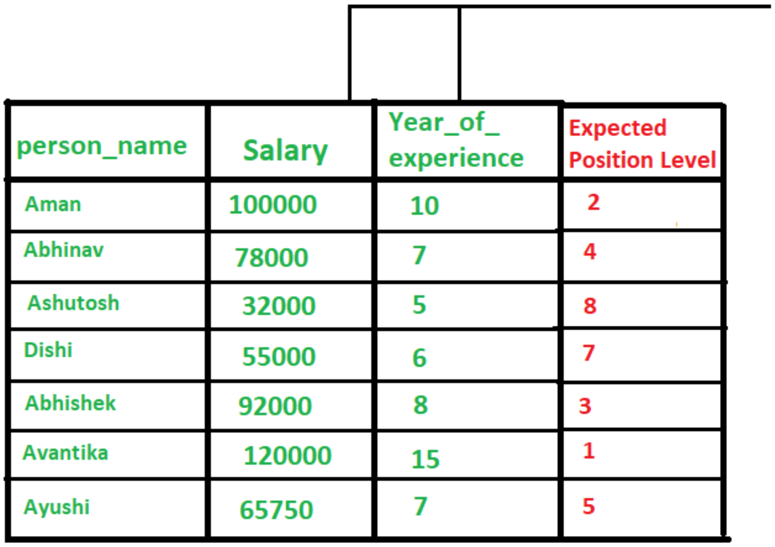
Normalization

- Also called data pre-processing, this is one of the crucial techniques for **data transformation in data mining.**
- Here, the data is transformed so that it falls under a given range.
- When attributes are on different ranges or scales, data modelling and mining can be difficult. Normalization helps in applying data mining algorithms and extracting data faster.
- Data normalization involves converting all data variable into a given range



person_name	Salary	Year_of_ experience	Expected Position Level
Aman	100000	10	2
Abhinav	78000	7	4
Ashutosh	32000	5	8
Dishi	55000	6	7
Abhishek	92000	8	3
Avantika	120000	15	1
Ayushi	65750	7	5

The attributes salary and year_of_experience are on different scale and hence attribute salary can take high priority over attribute year_of_experience in the model.

The popular normalization methods are:

1. Min-max normalization
2. Decimal scaling
3. Z-score normalization

1. Min-Max Normalization:

It is the linear transformation of the original unstructured data. It scales the data from 0 to 1. It is calculated by the following formula:

$$v' = \frac{v\text{−minF}}{\text{maxF−minF}}(new\_max_F - new\_min_F) + new\_min_F \ ,$$

where V’ = equivalent value for V
new_min .. new_max is new range
min..max is Old range for the current value of feature *F*.

Assume that the minimum and maximum values for the feature F are $50,000 and $100,000 correspondingly. It needs to range *F* from 0 to 1. In accordance with min-max normalization, *v* = $80,000 is transformed to:

$$v' = \frac{80{,}000 - 50{,}000}{100{,}000 - 50{,}000} + (1 - 0) + 0 = \frac{3}{5} = 0{,}6$$

- As you can see this technique enables to interpret the data easily.
- There are no large numbers, only concise data that do not require further transformation and can be used in decision-making process immediately.
- This transforms the original data linearly.

2. Z-Score Normalization:

- It is also called zero-mean normalization.
- The essence of this technique is the data transformation by the values conversation to a common scale where an average number equals zero and a standard deviation is one.
- A value is normalized to ′ under the formula:

$$v' = \frac{\mathrm{v}-\overline{F}}{\sigma_F},$$

Here is the mean and
σ_F'is the standard deviation of feature *F*.

Example:
On the supposition that the mean of feature is $65,000 and its standard deviation is $ 18,000. Applying the z-score normalization we get the following mean of the value equals to $85,800:

(85,800-65,000) /18,000 = 1.1.16

Z-Scores will help in situation like

- Sometimes we want to do more than summarize a bunch of scores.
- Sometimes we want to talk about particular scores within the bunch.
- We may want to tell other people about whether or not a score is above or below average.
- We may want to tell other people how far away a particular score is from average.
- We might also want to compare scores from different bunches of data. We will want to know which score is better.

Z-Scores tell us

- whether a particular score is equal to the mean, below the mean or above the mean of a bunch of scores.
- They can also tell us how far a particular score is away from the mean.
- Is a particular score close to the mean or far away?

If a Z-Score….

- Has a value of 0, it is equal to the group mean.
- Is positive, it is above the group mean.
- Is negative, it is below the group mean.

- Is equal to +1, it is 1 Standard Deviation above the mean.
- Is equal to +2, it is 2 Standard Deviations above the mean.
- Is equal to -1, it is 1 Standard Deviation below the mean.
- Is equal to -2, it is 2 Standard Deviations below the mean.

Z-Scores Can Help Us Understand…

- How typical a particular score is within bunch of scores.
- If data are normally distributed, approximately 95% of the data should have Z-score between -2 and +2.
- Z-scores that do not fall within this range may be less typical of the data in a bunch of scores.
- Thus Z score will help to identify the outlier

Z-Scores Can Help Us Compare…

- Individual scores from different bunches of data. We can use Z-scores to standardize scores from different groups of data. Then we can compare raw scores from different bunches of data.

Advantages of the z score

- The z-score is very useful when we are understanding the data. Some of the useful facts are mentioned below;
  The z-score is a very useful statistic of the data due to the following facts;
  It allows a data administrator to understand the probability of a score occurring within the normal distribution of the data.

- The z-score enables a data administrator to compare two different scores that are from different normal distributions of the data.

Is a higher or lower Z score better?

- Suppose we have data from two persons. Person A has a high Z score value and person B have low Z Score value. In this case, the higher Z-score indicates that Person A is far away from person B.

What does a negative and a positive z score mean?

- A negative z-score indicates that the data point is below the mean.
  A positive z-score indicates that the data point is above the mean.

Why is the mean of Z scores is 0?

- The standard deviation of the z-scores is always 1 and similarly, the mean of the z-scores is always 1.
  Z-scores values above the 0 represent that sample values are above the mean.
  z-scores values below the 0 represent that sample values are below the mean.

In the case of squared z-scores, the sum of the squared z-scores is always equal to the number of z-score values.

What is the meaning of the high Z score and low Z score?

- Suppose we have a high z-score value then it means a very low probability of data above this z-score.
- Suppose we have a low z-score value then it means a very low probability of data below this z-score.

Comparison of Min-Max Normalization and Z-Score Normalization

Min-max normalization	**Z-score normalization**
Not very well efficient in handling the outliers	Handles the outliers in a good way.
Min-max Guarantees that all the features will have the exact same scale.	Helpful in the normalization of the data but not with the ***exact*** same scale.

https://www.statisticshowto.com/probability-and-statistics/z-score/
https://www.codecademy.com/articles/normalization
https://statistics.laerd.com/statistical-guides/standard-score-1.php (For extra reading)

3. Decimal Scaling:

- It normalizes the values of an attribute by changing the position of their decimal points
- The number of points by which the decimal point is moved can be determined by the absolute maximum value of attribute A.
- A value, v, of attribute A is normalized to v' by computing

$$v' = \frac{v}{10^j}$$

- where j is the smallest integer such that Max(|v'|) < 1.

For **example**:

- Suppose: Values of an attribute P varies from -99 to 99.
- The maximum absolute value of P is 99.
- For normalizing the values we divide the numbers by 100 (i.e., j = 2) or (number of integers in the largest number) so that values come out to be as 0.98, 0.97 and so on.