EXPLORATORY DATA ANALYSIS FOR MACHINE LEARNING

Dennis Kweku Botwe

About the data

This telco company's data set, which stems from the IBM sample set collection. The company provides home and internet services to 7043 customers in California. The challenge is to help the company analyze all relevant customer data to develop focused customer retention programs. The data is obtained from Kaggle (*https://www.kaggle.com/blastchar/telco-customer-churn*)

Plan for Data exploration

Here are the steps used to establish a baseline model to determine whether there is any strong correlation between the different factors and the target variable:

1. Data Overview
2. Data Cleaning and Engineering
3. Exploratory Data Analysis
4. Further Feature engineering
5. Hypothesis Testing
6. Conclusion

Data Overview

- The dataset consists of 7043 columns and 21 rows.
- Churn is the target variable to predict. In this column, 'Yes' means the customer churned and 'No' means the customer is still with the company.
- Other attributes which are in the dataset are:
  customerID, gender, SeniorCitizen, Partner, Dependents, tenure, PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges, and TotalCharges.
- The dataset consists of both numeric and categorical values and there are no missing valu es in the dataset based on count of non-null as shown in the image below.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   customerID        7043 non-null   object
 1   gender            7043 non-null   object
 2   SeniorCitizen     7043 non-null   int64
 3   Partner           7043 non-null   object
 4   Dependents        7043 non-null   object
 5   tenure            7043 non-null   int64
 6   PhoneService      7043 non-null   object
 7   MultipleLines     7043 non-null   object
 8   InternetService   7043 non-null   object
 9   OnlineSecurity    7043 non-null   object
 10  OnlineBackup      7043 non-null   object
 11  DeviceProtection  7043 non-null   object
 12  TechSupport       7043 non-null   object
 13  StreamingTV       7043 non-null   object
 14  StreamingMovies   7043 non-null   object
 15  Contract          7043 non-null   object
 16  PaperlessBilling  7043 non-null   object
 17  PaymentMethod     7043 non-null   object
 18  MonthlyCharges    7043 non-null   float64
 19  TotalCharges      7043 non-null   object
 20  Churn             7043 non-null   object
dtypes: float64(1), int64(2), object(18)
memory usage: 1.1+ MB
```

Data Cleaning and Engineering

Categorical columns

Any categorical variable that has more than two unique values are dealt with Label Encoding. Five columns were newly encoded and they are PaperlessBilling, Partner, Dependents, Churn and PhoneService.

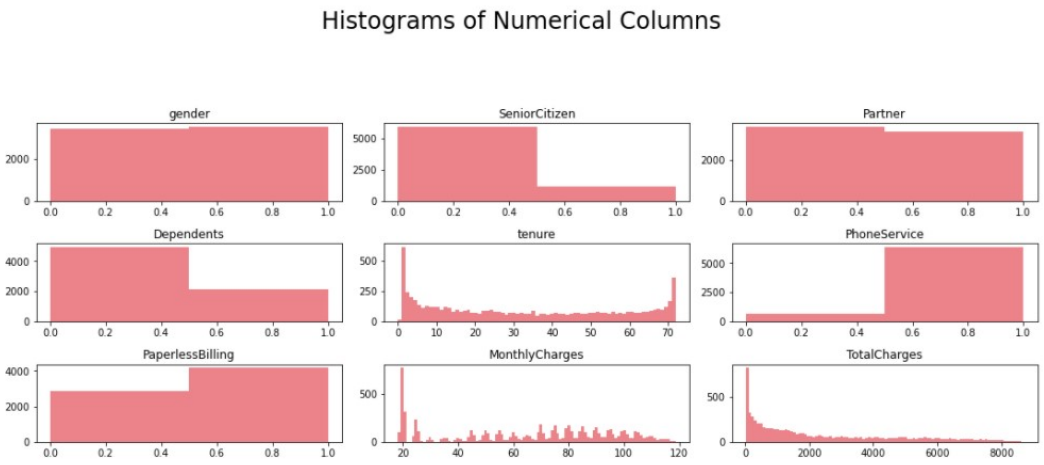
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   customerID        7043 non-null   object
 1   gender            7043 non-null   int32
 2   SeniorCitizen     7043 non-null   int64
 3   Partner           7043 non-null   int32
 4   Dependents        7043 non-null   int32
 5   tenure            7043 non-null   int64
 6   PhoneService      7043 non-null   int32
 7   MultipleLines     7043 non-null   object
 8   InternetService   7043 non-null   object
 9   OnlineSecurity    7043 non-null   object
 10  OnlineBackup      7043 non-null   object
 11  DeviceProtection  7043 non-null   object
 12  TechSupport       7043 non-null   object
 13  StreamingTV       7043 non-null   object
 14  StreamingMovies   7043 non-null   object
 15  Contract          7043 non-null   object
 16  PaperlessBilling  7043 non-null   int32
 17  PaymentMethod     7043 non-null   object
 18  MonthlyCharges    7043 non-null   float64
 19  TotalCharges      7043 non-null   float64
 20  Churn             7043 non-null   int32
dtypes: float64(2), int32(6), int64(2), object(11)
memory usage: 990.5+ KB
```

Exploratory Data Analysis

Exploring and visualizing the dataset by doing distribution of independent variables to better understand the patterns in the data and to potentially form some hypothesis.

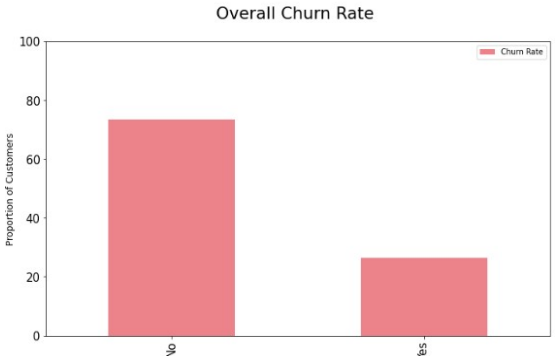
Numeric Data

Plotting histogram of numeric Columns:



From the plot it can be said that:

- Almost half of the customers in the dataset are female while the other half are male.
- Most of the customers in the dataset are younger people.
- Not many customers seem to have dependents while almost half of the customers have a partner.
- There are a lot of new customers in the organization (less than 10 months old) followed by a loyal customer segment that stays for more than 70 months on average.
- The dataset.describe() gives the insight that Monthly charges span anywhere between $18 to $118 per customer with a huge proportion of customers on $20 segment.
- Also, the overall churn rate shows that 73.5% of the customers are still in contract with the company and 26.5% have churn as shown in the graph below.



Further Engineering for the Categorical Columns

After applying get_dummies() to the dataset, 21 new columns were added to the dataset.

```
Number of rows: 7043
Number of columns: 42
```

Hypothesis testing

Hypothesis about this data:

1. **Null Hypothesis:** There is no relationship between Monthly charges with respect to the Churn rate.
   Alternate Hypothesis: There is a relationship between Monthly charges with respect to the Churn rate.

2. **Null Hypothesis:** There is no relationship between gender and churn.
   Alternate Hypothesis: There is a relationship between gender and churn.

3. **Null Hypothesis:** There is no relationship between Paperless billings and the churn rate.
   Alternate Hypothesis: There is a relationship between Paperless billings and the churn rate

1. For the first hypothesis before performing the test, the significance level was set to 95% i.e., p-value of 0.05. From the given the dataset the hypothesis test was performed and the result found was: The Pearson Correlation Coefficient, 0.19335642223784708 with a P-value of P = 2.7066456068899293e-60. The p-value is much less then the significance level and hence I decided to reject the null hypothesis that there exists no relationship between the variable Monthly Charges in relation to the target variable Churn.

2. For the second hypothesis, The Pearson Correlation Coefficient was -0.00861209507899 7791 with a P-value of P = 0.4699045390984688
   The data has weak negative correlation, and it's significant as p-value is almost closer to 0.05

3. For the third hypothesis, The Pearson Correlation Coefficient was 0.19182533166646779 with a P-value of P = 2.356554320584462e-59
   The data has weak positive correlation, and it's significant as p-value is a lot lesser than 0.05

Conclusion

- The dataset does not have any missing data values.
- Most of the customers seem to have phone service with Monthly charges spanning between $18 to $118 per customer.
- After the transformation, 22 new rows were added. Making it a total of 42 columns and 7043 rows.
- The Churn column contains two value counts with less 'Yes' than 'No'. This interprets less people are likely to churn.