Chapter 7

Analysis Subjects and Multiple Observations

7.1 Introduction

In *Chapter 5 – The Origin of Data*, we explored possible data sources from a technical point of view, and in *Chapter 6 – Data Models*, we discussed the data models and data structures that we might encounter when accessing data for analytics.

In this chapter we will cover the basic structure of our analysis table.

In the following sections we will look at two central elements of analytic data structures:

- the identification and definition of the analysis subject
- the determination of whether multiple observations per analysis subject exist and how they will be handled

Finally, we will also see that in some analysis tables, individual analysis subjects are not present, but aggregates of these subjects are analyzed.

7.2 Analysis Subject

Definition

Analysis subjects are entities that are being analyzed, and the analysis results are interpreted in their context. Analysis subjects are therefore the basis for the structure of our analysis tables.

The following are examples of analysis subjects:

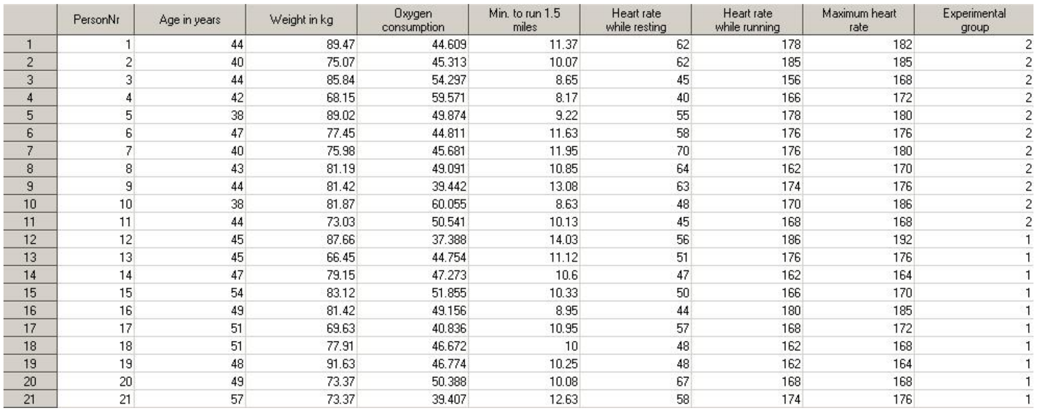
- Persons: Depending on the domain of the analysis, the analysis subjects have more specific names such as patients in medical statistics, customers in marketing analytics, or applicants in credit scoring.
- Animals: Piglets, for example, are analyzed in feeding experiments; rats are analyzed in pharmaceutical experiments.
- Parts of the body system: In medical research analysis subjects can also be parts of the body system such as arms (the left arm compared to the right arm), shoulders, or hips. Note that from a statistical point of view, the validity of the assumptions of the respective analysis methods has to be checked if dependent observations per person are used in the analysis.
- Things: Such as cash machines in cash demand prediction, cars in quality control in the automotive industry, or products in product analysis.
- Legal entities: Such as companies, contracts, accounts, and applications.
- Regions or plots in agricultural studies, or reservoirs in the maturity prediction of fields in the oil and gas industry.

Analysis subjects are the heart of each analysis because their attributes are measured, processed, and analyzed. In deductive (inferential) statistics the features of the analysis subjects in the sample are used to infer the properties of the analysis subjects of the population. Note that we use feature and attribute interchangeably here.

Representation in the data set

When we look at the analysis table that we want to create for our analysis, the analysis subjects are represented by rows, and the features that are measured per analysis subject are represented by columns. See Table 7.1 for an illustration.

Table 7.1: Results of ergonometric examinations for 21 runners



	PersonNr	Age in years	Weight in kg	Oxygen consumption	Min. to run 1.5 miles	Heart rate while resting	Heart rate while running	Maximum heart rate	Experimental group
1	1	44	89.47	44.609	11.37	62	178	182	2
2	2	40	75.07	45.313	10.07	62	185	185	2
3	3	44	85.84	54.297	8.65	45	156	168	2
4	4	42	68.15	59.571	8.17	40	166	172	2
5	5	38	89.02	49.874	9.22	55	178	180	2
6	6	47	77.45	44.811	11.63	58	176	176	2
7	7	40	75.98	45.681	11.95	70	176	180	2
8	8	43	81.19	49.091	10.85	64	162	170	2
9	9	44	81.42	39.442	13.08	63	174	176	2
10	10	38	81.87	60.055	8.63	48	170	186	2
11	11	44	73.03	50.541	10.13	45	168	168	2
12	12	45	87.66	37.388	14.03	56	186	192	1
13	13	45	66.45	44.754	11.12	51	176	176	1
14	14	47	79.15	47.273	10.6	47	162	164	1
15	15	54	83.12	51.855	10.33	50	166	170	1
16	16	49	81.42	49.156	8.95	44	180	185	1
17	17	51	69.63	40.836	10.95	57	168	172	1
18	18	51	77.91	46.672	10	48	162	168	1
19	19	48	91.63	46.774	10.25	48	162	164	1
20	20	49	73.37	50.388	10.08	67	168	168	1
21	21	57	73.37	39.407	12.63	58	174	176	1

In this table 21 runners have been examined, and each one is represented by one row in the analysis table. Features such as age, weight, and runtime, have been measured for each runner, and each feature is represented by a single column. Analyses, such as calculating the mean age of our population or comparing the runtime between experimental group 1 and 2, can directly start from this table.

Analysis subject identifier

A column PersonNr has been added to the table to identify the runner. Even if it is not used for analysis, the presence of an ID variable for the analysis subjects is important for the following reasons:

- data verifications and plausibility checks, if the original data in database queries or data forms have to be consulted
- the identification of the analysis subject if additional data per subject has to be added to the table
- if we work on samples and want to refer to the sampled analysis subject in the population

Also note that in some cases it is illegal, and in general it is against good analysis practice to add people's names, addresses, social security numbers, and phone numbers to analysis tables. The statistician is interested in data on analysis subjects, not in the personal identification of analysis subjects. If an anonymous subject number is not available, a surrogate key with an arbitrary numbering system has to be created for both the original data and the analysis data. The statistician in that case receives only the anonymous analysis data.

7.3 Multiple Observations

General

The analysis table in Table 7.1 is simple in that we have only one observation per analysis subject. It is therefore straightforward to structure the analysis table in this way.

There are, however, many cases where the situation becomes more complex; namely, when we have multiple observations per analysis subject.

Examples

- In the preceding example we will have multiple observations when each runner does more than one run, such as a second run after taking an isotonic drink.
- A dermatological study in medical research where different creams are applied to different areas of the skin.
- Evaluation of clinical parameters before and after surgery.
- An insurance customer with insurance contracts for auto, home, and life insurance.
- A mobile phone customer with his monthly aggregated usage data for the last 24 months.
- A daily time series of overnight stays for each hotel.

In general there are two reasons why multiple observations per analysis subject can exist:

- repeated measurements over time
- multiple observations because of hierarchical relationships

We will now investigate the properties of these two types in more detail.

7.3.1 Repeated Measurements over Time

Repeated measurements over time are obviously characterized by the fact that for the same analysis subject the observation is repeated over time. From a data model point of view, this means that we have a one-to-many relationship between the analysis subject entity and a time-related entity.

Note that we are using the term repeated measurement where observations are recorded repeatedly. We are not necessarily talking about measurements in the sense of numeric variables per observation of the same analysis subject—only the presence or absence of an attribute (yes or no) would be noted on each of X occasions.

The simplest form of repeated measurements is the two-observations-per-subject case. This case happens most often when comparing observations before and after a certain event and we are interested in the difference or change in certain criteria (pre-test and post-test). Examples of such an event include the following:

- giving a certain treatment or medication to patients
- execution of a marketing campaign to promote a certain product

If we have two or more repetitions of the measurement, we will get a measurement history or a time series of measurements:

- Patients in a clinical trial make quarterly visits to the medical center where laboratory values and vital signs values are collected. A series of measurement data such as the systolic and diastolic blood pressure can be analyzed over time.
- The number and duration of phone calls of telecommunications customers are available on a weekly aggregated basis.
- The monthly aggregated purchase history for retail customers.
- The weekly total amount of purchases using a credit card.
- The monthly list of bank branches visited by a customer.

The fact that we do not have only multiple observations per analysis subject, but ordered repeated observations, allows us to analyze their course over time such as by looking at trends. In Chapters 18–20 we will explore in detail how this information can be described and retrieved per analysis subject.

7.3.2 Multiple Observations because of Hierarchical Relationships

If we have multiple observations for an analysis subject because the subject has logically related child hierarchies, we call this *multiple observations because of hierarchical relationships*. The relation to the entity relationship diagrams that we introduced in Chapter 6 is that here we have so-called one-to-many relationships between the analysis subject and its child hierarchy. The following are examples:

- One insurance customer can have several types of insurance contracts (auto insurance, home insurance, life insurance). He can also have several contracts of the same type, e.g., if he has more than one car.
- A telecommunications customer can have several contracts; for each contract, one or more lines can be subscribed. (In this case we have a one-to-many relationship between the customer and contract and another one-to-many relationship between the contract and the line.)

- In one household, one or more persons can each have several credit cards.
- Per patient both eyes are investigated in an ophthalmological study.
- A patient can undergo several different examinations (laboratory, x-ray, vital signs) during one visit.

7.3.3 Multiple Observations Resulting from Combined Reasons

Multiple observations per analysis subject can also occur as a result of a combination of repeated measures over time and multiple observations because of hierarchical relationships:

- Customers can have different account types such as a savings account and a checking account. And for each account a transaction history is available.
- Patients can have visits at different times. At each visit, data from different examinations are collected.

If from a business point of view a data mart based on these relationships is needed, data preparation gets more complex, but the principles that we will see in “Data Mart Structures” remains the same.

7.3.4 Redefinition of the Analysis Subject Level

In some cases, the question arises whether we have useful multiple observations per analysis subject or whether we have to (are able to) redefine the analysis subject. Redefining the analysis subject means that we move from a certain analysis subject level, such as patient, to a more detailed one, such as shoulder.

The problem of redefining the analysis subject is that we then have dependent measures that might violate the assumptions of certain analysis methods. Think of a dermatological study where the effect of different creams applied to the same patient can depend on the skin type and are therefore not independent of each other. The decision about a redefinition of the analysis subject level requires domain-specific knowledge and a consideration of the statistically appropriate analysis method.

Besides the statistical correctness, the determination of the correct analysis subject level also depends on the business rationale of the analysis. The decision whether to model telecommunication customers on the customer or on the contract level depends on whether marketing campaigns or sales actions are planned and executed on the customer or contract level.

Note that so far we have only identified the fact that multiple observations can exist and the causal origins. We have not investigated how they can be considered in the structure of the analysis table. We will do this in the following section.

7.4 Data Mart Structures

In order to decide how to structure the analysis table for multiple observations, we will introduce the two most important structures for an analysis table, the one-row-per-subject data mart and the multiple-rows-per-subje*c*t data mart. In Chapters 8 and 9, respectively, we will discuss their properties, requirements, and handling of multiple observations in more detail.

7.4.1 One-Row-per-Subject Data Mart

In the one-row-per-subject data mart, all information per analysis subject is represented by one row. Features per analysis subject are represented by a column. When we have no multiple observations per analysis subject, the creation of this type of data mart is straightforward—the value of each variable that is measured per analysis subject is represented in the corresponding column. We saw this in the first diagram in Chapter 6. The one-row-per-subject data mart usually has only one ID variable, namely that of identifying the subjects.

Table 7.2: Content of CUSTOMER table

CustID	Birthdate	Gender
1	16.05.1970	Male
2	19.04.1964	Female

Table 7.3: Content of ACCOUNT table

AccountID	CustID	Type	OpenDate
1	1	Checking	05.12.1999
2	1	Savings	12.02.2001
3	2	Savings	01.01.2002
4	2	Checking	20.10.2003
5	2	Savings	30.09.2004

In the case of the presence of multiple observations per analysis subject, we have to represent them in additional columns. Because we are creating a one-row-per-subject data mart, we cannot create additional rows per analysis subject. See the following example.

Table 7.4: One-row-per-subject data mart for multiple observations

CustID	Birthdate	Gender	Number of Accounts	Proportion of Checking Accounts	Opendate of oldest account
1	16.05.1970	Male	2	50 %	05.12.1999
2	19.04.1964	Female	3	33 %	01.01.2002

Table 7.4 is the one-row-per-subject representation of Tables 7.2 and 7.3. We see that we have only two rows because we have only two customers. The variables from the CUSTOMER table have simply been copied to the table. When aggregating data from the ACCOUNT table, however, we experience a loss of information. We will discuss that in Chapter 8. Information from the underlying hierarchy of the ACCOUNT table has been aggregated to the customer level by completing the following tasks:

- counting the number of accounts per customer
- calculating the proportion of checking accounts
- identifying the open date of the oldest account

We have used simple statistics on the variables of ACCOUNT in order to aggregate the data per subject. More details about bringing all information into one row will be discussed in detail in *Chapter 8 – The One-Row-per-Subject Data Mart*.

7.4.2 The Multiple-Rows-per-Subject Data Mart

In contrast to the one-row-per-subject data mart, one subject can have multiple rows. Therefore, we need one ID variable that identifies the analysis subject and a second ID variable that identifies multiple observations for each subject. In terms of data modeling we have the child table of a one-to-many relationship with the foreign key of its master entity. If we also have information about the analysis subject itself, we have to repeat this with every observation for the analysis subject. This is also called de-normalizing.

- In the case of multiple observations because of hierarchical relationships, ID variables are needed for the analysis subject and the entities of the underlying hierarchy. See the following example of a multiple-rows-per-subject data mart. We have an ID variable CUSTID for CUSTOMER and an ID variable for the underlying hierarchy of the ACCOUNT table. Variables of the analysis subject such as birth date and gender are repeated with each account. In this case we have a de-normalized table as we explained in Chapter 6.

Table 7.5: Multiple-rows-per-subject data mart as a join of the CUSTOMER and ACCOUNT tables

CustID	Birthdate	Gender	AccountID	Type	OpenDate
1	16.05.1970	Male	1	Checking	05.12.1999
1	16.05.1970	Male	2	Savings	12.02.2001
2	19.04.1964	Female	3	Savings	01.01.2002
2	19.04.1964	Female	4	Checking	20.10.2003
2	19.04.1964	Female	5	Savings	30.09.2004

- In the case of repeated observations over time the repetitions can be enumerated by a measurement variable such as a time variable or, if we measure the repetitions only on an ordinal scale, by a sequence number. See the following example with PATNR as the ID variable for the analysis subject PATIENT. The values of CENTER and TREATMENT are repeated per patient because of the repeated measurements of CHOLESTEROL and TRIGLYCERIDE at each VISITDATE.

Table 7.6: Multiple-rows-per-subject data mart as a join of the CUSTOMER and ACCOUNT tables

PATNR	CENTER	TREATMENT	MEASUREMENT	VISITDATE	CHOLESTEROL	TRIGLYCERIDE
1	VIENNA	A	1	15JAN2002	220	220
1	VIENNA	A	2	20JUL2002	216	216
1	VIENNA	A	3	07JAN2002	205	205
2	SALZBURG	B	1	15APR2001	308	308
2	SALZBURG	B	2	01OCT2001	320	320

7.4.3 Summary of Data Mart Types

Table 7.7 summarizes how different data mart structures can be created, depending on the structure of the source data.

Table 7.7: Data mart types

	Data mart structure that is needed for the analysis	
Structure of the source data: "Multiple observations per analysis subject exist?"	**One-row-per-subject data mart**	**Multiple-rows-per-subject data mart**
NO	Data mart with one row per subject is created.	(Key-value table can be created.)
YES	Information of multiple observations has to be aggregated per analysis subject (see also Chapter 8).	Data mart with one-row-per-multiple observations is created. Variables at the analysis subject level are duplicated for each repetition (see also Chapter 9).

7.4.4 Using Both Data Mart Structures

There are analyses where data need to be prepared in both versions: the one-row-per-subject data mart and the multiple-rows-per subject data mart.

Consider the case where we have measurements of credit card activity per customer on a monthly basis.

- In order to do a segmentation analysis or prediction analysis on customer level we need the data in the form of a one-row-per-subject data mart.
- In order to analyze the course of the monthly transaction sum per customer over time, however, we need to prepare a multiple-rows-per-analysis subject data mart. In this data mart we will create line plots and calculated trends.
- The visual results of the line plots are input for the analyst to calculate derived variables for the one-row-per-subject data mart. Trends in the form of regression coefficients on the analysis subject level are calculated on the basis of the multiple-rows-per-subject data mart. These coefficients are then added to the one-row-per-subject data mart.

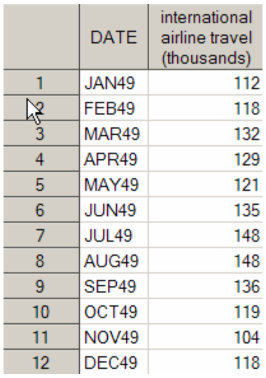
In the next two chapters we will take a closer look at the properties of one- and multiple-rows-per subject data marts. In *Chapter 14 – Transposing One- and Multiple-Rows-per-Subject Data Structures*, we will see how we can switch between different data mart structures.

7.5 No Analysis Subject Available?

General

In the preceding sections we dealt with cases where we were able to identify an analysis subject. We saw data tables where data for patients or customers were stored.

There are, however, analysis tables where we do not have an explicit analysis subject. Consider an example where we have aggregated data on a monthly level—for example, the number of airline travelers, which can be found in the SASHELP.AIR data set. This table is obviously an analysis table, which can be used directly for time series analysis. We do not, however, find an analysis subject in our preceding definition of it.



	DATE	international airline travel (thousands)
1	JAN49	112
2	FEB49	118
3	MAR49	132
4	APR49	129
5	MAY49	121
6	JUN49	135
7	JUL49	148
8	AUG49	148
9	SEP49	136
10	OCT49	119
11	NOV49	104
12	DEC49	118

We, therefore, have to refine the definition of analysis subjects and multiple observations. It is possible that in analysis tables we consider data on a level where information of analysis subjects is aggregated. The types of aggregations are in most cases counts, sums, or means.

Example

We will look at an example from the leisure industry. We want to analyze the number of overnight stays in Vienna hotels. Consider the following three analysis tables:

- Table that contains the monthly number of overnight stays per HOTEL
- Table that contains the monthly number of overnight stays per CATEGORY (5 stars, 4 stars …)
- Table that contains the monthly number of overnight stays in VIENNA IN TOTAL

The first table is a typical multiple-rows-per-subject table with a line for each hotel and month. In the second table we have lost our analysis subjects because we have aggregated, or summed, over them. The hotel category could now serve as a new "analysis subject," but it is more an analysis level than an analysis subject. Finally, the overnight stays in Vienna in total are on the virtual analysis level 'ALL,' and we have only one analysis subject, 'VIENNA'.

Longitudinal Data Structures

We have seen that as soon we start aggregating over an analysis subject, we come to analysis levels where the definition of an analysis subject is not possible or does not make sense. The aggregated information per category is often considered over time, so we have multiple observations per category over time.

These categories can be an aggregation level. We have seen the example of hotel categorization as 5 stars, 4 stars, and so on. The category at the highest level can also be called the ALL group, similar to a grand total.

Note that categorizations do not need to necessarily be hierarchical. They can also be two alternative categorizations such as hotel classification, as we saw earlier, and regional district in the preceding example. This requires that the analysis subject HOTEL have the properties classification and region, which allow aggregation of their number of overnight stays.

Data structures where aggregated data over time are represented either for categories or the ALL level are called *longitudinal data structures* or *longitudinal data marts*.

Strictly speaking the multiple-rows-per-subject data mart with repeated observations over time per analysis subject can also be considered a longitudinal data mart. The analytical methods that are applied to these data marts do not differ. The only difference is that in the case of the multiple-rows-per-subject data mart we have dependent observations per analysis subject, and in the case of longitudinal data structures we do not have an analysis subject in the classic sense.