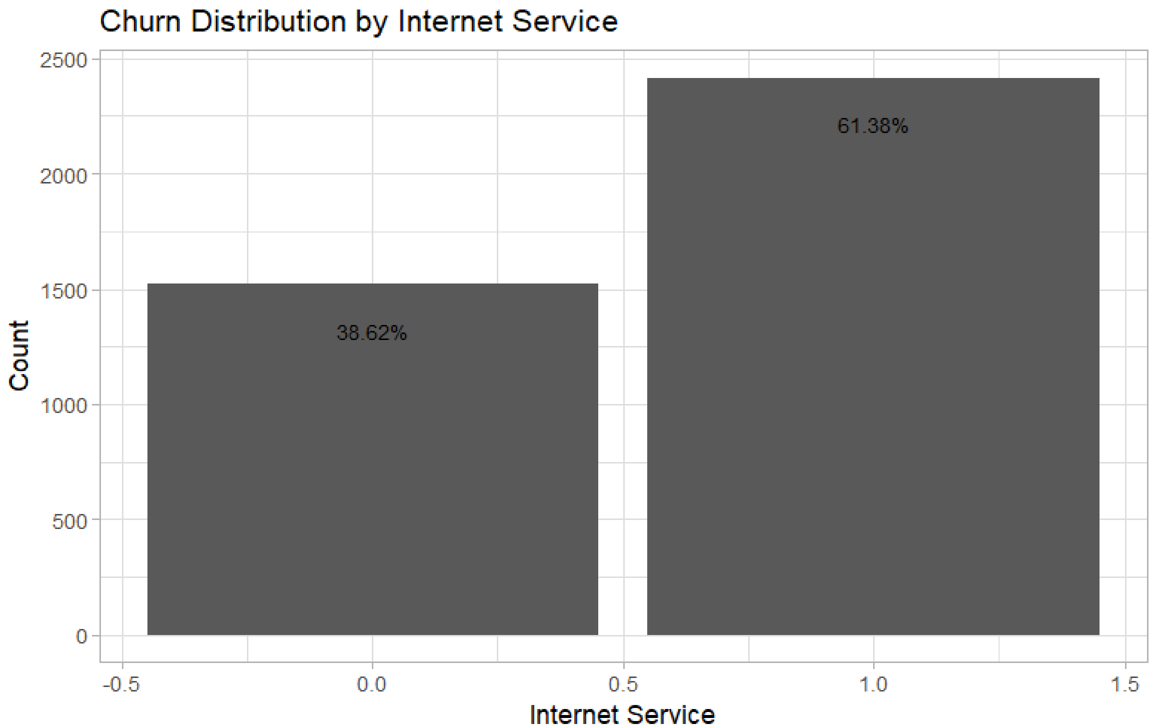


Figure 6. Churn Distribution By Internet Service

This is a bar graph of the distribution of the churn due to different types of Internet services. The bars represent different types of Internet service, and the heights of the bars correspond to the count of customers using each of those services. The percentage on the bars shows the proportion of churning in each category of Internet service. For instance, the left bar shows that 38.62% of the total customers belong to a certain Internet service type, while the right bar represents 61.38% for another type of service. This view will help in directly comparing the churn proportions across different services.

4.2. Correlation Analysis

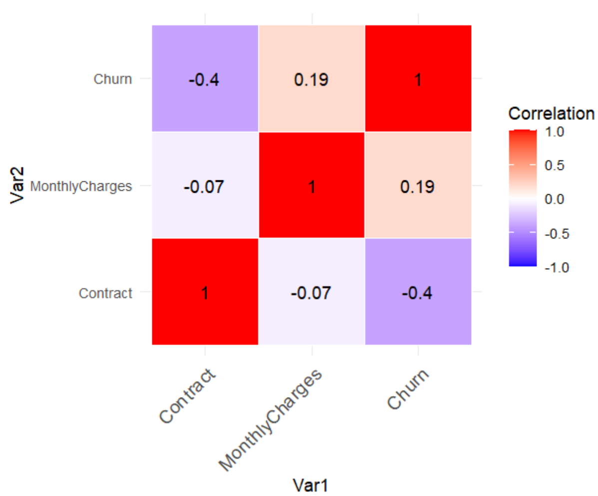


Figure 7. Monthly Charges - Churn - Contract

The above figure demonstrates a higher correlation between Contract and Churn versus Monthly Charges and Churn. The correlation is a significantly negative correlation of 0.4 , indicating a moderate strength of association between contract type and churn. The negative sign indicates an inverse relationship, indicating that as the value of one variable rises, the value of the other tends to fall. In concrete terms, the churn rate will decrease if one subscribes to a long term contract (e.g., one or two years). Moreover, figure 8 presents a more fine-grained illustration of the correlation between Churn and Contract status. The p-value is nearing zero, which further suggests that the correlation possesses high statistical significance. Thus, the analysis indicates that type of contract is an important characteristic in its relationship with churn, customers with shorter term contracts were more likely to churn compared to those who had long-term contracts.

```
+   coord_fixed()
> churn_correlation_result <- cor.test(churn_clean$Contract,churn_clean$Churn)
> print(churn_correlation_result)

	Pearson's product-moment correlation

data:  churn_clean$Contract and churn_clean$Churn
t = -36.175, df = 7030, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.4156741 -0.3762600
sample estimates:
       cor
-0.3961495

> |
```

Figure 8. Contract Churn Correlation

4.3. Hypothesis Testing

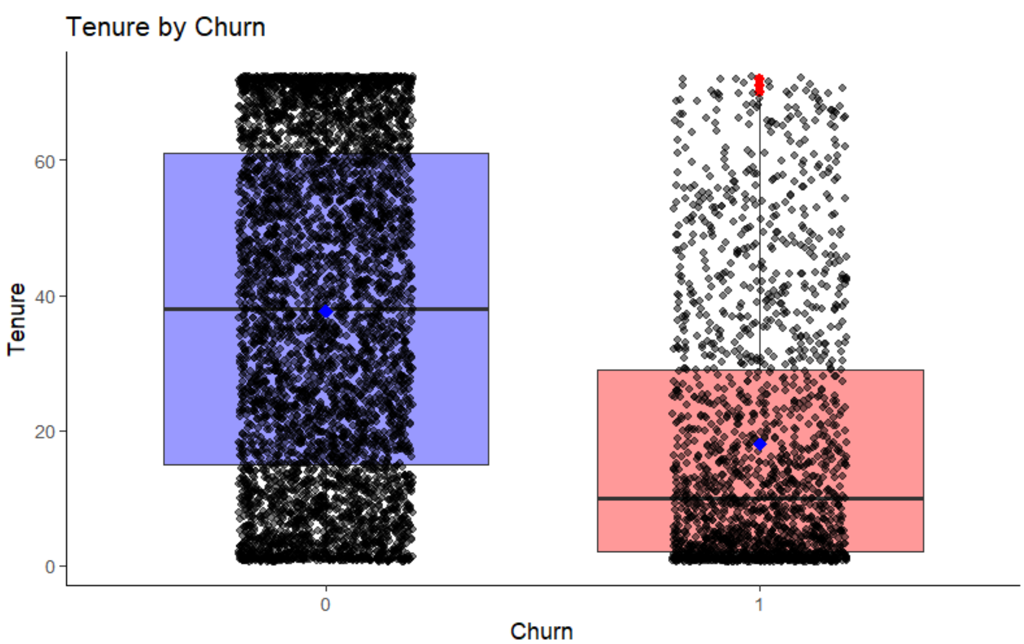


Figure 9. Tenure By Churn

The box plot presented in this report shows the distribution of customer tenure by churn status (Churn). The x-axis represents churn status (0 = no churn; 1 = churn), and the y-axis shows tenure in months.

For customers that did not churn (Churn = 0), the boxplot reveals a wide distribution of tenure with an interquartile range (IQR) between roughly 20 to 60 months; the 50th percentile, or median, is near to 40 months, which indicates that at least half of the customers, on average, will have been with the company longer than 40 months.

```
> t.testing <- t.test(hypertension_group, no_hypertension_group)
> print(t.testing)

        Welch Two Sample t-test

data:  hypertension_group and no_hypertension_group
t = -34.972, df = 4045.5, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -20.77364 -18.56811
sample estimates:
mean of x mean of y
 17.97913  37.65001

>
```

Figure 10. T-Test

For customers that churned (Churn = 1), the distribution of tenure is much more concentrated at the lower end of the tenure range. The IQR for customers that churned is much shorter, suggesting most churned customers spent at most 20 months with the company on average and, most likely, less than 20 months; the median is near to 15 months. This suggests that customers with shorter tenure were more likely to churn than those that had spent longer with the company.

Additionally, the plot presents individual terms (dictated by the black dots) on the box plot. The churned group contains a few red dots indicating outlier customers who spent an unusually high amount of time with the company, but still left. In this case, the blue dots located within the box represents the mean time for each group of customers. Overall, the box plot has a clear distinction between customers who did or did not churn. There is a clear statistical association (as noted by the box plot and data point location of each group) between customer tenure and likelihood to churn; the longer the time spent as a customer one was less likely to churn.

4.4. Regression Analysis

Below is the summary output using a linear regression model in R, with TotalCharges as the dependent variable and MonthlyCharges and tenure as independent variables. It can be observed that the model explains about 89.5% of the variation in TotalCharges, as indicated by an Adjusted R-squared value of 0.895. Both predictors-monthly charges and tenure-are significant with positive coefficients, which postulates that an increase in either of the two increases total charges. The residual standard error is 734.5, which represents an average distance that the observed values fall from the regression line.

```
> summary(regression_model)

Call:
lm(formula = TotalCharges ~ MonthlyCharges + tenure, data = churn_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-1942.3  -465.4   -94.7   494.0  1911.8

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -2162.4319    21.9899  -98.34   <2e-16 ***
MonthlyCharges    35.8789     0.3005  119.42   <2e-16 ***
tenure            65.4141     0.3683  177.62   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 734.5 on 7029 degrees of freedom
Multiple R-squared:  0.895,     Adjusted R-squared:  0.895
F-statistic: 2.997e+04 on 2 and 7029 DF,  p-value: < 2.2e-16

>
```

Figure 11. Statistical Summary

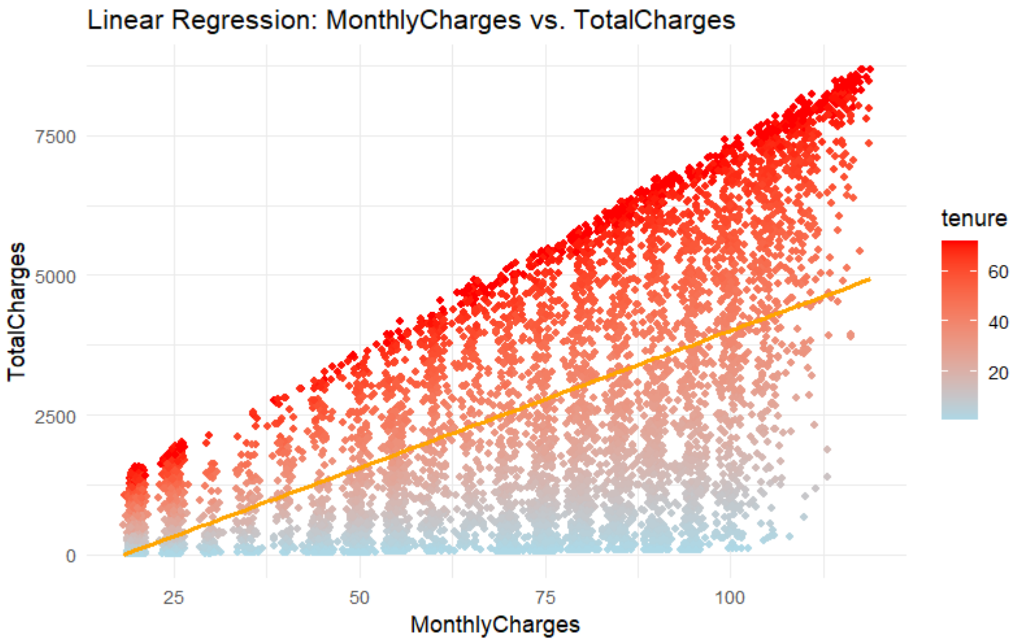


Figure 12. Monthly Charges VS Total Tenure Charges

This scatter plot presents an illustration of the relationship between MonthlyCharges (x-axis) and TotalCharges (y-axis), and a linear regression line (in yellow) is fitted to the model. Every point on this plot corresponds to a customer, and the color of the point indicating the customer's tenure (in months) as illustrated by the gradient color on the right of the plot that runs from blue (short tenure) to red (long tenure).

The regression line shows a positive relationship between MonthlyCharges and TotalCharges, which is of course expected, as TotalCharges are the cumulative sums of MonthlyCharges over time. As MonthlyCharges increase, TotalCharges will also generally increase, however the spread of points around the regression line indicates there is variability about total charges based on customer tenures.

Customers with longer tenures (indicated by red points) generally have higher clustering in TotalCharges and more spread on the MonthlyCharges axis, which indicates longer tenured customers have higher total charges accrued over time. Shorter tenured customers (represented by blue points) have lower total charges, even if the MonthlyCharges are similar to longer tenured customers.

The gradient effect clearly indicates how tenure affects the charged accumulated to total charges, as longer tenured customers are expected to have higher raters of total charges accrued, even though the levels of monthly billing are relatively similar. The positive slope of the regression line demonstrates that there is a strong linear correlation with MonthlyAmounts and TotalCharges.

5. Conclusion

All the inferential statistical query posted in this report has been answered by the findings. The key objective was to focus on the aspects that affect customers' churn in a service provider company. Analysis showed that there was enough evidence to prove that there exists a significant relationship between the contract duration and also with the problem of churning. Churns were most likely to happen when customers had month-to-month contracts. Correlation analysis showed the moderate negative correlation between the contract type. and churn. The hypotheses testing supported that longer-tenure customers are less likely to churn, therefore justifying how customer tenure influences churn rates. Also, a linear regression model showed that monthly charges and tenure significantly predict total charges, with the interactive relationship between these variables in dictating how much a customer spends. On the whole, the result portrays the important role which contract duration can play in modeling customer's churn. dynamical, with longer-term commitments associated with reduced churn rates. The statistical significance of the correlation and regression coefficients underlines how robust these findings are. The stakeholder perspective-as these shine in the context of marketing and customer support-is the ability to use this association between contract duration and churn rate in formulating specific strategies toward customer retention. They benefit from actionable insights that drive data-informed decision-making. This will increase customer satisfaction, loyalty, and the long-term financial viability of the company.

References

Amin, A., Al-Obeidat, F., Shah, B., Adnan, A., Loo, J., and Anwar, S. Customer churn prediction in telecommunication industry using data certainty. *Journal of Business Research*, 94:290–301, 2019.

Cover, T. M. Hypothesis testing with finite statistics. *The Annals of Mathematical Statistics*, 40(3):828–835, 2016.

Day, G. S. *The market driven organization: understanding, attracting, and keeping valuable customers*. Simon and Schuster, 1999.

Kavitha, S., Varuna, S., and Ramya, R. A comparative analysis on linear regression and support vector regression. In *2016 online international conference on green engineering and technologies (IC-GET)*, pp. 1–5. IEEE, 2016.

Klein, J. P., Moeschberger, M. L., Klein, J. P., and Moeschberger, M. L. Hypothesis testing. *Survival analysis: techniques for censored and truncated data*, pp. 201–242, 2003.

Makowski, D., Ben-Shachar, M. S., Patil, I., and Lüdecke, D. Methods and algorithms for correlation analysis in r. *Journal of Open Source Software*, 5(51):2306, 2020.

Morgenthaler, S. Exploratory data analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1):33–44, 2009.

Senthilnathan, S. Usefulness of correlation analysis. *Available at SSRN 3416918*, 2019.

Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., and Chatzisavvas, K. C. A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*, 55:1–9, 2015.

Williamson, J. Profit, growth and sales maximization. *Economica*, pp. 1–16, 1966.

6. Appendix: R Code

```
install.packages("psych")
library(psych)

library(plyr)
library(rpart.plot)
library(caret)
library(gridExtra)
library(tidyverse)
library(rsample)
library(e1071)
library(GGally)
library(data.table)
library(DT)

library(readr)
library(ggplot2)
library(dplyr)
library(tidyr)
library(corrplot)

library(generalhoslem)

setwd("C:/Users/yam27/OneDrive/Documents/Applied Stats")
getwd()
churn <- read.csv("customer_churn_data.csv")
glimpse(churn)
```

```
summary(churn)

sapply(churn, function(x) sum(is.na(x)))

churn[is.na(churn$TotalCharges),]

sum(is.na(churn$TotalCharges))/nrow(churn)

nrow(churn)

remove_clean <- na.omit(churn)
churn_clean <- na.omit(churn)
sapply(churn_clean, function(x) sum(is.na(x)))

# Summary for tenure
summary(churn_clean$tenure)

# Summary for monthly charges
summary(churn_clean$MonthlyCharges)

# Summary for total charges
summary(churn_clean$TotalCharges)

churn_clean$InternetService <-
  as.numeric(mapvalues(churn_clean$InternetService, from=c("No", "DSL", "Fiber optic"), to
      =c("0", "1", "2")))
churn_clean$InternetService

glimpse(churn_clean)

# describe(churn_clean$Churn)
describe(churn_clean)
describe(churn_clean$tenure)
glimpse(churn_clean)
str(churn_clean)
head(churn_clean)

Churn <- as.factor(churn_clean$Churn)

#----------------EDA--------------#

#1 contract type customer are less likely to churn

library(ggplot2)

# Create the base plot with Contract on x-axis and fill based on Churn
plotGraph <- ggplot(churn_clean, aes(x = Contract, fill = Churn)) +
  geom_bar() +
  geom_text(
    aes(y = after_stat(count) - 200,
        label = paste0(round(after_stat(prop.table(count)), 4) * 100, '%')),
    stat = 'count',
    position = position_dodge(width = 0.9),
    size = 3
  ) +
  scale_fill_manual(values = c("#F8766D", "#00BFC4")) +
  theme_minimal() +
  labs(
    title = "Churn Distribution by Contract Type",
    x = "Contract Type",
    y = "Count"
  )

# Plot the graph
plot(plotGraph)
```

```
#2. Is there a significant difference in monthly charges for customers with different
    tenure lengths?

library(ggplot2)
library(dplyr)

churn_clean <- churn_clean %>%
  mutate(tenure_group = case_when(
    tenure <= 20 ~ "0-20",
    tenure <= 40 ~ "21-40",
    tenure <= 60 ~ "41-60",
    TRUE ~ "61+"
  ))

anova_result <- lm(MonthlyCharges ~ tenure_group, data = churn_clean)

summary(anova_result)

posthoc_result <- TukeyHSD(aov(anova_result))

print(posthoc_result)

ggplot(churn_clean, aes(x = tenure_group, y = MonthlyCharges)) +
  geom_boxplot(aes(fill = tenure_group), color = "darkblue") +
  scale_fill_brewer(palette = "Set3") +
  labs(title = "Monthly Charges by Tenure Group",
       x = "Tenure Group",
       y = "Monthly Charges") +
  theme_classic()

#3. Which type of internet service influence more on churn rate ? (EDA)

library(ggplot2)

Graph <- ggplot(churn_clean, aes(x = InternetService, fill = Churn)) +
  geom_bar(position = "dodge") +
  geom_text(
    aes(y = after_stat(count) - 200,
        label = paste0(round(after_stat(prop.table(count)), 4) * 100, '%')),
    stat = 'count',
    position = position_dodge(width = 0.9),
    size = 3
  ) +
  scale_fill_manual(values = c("Yes" = "#FF5733", "No" = "#33FF57")) +
  theme_light() +
  labs(
    title = "Churn Distribution by Internet Service",
    x = "Internet Service",
    y = "Count"
  )

plot(Graph)

#----------------correlation analysis-------------#
#Is there a correlation between the customers Contract type and churn rate?

library(ggplot2)
library(reshape2)
library(corrplot)
library(dplyr)
library(plyr)

churn_clean$Churn <- as.numeric(mapvalues(churn_clean$Churn, from=c("No", "Yes"), to=c("0"
    , "1")))
```

```
churn_clean$Contract <- as.numeric(mapvalues(churn_clean$Contract, from=c("Month-to-month", "One year", "Two year"), to=c("0", "1", "2")))

Correlation_matrix <- cor(churn_clean[,c("Contract", "Churn")])

corr_data <- melt(Correlation_matrix)
ggplot(data = corr_data, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0, limit = c(-1, 1), space = "Lab", name="Correlation") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, size = 12, hjust = 1)) +
  coord_fixed()

summary(Correlation_matrix)

churn_correlation_result <- cor.test(churn_clean$Contract, churn_clean$Churn)
print(churn_correlation_result)

extended_correlation_matrix <- churn_clean %>%
  dplyr::select(Contract, MonthlyCharges, Churn) %>%
  cor()

extended_corr_data <- melt(extended_correlation_matrix)
ggplot(data = extended_corr_data, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0, limit = c(-1, 1), space = "Lab", name="Correlation") +
  geom_text(aes(label = round(value, 2)), color = "black", size = 4) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, size = 12, hjust = 1)) +
  coord_fixed()

#-----------------Hypothesis  testing---------------------#
# 6. Customers who have been with the company for a longer time are less likely to leave,

library(ggplot2)

hypertension_group <- churn_clean$tenure[churn_clean$Churn == "1"]
no_hypertension_group <- churn_clean$tenure[churn_clean$Churn == "0"]

t.testing <- t.test(hypertension_group, no_hypertension_group)
print(t.testing)

ggplot(churn_clean, aes(x = as.factor(Churn), y = tenure, fill = as.factor(Churn))) +
  geom_boxplot(outlier.color = "red", outlier.shape = 16, outlier.size = 2) +
  geom_jitter(width = 0.2, alpha = 0.5, color = "black") +
  stat_summary(fun = mean, geom = "point", shape = 18, size = 3, color = "blue", fill = "blue") + labs(title = "Tenure by Churn",
       x = "Churn",
       y = "Tenure") +
  scale_fill_manual(values = c("1" = "#FF9999", "0" = "#9999FF")) +  # Use soft color palette
  theme_classic() +
  theme(legend.position = "none")  # Remove legend since it's not necessary

#--------------------Linear Regression-------------#
#8. Is there any linear relationship between Monthly service Charges
#and the total service charges that Customer are paying

# Perform linear regression
regression_model <- lm(TotalCharges ~ MonthlyCharges + tenure, data = churn_clean)

# Summary of the regression model
summary(regression_model)
```

```
# Plotting the regression line
ggplot(churn_clean, aes(x = MonthlyCharges, y = TotalCharges, color = tenure)) +
  scale_color_gradient(low = "lightblue", high = "red") +  # Gradient from light blue to
      dark blue
  labs(title = "Scatter Plot: MonthlyCharges vs. TotalCharges",
       x = "MonthlyCharges",
       y = "TotalCharges") +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "orange") +
  labs(title = "Linear Regression: MonthlyCharges vs. TotalCharges",
       x = "MonthlyCharges",
       y = "TotalCharges") +
  theme_minimal()
```

.