Preparando a base de dados

Caio Stabel, Guilherme Sant'anna & Rodrigo Giannotti

```r
tables = c()

table_names <- c('account',
                 'card',
                 'client',
                 'disp',
                 'district',
                 'loan',
                 'order',
                 'trans')

folder <- 'czech_data'

i = 1
table_paths = c()

for (item in table_names) {
  table_paths[i] <-  paste0(folder, '/', item, '.asc')
  i = i+1
}
```

Antes de qualquer trabalho poder ser feito primeiro precisamos nos certificar que as tabelas sejam corretamente lidas pelo R e que as variáveis estejam em tipos não só coerentes como compreensíveis i.e. não em tcheco. Isso será feito neste script. A falta de comentários se da pois tudo o que está sendo feito segue as premissas dadas nas instruções.

TB_ACCOUNT

```
## Parsed with column specification:
## cols(
##   account_id = col_double(),
##   district_id = col_double(),
##   frequency = col_character(),
##   date = col_double()
## )
```

```
tb_account %<>% mutate(date = ymd(date  + 19000000))
tb_account %<>% rename(account_open_date = date)
```

Frequency

```
tb_account %<>% mutate(frequency = if_else(
  frequency == "POPLATEK MESICNE",'monthly',if_else(
    frequency == "POPLATEK TYDNE",'weekly','transaction'
    )))
```

```
## # A tibble: 6 x 4
##   account_id district_id frequency account_open_date
##        <dbl>       <dbl> <chr>     <date>
## 1        576          55 monthly   1993-01-01
## 2       3818          74 monthly   1993-01-01
## 3        704          55 monthly   1993-01-01
## 4       2378          16 monthly   1993-01-01
## 5       2632          24 monthly   1993-01-02
## 6       1972          77 monthly   1993-01-02
```

TB_CLIENT

```
## Parsed with column specification:
## cols(
##   client_id = col_double(),
##   birth_number = col_double(),
##   district_id = col_double()
## )
```

Sex / Birth Date

```
tb_client %<>%  mutate(sex = ifelse(
  (birth_number %% 10000) >= 5000, 'F', 'M'
  ), birth_number = ifelse(
  (birth_number %% 10000) >= 5000, birth_number - 5000, birth_number
  ),
  birth_date = ymd(birth_number + 19000000))
```

```
tb_client %>% head()
```

```
## # A tibble: 6 x 5
##   client_id birth_number district_id sex   birth_date
##       <dbl>        <dbl>       <dbl> <chr> <date>
## 1         1       701213          18 F     1970-12-13
## 2         2       450204           1 M     1945-02-04
## 3         3       401009           1 F     1940-10-09
## 4         4       561201           5 M     1956-12-01
## 5         5       600703           5 F     1960-07-03
## 6         6       190922          12 M     1919-09-22
```

TB_DISP

```
## Parsed with column specification:
## cols(
##   disp_id = col_double(),
##   client_id = col_double(),
##   account_id = col_double(),
##   type = col_character()
## )
```

```
tb_disp %<>% rename(disp_type = type)
```

```
## # A tibble: 6 x 4
##   disp_id client_id account_id disp_type
##     <dbl>     <dbl>      <dbl> <chr>
## 1       1         1          1 OWNER
## 2       2         2          2 OWNER
## 3       3         3          2 DISPONENT
## 4       4         4          3 OWNER
## 5       5         5          3 DISPONENT
## 6       6         6          4 OWNER
```

TB_ORDER

```
## Parsed with column specification:
## cols(
##   order_id = col_double(),
##   account_id = col_double(),
##   bank_to = col_character(),
##   account_to = col_double(),
##   amount = col_double(),
##   k_symbol = col_character()
## )
```

```
tb_order %<>% rename(order_bank = bank_to)
tb_order %<>% rename(order_account_to = account_to)
tb_order %<>% rename(order_amount = amount)
```

K_symbol

```
tb_order %<>% mutate(k_symbol = if_else(
  k_symbol == "POJISTNE", 'insurrance payment', if_else(
  k_symbol == "SIPO", 'household payment', if_else(
  k_symbol == "LEASING", 'leasing payment', if_else(
  k_symbol == "UVER", 'loan payment',
  'not informed'
)))))

tb_order %<>% rename(order_k_symbol = k_symbol)
```

```
## # A tibble: 6 x 6
##   order_id account_id order_bank order_account_to order_amount order_k_symbol
##      <dbl>      <dbl> <chr>                 <dbl>        <dbl> <chr>
## 1    29401          1 YZ                 87144583        2452  household payment
## 2    29402          2 ST                 89597016        3373. loan payment
## 3    29403          2 QR                 13943797        7266  household payment
## 4    29404          3 WX                 83084338        1135  household payment
## 5    29405          3 CD                 24485939         327  not informed
## 6    29406          3 AB                 59972357        3539  insurrance payme~
```

TB_TRANSACTION

```
## Parsed with column specification:
## cols(
##   trans_id = col_double(),
##   account_id = col_double(),
##   date = col_double(),
##   type = col_character(),
##   operation = col_character(),
##   amount = col_double(),
##   balance = col_double(),
##   k_symbol = col_character(),
##   bank = col_character(),
##   account = col_double()
## )
```

```
tb_trans %<>% rename(trans_date = date)
tb_trans %<>% mutate(trans_date = ymd(trans_date  + 19000000))
tb_trans %<>% rename(trans_bank = bank)
tb_trans %<>% rename(trans_account = account)
tb_trans %<>% rename(trans_amount = amount)
tb_trans %<>% rename(trans_balance = balance)
```

Type

```
tb_trans %<>% mutate(type = if_else(
 type == "PRIJEM", 'credit', if_else(
 type == "VYDAJ",'withdrawal',
 NULL
)))

tb_trans %<>% rename(trans_type = type)
```

Operation

```
tb_trans %<>% mutate(operation = if_else(
  operation == "VYBER KARTOU", 'credit withdrawal', if_else(
  operation == "VKLAD", 'credit cash', if_else(
  operation == "PREVOD Z UCTU", 'collection', if_else(
  operation == "VYBER", 'withdrawal', if_else(
  operation == "PREVOD NA UCET", 'remittance',
  NULL
))))))

tb_trans %<>% rename(trans_operation = operation)
```

K_symbol

```
tb_trans %<>% mutate(k_symbol = if_else(
  k_symbol == "POJISTNE", 'insurrance payment', if_else(
  k_symbol == "SLUZBY", 'statement payment', if_else(
  k_symbol == "UROK", 'interest credited', if_else(
  k_symbol == "SANKC. UROK", 'sanction interest', if_else(
  k_symbol == "SIPO", 'household', if_else(
  k_symbol == "DUCHOD", 'pension', if_else(
  k_symbol == "UVER", 'loan payment',
  'not informed'
))))))))

tb_trans %<>% rename(trans_k_symbol = k_symbol)
```

```
## # A tibble: 6 x 10
##   trans_id account_id trans_date trans_type trans_operation trans_amount
##      <dbl>      <dbl> <date>     <chr>      <chr>                  <dbl>
## 1   695247       2378 1993-01-01 credit     credit cash              700
## 2   171812        576 1993-01-01 credit     credit cash              900
## 3   207264        704 1993-01-01 credit     credit cash             1000
## 4  1117247       3818 1993-01-01 credit     credit cash              600
## 5   579373       1972 1993-01-02 credit     credit cash              400
## 6   771035       2632 1993-01-02 credit     credit cash             1100
## # ... with 4 more variables: trans_balance <dbl>, trans_k_symbol <chr>,
## #   trans_bank <chr>, trans_account <dbl>
```

TB_LOAN

```
## Parsed with column specification:
## cols(
##   loan_id = col_double(),
##   account_id = col_double(),
##   date = col_double(),
##   amount = col_double(),
##   duration = col_double(),
##   payments = col_double(),
##   status = col_character()
## )
```

```
tb_loan %<>% mutate(date = ymd(date + 19000000))
tb_loan %<>% rename(loan_date = date)
tb_loan %<>% rename(loan_amount = amount)
tb_loan %<>% rename(loan_duration = duration)
tb_loan %<>% rename(loan_payments = payments)
```

Status

```
tb_loan %<>% mutate(loan_status_desc = if_else(
  status == "A", 'no problems', if_else(
  status == "B", 'not payed', if_else(
  status == "C", 'OK so far', if_else(
  status == "D", 'client in debt',
  NULL
)))))

tb_loan %<>% rename(loan_status = status)
```

```
## # A tibble: 6 x 8
##   loan_id account_id loan_date  loan_amount loan_duration loan_payments
##     <dbl>      <dbl> <date>           <dbl>         <dbl>         <dbl>
## 1    5314       1787 1993-07-05       96396            12          8033
## 2    5316       1801 1993-07-11      165960            36          4610
## 3    6863       9188 1993-07-28      127080            60          2118
## 4    5325       1843 1993-08-03      105804            36          2939
## 5    7240      11013 1993-09-06      274740            60          4579
## 6    6687       8261 1993-09-13       87840            24          3660
## # ... with 2 more variables: loan_status <chr>, loan_status_desc <chr>
```

TB_CARD

```
## Parsed with column specification:
## cols(
##   card_id = col_double(),
##   disp_id = col_double(),
##   type = col_character(),
##   issued = col_character()
## )
```

```
tb_card %<>% mutate(issued = as_date(ymd_hms(issued)))
tb_card %<>% rename(card_issue_date = issued)
```

```
## # A tibble: 6 x 4
##   card_id disp_id type    card_issue_date
##     <dbl>   <dbl> <chr>   <date>
## 1    1005    9285 classic 1993-11-07
## 2     104     588 classic 1994-01-19
## 3     747    4915 classic 1994-02-05
## 4      70     439 classic 1994-02-08
## 5     577    3687 classic 1994-02-15
## 6     377    2429 classic 1994-03-03
```

TB_DISTRICT

```
## Parsed with column specification:
## cols(
##   A1 = col_double(),
##   A2 = col_character(),
##   A3 = col_character(),
##   A4 = col_double(),
##   A5 = col_double(),
##   A6 = col_double(),
##   A7 = col_double(),
##   A8 = col_double(),
##   A9 = col_double(),
##   A10 = col_double(),
##   A11 = col_double(),
##   A12 = col_character(),
##   A13 = col_double(),
##   A14 = col_double(),
##   A15 = col_character(),
##   A16 = col_double()
## )
```

Column names

```
new_names <- c(
  'district_id',
  'district_name',
  'district_region',
  'district_inhabitants',
  'district_s_cities',
  'district_m_cities',
  'district_g_cities',
  'district_gg_cities',
  'district_ncities',
  'district_urban_rate',
  'district_avg_sal',
  'district_unemployment95',
  'district_unemployment96',
  'district_entrepeneur_rate',
  'district_crimes95',
  'district_crimes96'
)

colnames(tb_district) <- new_names
```

```
## # A tibble: 6 x 16
##   district_id district_name district_region district_inhabi~ district_s_citi~
##         <dbl> <chr>         <chr>                      <dbl>            <dbl>
## 1           1 Hl.m. Praha   Prague                   1204953                0
## 2           2 Benesov       central Bohemia            88884               80
## 3           3 Beroun        central Bohemia            75232               55
## 4           4 Kladno        central Bohemia           149893               63
## 5           5 Kolin         central Bohemia            95616               65
## 6           6 Kutna Hora    central Bohemia            77963               60
## # ... with 11 more variables: district_m_cities <dbl>, district_g_cities <dbl>,
## #   district_gg_cities <dbl>, district_ncities <dbl>,
## #   district_urban_rate <dbl>, district_avg_sal <dbl>,
## #   district_unemployment95 <chr>, district_unemployment96 <dbl>,
## #   district_entrepeneur_rate <dbl>, district_crimes95 <chr>,
## #   district_crimes96 <dbl>
```