PZSIB Batch 4

DATAFRAME BASICS AND DATA CLEANSING

noviyantidesi01@gmail.com

TABLE OF CONTENT

IMPORT LIBRARIES

```
import warnings
warnings.filterwarnings('ignore')

import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
```

ELABORATE

1

Warnings

We can configure the interpreter to ignore or hide the warnings, so they are not displayed in the program output.

2

Numpy

For perfoming matheatical calculations by simplifying the same from its predefined functions.

3

Pandas

Data manipulation library that helps play with DataFrames.

4

Matplotlib

This is one of the most common libraries used for visualization closest to the python backend.

5

Seaborn

Another visualization library with better representation, look, and feel but built on top of mathplotlib.

READ DATASET

1. https://www.kaggle.com/datasets/blastchar/telco-customer-churn

2.
```
df = pd.read_csv("/content/WA_Fn-UseC_-Telco-Customer-Churn.csv")
df
```

3.
```
df.info()
```

4.
```
df['TotalCharges'] = pd.to_numeric(df['TotalCharges'], errors='coerce')
df.info()
```

5.
```
df.isnull().sum()
```

ELABORATE

1

- **Open** link
- **Download dataset** dengan domain CSV tersebut.

2

- **Upload** file ke Goggle Colab
- Gunakan fungsi **read_csv()** dari Pandas untuk membaca **file CSV** dan mengonversinya **menjadi dataframe**.
- **Print dataframe**, sehingga menampilkan data dalam bentuk **tabel yang menggambarkan struktur data** dalam file CSV tersebut. Berikut data yang dihasilkan

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	...	No	No	No	No	Month-to-month	Yes	Electronic check	29.85	29.85	No
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	...	Yes	No	No	No	One year	No	Mailed check	56.95	1889.5	No
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	...	No	No	No	No	Month-to-month	Yes	Mailed check	53.85	108.15	Yes
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	...	Yes	Yes	No	No	One year	No	Bank transfer (automatic)	42.30	1840.75	No
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	...	No	No	No	No	Month-to-month	Yes	Electronic check	70.70	151.65	Yes
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
7038	6840-RESVB	Male	0	Yes	Yes	24	Yes	Yes	DSL	Yes	...	Yes	Yes	Yes	Yes	One year	Yes	Mailed check	84.80	1990.5	No
7039	2234-XADUH	Female	0	Yes	Yes	72	Yes	Yes	Fiber optic	No	...	Yes	No	Yes	Yes	One year	Yes	Credit card (automatic)	103.20	7362.9	No
7040	4801-JZAZL	Female	0	Yes	Yes	11	No	No phone service	DSL	Yes	...	No	No	No	No	Month-to-month	Yes	Electronic check	29.60	346.45	No
7041	8361-LTMKD	Male	1	Yes	No	4	Yes	Yes	Fiber optic	No	...	No	No	No	No	Month-to-month	Yes	Mailed check	74.40	306.6	Yes
7042	3186-AJIEK	Male	0	No	No	66	Yes	No	Fiber optic	Yes	...	Yes	Yes	Yes	Yes	Two year	Yes	Bank transfer (automatic)	105.65	6844.5	No

7043 rows × 21 columns

ELABORATE

3

- **Menampilkan informasi tentang dataframe** "df" yang telah dibaca sebelumnya dari file CSV.
- Informasi yang dibaca seperti, **jumlah baris** dan **kolom**, **tipe data kolom, jumlah nilai non-null**, serta **penggunaan memori oleh dataframe**.
- Informasi yang didapatkan antara lain, **7043 rows** dan **21 columns**, terdapat **1 type Float**, **2 type Integer**, **18 type object**, serta **tidak terdapat null dalam setiap columns**.
- Informasi **TotalCharges** masih dalam type **object**, jika dilihat dari dataframe perlu adanya perubahan menjadi type **float**.

```
<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
```

ELABORATE

4

- **TotalCharges**, **diubah menjadi tipe data numerik** menggunakan fungsi **pd.to_numeric** dari Pandas. Parameter **errors='coerce'** digunakan untuk **mengatasi kesalahan** yang mungkin terjadi **saat mengkonversi**.
- **Run** mengenai **informasi terbaru** dari **dataframe**. Dalam hal ini, **Non-null count awal berjumlah 7043** sedangkan ketika sudah diubah ke dalam bentuk numerik **menjadi 7032 non-null**.

```
<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      7032 non-null   float64
 20  Churn             7043 non-null   object
dtypes: float64(2), int64(2), object(17)
memory usage: 1.1+ MB
```

5

- **df.isnull().sum()** digunakan untuk **mengitung jumah nilai yang hilang (NaN atau null)** dalam setiap kolom dataframe.
- Dalam **kolom TotalCharges** terdapat **11 data hilang atau NaN.**

```
customerID          0
gender              0
SeniorCitizen       0
Partner             0
Dependents          0
tenure              0
PhoneService        0
MultipleLines       0
InternetService     0
OnlineSecurity      0
OnlineBackup        0
DeviceProtection    0
TechSupport         0
StreamingTV         0
StreamingMovies     0
Contract            0
PaperlessBilling    0
PaymentMethod       0
MonthlyCharges      0
TotalCharges       11
Churn               0
dtype: int64
```

MISSING VALUES CHECKING

TotalCharges mempunyai NaN sejumlah 11, karena NaN tersebut termasuk ke dalam kategori kecil maka dapat dilakukan penghapusan secara langsung. **Batasan penghapusan data sebaiknya dilakukan ketika NaN kurang dari 10% dari data yang tersedia**. Penghapusan data dapat dilakukan dengan menggunakan code berikut :

df.dropna(subset=['TotalCharges'], inplace=True)

Setelah itu, deteksi jumlah NaN yang ada di dalam Dataframe sehingga di dapatkan output seperti gambar disamping. Artinya, **sudah tidak ada NaN di dalam kolom TotalCharges**.

```
customerID          0
gender              0
SeniorCitizen       0
Partner             0
Dependents          0
tenure              0
PhoneService        0
MultipleLines       0
InternetService     0
OnlineSecurity      0
OnlineBackup        0
DeviceProtection    0
TechSupport         0
StreamingTV         0
StreamingMovies     0
Contract            0
PaperlessBilling    0
PaymentMethod       0
MonthlyCharges      0
TotalCharges        0
Churn               0
dtype: int64
```

CATEGORICAL DATA ENCODING

Berdasarkan informasi dataframe **terdapat 18 object**, tetapi **yang dapat dikategorikan hanya berjumlah 17 kolom**, **1 kolom yang tidak dapat dikategorikan yaitu customerID**. **Batasan** pada dataframe ini yaitu **tidak memprediksi kasus data yang digunakan**. Oleh karena itu, **metode categorical Data Encoding yang digunakan adalah Label Encoding**. Label Encoding ini cocok digunakan karena dalam **setiap kolom hanya mempunyai 2-3 variabel** saja, serta jika menggunakan Label Encoding **tabel output yang dikeluarkan tetap sama dengan data sebelumnya**, yaitu sebanyak **21 kolom**.

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
0	7590-VHVEG	0	0	1	0	1	0	1	0	0	...	0	0	0	0	0	1	2	29.85	29.85	0
1	5575-GNVDE	1	0	0	0	34	1	0	0	2	...	2	0	0	0	1	0	3	56.95	1889.50	0
2	3668-QPYBK	1	0	0	0	2	1	0	0	2	...	0	0	0	0	0	1	3	53.85	108.15	1
3	7795-CFOCW	1	0	0	0	45	0	1	0	2	...	2	2	0	0	1	0	0	42.30	1840.75	0
4	9237-HQITU	0	0	0	0	2	1	0	1	0	...	0	0	0	0	0	1	2	70.70	151.65	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
7038	6840-RESVB	1	0	1	1	24	1	2	0	2	...	2	2	2	2	1	1	3	84.80	1990.50	0
7039	2234-XADUH	0	0	1	1	72	1	2	1	0	...	2	0	2	2	1	1	1	103.20	7362.90	0
7040	4801-JZAZL	0	0	1	1	11	0	1	0	2	...	0	0	0	0	0	1	2	29.60	346.45	0
7041	8361-LTMKD	1	1	1	0	4	1	2	1	0	...	0	0	0	0	0	1	3	74.40	306.60	1
7042	3186-AJIEK	1	0	0	0	66	1	0	1	2	...	2	2	2	2	2	1	0	105.65	6844.50	0

7032 rows × 21 columns

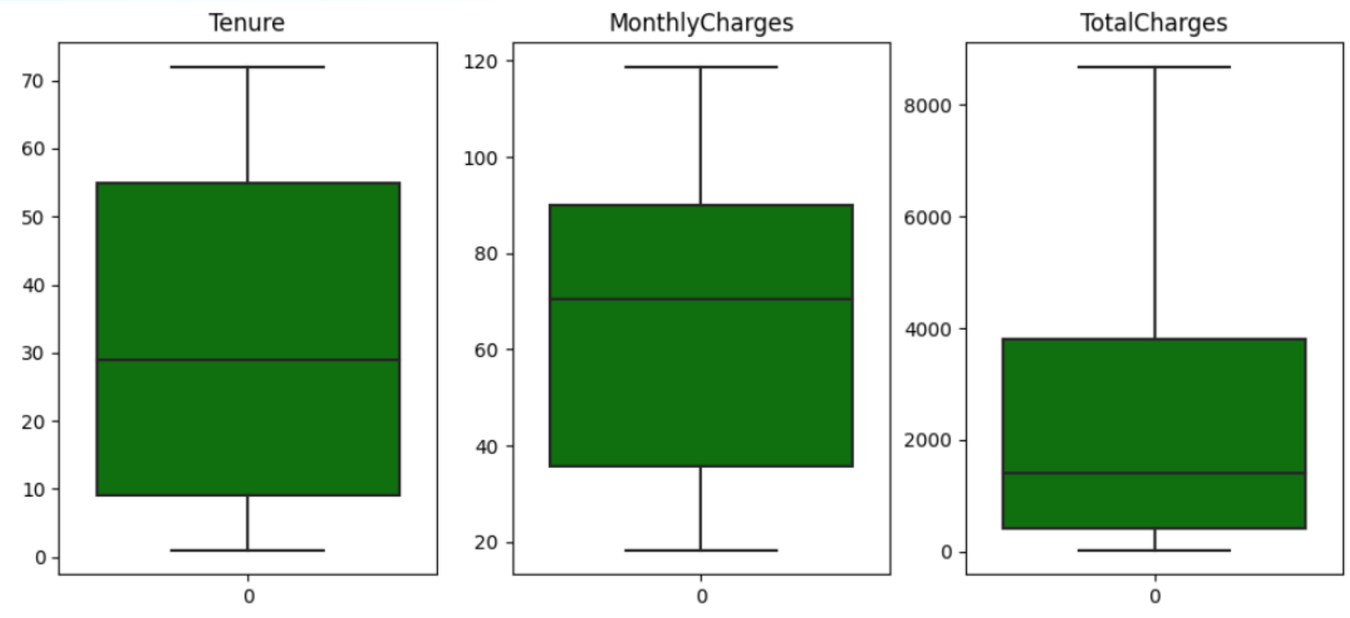
ANOMALIES AND OUTLIER HANDLING

Untuk mengetahui adanya outlier atau tidak, **perlu dilakukan pembacaan dataframe** seperti, count, mean, std, dan lainnya. **Dataframe yang digunakan ber-type float/numerik,** yaitu tenure, MonthlyCharges, dan Total Charges.

	tenure	MonthlyCharges	TotalCharges
count	7032.000000	7032.000000	7032.000000
mean	32.421786	64.798208	2283.300441
std	24.545260	30.085974	2266.771362
min	1.000000	18.250000	18.800000
25%	9.000000	35.587500	401.450000
50%	29.000000	70.350000	1397.475000
75%	55.000000	89.862500	3794.737500
max	72.000000	118.750000	8684.800000

Dari data sebelumnya, dapat dilakukan **visualisasi menggunakan boxplot**. Boxplot tersebut kemudian **dibagi menjadi tiga bagian** karena **jika digabungkan, nilai dari boxplot tenure dan MonthlyCharges tidak sebanding dengan TotalCharges**, sehingga **boxplot tidak dapat terlihat jelas** khususnya pada data tenure dan TotalCharges.

Dapat dilihat secara seksama **gambar boxplot di bawah ini, tidak ditemukan adanya outlier** pada setiap data yang ditampilkan. Oleh karena itu, dapat disimpulkan bahwa **tidak diperlukan penanganan lebih lanjut** terhadap data tersebut **untuk mengatasi nilai-nilai ekstrem yang berada di luar rentang normal.**



CODE PYTHON

Assignment 5

https://colab.research.google.com/drive/1OZ-4mMdttdK6NgGxO9ZBbZ_rvvKUeaJS?usp=sharing

Assignment 6

https://colab.research.google.com/drive/1UpFmxBqGP4b4S6e8brgYP2Wfmq-1LZJ8?usp=sharing

THOUGHTS AND SUGGESTIONS ARE WELCOMED!


noviyantidesi01@gmail.com


www.linkedin.com/in/desi-noviyanti/


github.com/desinoviyantii