Telco Customer Churn Analysis

Victoire Migashane

2025-02-07

Question: What is causing Churn?

Data Dictionary

The data comes from Kaggle.com website: Click here to get full data

Feature	Definition
`CustomerID`	Unique customer identifier.
`Count`	Count of records for the customer (always 1).
`Country`	Customer's country.
`State`	Customer's state.
`City`	Customer's city.
`Zip Code`	Customer's postal code.
`Lat Long`	Customer's latitude and longitude.
`Latitude`	Customer's latitude.
`Longitude`	Customer's longitude.
`Gender`	Customer's gender (Male/Female).
`Senior Citizen`	Senior citizen status (1 = Yes, 0 = No).
`Partner`	Whether the customer has a partner (Yes/No).
`Dependents`	Whether the customer has dependents (Yes/No).
`Tenure Months`	Number of months with the company.
`Phone Service`	Whether the customer has phone service (Yes/No).
`Multiple Lines`	Whether the customer has multiple phone lines (Yes/No/No service).
`Internet Service`	Type of internet service (DSL, Fiber, None).
`Online Security`	Whether the customer has online security (Yes/No/No service).
`Online Backup`	Whether the customer has online backup (Yes/No/No service).
`Device Protection`	Whether the customer has device protection (Yes/No/No service).
`Tech Support`	Whether the customer has tech support (Yes/No/No service).
`Streaming TV`	Whether the customer has streaming TV (Yes/No/No service).
`Streaming Movies`	Whether the customer has streaming movies (Yes/No/No service).
`Contract`	Type of contract (Month-to-month, 1 year, 2 years).
`Paperless Billing`	Whether the customer has paperless billing (Yes/No).
`Payment Method`	Payment method (Bank transfer, Credit card, etc.).
`Monthly Charges`	Monthly charge for the service.
`Total Charges`	Total charges during the tenure.
`Churn Label`	Whether the customer churned (Yes/No).
`Churn Value`	Churn indicator (1 = Churned, 0 = Active).
`Churn Score`	Churn prediction score.
`CLTV`	Customer lifetime value (future revenue estimate).
`Churn Reason`	Reason for churn (if applicable).

Wrangle

Dataset Summary

```
# Check the dimensions of the dataset
print(dim(df))  # The dataset contains 7,043 rows and 33 columns
```

```
[1] 7043   33
```

```
# Identify numeric vs. non-numeric (categorical) columns
numeric_cols <- df %>% select(where(is.numeric))  # Select numeric columns
numeric_cols %>% ncol()  # 10 columns are numeric
```

```
[1] 10
```

```
non_numeric_cols <- df %>% select(!where(is.numeric))  # Select non-numeric columns
non_numeric_cols %>% ncol()  # 23 columns are non-numeric (categorical/text)
```

```
[1] 23
```

```
# Display the column names of the dataset
colnames(df)
```

```
 [1] "CustomerID"        "Count"             "Country"
 [4] "State"             "City"              "Zip Code"
 [7] "Lat Long"          "Latitude"          "Longitude"
[10] "Gender"            "Senior Citizen"    "Partner"
[13] "Dependents"        "Tenure Months"     "Phone Service"
[16] "Multiple Lines"    "Internet Service"  "Online Security"
[19] "Online Backup"     "Device Protection" "Tech Support"
[22] "Streaming TV"      "Streaming Movies"  "Contract"
[25] "Paperless Billing" "Payment Method"    "Monthly Charges"
[28] "Total Charges"     "Churn Label"       "Churn Value"
[31] "Churn Score"       "CLTV"              "Churn Reason"
```

```
# Get descriptive statistics for numeric columns
summary(numeric_cols)
```

```
     Count      Zip Code        Latitude       Longitude      Tenure Months
 Min.   :1   Min.   :90001   Min.   :32.56   Min.   :-124.3   Min.   : 0.00
 1st Qu.:1   1st Qu.:92102   1st Qu.:34.03   1st Qu.:-121.8   1st Qu.: 9.00
 Median :1   Median :93552   Median :36.39   Median :-119.7   Median :29.00
 Mean   :1   Mean   :93522   Mean   :36.28   Mean   :-119.8   Mean   :32.37
 3rd Qu.:1   3rd Qu.:95351   3rd Qu.:38.22   3rd Qu.:-118.0   3rd Qu.:55.00
 Max.   :1   Max.   :96161   Max.   :41.96   Max.   :-114.2   Max.   :72.00

 Monthly Charges  Total Charges     Churn Value      Churn Score
 Min.   : 18.25   Min.   :  18.8   Min.   :0.0000   Min.   :  5.0
 1st Qu.: 35.50   1st Qu.: 401.4   1st Qu.:0.0000   1st Qu.: 40.0
```

```
 Median : 70.35   Median :1397.5   Median :0.0000   Median : 61.0
 Mean   : 64.76   Mean   :2283.3   Mean   :0.2654   Mean   : 58.7
 3rd Qu.: 89.85   3rd Qu.:3794.7   3rd Qu.:1.0000   3rd Qu.: 75.0
 Max.   :118.75   Max.   :8684.8   Max.   :1.0000   Max.   :100.0
                  NA's   :11
      CLTV
 Min.   :2003
 1st Qu.:3469
 Median :4527
 Mean   :4400
 3rd Qu.:5380
 Max.   :6500
```

```r
# Count the number of missing values in each column
colSums(is.na(df))  # 'Total Charges' has 11 missing values, 'Churn Reason' has 5,174 missing values
```

```
      CustomerID            Count          Country            State
               0                0                0                0
            City         Zip Code         Lat Long         Latitude
               0                0                0                0
       Longitude           Gender   Senior Citizen          Partner
               0                0                0                0
      Dependents    Tenure Months    Phone Service   Multiple Lines
               0                0                0                0
Internet Service  Online Security    Online Backup Device Protection
               0                0                0                0
    Tech Support     Streaming TV Streaming Movies         Contract
               0                0                0                0
Paperless Billing   Payment Method  Monthly Charges     Total Charges
               0                0                0               11
     Churn Label      Churn Value      Churn Score             CLTV
               0                0                0                0
    Churn Reason
            5174
```

Observations

- **Total Rows**: 7,043
- **Total Columns**: 33
- **Numeric Columns**: 10
- **String (Object) Columns**: 23
- **Missing Values**:
  - `Total Charges`: 11 null values
  - `Churn Reason`: 5,174 null values
- **Numeric Data**: Descriptive statistics available

Preparation

Data Cleaning

```
# Check the number of unique values in specific columns
unique(df$Count)  # Only 1 unique value, so this column will be removed
```

```
[1] 1
```

```
unique(df$Country)  # Only 1 unique value, so this column will be removed
```

```
[1] "United States"
```

```
unique(df$State)  # Only 1 unique value, so this column will be removed
```

```
[1] "California"
```

```
# Define columns to remove
removed_cols <- c("Count", "Country", "State", "Lat Long", "Latitude", "Longitude", "Churn Reason")

# Remove the unwanted columns
df <- df %>% select(-removed_cols)
```

```
Warning: Using an external vector in selections was deprecated in tidyselect 1.1.0.
i Please use ‘all_of()‘ or ‘any_of()‘ instead.
  # Was:
  data %>% select(removed_cols)

  # Now:
  data %>% select(all_of(removed_cols))

See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
This warning is displayed once every 8 hours.
Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
generated.
```

```
# Remove columns that still have missing values
df <- df[, colSums(is.na(df)) == 0]

# Check the new size of the dataframe
df %>% dim()  # Displays the updated number of rows and columns
```

```
[1] 7043   25
```

```
# Replace spaces in column names with underscores for easier referencing
colnames(df) <- gsub(" ", "_", colnames(df))

# Identify numeric vs. non-numeric (categorical) columns
numeric_cols <- df %>% select(where(is.numeric))  # Select numeric columns
numeric_cols %>% ncol()  # 10 columns are numeric
```

```
[1] 6
```

```
non_numeric_cols <- df %>% select(!where(is.numeric))  # Select non-numeric columns
non_numeric_cols %>% ncol()  # 23 columns are categorical (string/objects)
```

```
[1] 19
```

Data Cleaning Summary

Dropped Columns:

- **`count`**: Only had one value (1).
- **`Country`**: Only had one value (United States).
- **`State`**: Only had one value (California).
- **`Lat Long`**, **`Latitude`**, **`Longitude`**: Not analyzing location at the moment.
- **`Churn Reason`**: Had too many missing values.

New Dataset Summary:

- **Total Rows**: 7,043
- **Total Columns**: 25
- **Numeric Columns**: 6
- **Categorical Columns**: 19

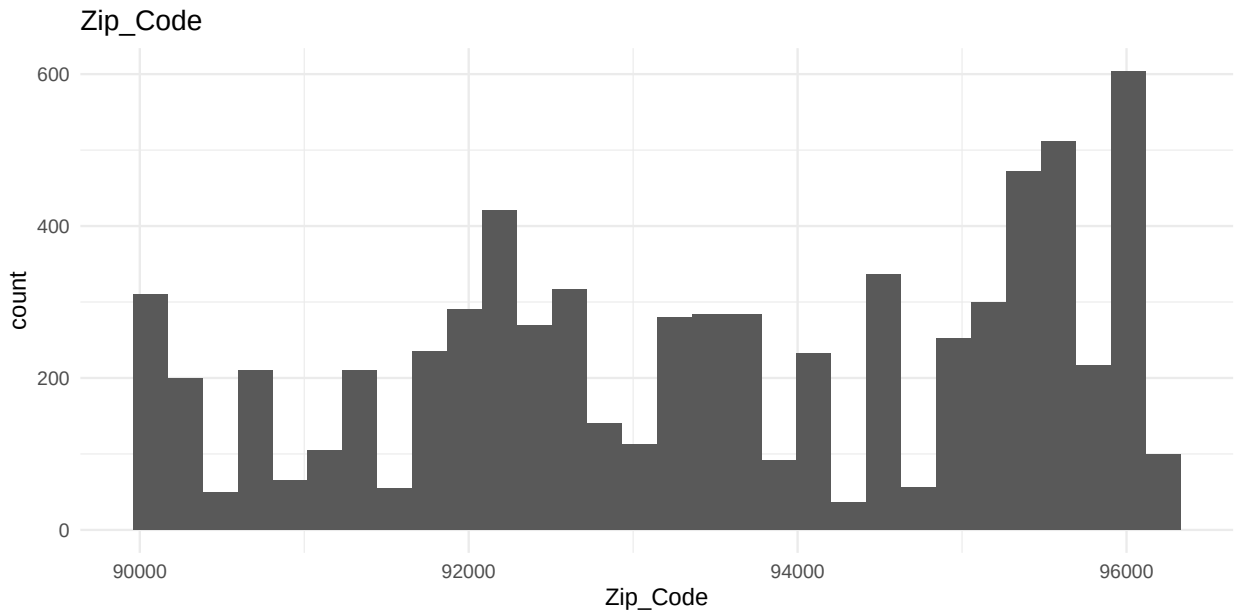
Visualize the columns to try and understand them better

The start by looking at the distributions of the 6 numerical comuns

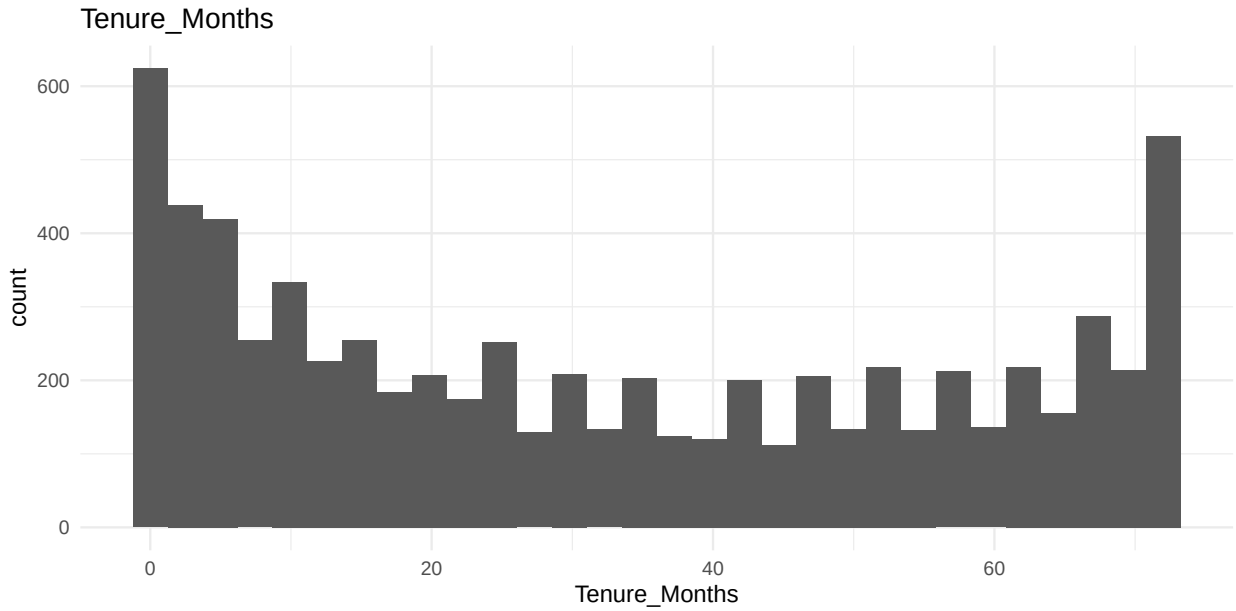
```
# Look at the distribution of the numeric columns
for (col in colnames(numeric_cols)) {
  fig <- ggplot(df, aes_string(x = col)) +
  geom_histogram() +
  labs(title = col) +
  theme(plot.title = element_text(hjust = 0.5)) + # center the title
  theme_minimal()

  print(fig)
}
```

```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with `aes()`.
## i See also `vignette("ggplot2-in-packages")` for more information.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

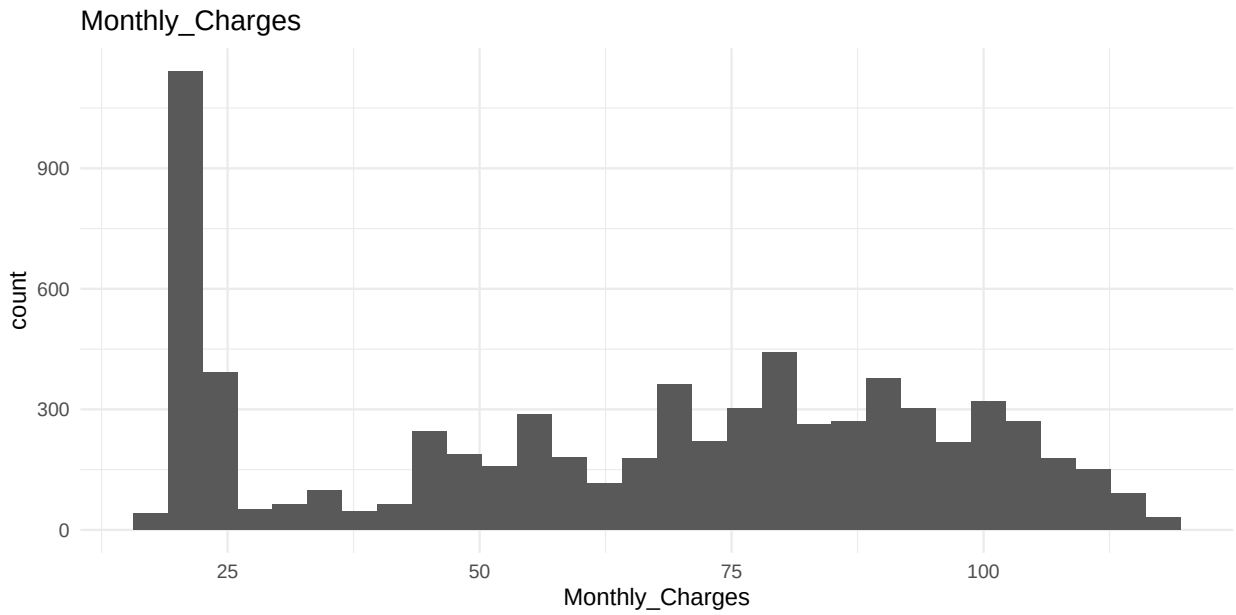
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



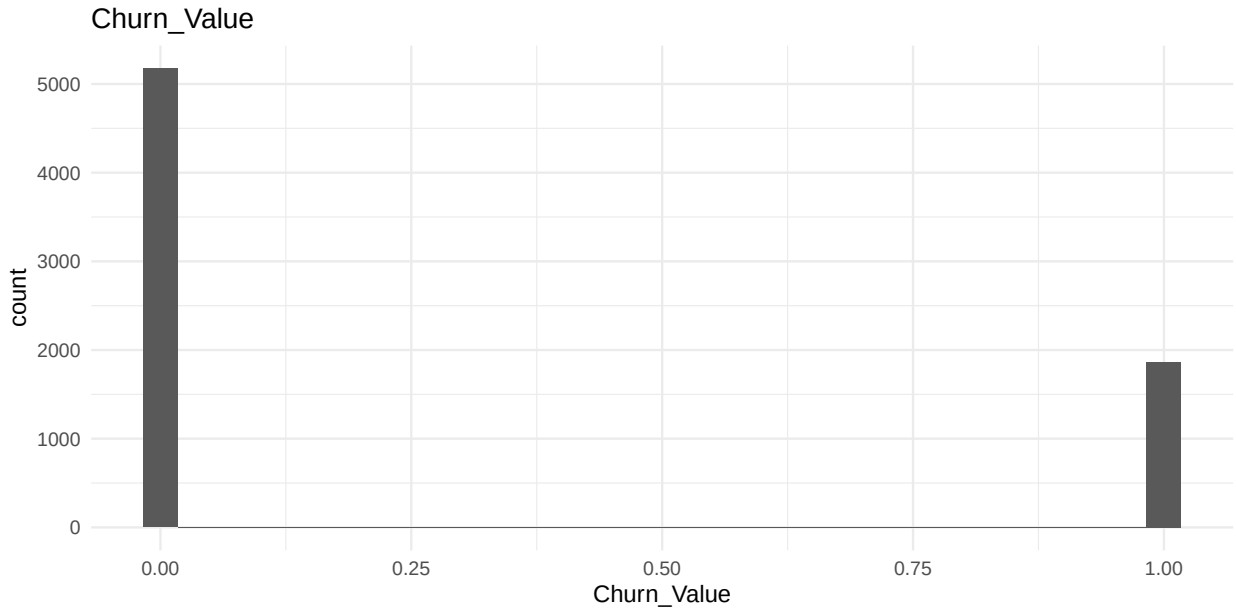

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



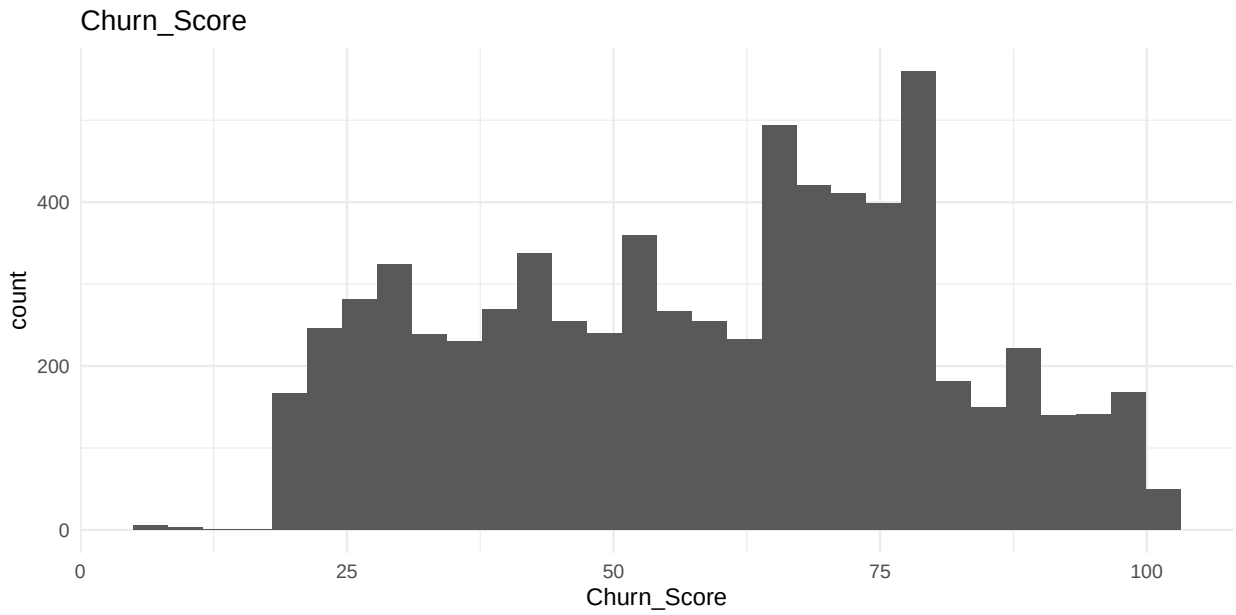

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



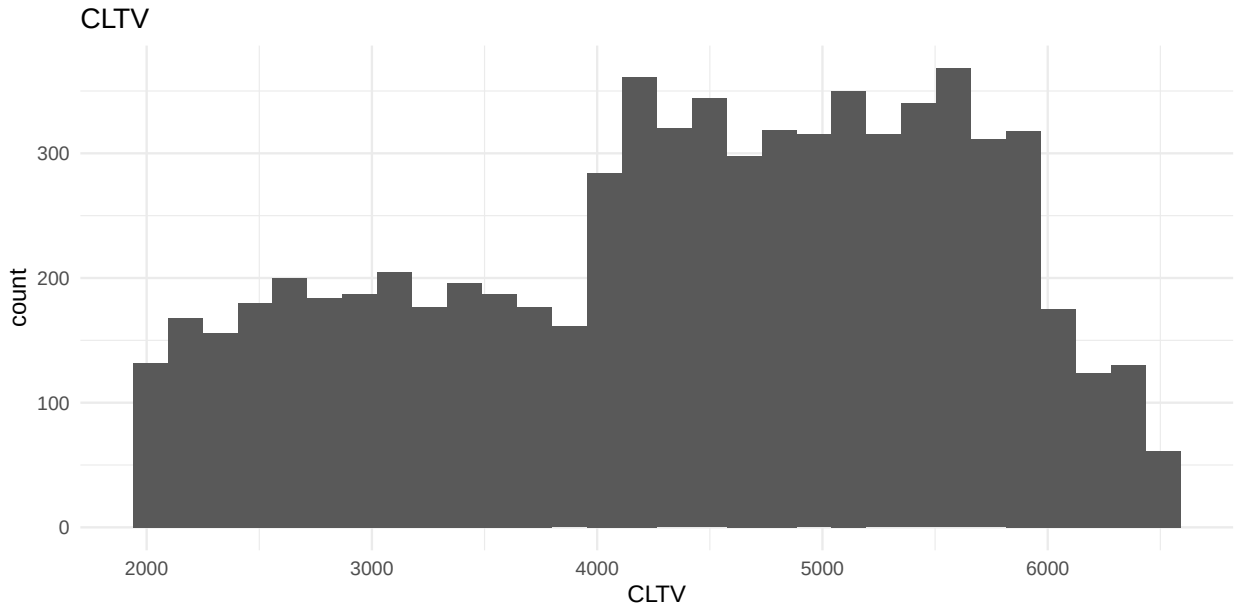

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

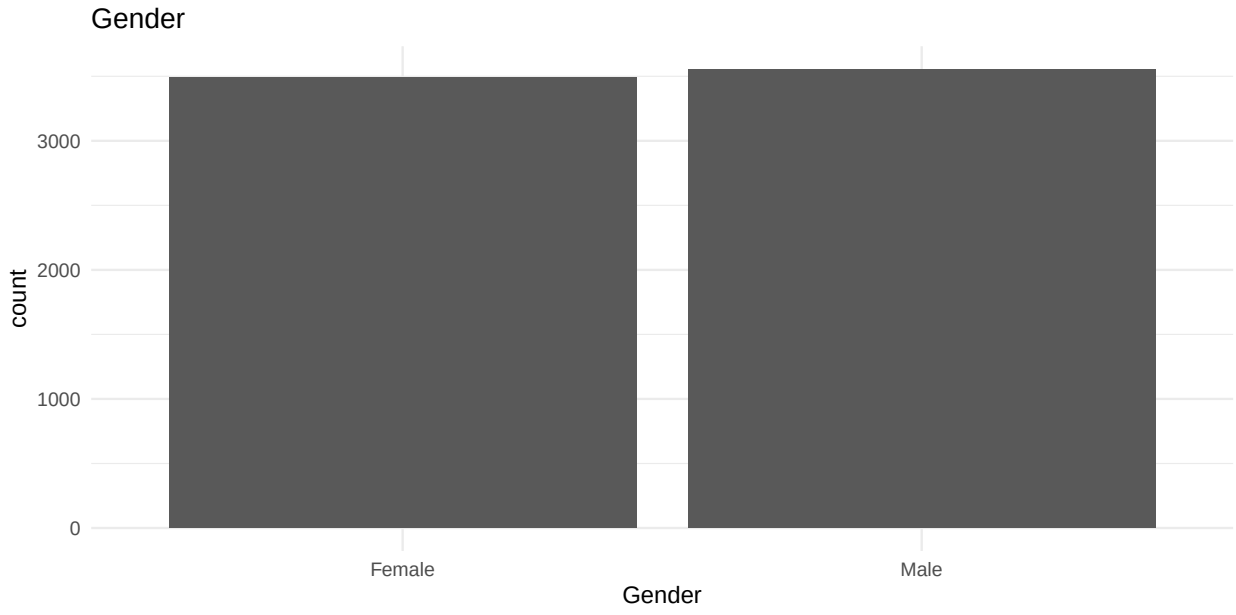


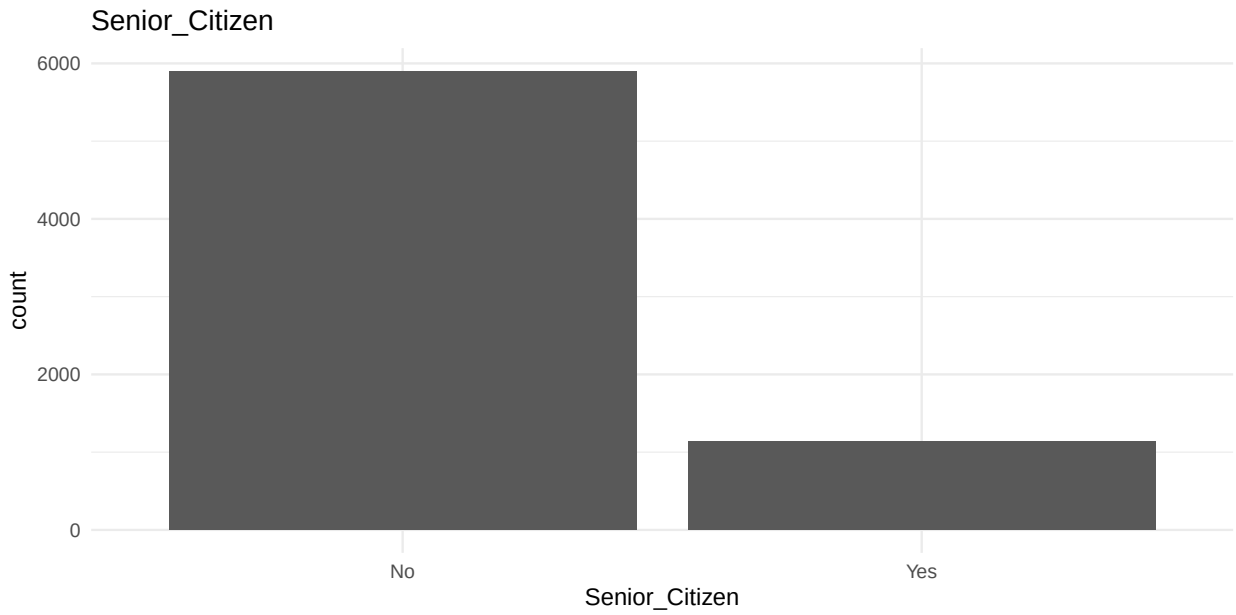
Observations

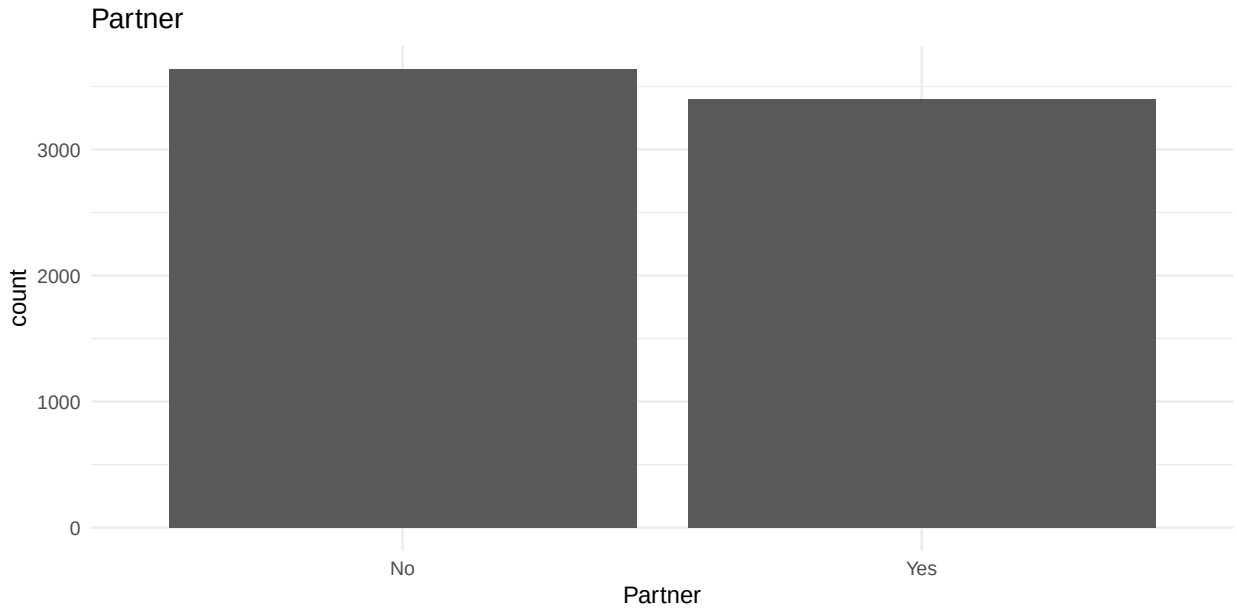
- **`Zip Code`**: Not very significant in the distribution.
- **`Tenure Months`**: Displays a distribution where outcomes are more likely to occur at the extremes.
- **`Monthly Charges`**: Appears to have a somewhat bell-shaped or normal distribution.
- **`Churn Value`**: About a third of all customers have churned in this dataset.

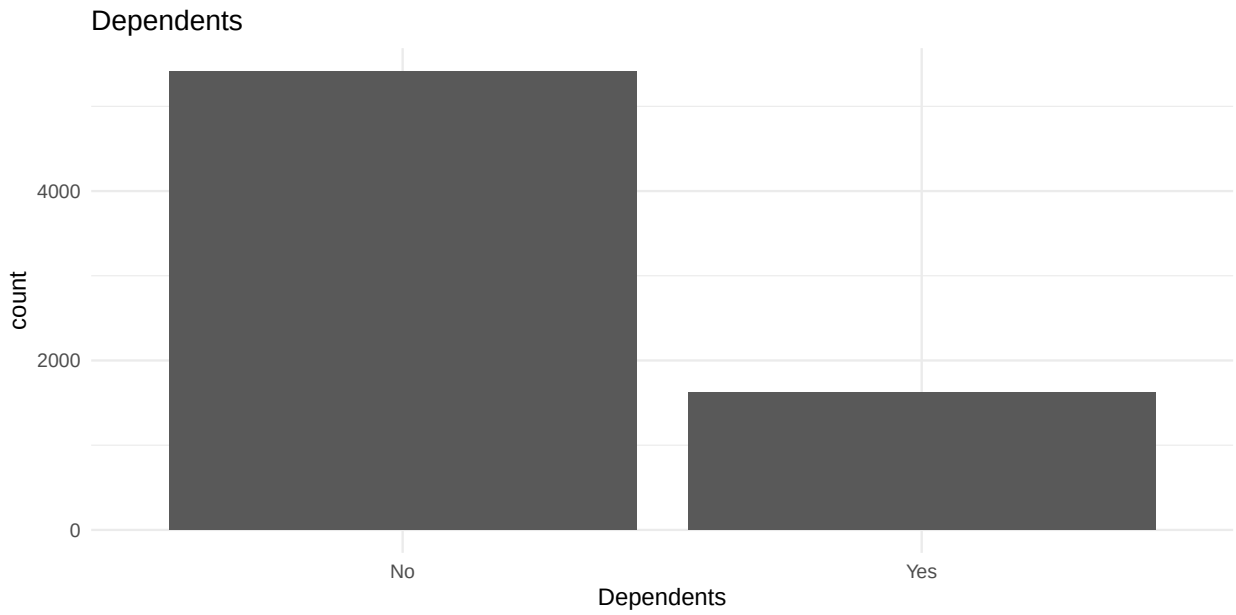
- **`Churn Score`**: Shows some characteristics of a normal distribution but is not fully normal.
- **`CLTV`**: Shows some characteristics of a normal distribution but is not fully normal.

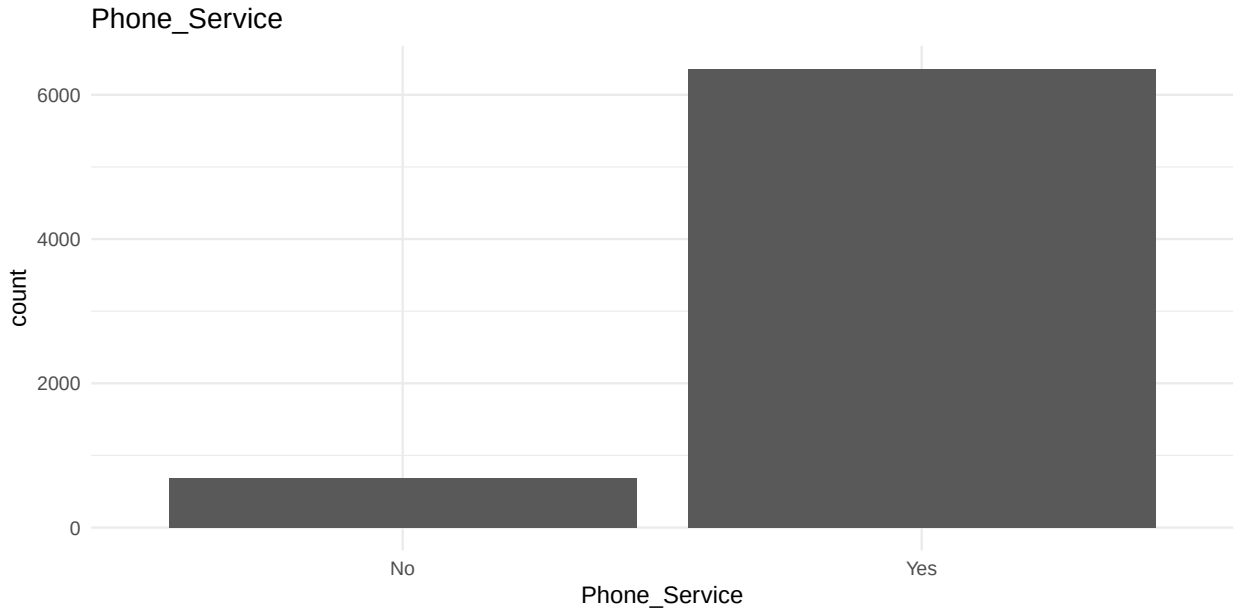
```
for (col in colnames(non_numeric_cols)) {
  if (length(unique(df[[col]])) < 10) {
    fig <- ggplot(df, aes(x = .data[[col]])) +
      geom_bar() +  # Use geom_bar() for categorical data
      labs(title = col) +
      theme(plot.title = element_text(hjust = 0.5)) +
      theme_minimal()

    print(fig)
  }
}
```

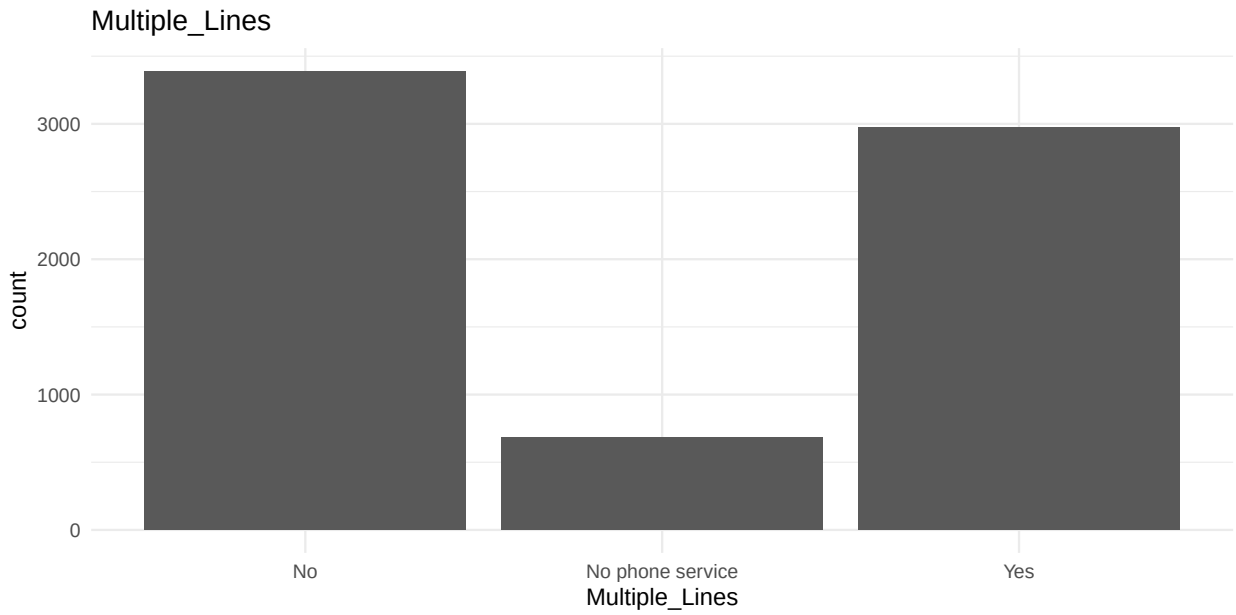


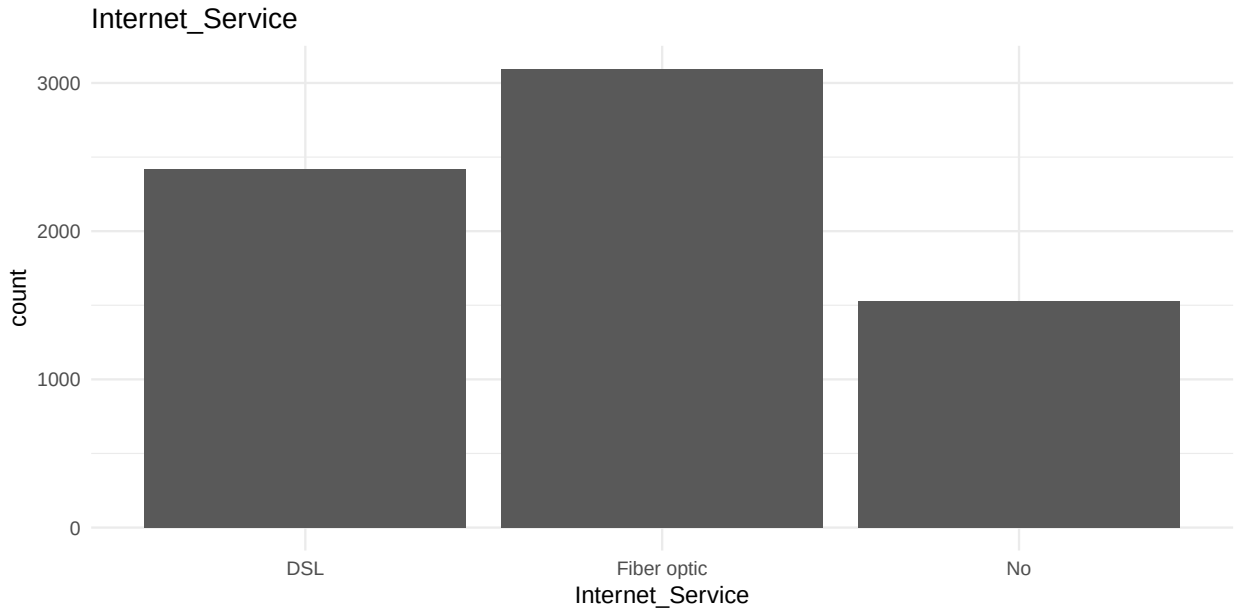

Senior_Citizen
6000
4000
2000
0
count
No
Yes
Senior_Citizen

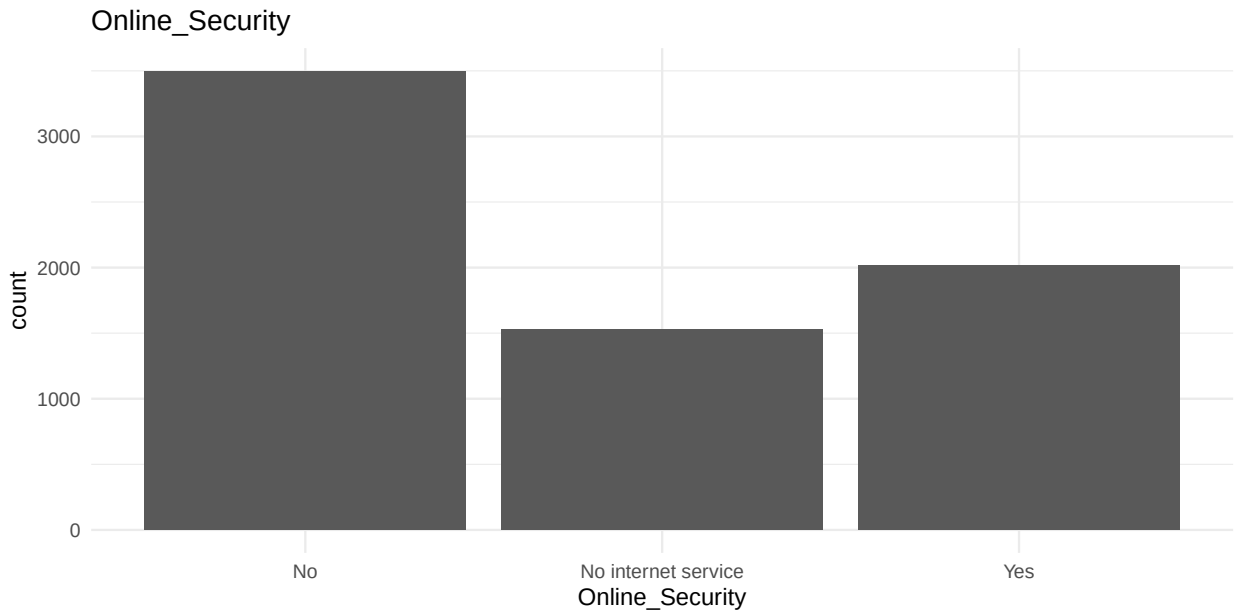

Partner
3000
2000
1000
0
count
No
Yes
Partner

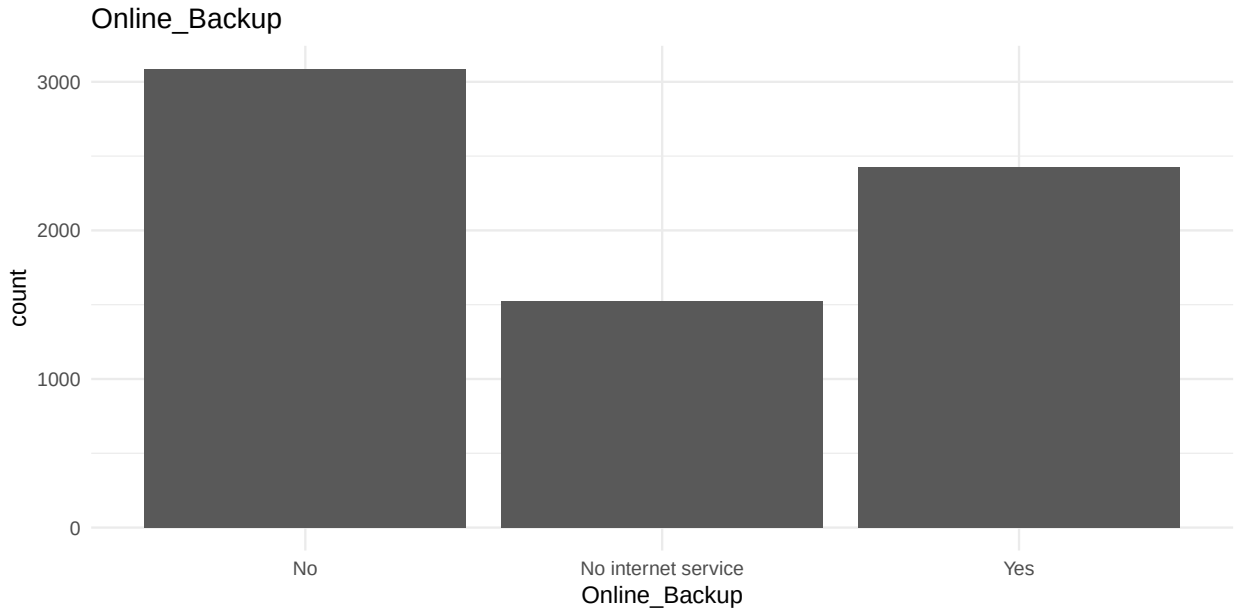

Dependents
4000
2000
0
count
No
Yes
Dependents

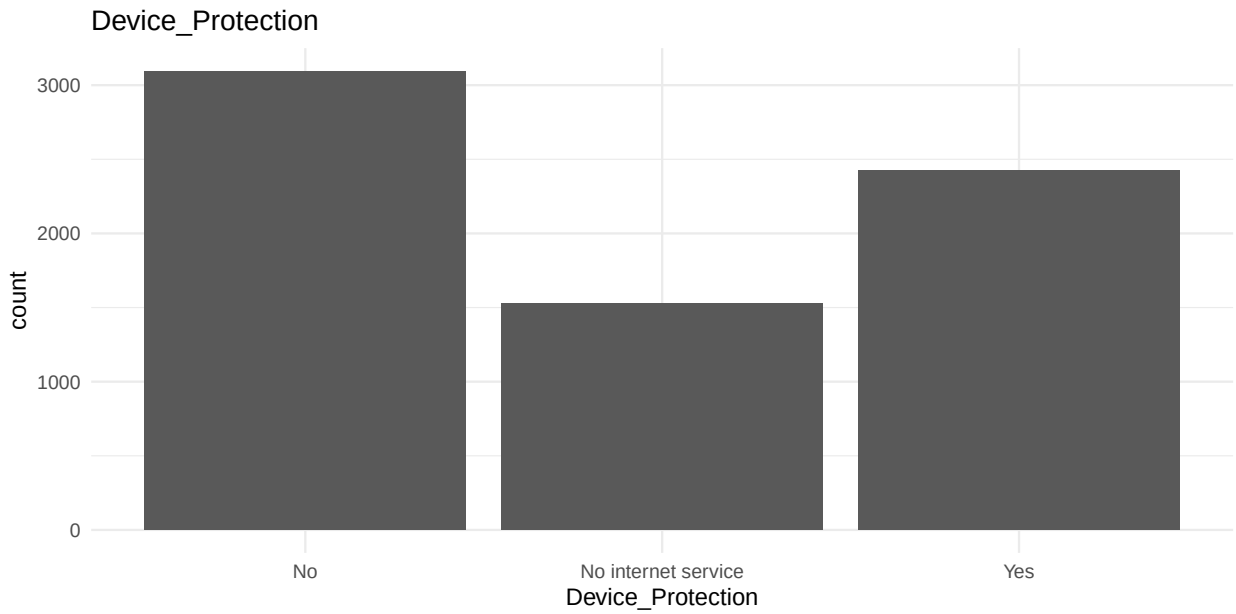

Phone_Service
6000
4000
2000
0
count
No
Yes
Phone_Service

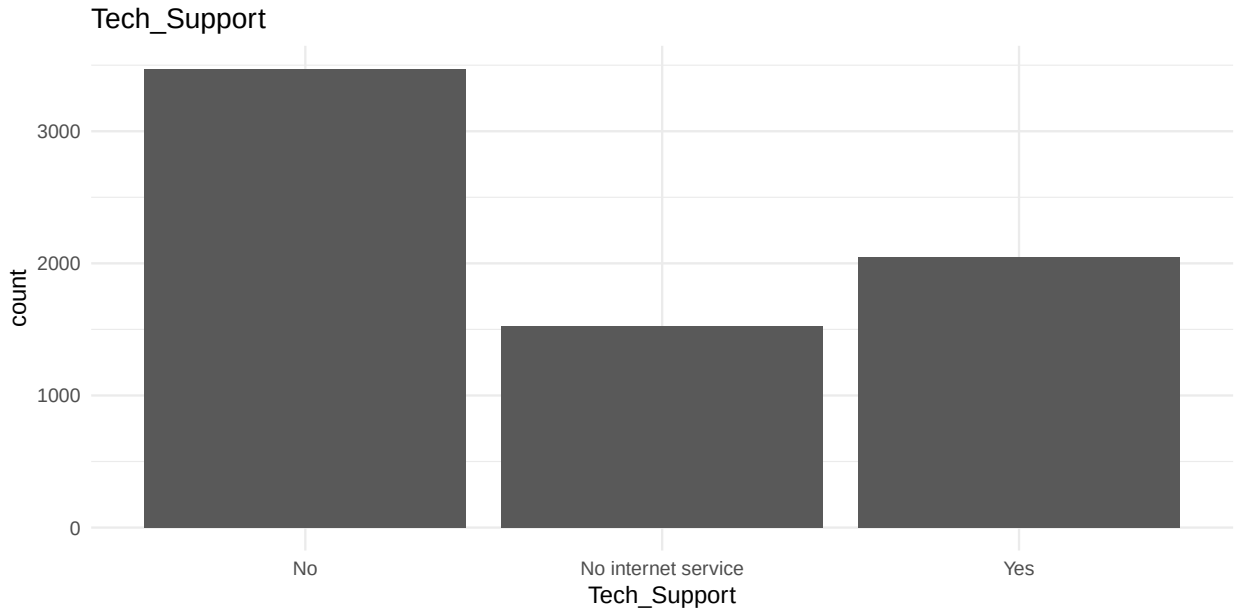

Multiple_Lines
3000
2000
1000
0
count
No
No phone service
Yes
Multiple_Lines

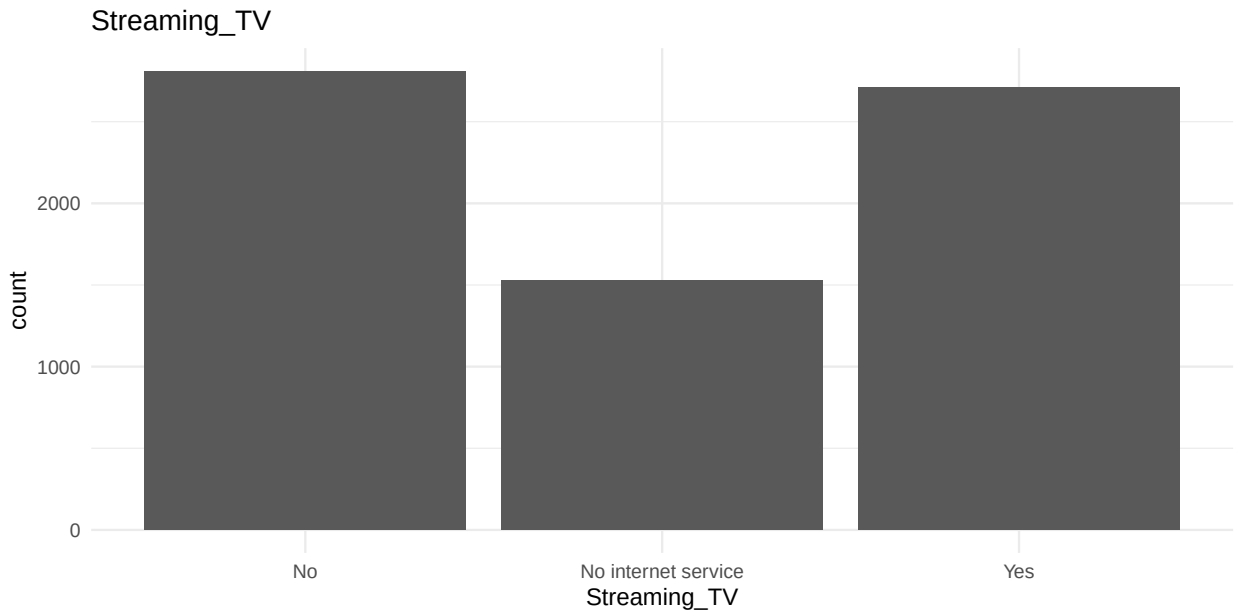

Internet_Service
3000
2000
1000
0
count
DSL
Fiber optic
No
Internet_Service

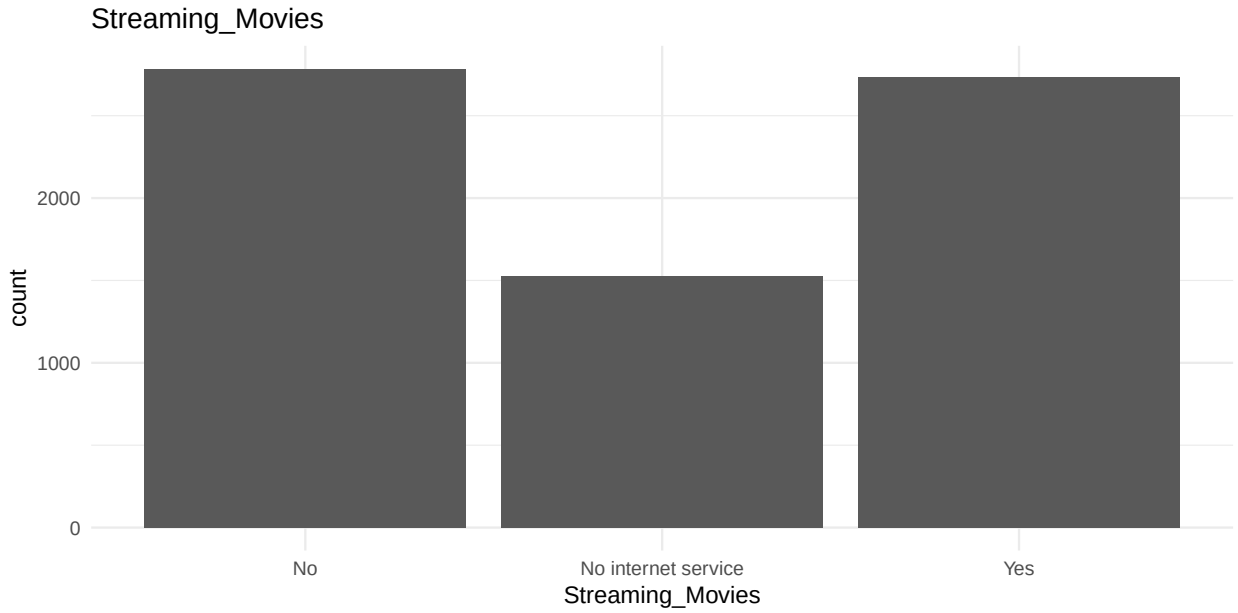

Online_Security
3000
2000
1000
0
count
No
No internet service
Yes
Online_Security

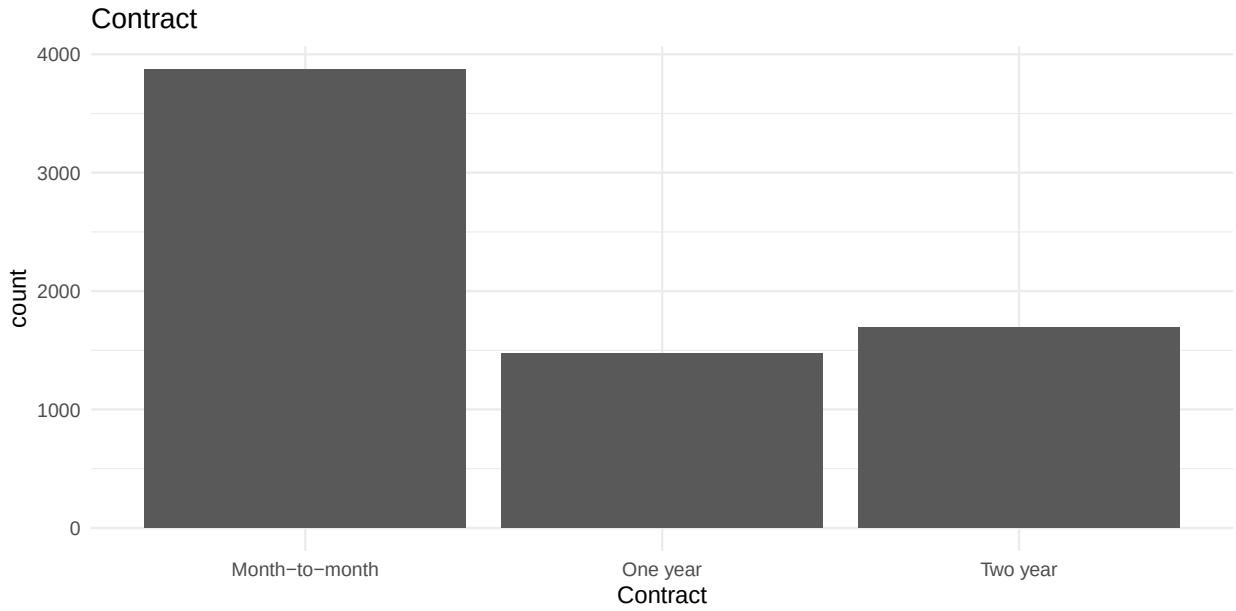

Online_Backup
3000
2000
1000
0
count
No
No internet service
Yes
Online_Backup

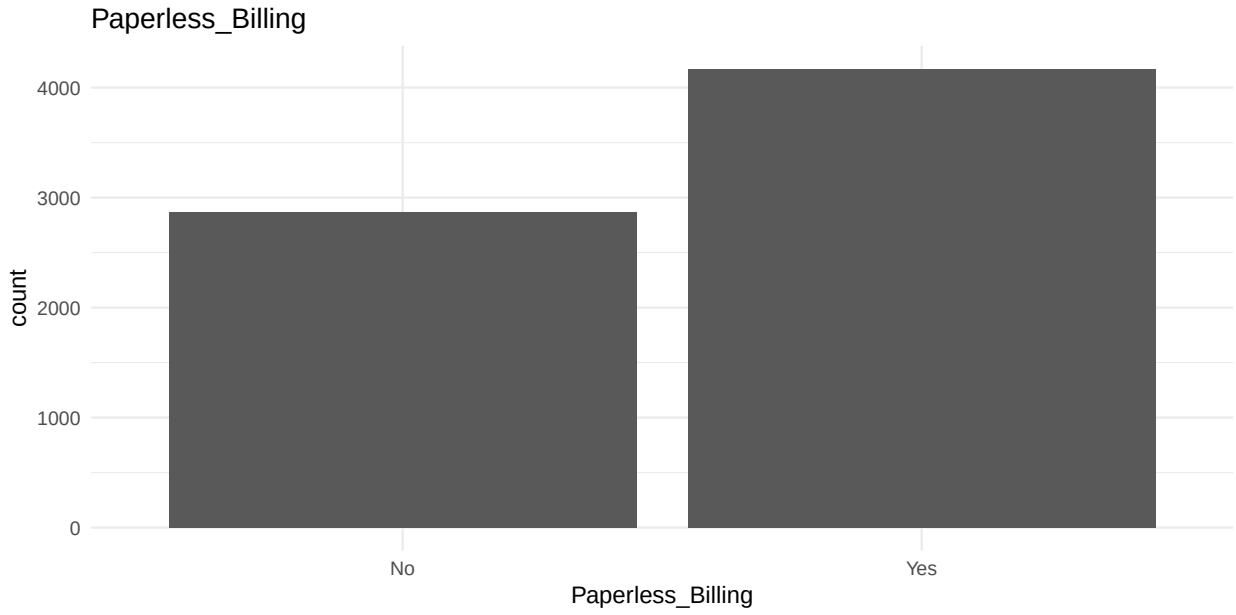

Device_Protection
3000
2000
1000
0
count
No
No internet service
Yes
Device_Protection

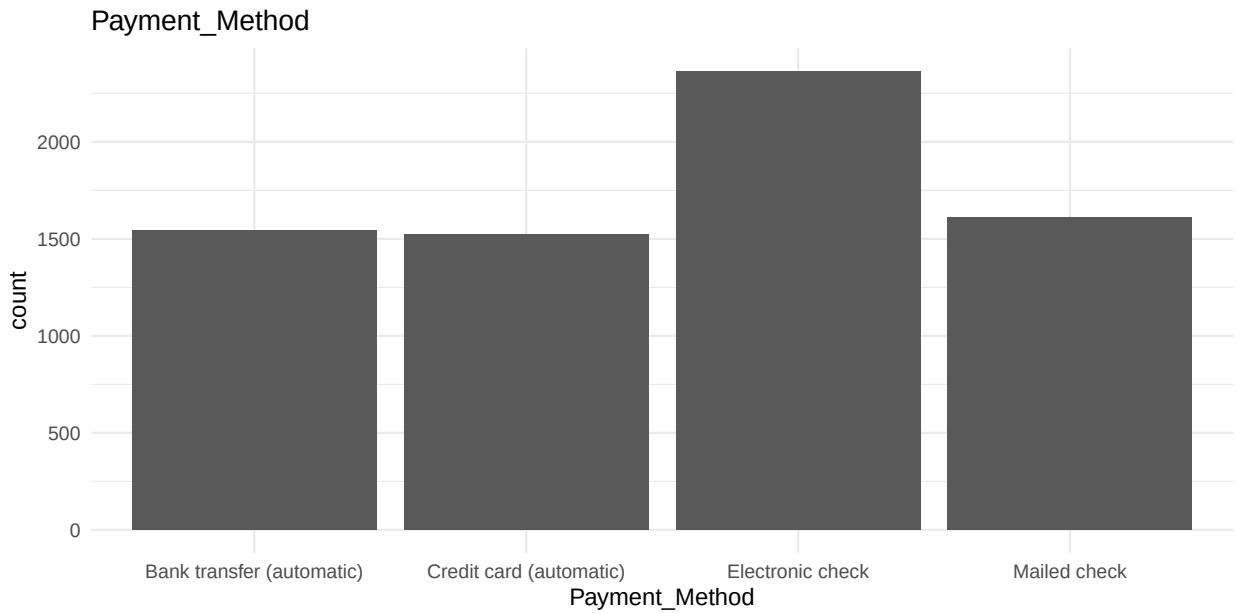

Tech_Support
3000
2000
1000
0
count
No
No internet service
Yes
Tech_Support

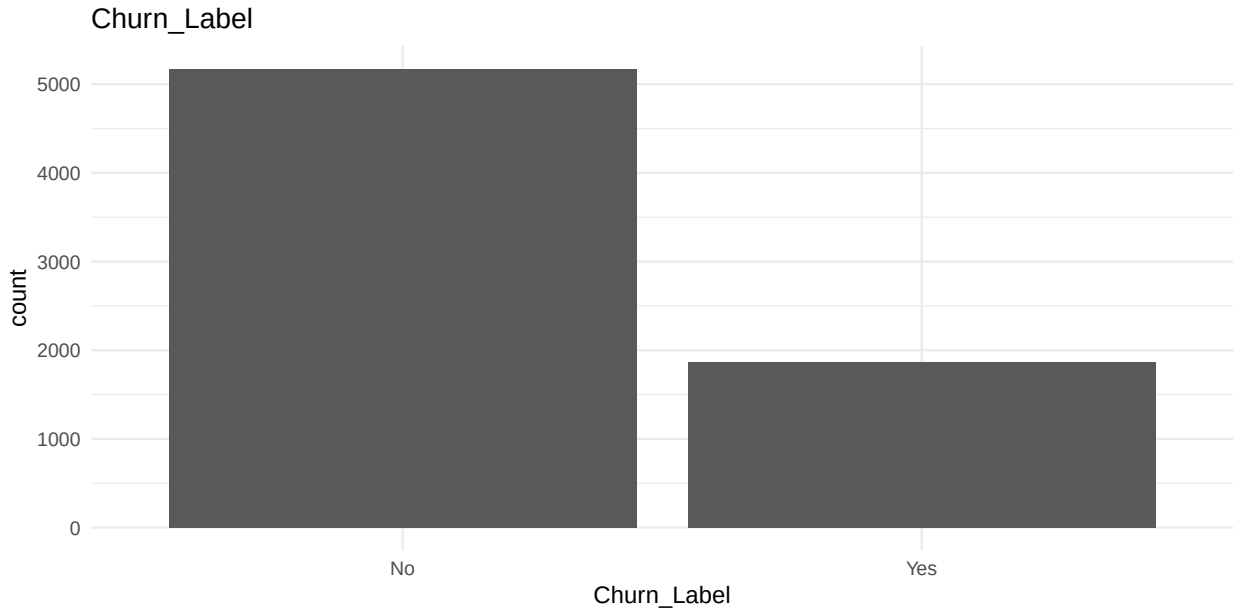

Streaming_TV
2000
1000
0
count
No
No internet service
Yes
Streaming_TV


Streaming_Movies
2000
1000
0
count
No
No internet service
Yes
Streaming_Movies


Contract
4000
3000
2000
1000
0
count
Month−to−month
One year
Two year
Contract


Paperless_Billing
4000
3000
2000
1000
0
count
No
Yes
Paperless_Billing





Observations

- **`Gender`**: Customers are almost evenly split between both genders.
- **`Senior Citizen`**: The majority of customers are not senior citizens.
- **`Phone Service`**: Most customers have phone service.
- **`Multiple Lines`**: Most customers do not have multiple lines, but a significant portion does. Some have no phone service.
- **`Internet Service`**: Most customers prefer fiber optic internet.

- **`Online Security`**: Customers show little interest in online security services.
- **`Online Backup`**: Low interest, possibly for similar reasons as online security.
- **`Device Protection`**: Similar to online security, customers show low interest.
- **`Tech Support`**: About half of the customers with internet service receive tech support.
- **`Contract`**: The majority of customers are on a month-to-month contract.
- **`Paperless Billing`**: Most customers prefer paperless billing.
- **`Payment Method`**: Electronic checks are the most commonly used payment method.
- **`Churn Label`**: About a third of all customers have churned in this dataset.

Exploritory Analysis

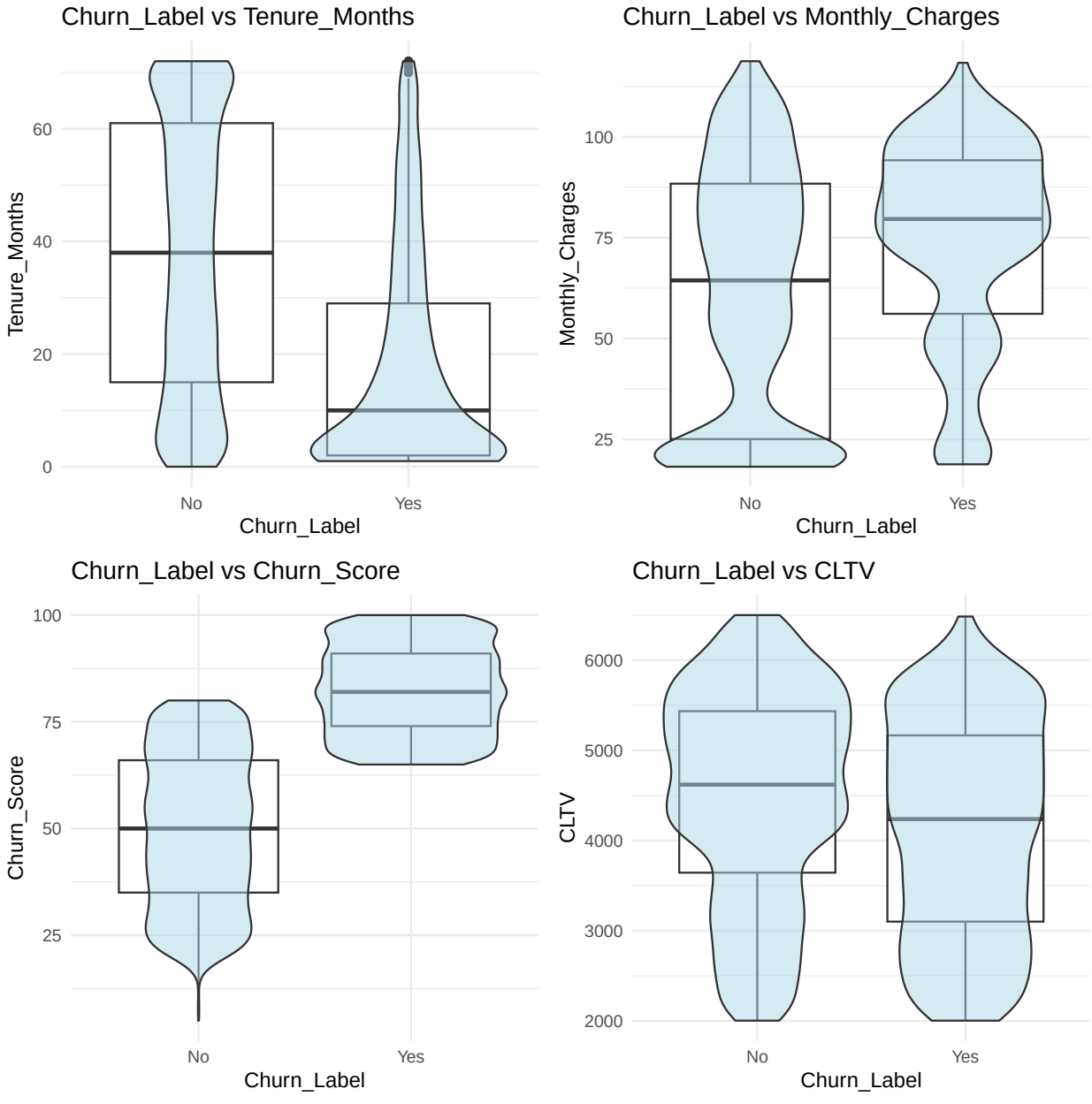
Features of Interest dictionary

Feature	Definition
`Churn Label`	Whether the customer churned (Yes/No).
`Gender`	Customer's gender (Male/Female).
`Churn Score`	Churn prediction score.
`Senior Citizen`	Senior citizen status (1 = Yes, 0 = No).
`Contract`	Type of contract (Month-to-month, 1 year, 2 years).
`Paperless Billing`	Whether the customer has paperless billing (Yes/No).
`Payment Method`	Payment method (Bank transfer, Credit card, etc.).
`Monthly Charges`	Monthly charge for the service.
`CLTV`	Customer lifetime value (future revenue estimate).
`Phone Service`	Whether the customer has phone service (Yes/No).
`Multiple Lines`	Whether the customer has multiple phone lines (Yes/No/No service).
`Internet Service`	Type of internet service (DSL, Fiber, None).
`Tech Support`	Whether the customer has tech support (Yes/No/No service).
`Tenure Months`	Number of months with the company.

I will visualize each feature (y-axis) with my targe variable churn (x-axis) using box plot, violine plot, bar plot to full see the how each feature relate with my target variable.

```
# Define numeric features to evaluate
num_features <- c("Tenure_Months", "Monthly_Charges", "Churn_Score", "CLTV")

# Create an empty list to store plots
plot_list <- list()

# Generate boxplot and violin plot for each numeric feature
for (feature in num_features) {
  p <- ggplot(df, aes(x = factor(Churn_Label), y = .data[[feature]])) +
    geom_boxplot() +
    geom_violin(alpha = 0.5, fill = "lightblue") +
```

```
    labs(title = paste("Churn_Label vs", feature), x = "Churn_Label", y = feature) +
    theme_minimal()
  plot_list <- append(plot_list, list(p))  # Store each plot in the list
}

# Arrange and display two plots per page
for (i in seq(1, length(plot_list), by = 2)) {
  grid.arrange(plot_list[[i]], plot_list[[i + 1]], ncol = 2)
}
```



Observations

- **Tenure Month vs Churn Label**:
  Like we originally saw in the histogram, the `No` results are very balanced with a U-shape, showing outcomes happening at the extremes. However, the `Yes` category shows a large concentration of churn between 0 and 30 months as a Telco customer. This is significant because it could indicate that most customers churn before reaching three years with Telco. **This brings up the question: Why are they leaving the company so soon?**

- **Monthly Charges vs Churn Label**:
  The focus of this plot is the violin plot, which provides insight into data density. Looking at the shape for `No`, we see that results are spread over a wide range of data points, and the box plot confirms this. The `Yes` category, however, shows a large concentration of data points at the higher end of monthly charges. This is very informative because it could suggest that **customers churn due to high monthly costs**.

- **Churn Score vs Churn Label**:
  The churn score is assigned to every customer based on historical data and demographics, representing Telco's prediction of the likelihood that a customer will churn. This plot shows how accurate Telco's predictions are, as there is little overlap between high and low churn scores. This suggests that **the higher the churn score, the higher the likelihood of churn**.

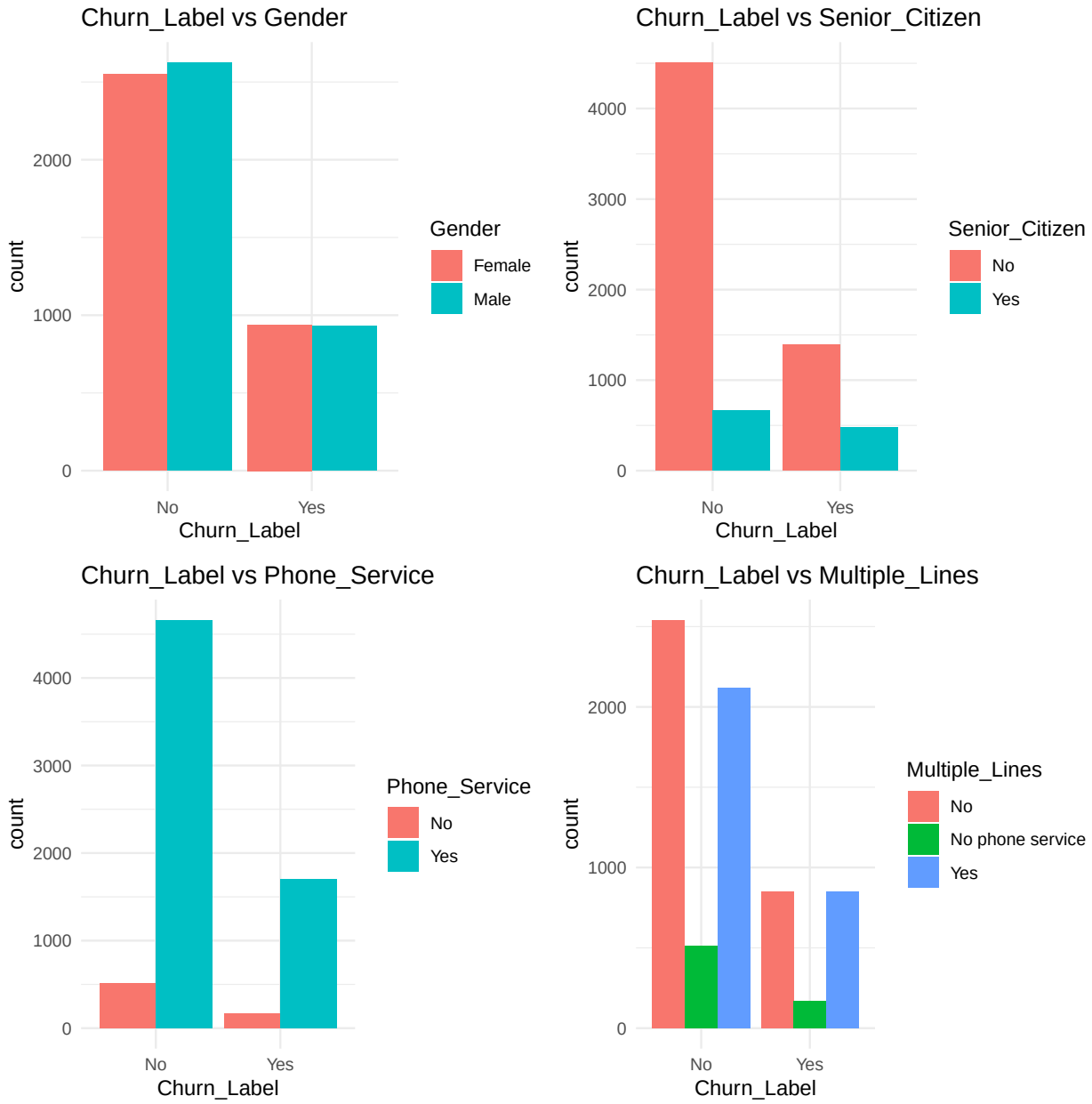
- **CLTV vs Churn Label**:
  CLTV represents a customer's estimated lifetime revenue. It is also a prediction made by Telco's algorithms to understand the value of their customers. While there is not a significant difference between the two groups, one might argue that **customers who churn tend to have high lifetime revenue expectations from Telco**. However, the difference between medians is not substantial, so hypothesis testing could determine whether this difference is statistically significant.

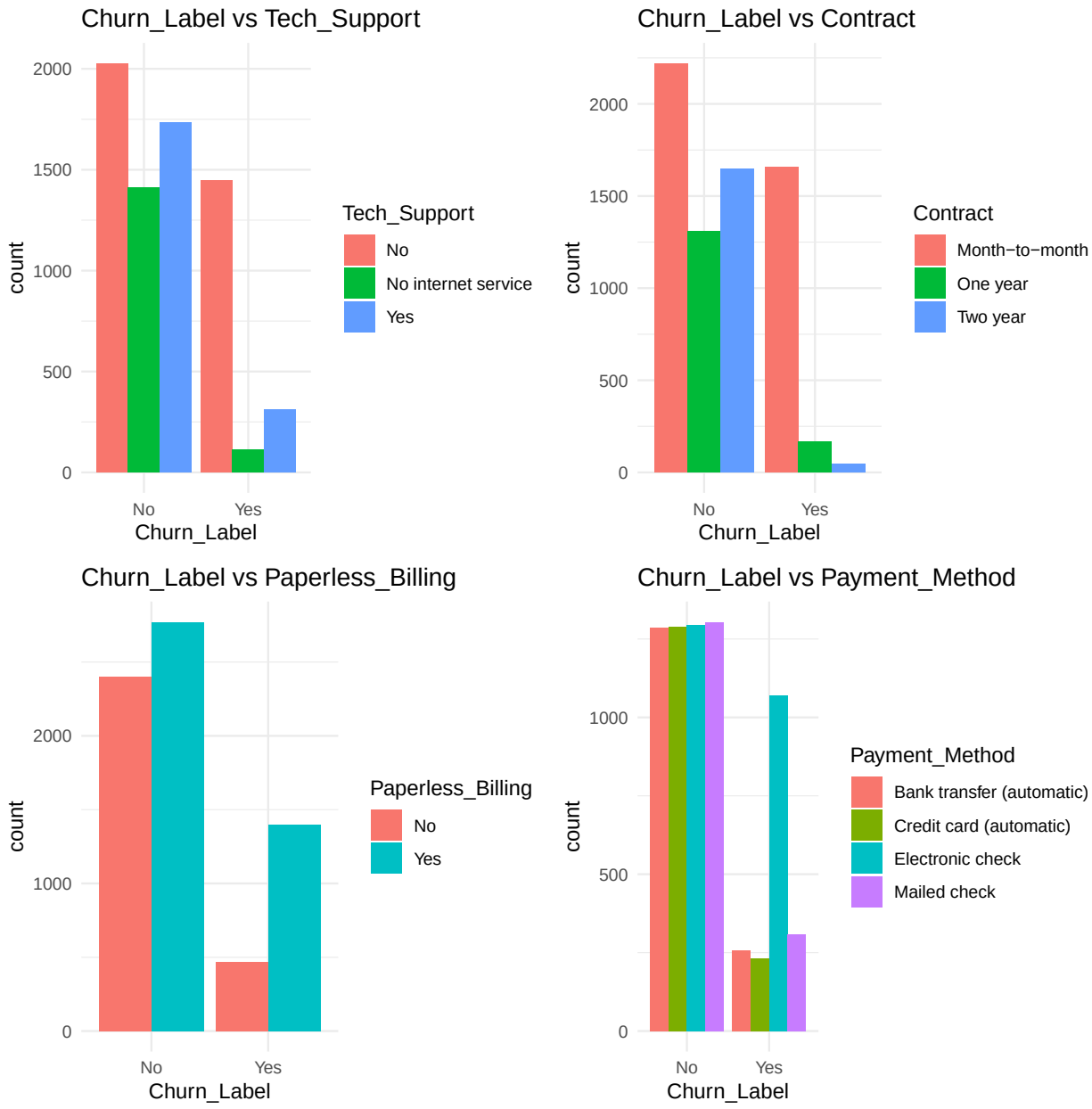
Important Features

Feature	Definition
`Churn Score`	Churn prediction score.
`Monthly Charges`	Monthly charge for the service.
`CLTV`	Customer lifetime value (future revenue estimate).
`Tenure Months`	Number of months with the company.

```
# Create an empty list to store plots
plot_list <- list()

# Define categorical features to evaluate
cat_features <- c("Gender", "Senior_Citizen", "Phone_Service", "Multiple_Lines", "Tech_Support", "Contra

# Generate bar plots for each categorical feature
for (feature in cat_features) {
  p <- ggplot(df, aes(x = factor(Churn_Label), fill = factor(.data[[feature]]))) +
    geom_bar(position = "dodge") +
    labs(title = paste("Churn_Label vs", feature), x = "Churn_Label", fill = feature) +
    theme_minimal()
  plot_list <- append(plot_list, list(p))  # Store each plot in the list
}
```

```r
# Arrange and display two plots per page
for (i in seq(1, length(plot_list), by = 2)) {
  grid.arrange(plot_list[[i]], plot_list[[i + 1]], ncol = 2)
}
```





Observations

- **Counts of Churn Label vs Gender**:
  There are no significant gender differences related to customer churn. **More customers stay with Telco than those who leave.**

- **Counts of Churn Label vs Senior Citizen**:
  The majority of customers are not seniors. Non-senior citizens churn more than senior citizens, but this might be due to their larger presence in the dataset. Senior citizens churn at about the same volume as those who don't. **I don't think there is much significance here.**

- **Counts of Churn Label vs Phone Service**:
  Only a small number of customers with phone service churn compared to those without it. This raises

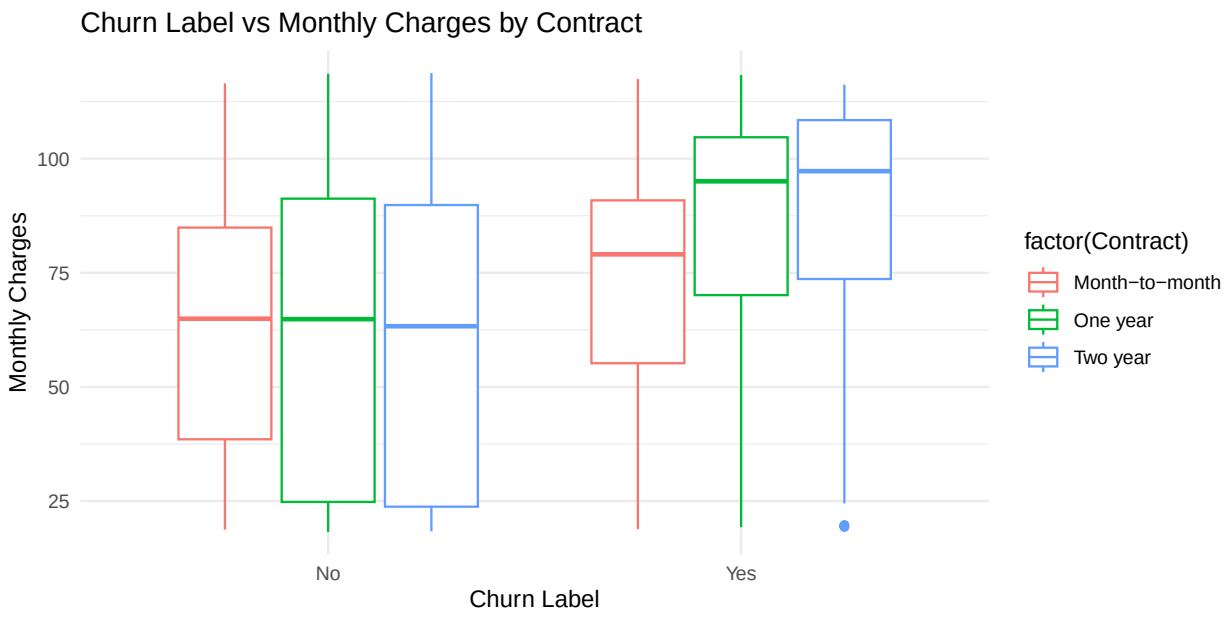
questions about customer satisfaction. **Are customers churning due to poor phone service coverage?**

- **Counts of Churn Label vs Multiple Lines**:
  Customers who churn with multiple lines are about the same as those who churn with only a single line. **It doesn't seem like the number of lines is a major factor in churn.**
- **Counts of Churn Label vs Tech Support**:
  It is very clear that customers who churn have overwhelmingly received no technical support, while the opposite is true for those who stay. **Why didn't these customers receive any technical support?**
- **Counts of Churn Label vs Contract**:
  Month-to-month customers are churning at an extreme rate. **This could be a significant feature in predicting churn.**
- **Counts of Churn Label vs Paperless Billing**:
  Paperless billing appears to play some role in customer churn, but its significance is unclear. **I think hypothesis testing would help a lot in understanding how important this feature is.**
- **Counts of Churn Label vs Payment Method**:
  Customers who use electronic checks churn at an extreme rate compared to every other group. **What is it about electronic checks that pushes customers away?**

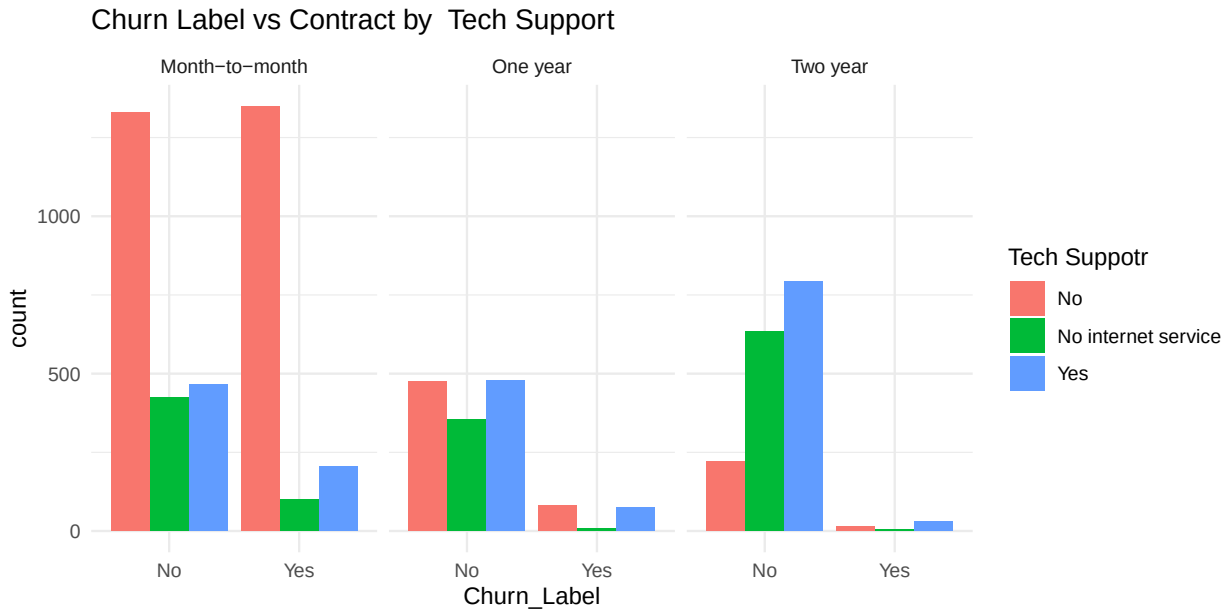
Important Features

Feature	Definition
`Contract`	Type of contract (Month-to-month, 1 year, 2 years).
`Payment Method`	Payment method (Bank transfer, Credit card, etc.).
`Phone Service`	Whether the customer has phone service (Yes/No).
`Tech Support`	Whether the customer has tech support (Yes/No/No service).

```
# Boxplot of Monthly_Charges vs Churn_Label, colored by Internet_Service
ggplot(df, aes(x = factor(Churn_Label), y = Monthly_Charges, color = factor(Contract))) +
  geom_boxplot() +
  labs(title = "Churn Label vs Monthly Charges by Contract",
       x = "Churn Label", y = "Monthly Charges") +
  theme_minimal()
```

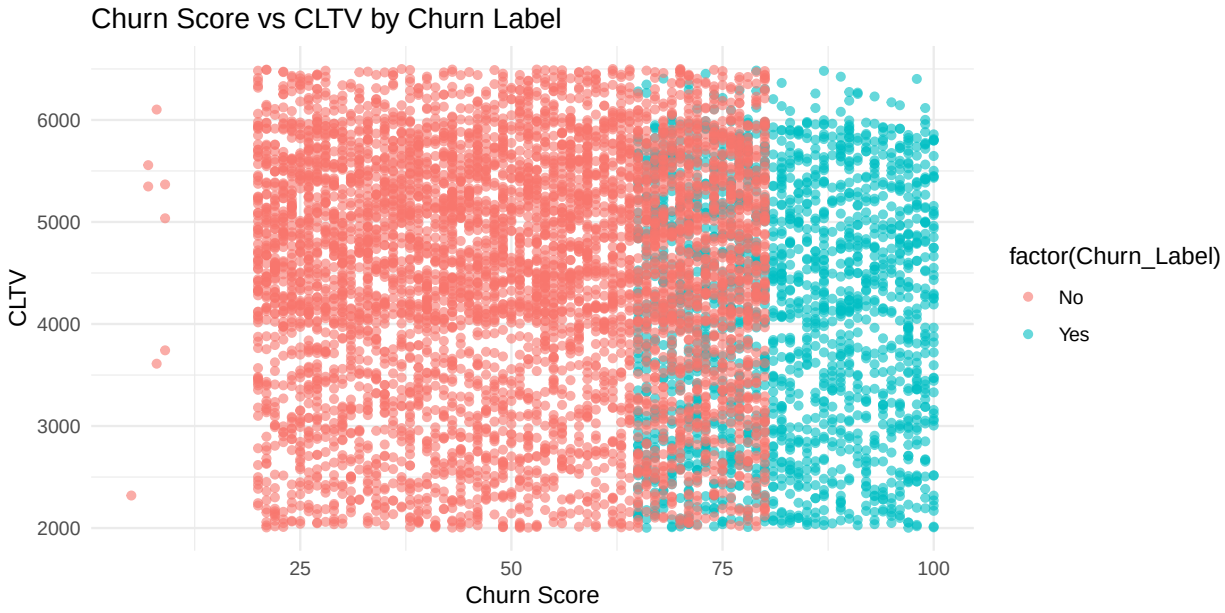



```
# Bar plot of Contract vs Churn_Label, faceted by Internet_Service
ggplot(df, aes(x = factor(Churn_Label), fill = factor( Tech_Support))) +
  geom_bar(position = "dodge") +
  facet_wrap(~ Contract) +
  labs(title = "Churn Label vs Contract by  Tech Support",
       x = "Churn_Label", fill = "Tech Suppotr") +
  theme_minimal()
```




```
# Scatter plot of Monthly_Charges vs CLTV, colored by Churn_Label
ggplot(df, aes(x = Churn_Score, y = CLTV, color = factor(Churn_Label))) +
  geom_point(alpha = 0.6) +
  labs(title = "Churn Score vs CLTV by Churn Label",
```

```
      x = "Churn Score", y = "CLTV") +
theme_minimal()
```



Observations

- **Churn Score vs CLTV by Churn Label**:
  The data suggests that all customers pay roughly the same monthly charges, but longer tenure leads to higher lifetime revenue (CLTV) for Telco. A key concern is month-to-month customers who churn at a higher rate despite paying lower monthly charges collectively. This raises a possible **income issue**—
  Insight: *These customers may not have high earnings, making retention more challenging.*
- **Churn Label vs Contract by Tech Support**:
  Customers with technical support appear to churn less, suggesting that support services play a crucial role in customer satisfaction and retention.
  Takeaway: *Providing better tech support could reduce churn rates.*
- **Churn Score vs CLTV by Churn Label**:
  It's unclear how Telco determines Churn Score and CLTV, but both seem to be **strong predictors of customer churn**. The relationship between expected lifetime revenue and churn score is highly correlated with churn behavior**.**
  Next Step:** Understanding how these values are calculated could provide deeper insights into churn patterns and predictive accuracy.

Conclusion

Below are the most predictive feature I have found for churn:

- Contract
- Payment Method
- Phone Service
- Tech Support
- Churn Score

- Monthly Charges
- CLTV
- Tenure Months