Asesmen 5

DataFrame Basics and Data Cleansing

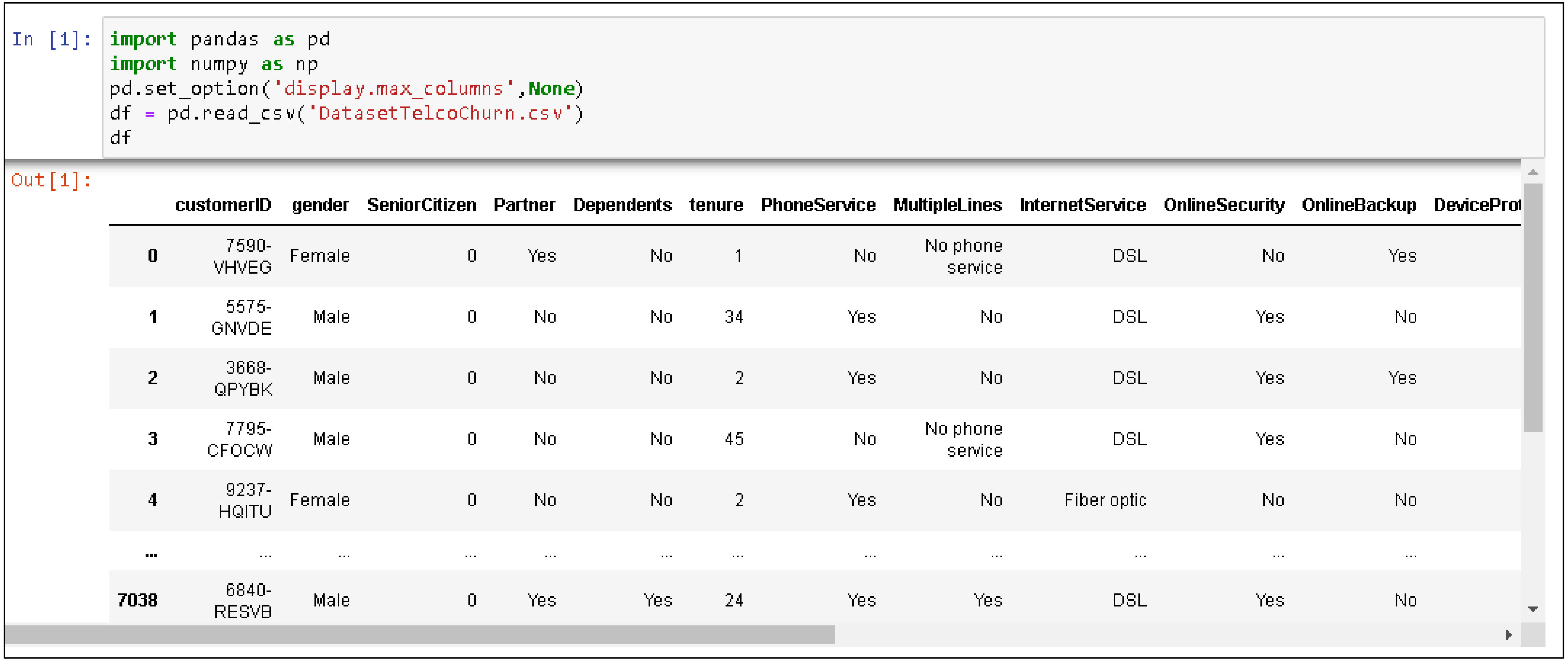
Shellya Nur Atqiya

https://drive.google.com/file/d/1osWDSIyOTVdxIaFpN5GebDLdGcsfLF4_/view?usp=sharing

DataFrame Basics and Data Cleansing

Missing Values | Categorical Data Encoding | Anomalies & Outlier

Import Dataset dan Library yang diperlukan



In [1]:

```
import pandas as pd
import numpy as np
pd.set_option('display.max_columns',None)
df = pd.read_csv('DatasetTelcoChurn.csv')
df
```

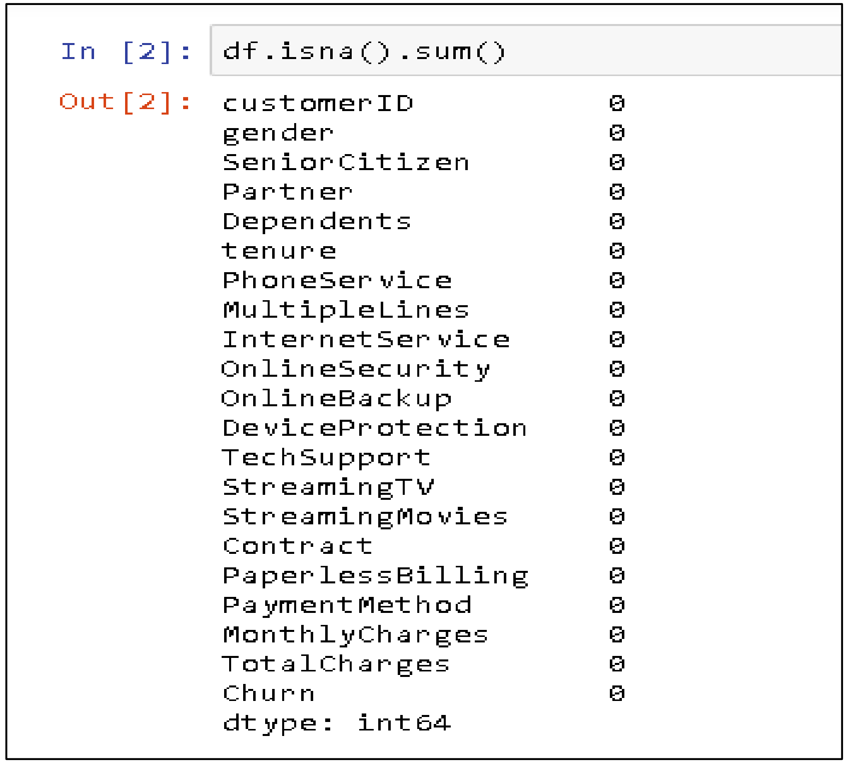
Out[1]:

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProt
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	Yes	
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	No	
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	Yes	
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	No	
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	No	
...	...	...	...	...	...	...	...	...	...	...	...	
7038	6840-RESVB	Male	0	Yes	Yes	24	Yes	Yes	DSL	Yes	No	

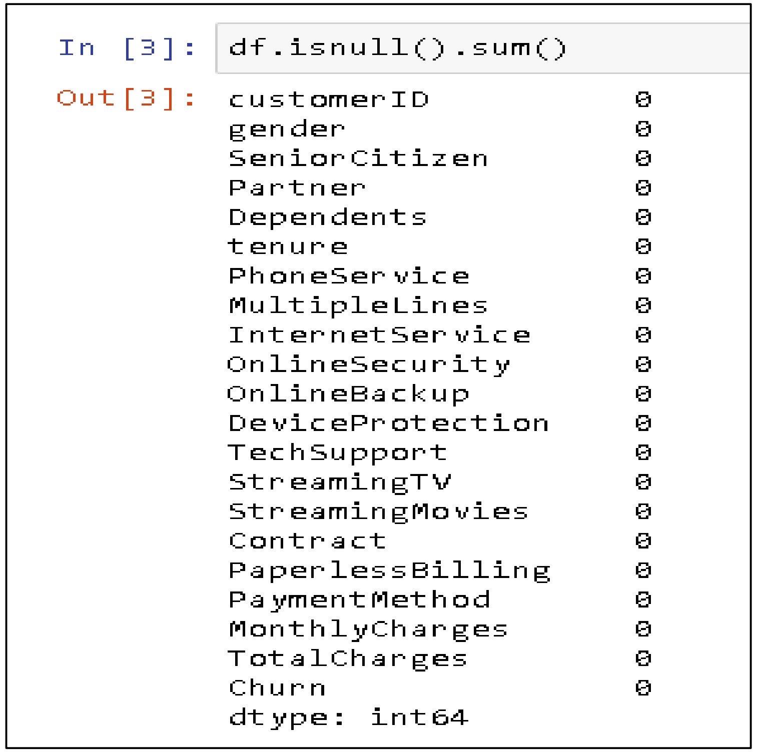
DataFrame Basics and Data Cleansing

Missing Values | Categorical Data Encoding | Anomalies & Outlier

Cek missing value dengan isna() dan isnull()
yaitu data yang Not Available dan Null



```
In [2]: df.isna().sum()

Out[2]: 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
```



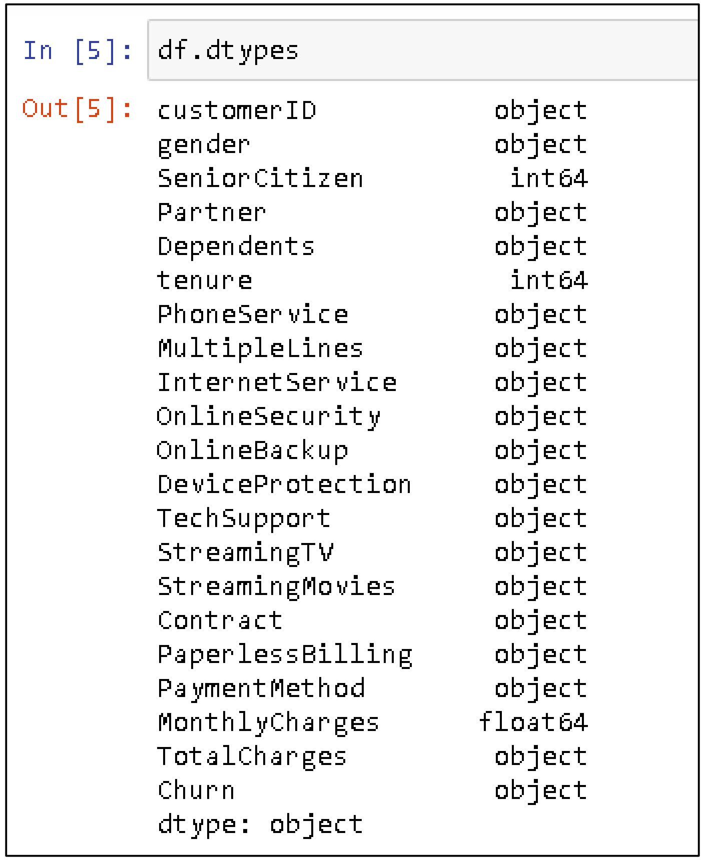
```
In [3]: df.isnull().sum()

Out[3]: 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
```

Berdasarkan output tersebut terlihat bahwa tidak ada kolom yang memiliki
missing value dalam bentuk NA dan Null

DataFrame Basics and Data Cleansing

Missing Values | Categorical Data Encoding | Anomalies & Outlier

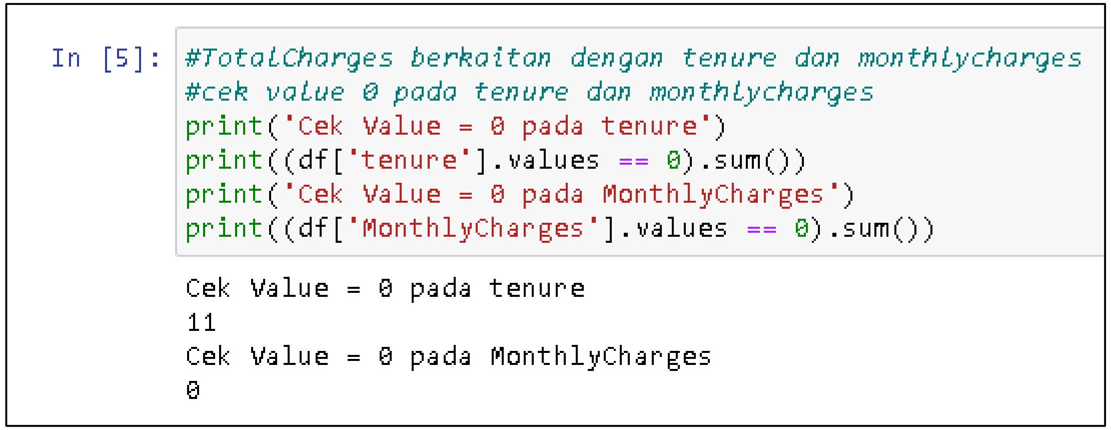
Cek tipe data tiap kolom dengan dtypes.



```
In [5]: df.dtypes

Out[5]: customerID          object
        gender              object
        SeniorCitizen        int64
        Partner             object
        Dependents          object
        tenure               int64
        PhoneService        object
        MultipleLines       object
        InternetService     object
        OnlineSecurity      object
        OnlineBackup        object
        DeviceProtection    object
        TechSupport         object
        StreamingTV         object
        StreamingMovies     object
        Contract            object
        PaperlessBilling    object
        PaymentMethod       object
        MonthlyCharges     float64
        TotalCharges        object
        Churn               object
        dtype: object
```

Ditemukan kejanggalan pada tipe data kolom TotalCharges yang seharusnya berupa float64 namun malah berupa object.

Crosscheck dengan kolom yang berkaitan dengan TotalCharges, yaitu tenure dan MonthlyCharges. Sebelumnya tidak ditemukan missing value NA dan Null, cek apakah terdapat value = 0.



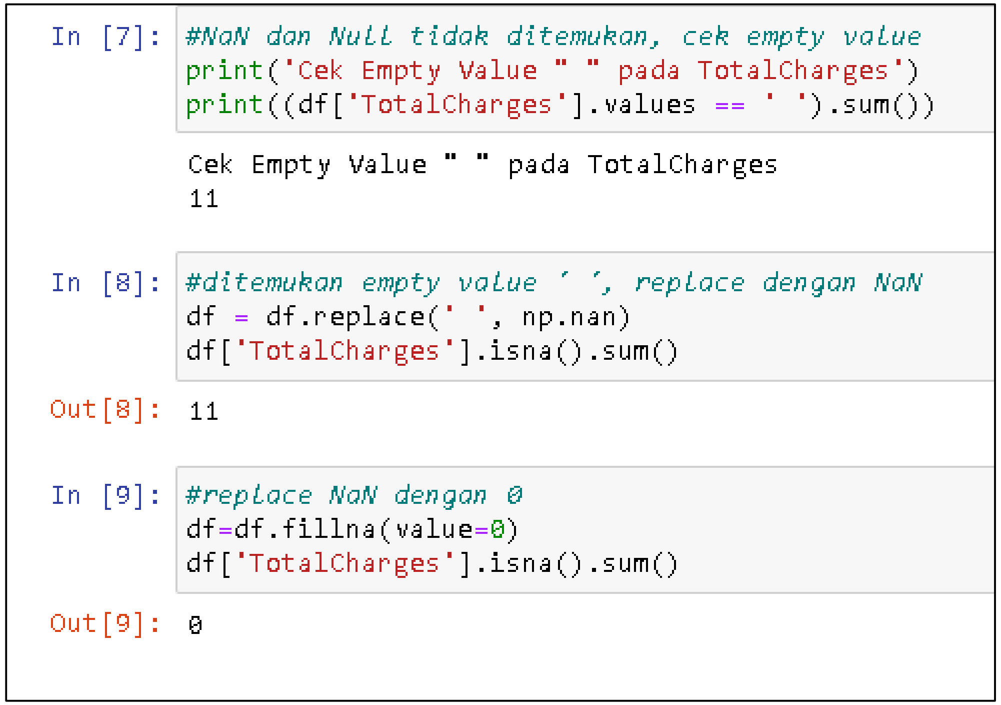
```
In [5]: #TotalCharges berkaitan dengan tenure dan monthlycharges
        #cek value 0 pada tenure dan monthlycharges
        print('Cek Value = 0 pada tenure')
        print((df['tenure'].values == 0).sum())
        print('Cek Value = 0 pada MonthlyCharges')
        print((df['MonthlyCharges'].values == 0).sum())

        Cek Value = 0 pada tenure
        11
        Cek Value = 0 pada MonthlyCharges
        0
```

Ditemukan missing values (value = 0) di kolom Tenure. Besar kemungkinan missing values pada kolom TotalCharges juga berjumlah 11

DataFrame Basics and Data Cleansing

Missing Values | Categorical Data Encoding | Anomalies & Outlier

Cek empty values “ “ pada kolom TotalCharges



```
In [7]: #NaN dan Null tidak ditemukan, cek empty value
        print('Cek Empty Value " " pada TotalCharges')
        print((df['TotalCharges'].values == ' ').sum())

        Cek Empty Value " " pada TotalCharges
        11

In [8]: #ditemukan empty value ' ', replace dengan NaN
        df = df.replace(' ', np.nan)
        df['TotalCharges'].isna().sum()

Out[8]: 11

In [9]: #replace NaN dengan 0
        df=df.fillna(value=0)
        df['TotalCharges'].isna().sum()

Out[9]: 0
```

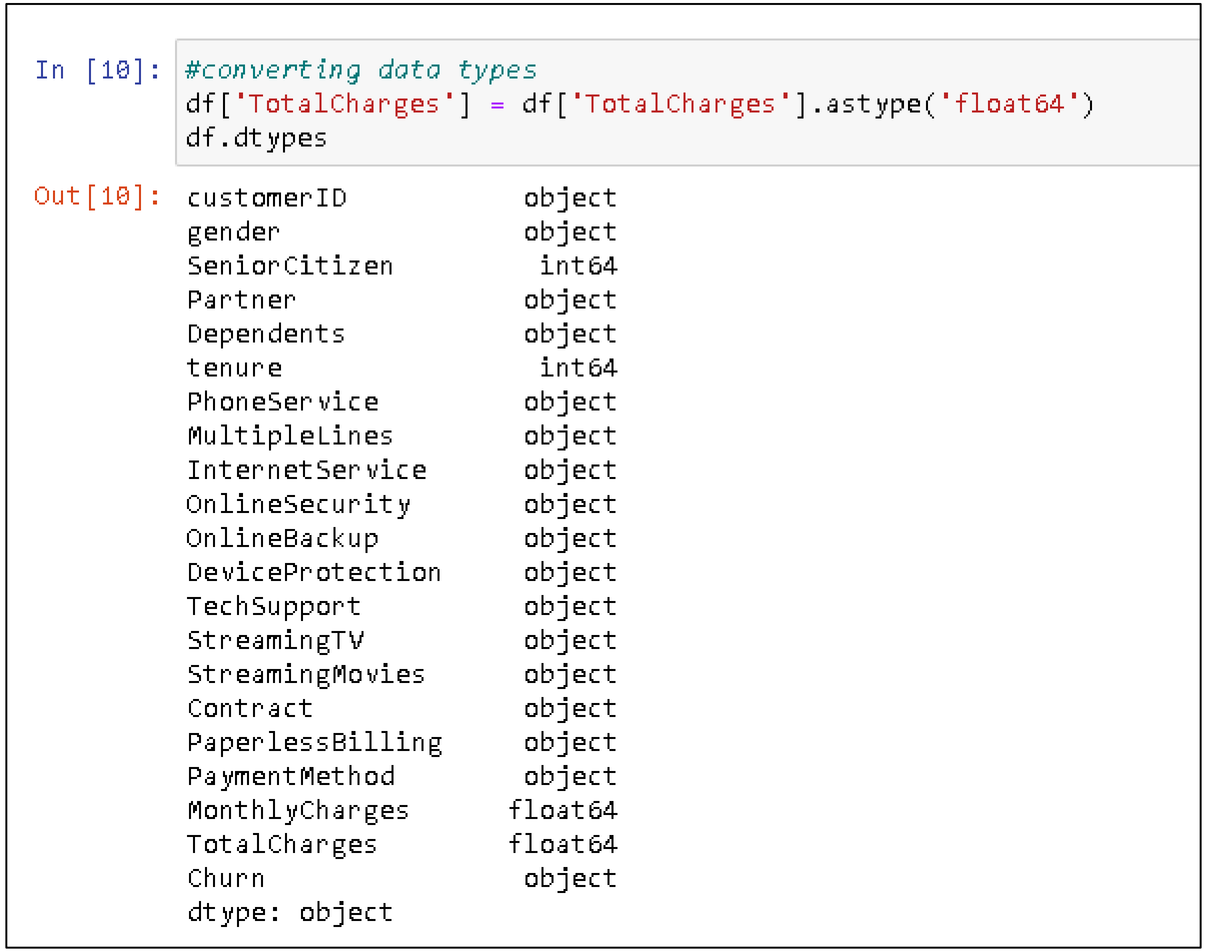
Ditemukan 11 empty values, ubah empty values tersebut menjadi NA dengan replace() dan fungsi numpy.nan

Ubah NA menjadi data bernilai 0 dengan fillna() sehingga tidak ada lagi missing value agar tipe data kolom TotalCharges dapat diubah menjadi float

DataFrame Basics and Data Cleansing

Missing Values | Categorical Data Encoding | Anomalies & Outlier

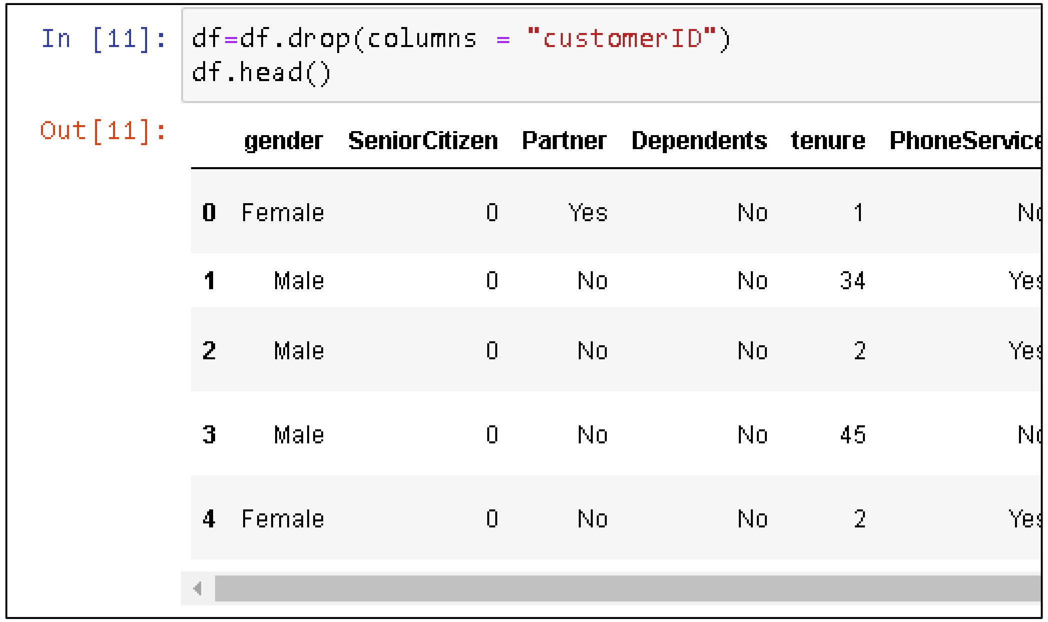
Ubah tipe data kolom TotalCharges menjadi tipe data float64 menggunakan astype()



```
In [10]: #converting data types
         df['TotalCharges'] = df['TotalCharges'].astype('float64')
         df.dtypes
```

```
Out[10]: customerID           object
         gender               object
         SeniorCitizen         int64
         Partner              object
         Dependents           object
         tenure                int64
         PhoneService         object
         MultipleLines        object
         InternetService      object
         OnlineSecurity       object
         OnlineBackup         object
         DeviceProtection     object
         TechSupport          object
         StreamingTV          object
         StreamingMovies      object
         Contract             object
         PaperlessBilling     object
         PaymentMethod        object
         MonthlyCharges      float64
         TotalCharges        float64
         Churn                object
         dtype: object
```

Hapus kolom customerID dengan drop()



```
In [11]: df=df.drop(columns = "customerID")
         df.head()
```

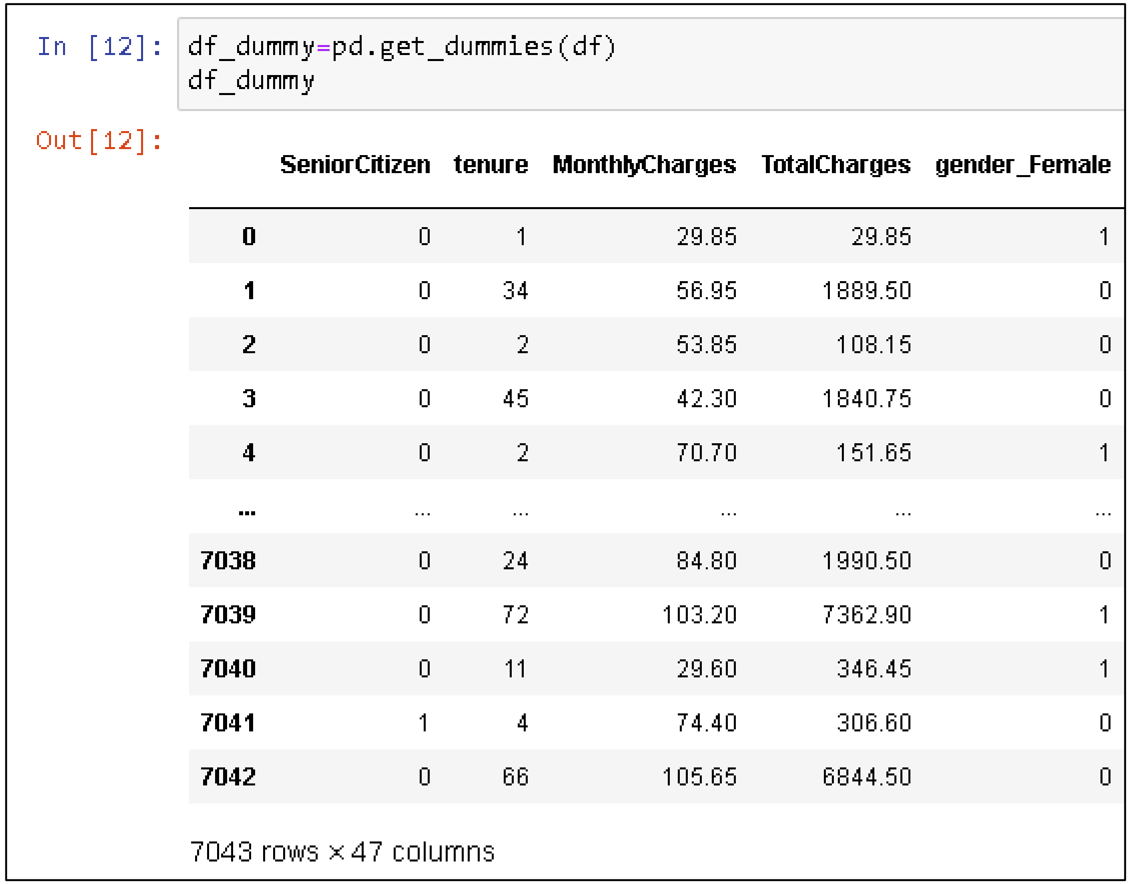
Out[11]:

	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService
0	Female	0	Yes	No	1	No
1	Male	0	No	No	34	Yes
2	Male	0	No	No	2	Yes
3	Male	0	No	No	45	No
4	Female	0	No	No	2	Yes

DataFrame Basics and Data Cleansing

Missing Values | Categorical Data Encoding | Anomalies & Outlier

Lihat Categorical Data Encoding dengan library pandas.get_dummies()



In [12]:
```
df_dummy=pd.get_dummies(df)
df_dummy
```

Out[12]:

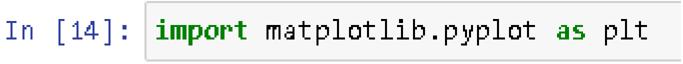
	SeniorCitizen	tenure	MonthlyCharges	TotalCharges	gender_Female
0	0	1	29.85	29.85	1
1	0	34	56.95	1889.50	0
2	0	2	53.85	108.15	0
3	0	45	42.30	1840.75	0
4	0	2	70.70	151.65	1
...	...	...	...	...	...
7038	0	24	84.80	1990.50	0
7039	0	72	103.20	7362.90	1
7040	0	11	29.60	346.45	1
7041	1	4	74.40	306.60	0
7042	0	66	105.65	6844.50	0

7043 rows × 47 columns

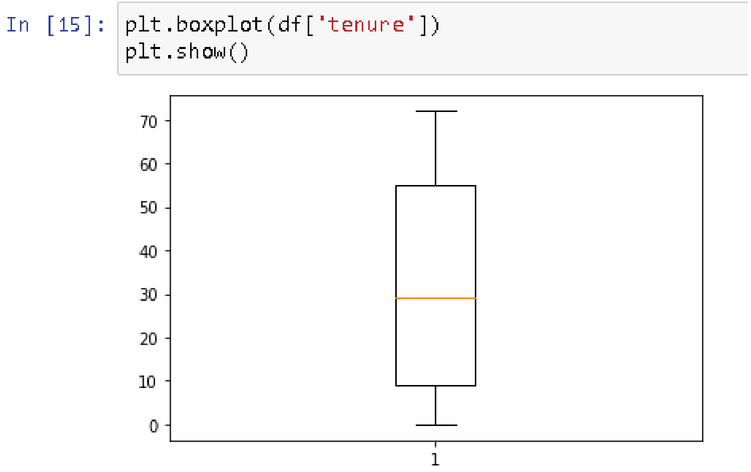
DataFrame Basics and Data Cleansing

Missing Values | Categorical Data Encoding | **Anomalies & Outlier**

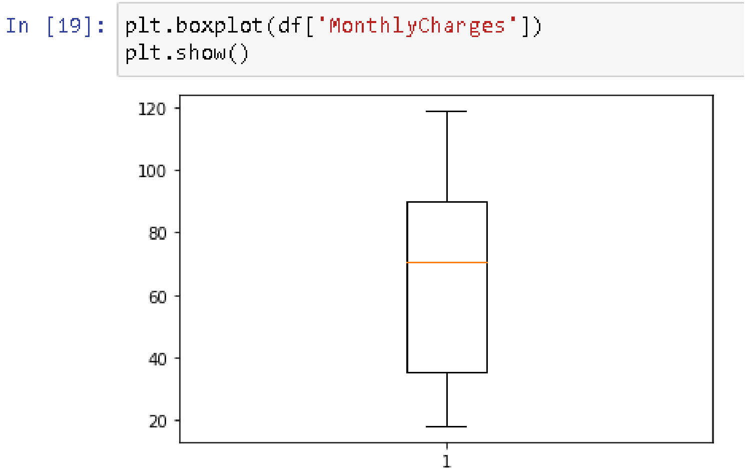
Import Library matplotlib.pyplot untuk membuat boxplot dengan plt.boxplot()



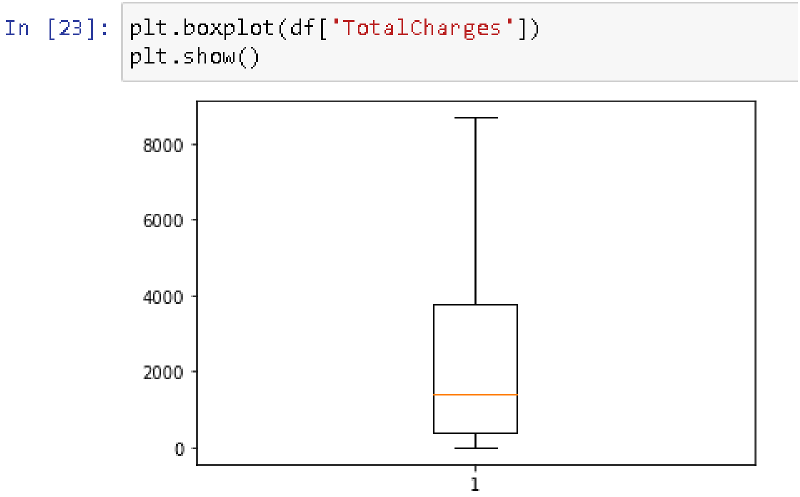
```
In [14]: import matplotlib.pyplot as plt
```



```
In [15]: plt.boxplot(df['tenure'])
         plt.show()
```



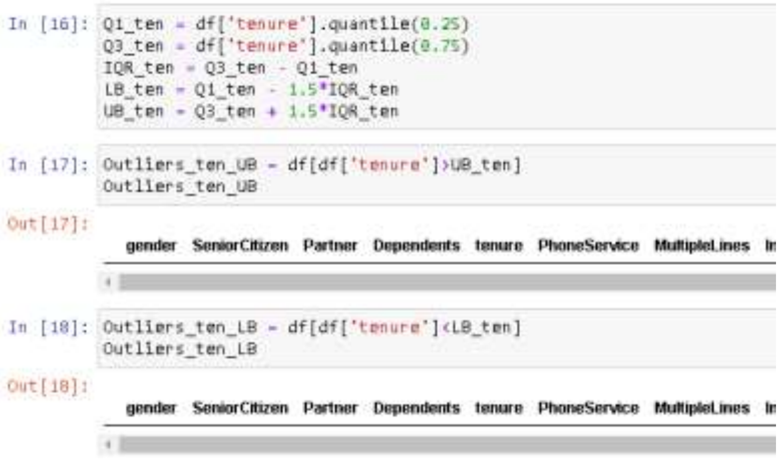
```
In [19]: plt.boxplot(df['MonthlyCharges'])
         plt.show()
```



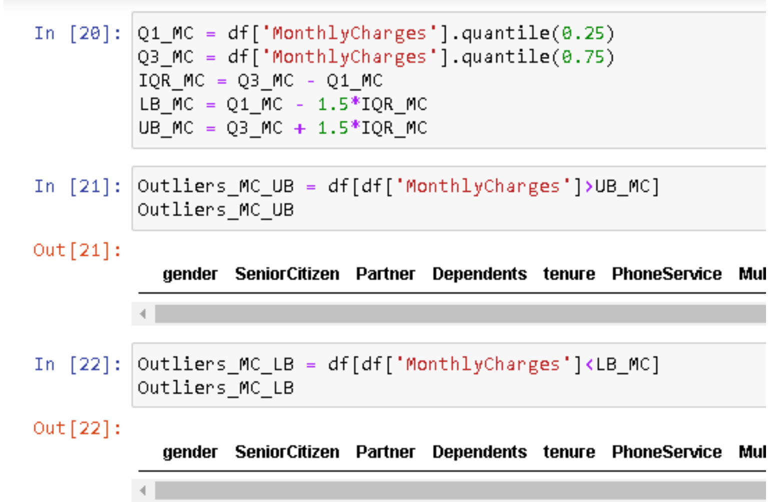
```
In [23]: plt.boxplot(df['TotalCharges'])
         plt.show()
```

DataFrame Basics and Data Cleansing

Missing Values | Categorical Data Encoding | **Anomalies & Outlier**

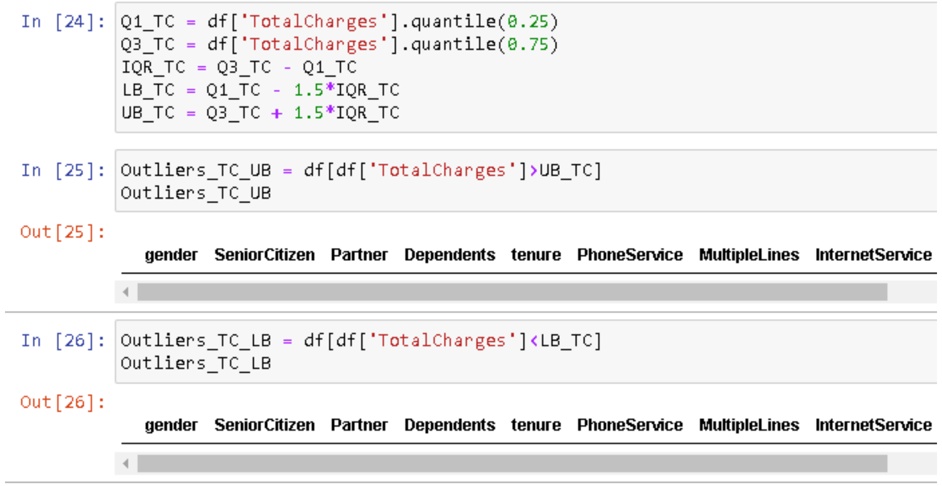


```
In [16]: Q1_ten = df['tenure'].quantile(0.25)
         Q3_ten = df['tenure'].quantile(0.75)
         IQR_ten = Q3_ten - Q1_ten
         LB_ten = Q1_ten - 1.5*IQR_ten
         UB_ten = Q3_ten + 1.5*IQR_ten

In [17]: Outliers_ten_UB = df[df['tenure']>UB_ten]
         Outliers_ten_UB

Out[17]:
           gender  SeniorCitizen  Partner  Dependents  tenure  PhoneService  MultipleLines  In

In [18]: Outliers_ten_LB = df[df['tenure']<LB_ten]
         Outliers_ten_LB

Out[18]:
           gender  SeniorCitizen  Partner  Dependents  tenure  PhoneService  MultipleLines  In
```



```
In [20]: Q1_MC = df['MonthlyCharges'].quantile(0.25)
         Q3_MC = df['MonthlyCharges'].quantile(0.75)
         IQR_MC = Q3_MC - Q1_MC
         LB_MC = Q1_MC - 1.5*IQR_MC
         UB_MC = Q3_MC + 1.5*IQR_MC

In [21]: Outliers_MC_UB = df[df['MonthlyCharges']>UB_MC]
         Outliers_MC_UB

Out[21]:
           gender  SeniorCitizen  Partner  Dependents  tenure  PhoneService  Mul

In [22]: Outliers_MC_LB = df[df['MonthlyCharges']<LB_MC]
         Outliers_MC_LB

Out[22]:
           gender  SeniorCitizen  Partner  Dependents  tenure  PhoneService  Mul
```

Periksa outlier



```
In [24]: Q1_TC = df['TotalCharges'].quantile(0.25)
         Q3_TC = df['TotalCharges'].quantile(0.75)
         IQR_TC = Q3_TC - Q1_TC
         LB_TC = Q1_TC - 1.5*IQR_TC
         UB_TC = Q3_TC + 1.5*IQR_TC

In [25]: Outliers_TC_UB = df[df['TotalCharges']>UB_TC]
         Outliers_TC_UB

Out[25]:
           gender  SeniorCitizen  Partner  Dependents  tenure  PhoneService  MultipleLines  InternetService

In [26]: Outliers_TC_LB = df[df['TotalCharges']<LB_TC]
         Outliers_TC_LB

Out[26]:
           gender  SeniorCitizen  Partner  Dependents  tenure  PhoneService  MultipleLines  InternetService
```

Tidak ditemukan adanya outlier, artinya tidak ada data yang perlu dihapus karena sudah berdistribusi normal.

Asesmen 5

DataFrame Basics and Data Cleansing

Data sudah bersih dari missing value dan juga outlier.

Terima kasih!