

## About the Dataset

The [Berka Dataset](#), or the PKDD'99 Financial Dataset, is a collection of real anonymized financial information from a Czech bank, used for PKDD'99 Discovery Challenge. The dataset can be accessed from my [GitHub page](#).

In the dataset, 8 raw files include 8 tables:

- **account** (4500 objects in the file ACCOUNT.ASC)—each record describes static characteristics of an account.
- **client** (5369 objects in the file CLIENT.ASC)—each record describes characteristics of a client.
- **disposition** (5369 objects in the file DISP.ASC)—each record relates together a client with an account i.e. this relation describes the rights of clients to operate accounts.
- **permanent order** (6471 objects in the file ORDER.ASC)—each record describes characteristics of a payment order.
- **transaction** (1056320 objects in the file TRANS.ASC)—each record describes one transaction on an account.
- **loan** (682 objects in the file LOAN.ASC)—each record describes a loan granted for a given account.
- **credit card** (892 objects in the file CARD.ASC)—each record describes a credit card issued to an account.
- **demographic data** (77 objects in the file DISTRICT.ASC)—each record describes demographic characteristics of a district.

loan	
loan_id	int
account_id	int
date	date
amount	int
duration	int
payments	decimal
status	varchar

order	
order_id	int
account_id	int
bank_to	varchar
account_to	int
amount	decimal
k_symbol	varchar

trans	
trans_id	int
account_id	int
date	date
type	varchar
operation	varchar
amount	int
balance	int
k_symbol	varchar
bank	varchar
account	int

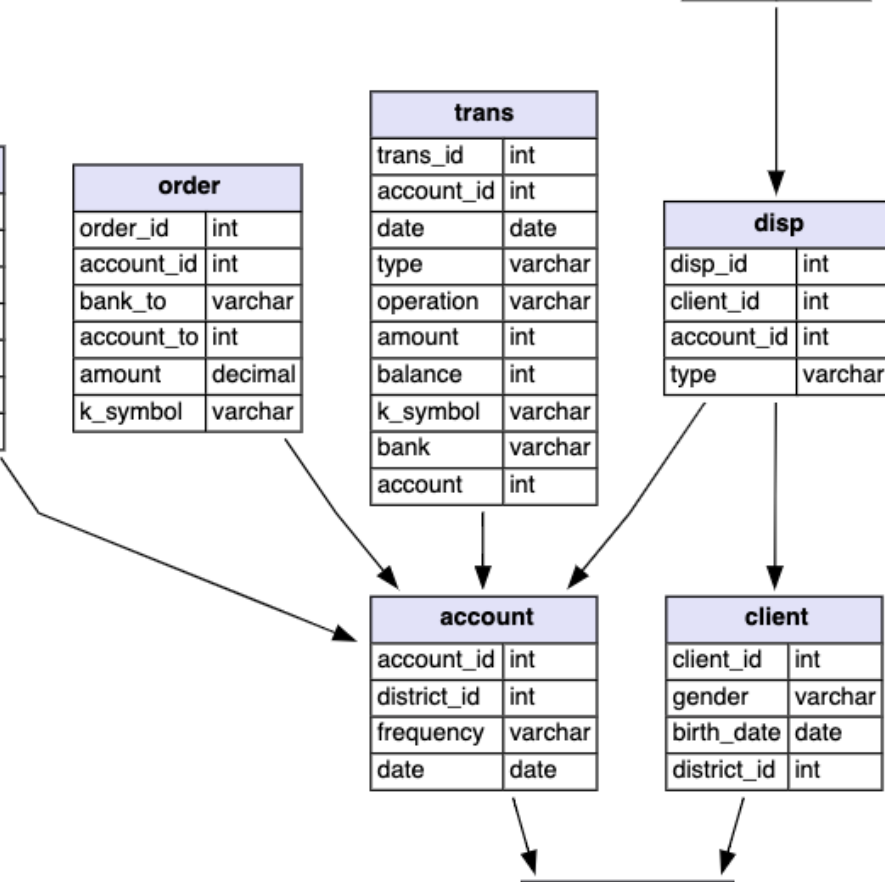
card	
card_id	int
disp_id	int
type	varchar
issued	date

disp	
disp_id	int
client_id	int
account_id	int
type	varchar

account	
account_id	int
district_id	int
frequency	varchar
date	date

client	
client_id	int
gender	varchar
birth_date	date
district_id	int

district	
district_id	int
A2	varchar
A3	varchar
A4	int
A5	int
A6	int
A7	int
A8	int
A9	int
A10	decimal
A11	int
A12	decimal
A13	decimal
A14	int
A15	int
A16	int



## Table Relationships

- Each account has both static characteristics (e.g. date of creation, address of the branch) given in relation “account” and dynamic characteristics (e.g. payments debited or credited, balances) given in the relations “permanent order” and “transaction”.
- Relation “client” describes the characteristics of persons who can manipulate the accounts.
- One client can have more accounts, more clients can manipulate with a single account; clients and accounts are related together in relation “disposition”.
- Relations “loan” and “credit card” describe some services which the bank offers to its clients.
- More than one credit card can be issued to an account,
- At most one loan can be granted for an account.
- Relation “demographic data” gives some publicly available information about the districts (e.g. the unemployment rate); additional information about the clients can be deduced from this.

### Account

COLUMN	DESCRIPTION	NOTES
account_id	identification of the account	
district_id	location of the branch	
date	Date of create of the account	In the form of YYMMDD
frequency	Frequency of issuance of statements	"POPLATEK MESICNE" stands for monthly issuance "POPLATEK TYDNE" stands for weekly issuance "POPLATEK PO OBRATU" stands for issuance after transaction

### Client

COLUMN	DESCRIPTION	NOTES
client_id	Client identifier	
birth_number	Birthday and sex	the number is in the form YYMMDD for men, the number is in the form YYMM+50DD for women, where YYMMDD is the date of birth
district_id	Address of the client	In the form of YYMMDD

### Disposition

COLUMN	DESCRIPTION	NOTES
disp_id	Record identifier	
client_id	Identification of a client	
account_id	Identification of an account	
type	Type of disposition (owner/user)	Only owner can issue permanent orders and ask for a loan

### Orders (Permanent Orders - Debit Only)

COLUMN	DESCRIPTION	NOTES
order_id	Record identifier	
account_id	account, the order is issued for	
bank_to	bank of the recipient	each bank has unique two-letter code
account_to	account of the recipient	
amount	debited amount	
K_symbol	type of the payment	"POJISTNE" stands for insurance payment "SIPO" stands for household payment "LEASING" stands for leasing "UVER" stands for loan payment

## Transactions

COLUMN	DESCRIPTION	NOTES
trans_id	record identifier	
account_id	account, the transation deals with	
date	date of transaction	in the form YYMMDD
type	+/- transaction	"PRIJEM" stands for credit "VYDAJ" stands for withdrawal
operation	mode of transaction	"VYBER KARTOU" credit card withdrawal "VKLAD" credit in cash "PREVOD Z UCTU" collection from another bank "VYBER" withdrawal in cash "PREVOD NA UCET" remittance to another bank
amount		
balance	balance after transaction	"POJISTNE" stands for insurrance payment "SLUZBY" stands for payment for statement "UROK" stands for interest credited "SANKC. UROK" sanction interest if negative balance "SIPO" stands for household "DUCHOD" stands for old-age pension "UVER" stands for loan payment
k_symbol	characterization of the transaction	
bank	bank of the partner	each bank has unique two-letter code
account	account of the partner	

## Loan

COLUMN	DESCRIPTION	NOTES
loan_id	Record identifier	
account_id	identification of the account	
date	date when the loan was granted	in the form YYMMDD
amount		
duration	duration of the loan	
payments	monthly payments	
status	status of paying off the loan	'A' stands for contract finished, no problems, 'B' stands for contract finished, loan not payed, 'C' stands for running contract, OK so far, 'D' stands for running contract, client in debt

## Credit Card

COLUMN	DESCRIPTION	NOTES
card_id	Record identifier	
disp_id	disposition to an account	
type	type of card	possible values are "junior", "classic", "gold"
issued	issue date	in the form YYMMDD

## Demographic Data

COLUMN	DESCRIPTION
A1 = district_id	District code
A2	District name
A3	Region
A4	no. Of inhabitants
A5	no. of municipalities with inhabitants < 499
A6	no. of municipalities with inhabitants 500-1999
A7	no. of municipalities with inhabitants 2000-9999
A8	no. of municipalities with inhabitants >10000
A9	no. of cities
A10	ratio of urban inhabitants
A11	average salary
A12	unemployment rate '95
A13	unemployment rate '96
A14	no. of entrepreneurs per 1000 inhabitants
A15	no. of committed crimes '95
A16	no. of committed crimes '96