Chapter 4
Types of Data Manipulations

Study Exercise – Learning the most famous data transformation methods

The most famous data transformations are listed in the following: **Normalization**, **Standardization**, **Log Transformation**, **Feature Extraction**, **Attribute Construction**, **Binary coding**, **Ranking Transformation**, **Discretization**, **Smoothing**, **Aggregation**, and **Binning**.

For each of the listed Data Transformation methods, answer the following questions.

1. What does the method do to the values of the data?
2. What is the objective of the method?
3. Does the method also involve changing the structure of the data (Data Wrangling)? Explain.
4. Describe a situation in which the method would be useful.
5. Does the method have names that are used commonly?

Once you've answered the preceding quesitons about each mentioned method, then answer the following quesitons.

- **Q1**: What's the difference between *Normalization* and *Standardization*? Describe two different situations in that one would be preferable to the other.
- **Q2**: What are the differences and similarities between *Attribute Construction* and *Feature Extraction*?

- **Q3**: *Binary Coding*, and *Ranking Transformation* have something in common that is in contrast with *Discretization*, what is that?
- **Q4**: What is the relationship between *Smoothing* and *Aggregation*?
- **Q5**: what is the relationship between *Binning* and *Discretization*?

Answers to method-specific questions

Normalization

What does the method do to the values of the data? The values of a whole population are rescaled with the following goals: 1) the largest value in the population is represented with a prescribed new maximum; the prescribed new maximum is usually one. 2) the smallest value in the population is represented with a prescribed new minimum; the prescribed new minimum is usually zero. 3) the relative difference between the values of the population is intact; for example, if an individual value was twice another individual value, this will stay the same.
What is the objective of the method? Rescaling the value in the population. For instance, if the height of a population of people is between 140 to 210 centimeters, using normalization we can rescale the values so they are between, say zero and one.
Does the method also involve changing the structure of the data (Data Wrangling)? Explain. No.
Describe a situation in which the method would be useful.

Most clustering algorithms are sensitive to the scale of the attributes and attributes with larger scales will tend to be given more weight. Normalization of all the attributes will make sure that each attribute will have equal weight.
Does the method have other names that are used commonly? Sometimes normalization is referred to as standardization however that is a common mistake and should be avoided.

Standardization

What does the method do to the values of the data? The values of a whole population are rescaled so the standard deviation of the rescaled population is equal to 1.
What is the objective of the method? Making sure the standard deviation of the rescaled attribute is equal to 1.
Does the method also involve changing the structure of the data (Data Wrangling)? Explain. No.
Describe a situation in which the method would be useful. When performing Principal Component Analysis (PCA) the attributes that have larger variations will be more important in the shaping of the results, so performing standardization will make sure that all attributes will have the same weight in shaping the results.
Does the method have names that are used commonly? No.

Log Transformation

What does the method do to the values of the data?

If the original value is x, the log transformation would mean that we would replace x with log(x). We normally use the natural logarithm in log transformation, but we won't have to.

What is the objective of the method?

There are phenomena such as wealth and salary in that the measured value of a small portion of the population is exponentially higher than the rest of the population. The fact that a portion of the population is represented with values too different from the rest of the population creates a bias in certain data analytics; log transformation rescales the values so this issue is taken care of.

Does the method also involve changing the structure of the data (Data Wrangling)? Explain.

No.

Describe a situation in which the method would be useful.

When salary is an independent attribute in a regression model, log transformation will help the model significantly.

Does the method have names that are used commonly?

No.

Feature Extraction

What does the method do to the values of the data?

The process of encoding our knowledge of the data objects and the patterns we expect from them into a computer program that transforms the data into a higher level of representation that we know will be more useful for the analytics we have in mind.
What is the objective of the method? Push the useful information in the data to the surface so our models can pick up on them easier.
Does the method also involve changing the structure of the data (Data Wrangling)? Explain. Yes, the structure of the data will change often.
Describe a situation in which the method would be useful. See Feature Extraction in action – An example with data from chapter 3.
Does the method have names that are used commonly? No.

Attribute Construction

What does the method do to the values of the data? The process of encoding our knowledge of the data objects and the patterns we expect from them into a formula that mixes two or more attributes into one that we know will be more useful for the analytics we have in mind.
What is the objective of the method?

Make the important pattern more accessible to the following HLCUs: Human language and Models.
Does the method also involve changing the structure of the data (Data Wrangling)? Explain. No! Just the number of attributes decreases.
Describe a situation in which the method would be useful. For instance, answering the question of if an individual has a healthy weight depends mostly on their weight and height. Physicians can give a chart so the healthy weight can be determined based on the people's height, or they can construct a new attribute based on both weight and height so the individual can know where they stand by calculating the new attribute. That new attribute is known as Body Mass Index (BMI).
Does the method have names that are used commonly? No.

Binary coding

What does the method do to the values of the data? This data transformation replaces categorical attributes with one or more binary attributes.
What is the objective of the method? Transforming categorical values into numerical ones. Essentially, making sure that the data is presented with numbers.
Does the method also involve changing the structure of the data (Data Wrangling)? Explain.

No! Just the number of attributes increases.
Describe a situation in which the method would be useful. Whenever we have categorical attributes that we need to use for algorithms that only accept numbers.
Does the method have names that are used commonly? No.

Ranking Transformation

What does the method do to the values of the data? This data transformation replaces categorical ordinal attributes with the ranking of categories.
What is the objective of the method? Transforming categorical ordinal values into numerical ones.
Does the method also involve changing the structure of the data (Data Wrangling)? Explain. No.
Describe a situation in which the method would be useful. Whenever we have categorical attributes that we need to use for algorithms that only accept numbers.
Does the method have names that are used commonly? It may also be referred to as just ranking too.

Discretization

What does the method do to the values of the data? This data transformation replaces numerical attributes with categorical ones.
What is the objective of the method? Making sure the numerical attributes are represented as categories.
Does the method also involve changing the structure of the data (Data Wrangling)? Explain. No.
Describe a situation in which the method would be useful. When the categorical attributes are preferred for the analytic goals instead of numerical ones. For instance, when presenting values to a human audience, they prefer categories over numbers.
Does the method have names that are used commonly? Binning.

Smoothing

What does the method do to the values of the data? The data transformation is used specifically for time series data where the consecutive data points have a close relationship with one another (autocorrelation effect). Using this relationship the data is transformed so the noise component is removed from the time series data.

What is the objective of the method? Removing noise from time series.
Does the method also involve changing the structure of the data (Data Wrangling)? Explain. No.
Describe a situation in which the method would be useful. Visualizing a time series and showing the smoothed trend instead of the actual data can help the audience understand the trend.
Does the method have names that are used commonly? No.

Aggregation

What does the method do to the values of the data? Restructure the data so the aggregation level is larger, for instance instead of hourly recordings, by aggregating over every day (summation or average as appropriate), we can get the daily aggregation level.
What is the objective of the method? As a data transformation tool, its goal is to remove noise from data, however, aggregation is also a data wrangling too.
Does the method also involve changing the structure of the data (Data Wrangling)? Explain.

Yes. Aggregation is a data-wrangling tool that may be used as a smoothing tool, but one has to reckon with the fact that the definition of the data object will change due to aggregation.
Describe a situation in which the method would be useful. Similar to the answer for smoothing.
Does the method have names that are used commonly? No.

Binning

What does the method do to the values of the data? See the answer for discretization. Binning is the same as discretization, however, their goal is different.
What is the objective of the method? Deal with noise.
Does the method also involve changing the structure of the data (Data Wrangling)? Explain. No.
Describe a situation in which the method would be useful. When there is noise in the data, binding transform the numerical attribute into a categorical one and thereby removes them drastically.
Does the method have names that are used commonly? Discretization.

Answers to general questions

Q1: What's the difference between Normalization and Standardization? Describe two different situations in that one would be preferable to the other.

Normalization attempts to control the range of the data and make sure the rescaled attribute stay within a prescribed range, however, standardization attempts to control the standard deviation of an attribute, making sure it is equal to 1.

It is best to use standardization for PCA, and normalization for clustering analysis.

Q2: What are the differences and similarities between Attribute Construction and Feature Extraction?

Both require a knowledge domain to be created.

Q3: Binary Coding and Ranking Transformation have something in common that is in contrast with Discretization, what is that?

They both attempt to transform categorical attributes into numerical ones. Discretization is the opposite.

Q4: What is the relationship between Smoothing and Aggregation?

They are both done for dealing with noise. However, aggregation may also be used as a data-wrangling method.

Q5: what is the relationship between Binning and Discretization?

Same data transformation with two separate goals; only the name is different. We use the name binding when the goal is to deal with noise, and we use the name discretization when the goal is to transform a numerical attribute into a categorical one.