DSC1107_MONFERO_SA1

John Benedict A. Monfero

March 18, 2025

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.4     ✓ tidyr     1.3.1
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
##
## Attaching package: 'kableExtra'
##
##
## The following object is masked from 'package:dplyr':
##
##     group_rows
##
##
##
## Attaching package: 'cowplot'
##
##
## The following object is masked from 'package:lubridate':
##
##     stamp
##
##
##
## Attaching package: 'stat471'
##
##
## The following object is masked from 'package:FNN':
##
##     knn
```

Objective:

> The purpose of this summative assessment is to evaluate students' ability to apply data mining techniques, data visualization, data wrangling, and predictive modeling using R. Students will work with a provided dataset to perform exploratory data analysis, data transformation, model tuning, and regression-based methods.

> **Dataset**: Download the provided dataset `customer_churn.csv` , which contains customer demographics, service usage data, and churn labels.

Unit 1: R for Data Mining

Intro to Modern Data Mining

> Load the dataset and provide an overview of its structure (e.g., dimensions, missing values, types of variables).

```
data <- read.csv("customer_churn.csv")
```

```
dim(data)
```

```
## [1] 10000    12
```

```
cat("There are about 10,000 observations in the dataset, each `CustomerID` concerns about 11 things (n - 1 = 11)")
```

```
## There are about 10,000 observations in the dataset, each `CustomerID` concerns about 11 things (n - 1 = 11)
```

Handle missing values appropriately.

```
# Check individually, ideally all of columns (features) shall have none N/A entries
data %>%
  summarise(
    sum(is.na(CustomerID)),
    sum(is.na(Gender)),
    sum(is.na(SeniorCitizen)),
    sum(is.na(Partner)),
    sum(is.na(Dependents)),
    sum(is.na(Tenure)),
    sum(is.na(PhoneService)),
    sum(is.na(InternetService)),
    sum(is.na(Contract)),
    sum(is.na(MonthlyCharges)),
    sum(is.na(TotalCharges)),
    sum(is.na(Churn))
  )
```

```
##   sum(is.na(CustomerID)) sum(is.na(Gender)) sum(is.na(SeniorCitizen))
## 1                      0                  0                         0
##   sum(is.na(Partner)) sum(is.na(Dependents)) sum(is.na(Tenure))
## 1                   0                      0                 0
##   sum(is.na(PhoneService)) sum(is.na(InternetService)) sum(is.na(Contract))
## 1                        0                           0                    0
##   sum(is.na(MonthlyCharges)) sum(is.na(TotalCharges)) sum(is.na(Churn))
## 1                          0                        0                 0
```

```
# Overall, we want each observations
if (sum(is.na(data)) == 0) {
  cat("None observations who has n/a entries about their features in the given dataset")
}
```

```
## None observations who has n/a entries about their features in the given dataset
```

Explain why data mining is important for this dataset.

```
cat("According to IBM Technology (2023), Data Mining became a fundamental strategy nowadays; as we can convert any bus
siness-driven data into meaningful patterns and guide the grant users to have better basis and desicions about their m
arket and trends they will need. With Data Mining, we perform such processing data, indentifying patterns, and regress
ion lines in the specific amounts of variables and observations provided by our dataset.",  "\n\n\n")
```

```
## According to IBM Technology (2023), Data Mining became a fundamental strategy nowadays; as we can convert any bussi
ness-driven data into meaningful patterns and guide the grant users to have better basis and desicions about their mar
ket and trends they will need. With Data Mining, we perform such processing data, indentifying patterns, and regressio
n lines in the specific amounts of variables and observations provided by our dataset.
```

```
cat("To further understand, the dataset `customer_churn.csv` contains 12 variables which appears to determine knowing
the customer churn analysis; rather focuses on why customers may want to leave the services offered by a company.",
"\n\n\n")
```

```
## To further understand, the dataset `customer_churn.csv` contains 12 variables which appears to determine knowing th
e customer churn analysis; rather focuses on why customers may want to leave the services offered by a company.
```

```
cat("As previously mentioned, Data Mining for the dataset `customer_churn.csv` opens opportunity to create data visual
izations, neccesary tuning models, and compiling regression-based methods (from tidy enough or not yet) into meaningfu
l outcomes that can be the reference for desicion making skills within the context concerned.")
```

```
## As previously mentioned, Data Mining for the dataset `customer_churn.csv` opens opportunity to create data visualiz
ations, neccesary tuning models, and compiling regression-based methods (from tidy enough or not yet) into meaningful
outcomes that can be the reference for desicion making skills within the context concerned.
```

```
# Understand the data set by having a glimpse of its multidimensional characteristics and explain each variable's pote
ntial
glimpse(data)
```

```
## Rows: 10,000
## Columns: 12
## $ CustomerID      <chr> "CUST00001", "CUST00002", "CUST00003", "CUST00004", "C…
## $ Gender          <chr> "Male", "Male", "Male", "Female", "Male", "Female", "F…
## $ SeniorCitizen   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, …
## $ Partner         <chr> "No", "No", "Yes", "Yes", "No", "No", "Yes", "Yes", "Y…
## $ Dependents      <chr> "No", "No", "No", "Yes", "No", "Yes", "No", "Yes", "Ye…
## $ Tenure          <int> 65, 26, 54, 70, 53, 45, 35, 20, 48, 33, 33, 39, 6, 51,…
## $ PhoneService    <chr> "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes"…
## $ InternetService <chr> "Fiber optic", "Fiber optic", "Fiber optic", "DSL", "D…
## $ Contract        <chr> "Month-to-month", "Month-to-month", "Month-to-month", …
## $ MonthlyCharges  <dbl> 20.04, 65.14, 49.38, 31.19, 103.86, 87.34, 119.91, 69.…
## $ TotalCharges    <dbl> 1302.60, 1693.64, 2666.52, 2183.30, 5504.58, 3930.30, …
## $ Churn           <chr> "No", "No", "No", "No", "Yes", "Yes", "Yes", "Yes", "N…
```

- *CustomerID* (`chr`) – A unique identifier for each customer. It doesn't have predictive value but is useful for tracking individuals.
- *Gender* (`chr`) – Indicates whether the customer is "Male" or "Female". This can be used to analyze if gender has an impact on churn.
- *SeniorCitizen* (`int`) – A binary variable where 1 indicates that the customer is a senior citizen (typically 65+ years old) and 0 means they are not. This helps in understanding if age influences churn behavior.
- *Partner* (`chr`) – "Yes" if the customer has a spouse or partner, "No" otherwise. This may indicate social stability, which could influence customer loyalty.
- *Dependents* (`chr`) – "Yes" if the customer has dependents (such as children or elderly family members), "No" otherwise. Customers with dependents may have different service needs and spending behaviors.
- *Tenure* (`int`) – The number of months the customer has been with the service provider. Longer tenure often correlates with customer loyalty.
- *PhoneService* (`chr`) – "Yes" if the customer has a phone service, "No" otherwise. This shows whether the customer subscribes to the provider's phone services.
- *InternetService* (`chr`) – Type of internet service the customer has, such as "Fiber optic", "DSL", or "No" (if the customer does not have internet service). Different internet service types may affect churn rates.
- *Contract* (`chr`) – The type of contract the customer has, such as "Month-to-month", "One year", or "Two year". Longer contracts usually indicate customer retention, while month-to-month contracts are more flexible but may lead to higher churn.
- *MonthlyCharges* (`dbl`) – The amount the customer is billed each month for their services. High monthly charges may lead to churn if customers feel the cost is too high.
- *TotalCharges* (`dbl`) – The total amount charged to the customer during their tenure. This gives insight into a customer's lifetime value.
- *Churn* (`chr`) – The target variable indicating whether the customer has left the service ("Yes") or remains active ("No"). This is the outcome you want to predict or analyze.

Data Transformation

> Convert categorical variables into factor variables.

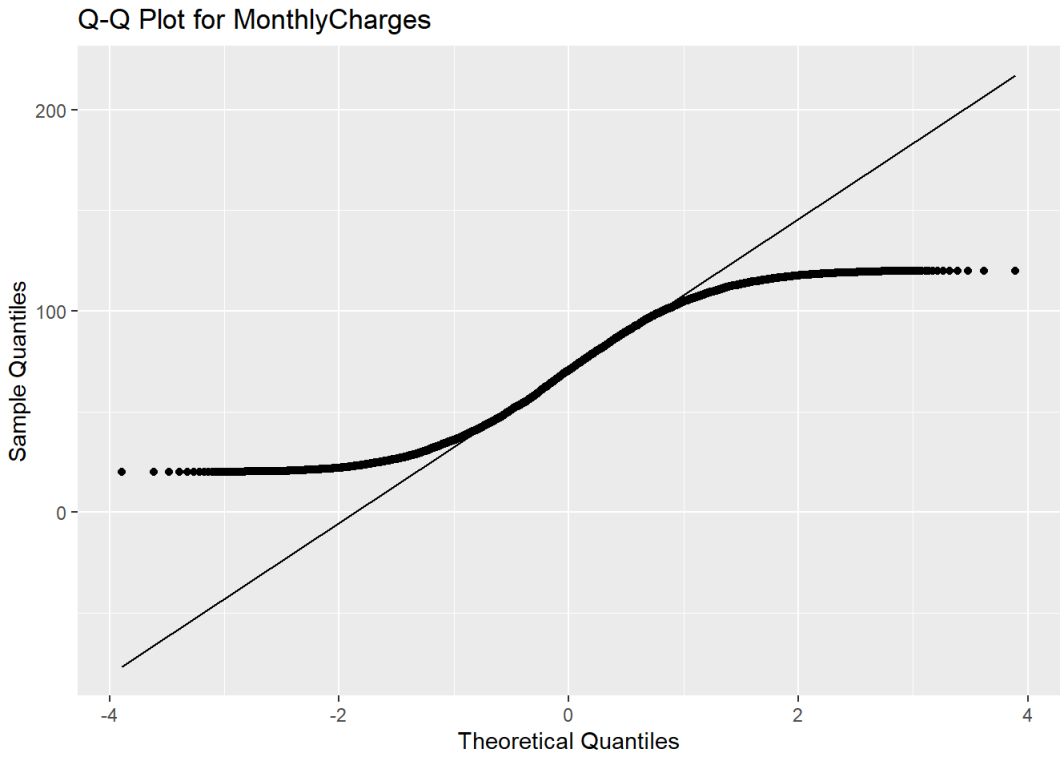
```
data <- data %>%
  mutate(
    Gender = as.factor(Gender),
    Partner = as.factor(Partner),
    Dependents = as.factor(Dependents),
    PhoneService = as.factor(PhoneService),
    InternetService = as.factor(InternetService),
    Contract = as.factor(Contract),
    Churn = as.factor(Churn) # VERY IMPORTANT : MAIN TARGET VARIABLE
  )

glimpse(data)
```

```
## Rows: 10,000
## Columns: 12
## $ CustomerID      <chr> "CUST00001", "CUST00002", "CUST00003", "CUST00004", "C…
## $ Gender          <fct> Male, Male, Male, Female, Male, Female, Female, Female…
## $ SeniorCitizen   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, …
## $ Partner         <fct> No, No, Yes, Yes, No, No, Yes, Yes, Yes, No, No, No, N…
## $ Dependents      <fct> No, No, No, Yes, No, Yes, No, Yes, Yes, No, No, No, No…
## $ Tenure          <int> 65, 26, 54, 70, 53, 45, 35, 20, 48, 33, 33, 39, 6, 51,…
## $ PhoneService    <fct> Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, No, Yes, Yes, …
## $ InternetService <fct> Fiber optic, Fiber optic, Fiber optic, DSL, DSL, Fiber…
## $ Contract        <fct> Month-to-month, Month-to-month, Month-to-month, One ye…
## $ MonthlyCharges  <dbl> 20.04, 65.14, 49.38, 31.19, 103.86, 87.34, 119.91, 69.…
## $ TotalCharges    <dbl> 1302.60, 1693.64, 2666.52, 2183.30, 5504.58, 3930.30, …
## $ Churn           <fct> No, No, No, No, Yes, Yes, Yes, Yes, No, No, Yes, No, N…
```

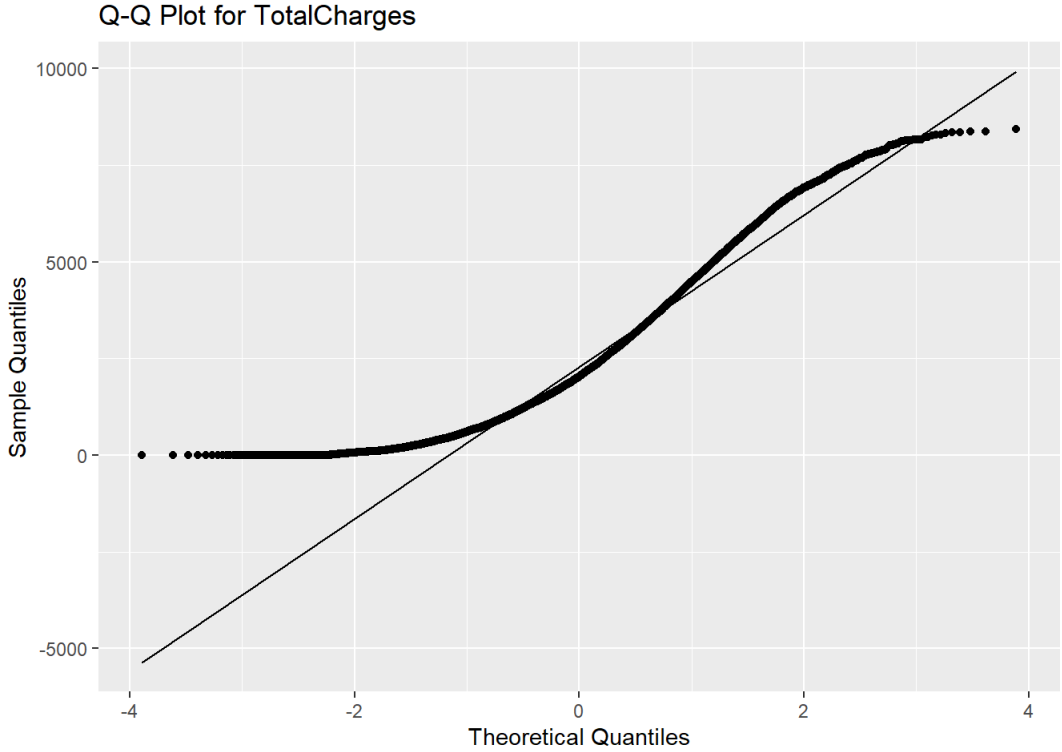
Normalize or standardize numerical features where necessary.

```
# Select only columns with `dbl` type
dbl_columns <- data %>% select(where(is.double))

# Create a function to generate Q-Q plots
generate_qq_plots <- function(column_name, column_data) {
  ggplot(data = NULL, aes(sample = column_data)) +
    stat_qq() +
    stat_qq_line() +
    labs(
      title = paste("Q-Q Plot for", column_name),
      x = "Theoretical Quantiles",
      y = "Sample Quantiles"
    )
}

# Apply function to all double columns
qq_plots <- dbl_columns %>%
  imap(~ generate_qq_plots(.y, .x))  # `.imap` from purrr allows passing column name and data

# Print Q-Q plots
qq_plots
```

```
## $MonthlyCharges
```



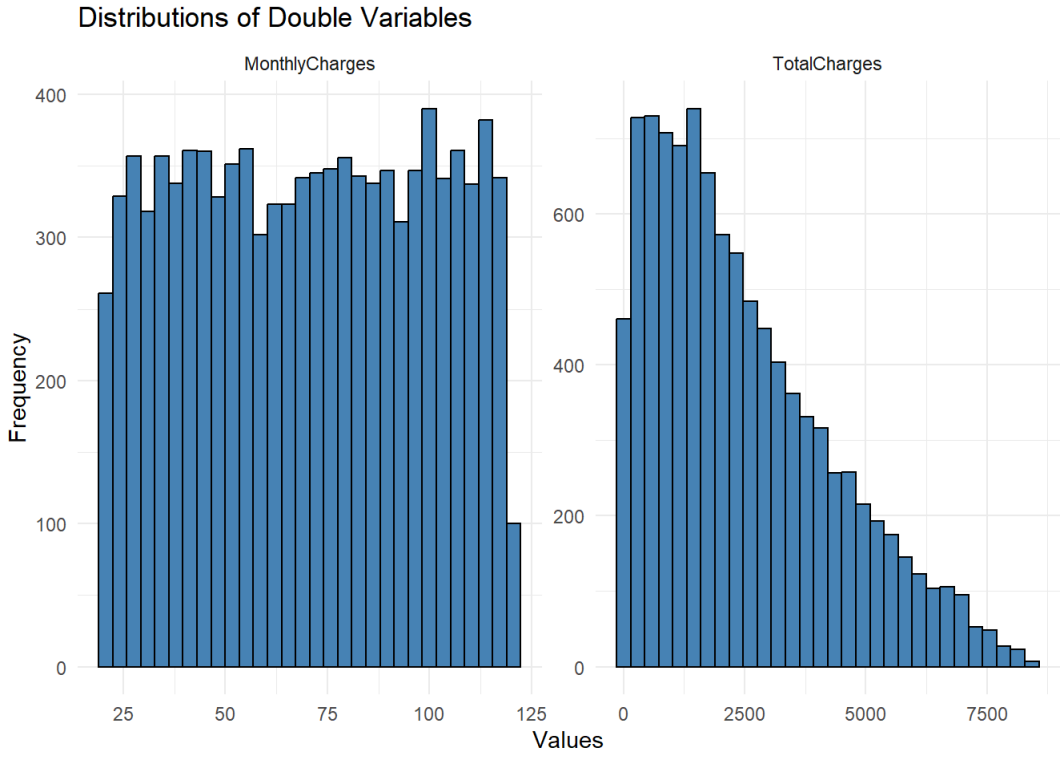

```
##
## $TotalCharges
```




```
cat("Both plot are suggesting non-normal distributions: For variable `Monthly Charges` the Q-Q Plot suggests that the distribution is remains heavy tailed on both standard deviations 2; while variable `Total Charges`, has the peak was t end to happened on the right tailed skewness before the expected mean of the distribution. Hence both of them does not follow normality.")
```

```
## Both plot are suggesting non-normal distributions: For variable `Monthly Charges` the Q-Q Plot suggests that the di stribution is remains heavy tailed on both standard deviations 2; while variable `Total Charges`, has the peak was ten d to happened on the right tailed skewness before the expected mean of the distribution. Hence both of them does not f ollow normality.
```

```
dbl_columns %>%
  pivot_longer(everything()) %>%
  ggplot(aes(x = value)) +
  geom_histogram(bins = 30, fill = "steelblue", color = "black") +
  facet_wrap(~ name, scales = "free") +
  labs(title = "Distributions of Double Variables", x = "Values", y = "Frequency") +
  theme_minimal()
```




```
cat("Likewise, here is the result of the histogram perceiving the same pattern implied by the Q-Q Plot earlier, but sh
owing more visually how does little kurtosis, and right skewwed is very notiable on each feature.")
```

```
## Likewise, here is the result of the histogram perceiving the same pattern implied by the Q-Q Plot earlier, but show
ing more visually how does little kurtosis, and right skewwed is very notiable on each feature.
```

> Transform the Data: If you suspect skewness or heavy tails, consider applying a transformation

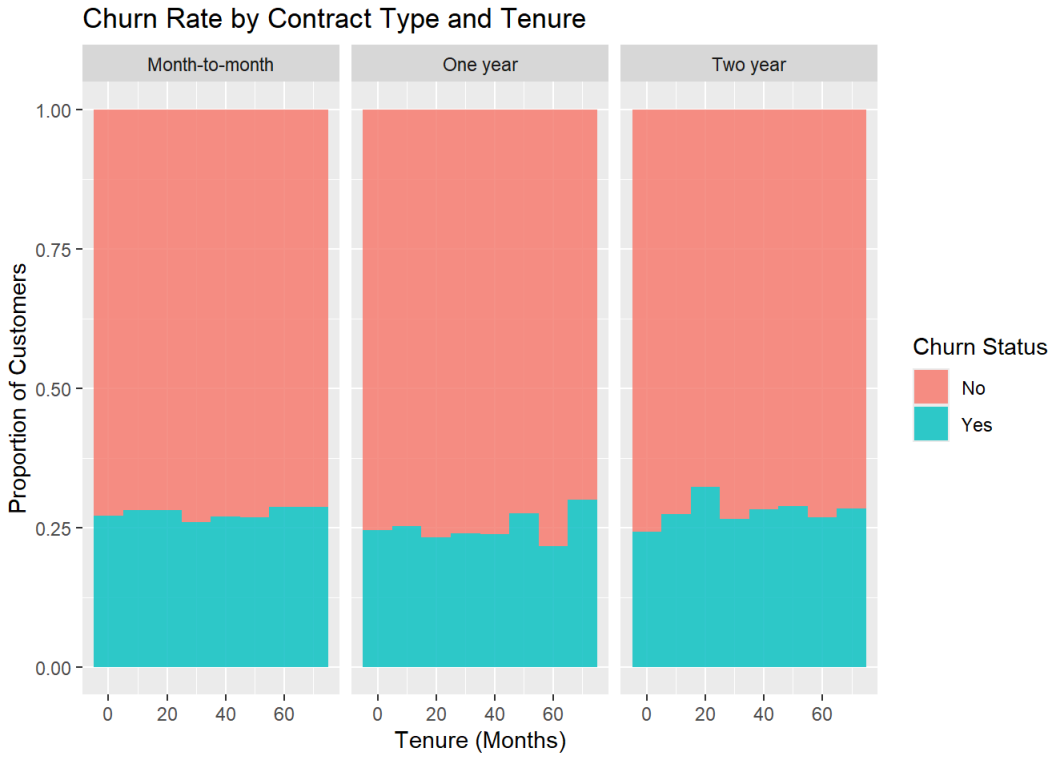
```
# Normalize using Min-Max Scaling
normalized_data <- data %>%
  mutate(
    MonthlyCharges_scaled = (MonthlyCharges - min(MonthlyCharges)) / (max(MonthlyCharges) - min(MonthlyCharges)),
    TotalCharges_scaled = (TotalCharges - min(TotalCharges, na.rm = TRUE)) / (max(TotalCharges, na.rm = TRUE) - min(To
talCharges, na.rm = TRUE))
  )
```

Data Visualization

> Create at least three meaningful visualizations to explore relationships in the data (e.g., churn rate by tenure, service type, or monthly charges).

> Provide insights based on the visualizations.

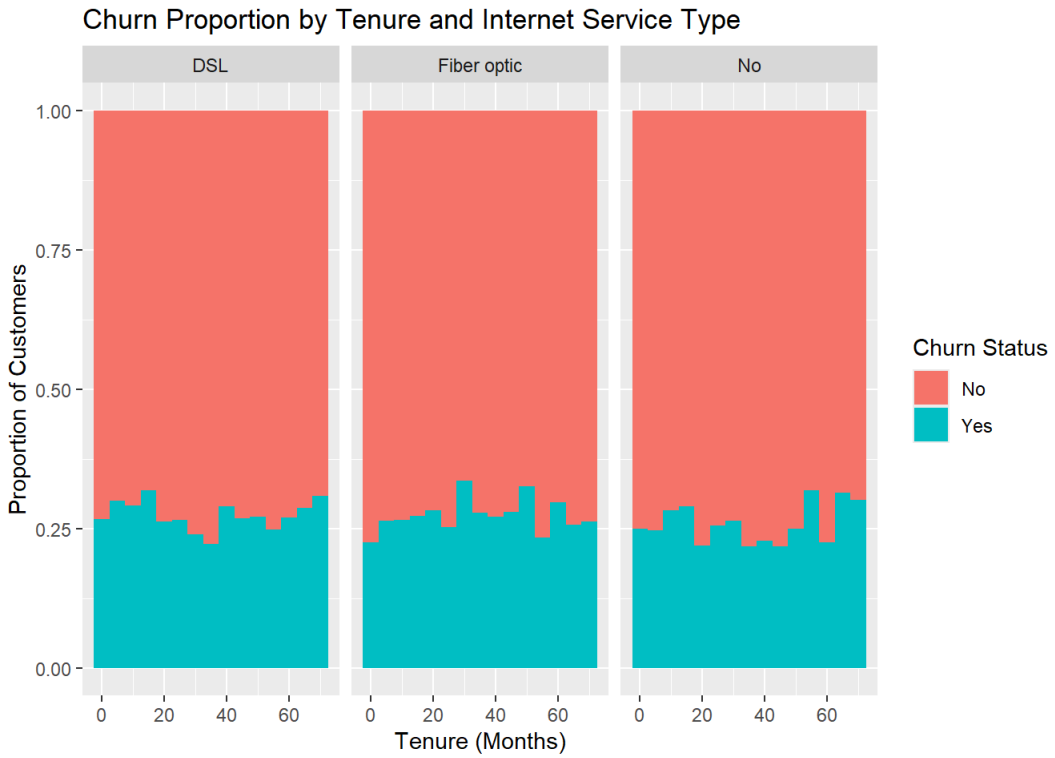
```
data %>%
  ggplot(aes(x = Tenure, fill = Churn)) +
  geom_histogram(binwidth = 10, position = "fill", alpha = 0.80) +
  facet_wrap(~Contract) +
  labs(
    title = "Churn Rate by Contract Type and Tenure",
       x = "Tenure (Months)",
       y = "Proportion of Customers",
       fill = "Churn Status"
    )
```




```
cat("The first multivariable visualization suggests that the churn rate is rather relatively consistent across differe
nt contract types {Month-to-Month; One Year; Two Year}, with approximately 75% of customers retaining their subscripti
on (No churn) and 25% opting to leave (Yes churn), irrespective of their tenure length, the proportion remains constan
t. This indicates that other factors, beyond the duration of tenure, might play a better and more significant role in
determining customer churn.")
```

```
## The first multivariable visualization suggests that the churn rate is rather relatively consistent across different
contract types {Month-to-Month; One Year; Two Year}, with approximately 75% of customers retaining their subscription
(No churn) and 25% opting to leave (Yes churn), irrespective of their tenure length, the proportion remains constant.
This indicates that other factors, beyond the duration of tenure, might play a better and more significant role in det
ermining customer churn.
```

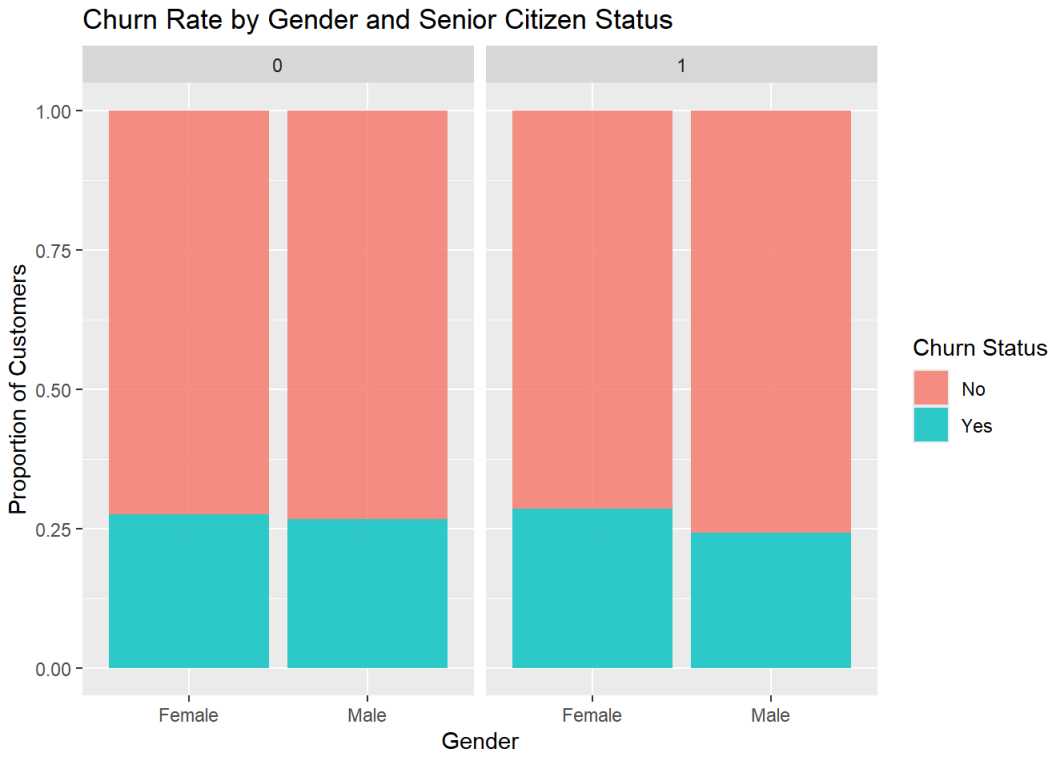
```
data %>%
  ggplot(aes(x = Tenure, fill = Churn)) +
  geom_histogram(binwidth = 5, position = "fill") +
  facet_wrap(~InternetService) +
  labs(
    title = "Churn Proportion by Tenure and Internet Service Type",
    x = "Tenure (Months)",
    y = "Proportion of Customers",
    fill = "Churn Status"
  )
```




```
cat("The data visualization suggests that the churn rate is rather independent across different Internet Service Types
{DSL; Fiber Optic; No}, with approximately 70-75% of customers retaining their subscription (No churn) and 25-30% opti
ng to leave (Yes churn), very weak influenced by their tenure length, the proportion remains independent. This indicat
es that other factors or other interactions of variables, beyond the duration of tenure, might play a better and more
significant role in determining customer churn.")
```

```
## The data visualization suggests that the churn rate is rather independent across different Internet Service Types
{DSL; Fiber Optic; No}, with approximately 70-75% of customers retaining their subscription (No churn) and 25-30% opti
ng to leave (Yes churn), very weak influenced by their tenure length, the proportion remains independent. This indicat
es that other factors or other interactions of variables, beyond the duration of tenure, might play a better and more
significant role in determining customer churn.
```

```
data %>%
  ggplot(aes(x = Gender, fill = Churn)) +
  geom_bar(position = "fill", alpha = 0.8) +
  facet_wrap(~SeniorCitizen) +
  labs(
    title = "Churn Rate by Gender and Senior Citizen Status",
    x = "Gender",
    y = "Proportion of Customers",
    fill = "Churn Status"
  )
```



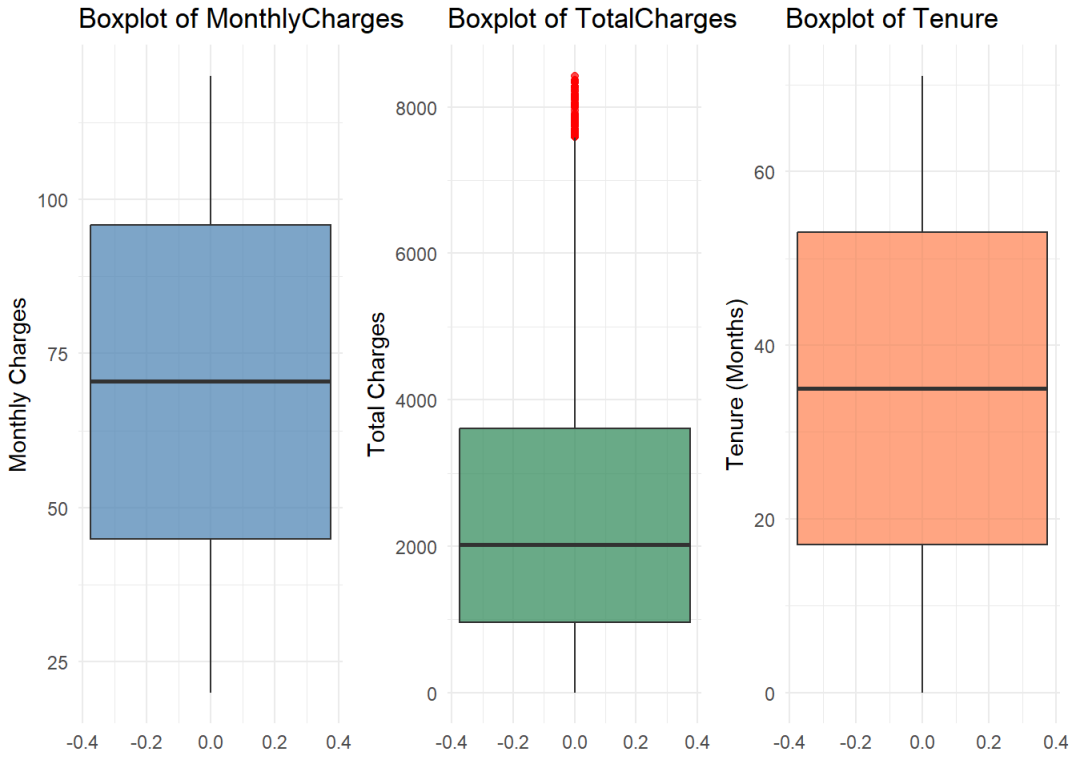

```
cat("In this visualization, there are exactly no difference of churn rate whenever their customer is not a Senior Citi
zen, only on the other hand, slightly more females but not to an alarming proportions, they tend to cut subscriptions
(Churn Yes) when they are a Senior Citizen.")
```

```
## In this visualization, there are exactly no difference of churn rate whenever their customer is not a Senior Citize
n, only on the other hand, slightly more females but not to an alarming proportions, they tend to cut subscriptions (C
hurn Yes) when they are a Senior Citizen.
```

Data Wrangling

Use Boxplot to detect outliers visually

```r
# Boxplot for MonthlyCharges
boxplot1 <- data %>%
  ggplot(aes(y = MonthlyCharges)) +
  geom_boxplot(fill = "steelblue", alpha = 0.7, outlier.color = "red") +
  labs(
    title = "Boxplot of MonthlyCharges",
    y = "Monthly Charges"
  ) +
  theme_minimal()

# Boxplot for TotalCharges
boxplot2 <- data %>%
  ggplot(aes(y = TotalCharges)) +
  geom_boxplot(fill = "seagreen", alpha = 0.7, outlier.color = "red") +
  labs(
    title = "Boxplot of TotalCharges",
    y = "Total Charges"
  ) +
  theme_minimal()

# Boxplot for Tenure
boxplot3 <- data %>%
  ggplot(aes(y = Tenure)) +
  geom_boxplot(fill = "coral", alpha = 0.7, outlier.color = "red") +
  labs(
    title = "Boxplot of Tenure",
    y = "Tenure (Months)"
  ) +
  theme_minimal()

# Combine the boxplots using cowplot
combined_boxplots <- plot_grid(
  boxplot1, boxplot2, boxplot3,
  nrow = 1,
  ncol = 3 # Arrange plots horizontally
)

# Display the combined boxplots
print(combined_boxplots)
```




```r
cat("This suggests that there are some outliers observed in the variable `TotalCharges` from the provided dataset")
```

```
## This suggests that there are some outliers observed in the variable `TotalCharges` from the provided dataset
```

> Filter data to remove outliers. Recall the Concept of IQR (Inter-Quartile Range)

```
# Calculate the IQR for TotalCharges
Q1 <- quantile(data$TotalCharges, 0.25, na.rm = TRUE)  # First quartile (25th percentile)
Q3 <- quantile(data$TotalCharges, 0.75, na.rm = TRUE)  # Third quartile (75th percentile)
IQR <- Q3 - Q1  # Interquartile range

# Define the lower and upper bounds for outliers
lower_bound <- Q1 - 1.5 * IQR
upper_bound <- Q3 + 1.5 * IQR

# Filter the data to remove outliers
data_no_outliers <- data %>%
  filter(TotalCharges >= lower_bound & TotalCharges <= upper_bound)

# Display the cleaned dataset
glimpse(data_no_outliers)
```

```
## Rows: 9,928
## Columns: 12
## $ CustomerID      <chr> "CUST00001", "CUST00002", "CUST00003", "CUST00004", "C…
## $ Gender          <fct> Male, Male, Male, Female, Male, Female, Female, Female…
## $ SeniorCitizen   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, …
## $ Partner         <fct> No, No, Yes, Yes, No, No, Yes, Yes, Yes, No, No, No, N…
## $ Dependents      <fct> No, No, No, Yes, No, Yes, No, Yes, Yes, No, No, No, No…
## $ Tenure          <int> 65, 26, 54, 70, 53, 45, 35, 20, 48, 33, 33, 39, 6, 51,…
## $ PhoneService    <fct> Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, No, Yes, Yes, …
## $ InternetService <fct> Fiber optic, Fiber optic, Fiber optic, DSL, DSL, Fiber…
## $ Contract        <fct> Month-to-month, Month-to-month, Month-to-month, One ye…
## $ MonthlyCharges  <dbl> 20.04, 65.14, 49.38, 31.19, 103.86, 87.34, 119.91, 69.…
## $ TotalCharges    <dbl> 1302.60, 1693.64, 2666.52, 2183.30, 5504.58, 3930.30, …
## $ Churn           <fct> No, No, No, No, Yes, Yes, Yes, Yes, No, No, Yes, No, N…
```

Review

> In this first chapter, we are able to demonstrate the ability and apply data mining techniques: such as displaying powerful data visualization, and some tidying some data wrangling and transformation.

> Use relevant and accurate type of plots (by knowing data types first) and algorithms that are needed in order to normalized (Min-Max Scale), and remove any outliers (IQR).

Save the tidy dataset for future use

```
# Save the data_no_outliers dataset as a CSV file
write.csv(data_no_outliers, file = "CUSTOMER_CHURN_TIDY.CSV", row.names = FALSE)
```