

Telco Consumer Churn Data Cleaning Report

By: Rachel Krisyanti

In the following, I provide a report on the results of data cleaning on the Telco Customer Churn dataset.

a. Data Information

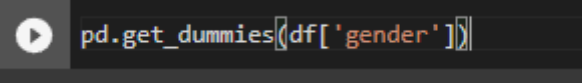
The datasets used are data entitled Telco Consumer Churn. That is data used for a focused customer retention program. The data can be used to : “Predict behavior to retain customers. You can analyze all relevant customer data and develop focused retention programs.” (IBM Sample Data Sets).

This dataset consists of 21 columns where each column consists of 7043 rows. The columns are as described below.

No.	Column Name	Column Description
1	Customer ID	Contains customer ID number, consist of 7043 unique value
2	Gender	Whether the customer is a male of a female
3	Senior Citizen	Whether the customer is a senior citizen or not
4	Partner	Whether the customer has a partner or not
5	Dependents	Whether the customer has dependents or not
6	Tenure	Number of months the customer has stayed with the company
7	Phone Service	Whether the customer has a phone service or not
8	Multiple Lines	Whether the customer has multiple lines or not
9	Internet Service	Customer's internet service provider (DSL, Fiber optic, No)
10	Online Security	Whether the customer has online security or not (Yes, No, No Internet Service)
11	Online Backup	Whether the customer has online backup or not (Yes, No, No Internet Service)
12	Device Protection	Whether the customer has device protection or not (Yes, No, No Internet Service)
13	Tech Support	Whether the customer has tech support or not (Yes, No, No internet service)
14	StreamingTV	Whether the customer has streaming TV or not (Yes, No, No internet service)
15	StreamingMovies	Whether the customer has streaming movies or not (Yes, No, No internet service)
16	Contract	The contract term of the customer (Month-to-month, One year, Two year)
17	PaperlessBilling	Whether the customer has paperless billing or not (Yes, No)
18	PaymentMethod	The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
19	MonthlyCharges	The amount charged to the customer monthly
20	TotalCharges	The total amount charged to the customer
21	Churn	Whether the customer churned or not (Yes or No)

b. Details on Data Cleaning (column name, what you did, reasoning, code screenshot)

No.	Column name	Treatment	Reason
1	Total Charges	Perform imputation by filling in the missing value with the median value.	Because after checking with the code df.info () there are 11 lines with no value. Then check the distribution of the data using the Box Plot diagram, found the data in the form of skewed. Then it is considered to fill in the median value so that it is not biased with the mean value which will have a wide range considering the skewed data distribution.
	Code screenshot		
	<pre>[12] df['TotalCharges'] = pd.to_numeric(df['TotalCharges'], errors='coerce') df = df.replace(np.nan, 0, regex=True) print(df) print(df.dtypes)</pre> <p>1) Code to change string data type to float.</p> <pre># Mean imputation on Rating column med_TotalCharges = df[df['TotalCharges']!=0]['TotalCharges'].median() df['TotalCharges'] = df['TotalCharges'].apply(lambda x: med_TotalCharges if x== 0 else x) df</pre> <p>2) Code to perform imputation</p>		
2	Customer ID	Drop the column	Because it should not effect the analysis we would conduct.
	Code screenshot		
	<pre>[45] df = df.drop(columns=['customerID'])</pre>		
3	Gender	Perform categorical data encoding (One Hot Encoding)	Because the data is and object data type, so it is necessary to do categorical data encoding so that it can be used when modelling. And performed one hot encoding because the data in the column is not ordinal or provides level information.
	Partner		
	Dependents		
	Phone Service		
	Multiple Lines		
	Internet Services		
	Online Security		
	Online Backup		
	Device Protection		
	Tech Support		
	Streaming TV		
	Streaming Movies		
	Contract		
	Paperless Billing		

	Payment Method		
	Churn		
	Code screenshot (given one code screenshot to represent all columns as an example)		
			

- c. Link To Google Colab Notebook

<https://colab.research.google.com/drive/1c1zpasufkMC4MLyAptH8gp2ps5NoC3oG#scrollTo=WCO0Cj6XHi5J>