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 analying, cleaning, exploring, and manupulating 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
```

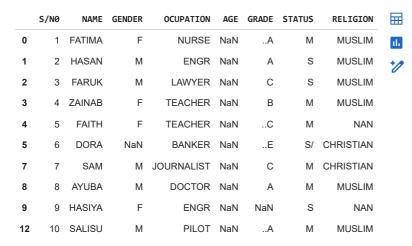
Uploading the dataset

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

	S/N0	NAME	GENDER	OCUPATION	AGE	GRADE	STATUS	RELIGION	
0	1	FATIMA	F	NURSE	NaN	A	М	MUSLIM	11.
1	2	HASAN	М	ENGR	NaN	Α	S	MUSLIM	+/
2	3	FARUK	М	LAWYER	NaN	С	S	MUSLIM	
3	4	ZAINAB	F	TEACHER	NaN	В	М	MUSLIM	
4	5	FAITH	F	TEACHER	NaN	C	М	NAN	
5	6	DORA	NaN	BANKER	NaN	E	S/	CHRISTIAN	
6	6	DORA	NaN	BANKER	NaN	E	S/	CHRISTIAN	
7	7	SAM	М	JOURNALIST	NaN	С	М	CHRISTIAN	
8	8	AYUBA	М	DOCTOR	NaN	Α	М	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	М	PILOT	NaN	A	М	MUSLIM	

Cleaning the duplicates data

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



cleaning unnecessary columns

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

	S/N0	NAME	GENDER	OCUPATION	AGE	GRADE	RELIGION	
0	1	FATIMA	F	NURSE	NaN	A	MUSLIM	11.
1	2	HASAN	М	ENGR	NaN	Α	MUSLIM	+/
2	3	FARUK	М	LAWYER	NaN	С	MUSLIM	
3	4	ZAINAB	F	TEACHER	NaN	В	MUSLIM	
4	5	FAITH	F	TEACHER	NaN	C	NAN	
5	6	DORA	NaN	BANKER	NaN	E	CHRISTIAN	
7	7	SAM	М	JOURNALIST	NaN	С	CHRISTIAN	
8	8	AYUBA	М	DOCTOR	NaN	Α	MUSLIM	
9	9	HASIYA	F	ENGR	NaN	NaN	NAN	
12	10	SALISU	М	PILOT	NaN	A	MUSLIM	

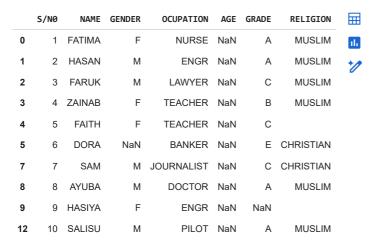
deleting messy on grade column

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

	S/N0	NAME	GENDER	OCUPATION	AGE	GRADE	RELIGION	
0	1	FATIMA	F	NURSE	NaN	Α	MUSLIM	ılı
1	2	HASAN	М	ENGR	NaN	Α	MUSLIM	+/
2	3	FARUK	М	LAWYER	NaN	С	MUSLIM	_
3	4	ZAINAB	F	TEACHER	NaN	В	MUSLIM	
4	5	FAITH	F	TEACHER	NaN	С	NAN	
5	6	DORA	NaN	BANKER	NaN	Е	CHRISTIAN	
7	7	SAM	М	JOURNALIST	NaN	С	CHRISTIAN	
8	8	AYUBA	М	DOCTOR	NaN	Α	MUSLIM	
9	9	HASIYA	F	ENGR	NaN	NaN	NAN	
12	10	SALISU	М	PILOT	NaN	Α	MUSLIM	

Replacing NAN with nothing on religion column

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



Cleaning any column that has only NaN

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

