Data Science | Digital Skill Fair 39.0

MEDICAL DATA ANALYTICS

BY : MARIA PUTRI FREDELLA GULTOM | DATA SCIENCE

HI!

I’m Maria

Being a data enthusiast, I'm in the process of an analysis of medical patient data.

The goal is to simplify how we approach treatment plans down the line. For this project, I'm employing **Exploratory Data Analysis (EDA)** to reshape the data so its organization is easier to grasp, to locate any present challenges, and to create visual representations of the findings.

ibimbing

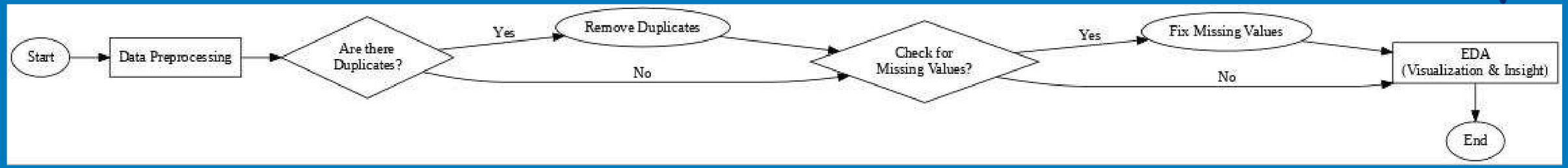
Data Science | Digital Skill Fair 39.0

Project Goals

to analyze the medical patients data for identify potential data quality issues requiring preprocessing, and to gain initial insights into the characteristics of the patient population.

Learn More

Flowchart


Start
Data Preprocessing
Are there Duplicates?
Yes
Remove Duplicates
No
Check for Missing Values?
Yes
Fix Missing Values
No
EDA (Visualization & Insight)
End

DATA OVERVIEW

	id	full_name	age	gender	smoking_status	glucose_levels	condition
0	1	User0001	21.0	male	0	1.234568e+15	Pneumonia
1	2	User0002	30.0	male	0	7.234467e+15	Diabetic
2	3	User0003	18.0	male	0	9.876543e+15	Pneumonia
3	4	User0004	20.0	male	0	4.890123e+15	Pneumonia
4	5	User0005	76.0	male	0	1.321099e+15	Diabetic

	id	full_name	age	gender	smoking_status	glucose_levels	condition
2015	2016	User2146	21.0	male	0	8.901235e+15	Diabetic
2016	2017	User2147	33.0	male	0	2.765432e+15	Pneumonia
2017	2018	User2148	20.0	male	0	9.901235e+15	Diabetic
2018	2019	User2149	47.0	male	0	5.543211e+15	Diabetic
2019	2020	User2150	19.0	male	0	2.109877e+15	Diabetic

The dataset consists of 2020 rows and 7 columns, each rows represents a unique medical patient. Then, the columns include **id, full name, age, gender, smoking status, glucose levels, and condition.** This structure format allow us to analyze demographic characteristics of the patient population.

ABOUT DATASET

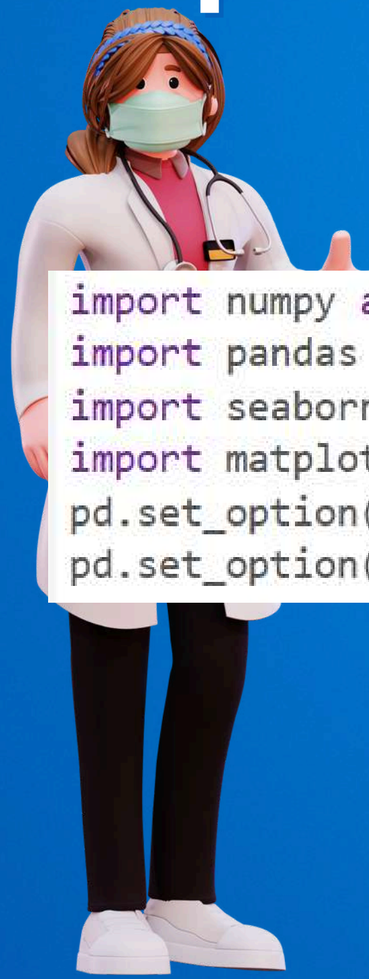
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2020 entries, 0 to 2019
Data columns (total 7 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   id              2020 non-null   int64
 1   full_name       2020 non-null   object
 2   age             1639 non-null   float64
 3   gender          2020 non-null   object
 4   smoking_status  2020 non-null   int64
 5   glucose_levels  1519 non-null   float64
 6   condition       2020 non-null   object
dtypes: float64(2), int64(2), object(3)
memory usage: 110.6+ KB
```

1. **Total Data** :
   - Dataset contains 2020 entries (rows) with index 0 until 2019.
2. **Column and Type of Data** :
   - Contains 7 columns :
     - id: 2020 non-null, type int64.
     - full name: 2020 non-null, type object (string/text).
     - age: 1639 non-null, type float64 (having missing values).
     - gender: 2020 non-null, type object (string/text).
     - smoking status: 2020 non-null, type int64.
     - glucose levels: 1519 non-null, type float64 (having missing values).
     - condition: 2020 non-null, type object (string/text).
3. **Missing Values** :
   - The age column (381 missing entries) and glucose levels column (501 missing entries) from 2020 entries data.
4. **Memory Usage** :
   - Dataset having almost 110.6+ KB.

Data Science | Digital Skill Fair 39.0

DATA PREPROCESSING

Import Libraries



```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
pd.set_option("display.max_columns", None)
pd.set_option("display.max_rows", None)
```

- numpy is used for numerical computations, especially arrays and matrices.
- pandas is used for provides data structures and functions for data manipulation and analysis.
- seaborn is a statical data visualization library to generate complex with less code.
- matplotlib.pyplot is used for creating static, animated, and interactive plots.
- pd.set_option() is configuring Pandas display settings, used to show all columns and rows of a DataFrame.

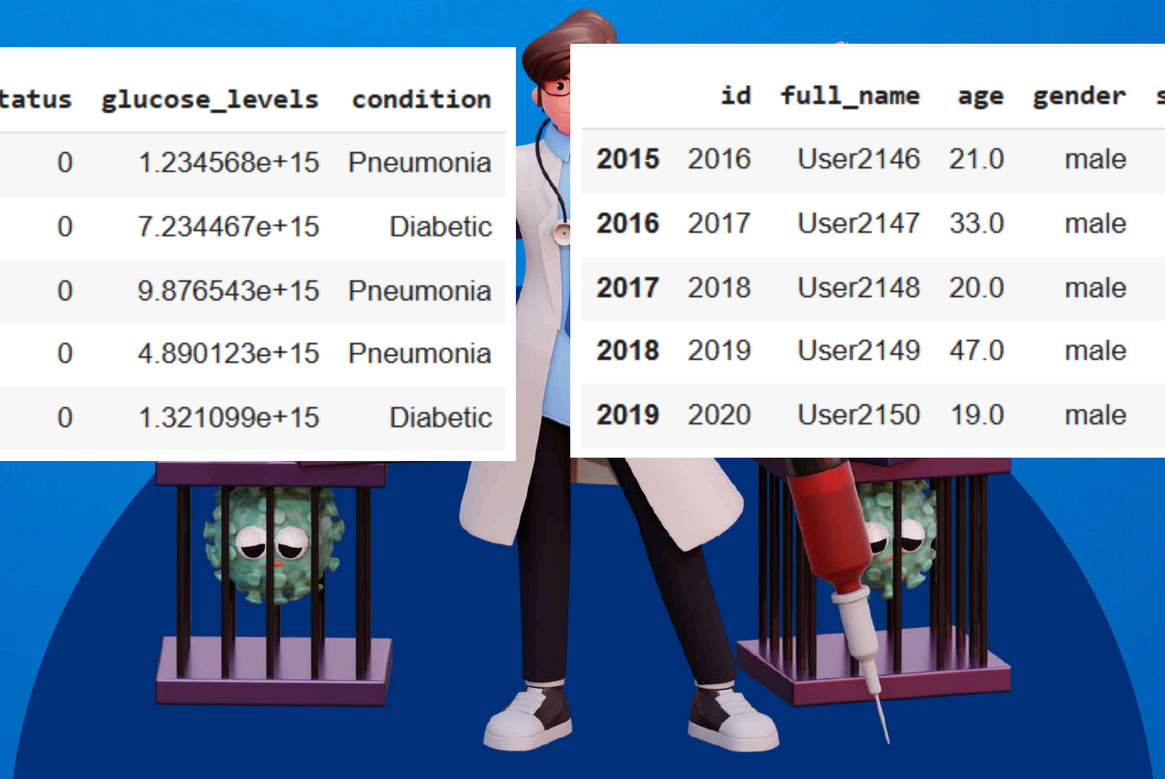
Top 5 rows of data

data.head()
show the first 5 rows of the dataset

	id	full_name	age	gender	smoking_status	glucose_levels	condition
0	1	User0001	21.0	male	0	1.234568e+15	Pneumonia
1	2	User0002	30.0	male	0	7.234467e+15	Diabetic
2	3	User0003	18.0	male	0	9.876543e+15	Pneumonia
3	4	User0004	20.0	male	0	4.890123e+15	Pneumonia
4	5	User0005	76.0	male	0	1.321099e+15	Diabetic

data.tail()
show the last 5 rows of the dataset

	id	full_name	age	gender	smoking_status	glucose_levels	condition
2015	2016	User2146	21.0	male	0	8.901235e+15	Diabetic
2016	2017	User2147	33.0	male	0	2.765432e+15	Pneumonia
2017	2018	User2148	20.0	male	0	9.901235e+15	Diabetic
2018	2019	User2149	47.0	male	0	5.543211e+15	Diabetic
2019	2020	User2150	19.0	male	0	2.109877e+15	Diabetic



Information about the Medical Data

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2020 entries, 0 to 2019
Data columns (total 7 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   id               2020 non-null   int64  
 1   full_name        2020 non-null   object 
 2   age              1639 non-null   float64
 3   gender           2020 non-null   object 
 4   smoking_status   2020 non-null   int64  
 5   glucose_levels   1519 non-null   float64
 6   condition        2020 non-null   object 
dtypes: float64(2), int64(2), object(3)
memory usage: 110.6+ KB
```

The following information to observed :

1. Data contains 7 columns.
2. 2020 data entries.
3. The age column (381 missing entries) and glucose levels column (501 missing entries) from 2020 entries data.

data.info()

Data Science | Digital Skill Fair 39.0

DATA CLEANING

Checking Duplicate

```
In [22]:
 len(data)

Out[22]:
2020

In [23]:
 len(data.drop_duplicates())

Out[23]:
2020

In [24]:
 len(data.drop_duplicates()) / len(data)

Out[24]:
1.0
```

The dataset contains 2020 in total

This indicates that there is not a duplicate row in the dataset

If the value of ratio is <1, it means the presence of duplicate data within the dataset

Missing value

data.isna().sum()

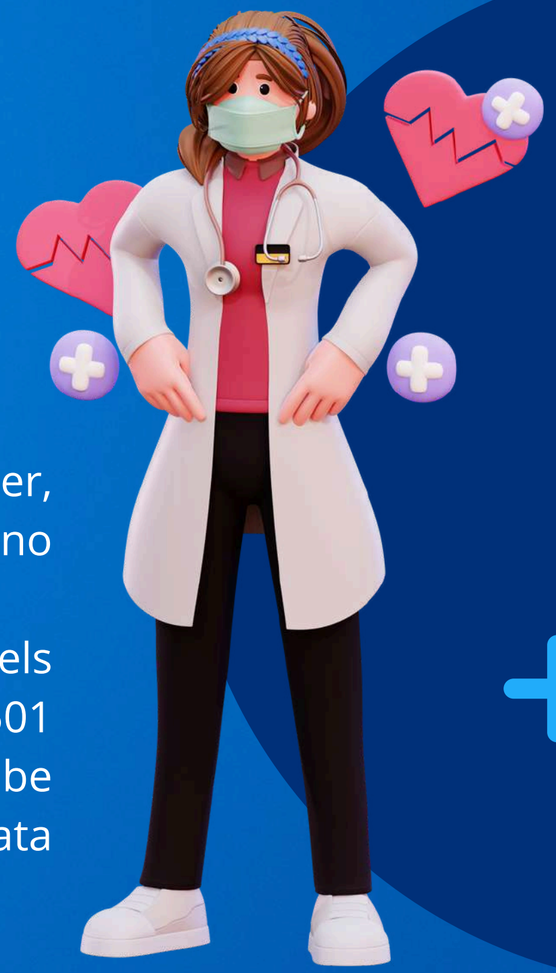
	0
id	0
full_name	0
age	381
gender	0
smoking_status	0
glucose_levels	501
condition	0

data.isnull().sum()

	0
id	0
full_name	0
age	381
gender	0
smoking_status	0
glucose_levels	501
condition	0

Based on the output :

- The column id, full name, gender, smoking status, condition have no missing values.
- The age and glucose levels column contain 381 and 501 missing values, which should be addresed during the data preprocessing stage.



Missing value handling

```
# Handling missing value for EDA, without splitting
for column in data.columns:
    if data[column].dtype == 'object':
        data[column] = data[column].fillna(data[column].mode()[0])
    else:
        data[column] = data[column].fillna(data[column].median())
```

Handling missing data :

- Columns with categorical (object) data types are filled with the mode, which is the most frequent value in the column.
- Columns with numerical data types are filled with the median, which is the middle value of the data distribution.

Missing value handling

	0
id	0
full_name	0
age	0
gender	0
smoking_status	0
glucose_levels	0
condition	0

dtype: int64

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2020 entries, 0 to 2019
Data columns (total 7 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   id               2020 non-null   int64
 1   full_name        2020 non-null   object
 2   age              2020 non-null   float64
 3   gender           2020 non-null   int64
 4   smoking_status   2020 non-null   int64
 5   glucose_levels   2020 non-null   float64
 6   condition        2020 non-null   object
dtypes: float64(2), int64(3), object(2)
memory usage: 110.6+ KB
```

Because the age and glucose levels columns are numerical, missing values in the column are replaced with the median, and the process of removing NULL, values has been successfully completed.

ibimbing

Data Science | Digital Skill Fair 39.0

EXPLORATORY DATA ANALYST

Statistical Summary

	age	smoking_status
count	1639.000000	2020.000000
mean	51.458816	0.095050
std	20.912764	0.293356
min	18.000000	0.000000
25%	33.000000	0.000000
50%	50.000000	0.000000
75%	69.000000	0.000000
max	89.000000	1.000000

	gender
count	2020.000000
mean	3.852475
std	0.354716
min	3.000000
25%	4.000000
50%	4.000000
75%	4.000000
max	4.000000

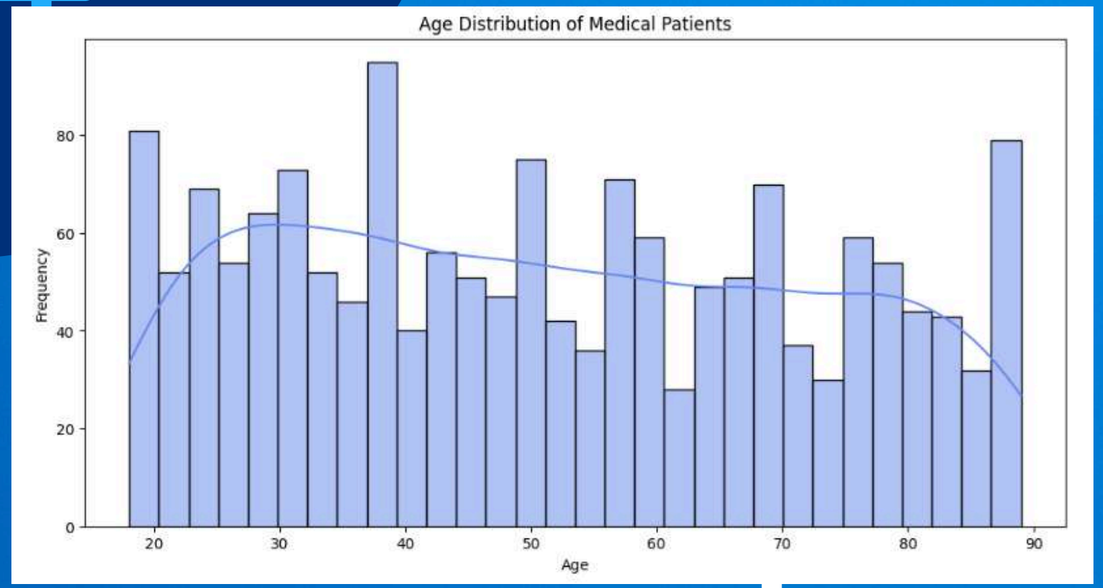
```
gender
4    1722
3     298
```

```
smoking_status
0    1828
1     192
```

Categorical Distribution (gender and smoking status)

1. Gender
   - Male (4) : 1722 patients (85.2%)
   - Female (3) : 298 patients (14.8%)
2. Smoking Status
   - Smoker (1) : 192 patients (9.5%)
   - Non smoker (0) : 1828 patients (90.5%)

Age Distribution

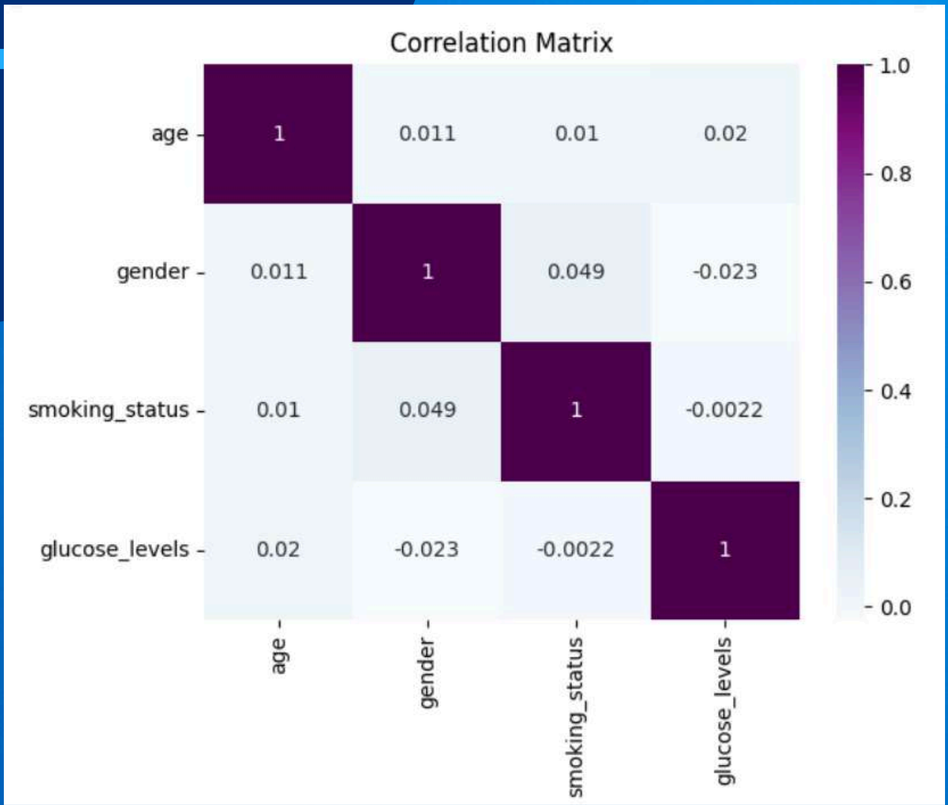


Age Distribution of Medical Patients The chart on the side showing the age ditribution of the medical patients. From that chart, can coclude that:

- Most patients were between 30 until 80 years old, with the highest frequensy around the 35 years old.
- There were also a number of very young patients, even below the age of 18, the although they were fewer than thos in the young adult age group.

This distribution suggests that the majority of medical patients were in their productive or young adult age.

Correlation Matrix



Based on the correlation matrix above, there is **a strong negative correlation** between gender and smoking status, with a coefficient of **0.049**. This suggests that **gender had a insignificant impact on smoking status chances**.

Additionally, the value of **1** in the matrix represents a **perfect correlation of a variable with itself** — such as smoking status with smoking status or age with age. This is standard in any correlation matrix, as each variable is always perfectly correlated with itself.

CONCLUSION

ID

The id column contains unique identifiers for each patient. There are no immediately obvious missing values in the dataset.

Full Name

This column contains the full names of the patients, but in this dataset seems consistent with a combination of “user” and a numerical identifier.

Age

Patients age distribution tends to be normal with most 19 - 89 years old.

Gender

- Male (4) : 1722 patients (85.2%)
- Female (3) : 298 patients (14.8%)

Smoking status

The values are binary, with ‘0’ and ‘1’ observed. ‘0’ represents non-smokers, and ‘1’ represents smokers. The majority of the visible entries are ‘0’ (90.5%)

Glucose levels

This column contains numerical values, a misunderstanding of the unit of measurement, or a significant outlier issue that needs to be investigated and potentially corrected.

Condition

This column contains categorical values indicating the medical condition diagnosed for each patient.

Data Science | Digital Skill Fair 39.0

THANK YOU!!

View Repository on GitHub :
https://github.com/Mariagltm/MedicalConditions

Contact Person :

mariagltm11@gmail.com

www.linkedin.com/in/maria-putri-fredella-gultom