# Dataset Description

## clients

This table contains clients which have had accounts with the bank. The table displays the current characteristics of the client.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Field name | Key | Datatype | Description | Possible values |
| client\_id | Y | integer | Unique ID of the client company | Positive integers |
| status |  | String | Status of the client; Active if it has at least one active account. | “ACTIVE” and “INACTIVE” |
| sector |  | integer | Sector code of the client company | 1 to 10 |

## transactions

This table contains any movement to and from the client’s current account. The data contains any movement between 2018-01-01 and 2021-12-31.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Field name | Key | Datatype | Description | Possible values |
| client\_id |  | integer | Unique ID of the client company | Positive integers |
| amount |  | double | Amount of the transaction (in €) | Positive values |
| date |  | date | Date of the transaction | Any date in the years 2018 to 2021 |
| direction |  | string | Direction of the transaction. | “in” or “out” |

## credit\_lines

This table contains information on the credit lines given by the bank to the corporate customers. The data contains any valid information between 2018-01-01 and 2021-12-31.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Field name | Key | Datatype | Description | Possible values |
| client\_id | Y | integer | Unique ID of the client company | Positive integers |
| credit\_line\_number | Y | integer | Unique ID of the credit line | Positive integers |
| available\_credit\_line |  | double | Credit line given by the bank (in €) | Positive values |
| active\_credit\_line |  | boolean | Whether the credit line is active or not. | True or False |
| Last\_update | Y | date | Date the information was updated. The information on the credit line is valid from this day onwards. | Any date in the years 2018 to 2021 |

A client can have none, one or many credit lines, denoted by “credit\_line\_number”. The credit line ID does not have to be a consecutive number, this means a client can have 3 credit lines with credit line numbers 1, 3 and 5. If the information on a credit line of a client has had an update, only the newest record is valid.

# Test

From the test data (csv files) execute the following exercises in your choice of data processing language (SQL, Python (any of the mayor data science libraries like Pandas), Pyspark or a Spark API in another language, or R).

## Exercise 1

Create a table with how many transactions and their volume (€) each client has had per month and per direction. Only keep *Active* clients. **Solution required: Table delivered in Excel**

## Exercise 2

Create a table with how many transactions in any direction as well as *net transaction sum* each sector has had per year. The *net transaction sum* is the sum of all incoming transactions minus all outgoing transactions. Consider all clients regardless of status. Mark in a Boolean field sectors with a net negative (TRUE; more outgoing transactions than incoming). **Solution required: Table delivered in Excel**

## Exercise 3

Get the most current information on each credit line of each client. **Solution required: Table delivered in Excel**

## Exercise 4

Get the sum of available activecredit lines per activeclient as of the 31.12.2020. **Solution required: Top 10 via the highest sum in an Excel table**

## Exercise 5

Get *active* clients who have a higher volume (€) of outgoing transactions more than 75% of their *active* credit lines.

* Consider only transactions of the year 2021.
* Ignore clients without outgoing transactions or without credit lines.
* You may use tables/datasets created in other exercises.
* Use most current credit line data.

**Solution required: Table delivered in Excel**