Missing Data and Duplicates removal

Import modules.

In [1]:
```
import pandas as pd
```

Load the survey dataset into a dataframe.

In [2]:
```
df = pd.read_csv("survey_data.csv")
df.head()
```

Out[2]:

	Respondent	MainBranch	Hobbyist	OpenSourcer	OpenSource	Employment	Country	Student	EdLevel	UndergradMajor	...	WelcomeChange	SONewContent	Age	Gender	Trans	Sexuality	Et
0	4	I am a developer by profession	No	Never	The quality of OSS and closed source software ...	Employed full-time	United States	No	Bachelor's degree (BA, BS, B.Eng., etc.)	Computer science, computer engineering, or sof...	...	Just as welcome now as I felt last year	Tech articles written by other developers;Indu...	22.0	Man	No	Straight / Heterosexual	White or of Eu
1	9	I am a developer by profession	Yes	Once a month or more often	The quality of OSS and closed source software ...	Employed full-time	New Zealand	No	Some college/university study without earning ...	Computer science, computer engineering, or sof...	...	Just as welcome now as I felt last year	NaN	23.0	Man	No	Bisexual	White or of Eu
2	13	I am a developer by profession	Yes	Less than once a month but more than once per ...	OSS is, on average, of HIGHER quality than pro...	Employed full-time	United States	No	Master's degree (MA, MS, M.Eng., MBA, etc.)	Computer science, computer engineering, or sof...	...	Somewhat more welcome now than last year	Tech articles written by other developers;Cour...	28.0	Man	No	Straight / Heterosexual	White or of Eu
3	16	I am a developer by profession	Yes	Never	The quality of OSS and closed source software ...	Employed full-time	United Kingdom	No	Master's degree (MA, MS, M.Eng., MBA, etc.)	NaN	...	Just as welcome now as I felt last year	Tech articles written by other developers;Indu...	26.0	Man	No	Straight / Heterosexual	White or of Eu
4	17	I am a developer by profession	Yes	Less than once a month but more than once per ...	The quality of OSS and closed source software ...	Employed full-time	Australia	No	Bachelor's degree (BA, BS, B.Eng., etc.)	Computer science, computer engineering, or sof...	...	Just as welcome now as I felt last year	Tech articles written by other developers;Indu...	29.0	Man	No	Straight / Heterosexual	Hisp Latino/Latina;Mu

5 rows × 85 columns

Finding duplicates

Find how many duplicate rows (count) exist in the dataframe.

In [3]:
```
# your code goes here
df.duplicated().sum()
```

Out[3]:
```
154
```

Removing duplicates

Remove the duplicate rows from the dataframe. Update in actual Dataframe

In [4]:
```
# your code goes here
df.drop_duplicates(inplace=True)
```

Verify if duplicates were actually dropped.

In [5]:
```
# your code goes here
df.duplicated().sum()
```

Out[5]:
```
0
```

Finding Missing values

Find the missing values for all columns.

In [6]:
```
# your code goes here
df.isna().sum()
```

Out[6]:

Respondent	0
MainBranch	0
Hobbyist	0
OpenSourcer	0
OpenSource	81
Employment	0
Country	0
Student	51
EdLevel	112
UndergradMajor	737
EduOther	164
OrgSize	96
DevType	65
YearsCode	9
Age1stCode	13
YearsCodePro	16
CareerSat	0
JobSat	1
MgrIdiot	493
MgrMoney	497
MgrWant	493
JobSeek	0
LastHireDate	0
LastInt	413
FizzBuzz	37
JobFactors	3
ResumeUpdate	39
CurrencySymbol	0
CurrencyDesc	0
CompTotal	809
...	
Containers	82
BlockchainOrg	2322
BlockchainIs	2610
BetterLife	98
ITperson	35
OffOn	38
SocialMedia	293
Extraversion	20
ScreenName	507
SOVisit1st	325
SOVisitFreq	5
SOVisitTo	1
SOFindAnswer	3
SOTimeSaved	50
SOHowMuchTime	1917
SOAccount	1
SOPartFreq	1128
SOJobs	6
EntTeams	5
SOComm	0
WelcomeChange	85
SONewContent	1965
Age	287
Gender	73
Trans	123
Sexuality	542
Ethnicity	675
Dependents	140
SurveyLength	19

```
SurveyEase          14
Length: 85, dtype: int64
```

Find out how many rows are missing in the column 'WorkLoc'

In [7]:
```
# your code goes here
df["WorkLoc"].isna().sum()
```

Out[7]:
```
32
```

Imputing missing values

Find the value counts and unique values for the column WorkLoc

In [9]:
```
# your code goes here
print(df["WorkLoc"].value_counts())
print(df.WorkLoc.unique())
```

```
Office                                          6806
Home                                            3589
Other place, such as a coworking space or cafe   971
Name: WorkLoc, dtype: int64
['Home' 'Office' 'Other place, such as a coworking space or cafe' nan]
```

Impute (replace) all the empty rows in the column WorkLoc with the value NONE

In [12]:
```
# your code goes here
df['WorkLoc'].fillna('NONE', inplace = True)
```

After imputation there should ideally not be any empty rows in the WorkLoc column.

In [14]:
```
df["WorkLoc"].isna().sum()
```

Out[14]:
```
0
```