

Data Manipulation and Cleaning work

Account Table

1. Convert the Date attribute into a yyyy-mm-dd by adding **24** in year format in Excel or SQL

- 1993 -> 2017
- 1994 -> 2018
- 1995 -> 2019
- 1996 -> 2020
- 1997 -> 2021

SQL Method:

```
SELECT DATE_ADD(date, INTERVAL 24 YEAR) AS NewDate
```

```
FROM account;
```

2. Replace in frequency attribute “POPLATEK MESICNE” AS **Monthly Issuance**, “POPLATEK TYDNE” AS **Weekly Issuance**, and “POPLATEK POBRATU” AS **Issuance After a Transaction** in Excel or create a case statement in SQL.

A.

```
select
  case frequency
    when 'POPLATEK MESICNE' then 'Monthly Issuance'
    when 'POPLATEK TYDNE' then 'Weekly Issuance'
    when 'POPLATEK PO OBRATU' then 'Issuance After a Transaction'
    else frequency
  end as new_frequency
from account;
```

3. Create a Custom Column Card_Assigned and assign below :

- Silver -> Monthly issuance
- Diamond - weekly issuance
- Gold - Issuance after a transaction

Creating the custom column:

```
ALTER TABLE account:
```

```
ADD `Card_Assigned` VARCHAR(50);
```

```
SELECT
```

```
  Card_Assigned,
```

CASE Card_Assigned

WHEN 'Silver' THEN 'Monthly Issuance'

WHEN 'Diamond' THEN 'Weekly Issuance'

WHEN 'Gold' THEN 'Issuance after a transaction'

ELSE "

END AS Custom_Column

FROM account;

CARD Table:

1. Replace type attribute value “junior” as Silver, “Classic” as Gold, And “Gold” as Diamond by using replace in Excel or by using update in SQL.

Solution.

In Excel, you can use the "Find and Replace" feature to replace the "type" attribute values. Here are the steps:

- Open your Excel file and navigate to the column containing the "type" attribute values.
- Select the cells in the column that you want to replace.
- Press "Ctrl + H" or go to the "Home" tab and click on "Find & Select" and then "Replace."
- In the "Find what" field, enter the value you want to find (e.g., "junior").
- In the "Replace with" field, enter the new value you want to assign (e.g., "Silver").
- Click "Replace All" to replace all occurrences of "junior" with "Silver" in the selected cells.
- Repeat above process for each value you want to replace.
- Follow the same steps for the other replacements:

To replace "Classic" with "Gold":

Find what: Classic

Replace with: Gold

To replace "Gold" with "Diamond":

Find what: Gold

Replace with: Diamond

2. Convert issued attribute into yyyy-mm-dd adding 23 in year.

UPDATE card

SET issued = DATE_ADD(STR_TO_DATE(issued, '%y%m%d %H:%i:%s'), INTERVAL 23 YEAR);

CLIENT Table

1. Convert bith_number attribute to yyyy-mm-dd format and also create another column named sex by applying in bith_number 0 for females and 1 for males.

(=if(mod(bith_number,2)=0, “Female”, “Male”) in excel or using case statement in SQL.

For **Male** its in **YYMMDD** format and for female it is **YYMM+50DD** for **Women**

SELECT

CASE

WHEN MOD(birth_number, 2) = 0 THEN
DATE_FORMAT(STR_TO_DATE(CONCAT('20', SUBSTRING(birth_number, 1, 6)), '%Y%m%d'), '%Y-%m-%d')

ELSE DATE_FORMAT(STR_TO_DATE(CONCAT('20', SUBSTRING(birth_number, 1, 4), '+50', SUBSTRING(birth_number, 5, 2)), '%Y%m+%d'), '%Y-%m-%d')

END AS birth_date,

CASE

WHEN MOD(birth_number, 2) = 0 THEN 'Female'

ELSE 'Male'

END AS sex

FROM client;

DISTRICT Table

1. Change all column names and delete the attributes **a12** and

2.

district.csv

Demographic

COLUMN NAME	TYPE	DESCRIPTION
# a1 ⓘ	integer	district code
T a2 ⓘ	string	district name
T a3 ⓘ	string	region
# a4 ⓘ	integer	no. of inhabitants
# a5 ⓘ	integer	no. of municipalities with inhabitants < 499
# a6 ⓘ	integer	no. of municipalities with inhabitants 500-1999
# a7 ⓘ	integer	no. of municipalities with inhabitants 2000-9999
# a8 ⓘ	integer	no. of municipalities with inhabitants > 10000
# a9 ⓘ	integer	no. of cities
# a10 ⓘ	decimal	ratio of urban inhabitants
# a11 ⓘ	integer	average salary
# a12 ⓘ	decimal	unemployment rate '95
# a13 ⓘ	decimal	unemployment rate '96

ALTER TABLE district

CHANGE COLUMN A1 `a1` INT COMMENT 'districtcode',

CHANGE COLUMN A2 `a2` VARCHAR(50) COMMENT 'districtname',

CHANGE COLUMN A3 `a3` VARCHAR(50) COMMENT 'region',

CHANGE COLUMN A4 `a4` INT COMMENT 'no of inhabitants',

CHANGE COLUMN A5 `a5` INT COMMENT 'no of municipalities with inhabitants<499',

CHANGE COLUMN A6 `a6` INT COMMENT 'no of municipalities with
inhabitants>500<1999',

CHANGE COLUMN A7 `a7` INT COMMENT 'no of municipalities with
inhabitants>2000<9999',

CHANGE COLUMN A8 `a8` INT COMMENT 'no of municipalities with
inhabitants>10000',

```

CHANGE COLUMN A9 `a9` INT COMMENT 'no of cities',
CHANGE COLUMN A10 `a10` DECIMAL COMMENT 'no of urban inhabitants',
CHANGE COLUMN A11 `a11` INT COMMENT 'avg salary',
CHANGE COLUMN A12 `a12` DECIMAL COMMENT 'unemployment rate 1995',
CHANGE COLUMN A13 `a13` DECIMAL COMMENT 'unemployment rate 1996',
CHANGE COLUMN A14 `a14` INT COMMENT 'no of entrepreneurs per 1000 inhabitants',
CHANGE COLUMN A15 `a15` INT COMMENT 'no of committed crimes 1995',
CHANGE COLUMN A16 `a16` INT COMMENT 'no of committed crimes 1996';

```

LOAN Table

1. Convert the Date Attribute into yyyy-mm-dd format adding 23 in year.

```

SELECT DATE_ADD(date, INTERVAL 23 YEAR) AS NewDate
FROM loan;

```

2. Convert Status Attribute value “A” as Contract Finished, “B” as Loan Not Paid, “C” as Running Contract, and “D” Client in debt.

Solution.

```

SELECT
    CASE Status
        WHEN 'A' THEN 'Contract Finished'
        WHEN 'B' THEN 'Loan Not Paid'
        WHEN 'C' THEN 'Running Contract'
        WHEN 'D' THEN 'Client in debt'
        ELSE 'Unknown Status'
    END AS StatusDescription
FROM loan;

```

In th Transactions Table do the following , whosoever count is highest sort it in descending order and change the year from 2022,2021,2020 and so on

--DATA TRANSFORMATION

/*

2021 -> 2017

2020 -> 2018

2019 -> 2019 -- NO CHANGE

2018 -> 2020

2017 -> 2021

2016 -> 2022

UPDATE TRANSACTIONS

SET BANK = 'Sky Bank' WHERE BANK IS NULL AND YEAR(DATE) = 2022;

UPDATE TRANSACTIONS

SET BANK = 'DBS Bank' WHERE BANK IS NULL AND YEAR(DATE) = 2021;

Solution:

UPDATE TRANSACTIONS

SET DATE = CASE

WHEN YEAR(DATE) = 2021 THEN DATE_SUB(DATE, INTERVAL 4 YEAR)

WHEN YEAR(DATE) = 2020 THEN DATE_SUB(DATE, INTERVAL 2 YEAR)

WHEN YEAR(DATE) = 2018 THEN DATE_SUB(DATE, INTERVAL 2 YEAR)

WHEN YEAR(DATE) = 2017 THEN DATE_ADD(DATE, INTERVAL 4 YEAR)

WHEN YEAR(DATE) = 2016 THEN DATE_ADD(DATE, INTERVAL 6 YEAR)

ELSE DATE

END;

-- Step 2: Update 'Sky Bank' for NULL BANK and YEAR(DATE) = 2022

UPDATE TRANSACTIONS

SET BANK = 'Sky Bank'

WHERE BANK IS NULL AND YEAR(DATE) = 2022;

-- Step 3: Update 'DBS Bank' for NULL BANK and YEAR(DATE) = 2021

UPDATE TRANSACTIONS

SET BANK = 'DBS Bank'

WHERE BANK IS NULL AND YEAR(DATE) = 2021;