UNDERSTANDING TELCO CUSTOMER CHURN

DATE: 27TH OF JANUARY, 2024

INTRODUCTION
This is a Telecommunication company that provides some services to its customers, some customers have left, and we are here to analyze the case and figure out the reasons behind their churn using the provided features.

OBJECTIVE: Investigate and identify the underlying factors contributing to customer churn within the Telecommunication company by conducting a comprehensive analysis of the provided features.

DATASET DESCRIPTION

- **Churn:** customers who left within the last month.

- **Services that each customer has signed up for:** Phone, Multiple lines, Internet, Online Security, Online backup, Device protection, Tech support, and Streaming TV and movies.

- **Customer account information:** how long they've been a customer, Contract, Payment method, Paperless billing, Monthly charges, and Total charges.

- **Demographic info about customers:** Gender, Age range, and if they have Partners and Dependents.

IMPORTING LIBRARIES
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

READING THE DATASET

1. Read the csv file and stores the content.
   df = pd.read_csv("Telco-Customer-Churn.csv")

2. Displayed the content of the data frame
   df

DATA EXPLORATION

1. Determined the number of rows and column
   df.shape

2. Listed all the columns, non null count and data type.
   df.info()

3. Previewed the first five column of the dataset
   df.head()

4. Generated descriptive statistics
   df.describe()

5. Retrieved a random sample of rows
   df.sample()

6. Displayed the data types of each column
   df.dtypes

7. Retrieved the column labels (column names)
   df.columns

Determined the unique values in all the columns

1. df[“customerID”].unique()
2. df[“gender”].unique()
3. df["SeniorCitizen"].unique()
4. df[“Partner”].unique()
5. df[“Dependents”].unique()
6. df["tenure"].unique()
7. df[“PhoneService”] .unique()
8. df[“MultipleLines”] .unique()
9. df[“InternetService”] .unique()
10. df[“OnlineSecurity”] .unique()
11. df[“OnlineBackup”] .unique()
12. df[“DeviceProtection”] .unique()
13. df[“TechSupport”] .unique()
14. df[“StreamingTV”] .unique()
15. df[“StreamingMovies”].unique()
16. df[“Contract”].unique()
17. df[“PaperlessBilling”].unique()
18. df[“PaymentMethod”].unique()
19. df[“MonthlyCharges”].unique()
20. df[“TotalCharges”].unique()
21. df[“Churn”].unique()

DATA CLEANING

1. Determined the sum of the null values
   df.isnull().sum()

2. Determined the sum of the duplicate values
   df.duplicated().sum()

3. Replace 1 with Yes and 0 with No in the senior citizen column

   df["SeniorCitizen"] = df["SeniorCitizen"].replace(1,"yes")
   df["SeniorCitizen"] = df["SeniorCitizen"].replace(0,"no")

df["SeniorCitizen"].unique()

4. Correct the data type of “TotalCharges” from Object to Float
   df["TotalCharges"]
   df["TotalCharges"] = pd.to_numeric(df["TotalCharges"], errors="coerce")

   There are 11 missing values, this is so small ratio to the data size, so dropping them will not affect the data analysis.

5. Removed Null values from TotalCharges column
   df.dropna(subset=["TotalCharges"], inplace=True)

   df["TotalCharges"].isnull().sum()

EXPLORATORY DATA ANALYSIS

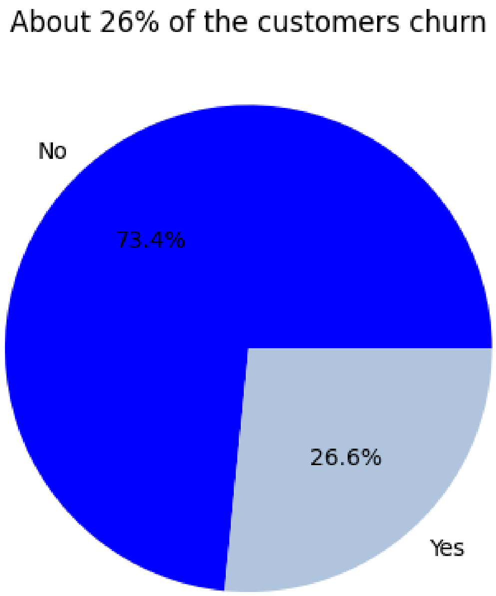
1. What is the overall churn rate in the company?

   Count of each unique values in the "Churn" column

   churn_rate = df["Churn"].value_counts()
   churn_rate

   churn_rate = df["Churn"].value_counts()
   plt.pie(churn_rate.values, labels= churn_rate.index, colors = ["blue", "lightsteelblue"], autopct="%1.1f%%")
   plt.title("About 26% of the customers churn")
   plt.show()

   Insight: About a quarter of the customers left last month.


About 26% of the customers churn
No
73.4%
26.6%
Yes

2. What is the gender ratio in the company?

   gender_rate = df["gender"].value_counts()

- Created a pieplot

  plt.pie(gender_rate.values, labels= gender_rate.index, colors = ["blue", "lightsteelblue"], autopct="%1.1f%%")

- Added a circle at the center to transform it in a donut chart

  my_circle = plt.Circle((0,0), 0.4, color='white')

  gcf: This is a Matplotlib function that stands for "get current figure". The gcf() function is used to get a reference to the current figure being worked on or to create a new figure if none exists.

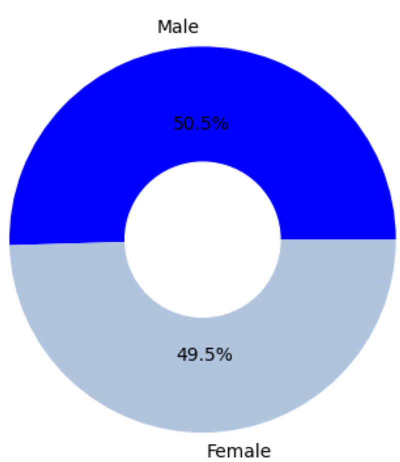
  p = plt.gcf()

  gca(): Stands for "get current axes." In Matplotlib, the axes are the region of the plot where data is visualized. The gca() function is used to get a reference to the current axes of the current figure.

  add_artist(my_circle): This is a method used on the current axes. my_circle is assumed to be a circular object created earlier, possibly using plt.Circle().

  p.gca().add_artist(my_circle)

  plt.show()



3. How does gender affect churn rate? is there a significant difference between males and females' churn rates?

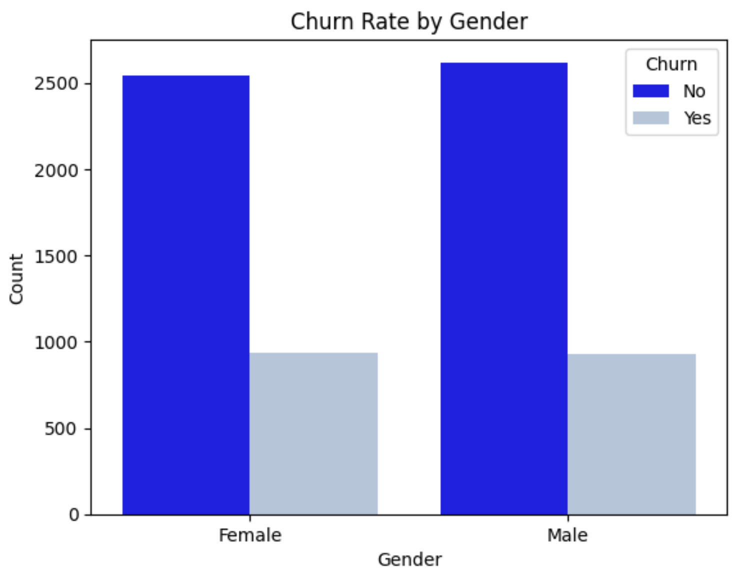
   sns.countplot(data= df, x = "gender", hue="Churn", palette={"lightsteelblue", "blue"})

   plt.title("Churn Rate by Gender")

   plt.xlabel("Gender")

   plt.ylabel("Count")

   plt.show()



   Insight: Both males and females are almost the the same ratio

4. Does being a senior citizen influence the likelihood of churn?

   sns.countplot(x = "SeniorCitizen", hue="Churn", data= df)

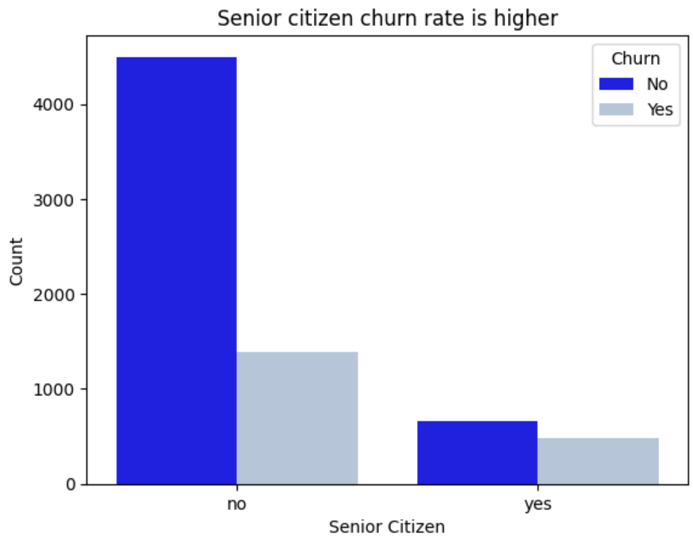
   sns.countplot(data= df, x = "SeniorCitizen", hue="Churn", palette={"lightsteelblue", "blue"})

   plt.title("Senior citizen churn rate is higher")

   plt.xlabel("Senior Citizen")

   plt.ylabel("Count")

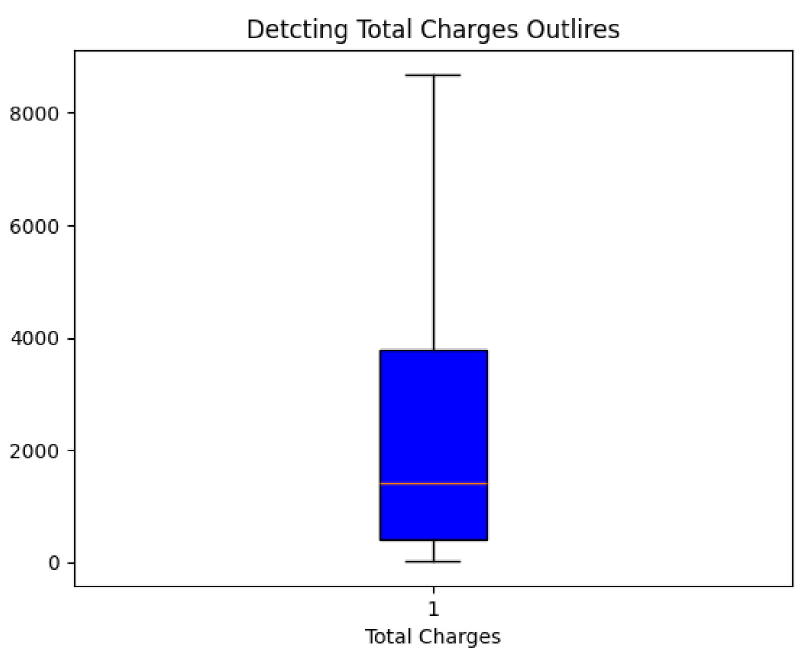
   plt.show()



it can be noticed that the ratio of senior citizens among the customers is lower, and the likelihood of churn is higher for the senior citizens compared to non-senior citizens.

5. How much charges do the customers mostly pay?

```
plt.boxplot(df["TotalCharges"], patch_artist=True, boxprops=dict(facecolor='blue'))
plt.title("Detcting Total Charges Outlires")
plt.xlabel("Total Charges")
```



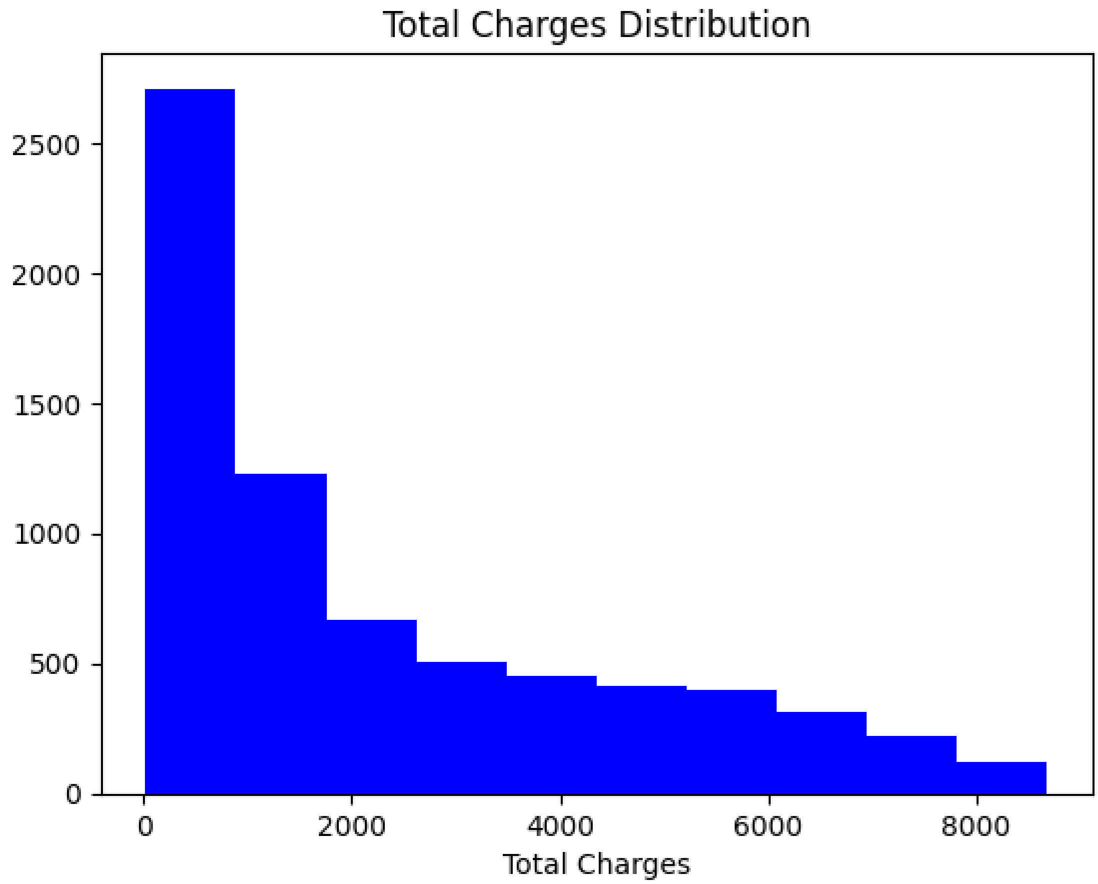
6. Describe the distribution of the Total Charge customers.

   plt.hist(df["TotalCharges"])

   plt.title("Total Charges Distribution")

   plt.xlabel("Total Charges")

   Insights: There are no outliers.



Insight: The distribution seems skewed, indicating that the majority of our customers pay less than 2000 in charges. As charges increase, the number of customers decreases.

7. Are customers with partners less likely to churn compared to those without parteners?

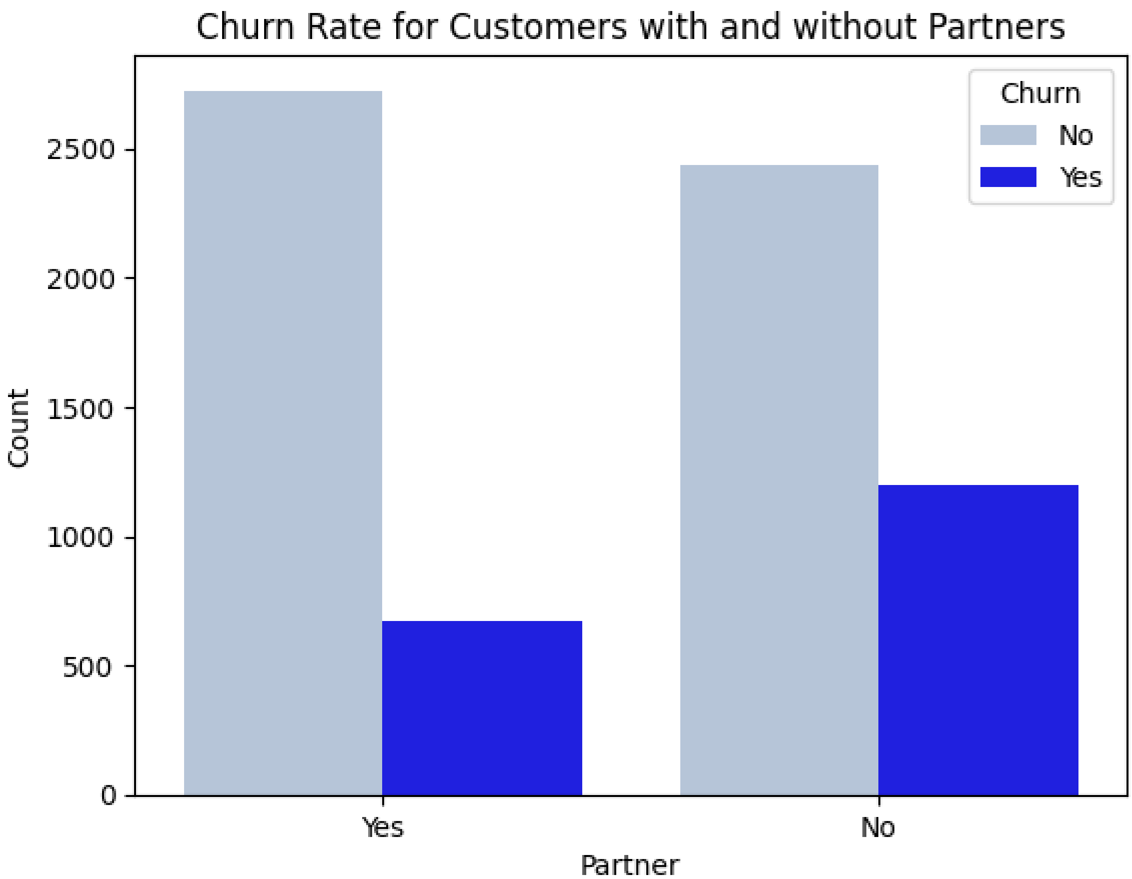
   sns.countplot(x= "Partner", hue= "Churn", data =df)

   plt.title('Churn Rate for Customers with and without Partners')

   plt.xlabel("Partner")

   plt.ylabel("Count")

   plt.show()

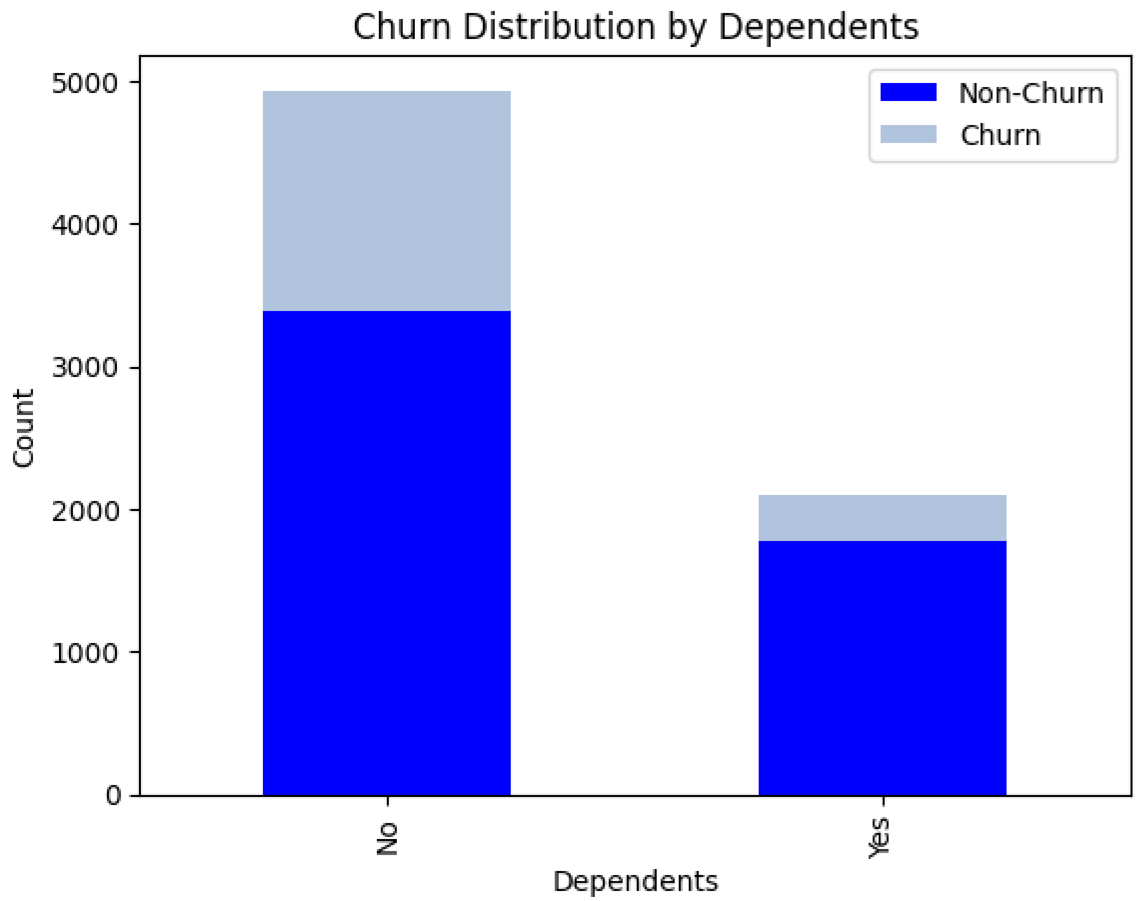


Insight: Observing the visualization, it becomes apparent that, despite a lower ratio of customers without partners, they exhibit a higher likelihood of churn compared to customers with partners.

```
df["Dependents"].unique()
```

8. Dose having dependents affect the churn rate?

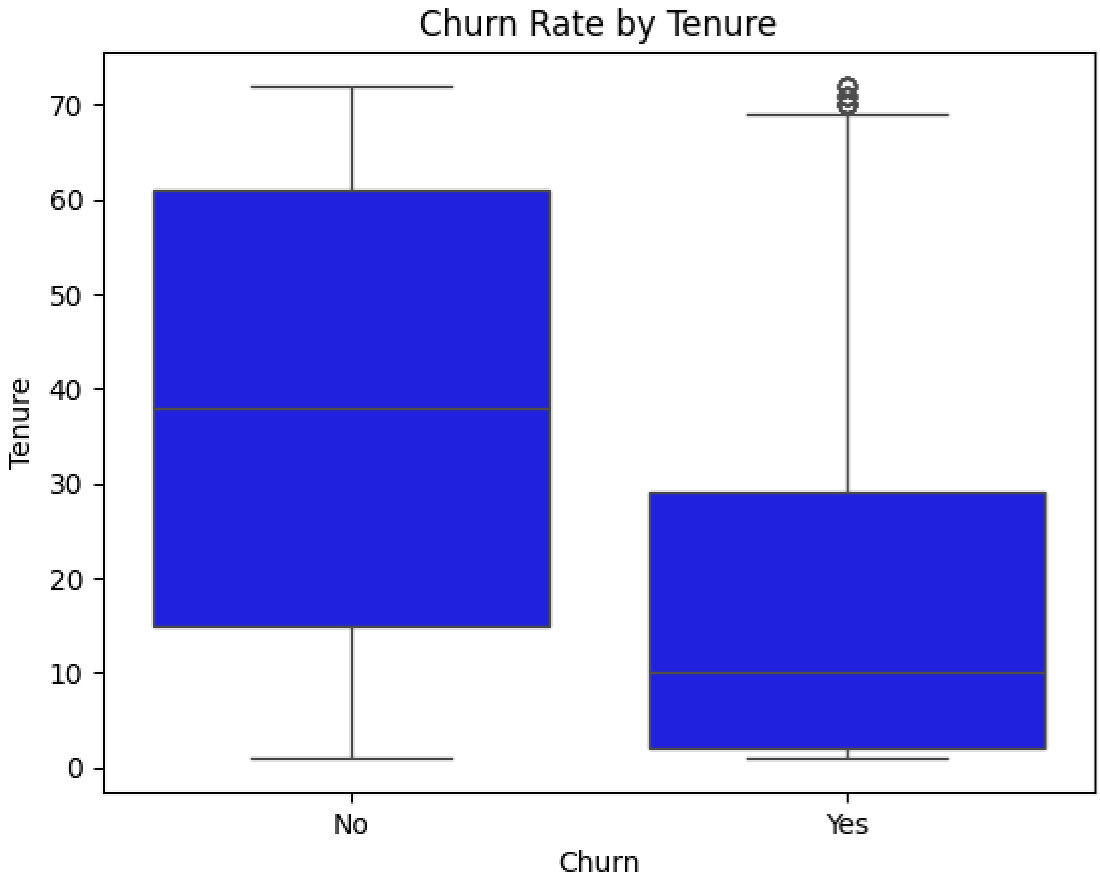
```
df_grouped = df.groupby(["Dependents", "Churn"]).size().unstack()
df_grouped.plot(kind= "bar", stacked=True)

plt.title("Churn Distribution by Dependents")
plt.xlabel("Dependents")
plt.ylabel("Count")
plt.legend(["Non-Churned", "Churned"])
```



Insights: it's noted that most of our customers don't have dependents, and they also more likely to leave compared to customers with dependents.

9. Is there a correleation between tenure and churn rate?

```
df["tenure"].unique()

sns.boxplot(x= "Churn", y="tenure", data = df)

plt.title("Churn Rate by Tenure")

plt.xlabel("Churn")

plt.ylabel("Tenure")

plt.show()
```

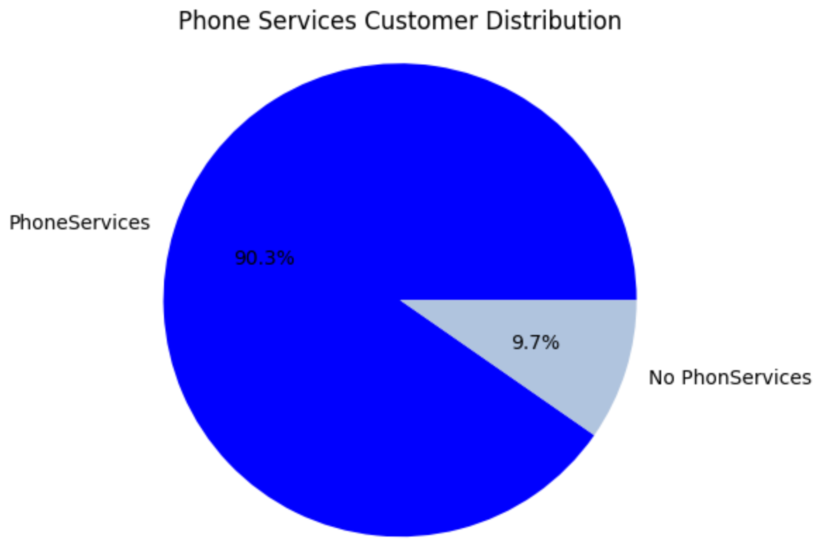


Insights

- It appears that The shorter the duration customers spend with the company the higher their chances of churning.
- The average tenre rate of churned customers is about 10 years.

10. What is the distribution of customers that have assess to phone services

```
df["PhoneService"].value_counts()


phone_values = df["PhoneService"].value_counts().values
labels = ["With PhoneServices", "Without PhonServices"]
sizes = [6352, 680]
plt.pie(sizes, labels= labels, colors = ["blue", "lightsteelblue"], autopct="%1.1f%%")
plt.axis("equal")
plt.title("PhoneService precense or Absence")
plt.show()
```



Insights:

The largest portion of the customers have phone services.

11. Does the presence or absence of phone service impact churn behavior?

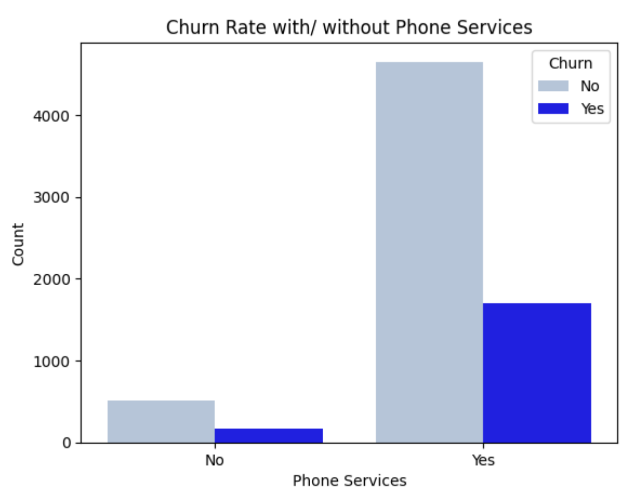
sns.countplot(x= "PhoneService", hue="Churn", data= df)

plt.title("Churn Rate with/ without Phone Services")

plt.xlabel("Phone Services")

plt.ylabel("Count")

plt.show()

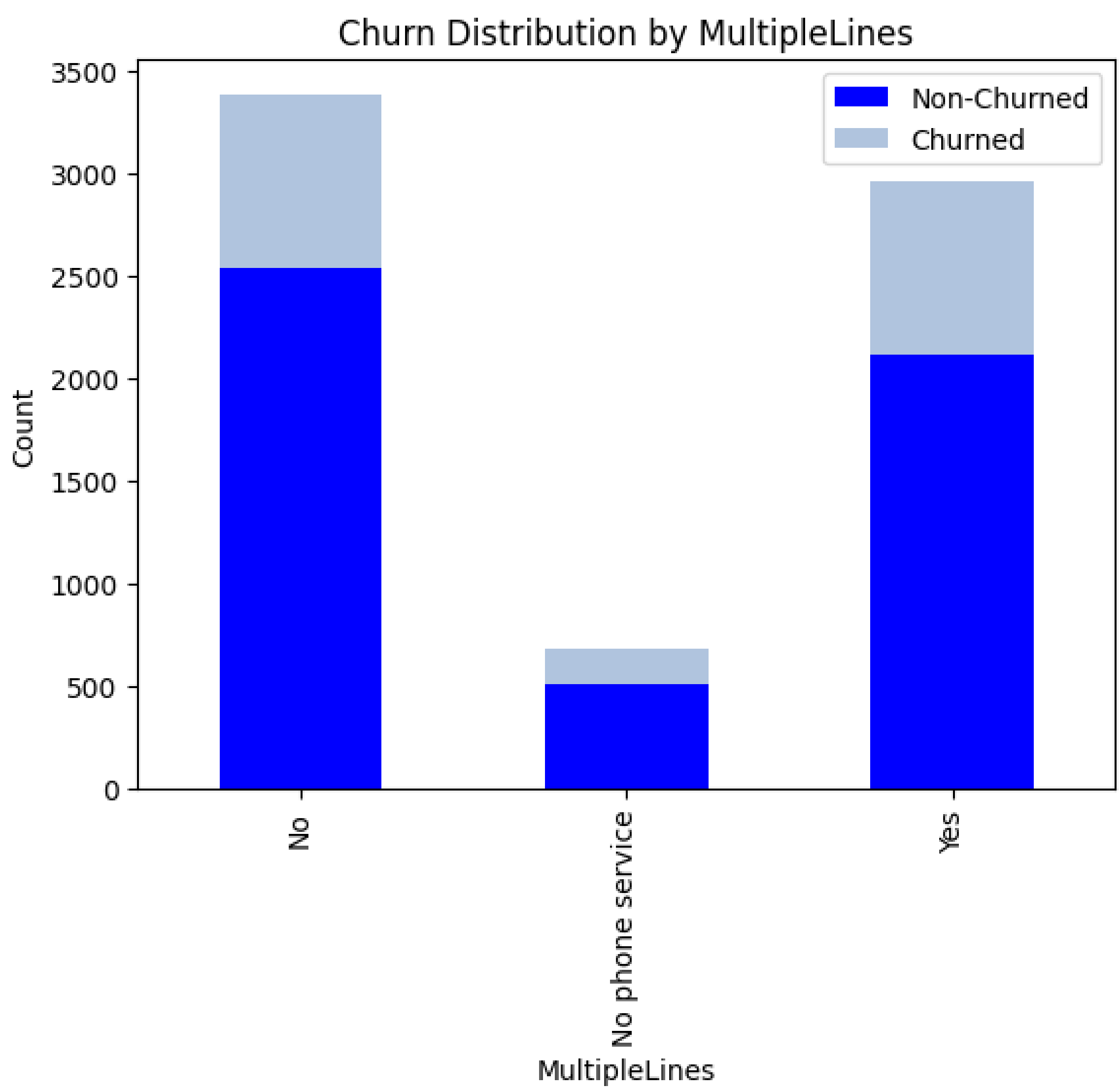


Insights:

The churn rate is not affected by phone services.

12. How does having multiple lines or not affect the churn rate?

```
df["MultipleLines"].unique()


df_grouped = df.groupby(["MultipleLines", "Churn"]).size().unstack()
df_grouped.plot(kind= "bar", stacked=True)
plt.title("Churn Distribution by MultipleLines")
plt.xlabel("MultipleLines")
plt.ylabel("Count")
plt.legend(["Non-Churned", "Churned"])
```



Insight:

- The portion of customers who have phone services but without multiple lines is a little higher than those who have multiple lines
- Little proportion of customers who don't have a phone service churn.
- There is no significant difference in churning rate between customers with Multiple lines and those without multiple lines.

13. Which Internet Service Provider has the highest churn rate?

df["InternetService"].unique()

df["Churn_N"] = df["Churn"].map({"Yes":1, "No": 0})

internet_churn = df.groupby("InternetService")["Churn_N"].mean().reset_index()

internet_churn = internet_churn.sort_values("Churn_N", ascending=False)

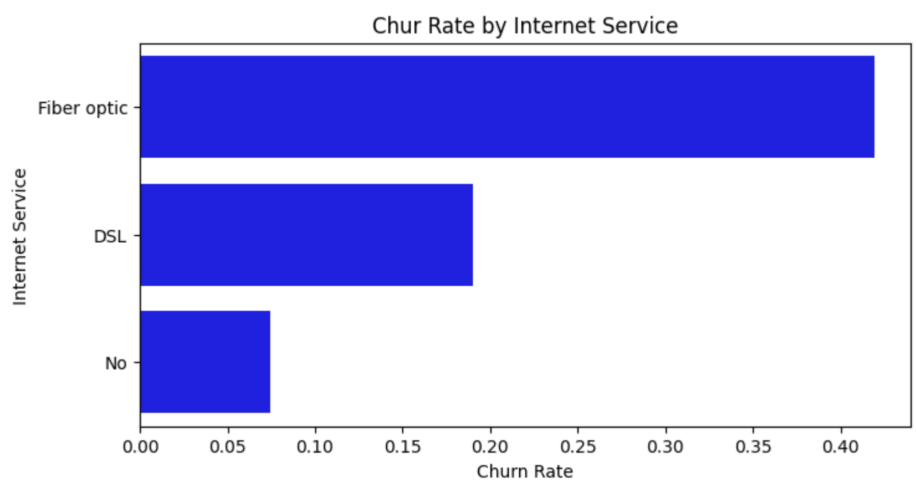
plt.figure(figsize=(8,4))

sns.barplot(x="Churn_N", y="InternetService", data= internet_churn, color = "lightblue")

plt.xlabel("Churn Rate")

plt.ylabel("Internet Service")

plt.title("Chur Rate by Internet Service")

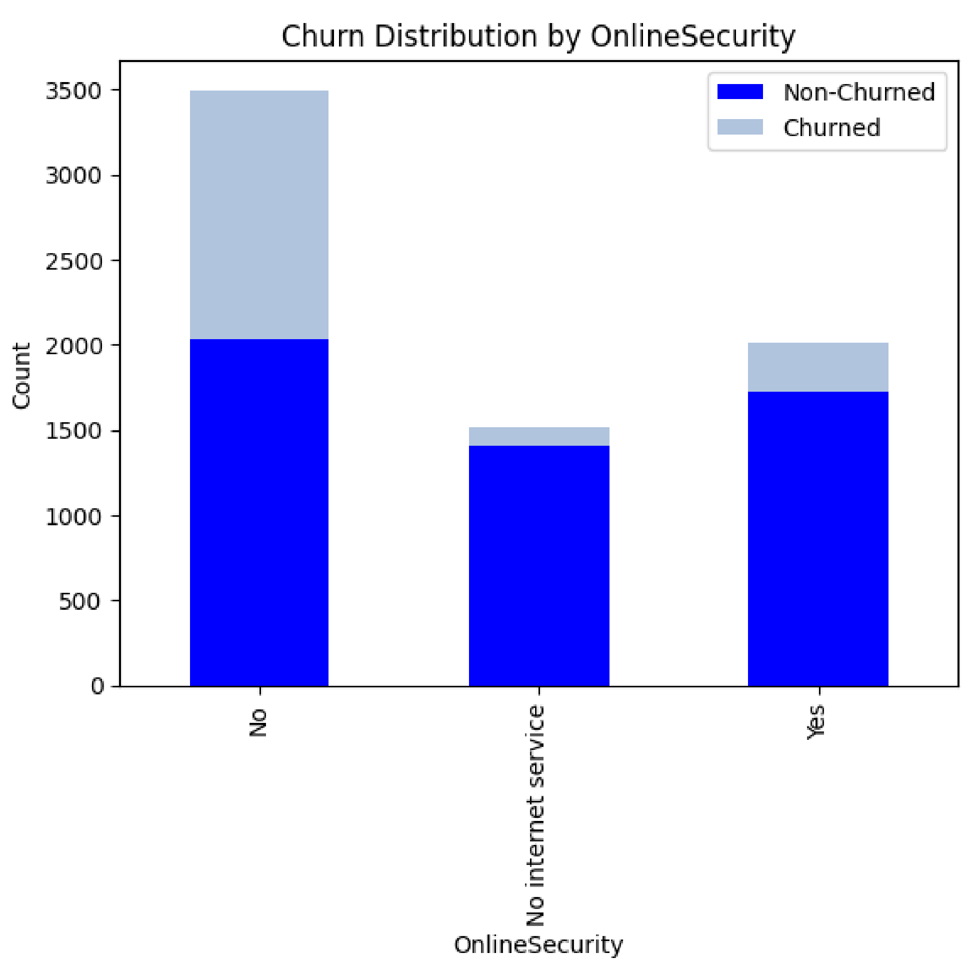


Insight:

Fiber optic has the highest churn rate

14. Do customers with online security tend to churn less frequently?

```
df["Churn"].unique()
df.columns
### Do customers with online security tend to churn less frequently?
df_grouped = df.groupby(["OnlineSecurity", "Churn"]).size().unstack()
df_grouped.plot(kind= "bar", stacked=True)
plt.title("Churn Distribution by OnlineSecurity")
plt.xlabel("OnlineSecurity")
plt.ylabel("Count")
plt.legend(["Non-Churned", "Churned"])
```



Insight:

The larger ratio of customers don't have Online security service, and it appears that they tend to churn more than the customers with online security.

SUMMARY

Features that influence churn rate:

- SeniorCitizen: yes
- Partner: no
- Dependents: no
- Tenure: low tenure
- InternetService: fiber optic
- OnlineSecurity: no

Features that don't influence churn rate:

- Gender
- PhoneService
- MultipleLines