Quantifying missing data

Missing data refers to the absence of a value for observations and is a common occurrence in most datasets. Sci kit-learn, the open source Python library for machine learning, does not support missing values as input for machine learning models, so we need to convert these values into numbers. To select the missing data imputation technique, it is important to know about the amount of missing information in our variables. In this recipe, we will learn how to identify and quantify missing data using pandas and how to make plots with the percentages of missing data per variable.

```
# import the required python libraries
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# to display total no of columns present in dataset
pd.set_option('display.max_columns',None)

# we will use the selected variables from a dataset
cols ['AGE', 'NUMCHLD', 'INCOME', 'WEALTH1', 'MBCRAFT',
'MBGARDEN','MBBOOKS','MBCOLECT','MAGFAML','MAGFEM',
'MAGMALE']

data = pd.read_csv('data/cup98LRN.txt',usecols=cols)
data.head()
```

@datascience_2.0

output:

AGE	NUMCHLD	INCOME	WEALTH1	MBCRAFT	MBGARDEN	MBBOOKS	MBCOLECT	MAGFAML	MAGFEM	MAGMALE
60.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
46.0	1.0	6.0	9.0	0.0	0.0	3.0	1.0	1.0	1.0	0.0
NaN	NaN	3.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
70.0	NaN	1.0	4.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
78.0	1.0	3.0	2.0	1.0	0.0	9.0	0.0	4.0	1.0	0.0

 # let's calculate the missing values in each column or variable

data.isnull().sum()

output:

```
AGE          23665
NUMCHLD      83026
INCOME       21286
WEALTH1      44732
MBCRAFT      52854
MBGARDEN     52854
MBBOOKS      52854
MBCOLECT     52914
MAGFAML      52854
MAGFEM       52854
MAGMALE      52854
dtype: int64
```

 # let's quantify the percentage of missing values in each variable

data.isnull().mean()

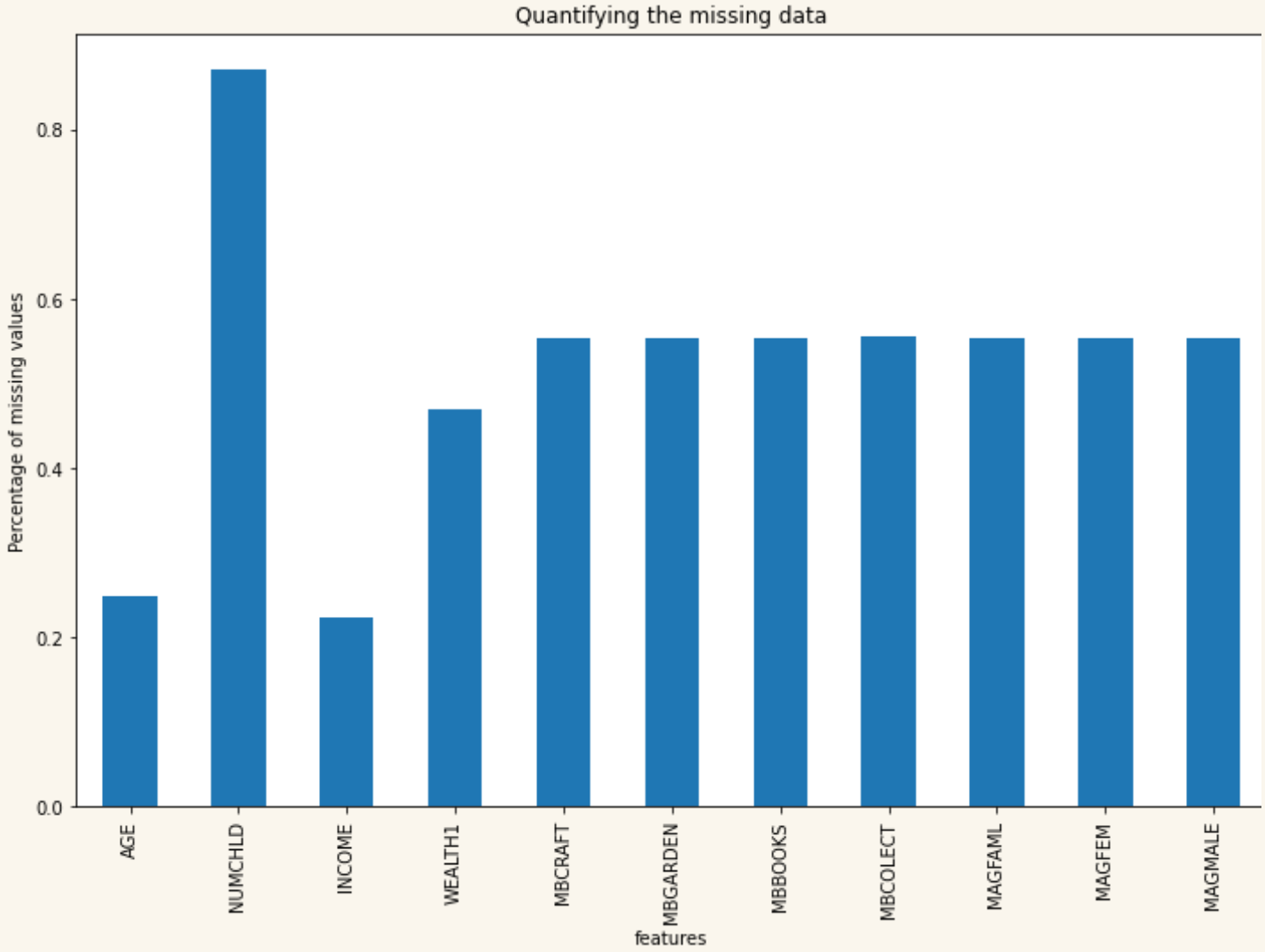
output:

```
AGE          0.248030
NUMCHLD      0.870184
INCOME       0.223096
WEALTH1      0.468830
MBCRAFT      0.553955
MBGARDEN     0.553955
MBBOOKS      0.553955
MBCOLECT     0.554584
MAGFAML      0.553955
MAGFEM       0.553955
MAGMALE      0.553955
dtype: float64
```

```
# we can also plot the missing values

data.isnull().mean().plot.bar(figsize=(12,8))
plt.ylabel('Percentage of missing values')
plt.xlabel('features')
plt.title('Quantifying the missing data')
```

output:

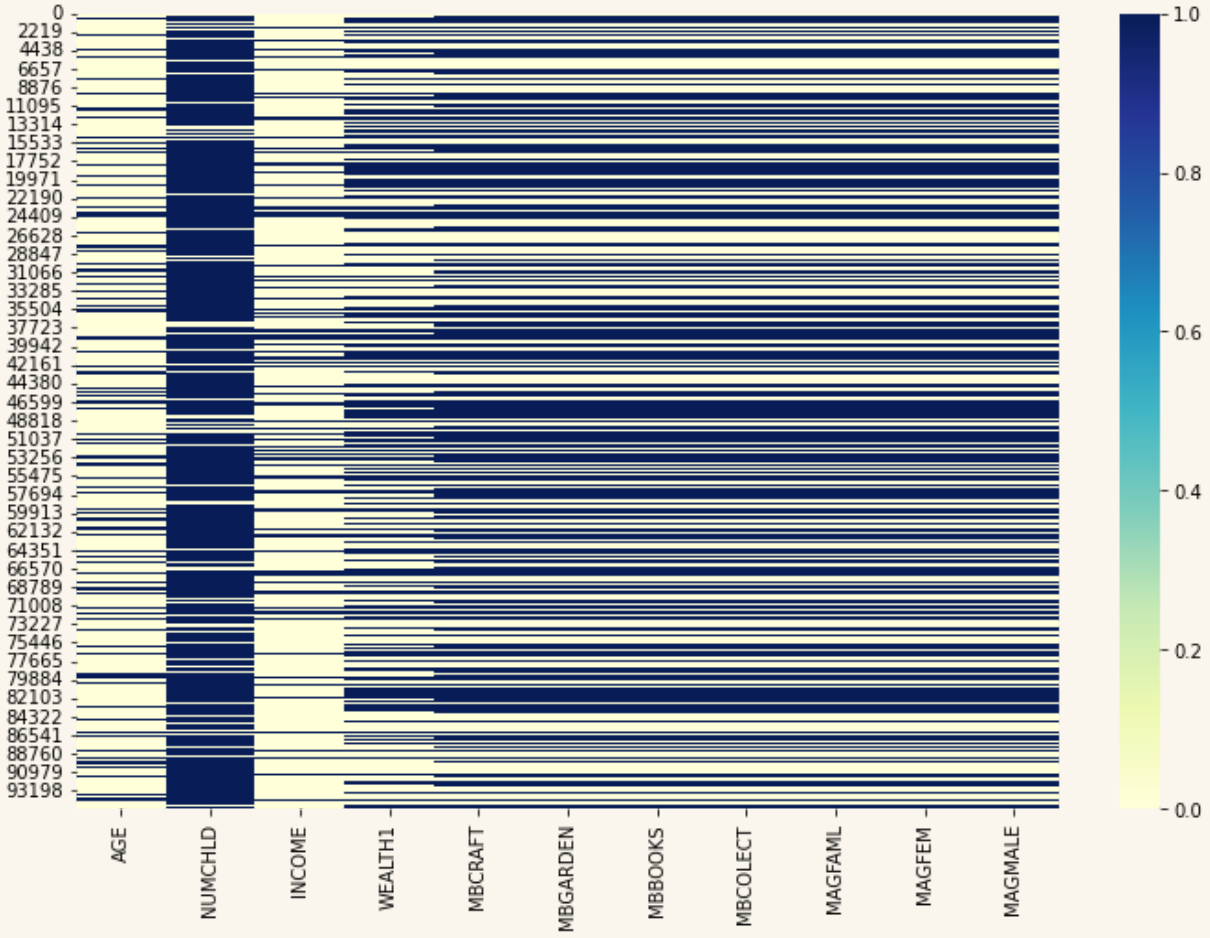


Tip: We can change the figure size using the figsize argument within pandas **plot.bar**() and we can add x and y labels and a title with the **plt.xlabel**() , **plt.ylabel**(), and **plt.title**() methods from Matplotlib to enhance the aesthetics of the plot.

@datascience_2.0

```
# we can aslo observe the missing values with the help of heat map
plt.figure(figsize=(12,8))
sns.heatmap(data.isnull(),cmap="YlGnBu")
```

output:



Observation: You can observe the missing values in blue color representation.