Credit Risk DA Project

Database Connection

Download the DBeaver SQL client to connect to the MySQL database:

- https://dbeaver.io/

Follow the documentation to set up a connection to the database:

- https://dbeaver.com/docs/wiki/Create-Connection/

The database is hosted on AWS, here are the connection details:

- Endpoint: home-credit-default-risk.c7rizeij2t53.ap-southeast-1.rds.amazonaws.com
- Port: 3306
- Database: credit
- Login User: student
- Login Password: student

Overview

Consider you are asked to review a list of loan applications. The given “credit” database contains data on the loan applicant and their historical loan behavior. There are many columns in the database, you **don’t need to use all the columns**, We will provide a list of useful column descriptions for you.

Cautions

Missing Values:

There are columns with missing values. You need to handle them during your analysis. There are multiple ways we can handle missing values: 4 Ways to Replace NULL with a Different Value in MySQL

Discretization:

Discretization means we want to convert numbers into bins, for example, age to age groups or income to income groups. There are mainly 2 reasons for this:

- It is easier to see patterns with a group of values. For example, it is better to say people older than 20 are richer than people younger than 20, instead of saying people aged 20 are richer than people aged 21.
- We want to avoid biased statistics. If we apply group by aggregation directly on a number column like age, the average statistics can be biased. For example, if there is only 1 person aged 59, then the average income of people aged 59 only represents that 1 person in the dataset.

We can do it with the CASE Function in MySQL:
MySQL CASE Function
During the analysis, you can consider converting some factors into groups.

Task 1 Run SQL via DBeaver

Follow the documentation to open the “SQL Editor”:

- https://dbeaver.com/docs/wiki/SQL-Editor/

Run SQL to examine the number of rows in each table:

Table	**Count**
application	307,511
bureau	1,716,428

Loan Applications

The “application” table stores the loan applications. This includes:

- The demographic of the loan applicants
- The loan size or purposes
- The applicant’s credit score
- Is the loan applicant has a payment difficulties with the loan.

SK_ID_CURR	ID of the loan in our sample
TARGET	Target variable, this is the **future information**. Will this loan applicant has payment difficulties? (1: client with payment difficulties: he/she had late payment more than X days, 0: no payment difficulties)
CODE_GENDER	Gender of the client
FLAG_OWN_CAR	Flag if the client owns a car
FLAG_OWN_REALTY	Flag if the client owns a house or flat
CNT_CHILDREN	Number of children the client has
AMT_INCOME_TOTAL	Income of the client

AMT_CREDIT	Credit amount of the loan
AMT_ANNUITY	Loan annuity
AMT_GOODS_PRICE	For consumer loans it is the price of the goods for which the loan is given
NAME_TYPE_SUITE	Who was accompanying client when he was applying for the loan
NAME_INCOME_TYPE	Clients income type (businessman, working, maternity leave,…)
NAME_EDUCATION_TYPE	Level of highest education the client achieved
NAME_FAMILY_STATUS	Family status of the client
NAME_HOUSING_TYPE	What is the housing situation of the client (renting, living with parents, ...)
DAYS_BIRTH	Client's age in days at the time of application
DAYS_EMPLOYED	How many days before the application the person started current employment
OCCUPATION_TYPE	What kind of occupation does the client have
EXT_SOURCE_1	Normalized credit score from an external data source
EXT_SOURCE_2	Normalized credit score from an external data source
EXT_SOURCE_3	Normalized credit score from an external data source

Task 2 What is a Credit Score

In the “application” table above there are 3 credit score columns. Research online to see what is a credit score and why we need it. (Note that the scores in the database are normalized, which means they are scaled to the 0 to 1 range)

- A credit score is a number from 300 to 850 that depicts a consumer’s creditworthiness.
- The higher the score, the better a borrower looks to potential lenders.
- Credit score is based on credit history: number of open accounts, total levels of debt, repayment history, and other factors.
- Lenders use credit scores to evaluate the probability that an individual will repay loans in a timely manner. It plays a key role in a lender’s decision to offer you credit.
- The average FICO Score range is often used:
    1. Excellent: 800–850
    2. Very Good: 740–799
    3. Good: 670–739
    4. Fair: 580–669
    5. Poor: 300–579

6.

Task 3 Understand Credit Amount and Annuity

What are Credit Amount and Annuity? Fill in your answer below:

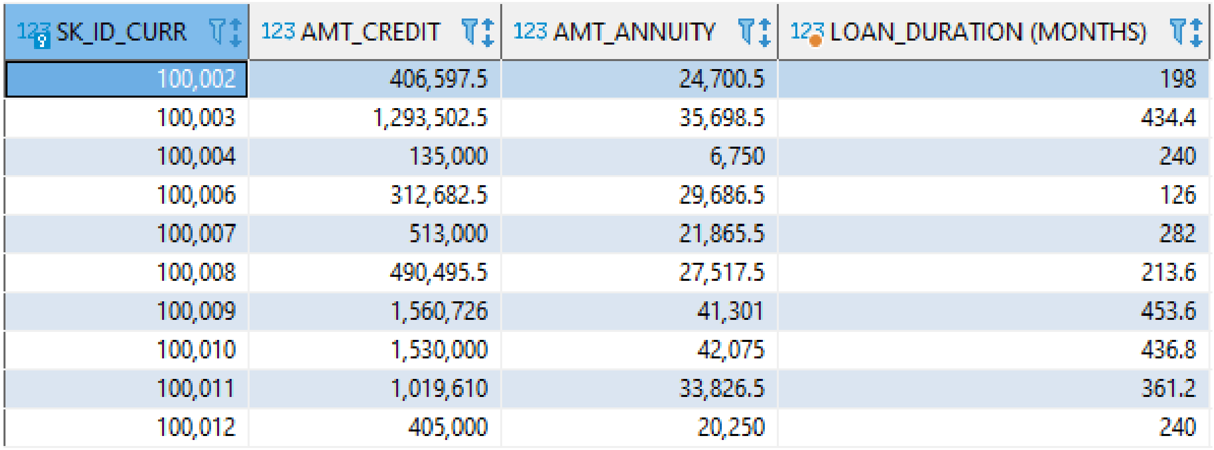
Credit Amount	• Credit Amount means the maximum amount that Lender is committed to lend • The amount of money loaned, according to your needs at any given time.
Annuity	• An annuity is a series of payments made at equal intervals • Examples of annuities are regular deposits to a savings account, monthly home mortgage payments, monthly insurance payments and pension payments. • Annuities can be classified by the frequency of payment dates. • The payments (deposits) may be made weekly, monthly, quarterly, yearly, or at any other regular interval of time.

Task 4 Deduce the Loan Duration

Given the information from Task 4, we should be able to deduce the Loan Duration for each application. Loan duration describes how many periods (months) the applicant will need to pay back their loans.

Paste the SQL and part of the results below:

```
SELECT
      SK_ID_CURR,
      AMT_CREDIT,
      AMT_ANNUITY,
      ROUND(AMT_CREDIT/AMT_ANNUITY, 1) * 12 AS "LOAN_DURATION (MONTHS)"
FROM application
```



SK_ID_CURR	AMT_CREDIT	AMT_ANNUITY	LOAN_DURATION (MONTHS)
100,002	406,597.5	24,700.5	198
100,003	1,293,502.5	35,698.5	434.4
100,004	135,000	6,750	240
100,006	312,682.5	29,686.5	126
100,007	513,000	21,865.5	282
100,008	490,495.5	27,517.5	213.6
100,009	1,560,726	41,301	453.6
100,010	1,530,000	42,075	436.8
100,011	1,019,610	33,826.5	361.2
100,012	405,000	20,250	240

Task 5 Are there any factors in the application table affecting the Credit Scores?

In the “application” table try to explore if there are any columns affecting the credit score. For example, is gender a factor?

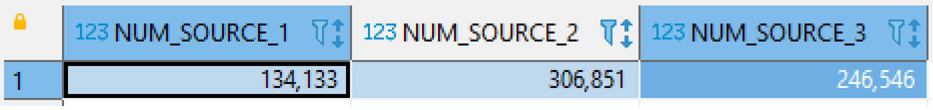
Do the analysis of at least 3 factors for 3 different credit scores, it is expected to see different results for different credit scores, for example, a factor might affect EXT_SOURCE_1 but not EXT_SOURCE_3.

Please explain your findings with SQL statements and results:

1. PRE-ANALYSIS

A. To calculate the number of rows with data in EXT_SOURCE_1, EXT_SOURCE_2, EXT_SOURCE_3 columns from **application** table

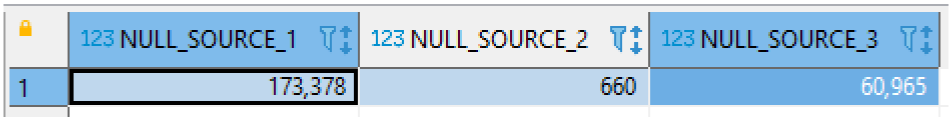
```
SELECT
       COUNT(EXT_SOURCE_1) AS "NUM_SOURCE_1",
       COUNT(EXT_SOURCE_2) AS "NUM_SOURCE_2",
       COUNT(EXT_SOURCE_3) AS "NUM_SOURCE_3"
       FROM application
```



	NUM_SOURCE_1	NUM_SOURCE_2	NUM_SOURCE_3
1	134,133	306,851	246,546

B. To calculate the number of NULL values in EXT_SOURCE 1, EXT_SOURCE_2, EXT_SOURCE_3 columns from **application** table

```
SELECT
(SELECT COUNT(IFNULL(EXT_SOURCE_1, 'N/A')) FROM application where EXT_SOURCE_1 IS NULL)
AS "NULL_SOURCE_1",
(SELECT COUNT(IFNULL(EXT_SOURCE_2, 'N/A')) FROM application WHERE EXT_SOURCE_2 IS NULL)
AS "NULL_SOURCE_2",
(SELECT COUNT(IFNULL(EXT_SOURCE_3, 'N/A')) FROM application WHERE EXT_SOURCE_3 IS NULL)
AS "NULL_SOURCE_3"
```



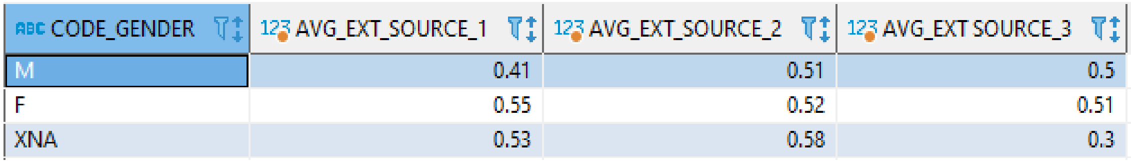
	NULL_SOURCE_1	NULL_SOURCE_2	NULL_SOURCE_3
1	173,378	660	60,965

Percentage of NULL %	EXT_SOURCE_1	EXT_SOURCE_2	EXT_SOURCE_3
	56.38%	0.21%	19.83%

2. Factors affecting different credit scores (normalized)

A. CODE_GENDER : Gender of the client

```
SELECT
      CODE_GENDER,
      ROUND(AVG(EXT_SOURCE_1),2) AS "AVG_EXT_SOURCE_1",
      ROUND(AVG(EXT_SOURCE_2),2) AS "AVG_EXT_SOURCE_2",
      ROUND(AVG(EXT_SOURCE_3),2) AS "AVG_EXT SOURCE_3"
FROM application
      GROUP BY CODE_GENDER
```



CODE_GENDER	AVG_EXT_SOURCE_1	AVG_EXT_SOURCE_2	AVG_EXT SOURCE_3
M	0.41	0.51	0.5
F	0.55	0.52	0.51
XNA	0.53	0.58	0.3

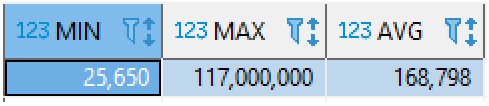
Result :

- Observation shows that on average, female client has better credit scores than male client.

B. AMT_INCOME_TOTAL : the total income of the client

- **To check the min, max and average income of the clients in the database.**

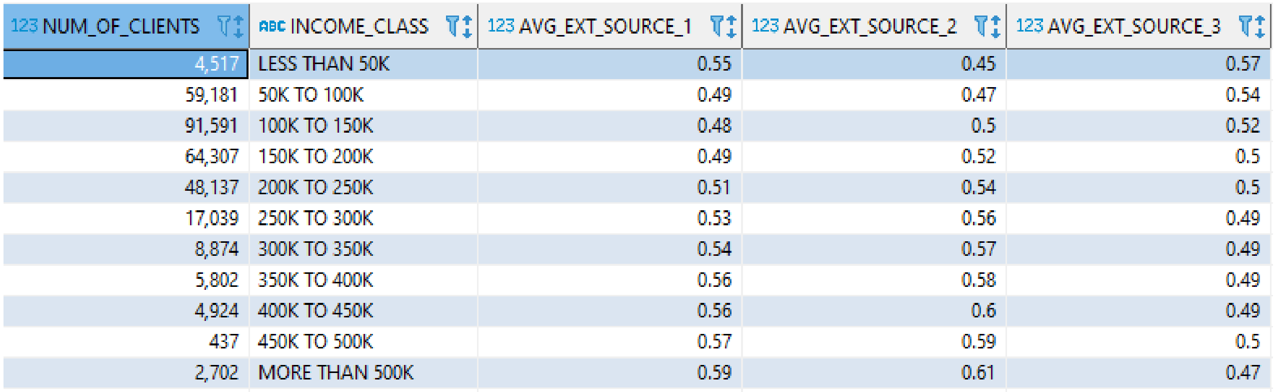
```
SELECT
      min(AMT_INCOME_TOTAL),
      max(AMT_INCOME_TOTAL),
      ROUND(avg(AMT_INCOME_TOTAL))
FROM application
   ORDER BY AMT_INCOME_TOTAL
```



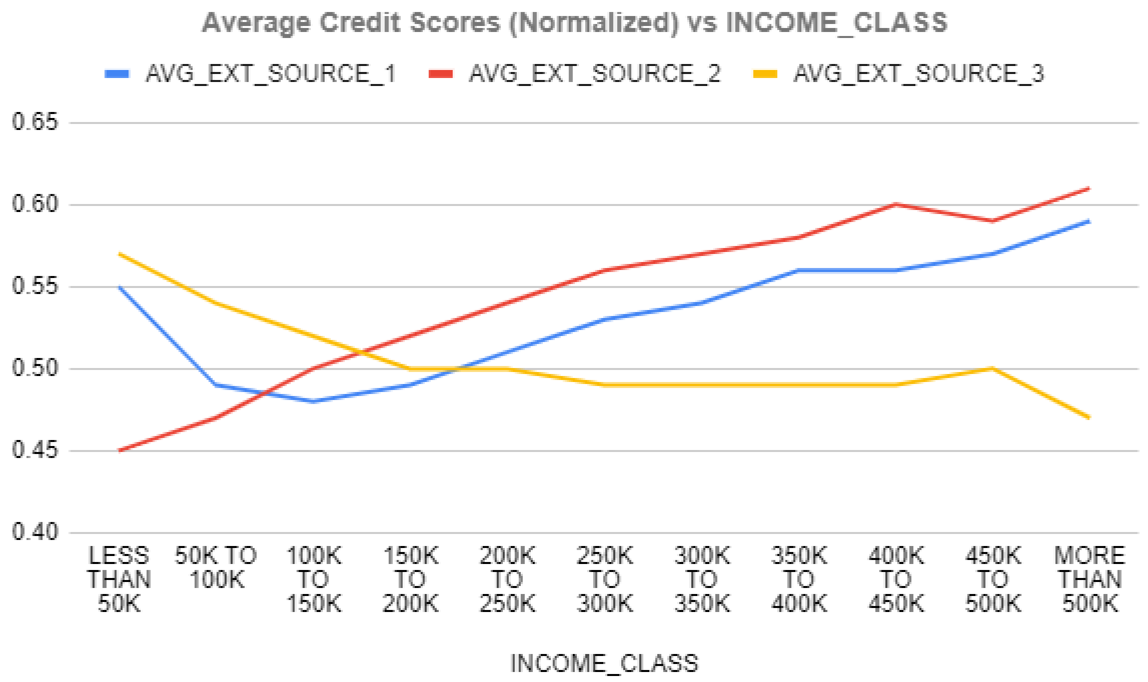
MIN	MAX	AVG
25,650	117,000,000	168,798

- **To analyze if the credit scores are affected by the total income of the client.**

```
SELECT COUNT(AMT_INCOME_TOTAL) AS ‘NUM_OF_CLIENTS’,
CASE
WHEN AMT_INCOME_TOTAL <= 50000 THEN "LESS THAN 50K"
WHEN AMT_INCOME_TOTAL > 50000 and AMT_INCOME_TOTAL <= 100000 THEN "50K TO 100K"
WHEN AMT_INCOME_TOTAL > 100000 and AMT_INCOME_TOTAL <= 150000 THEN "100K TO 150K"
WHEN AMT_INCOME_TOTAL > 150000 and AMT_INCOME_TOTAL <= 200000 THEN "150K TO 200K"
WHEN AMT_INCOME_TOTAL > 200000 AND AMT_INCOME_TOTAL <= 250000 THEN "200K TO 250K"
WHEN AMT_INCOME_TOTAL > 250000 AND AMT_INCOME_TOTAL <= 300000 THEN "250K TO 300K"
WHEN AMT_INCOME_TOTAL > 300000 AND AMT_INCOME_TOTAL <= 350000 THEN "300K TO 350K"
WHEN AMT_INCOME_TOTAL > 350000 AND AMT_INCOME_TOTAL <= 400000 THEN "350K TO 400K"
WHEN AMT_INCOME_TOTAL > 400000 AND AMT_INCOME_TOTAL <= 450000 THEN "400K TO 450K"
WHEN AMT_INCOME_TOTAL > 450000 AND AMT_INCOME_TOTAL <= 500000 THEN "450K TO 500K"
WHEN AMT_INCOME_TOTAL > 500000 THEN "MORE THAN 500K"
END AS "INCOME_CLASS",
      ROUND(AVG(EXT_SOURCE_1),2) AS AVG_EXT_SOURCE_1,
      ROUND(AVG(EXT_SOURCE_2),2) AS AVG_EXT_SOURCE_2,
      ROUND(AVG(EXT_SOURCE_3),2) AS AVG_EXT_SOURCE_3
FROM application
GROUP BY INCOME_CLASS
ORDER BY AMT_INCOME_TOTAL
```



NUM_OF_CLIENTS	INCOME_CLASS	AVG_EXT_SOURCE_1	AVG_EXT_SOURCE_2	AVG_EXT_SOURCE_3
4,517	LESS THAN 50K	0.55	0.45	0.57
59,181	50K TO 100K	0.49	0.47	0.54
91,591	100K TO 150K	0.48	0.5	0.52
64,307	150K TO 200K	0.49	0.52	0.5
48,137	200K TO 250K	0.51	0.54	0.5
17,039	250K TO 300K	0.53	0.56	0.49
8,874	300K TO 350K	0.54	0.57	0.49
5,802	350K TO 400K	0.56	0.58	0.49
4,924	400K TO 450K	0.56	0.6	0.49
437	450K TO 500K	0.57	0.59	0.5
2,702	MORE THAN 500K	0.59	0.61	0.47



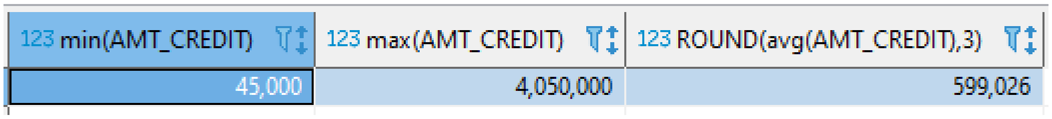
Result :

- Both AVG_EXT_SOURCE_1 and AVG_EXT_SOURCE_2 show upward trend, which indicate that group of clients with higher total income will have better credit scores.
- However, AVG_EXT_SOURCE_3 illustrate a downward trend instead. Which indicates that, based on that source, group of clients with higher total income have lower credit scores than those who earn less.
- AMT_INCOME_TOTAL is a factor which depends to EXT_SOURCE_1 and EXT_SOURCE_2 but according to observation it does not impose on EXT_SOURCE_3.

C. AMT_CREDIT : Credit amount of the loan

- **To check the min, max and avg of the total amount credit of the clients**

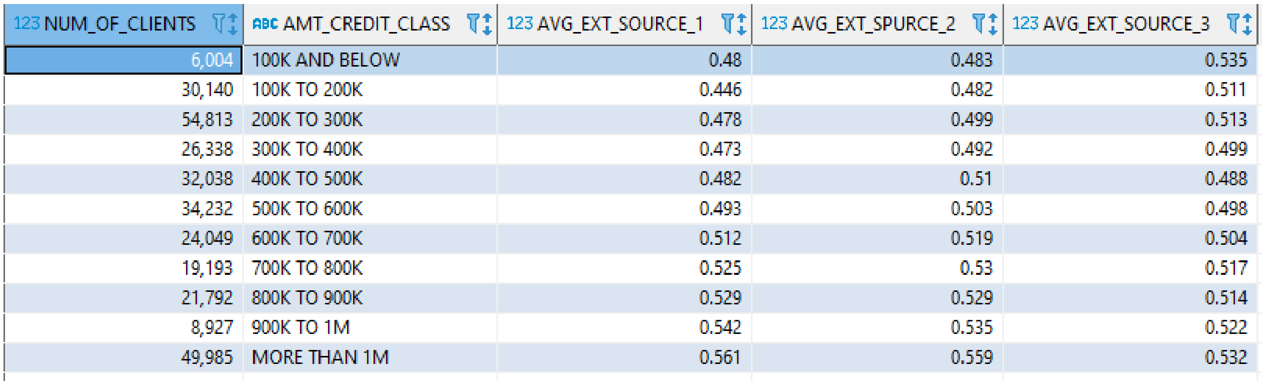
```
SELECT min(AMT_CREDIT), max(AMT_CREDIT), ROUND(avg(AMT_CREDIT),3)
FROM application
```



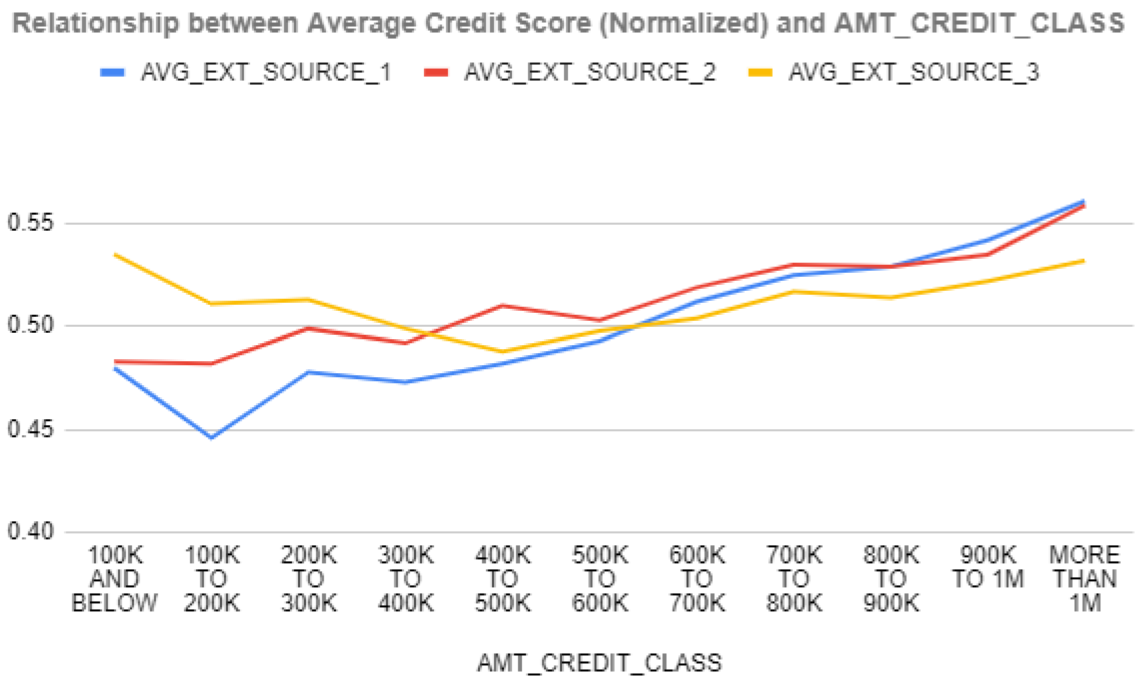
123 min(AMT_CREDIT)	123 max(AMT_CREDIT)	123 ROUND(avg(AMT_CREDIT),3)
45,000	4,050,000	599,026

- **To analyze if the total amount credit of the clients have significant affect to the normalized credit scores**

```
SELECT COUNT(AMT_CREDIT) AS "NUM_OF_CLIENTS",
CASE
      WHEN AMT_CREDIT <= 100000 THEN '100K AND BELOW'
      WHEN AMT_CREDIT > 100000 AND AMT_CREDIT <= 200000 THEN '100K TO 200K'
      WHEN AMT_CREDIT > 200000 AND AMT_CREDIT <= 300000 THEN '200K TO 300K'
      WHEN AMT_CREDIT > 300000 AND AMT_CREDIT <= 400000 THEN '300K TO 400K'
      WHEN AMT_CREDIT > 400000 AND AMT_CREDIT <= 500000 THEN '400K TO 500K'
      WHEN AMT_CREDIT > 500000 AND AMT_CREDIT <= 600000 THEN '500K TO 600K'
      WHEN AMT_CREDIT > 600000 AND AMT_CREDIT <= 700000 THEN '600K TO 700K'
      WHEN AMT_CREDIT > 700000 AND AMT_CREDIT <= 800000 THEN '700K TO 800K'
      WHEN AMT_CREDIT > 800000 AND AMT_CREDIT <= 900000 THEN '800K TO 900K'
      WHEN AMT_CREDIT > 900000 AND AMT_CREDIT <= 1000000 THEN '900K TO 1M'
      ELSE 'MORE THAN 1M'
END AS "AMT_CREDIT_CLASS",
ROUND(AVG(EXT_SOURCE_1),3) AS 'AVG_EXT_SOURCE_1',
ROUND(AVG(EXT_SOURCE_2),3) AS 'AVG_EXT_SPURCE_2',
ROUND(AVG(EXT_SOURCE_3),3) AS 'AVG_EXT_SOURCE_3'
FROM application
GROUP BY AMT_CREDIT_CLASS
ORDER BY AMT_CREDIT_CLASS
```



123 NUM_OF_CLIENTS	ABC AMT_CREDIT_CLASS	123 AVG_EXT_SOURCE_1	123 AVG_EXT_SPURCE_2	123 AVG_EXT_SOURCE_3
6,004	100K AND BELOW	0.48	0.483	0.535
30,140	100K TO 200K	0.446	0.482	0.511
54,813	200K TO 300K	0.478	0.499	0.513
26,338	300K TO 400K	0.473	0.492	0.499
32,038	400K TO 500K	0.482	0.51	0.488
34,232	500K TO 600K	0.493	0.503	0.498
24,049	600K TO 700K	0.512	0.519	0.504
19,193	700K TO 800K	0.525	0.53	0.517
21,792	800K TO 900K	0.529	0.529	0.514
8,927	900K TO 1M	0.542	0.535	0.522
49,985	MORE THAN 1M	0.561	0.559	0.532



Result :

- Clients with higher amount of credit usually have better credit scores
- Having more credit amount does not necessarily means a bad thing.
- It is easier to track the credit health of the client if they have more credit history.
- Having better credit scores resulted with higher amount of loan given, hence higher credit amount.

Task 6 Are there any factors in the application table affecting the Credit Amount?

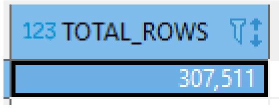
Who is going to lend more money than others? In this task, we want to see are there any factors affecting the credit amount. **Do the analysis of at least 3 factors**

Please explain your findings with SQL statements and results:

1. **Pre-Analysis :**

   A. To calculate the number of rows inside the credit amount column

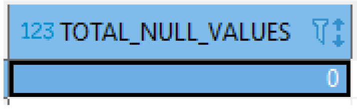
```
SELECT COUNT(AMT_CREDIT) AS "TOTAL_ROWS"
       FROM application
```



123 TOTAL_ROWS
307,511

   B. To calculate the number of NULL Values :

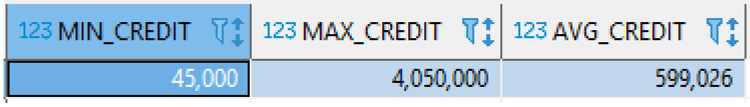
```
SELECT COUNT(IFNULL(AMT_CREDIT, 'N/A')) AS "TOTAL_NULL_VALUES"
FROM application
WHERE AMT_CREDIT IS NULL
```



123 TOTAL_NULL_VALUES
0

   C. To determine the min, max and avg AMT_CREDIT:

```
SELECT
       MIN(AMT_CREDIT) AS "MIN_CREDIT",
       MAX(AMT_CREDIT) AS "MAX_CREDIT",
       ROUND(AVG(AMT_CREDIT), 3) AS "AVG_CREDIT"
       FROM application
```

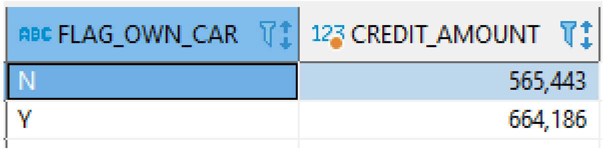


123 MIN_CREDIT	123 MAX_CREDIT	123 AVG_CREDIT
45,000	4,050,000	599,026

2. Factors Affecting the Credit Amount (AMT_CREDIT)

A. FLAG_OWN_CAR : Flag if the client owns a car

```
SELECT
       FLAG_OWN_CAR,
       ROUND(AVG(AMT_CREDIT)) AS "CREDIT_AMOUNT"
FROM application
GROUP BY FLAG_OWN_CAR
```



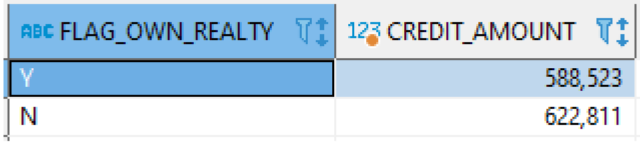
FLAG_OWN_CAR	CREDIT_AMOUNT
N	565,443
Y	664,186

Result:

- Clients who own car have higher amount of credit than clients who do not own car.
- Having own car will cause the client to have additional debt (car loan), thus higher CREDIT_AMOUNT.

B. FLAG_OWN_REALTY : Flag if the client owns a house or flat

```
SELECT
       FLAG_OWN_REALTY,
       ROUND(AVG(AMT_CREDIT)) AS "CREDIT_AMOUNT"
FROM application
          GROUP BY FLAG_OWN_REALTY
```



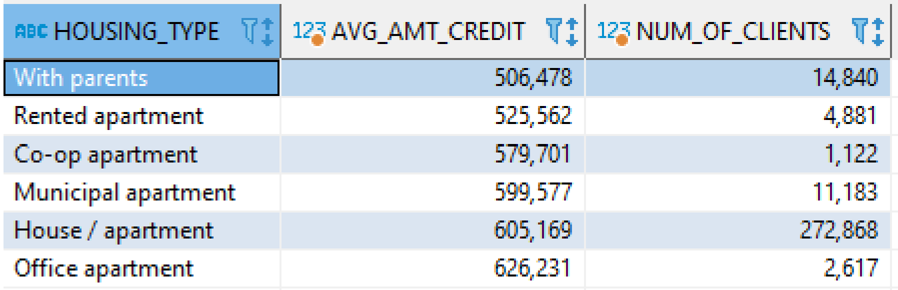
FLAG_OWN_REALTY	CREDIT_AMOUNT
Y	588,523
N	622,811

Result:

- Client who owns a house or flat has higher credit amount than client who does not.
- Having own house or flat will cause the client to have additional debt (mortgage, home loan), thus higher CREDIT_AMOUNT.

C. **NAME_HOUSING_TYPE :** Housing situation of the client

```
SELECT
	NAME_HOUSING_TYPE AS "HOUSING_TYPE",
	ROUND(AVG(AMT_CREDIT)) AS "AVG_AMT_CREDIT",
	COUNT(NAME_HOUSING_TYPE) AS "NUM_OF_CLIENTS"
FROM application
GROUP BY NAME_HOUSING_TYPE
ORDER BY AVG_AMT_CREDIT
```



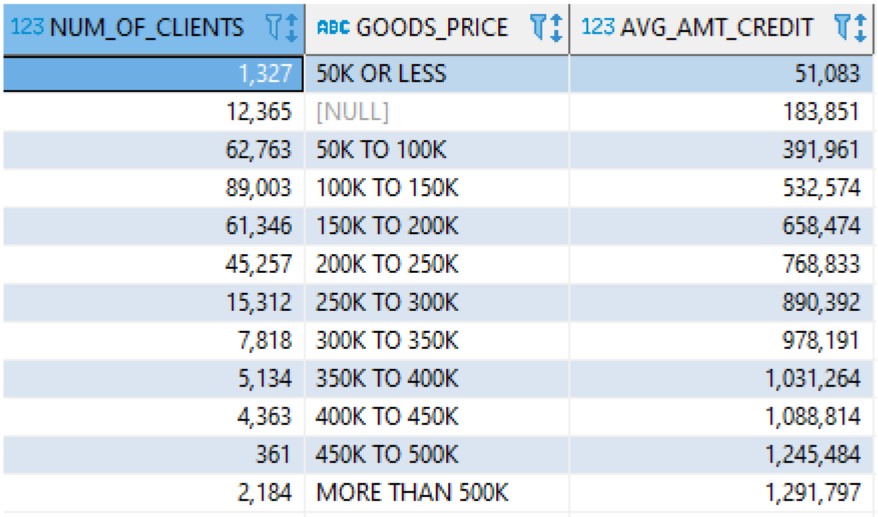
HOUSING_TYPE	AVG_AMT_CREDIT	NUM_OF_CLIENTS
With parents	506,478	14,840
Rented apartment	525,562	4,881
Co-op apartment	579,701	1,122
Municipal apartment	599,577	11,183
House / apartment	605,169	272,868
Office apartment	626,231	2,617

Result :

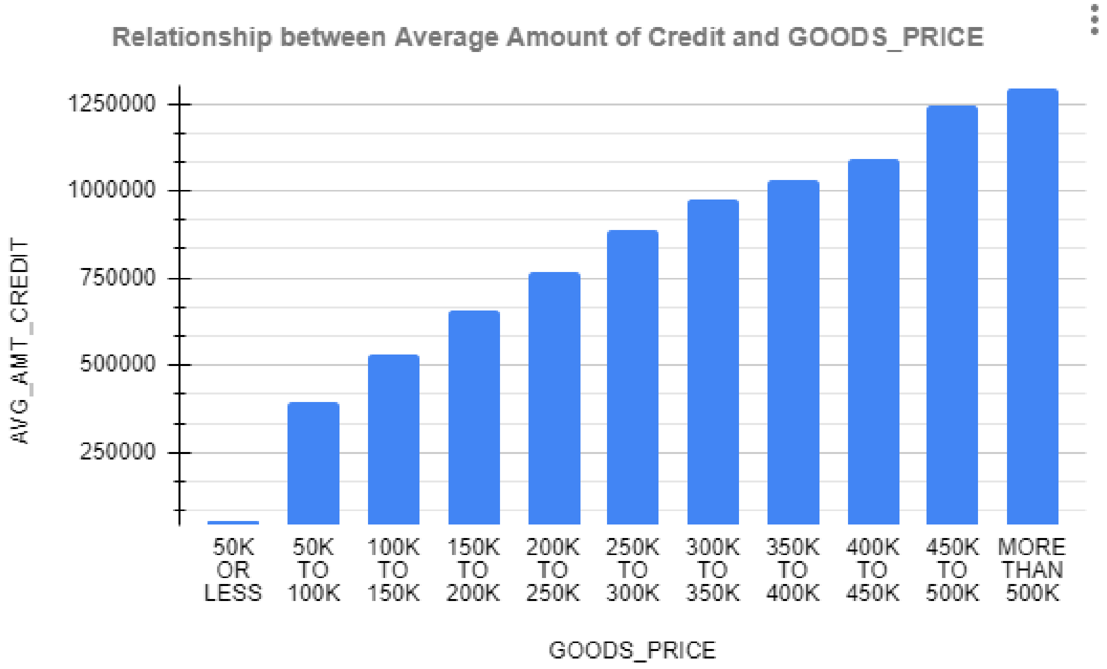
- The result correlates with FLAG_OWN_REALTY, where clients who do not own a house will have lower AVG_AMT_CREDIT – living with parent, renting an apartment.
- Meanwhile, clients who own a house or apartment, will have higher debt or higher AVG_AMT_CREDIT – co-op apartment, municipal apartment, owning a house / apartment and owning an office apartment.

D. AMT_GOODS_PRICE : For consumer loans it is the price of the goods for which the loan is given

```
SELECT COUNT(AMT_GOODS_PRICE) AS 'NUM_OF_CLIENTS',
CASE
      WHEN AMT_GOODS_PRICE <= 50000 THEN "50K OR LESS"
      WHEN AMT_GOODS_PRICE > 50000 and AMT_INCOME_TOTAL <= 100000 THEN "50K TO 100K"
      WHEN AMT_GOODS_PRICE > 100000 and AMT_INCOME_TOTAL <= 150000 THEN "100K TO 150K"
      WHEN AMT_GOODS_PRICE > 150000 and AMT_INCOME_TOTAL <= 200000 THEN "150K TO 200K"
      WHEN AMT_GOODS_PRICE > 200000 AND AMT_INCOME_TOTAL <= 250000 THEN "200K TO 250K"
      WHEN AMT_GOODS_PRICE > 250000 AND AMT_INCOME_TOTAL <= 300000 THEN "250K TO 300K"
      WHEN AMT_GOODS_PRICE > 300000 AND AMT_INCOME_TOTAL <= 350000 THEN "300K TO 350K"
      WHEN AMT_GOODS_PRICE > 350000 AND AMT_INCOME_TOTAL <= 400000 THEN "350K TO 400K"
      WHEN AMT_GOODS_PRICE > 400000 AND AMT_INCOME_TOTAL <= 450000 THEN "400K TO 450K"
      WHEN AMT_GOODS_PRICE > 450000 AND AMT_INCOME_TOTAL <= 500000 THEN "450K TO 500K"
      WHEN AMT_GOODS_PRICE > 500000 THEN "MORE THAN 500K"
END AS 'GOODS_PRICE',
ROUND(AVG(AMT_CREDIT)) AS AVG_AMT_CREDIT
FROM application
GROUP BY GOODS_PRICE
ORDER BY AVG_AMT_CREDIT
```



123 NUM_OF_CLIENTS	ABC GOODS_PRICE	123 AVG_AMT_CREDIT
1,327	50K OR LESS	51,083
12,365	[NULL]	183,851
62,763	50K TO 100K	391,961
89,003	100K TO 150K	532,574
61,346	150K TO 200K	658,474
45,257	200K TO 250K	768,833
15,312	250K TO 300K	890,392
7,818	300K TO 350K	978,191
5,134	350K TO 400K	1,031,264
4,363	400K TO 450K	1,088,814
361	450K TO 500K	1,245,484
2,184	MORE THAN 500K	1,291,797



Result :

- GOODS_PRICE is directly proportional to the AVG_AMT_CREDIT.
- Higher goods price results with higher credit amount.

Task 7 Are there any factors in the application table affecting the Payment Difficulties?

In the database, the TARGET column describes will there be a payment difficulty for a loan. We want to see if there are any factors in the application table that can be used to predict this future information. **Do the analysis of at least 3 factors**

Please explain your findings with SQL statements and results:

A. OCCUPATION_TYPE : What type of occupation does the client have

```sql
SELECT TARGET,
       COUNT(TARGET) AS NUM_OF_CLIENTS,
       OCCUPATION_TYPE
FROM application
GROUP BY TARGET, OCCUPATION_TYPE
ORDER BY
       OCCUPATION_TYPE ASC,
       NUM_OF_CLIENTS DESC
```

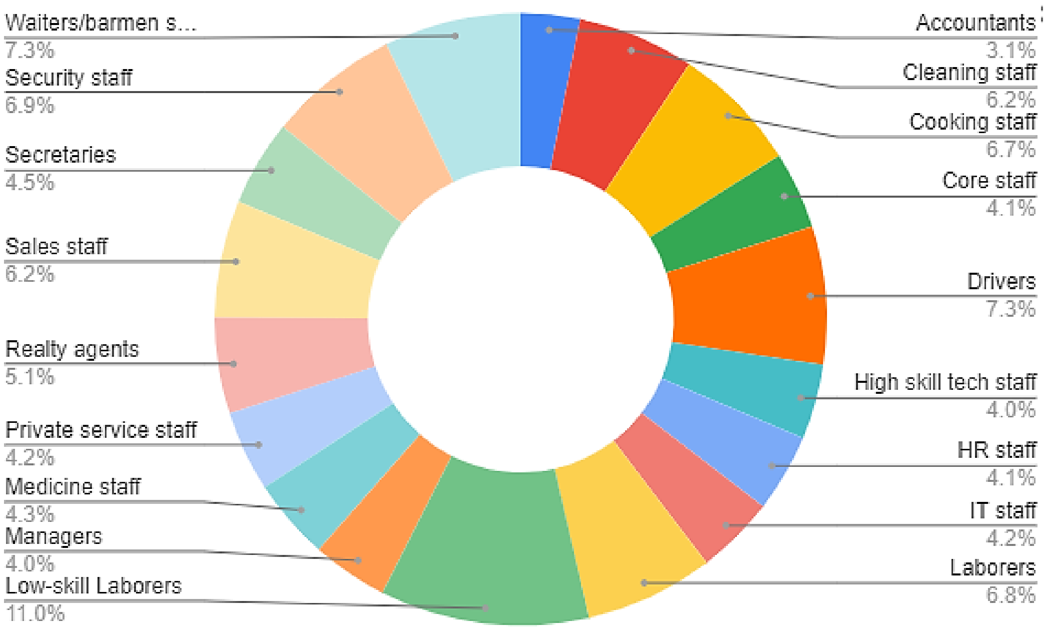
Notes : PAYMENT_DIFFICULTIES (%) indicates the percentage of clients from the specific occupation that have encountered payment difficulties before.

PAYMENT_DIFFICULTIES (%) is calculated separately on spreadsheets according to results from SQL queries.

TARGET 0 : No payment difficulties
TARGET 1 : Has payment difficulties

OCCUPATION_TYPE	TARGET	NUM_OF_CLIENTS	PAYMENT_DIFFICULTIES (%)
Accountants	0	9339	4.83
	1	474	
Cleaning staff	0	4206	9.61
	1	447	
Cooking staff	0	5325	10.44
	1	621	
Core staff	0	25832	6.30
	1	1738	
Drivers	0	16496	11.33
	1	2107	
High skill tech staff	0	10679	6.16
	1	701	
HR staff	0	527	6.39
	1	36	
IT staff	0	492	6.46
	1	34	
Laborers	0	49348	10.58
	1	5838	
Low-skill Laborers	0	1734	17.15
	1	359	
Managers	0	20043	6.21
	1	1328	

Medicine staff	0	7965	6.70
	1	572	
Private service staff	0	2477	6.60
	1	175	
Realty agents	0	692	7.86
	1	59	
Sales staff	0	29010	9.63
	1	3092	
Secretaries	0	1213	7.05
	1	92	
Security staff	0	5999	10.74
	1	722	
Waiters/barmen staff	0	1196	11.28
	1	152	



Result:

- Low-skill Laborers have the highest % of clients that encountered loan payment difficulties, followed by Drivers, Waiters/Barmen staff, Security staff, Laborers, Cooking Staff and Cleaning Staff.
- Accountants have the lowest % of clients that encountered loan payment difficulties, followed by High skill tech staff, Core staff, Managers, IT staff, HR staff, Private Service staff, Medicine staff, Secretaries, Realty agents and Sales staff.
- Professionals tend to have better pay, hence a smaller number of clients from this group have trouble paying their debt due to stable income

B. NAME_EDUCATION_TYPE : Level of highest education the clients achieved

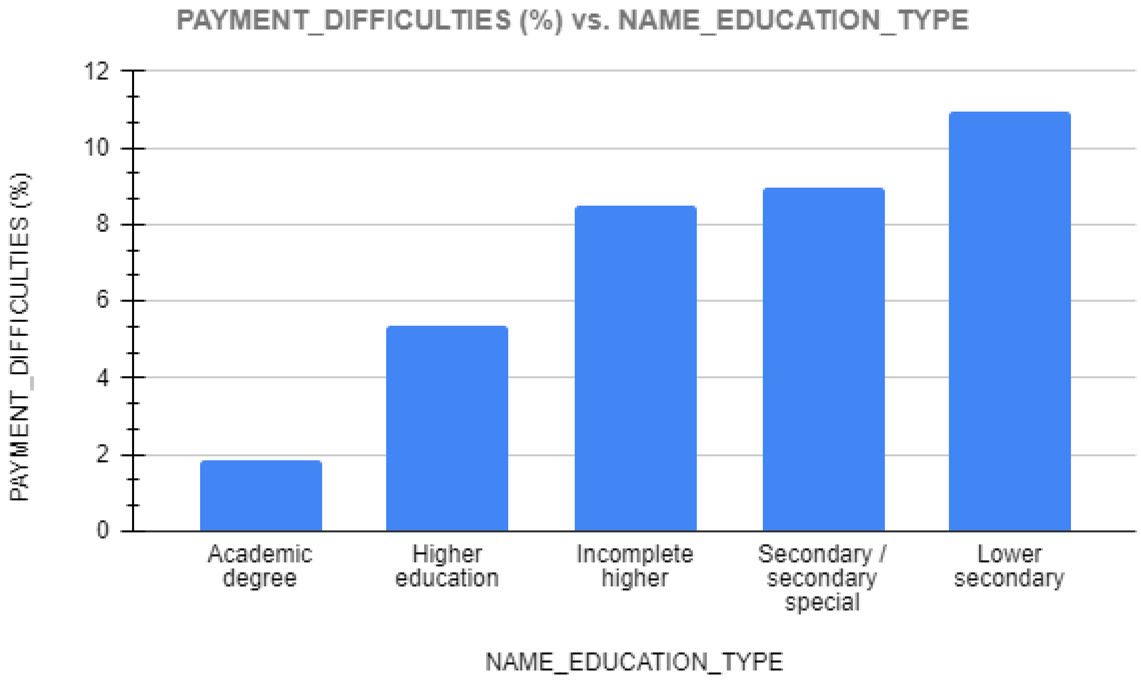
```
SELECT
       NAME_EDUCATION_TYPE,
       TARGET,
       COUNT(TARGET) AS 'NUM_OF_CLIENTS'
FROM application
GROUP BY
       NAME_EDUCATION_TYPE, TARGET
ORDER BY
       NAME_EDUCATION_TYPE,
       NUM_OF_CLIENTS DESC
```

Notes : PAYMENT_DIFFICULTIES (%) indicates the percentage of clients from the specific group (INCOME_CLASS) that have encountered payment difficulties before.

PAYMENT_DIFFICULTIES (%) is calculated separately on spreadsheets according to results from SQL queries.

TARGET 0 : No payment difficulties
TARGET 1 : Has payment difficulties

NAME_EDUCATION_TYPE	TARGET	NUM_OF_CLIENTS	PAYMENT_DIFFICULTIES (%)
Academic degree	0	161	1.829268293
	1	3	
Higher education	0	70854	5.355115344
	1	4009	
Incomplete higher	0	9405	8.48496643
	1	872	
Lower secondary	0	3399	10.92767296
	1	417	
Secondary / secondary special	0	198867	8.939928843
	1	19524	



Result:

- Observation shows that clients with highest level of education which is the Academic Degree has the lowest percentage of clients who had encountered payment difficulties.
- Whereas the clients that have an education level of lower secondary tend to have a greater number of clients facing difficulties in loan repayment.
- Results show that the clients' level of education has significantly affect their payment difficulties.

C. FLAG_OWN_REALTY : Flag if the clients own a house or flat

```sql
SELECT FLAG_OWN_REALTY, TARGET, COUNT(TARGET) AS NUM_OF_CLIENT
FROM application
GROUP BY TARGET, FLAG_OWN_REALTY
ORDER BY
      FLAG_OWN_REALTY DESC,
      NUM_OF_CLIENT DESC
```

FLAG_OWN_REALTY	TARGET	NUM_OF_CLIENT	PAYMENT_DIFFICULTIES (%)
Y	0	196329	7.96
	1	16983	
N	0	86357	8.32
	1	7842	

Notes : PAYMENT_DIFFICULTIES (%) indicates the percentage of clients from the specific group (FLAG_OWN_REALTY = Y or FLAG_OWN_REALTY = N) that have encountered payment difficulties before.

PAYMENT_DIFFICULTIES (%) is calculated separately on spreadsheets according to results from SQL queries.

TARGET 0 : No payment difficulties
TARGET 1 : Has payment difficulties

Result :

- Slightly higher percentage of clients have payment difficulties but not owning any realty.

Inference :

- Having their own realty (a house or flat) indicates that the clients have better credit score, which allows them to get higher loan from the loan provider and to be able buy their own realty.
- Better credit scores were resulted from good loan repayment history, hence those group of clients with realty has lower percentage of people that has payment difficulties.

D. NAME_INCOME_TYPE: Clients income type

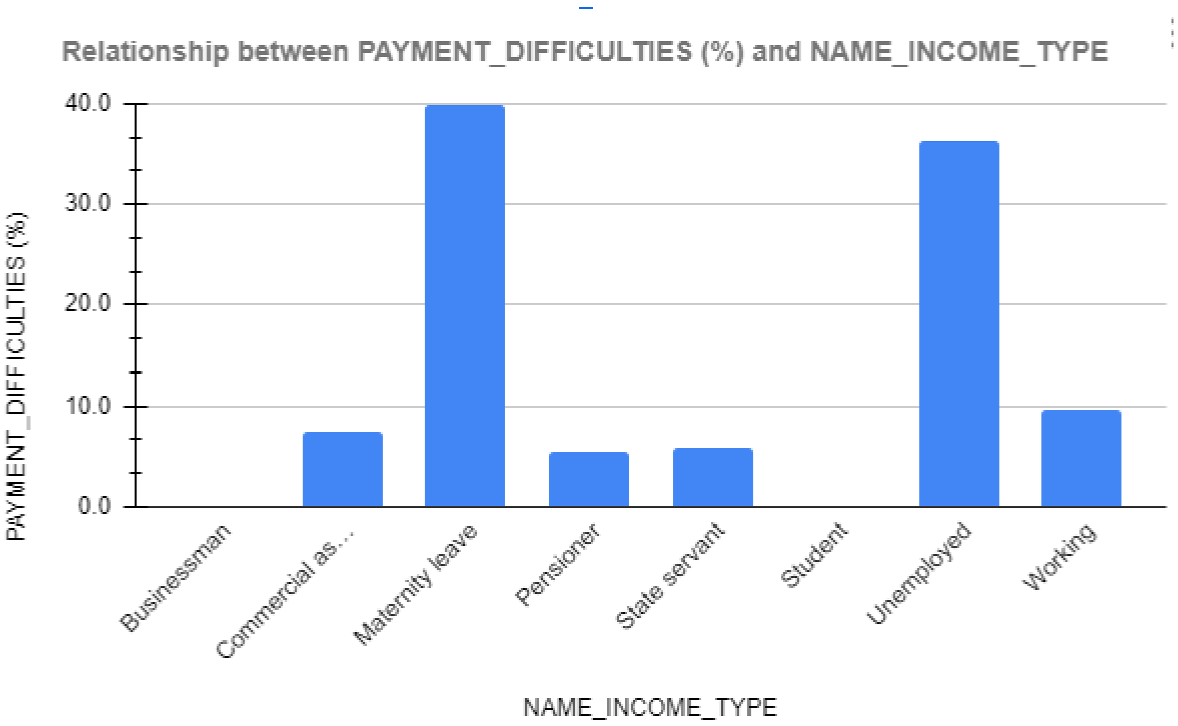
```
SELECT
      NAME_INCOME_TYPE,
      TARGET,
      COUNT(TARGET) AS 'NUM_OF_CLIENTS'
FROM application
GROUP BY
      NAME_INCOME_TYPE, TARGET
ORDER BY
      NAME_INCOME_TYPE ASC, NUM_OF_CLIENTS DESC
```

Notes : PAYMENT_DIFFICULTIES (%) indicates the percentage of clients from the specific group (FLAG_OWN_CAR = Y, or FLAG_OWN_CAR = N) that have encountered payment difficulties before.

PAYMENT_DIFFICULTIES (%) is calculated separately on spreadsheets according to results from SQL queries.

TARGET 0 : No payment difficulties
TARGET 1 : Has payment difficulties

NAME_INCOME_TYPE	TARGET	NUM_OF_CLIENTS	PAYMENT_DIFFICULTIES (%)
Businessman	0	10	0.0
	1	0	
Commercial associate	0	66257	7.5
	1	5360	
Maternity leave	0	3	40.0
	1	2	
Pensioner	0	52380	5.4
	1	2982	
State servant	0	20454	5.8
	1	1249	
Student	0	18	0.0
	1	0	
Unemployed	0	14	36.4
	1	8	
Working	0	143550	9.6
	1	15224	



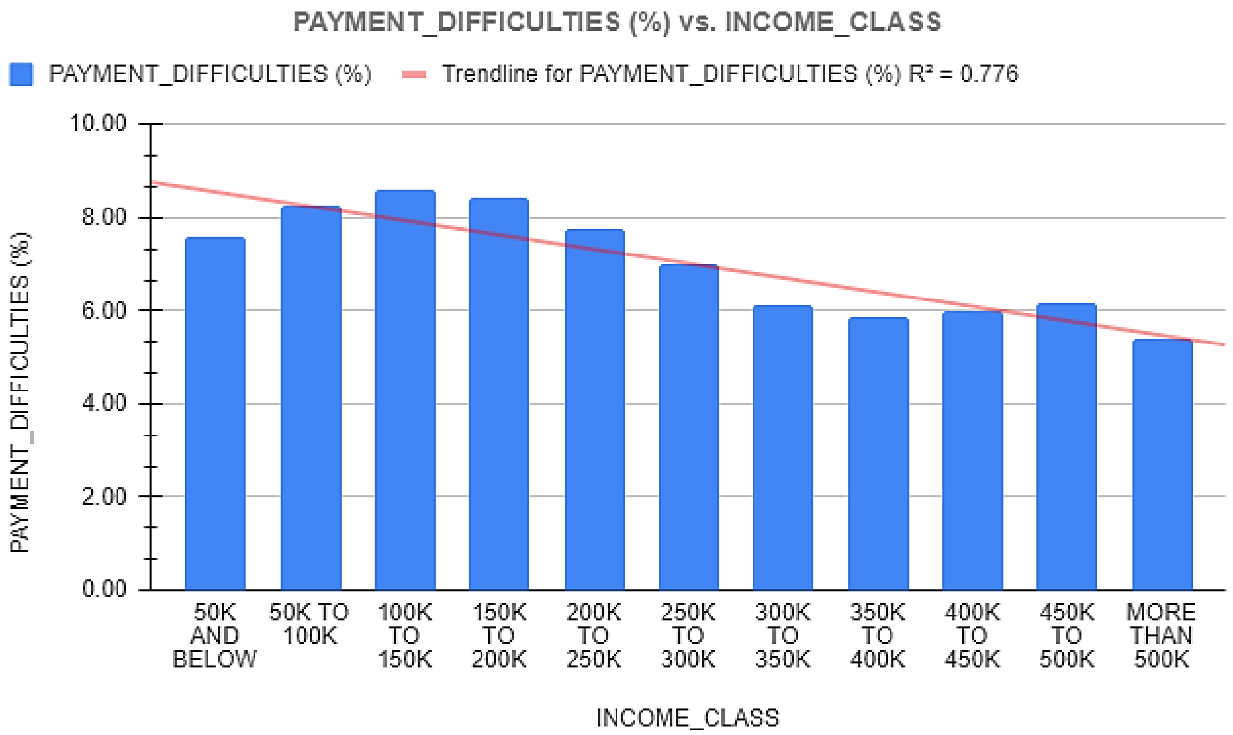
Result :

- 2 out of 5 clients who are on maternity leave have encountered payment difficulties which resulted with 40%.
- 36.4% of clients who are unemployed have difficulties in loan repayment.
- Businessman and students have 0% of payment difficulties.
- Commercial Associate, Pensioner, State Servant and Working clients have below than 10% of payment difficulties for each group.

E. AMT_INCOME_TOTAL : Total income of the client

```sql
SELECT
CASE
WHEN AMT_INCOME_TOTAL <= 50000 THEN "50K AND BELOW"
WHEN AMT_INCOME_TOTAL > 50000 AND AMT_INCOME_TOTAL <= 100000 THEN "50K TO 100K"
WHEN AMT_INCOME_TOTAL > 100000 AND AMT_INCOME_TOTAL <= 150000 THEN "100K TO 150K"
WHEN AMT_INCOME_TOTAL > 150000 AND AMT_INCOME_TOTAL <= 200000 THEN "150K TO 200K"
WHEN AMT_INCOME_TOTAL > 200000 AND AMT_INCOME_TOTAL <= 250000 THEN "200K TO 250K"
WHEN AMT_INCOME_TOTAL > 250000 AND AMT_INCOME_TOTAL <= 300000 THEN "250K TO 300K"
WHEN AMT_INCOME_TOTAL > 300000 AND AMT_INCOME_TOTAL <= 350000 THEN "300K TO 350K"
WHEN AMT_INCOME_TOTAL > 350000 AND AMT_INCOME_TOTAL <= 400000 THEN "350K TO 400K"
WHEN AMT_INCOME_TOTAL > 400000 AND AMT_INCOME_TOTAL <= 450000 THEN "400K TO 450K"
WHEN AMT_INCOME_TOTAL > 450000 AND AMT_INCOME_TOTAL <= 500000 THEN "450K TO 500K"
WHEN AMT_INCOME_TOTAL > 500000 THEN "MORE THAN 500K"
END AS "INCOME_CLASS",
       TARGET,
       COUNT(TARGET) AS "NUM_OF_CLIENT"
FROM application
GROUP BY INCOME_CLASS, TARGET
ORDER BY INCOME_CLASS
```

INCOME_CLASS	TARGET	NUM_OF_CLIENT	PAYMENT_DIFFICULTIES (%)
50K AND BELOW	0	4174	7.59
	1	343	
50K TO 100K	0	54299	8.25
	1	4882	
100K TO 150K	0	83697	8.62
	1	7894	
150K TO 200K	0	58875	8.45
	1	5432	
200K TO 250K	0	44409	7.74
	1	3728	
250K TO 300K	0	15846	7.00
	1	1193	
300K TO 350K	0	8329	6.14
	1	545	
350K TO 400K	0	5462	5.86
	1	340	
400K TO 450K	0	4629	5.99
	1	295	
450K TO 500K	0	410	6.18
	1	27	
MORE THAN 500K	0	2556	5.40
	1	146	



Notes : PAYMENT_DIFFICULTIES (%) indicates the percentage of clients from the specific group (INCOME_CLASS) that have encountered payment difficulties before.

PAYMENT_DIFFICULTIES (%) is calculated separately on spreadsheets according to results from SQL queries.

TARGET 0 : No payment difficulty
TARGET 1 : Has payment difficulty

Results :

- Higher amount of income results in lower percentage of clients having payment difficulty.
- According to observation, the group of clients with higher income has higher capability to repay their debt due to more financial stability to do so.

Previous/Other Loan Applications

In the previous section, we explored if the demographic data related to payment difficulties, this section we want to see if **historical loan behavior** affecting the payment difficulties.

The “bureau” table stores the other loans of the applicants from the other lenders.

“bureau” table:

SK_ID_CURR	ID of loan in our sample - one loan in our sample can have 0,1,2 or more related previous credits in credit bureau
SK_BUREAU_ID	Recoded ID of previous Credit Bureau credit related to our loan (unique coding for each loan application), The IDs of the “other loans”
CREDIT_DAY_OVERDUE	Number of days past due on CB credit at the time of application for related loan in our sample
AMT_CREDIT_MAX_OVERDUE	Maximal amount overdue on the Credit Bureau credit so far (at application date of loan in our sample)
CNT_CREDIT_PROLONG	How many times was the Credit Bureau credit prolonged
AMT_CREDIT_SUM	Current credit amount for the Credit Bureau credit
AMT_CREDIT_SUM_DEBT	Current debt on Credit Bureau credit
AMT_CREDIT_SUM_LIMIT	Current credit limit of credit card reported in Credit Bureau
AMT_CREDIT_SUM_OVERDUE	Current amount overdue on Credit Bureau credit
CREDIT_TYPE	Type of Credit Bureau credit (Car, cash,...)
DAYS_CREDIT_UPDATE	How many days before loan application did last information about the Credit Bureau credit come
AMT_ANNUITY	Annuity of the Credit Bureau credit

Task 8 Is the number of other loans affecting the payment difficulties?

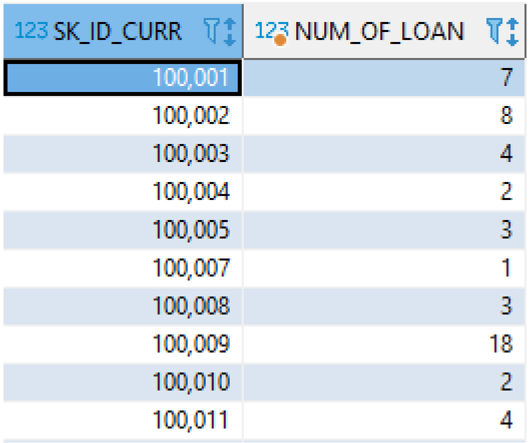
We want to see if loan applicants have other historical loans affecting their payment abilities.
Hints:

- You will need to count the number of loans for each SK_ID_CURR in the “bureau” table.
- Transform the counts into count groups (Discretization).
- Compute the relation between average other loan count to the TARGET

Paste the SQL and part of the results below:

1. **To count the number of loans for each SK_ID_CURR**

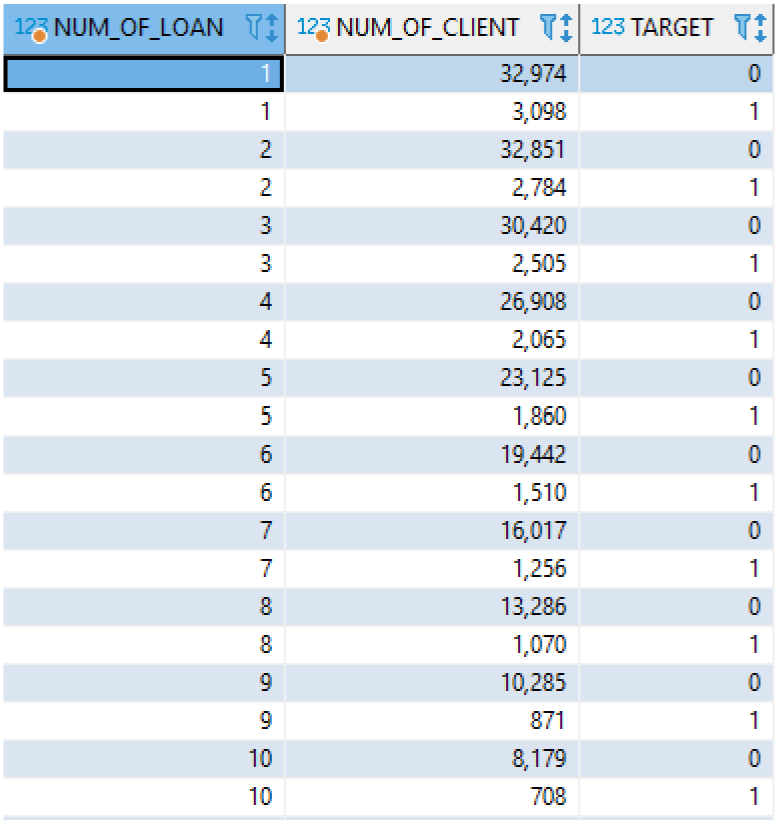
```
SELECT SK_ID_CURR,
COUNT(SK_ID_CURR) AS "NUM_OF_LOAN"
FROM bureau
GROUP BY SK_ID_CURR
```



SK_ID_CURR	NUM_OF_LOAN
100,001	7
100,002	8
100,003	4
100,004	2
100,005	3
100,007	1
100,008	3
100,009	18
100,010	2
100,011	4

2. **Transform the count into count groups (Discretization).**

```
WITH TEMP_TABLE AS (SELECT SK_ID_CURR, COUNT(SK_ID_CURR) AS "NUM_OF_LOAN"
FROM bureau
GROUP BY SK_ID_CURR
ORDER BY "NUM_OF_LOAN")

SELECT TEMP_TABLE.NUM_OF_LOAN,
       COUNT(TEMP_TABLE.SK_ID_CURR) AS "NUM_OF_CLIENT",
       application.TARGET
FROM TEMP_TABLE JOIN application
ON TEMP_TABLE.SK_ID_CURR = application.SK_ID_CURR
GROUP BY NUM_OF_LOAN, TARGET
ORDER BY NUM_OF_LOAN
```



NUM_OF_LOAN	NUM_OF_CLIENT	TARGET
1	32,974	0
1	3,098	1
2	32,851	0
2	2,784	1
3	30,420	0
3	2,505	1
4	26,908	0
4	2,065	1
5	23,125	0
5	1,860	1
6	19,442	0
6	1,510	1
7	16,017	0
7	1,256	1
8	13,286	0
8	1,070	1
9	10,285	0
9	871	1
10	8,179	0
10	708	1

# Results only show from row 1-20

3. Compute the relation between average other loan count to the TARGET

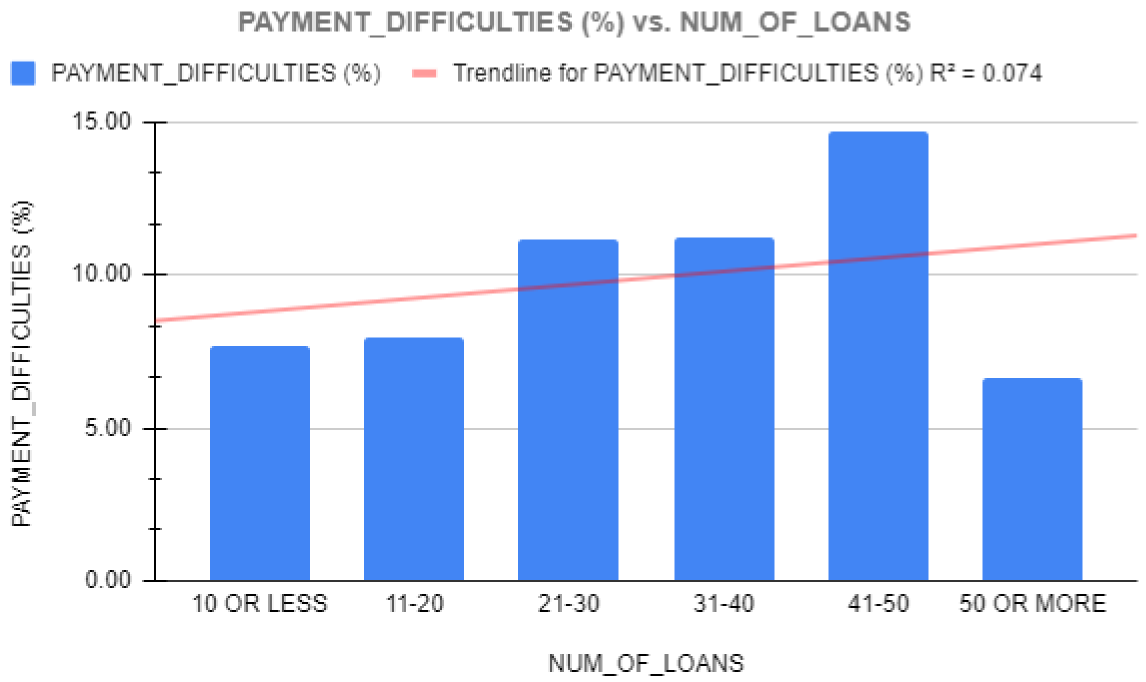
```
WITH TEMP AS (SELECT SK_ID_CURR, COUNT(SK_ID_CURR) AS "NUM_OF_LOAN"
FROM bureau
GROUP BY SK_ID_CURR
ORDER BY "NUM_OF_LOAN")

SELECT CASE
       WHEN TEMP.NUM_OF_LOAN <= 10 THEN '10 OR LESS'
       WHEN TEMP.NUM_OF_LOAN > 10 AND TEMP.NUM_OF_LOAN <= 20 THEN '11-20'
       WHEN TEMP.NUM_OF_LOAN > 20 AND TEMP.NUM_OF_LOAN <= 30 THEN '21-30'
       WHEN TEMP.NUM_OF_LOAN > 30 AND TEMP.NUM_OF_LOAN <= 40 THEN '31-40'
       WHEN TEMP.NUM_OF_LOAN > 40 AND TEMP.NUM_OF_LOAN <= 50 THEN '41-50'
       WHEN TEMP.NUM_OF_LOAN > 50 THEN '50 OR MORE'
END AS "NUM_OF_LOANS",
COUNT(TEMP.SK_ID_CURR) AS "NUM_OF_CLIENT",
       application.TARGET
FROM TEMP JOIN application
ON TEMP.SK_ID_CURR = application.SK_ID_CURR
GROUP BY NUM_OF_LOANS, TARGET
ORDER BY NUM_OF_LOANS
```

Notes : PAYMENT_DIFFICULTIES (%) indicates the percentage of clients from the specific group (INCOME_CLASS) that have encountered payment difficulties before.

PAYMENT_DIFFICULTIES (%) is calculated separately on spreadsheets according to results from SQL queries.

TARGET 0 : No payment difficulties
TARGET 1 : Has payment difficulties

NUM_OF_LOANS	NUM_OF_CLIENT	TARGET	PAYMENT_DIFFICULTY (%)
10 OR LESS	213487	0	7.67
	17727	1	
11-20	27443	0	7.93
	2365	1	
21-30	1968	0	11.15
	247	1	
31-40	182	0	11.22
	23	1	
41-50	29	0	14.71
	5	1	
50 OR MORE	14	0	6.67
	1	1	



Result :

- Percentage of payment difficulties show an upward trend as the number of loans made by the clients increases.
- Generally, higher number of loans often results with higher percentage of clients to encounter payment difficulties.
- However, it was not the case for the group of clients that have 50 of more loans to settle.

Task 9 FreeStyle

Now, conduct your own research and analysis to see what factors from the “application” and the “bureau” tables are affecting

- The Credit Scores
- The Payment Difficulty