

USE PANDAS TO CLEAN AND PREPROCESS A MESSY DATASET, DOCUMENTING THE STEPS TAKEN DURING THE CLEANING PROCESS.

Pandas is a python library used for working with datasets. it has functions for analyzing, cleaning, exploring, and manipulating data.

Importing the pandas library

importing the pandas

```
import pandas as pd
```

knowing the pandas version

checking the pandas version

```
print (pd.__version__)

1.5.3
```




Uploading the dataset

```
df = pd.read_excel("/content/sample_data/3mtt wk 5 dataset.xlsx")
df
```

	S/NØ	NAME	GENDER	OCUPATION	AGE	GRADE	STATUS	RELIGION
0	1	FATIMA	F	NURSE	NaN	..A	M	MUSLIM
1	2	HASAN	M	ENGR	NaN	A	S	MUSLIM
2	3	FARUK	M	LAWYER	NaN	C	S	MUSLIM
3	4	ZAINAB	F	TEACHER	NaN	B	M	MUSLIM
4	5	FAITH	F	TEACHER	NaN	..C	M	NAN
5	6	DORA	NaN	BANKER	NaN	..E	S/	CHRISTIAN
6	6	DORA	NaN	BANKER	NaN	..E	S/	CHRISTIAN
7	7	SAM	M	JOURNALIST	NaN	C	M	CHRISTIAN
8	8	AYUBA	M	DOCTOR	NaN	A	M	MUSLIM
9	9	HASIYA	F	ENGR	NaN	NaN	S	NAN
10	9	HASIYA	F	ENGR	NaN	NaN	S	NAN
11	9	HASIYA	F	ENGR	NaN	NaN	S	NAN
12	10	SALISU	M	PILOT	NaN	..A	M	MUSLIM




Cleaning the duplicates data

```
df = df.drop_duplicates()
df
```

	S/NØ	NAME	GENDER	OCUPATION	AGE	GRADE	STATUS	RELIGION	
0	1	FATIMA	F	NURSE	NaN	..A	M	MUSLIM	
1	2	HASAN	M	ENGR	NaN	A	S	MUSLIM	
2	3	FARUK	M	LAWYER	NaN	C	S	MUSLIM	
3	4	ZAINAB	F	TEACHER	NaN	B	M	MUSLIM	
4	5	FAITH	F	TEACHER	NaN	..C	M	NAN	
5	6	DORA	NaN	BANKER	NaN	..E	S/	CHRISTIAN	
7	7	SAM	M	JOURNALIST	NaN	C	M	CHRISTIAN	
8	8	AYUBA	M	DOCTOR	NaN	A	M	MUSLIM	
9	9	HASIYA	F	ENGR	NaN	NaN	S	NAN	
12	10	SALISU	M	PILOT	NaN	..A	M	MUSLIM	




✓ cleaning unnecessary columns

```
df = df.drop(columns = 'STATUS')
df
```

	S/NØ	NAME	GENDER	OCUPATION	AGE	GRADE	RELIGION	
0	1	FATIMA	F	NURSE	NaN	..A	MUSLIM	
1	2	HASAN	M	ENGR	NaN	A	MUSLIM	
2	3	FARUK	M	LAWYER	NaN	C	MUSLIM	
3	4	ZAINAB	F	TEACHER	NaN	B	MUSLIM	
4	5	FAITH	F	TEACHER	NaN	..C	NAN	
5	6	DORA	NaN	BANKER	NaN	..E	CHRISTIAN	
7	7	SAM	M	JOURNALIST	NaN	C	CHRISTIAN	
8	8	AYUBA	M	DOCTOR	NaN	A	MUSLIM	
9	9	HASIYA	F	ENGR	NaN	NaN	NAN	
12	10	SALISU	M	PILOT	NaN	..A	MUSLIM	




✓ deleting messy on grade column

```
df["GRADE"] = df["GRADE"].str.lstrip(".")
df
```

	S/NØ	NAME	GENDER	OCUPATION	AGE	GRADE	RELIGION	
0	1	FATIMA	F	NURSE	NaN	A	MUSLIM	
1	2	HASAN	M	ENGR	NaN	A	MUSLIM	
2	3	FARUK	M	LAWYER	NaN	C	MUSLIM	
3	4	ZAINAB	F	TEACHER	NaN	B	MUSLIM	
4	5	FAITH	F	TEACHER	NaN	C	NAN	
5	6	DORA	NaN	BANKER	NaN	E	CHRISTIAN	
7	7	SAM	M	JOURNALIST	NaN	C	CHRISTIAN	
8	8	AYUBA	M	DOCTOR	NaN	A	MUSLIM	
9	9	HASIYA	F	ENGR	NaN	NaN	NAN	
12	10	SALISU	M	PILOT	NaN	A	MUSLIM	




✓ Replacing NAN with nothing on religion column

```
df["RELIGION"] = df["RELIGION"].str.replace('NaN', '')
df
```

	S/NØ	NAME	GENDER	OCUPATION	AGE	GRADE	RELIGION	
0	1	FATIMA	F	NURSE	NaN	A	MUSLIM	
1	2	HASAN	M	ENGR	NaN	A	MUSLIM	
2	3	FARUK	M	LAWYER	NaN	C	MUSLIM	
3	4	ZAINAB	F	TEACHER	NaN	B	MUSLIM	
4	5	FAITH	F	TEACHER	NaN	C		
5	6	DORA	NaN	BANKER	NaN	E	CHRISTIAN	
7	7	SAM	M	JOURNALIST	NaN	C	CHRISTIAN	
8	8	AYUBA	M	DOCTOR	NaN	A	MUSLIM	
9	9	HASIYA	F	ENGR	NaN	NaN		
12	10	SALISU	M	PILOT	NaN	A	MUSLIM	

✓ Cleaning any column that has only NaN

```
df = df.drop(columns = 'AGE')
df
```

	S/NØ	NAME	GENDER	OCUPATION	GRADE	RELIGION	
0	1	FATIMA	F	NURSE	A	MUSLIM	
1	2	HASAN	M	ENGR	A	MUSLIM	
2	3	FARUK	M	LAWYER	C	MUSLIM	
3	4	ZAINAB	F	TEACHER	B	MUSLIM	
4	5	FAITH	F	TEACHER	C		
5	6	DORA	NaN	BANKER	E	CHRISTIAN	
7	7	SAM	M	JOURNALIST	C	CHRISTIAN	
8	8	AYUBA	M	DOCTOR	A	MUSLIM	
9	9	HASIYA	F	ENGR	NaN		
12	10	SALISU	M	PILOT	A	MUSLIM	

Next steps:

Generate code with df

 View recommended plots