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 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]:

```
df= pd.read_csv('stroke prediction.csv')
df.head()
```

Out[2]:

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	30669	Male	3.0	0	0	No	children	Rural	95.12	18.0	NaN	0
1	30468	Male	58.0	1	0	Yes	Private	Urban	87.96	39.2	never smoked	0
2	16523	Female	8.0	0	0	No	Private	Urban	110.89	17.6	NaN	0
3	56543	Female	70.0	0	0	Yes	Private	Rural	69.04	35.9	formerly smoked	0
4	46136	Male	14.0	0	0	No	Never_worked	Rural	161.28	19.1	NaN	0

In [3]:

```
df.shape
```

Out[3]:

```
(43400, 12)
```

In [4]:

```
df["stroke"].value_counts()
```

Out[4]:

```
0    42617
1      783
Name: stroke, dtype: int64
```

Discretization

In [5]:

```
### Creating Bins
intervals = [0,12,19,30,60,90]
categories = ['child','teenager','young_adult','middle_aged', 'senior_citizen']

###apply discretization using intervals
df['age_category'] = pd.cut(df['age'], bins = intervals, labels= categories)
df.head()
```

Out[5]:

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke	age_c
0	30669	Male	3.0	0	0	No	children	Rural	95.12	18.0	NaN	0	
1	30468	Male	58.0	1	0	Yes	Private	Urban	87.96	39.2	never smoked	0	midd
2	16523	Female	8.0	0	0	No	Private	Urban	110.89	17.6	NaN	0	
3	56543	Female	70.0	0	0	Yes	Private	Rural	69.04	35.9	formerly smoked	0	senio
4	46136	Male	14.0	0	0	No	Never_worked	Rural	161.28	19.1	NaN	0	t

In [6]:

```
df[['age','age_category']]
```

Out[6]:

	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

In [7]:

```
df['age_category'].value_counts()
```

Out[7]:

```
middle_aged       18653
senior_citizen    10511
young_adult        5725
child              5326
teenager           3185
Name: age_category, dtype: int64
```

In [8]:

```
pd.crosstab(df.age_category,df.stroke)
```

Out[8]:

stroke	0	1
age_category		
child	5325	1
teenager	3184	1
young_adult	5723	2
middle_aged	18453	200
senior_citizen	9932	579