import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

# Ignore warnings

import warnings

warnings.filterwarnings('ignore')

# Load the dataset

data = pd.read\_csv("BankChurners\_Cleaned.csv")

# General information about the dataset

print("Dataset Information:")

print(data.info())

# Display the first few rows of the dataset

print("\nFirst 5 Rows of the Dataset:")

print(data.head())

# Summary statistics

print("\nSummary Statistics:")

print(data.describe().T)

# Inspect data types

print("\nData Types:")

print(data.dtypes)

# Check the number of rows and columns

print("\nDataset Shape:")

print(data.shape)

# Check for missing values

missing\_values = data.isnull().sum()

missing\_percentage = (missing\_values / len(data)) \* 100

print("\nMissing Values Count:")

print(missing\_values)

print("\nPercentage of Missing Values:")

print(missing\_percentage)

# Check the number of unique values in each column

print("\nNumber of Unique Values in Each Column:")

print(data.nunique())

# Distribution of a sample numerical column (e.g., 'Customer\_Age' if present)

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

sns.histplot(data['Customer\_Age'], kde=True, color='skyblue', bins=10)

plt.title("Customer Age Distribution")

plt.xlabel("Age")

plt.ylabel("Frequency")

plt.show()

# Countplot for categorical columns (customize as per your dataset)

categorical\_columns = data.select\_dtypes(include=['object']).columns

for col in categorical\_columns:

plt.figure(figsize=(6, 3))

sns.countplot(x=col, data=data, palette='cool')

plt.title(f"Distribution of {col}")

plt.xlabel(col)

plt.ylabel("Count")

plt.xticks(rotation=45)

plt.show()

# Select only numeric columns for the correlation matrix

numerical\_data = data.select\_dtypes(include=['float64', 'int64'])

# Compute the correlation matrix

correlation\_matrix = numerical\_data.corr()

# Plot the heatmap

plt.figure(figsize=(12, 8))

sns.heatmap(correlation\_matrix, annot=True, fmt=".2f", cmap="coolwarm", cbar=True)

plt.title("Correlation Heatmap")

plt.show()

# Boxplot for age distribution by churn status (example column 'Attrition\_Flag' if present)

if 'Attrition\_Flag' in data.columns:

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

sns.boxplot(x='Attrition\_Flag', y='Customer\_Age', data=data, palette='muted')

plt.title("Customer Age Distribution by Attrition Flag")

plt.xlabel("Attrition Flag")

plt.ylabel("Age")

plt.show()

# Distribution of numerical columns

numerical\_columns = data.select\_dtypes(include=['float64', 'int64']).columns

for col in numerical\_columns:

plt.figure(figsize=(6, 3))

sns.histplot(data[col], kde=True, bins=20, color='blue')

plt.title(f"Distribution of {col}")

plt.xlabel(col)

plt.ylabel("Frequency")

plt.show()

# Display skewness for numerical columns

skewness = data[numerical\_columns].skew()

print("\nSkewness of Numerical Columns:")

print(skewness)

Output:













































































