1.0- Handling Missing values (1)

December 29, 2023

0.1 Missing Values

Missing values occurs in dataset when some of the informations is not stored for a variable There are 3 mechanisms

0.1.1 1 Missing Completely at Random, MCAR:

Missing completely at random (MCAR) is a type of missing data mechanism in which the probability of a value being missing is unrelated to both the observed data and the missing data. In other words, if the data is MCAR, the missing values are randomly distributed throughout the dataset, and there is no systematic reason for why they are missing.

For example, in a survey about the prevalence of a certain disease, the missing data might be MCAR if the survey participants with missing values for certain questions were selected randomly and their missing responses are not related to their disease status or any other variables measured in the survey.

0.1.2 2. Missing at Random MAR:

Missing at Random (MAR) is a type of missing data mechanism in which the probability of a value being missing depends only on the observed data, but not on the missing data itself. In other words, if the data is MAR, the missing values are systematically related to the observed data, but not to the missing data. Here are a few examples of missing at random:

Income data: Suppose you are collecting income data from a group of people, but some participants choose not to report their income. If the decision to report or not report income is related to the participant's age or gender, but not to their income level, then the data is missing at random.

Medical data: Suppose you are collecting medical data on patients, including their blood pressure, but some patients do not report their blood pressure. If the patients who do not report their blood pressure are more likely to be younger or have healthier lifestyles, but the missingness is not related to their actual blood pressure values, then the data is missing at random.

0.2 3. Missing data not at random (MNAR)

It is a type of missing data mechanism where the probability of missing values depends on the value of the missing data itself. In other words, if the data is MNAR, the missingness is not random and is dependent on unobserved or unmeasured factors that are associated with the missing values.

For example, suppose you are collecting data on the income and job satisfaction of employees in a company. If employees who are less satisfied with their jobs are more likely to refuse to report their income, then the data is not missing at random. In this case, the missingness is dependent on job satisfaction, which is not directly observed or measured.

0.3 Examples

[40]:
```python
import seaborn as sns
import numpy as np
import pandas as pd
```

[41]:
```python
df=sns.load_dataset('titanic')
```

[42]:
```python
df.head()
```

[42]:
```
   survived  pclass     sex   age  sibsp  parch     fare embarked  class  \
0         0       3    male  22.0      1      0   7.2500        S  Third
1         1       1  female  38.0      1      0  71.2833        C  First
2         1       3  female  26.0      0      0   7.9250        S  Third
3         1       1  female  35.0      1      0  53.1000        S  First
4         0       3    male  35.0      0      0   8.0500        S  Third

     who  adult_male deck  embark_town alive  alone
0    man        True  NaN  Southampton    no  False
1  woman       False    C    Cherbourg   yes  False
2  woman       False  NaN  Southampton   yes   True
3  woman       False    C  Southampton   yes  False
4    man        True  NaN  Southampton    no   True
```

[43]:
```python
## Check missing values
df.isnull().sum()
```

[43]:
```
survived         0
pclass           0
sex              0
age            177
sibsp            0
parch            0
fare             0
embarked         2
class            0
who              0
adult_male       0
deck           688
embark_town      2
alive            0
alone            0
dtype: int64
```

[44]:
```python
df.shape
```

[44]: (891, 15)

[45]:
```python
## Delete the rows or data point to handle missing values

df.shape
```

[45]: (891, 15)

[46]:
```python
df.dropna().shape
```

[46]: (182, 15)

[47]:
```python
## Column wise deletion
df.dropna(axis=1)
```

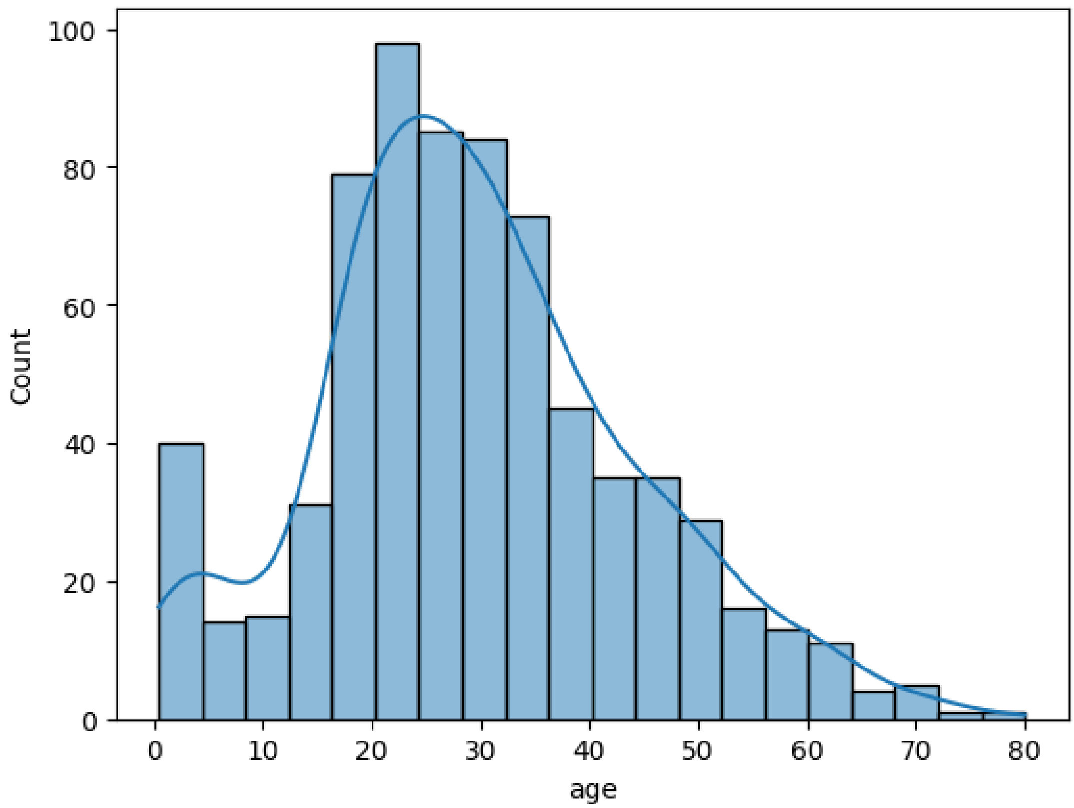
[47]:
```
     survived  pclass     sex  sibsp  parch     fare   class    who  \
0           0       3    male      1      0   7.2500   Third    man
1           1       1  female      1      0  71.2833   First  woman
2           1       3  female      0      0   7.9250   Third  woman
3           1       1  female      1      0  53.1000   First  woman
4           0       3    male      0      0   8.0500   Third    man
..        …       …     …      …      …        …     …      …
886         0       2    male      0      0  13.0000  Second    man
887         1       1  female      0      0  30.0000   First  woman
888         0       3  female      1      2  23.4500   Third  woman
889         1       1    male      0      0  30.0000   First    man
890         0       3    male      0      0   7.7500   Third    man

     adult_male alive  alone
0          True    no  False
1         False   yes  False
2         False   yes   True
3         False   yes  False
4          True    no   True
..          …     …      …
886        True    no   True
887       False   yes   True
888       False    no  False
889        True   yes   True
890        True    no   True

[891 rows x 11 columns]
```

0.4 Imputation Missing Values

0.4.1 1- Mean Value Imputation

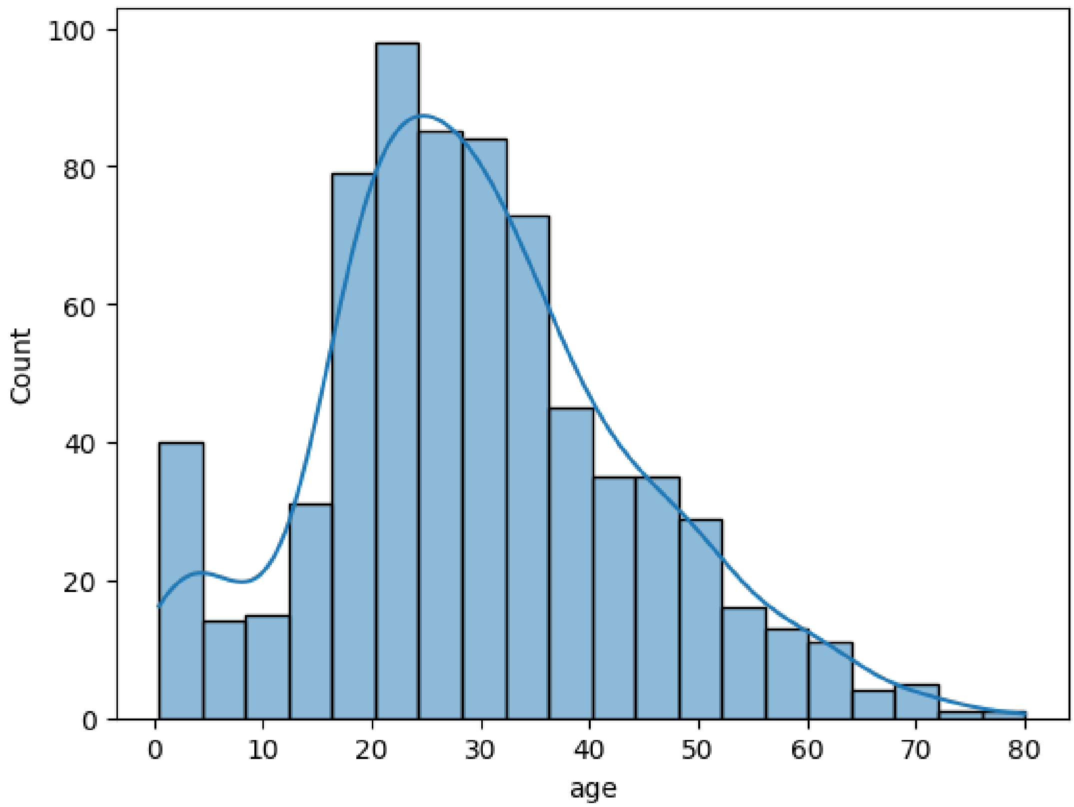
[48]:
```
sns.histplot(df['age'],kde=True)
```

[48]: <Axes: xlabel='age', ylabel='Count'>



[49]:
```
sns.histplot(df['age'],kde=True)
```

[49]: <Axes: xlabel='age', ylabel='Count'>



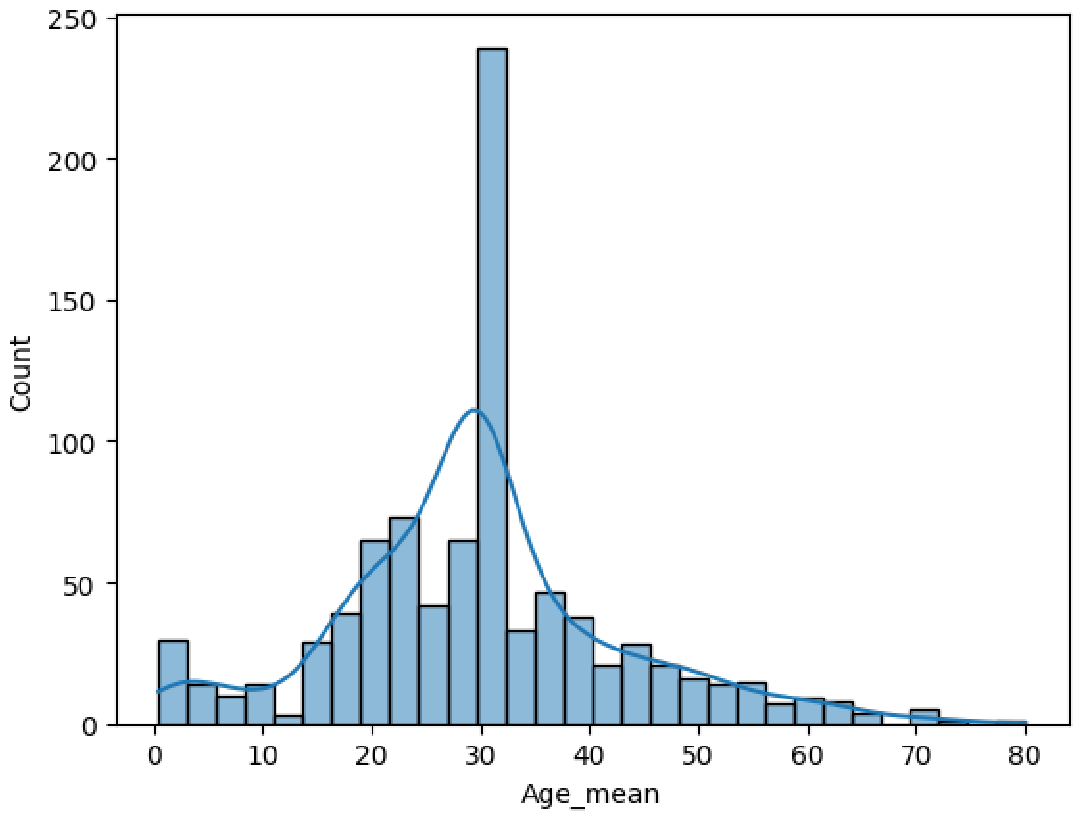
[50]:
```
df['Age_mean']=df['age'].fillna(df['age'].mean())
```

[51]:
```
df[['Age_mean','age']]
```

[51]:
```
      Age_mean   age
0    22.000000  22.0
1    38.000000  38.0
2    26.000000  26.0
3    35.000000  35.0
4    35.000000  35.0
..         …    …
886  27.000000  27.0
887  19.000000  19.0
888  29.699118   NaN
889  26.000000  26.0
890  32.000000  32.0

[891 rows x 2 columns]
```

[52]:
```
sns.histplot(df['Age_mean'],kde=True)
```

[52]: <Axes: xlabel='Age_mean', ylabel='Count'>



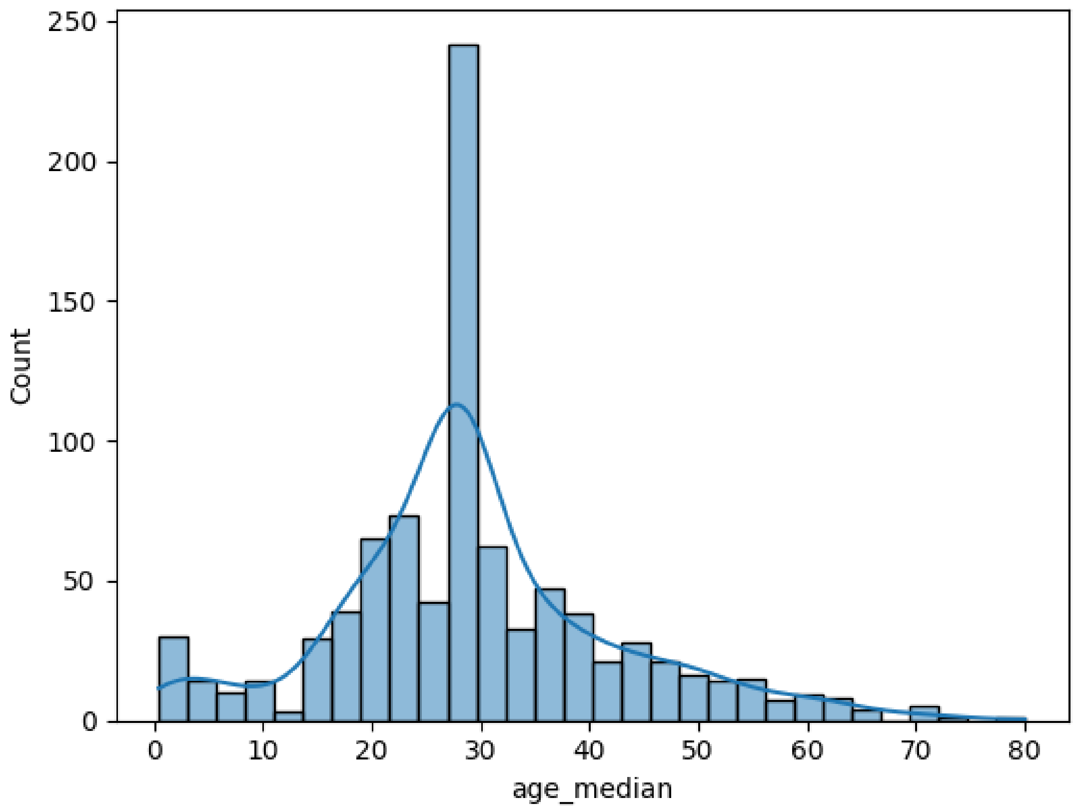
[53]:
```
## MEan Imputation Works Well when we have normally distributed data
```

0.4.2 2. Median Value Imputation- If we have outliers in the dataset use median

[54]:
```
df['age_median']=df['age'].fillna(df['age'].median())
```

[55]:
```
sns.histplot(df['age_median'],kde=True)
```

[55]: <Axes: xlabel='age_median', ylabel='Count'>



[55]:

[55]:

[56]:
```
df[['age_median','Age_mean','age']]
```

[56]:

```
     age_median   Age_mean   age
0          22.0  22.000000  22.0
1          38.0  38.000000  38.0
2          26.0  26.000000  26.0
3          35.0  35.000000  35.0
4          35.0  35.000000  35.0
..          …          …     …
886        27.0  27.000000  27.0
887        19.0  19.000000  19.0
888        28.0  29.699118   NaN
889        26.0  26.000000  26.0
890        32.0  32.000000  32.0

[891 rows x 3 columns]
```

0.4.3 3. Mode Imputation Technqiue–Categorical values

[60]:
```
df[df['embarked'].isnull()]
```

[60]:
```
     survived  pclass     sex   age  sibsp  parch  fare embarked  class  \
61          1       1  female  38.0      0      0  80.0      NaN  First
829         1       1  female  62.0      0      0  80.0      NaN  First

       who  adult_male deck embark_town alive  alone  Age_mean  age_median
61   woman       False    B         NaN   yes   True      38.0        38.0
829  woman       False    B         NaN   yes   True      62.0        62.0
```

[61]:
```
df[df['embarked'].isnull()]
```

[61]:
```
     survived  pclass     sex   age  sibsp  parch  fare embarked  class  \
61          1       1  female  38.0      0      0  80.0      NaN  First
829         1       1  female  62.0      0      0  80.0      NaN  First

       who  adult_male deck embark_town alive  alone  Age_mean  age_median
61   woman       False    B         NaN   yes   True      38.0        38.0
829  woman       False    B         NaN   yes   True      62.0        62.0
```

[66]:
```
df['embarked'].unique()
```

[66]:
```
array(['S', 'C', 'Q', nan], dtype=object)
```

[70]:
```
mode_value=df[df['embarked'].notna()]['embarked'].mode()[0]
mode_value
```

[70]:
```
'S'
```

[74]:
```
df['embarked_mode']=df['embarked'].fillna(mode_value)
```

[75]:
```
df[['embarked_mode','embarked']]
```

[75]:
```
    embarked_mode embarked
0               S        S
1               C        C
2               S        S
3               S        S
4               S        S
..            …        …
886             S        S
887             S        S
888             S        S
889             C        C
890             Q        Q
```

```
[891 rows x 2 columns]
```

[76]:
```python
df['embarked_mode'].isnull().sum()
```

[76]:
```
0
```

[77]:
```python
df['embarked'].isnull().sum()
```

[77]:
```
2
```

[ ]:

[ ]:

[ ]:

[ ]:

[ ]:

[ ]:

[ ]:

[ ]: