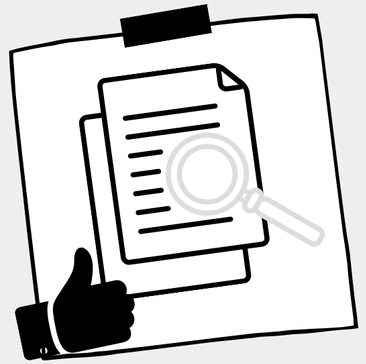
Erstyna Kathlya
Studi Independen Zenius

May 11,2022

Assignment Topic 5: Data Cleansing



Assignments

1. Missing Values Checking
2. Categorical Data Encoding
3. Anomalies and Outlier Handling

1. Missing Values Checking

Langkah 1: Mengecek Tipe Data

– Untuk mengecek apakah ada kolom yang tipe datanya belum sesuai
– Hasil: 1 kolom tipe datanya belum sesuai, yaitu kolom 'TotalCharges' yang seharusnya bertipe data float (decimal format)

Langkah 2: Mengubah Tipe Data

– Ubah tipe data 'TotalCharges' yang awalnya bertipe 'object' menjadi 'float64'. Pemilihan tipe data 'float64' dibandingkan 'int64' dilakukan untuk menghindari adanya pembulatan nilai di 'TotalCharges'. Kemudian, lakukan pengecekan apakah perubahan tipe data sudah berhasil dilakukan
– Hasil: Tipe data sudah sesuai

Hasil di Google Colab

Langkah 1

```
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
```

```
df['TotalCharges'] = pd.to_numeric
df.dtypes
```

```
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
```

Langkah 2

1. Missing Values Checking

Langkah 3

Mengecek Missing Values di Setiap Kolom

Hasil: terdapat 11 baris yang memiliki missing values pada kolom 'TotalCharges'

```
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       11
Churn               0
dtype: int64
```

Langkah 4

Mengecek Baris Mana Saja yang Memiliki Missing Values

Hasil: terdapat missing values pada baris indeks ke-488, 753, 936, 1082, 1340, 3331, 3826, 4380, 5218, 6670, dan 6754

File Edit View Insert Runtime Tools

Code + Text

78]

	936	5709-LVOEQ	Female
	1082	4367-NUYAO	Male
	1340	1371-DWPAZ	Female
	3331	7644-OMVMY	Male
	3826	3213-VVOLG	Male
	4380	2520-SGTTA	Female
	5218	2923-ARZLG	Male
	6670	4075-WKNIU	Female
	6754	2775-SEFEE	Male

11 rows × 21 columns

1. Missing Values Checking

Langkah 5

Ganti nilai yang hilang dengan nilai 0

Missing values terdeteksi pada baris yang memiliki nilai 'tenur' = 0.

Karena nilai 'TotalCharges' dihitung dari hasil perkalian 'tenur' dengan 'MonthlyCharges', maka nilai yang hilang diubah menjadi nilai 0

Setelah itu, beri nama / variabel baru untuk dataset yang missing value-nya sudah teratasi, di sini diberikan variabel yaitu **df_1.**

```
df_1 = df.fillna(value=0)
df_1
```

```
df_1[488:489]
```

Langkah 6

Lakukan pengecekan pada salah satu baris yang memiliki missing values

Pengecekan dilakukan untuk memastikan apakah nilai yang hilang (NaN) sudah berubah menjadi nilai 0

Diambil salah satu baris untuk dilakukan pengecekan, yaitu baris dengan indeks 488

Categorical Data Encoding 2

1 Menghapus Kolom 'customerID'

```
del df_1['customerID']
df_1
```

Kolom 'customerID' dihapus untuk menghindari hal yang tidak diinginkan pada saat proses encoding, misalkan penambahan kolom dengan jumlah yang banyak.
Hal ini disebabkan oleh data 'customerID' yang berbeda-beda, sehingga pada saat encoding, variabel dummy yang dihasilkan untuk kolom 'customerID' juga akan semakin banyak.
Hasil: tersisa 20 kolom

3	Male	0	No	No	45	No	No phone service	
4	Female	0	No	No	2	Yes	No	Fiber
...	...	...	...	...	...	...	...	
7038	Male	0	Yes	Yes	24	Yes	Yes	
7039	Female	0	Yes	Yes	72	Yes	Yes	Fiber
7040	Female	0	Yes	Yes	11	No	No phone service	
7041	Male	1	Yes	No	4	Yes	Yes	Fiber
7042	Male	0	No	No	66	Yes	No	Fiber

7043 rows × 20 columns

2 Membuat Variabel Dummy

Membuat variabel dummy merupakan salah satu proses dalam categorical data encoding. Dummy digunakan untuk data yang bertipe string / object
Dataset yang sudah dibuat variabel dummy diberikan dengan nama **df_dummy**
Hasil: dataset dengan 7043 baris dan 47 kolom

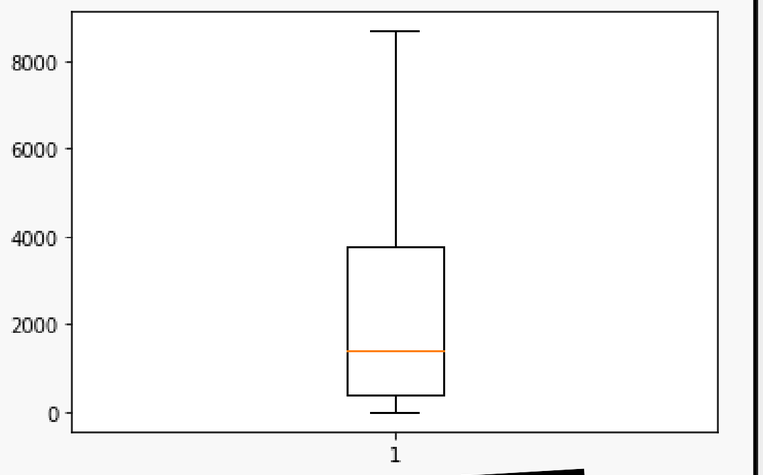
```
df_dummy = pd.get_dummies(df_1)
df_dummy
```

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

7043 rows × 47 columns

Anomalies and Outliers

Membuat Boxplot




```
import matplotlib.pyplot as plt
plt.boxplot(df_dummy['TotalCharges'])
plt.show()
```

Menghitung Outlier

```
Q1 = df_dummy['TotalCharges'].quantile(0.25)
Q3 = df_dummy['TotalCharges'].quantile(0.75)
IQR = Q3 - Q1
lower_bound = Q1 - 1.5*IQR
upper_bound = Q3 + 1.5*IQR

outliers = df_dummy[df_dummy['TotalCharges']>upper_bound]
outliers
```

SeniorCitizen	tenure	MonthlyCharges	TotalCharges

0 rows × 47 columns

Dari sini, dapat disimpulkan bahwa dataframe tidak memiliki outlier

Tidak ditemukan anomali pada dataframe ini karena tidak ditemukan kesalahan dalam input data maupun kesalahan transformasi data

Erstyna Kathlya
Studi Independen Zenius

May 11, 2022

Terima kasih!