

# pipeline

August 1, 2025

```
[1]: import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
# from sklearn.preprocessing import LabelEncoder
from preprocessing_functions import clean_categorical_title,
↪ update_missing_textual_values
```

```
[2]: popis = pd.read_excel("../..../datasets/Allianz2_kolone.xlsx")
```

```
[3]: popis.head(10)
```

```
[3]:
```

|   | City   | GeNder_oF_CustoMer | Age  | InCoMe  | ILLness |
|---|--------|--------------------|------|---------|---------|
| 0 | Dallas | Male               | 41.0 | 40367.0 | No      |
| 1 | Dallas | Male               | 54.0 | 45084.0 | No      |
| 2 | Dallas | male               | 42.0 | 52483.0 | No      |
| 3 | Dallas | Male               | 40.0 | 40941.0 | No      |
| 4 | Dallas | male               | 46.0 | 50289.0 | No      |
| 5 | Dallas | Female             | 36.0 | 50786.0 | No      |
| 6 | Dallas | Female             | 32.0 | 33155.0 | No      |
| 7 | Dallas | Male               | 39.0 | 30914.0 | No      |
| 8 | Dallas | Male               | 51.0 | 68667.0 | No      |
| 9 | Dallas | Female             | 30.0 | 50082.0 | No      |

```
[4]: popis = clean_categorical_title(popis, 'GeNder_oF_CustoMer')
```

```
[5]: popis.head(15)
```

```
[5]:
```

|   | City   | GeNder_oF_CustoMer | Age  | InCoMe  | ILLness |
|---|--------|--------------------|------|---------|---------|
| 0 | Dallas | Male               | 41.0 | 40367.0 | No      |
| 1 | Dallas | Male               | 54.0 | 45084.0 | No      |
| 2 | Dallas | Male               | 42.0 | 52483.0 | No      |
| 3 | Dallas | Male               | 40.0 | 40941.0 | No      |
| 4 | Dallas | Male               | 46.0 | 50289.0 | No      |
| 5 | Dallas | Female             | 36.0 | 50786.0 | No      |
| 6 | Dallas | Female             | 32.0 | 33155.0 | No      |
| 7 | Dallas | Male               | 39.0 | 30914.0 | No      |
| 8 | Dallas | Male               | 51.0 | 68667.0 | No      |
| 9 | Dallas | Female             | 30.0 | 50082.0 | No      |

|    |        |        |      |         |     |
|----|--------|--------|------|---------|-----|
| 10 | Dallas | Female | 48.0 | 41524.0 | Yes |
| 11 | Dallas | Male   | 47.0 | 54777.0 | No  |
| 12 | Dallas | Male   | 46.0 | 62749.0 | No  |
| 13 | Dallas | Female | 42.0 | 50894.0 | No  |
| 14 | Dallas | Female | 61.0 | 38429.0 | No  |

```
[6]: update_missing_textual_values(popis)
```

Najčešće korištena vrijednost u City je New York City  
Najčešće korištena vrijednost u GeNdeR\_oF\_CustoMer je Male  
Najčešće korištena vrijednost u ILLness je No