# DATA CLEANSING

RUMAISYA AZ-ZAHRA

# Data Set Telco Churn

**Data Set Telco Churn perlu dibersihkan agar outputnya tidak menjadi sampah (masih kotor). Dilakukan 3 tahap Data Cleansing sebagai berikut :**

**Missing Value Checking**

↓

**Categorical Data Encoding**

↓

**Anomalies and Outlier Handling**

LANGKAH PERTAMA

# Import Dataset csv ke google colab

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

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

Untuk Mengignore pemberitahuan yang bukan error sehingga memudahkan dalam pengerjaan.

**Dataset diimport menggunakan pandas csv**

```
[100] df = pd.read_csv('Telco_churn.csv')
      df
```

| | customerID | gender | SeniorCitizen | Partner | Dependents | tenure | PhoneService | MultipleLines | InternetService | OnlineSecurity | ... | DeviceProtection | TechSupport | S |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 7590-VHVEG | Female | 0 | Yes | No | 1 | No | No phone service | DSL | No | ... | No | No | |
| **1** | 5575-GNVDE | Male | 0 | No | No | 34 | Yes | No | DSL | Yes | ... | Yes | No | |
| **2** | 3668-QPYBK | Male | 0 | No | No | 2 | Yes | No | DSL | Yes | ... | No | No | |
| **3** | 7795-CFOCW | Male | 0 | No | No | 45 | No | No phone service | DSL | Yes | ... | Yes | Yes | |
| **4** | 9237-HQITU | Female | 0 | No | No | 2 | Yes | No | Fiber optic | No | ... | No | No | |

01

# MISSING VALUE CHEKING

## MENCARI KOLOM YANG MISSING (NAN)

# CEK MISSING DATA

```
[102] df.info()
```

```
<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)
```

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

```
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
```

```
df.describe()
```

| | SeniorCitizen | tenure | MonthlyCharges |
|---|---|---|---|
| **count** | 7043.000000 | 7043.000000 | 7043.000000 |
| **mean** | 0.162147 | 32.371149 | 64.761692 |
| **std** | 0.368612 | 24.559481 | 30.090047 |
| **min** | 0.000000 | 0.000000 | 18.250000 |
| **25%** | 0.000000 | 9.000000 | 35.500000 |
| **50%** | 0.000000 | 29.000000 | 70.350000 |
| **75%** | 0.000000 | 55.000000 | 89.850000 |
| **max** | 1.000000 | 72.000000 | 118.750000 |

**Tidak ada kolom yang missing. Namun, setelah ditinjau kolom TotalCharges error sehingga tidak muncul di df.describe() dan perlu diperbaiki**

# MEMPERBAIKI KOLOM ERROR

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

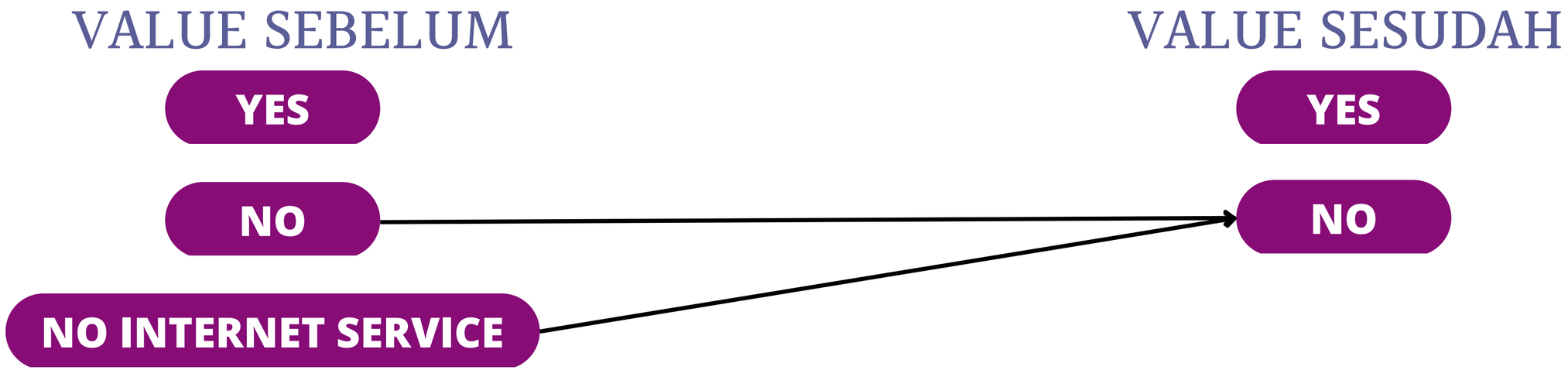
```
df.describe()
```

| | SeniorCitizen | tenure | MonthlyCharges | TotalCharges |
|---|---|---|---|---|
| count | 7043.000000 | 7043.000000 | 7043.000000 | 7032.000000 |
| mean | 0.162147 | 32.371149 | 64.761692 | 2283.300441 |
| std | 0.368612 | 24.559481 | 30.090047 | 2266.771362 |
| min | 0.000000 | 0.000000 | 18.250000 | 18.800000 |
| 25% | 0.000000 | 9.000000 | 35.500000 | 401.450000 |
| 50% | 0.000000 | 29.000000 | 70.350000 | 1397.475000 |
| 75% | 0.000000 | 55.000000 | 89.850000 | 3794.737500 |
| max | 1.000000 | 72.000000 | 118.750000 | 8684.800000 |

Data Total Charges Muncul

MENGELOMPOKAN VALUE YANG BERNILAI SAMA

# MENINJAU VALUE




Jika ditinjau lebih lanjut keduanya sama-sama bernilai "tidak melakukan" sehingga jawaban "No Internet service" bisa dikategorikan ke dalam jawaban / value "No" sehingga categorical data encodingnya akan menjadi lebih mudah.

MENGELOMPOKAN VALUE YANG BERNILAI SAMA

# MENINJAU VALUE

```
df['MultipleLines'].value_counts()
```

```
No                  3390
Yes                 2971
No phone service     682
Name: MultipleLines, dtype: int64
```

→ Cek terlebih dahulu jumlah jawaban / value secara keseluruh

```
df['MultipleLines'] = df['MultipleLines'].replace(['No phone service'], 'No')
```

| MultipleLines |
| --- |
| No phone service |
| No |

→

| MultipleLines |
| --- |
| No |
| No |

Replace jawaban Value pada kolom lainnya sesuai langkah-langkah ini

## MENGELOMPOKAN VALUE YANG BERNILAI SAMA

# MENINJAU VALUE

```
df.head()
```

| index | customerID | gender | SeniorCitizen | Partner | Dependents | tenure | PhoneService | MultipleLines | InternetService | OnlineSecurity | OnlineBackup | DeviceProtection | TechSupport | StreamingTV | StreamingMovies | Contract | PaperlessBilling | PaymentMethod | MonthlyCharges | TotalCharges |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 7590-VHVEG | Female | 0 | Yes | No | 1 | No | No | DSL | No | Yes | No | No | No | No | Month-to-month | Yes | Electronic check | 29.85 | 29.85 |
| 1 | 5575-GNVDE | Male | 0 | No | No | 34 | Yes | No | DSL | Yes | No | Yes | No | No | No | One year | No | Mailed check | 56.95 | 1889.5 |
| 2 | 3668-QPYBK | Male | 0 | No | No | 2 | Yes | No | DSL | Yes | Yes | No | No | No | No | Month-to-month | Yes | Mailed check | 53.85 | 108.15 |
| 3 | 7795-CFOCW | Male | 0 | No | No | 45 | No | No | DSL | Yes | No | Yes | Yes | No | No | One year | No | Bank transfer (automatic) | 42.3 | 1840.75 |
| 4 | 9237-HQITU | Female | 0 | No | No | 2 | Yes | No | Fiber optic | No | No | No | No | No | No | Month-to-month | Yes | Electronic check | 70.7 | 151.65 |

Hasil Kolom-kolom setelah dilakukan mereplace value yang bernilai sama agar memudahkan categorical data encoding

02

# CATEGORICAL DATA ENCODING

# MENGKATEGORIKAN DATA KATEGORI MENJADI NUMERIK

## LABEL ENCODING

Dilakukan pada kolom yang memiliki kategorikal sedikit yaitu kolom yang bervalue YES dan NO seperti Partner, Dependents, PhoneService, MultipleLines, OnlineSecurity, dan lainnya.

## ONE HOT ENCODING

Digunakan dalam kolom yang bervalue sedikit. Pada data case. encoding ini digunakan dalam kolom yang memiliki value 3 antara lain kolom InternetService dan Contract. Dari hal ini, sehingga kolom akan bertambah sesuai jumlah value.

## FREQUENCY ENCODING

Digunakan pada kolom yang memiliki value yang cukup banyak. Pada dataset ini, frequency encoding digunakan pada kolom PaymentMethod karena value yang dihasilkannya berjumlah 4.

# LABEL ENCODING

Kolom 'gender'

```
df['gender'] = df['gender'].astype('category').cat.codes
df.head()
```

| | customerID | gender | SeniorCitizen | Partner |
|---|---|---|---|---|
| 0 | 7590-VHVEG | 0 | 0 | Yes |
| 1 | 5575-GNVDE | 1 | 0 | No |
| 2 | 3668-QPYBK | 1 | 0 | No |
| 3 | 7795-CFOCW | 1 | 0 | No |

**0 = Female**
**1 = male**

Kolom 'Partner'

```
df['Partner'] = df['Partner'].astype('category').cat.codes
df.head()
```

| | customerID | gender | SeniorCitizen | Partner |
|---|---|---|---|---|
| 0 | 7590-VHVEG | 0 | 0 | 1 |
| 1 | 5575-GNVDE | 1 | 0 | 0 |
| 2 | 3668-QPYBK | 1 | 0 | 0 |
| 3 | 7795-CFOCW | 1 | 0 | 0 |

**0 = NO**
**1 = YES**

**Lakukan hal yang sama pada kolom yang memiliki value YES/NO**

# ONE HOT ENCODING

## Kolom 'InternetService'

```
df['InternetService'].value_counts()
```

```
Fiber optic     3096
DSL             2421
No              1526
Name: InternetService, dtype: int64
```

**Cek jumlah value Keseluruhan**

```
dummies_InternetService = pd.get_dummies(df['InternetService'],prefix='InternetService')
dummies_InternetService.head()
```

```
df = pd.concat([df, dummies_InternetService], axis=1)
df.head()
```

```
df = df.drop('InternetService',axis=1)
df.head()
```

**Jumlah kolom bertambah sesuai value**

| ling | PaymentMethod | MonthlyCharges | TotalCharges | Churn | ternetService_DSL | InternetService_Fiber optic | InternetService_No |
|---|---|---|---|---|---|---|---|
| Yes | Electronic check | 29.85 | 29.85 | No | 1 | 0 | 0 |
| No | Mailed check | 56.95 | 1889.50 | No | 1 | 0 | 0 |
| Yes | Mailed check | 53.85 | 108.15 | Yes | 1 | 0 | 0 |
| No | Bank transfer (automatic) | 42.30 | 1840.75 | No | 1 | 0 | 0 |
| Yes | Electronic check | 70.70 | 151.65 | Yes | 0 | 1 | 0 |

**Lakukan hal yang sama pada kolom yang memiliki value berjumlah 3**

# FREQUENCY ENCODING

## Kolom 'PaymentMethod'

```
df['PaymentMethod'].value_counts()
```

```
Electronic check             2365
Mailed check                 1612
Bank transfer (automatic)    1544
Credit card (automatic)      1522
Name: PaymentMethod, dtype: int64
```

**Cek jumlah value Keseluruhan**

```
freq_et = df['PaymentMethod'].value_counts().reset_index()
freq_et.rename(columns={"index": "PaymentMethod", "PaymentMethod": "freq_PaymentMethod"}, inplace = True)
freq_et['pct_PaymentMethod'] = round((freq_et['freq_PaymentMethod']/freq_et['freq_PaymentMethod'].sum())*100,2)
freq_et
```

| | PaymentMethod | freq_PaymentMethod | pct_PaymentMethod |
|---|---|---|---|
| 0 | Electronic check | 2365 | 33.58 |
| 1 | Mailed check | 1612 | 22.89 |
| 2 | Bank transfer (automatic) | 1544 | 21.92 |
| 3 | Credit card (automatic) | 1522 | 21.61 |

```
df = df.merge(freq_et[['PaymentMethod','pct_PaymentMethod']], on='PaymentMethod', how='inner')
df.head()
```

| ce_Fiber optic | InternetService_No | Contract_Month-to-month | Contract_One year | Contract_Two year | pct_PaymentMethod |
|---|---|---|---|---|---|
| 0 | 0 | 1 | 0 | 0 | 33.58 |
| 1 | 0 | 1 | 0 | 0 | 33.58 |
| 1 | 0 | 1 | 0 | 0 | 33.58 |

**Lakukan hal yang sama pada kolom yang memiliki value berjumlah 4 (cukup banyak)**

**Muncul sesuai persentase**

# HASIL AKHIR

## SETELAH DILAKUKAN ENCODING

## DATA KESELURUHAN MENJADI BERSIFAT NUMERIK

| index | customerID | gender | SeniorCitizen | Partner | Dependents | tenure | PhoneService | MultipleLines | OnlineSecurity | OnlineBackup | DeviceProtection | TechSupport | StreamingTV | StreamingMovies | PaperlessBilling | PaymentMethod | MonthlyCharges | TotalCharges | Churn | InternetService_DSL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 7590-VHVEG | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | Electronic check | 29.85 | 29.85 | 0 | 1 |
| 1 | 9237-HQITU | 0 | 0 | 0 | 0 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | Electronic check | 70.7 | 151.65 | 1 | 0 |
| 2 | 9305-CDSKC | 0 | 0 | 0 | 0 | 8 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | Electronic check | 99.65 | 820.5 | 1 | 0 |
| 3 | 7892-POOKP | 0 | 0 | 1 | 0 | 28 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | Electronic check | 104.8 | 3046.05 | 1 | 0 |
| 4 | 5129-JLPIS | 1 | 0 | 0 | 0 | 25 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | Electronic check | 105.5 | 2686.05 | 0 | 0 |

| ervice_Fiber optic | InternetService_No | Contract_Month-to-month | Contract_One year | Contract_Two year | pct_PaymentMethod | log_tenure | log_MonthlyCharges | log_Tota |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 1 | 0 | 0 | 33.58 | 0.693147 | 3.429137 | |
| 1 | 0 | 1 | 0 | 0 | 33.58 | 1.098612 | 4.272491 | |
| 1 | 0 | 1 | 0 | 0 | 33.58 | 2.197225 | 4.611649 | |
| 1 | 0 | 1 | 0 | 0 | 33.58 | 3.367296 | 4.661551 | |
| 1 | 0 | 1 | 0 | 0 | 33.58 | 3.258097 | 4.668145 | |

Kelanjutan Kolom

03

# ANOMALIES AND OUTLIER HANDLING

# PENGECEKAN DATA TIDAK NORMAL

```
df.describe()
```

```
df[['tenure']].describe()
```

| | tenure |
|---|---|
| count | 7043.000000 |
| mean | 32.371149 |
| std | 24.559481 |
| min | 0.000000 |
| 25% | 9.000000 |
| 50% | 29.000000 |
| 75% | 55.000000 |
| max | 72.000000 |

```
df[['MonthlyCharges']].describe()
```

| | MonthlyCharges |
|---|---|
| count | 7043.000000 |
| mean | 64.761692 |
| std | 30.090047 |
| min | 18.250000 |
| 25% | 35.500000 |
| 50% | 70.350000 |
| 75% | 89.850000 |
| max | 118.750000 |

```
df[['TotalCharges']].describe()
```

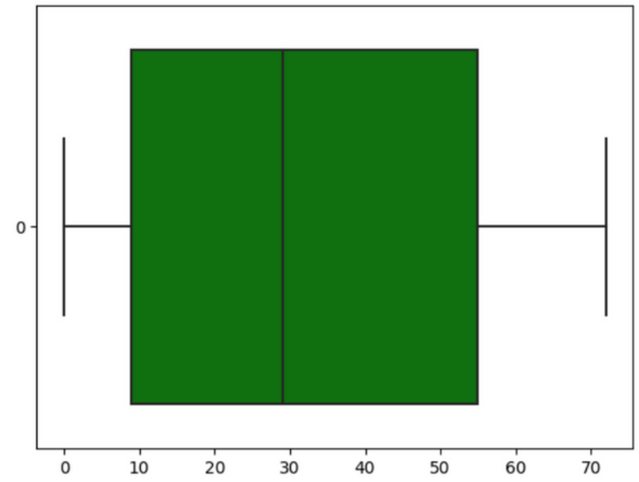
| | TotalCharges |
|---|---|
| count | 7032.000000 |
| mean | 2283.300441 |
| std | 2266.771362 |
| min | 18.800000 |
| 25% | 401.450000 |
| 50% | 1397.475000 |
| 75% | 3794.737500 |
| max | 8684.800000 |

Dari hasil tinjauan data yang tidak terlihat normal yaitu tenure, monthlycharges, dan total charges. Hal tersebut karena perbedaan antara nilai max yang besar, tetapi median, rata-rata, dsb kurang sesuai.

# VISUALISASI BOXPLOT

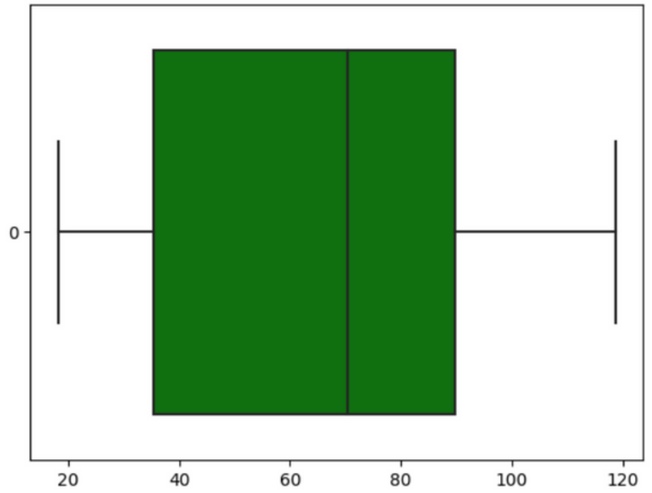
```
sns.boxplot(df['tenure'],color='green',orient='h')
```

```
<Axes: >
```



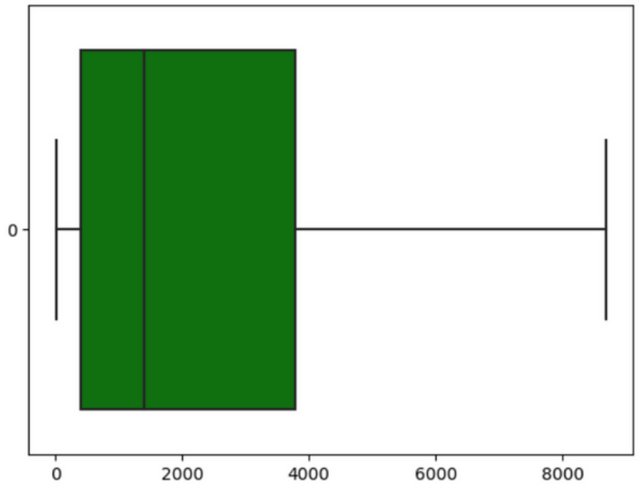

```
sns.boxplot(df['MonthlyCharges'],color='green',orient='h')
```

```
<Axes: >
```




```
sns.boxplot(df['TotalCharges'],color='green',orient='h')
```

```
<Axes: >
```



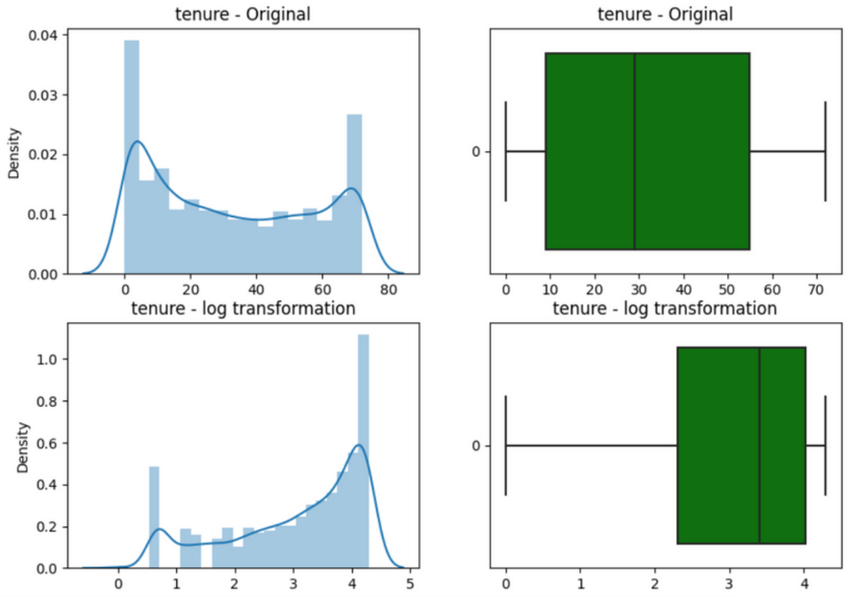

Tidak terdapat outlier pada boxplot tersebut karena dataset ini sudah baik (tidak terdapat kolom missing). Kemudian, nilai median dan maximalnya tidak terlalu jauh range perbedaannya meskipun kurang sesuai angka (numeriknya).

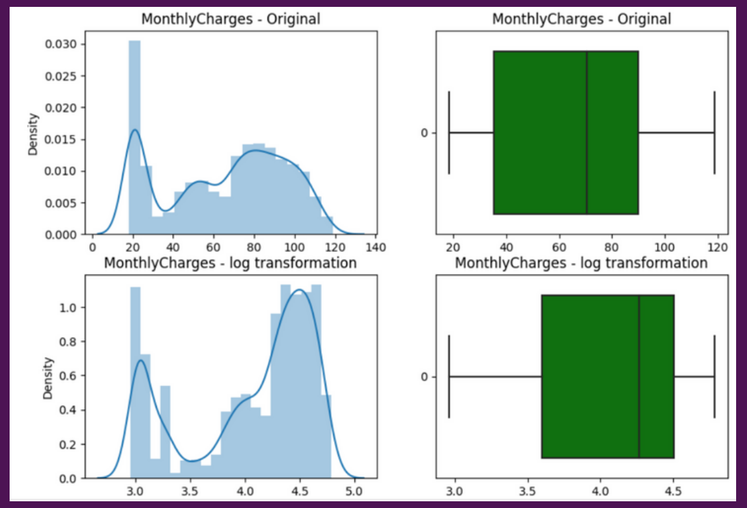
# LOG TRANSFORMATION

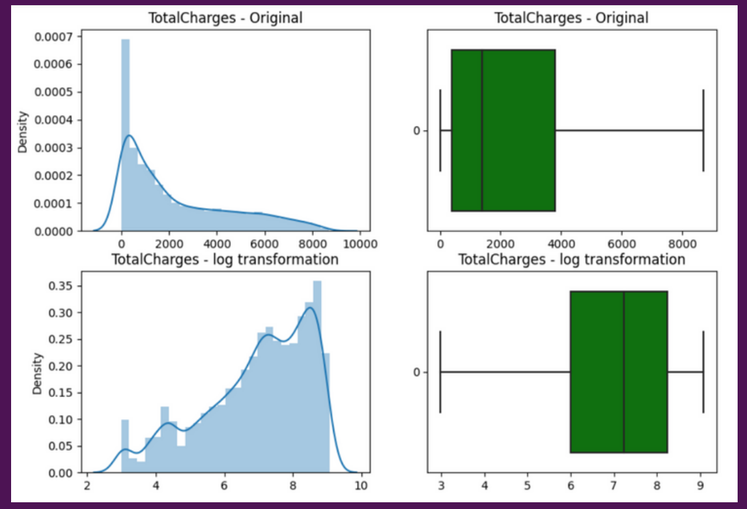
Kolom 'Tenure'

```
df['log_tenure'] = np.log(df['tenure']+1)
```

```
f,ax = plt.subplots(2,2,figsize=(10,7))

g = sns.distplot(df['tenure'],kde=True, ax=ax[0,0])
ax[0,0].set_title('tenure - Original')
ax[0,0].set_xlabel('')

g = sns.boxplot(df['tenure'],color='green',orient='h', ax=ax[0,1])
ax[0,1].set_title('tenure - Original')
ax[0,1].set_xlabel('')

g = sns.distplot(np.log(df['tenure']+1),kde=True, ax=ax[1,0])
ax[1,0].set_title('tenure - log transformation')
ax[1,0].set_xlabel('')

g = sns.boxplot(np.log(df['tenure']+1),color='green',orient='h', ax=ax[1,1])
ax[1,1].set_title('tenure - log transformation')
ax[1,1].set_xlabel('')
```




**Lakukan hal yang sama pada kolom MonthlyCharges dan TotalCharges**

# LINK GOOGLE COLLAB

**DATA CLEANSING**

https://colab.research.google.com/drive/1_ranKrkVVtfoMy5S8Dd7ANstHCXiuTfv

**TOPIC 6**

https://colab.research.google.com/drive/1Zf1_NGQNMPMCD91rAVnvOdYxj00I-dzc