Capstone Project 1: Data Wrangling

The data was downloaded from Kaggle. Each row represents a customer and each column contains customer's attributes described on the column Metadata. The "Churn" column is our target.

The data set includes information about :

- Customers who left within the last month – the column is called Churn
- 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

The data was then imported into a Pandas DataFrame for ease of data manipulation. The DataFrame contains 7043 rows (customers) and 21 columns (features).

Initially on checking the information of the DataFrame, no nulls were shown. On further inspection I saw the column 'TotalCharges' had a few rows with spaces, 11 rows to be exact. Tenure for all these 11 rows is 0, churn is "No" as well. One can interpret this as belonging to all new customers, if indeed these are the only rows with tenure = 0 too. Hence we can set 'TotalCharges' to 0, whenever tenure is 0. So I used the replace function to convert the spaces to 0. I also checked for the unique values and items of each column. The 'customerID' column is of no use to us, so I created a new DataFrame without the 'customerID' column using iloc. The new shape of the DataFrame is 7043 rows by 20 columns.

On the whole this data was very clean and required minimal data cleaning steps. As of now the DataFrame looks good to be used for further analysis. If any data manipulation is required in the later part of the project, it'll be done accordingly.