S1W1P2

Project Scenario, Roles, and Dataset

Customer Churn Analysis for Telecommunications Company

**Student Name:**

**Student ID:**

Table of Contents

[Data Preparation and Preprocessing 4](#_Toc172556163)

[Load and preprocess the dataset 4](#_Toc172556164)

[Handle missing data points and encode categorical variables. 4](#_Toc172556165)

[Perform feature scaling and normalization. 6](#_Toc172556166)

[Ensure data integrity and consistency. 7](#_Toc172556167)

[EDA 7](#_Toc172556168)

[References 12](#_Toc172556174)

**Table of Figures**

[Figure 1: Data loading 3](#_Toc172550046)

[Figure 2: Checking data issues 4](#_Toc172550047)

[Figure 3: Data encoding 5](#_Toc172550048)

[Figure 4: Standardization 6](#_Toc172550049)

[Figure 5: Descriptive statistics 6](#_Toc172550050)

[Figure 6:Tenure according to gender and Monthly charges by contract type 8](#_Toc172550051)

[Figure 7: Histogram for Tenure 9](#_Toc172550052)

[Figure 8: Churn by Contract and Monthly Charges for Internet Services 10](#_Toc172550053)

[Figure 9: Churn by senior citizen and gender distribution 13](#_Toc172550054)

# Data Preparation and Preprocessing

## Load and preprocess the dataset

To load the dataset from the system, read the CSV file by giving the file path from the system. Load dataset as data frame by importing “pandas” library. Also import some useful libraries like numpy, matplotlib, and seaborn.

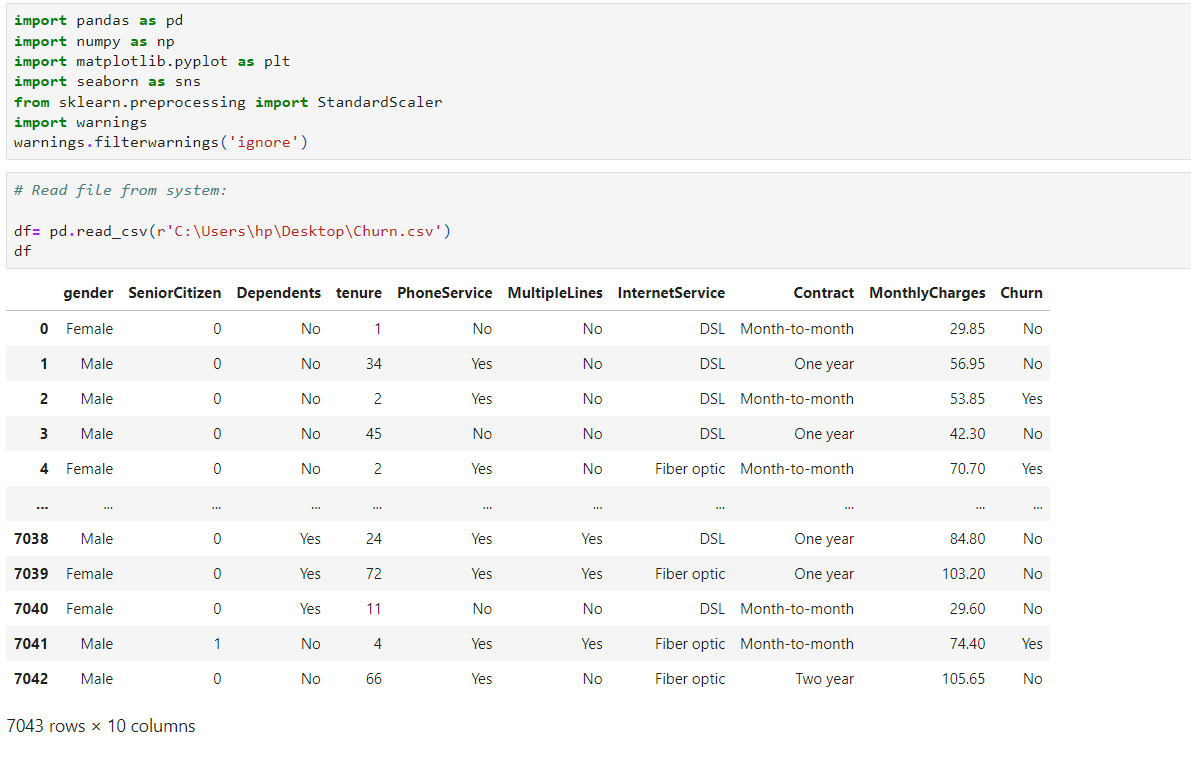


Figure 1: Data loading

## Handle missing data points and encode categorical variables.

There are no missing data or other issues in the given dataset. These are checked by using the “info” and “is null()” functions.

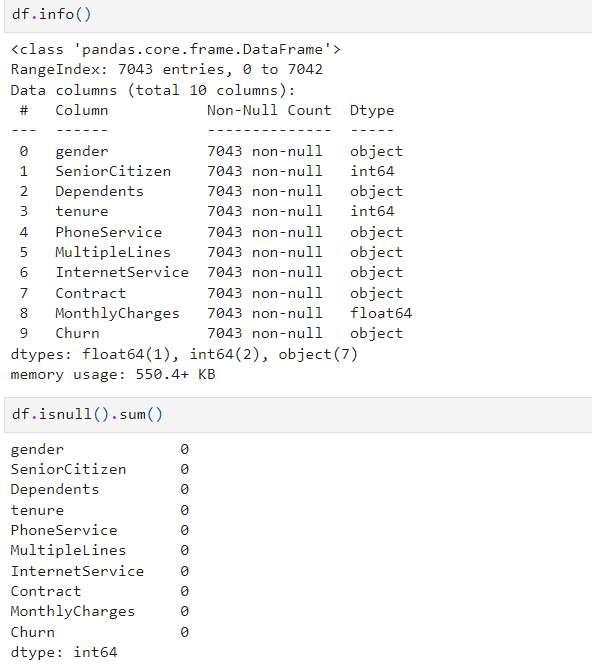


Figure 2: Checking data issues

To encode variables, using the “binary encoding method and label encoding methods”. The binary encoding method is used for “gender, Internet Service, phone services, multiple lines, churn, and dependents” due to having only two values but the “Contract” variable has three values so, used label encoding.

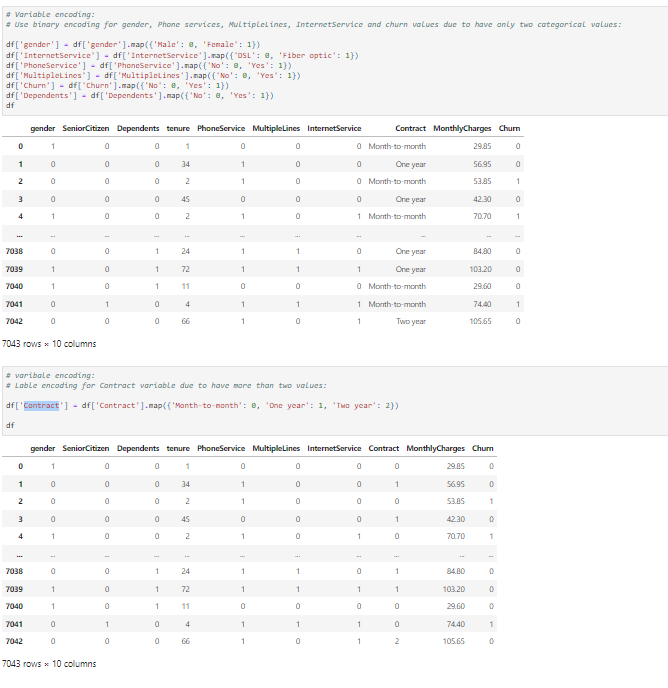


Figure 3: Data encoding

## Perform feature scaling and normalization.

The code standardizes the data by centering it around the mean and scaling it to unit variance. Standardsclaer() calculates the mean and standard deviation for each feature. fit\_transform(pdf) then applies this scaling, transforming the data to have a mean of 0 and a standard deviation of 1

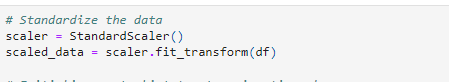
. 

Figure 4: Standardization

## Ensure data integrity and consistency.

Ensuring data integrity and consistency involves validating data formats, checking for outliers or anomalies, and confirming that all necessary columns or features are present and correctly formatted. This step ensures the reliability and accuracy of the dataset for subsequent analysis or modeling tasks.

## EDA

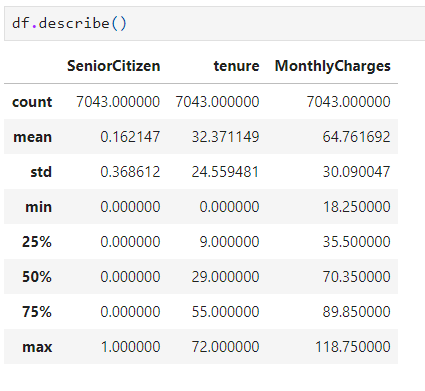


Figure 5: Descriptive statistics

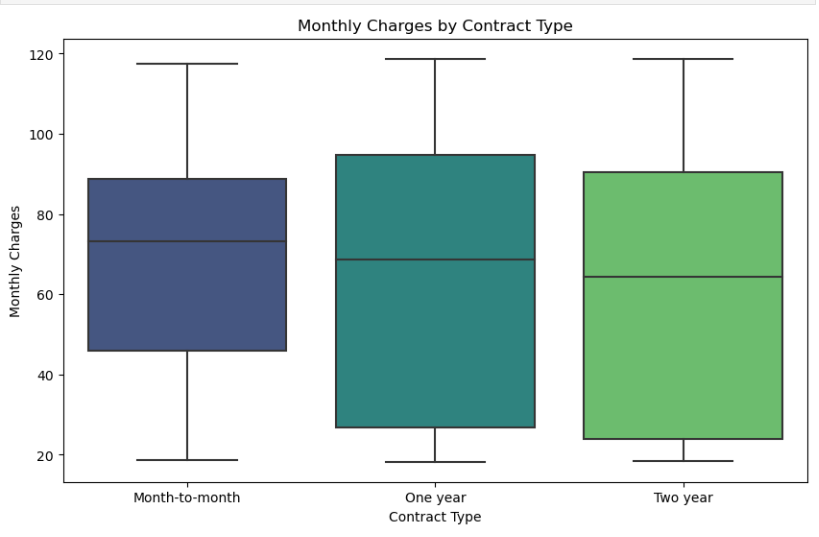
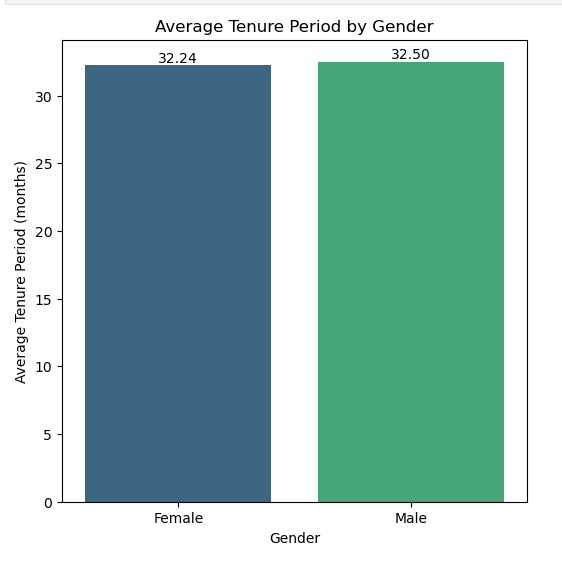


Figure 6:Tenure according to gender and Monthly charges by contract type

The above graphs are for “tenure according to gender” and another box plot graph for “relationship between monthly charges and contract”. Females served more tenure periods than males. “Month-to-month” charges are more costly than others.

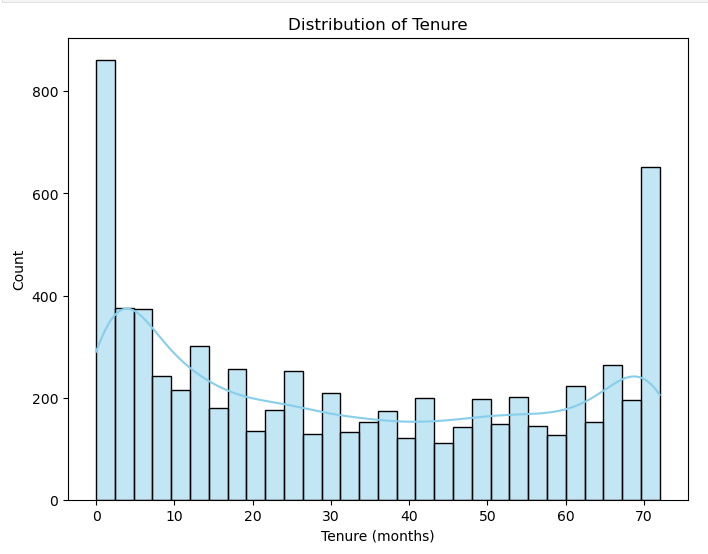


Figure 7: Histogram for Tenure

The above graphs show the “histogram for knowing the distribution of tenure”. According to this graph, it can be easily seen that 0 and 70 values have the highest count and others are normally distributed.

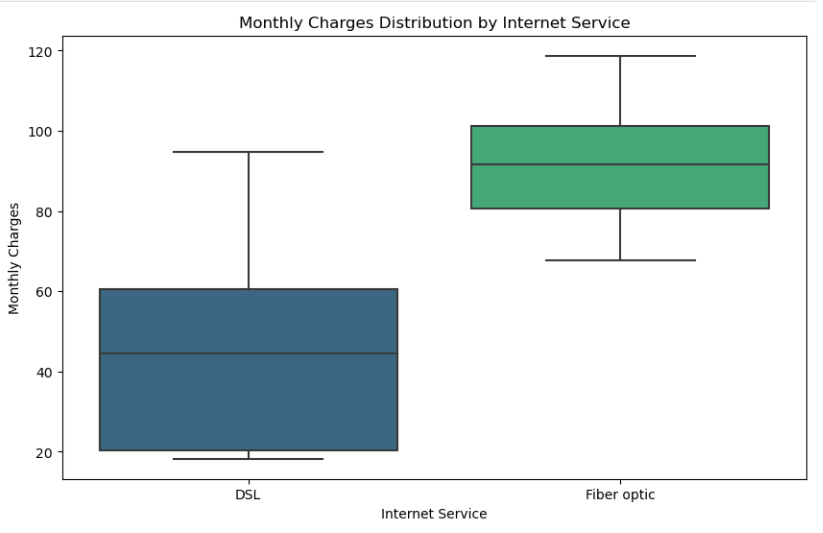
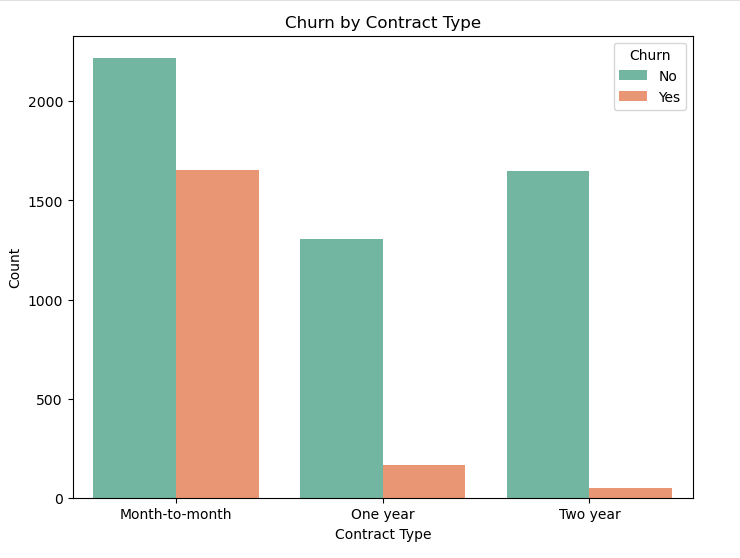


Figure 8: Churn by Contract and Monthly Charges for Internet Services

The above charts show the “churn by contract type” and “other shows the monthly charges distribution by internet service”. According to the churn graph, more customers are no churn in each contract type. The box plot shows that “Fiberoptic” has higher monthly charges than DSL.

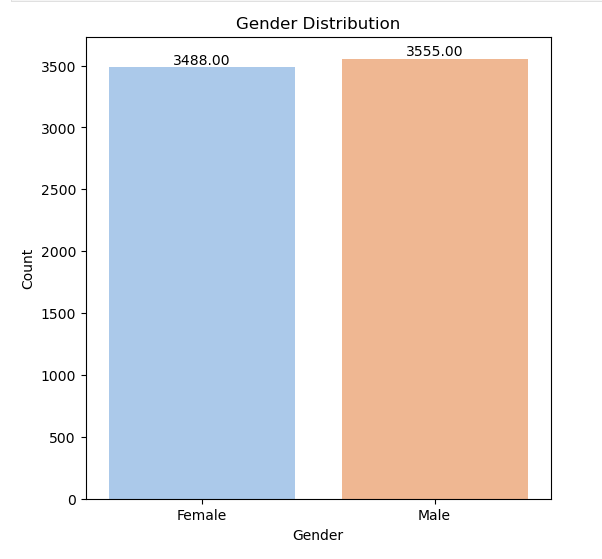
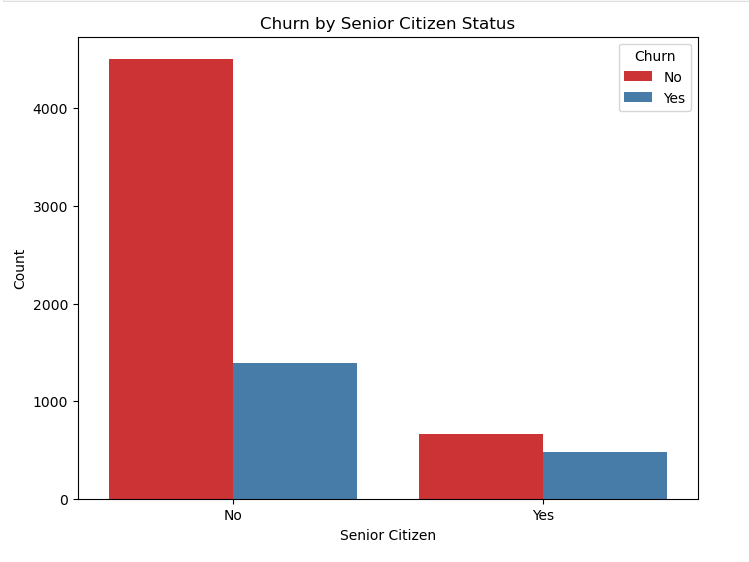


Figure 9: Churn by senior citizen and gender distribution

The above chart shows the “Churn behavior among senior citizens” and the other shows the “gender distribution”. According to this churn graph, it can be easily seen that most senior citizens are also not churnes as they like their telecom services. The gender distribution graph shows that “Male” customer is more connected to their telecom services (Mahadevan, 2022).

# References

‌Mahadevan, M. (2022). Step-by-Step Exploratory Data Analysis (EDA) using Python. [online] Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2022/07/step-by-step-exploratory-data-analysis-eda-using-python/> [Accessed 5 Jul. 2024].