Discretization

Discretization is the process of transforming continuous variables into discrete variables by creating a set of contiguous intervals that span the range of the variable's values. Discretization is also called **binning**, where bin is an alternative name for interval.

Discretization helps handle outliers and may improve value spread in skewed variables

Discretization helps handle outliers by placing these values into the lower or higher intervals, together with the remaining inlier values of the distribution. Thus, these outlier observations no longer differ from the rest of the values at the tails of the distribution, as they are now all together in the same interval / bucket. In addition, by creating appropriate bins or intervals, discretization can help spread the values of a skewed variable across a set of bins with equal number of observations.

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

In [2]:
```
stroke_data = pd.read_csv('stroke prediction.csv')
stroke_data.head()
```

Out[2]:

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_typ
0	30669	Male	3.0	0	0	No	children	Rura
1	30468	Male	58.0	1	0	Yes	Private	Urba
2	16523	Female	8.0	0	0	No	Private	Urba
3	56543	Female	70.0	0	0	Yes	Private	Rura
4	46136	Male	14.0	0	0	No	Never_worked	Rura

In [3]:
```
stroke_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 43400 entries, 0 to 43399
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   id                 43400 non-null  int64  
 1   gender             43400 non-null  object 
 2   age                43400 non-null  float64
 3   hypertension       43400 non-null  int64  
 4   heart_disease      43400 non-null  int64  
 5   ever_married       43400 non-null  object 
 6   work_type          43400 non-null  object 
 7   Residence_type     43400 non-null  object 
 8   avg_glucose_level  43400 non-null  float64
 9   bmi                41938 non-null  float64
 10  smoking_status     30108 non-null  object 
 11  stroke             43400 non-null  int64  
dtypes: float64(3), int64(4), object(5)
memory usage: 4.0+ MB
```

In [4]:
```
stroke_data['age'].value_counts()
```

Out[4]:
```
51.00    738
52.00    721
53.00    701
78.00    698
50.00    694
        ...
0.48      37
0.40      35
1.00      34
0.16      26
0.08      17
Name: age, Length: 104, dtype: int64
```

Creating Bins

In [5]:
```
intervals = [0,12,19,30,60,90]

categories=['child','teenager','young_adult','middle_aged', 'senior_citizen']
```

In [6]:
```
stroke_data['age_category']=pd.cut(x=stroke_data['age'],
                                   bins = intervals,
                                   labels= categories)
```

In [7]:
```
stroke_data.head()
```

Out[7]:

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_typ
0	30669	Male	3.0	0	0	No	children	Rura
1	30468	Male	58.0	1	0	Yes	Private	Urba
2	16523	Female	8.0	0	0	No	Private	Urba
3	56543	Female	70.0	0	0	Yes	Private	Rura
4	46136	Male	14.0	0	0	No	Never_worked	Rura

In [8]:
```
stroke_data[['age', 'age_category']]
```

Out[8]:

	age	age_category
0	3.0	child
1	58.0	middle_aged
2	8.0	child
3	70.0	senior_citizen
4	14.0	teenager
...	...	...
43395	10.0	child
43396	56.0	middle_aged
43397	82.0	senior_citizen
43398	40.0	middle_aged
43399	82.0	senior_citizen

43400 rows × 2 columns