NAME: DESMOND ELORM HONU

COURSE: DATA SCIENCE

ROLL NUMBER: 10211100281

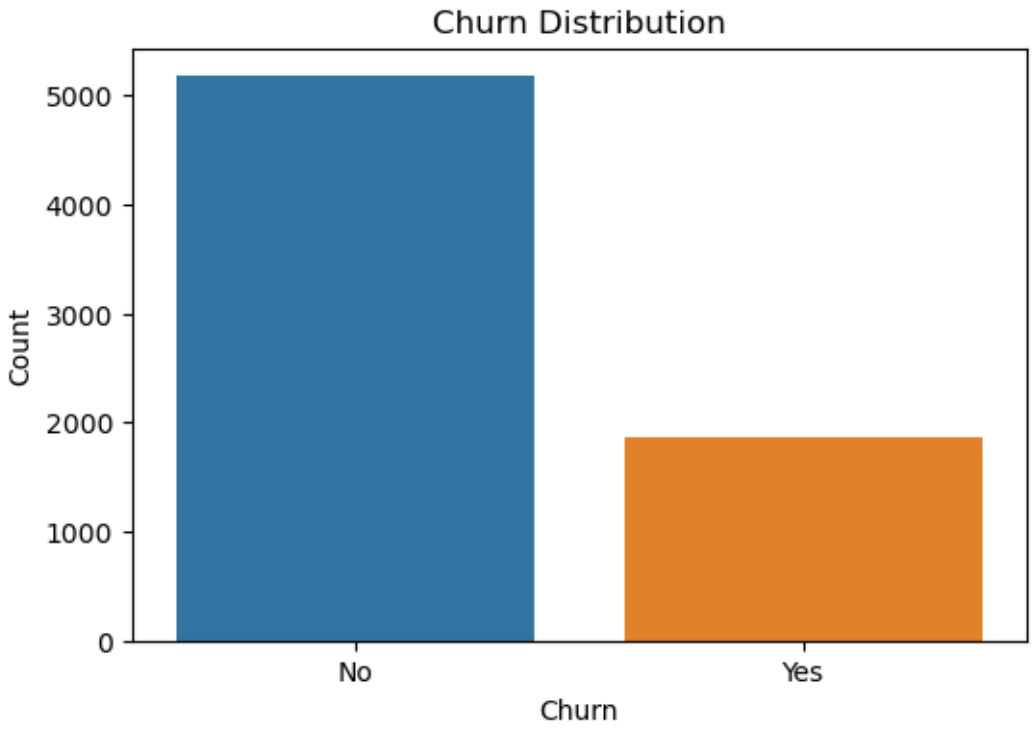
Untitled

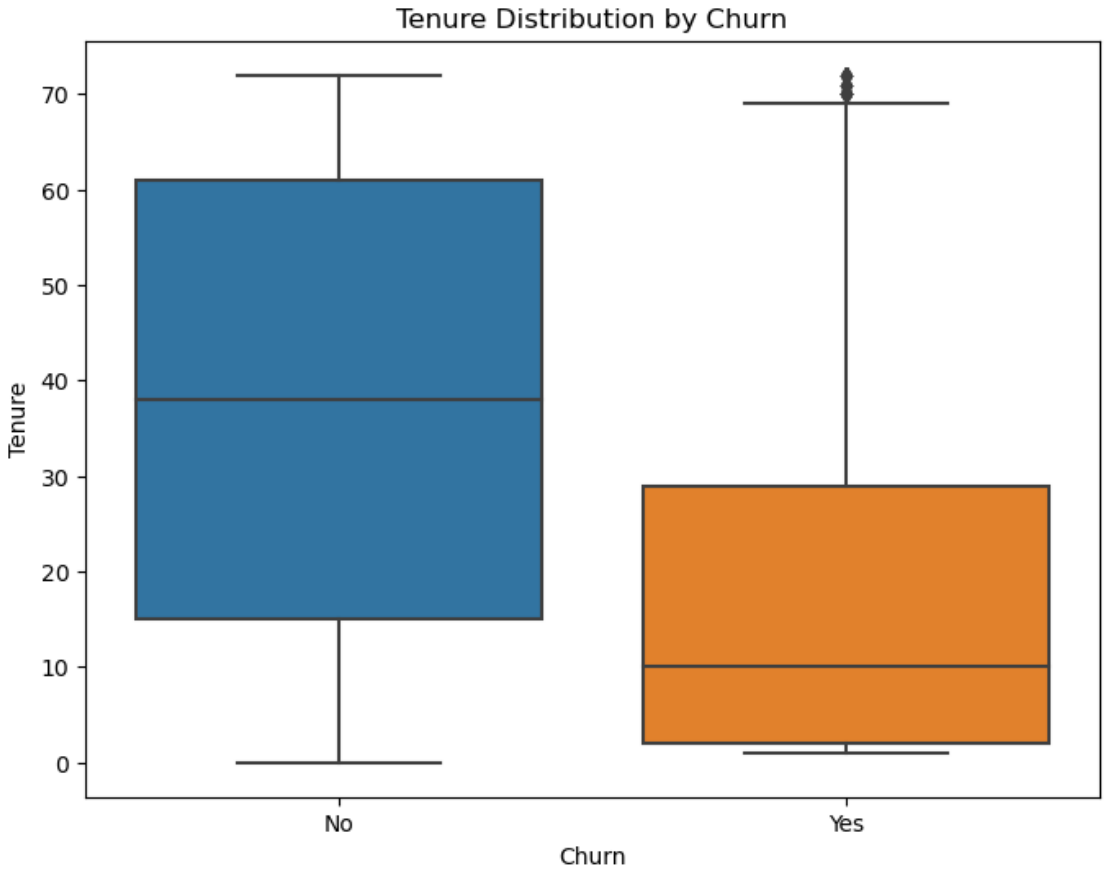
October 16, 2023

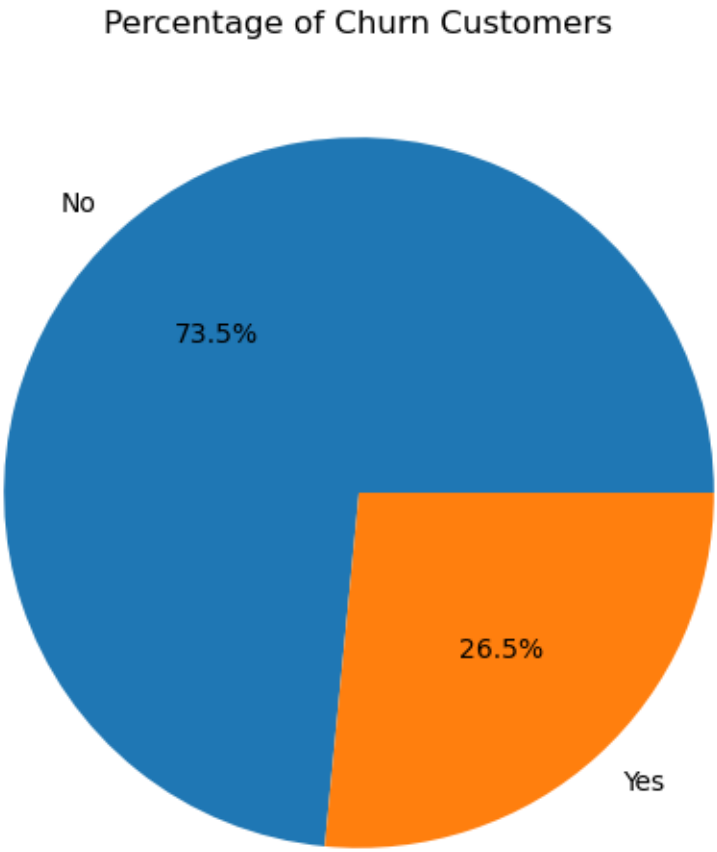
[9]:
```
# Import necessary libraries
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns


data = pd.read_csv("WA_Fn-UseC_-Telco-Customer-Churn.csv")

 #1. What is the target variable in this preamble?

target_variable = "Churn"

# 2. What type of machine learning problem is this?
#Explain.
#This is a binary classification problem because the target variable "Churn"␣
 ↪has two classes: "Yes" or "No."
#The goal is to predict whether a customer will churn (leave) or not.

# 3. Display the column names and data types of the dataset.
column_names = data.columns
data_types = data.dtypes

# 4. Check for missing values and handle them if there are any.
missing_values = data.isnull().sum()


# 5. Identify if there are duplicates and implement how to handle them.
duplicates = data.duplicated()


# 6. Are there categorical features that need to be transformed to numeric␣
 ↪values?
#If yes, how would you transform them?
# Identify categorical features and transform them using one-hot encoding or␣
 ↪label encoding.
```

```
# 7. What is the distribution of churned and non-churned customers in the dataset?
churn_distribution = data[target_variable].value_counts()

# 8. Define an outlier. Give a practical example.
 #An outlier is an observation that lies an abnormal distance from other values in a random sample.

# 9. Are there any outliers in the numerical features, and should they be removed or transformed?
#Identify outliers using methods like IQR or Z-score, and decide whether to remove or transform them.

# 10. Is there any pattern in Churn Customers based on gender?
churn_gender_pattern = data.groupby(['gender', 'Churn']).size().unstack()

# 11. What is the percentage of Churn Customers and customers that keep in with the active services?
churn_percentage = (data[target_variable].value_counts(normalize=True) * 100).round(2)

# 12. Implement 3 (three) data visualizations relevant to the study.
# Here are three example visualizations:
# a. Bar plot showing the distribution of Churn.
plt.figure(figsize=(6, 4))
sns.countplot(x=target_variable, data=data)
plt.title("Churn Distribution")
plt.xlabel("Churn")
plt.ylabel("Count")
plt.show()

# b. Box plot to visualize the distribution of tenure.
plt.figure(figsize=(8, 6))
sns.boxplot(x=target_variable, y='tenure', data=data)
plt.title("Tenure Distribution by Churn")
plt.xlabel("Churn")
plt.ylabel("Tenure")
plt.show()

# c. Pie chart to visualize the percentage of Churn Customers.
plt.figure(figsize=(6, 6))
plt.pie(churn_percentage, labels=churn_percentage.index, autopct='%1.1f%%')
plt.title("Percentage of Churn Customers")
plt.show()
```

```
Matplotlib is building the font cache; this may take a moment.
```


Churn Distribution
5000
4000
3000
2000
1000
0
Count
No
Yes
Churn


Tenure Distribution by Churn
70
60
50
40
30
20
10
0
Tenure
No
Yes
Churn


Percentage of Churn Customers
No
73.5%
26.5%
Yes

[ ]: