EXPLORATORY DATA ANALYSIS FOR MACHINE LEARNING

IBM Machine Learning - Project 1
Zidan Qurosey Sabilla

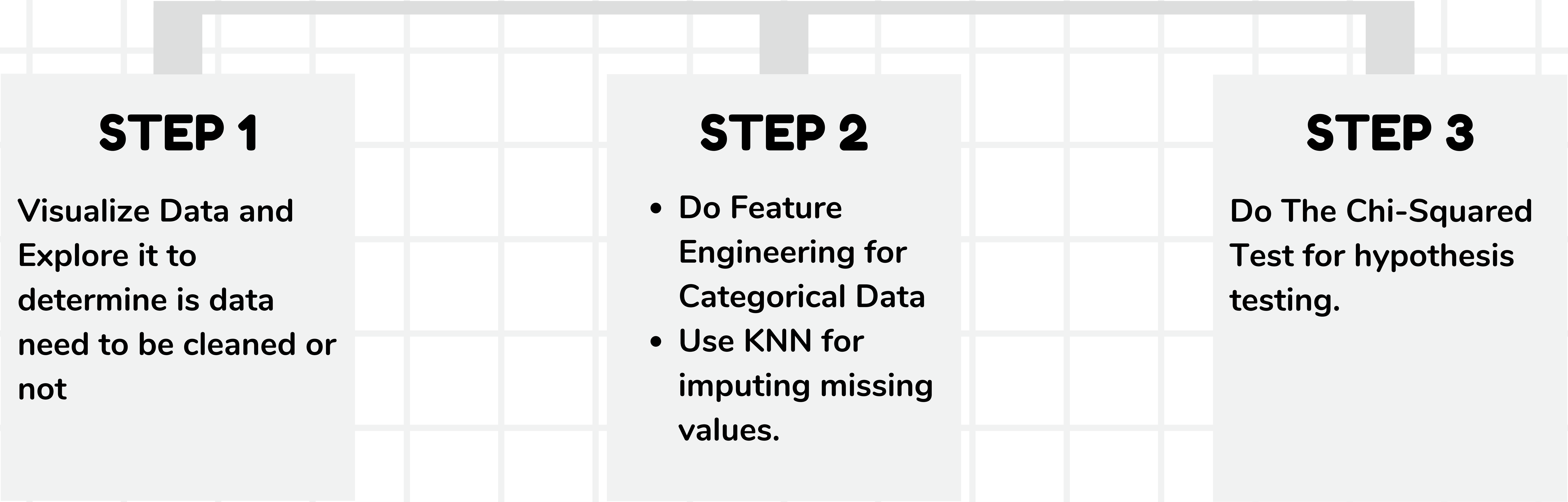
ABOUT THE DATA

- An ongoing outbreak of monkeypox, a viral disease, was confirmed in May 2022. The initial cluster of cases was found in the United Kingdom, where the first case was detected in London on 6 May 2022 in a patient with a recent travel history from Nigeria.
- This is a **SYNTHETIC** dataset generated based on a study published by thebmj: Clinical features and novel presentations of human monkeypox in a central London centre during the 2022 outbreak: descriptive case series.
- Dataset consists of a CSV which have a record of **25,000 Patients** with their corresponding features and a target variable indicating if the patient has monkeypox or not.
- Dataset contain 11 columns.

DATA DICTIONARY

Variable	Type	Description
Systemic Illness	Nominal	Type of illness
Rectal Pain	Boolean	Do they have Rectal Pain
Sore Throat	Boolean	Do they have Sore Throat
Penile Oedema	Boolean	Do they have Penile Oedema
Sexually Transmitted Infection	Boolean	Do they have any sexually transmitted infection
Oral Lesions	Boolean	Do they have Oral Lesions
Solitary Lesion	Boolean	Do they have Solitary Lesion
Swollen Tonsils	Boolean	Do they have Swollen Tonsils
HIV Infection	Boolean	Do they have HIV Infection

STRATEGY



STEP 1

Visualize Data and Explore it to determine is data need to be cleaned or not

STEP 2

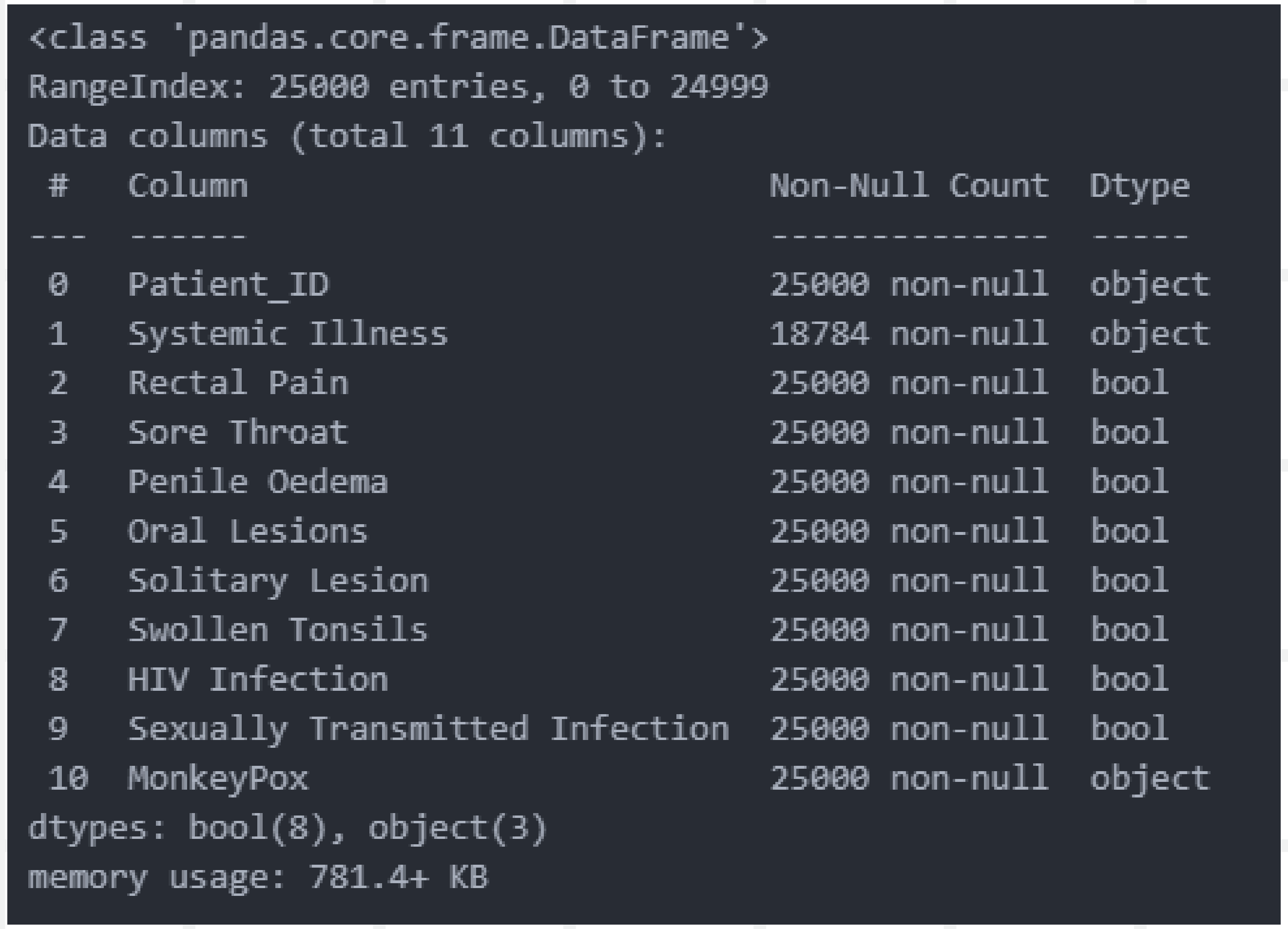
- Do Feature Engineering for Categorical Data
- Use KNN for imputing missing values.

STEP 3

Do The Chi-Squared Test for hypothesis testing.

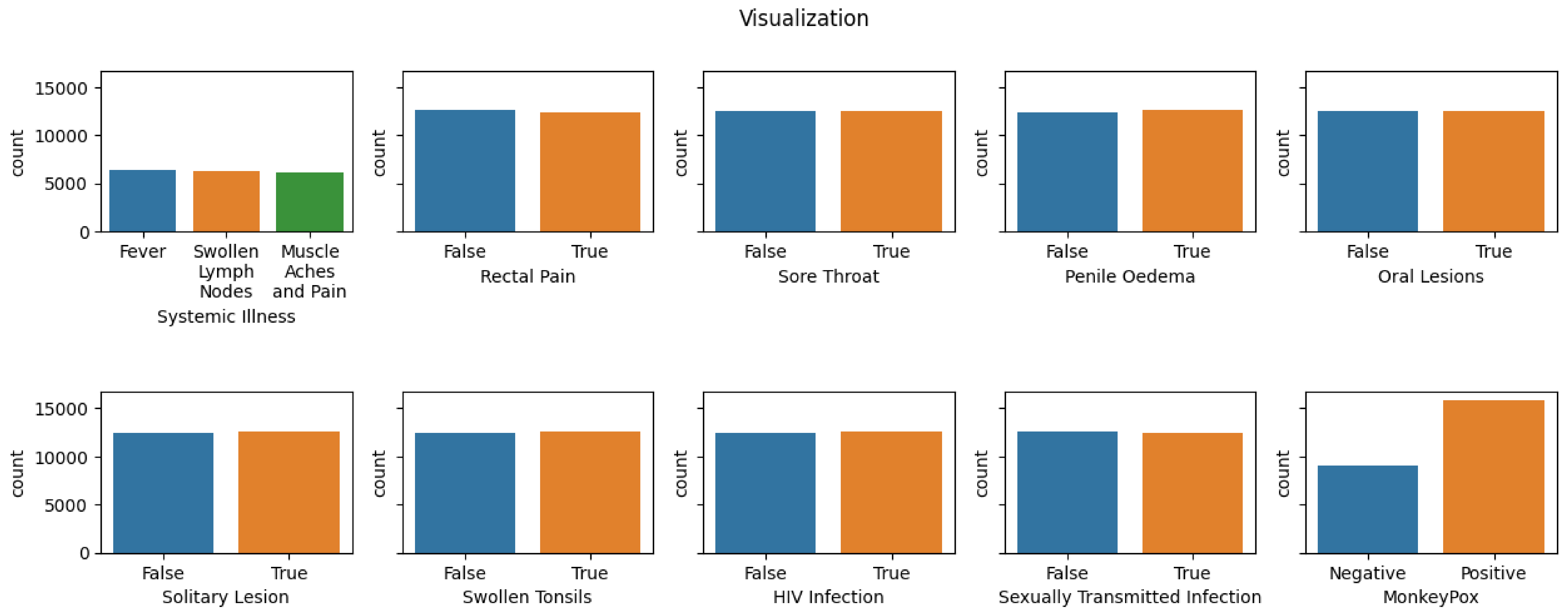
EXPLORATORY DATA

- There is only 1 variable that has missing values



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25000 entries, 0 to 24999
Data columns (total 11 columns):
 #   Column                           Non-Null Count  Dtype 
---  ------                           --------------  ----- 
 0   Patient_ID                       25000 non-null  object
 1   Systemic Illness                 18784 non-null  object
 2   Rectal Pain                      25000 non-null  bool  
 3   Sore Throat                      25000 non-null  bool  
 4   Penile Oedema                    25000 non-null  bool  
 5   Oral Lesions                     25000 non-null  bool  
 6   Solitary Lesion                  25000 non-null  bool  
 7   Swollen Tonsils                  25000 non-null  bool  
 8   HIV Infection                    25000 non-null  bool  
 9   Sexually Transmitted Infection  25000 non-null  bool  
 10  MonkeyPox                        25000 non-null  object
dtypes: bool(8), object(3)
memory usage: 781.4+ KB
```

EXPLORATORY DATA



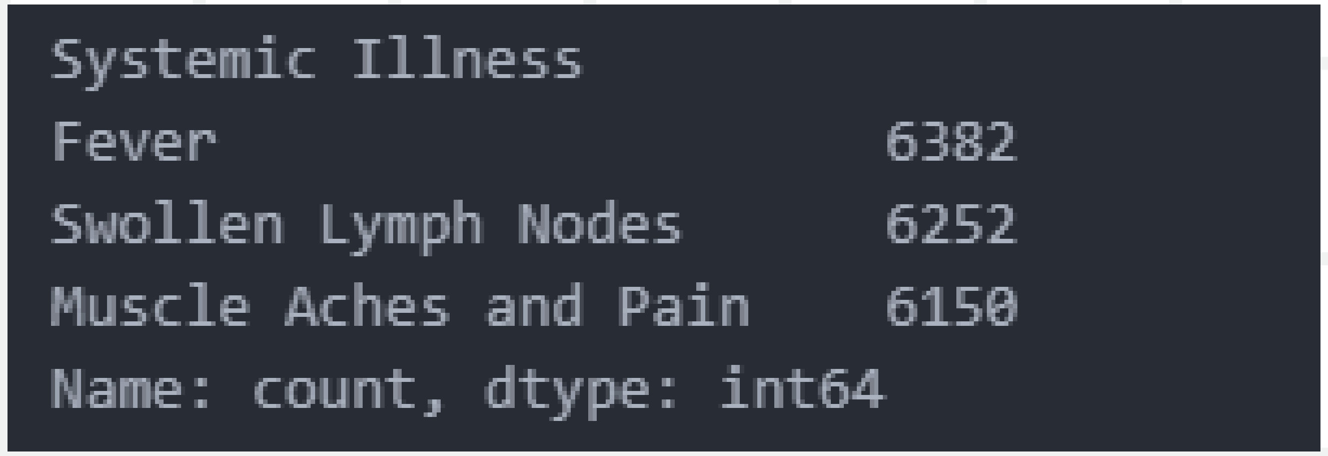
- It seems that MonkeyPox is imbalanced

FEATURE ENGINEERING

1. **Handling Missing Values**

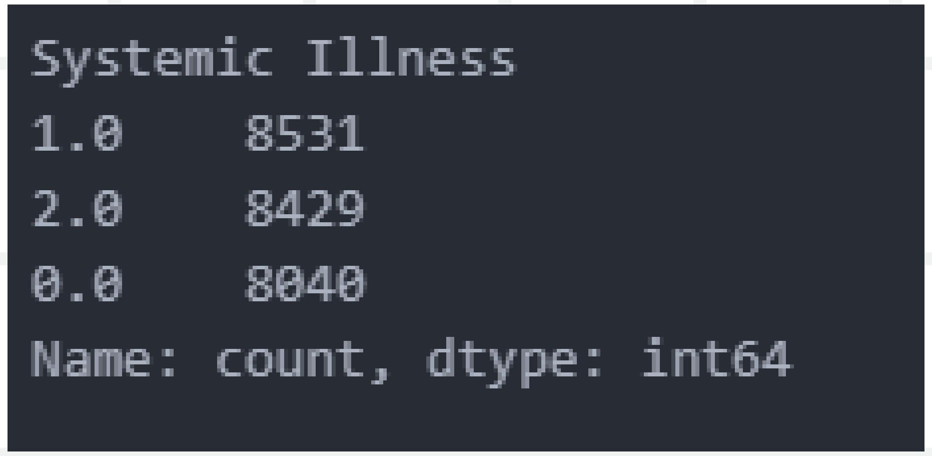
To handled missing values, Systemic Illnes, Will be using KKN for the imputer.

Before:



```
Systemic Illness
Fever                     6382
Swollen Lymph Nodes       6252
Muscle Aches and Pain     6150
Name: count, dtype: int64
```

After:



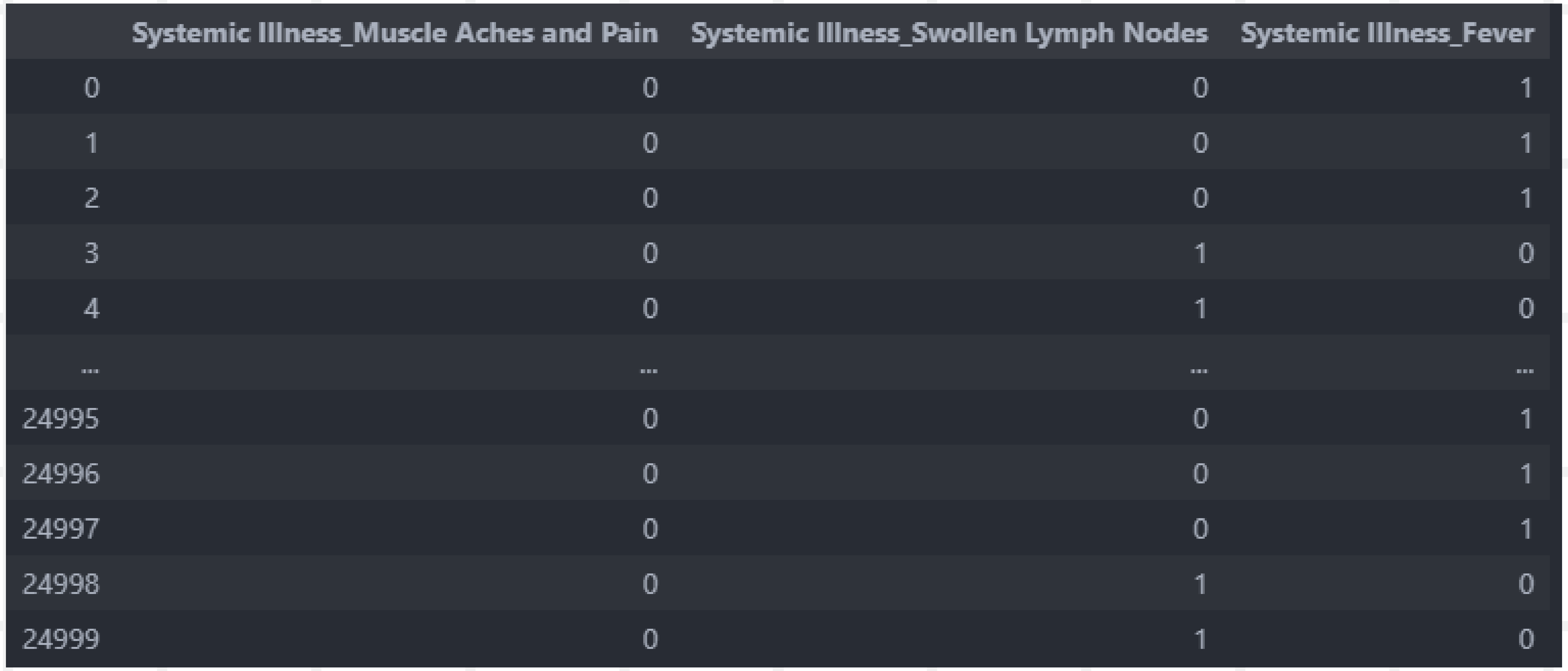
```
Systemic Illness
1.0    8531
2.0    8429
0.0    8040
Name: count, dtype: int64
```

0 : Muscle Aches and Pain
1 : Swollen Lymph Nodes
2 : Fever

FEATURE ENGINEERING

2. Dummies Variable or One Hot Encoding

For the variable Systemic Illness, Will be using pd.get_dummies for making dummies variable.

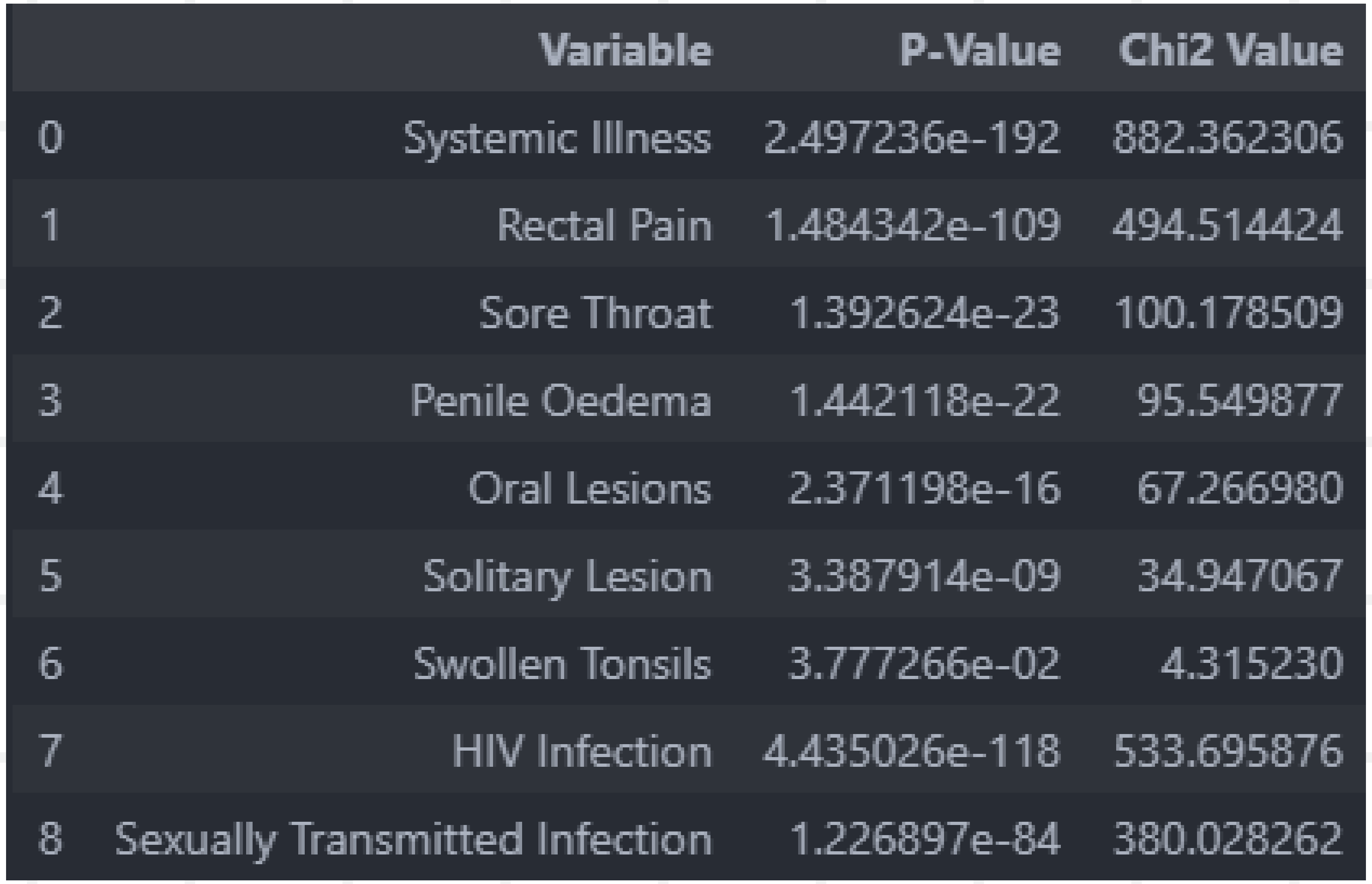


	Systemic Illness_Muscle Aches and Pain	Systemic Illness_Swollen Lymph Nodes	Systemic Illness_Fever
0	0	0	1
1	0	0	1
2	0	0	1
3	0	1	0
4	0	1	0
...	...	...	...
24995	0	0	1
24996	0	0	1
24997	0	0	1
24998	0	1	0
24999	0	1	0

HYPOTHESIS TESTING

- The Chi-Squared Test is chosen for hypothesis testing to determine if there is a significant association between an independent categorical variable and a dependent categorical variable.
- It evaluates whether observed data distribution deviates significantly from the expected distribution, assuming independence.
- The goal is to identify a statistically significant relationship between the categorical variables.
- Sample of hypothesis:
    - H_0 : There is no correlation between variable X and variable.
    - H_1 : There is correlation between variable X and variable Y

HYPOTHESIS TESTING



	Variable	P-Value	Chi2 Value
0	Systemic Illness	2.497236e-192	882.362306
1	Rectal Pain	1.484342e-109	494.514424
2	Sore Throat	1.392624e-23	100.178509
3	Penile Oedema	1.442118e-22	95.549877
4	Oral Lesions	2.371198e-16	67.266980
5	Solitary Lesion	3.387914e-09	34.947067
6	Swollen Tonsils	3.777266e-02	4.315230
7	HIV Infection	4.435026e-118	533.695876
8	Sexually Transmitted Infection	1.226897e-84	380.028262

- It seems that all variable independent (X) is rejected the H_0, so there is correlation between variable X and variable Y.
- I suggest for do Logistics Regression Analysis, Logistic regression is a data analysis technique that uses mathematics to uncover the relationship between two data factors.

CONCLUSION

As shown in analysis, logistics regression will be a good choice for this dataset to assess the extent of the influence of independent variables on the dependent variable.

Jupyter Notebook for this analysis can be found here : https://github.com/zidanqrs/IBM-Machine-Learning-Course/blob/main/1-Exploratory-Data-Analysis-for-ML/Project-1.ipynb