</talentlabs>

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	307511
bureau	1716428

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 three-digit number that rates one’s creditworthiness. The higher the score, the more likely we are to get approved for loans and for better rates.

A credit score is based on the credit history, which includes information like the number accounts, total levels of debt, repayment history, and other factors. Lenders use credit scores to evaluate one’s credit worthiness, or the likelihood that we will repay loans in a timely manner. Lenders use the credit score to determines whether to approve us for products like mortgages, personal loans, and credit cards, and what interest rates we will pay.

There are five main factors evaluated when calculating a credit score which are:

- **Payment history:** The payment history includes whether we've paid our bills on time. It takes into account how many late payments we've had, and how late they were.

- **Amounts owed:** Amounts owed is the percentage of credit we've used compared to the credit available to us, which is known as credit utilization.
- **Length of credit history:** Longer credit histories are considered less risky, as there is more data to determine payment history.
- **Credit mix:** A variety of credit types shows lenders we can manage various types of credit. It can include installment credit, such as car loans or mortgage loans, and revolving credit, such as credit cards.
- **New credit:** Lenders view new credit as a potential sign one may be desperate for credit. Too many recent applications for credit can negatively affect the credit score.

Credit scores matters because for lenders and creditors, it provide a quick and standardized way for them to assess the risk of lending money. They help lenders predict how likely a borrower is to repay a loan or credit card debt based on their credit history. Plus, credit scores influence the terms of the credit offered, including interest rates. Higher credit scores generally qualify for lower interest rates, which compensates for the lower risk perceived by the lender.

As for the consumers, a good credit score improves one's chances of being approved for loans and credit cards. It essentially reflects their creditworthiness, indicating to lenders that they are a responsible borrower. Knowing that they have a strong credit score can provide peace of mind and confidence in their financial stability and readiness for future opportunities.

Task 3 Understand Credit Amount and Annuity

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

Credit Amount	Credit Amount generally refers to the total sum of money involved in a credit transaction or available in a credit account. In context of loan, it refers to the total sum of money borrowed from a lender. In summary, credit amount deals with the amount of credit available or owed.
Annuity	An annuity is a financial product that provides a series of payments made at equal intervals. Annuities are commonly used for retirement savings and income purposes. In summary, an annuity is a financial product providing a series of payments over time, often used for retirement planning or investment purposes.

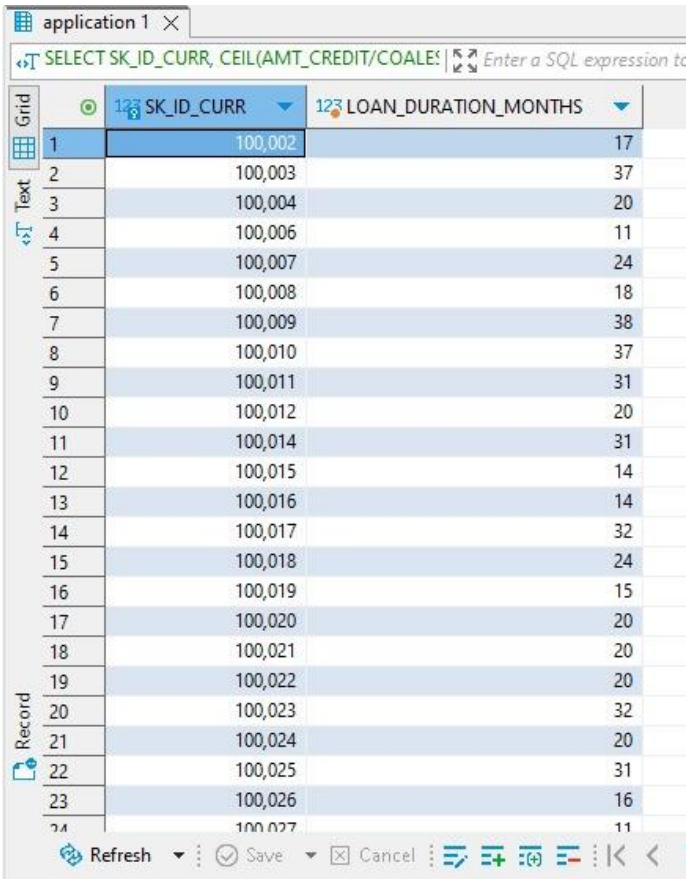
Task 4 Deduce the Loan Duration

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:

</talentlabs>

```
SELECT SK_ID_CURR,
CEIL(AMT_CREDIT/COALESCE(AMT_ANNUITY,1)) AS LOAN_DURATION_MONTHS
FROM application
```



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:

AMT_INCOME_TOTAL factor affecting **EXT_SOURCE_1**, **EXT_SOURCE_2**, **EXT_SOURCE_3**

```
WITH avg_values AS
    (SELECT
```

```
        AVG(EXT_SOURCE_1) AS avg_ext_source_1,
        AVG(EXT_SOURCE_2) AS avg_ext_source_2,
        AVG(EXT_SOURCE_3) AS avg_ext_source_3
    FROM application
    WHERE EXT_SOURCE_1 IS NOT NULL AND EXT_SOURCE_2 IS NOT NULL AND EXT_SOURCE_3 IS
NOT NULL)
SELECT
    CASE
        WHEN AMT_INCOME_TOTAL < 20000 THEN 'Under 20K'
        WHEN AMT_INCOME_TOTAL BETWEEN 20000 AND 49999 THEN '20K-49K'
        WHEN AMT_INCOME_TOTAL BETWEEN 50000 AND 79999 THEN '50K-79K'
        WHEN AMT_INCOME_TOTAL BETWEEN 80000 AND 109999 THEN '80K-109K'
        ELSE '110K+'
    END AS total_income,
    ROUND(AVG(COALESCE(EXT_SOURCE_1, avg_values.avg_ext_source_1)), 2) AS
credit_score_1,
    ROUND(AVG(COALESCE(EXT_SOURCE_2, avg_values.avg_ext_source_2)), 2) AS
credit_score_2,
    ROUND(AVG(COALESCE(EXT_SOURCE_3, avg_values.avg_ext_source_3)), 2) AS
credit_score_3
FROM
    application, avg_values
GROUP BY
    total_income
ORDER BY
    MIN(AMT_INCOME_TOTAL);
```

Results 1

WITH avg_values AS (SELECT AVG(EXT_SOURCE_ | Enter a SQL expression to filter results (use Ctrl+Space)

Grid | Text

	total_income	credit_score_1	credit_score_2	credit_score_3
1	20K-49K	0.52	0.45	0.55
2	50K-79K	0.51	0.47	0.54
3	80K-109K	0.5	0.48	0.52
4	110K+	0.51	0.53	0.5

Explanation:

Since there are null values in EXT_SOURCE_1, EXT_SOURCE_2 and EXT_SOURCE_3 columns, we decided to use COALESCE to fill the empty part with the existing average credit scores respectively.

The table shows the average EXT_SOURCE_1, EXT_SOURCE_2, and EXT_SOURCE_3 values for each credit score that are affected by factor such as client's total income.

As AMT_INCOME_TOTAL increases, the averages of EXT_SOURCE_1 and EXT_SOURCE_3 decrease or barely increase while EXT_SOURCE_2 slightly increase as the income bracket increases. This may suggest that AMT_INCOME_TOTAL (income) has a positive relationship with particularly

EXT_SOURCE_2 while having little to no correlation at all with EXT_SOURCE_1 and EXT_SOURCE_3.

In overall, if EXT_SOURCE_1 and EXT_SOURCE_3 decrease or barely increase as income increases, these sources are likely influenced by factors independent of income, suggesting they may capture other aspects of creditworthiness or risk.

EXT_SOURCE_2's steady increase with income suggests a more direct link between higher income and reduced risk, implying that this factor strongly weighs financial stability.

To sum up, the external sources appear to assess different dimensions of risk, with EXT_SOURCE_2 potentially focusing more on financial factors like income, while EXT_SOURCE_1 and EXT_SOURCE_3 are driven by other considerations.

CODE_GENDER factor affecting **EXT_SOURCE_1**, **EXT_SOURCE_2**, **EXT_SOURCE_3**

```
WITH avg_values AS
    (SELECT
        AVG(EXT_SOURCE_1) AS avg_ext_source_1,
        AVG(EXT_SOURCE_2) AS avg_ext_source_2,
        AVG(EXT_SOURCE_3) AS avg_ext_source_3
    FROM application
    WHERE EXT_SOURCE_1 IS NOT NULL AND EXT_SOURCE_2 IS NOT NULL AND EXT_SOURCE_3 IS
NOT NULL)
SELECT CODE_GENDER AS gender,
    ROUND(AVG(COALESCE(EXT_SOURCE_1, avg_values.avg_ext_source_1)), 2) AS
credit_score_1,
    ROUND(AVG(COALESCE(EXT_SOURCE_2, avg_values.avg_ext_source_2)), 2) AS
credit_score_2,
    ROUND(AVG(COALESCE(EXT_SOURCE_3, avg_values.avg_ext_source_3)), 2) AS
credit_score_3
FROM
    application, avg_values
WHERE
    CODE_GENDER != 'XNA'
GROUP BY
    CODE_GENDER;
```

application 1

WITH coalesce_fix_null_values AS (SELECT AVG(I | Enter a SQL expression to filter results (use Ctrl+Space)

Grid

	gender	credit_score_1	credit_score_2	credit_score_3
1	M	0.47	0.51	0.5
2	F	0.53	0.52	0.51

Text

Explanation:

</talentlabs>

The table shows the average EXT_SOURCE_1, EXT_SOURCE_2, and EXT_SOURCE_3 values for each credit score that are affected by factor such as client's gender.

'XNA' value in CODE_GENDER column means that the value is either missing or not available and there's only 4 of it out of the whole thing. This means that the missing data does not significantly affect the analysis hence we decided to exclude it from the calculation as shown above in WHERE condition.

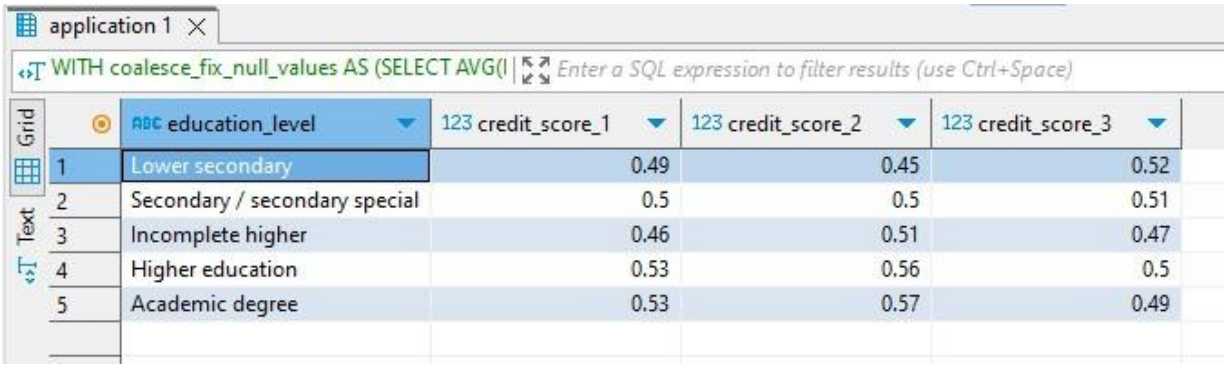
The result shows that the average for female gender, 'F' are higher compared to male, 'M' for all credit scores; EXT_SOURCE_1, EXT_SOURCE_2 and EXT_SOURCE_3.

In conclusion, females are perceived as having a lower credit risk compared to males, as indicated by their higher average scores across all external sources. This may be attributed to differences in financial behavior, such as more conservative borrowing habits, better repayment discipline, or other factors that contribute to a higher creditworthiness in the eyes of the external scoring models. This trend holds consistently across the three sources, implying a broad recognition of these gender-based risk assessments.

Thus, gender appears to be a significant factor influencing external risk assessments, with females generally receiving more favorable evaluations.

NAME_EDUCATION_TYPE factor affecting **EXT_SOURCE_1**, **EXT_SOURCE_2**, **EXT_SOURCE_3**

```sql
WITH avg_values AS
    (SELECT
        AVG(EXT_SOURCE_1) AS avg_ext_source_1,
        AVG(EXT_SOURCE_2) AS avg_ext_source_2,
        AVG(EXT_SOURCE_3) AS avg_ext_source_3
    FROM application
    WHERE EXT_SOURCE_1 IS NOT NULL AND EXT_SOURCE_2 IS NOT NULL AND EXT_SOURCE_3 IS
NOT NULL)
SELECT NAME_EDUCATION_TYPE AS education_level,
    ROUND(AVG(COALESCE(EXT_SOURCE_1, avg_values.avg_ext_source_1)), 2) AS
credit_score_1,
    ROUND(AVG(COALESCE(EXT_SOURCE_2, avg_values.avg_ext_source_2)), 2) AS
credit_score_2,
    ROUND(AVG(COALESCE(EXT_SOURCE_3, avg_values.avg_ext_source_3)), 2) AS
credit_score_3
FROM
    application, avg_values
GROUP BY
    NAME_EDUCATION_TYPE
ORDER BY
    CASE NAME_EDUCATION_TYPE
        WHEN 'Lower secondary' THEN 1
        WHEN 'Secondary / secondary special' THEN 2
        WHEN 'Incomplete higher' THEN 3
        WHEN 'Higher education' THEN 4
        ELSE 5
    END;
```



	education_level	credit_score_1	credit_score_2	credit_score_3
1	Lower secondary	0.49	0.45	0.52
2	Secondary / secondary special	0.5	0.5	0.51
3	Incomplete higher	0.46	0.51	0.47
4	Higher education	0.53	0.56	0.5
5	Academic degree	0.53	0.57	0.49

Explanation:

According to the result shown above for average credit score 1, 2 and 3, those with higher education and academic degrees have relatively higher credit scores across all three metrics Applicants with lower secondary education tend to have lower credit scores and this is especially true with credit score 1 and 2 but not so much for credit score 3. Interestingly, people with incomplete higher education show somewhat lower credit scores, especially for credit score 3 (0.47), suggesting that not completing higher education may affect creditworthiness negatively.

The result for both credit score 1 and 2 is as expected as applicants who have higher or finished education have better credit score than those who don't or has not completed it. Credit score 3 yield unexpected result however with applicants with the lowest education level having better credit score compared to other higher education level. This may suggest that credit score 3 is not heavily reliant on applicant's education level but capturing other aspects of creditworthiness or risk instead. Credit score 2's steady increase with education level suggests a more direct link between higher level of education and reduced risk, implying that this factor strongly weighs educational achievement.

As an overall view, education level appears to have a positive correlation with credit score, where individuals with higher education (particularly a degree) are associated with higher credit scores, which could suggest they have better financial stability or a lower likelihood of defaulting on loans. However, those with incomplete higher education or lower educational attainment have lower credit scores, which may imply a higher risk of payment difficulties.

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

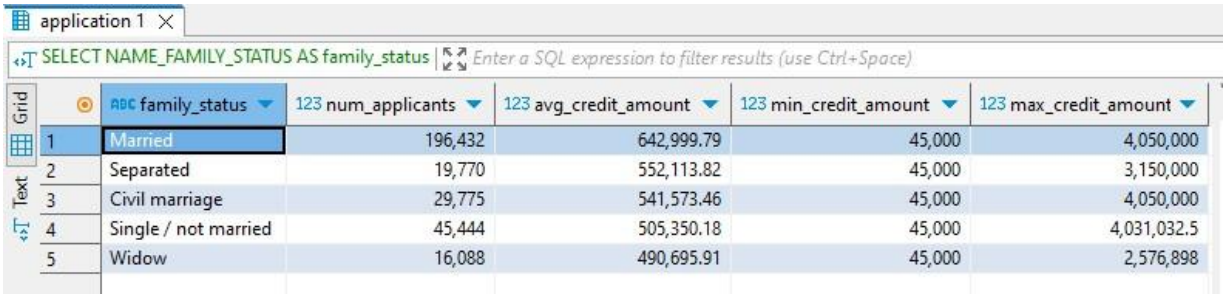
Who is going to lend more money than others? In this task, we want to see are there any factors affecting the credit amount. **Do the analysis of at least 3 factors**

Please explain your findings with SQL statements and results:

NAME_FAMILY_STATUS factor affecting **AMT_CREDIT**

```
SELECT
    NAME_FAMILY_STATUS AS family_status,
    COUNT(*) AS num_applicants,
    ROUND(AVG(AMT_CREDIT),2) AS avg_credit_amount,
```

```
    MIN(AMT_CREDIT) AS min_credit_amount,
    MAX(AMT_CREDIT) AS max_credit_amount
FROM
    application
WHERE
    NAME_FAMILY_STATUS != 'Unknown'
GROUP BY
    NAME_FAMILY_STATUS
ORDER BY
    avg_credit_amount DESC;
```



application 1

SELECT NAME_FAMILY_STATUS AS family_status | Enter a SQL expression to filter results (use Ctrl+Space)

Grid / Text

	family_status	num_applicants	avg_credit_amount	min_credit_amount	max_credit_amount
1	Married	196,432	642,999.79	45,000	4,050,000
2	Separated	19,770	552,113.82	45,000	3,150,000
3	Civil marriage	29,775	541,573.46	45,000	4,050,000
4	Single / not married	45,444	505,350.18	45,000	4,031,032.5
5	Widow	16,088	490,695.91	45,000	2,576,898

Explanation:

The result shown above displays clients who are married are more likely to get loans from banks/lenders based on the highest average of credit amount compared to other family status. This suggests that banks or lenders are more likely to lend higher amounts to married applicants.The reasoning could be that married individuals are seen as more financially stable, as they may have dual incomes or more established households. This also correlates with how they makes up the most numbers of applicants out of other family status.

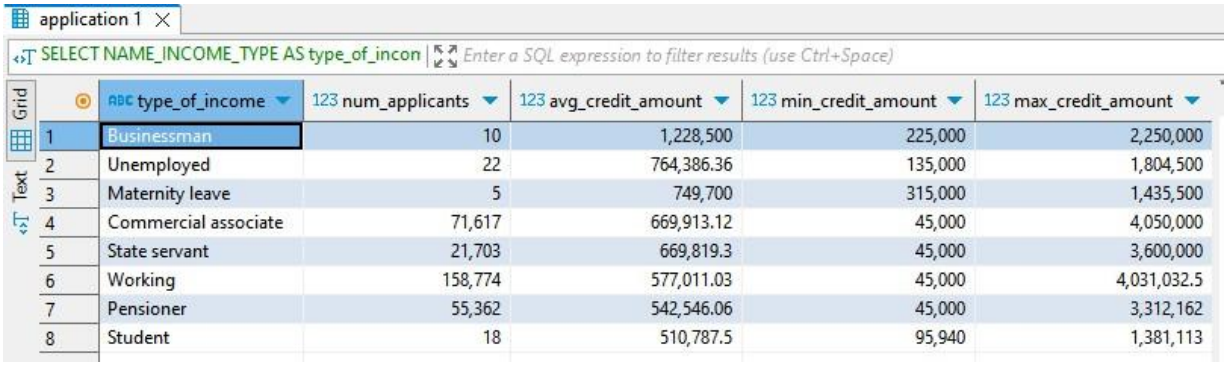
Furthermore, result shows that widowers who makes up the least number of applicants are also the one having the lowest average of credit amount. This could indicate that lenders view them as higher-risk due to possible financial strain after the death of their loved ones. They may be considered to have higher financial risks due to a single source of income.

Single/not married, Separated, and Civil Marriage applicants fall in the mid-range of the average credit amount. These applicants might receive loans based on different criteria, such as their financial status, family situation, or household income.

NAME_INCOME_TYPE factor affecting **AMT_CREDIT**

```
SELECT
    NAME_INCOME_TYPE AS type_of_income,
    COUNT(*) AS num_applicants,
    ROUND(AVG(AMT_CREDIT),2) AS avg_credit_amount,
    MIN(AMT_CREDIT) AS min_credit_amount,
    MAX(AMT_CREDIT) AS max_credit_amount
FROM
    application
GROUP BY
    NAME_INCOME_TYPE
```

```
ORDER BY
    avg_credit_amount DESC;
```



application 1

SELECT NAME_INCOME_TYPE AS type_of_incon | Enter a SQL expression to filter results (use Ctrl+Space)

Grid | Text

	type_of_income	num_applicants	avg_credit_amount	min_credit_amount	max_credit_amount
1	Businessman	10	1,228,500	225,000	2,250,000
2	Unemployed	22	764,386.36	135,000	1,804,500
3	Maternity leave	5	749,700	315,000	1,435,500
4	Commercial associate	71,617	669,913.12	45,000	4,050,000
5	State servant	21,703	669,819.3	45,000	3,600,000
6	Working	158,774	577,011.03	45,000	4,031,032.5
7	Pensioner	55,362	542,546.06	45,000	3,312,162
8	Student	18	510,787.5	95,940	1,381,113

Explanation:

Table shows that even though the number of applicants for businessman is relatively low, they receive the highest average credit amount. This may imply that banks view them as able to handle larger credit amounts, possibly due to the nature of their work and potential assets. Secondly, despite not having a steady income, unemployed individuals receive a higher average credit amount compared to most other categories. This could be surprising and might suggest that specific conditions such as previous financial standing or external factors allow them to qualify for larger loans.

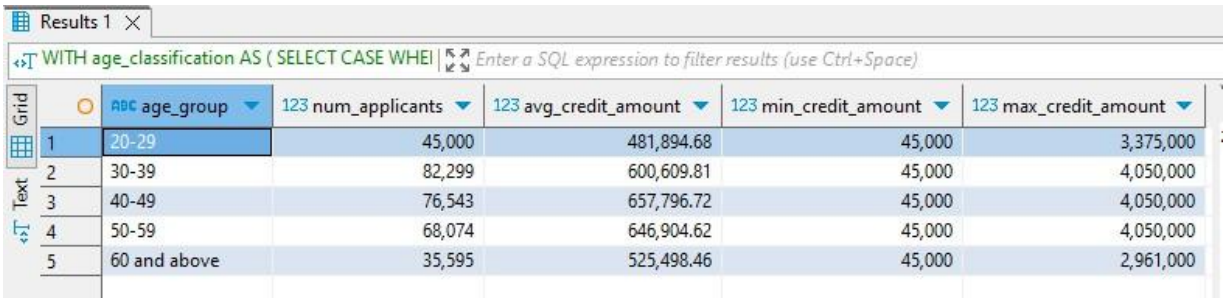
The maternity leave group also receives relatively high credit amounts, possibly due to their employment status or expectations to return to work, yet they make up a small portion of the total applicants. Followed by commercial associate and state servant group in terms of the average credit amount, they both have a middle range in both average credit amount and number of applicants, suggesting that their credit offerings and population size are more balanced.

While working people represent the largest portion of applicants, their average credit amount is relatively moderate compared to other groups like businessman. This indicates that while they are the largest customer base, they may not be receiving the highest loan amounts. For pensioner, despite making up a significant portion of applicants, pensioners tend to receive lower credit amounts, likely due to their fixed income. Lastly, students make up a small portion of applicants and receive the lowest average credit amount, which could be expected due to their limited earning capacity and financial experience.

DAYS_BIRTH factor affecting **AMT_CREDIT**

```
WITH age_classification AS (
    SELECT
        CASE
            WHEN FLOOR(ABS(DAYS_BIRTH)/365) BETWEEN 20 AND 29 THEN '20-29'
            WHEN FLOOR(ABS(DAYS_BIRTH)/365) BETWEEN 30 AND 39 THEN '30-39'
            WHEN FLOOR(ABS(DAYS_BIRTH)/365) BETWEEN 40 AND 49 THEN '40-49'
            WHEN FLOOR(ABS(DAYS_BIRTH)/365) BETWEEN 50 AND 59 THEN '50-59'
            ELSE '60 and above'
        END AS age_group,
        AMT_CREDIT
```

```
    FROM application)
SELECT
    age_group,
    COUNT(*) AS num_applicants,
    ROUND(AVG(AMT_CREDIT), 2) AS avg_credit_amount,
    MIN(AMT_CREDIT) AS min_credit_amount,
    MAX(AMT_CREDIT) AS max_credit_amount
FROM
    age_classification
GROUP BY
    age_group
ORDER BY
    age_group ASC;
```


Results 1

WITH age_classification AS (SELECT CASE WHEI | Enter a SQL expression to filter results (use Ctrl+Space)

Grid | Text

	age_group	num_applicants	avg_credit_amount	min_credit_amount	max_credit_amount
1	20-29	45,000	481,894.68	45,000	3,375,000
2	30-39	82,299	600,609.81	45,000	4,050,000
3	40-49	76,543	657,796.72	45,000	4,050,000
4	50-59	68,074	646,904.62	45,000	4,050,000
5	60 and above	35,595	525,498.46	45,000	2,961,000

Explanation:

According to the table above, age group 40-49 years old are more likely to get loans from banks or lenders. Since this group typically represents people who are already in a stable career positions, may have already purchased homes or other significant assets, this enables them to secure higher loans.

On the opposite, young adults in this age group are likely in the early stages of their careers. They may be taking smaller loans, either for starting their careers, purchasing their first car, or renting property. Their financial stability may not be fully established yet, which is why their average credit amount is the lowest compared to the older groups.

Despite age group 30-39 years old making up the most number of applicants, they falls in the middle in terms of their average credit amount. This is probably because people in their 30s have different financial needs, with some borrowing smaller amounts for things like cars or personal projects, while others taking larger loans like mortgages. Overall, the group tends to borrow moderate amounts on average, possibly due to varying incomes or being cautious with their borrowing.

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:

</talentlabs>

DAYS_EMPLOYED factor affecting **TARGET**

```
WITH employment_period AS (
    SELECT
        CASE
            WHEN FLOOR(ABS(DAYS_EMPLOYED)/365) <= 1 THEN '1 year and under'
            WHEN FLOOR(ABS(DAYS_EMPLOYED)/365) BETWEEN 2 AND 5 THEN '2-5 years'
            WHEN FLOOR(ABS(DAYS_EMPLOYED)/365) BETWEEN 6 AND 9 THEN '6-9 years'
            WHEN FLOOR(ABS(DAYS_EMPLOYED)/365) BETWEEN 10 AND 13 THEN '10-13 years'
            ELSE '14 years and above'
        END AS period_employed,
        TARGET
    FROM application)
SELECT
    period_employed,
    COUNT(*) AS num_applicants,
    COUNT(CASE WHEN TARGET = 1 THEN 1 END) AS YES_payment_difficulty,
    COUNT(CASE WHEN TARGET = 0 THEN 1 END) AS NO_payment_difficulty,
    ROUND(COUNT(CASE WHEN TARGET = 1 THEN 1 END) * 100.0 / COUNT(*), 2) AS
YES_percent_pd,
    ROUND(COUNT(CASE WHEN TARGET = 0 THEN 1 END) * 100.0 / COUNT(*), 2) AS
NO_percent_pd
FROM
    employment_period
GROUP BY
    period_employed
ORDER BY
    CASE period_employed
        WHEN '1 year and under' THEN 1
        WHEN '2-5 years' THEN 2
        WHEN '6-9 years' THEN 3
        WHEN '10-13 years' THEN 4
        ELSE 5
    END;
```

Results 1

WITH employment_period AS (SELECT CASE WI | Enter a SQL expression to filter results (use Ctrl+Space)

	period_employed	num_applicants	YES_payment_difficulty	NO_payment_difficulty	YES_percent_pd	NO_percent_pd
1	1 year and under	59,745	6,696	53,049	11.21	88.79
2	2-5 years	92,793	9,034	83,759	9.74	90.26
3	6-9 years	48,619	3,460	45,159	7.12	92.88
4	10-13 years	22,924	1,337	21,587	5.83	94.17
5	14 years and above	83,430	4,298	79,132	5.15	94.85

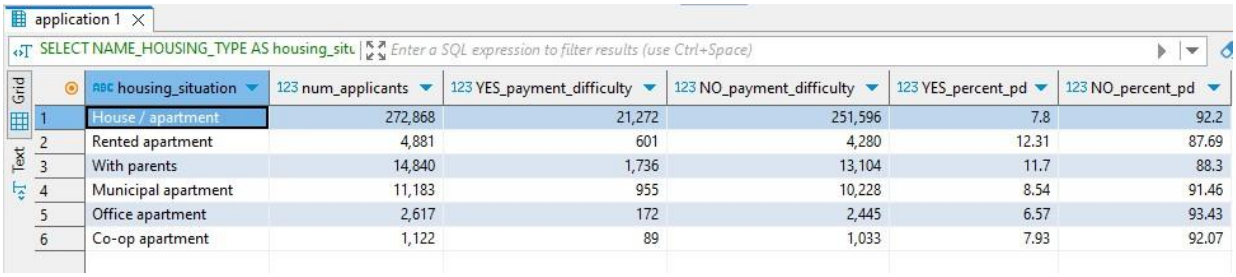
Explanation:

Based on the table above, those who have been employed for less than or equal to one year show a relatively high percentage of payment difficulties (11.2%) compared to other employment period. This suggest that newly employed individuals are more likely of having difficulty in loan repayment.

On the opposite, those who have been employed for 14 years or more have less difficulty in repaying the loan. This overall imply that longer employment periods generally correlate with better financial stability and loan repayment capabilities. It is safe to say that one's period of employment have a positive relationship with one's capability of repaying back the loan.

NAME_HOUSING_TYPE factor affecting **TARGET**

```
SELECT
    NAME_HOUSING_TYPE AS housing_situation,
    COUNT(*) AS num_applicants,
    COUNT(CASE WHEN TARGET = 1 THEN 1 END) AS YES_payment_difficulty,
    COUNT(CASE WHEN TARGET = 0 THEN 1 END) AS NO_payment_difficulty,
    ROUND(COUNT(CASE WHEN TARGET = 1 THEN 1 END) * 100.0 / COUNT(*), 2) AS
YES_percent_pd,
    ROUND(COUNT(CASE WHEN TARGET = 0 THEN 1 END) * 100.0 / COUNT(*), 2) AS
NO_percent_pd
FROM
    application
GROUP BY
    NAME_HOUSING_TYPE
ORDER BY
    num_applicants DESC;
```



application 1

SELECT NAME_HOUSING_TYPE AS housing_situ | Enter a SQL expression to filter results (use Ctrl+Space)

Grid | Text

	housing_situation	num_applicants	YES_payment_difficulty	NO_payment_difficulty	YES_percent_pd	NO_percent_pd
1	House / apartment	272,868	21,272	251,596	7.8	92.2
2	Rented apartment	4,881	601	4,280	12.31	87.69
3	With parents	14,840	1,736	13,104	11.7	88.3
4	Municipal apartment	11,183	955	10,228	8.54	91.46
5	Office apartment	2,617	172	2,445	6.57	93.43
6	Co-op apartment	1,122	89	1,033	7.93	92.07

Explanation:

Result above suggests that number of applicants have no correlation with client's ability to pay back the loan. We can see that those who live in a rented apartment have the biggest difficulty in paying back the loan considering the high percentage in payment difficulty This makes sense since they have to allocate their money/income into the rent as well as paying monthly commitments such as loan and other necessary spending.

Moreover, those who live in an office apartment can be seen to have better financial situation compared to other housing situations since they have the the least difficulty in repaying the loan. This may be the case as they didn't have to worry about extra spending going to the housing situation as their accommodation may be provided by the company they worked at.

FLAG_OWN_REALTY factor affecting **TARGET**

```
WITH property_ownership AS
    (SELECT
       CASE
           WHEN FLAG_OWN_REALTY = 'Y' THEN 'OWN_house_or_flat'
           ELSE 'NO_house_or_flat'
       END AS having_property,
       TARGET
    FROM application)
SELECT
    having_property,
    COUNT(CASE WHEN TARGET = 1 THEN 1 END) AS YES_payment_difficulty,
    COUNT(CASE WHEN TARGET = 0 THEN 1 END) AS NO_payment_difficulty,
    ROUND(COUNT(CASE WHEN TARGET = 1 THEN 1 END) * 100.0 / COUNT(*), 2) AS
YES_percent_pd,
    ROUND(COUNT(CASE WHEN TARGET = 0 THEN 1 END) * 100.0 / COUNT(*), 2) AS
NO_percent_pd
FROM
    property_ownership
GROUP BY
    having_property
```

Results 1

WITH have_property AS (SELECT CASE WHEN FL | Enter a SQL expression to filter results (use Ctrl+Space)

	having_property	YES_payment_difficulty	NO_payment_difficulty	YES_percent_pd	NO_percent_pd
1	OWN_house_or_flat	16,983	196,329	7.96	92.04
2	NO_house_or_flat	7,842	86,357	8.32	91.68

Explanation:

Based on the table above, we can see that those who owned a house or a flat have less difficulty in paying back the loan compared to those who doesn't own any. They also make up a big portion of the number of applicants. Bank/lenders may see them as financially stabled since owning a home or flat often correlates with greater financial stability, access to credit, lower housing costs, and a more secure and stable life situation. These factors combined help homeowners manage their loan repayments more effectively than those who do not own property.

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 7 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:

```sql
WITH loan_per_id AS
    (SELECT SK_ID_CURR, COUNT(SK_ID_CURR) AS loan_count
    FROM bureau
    GROUP BY SK_ID_CURR)
```

```sql
SELECT
    CASE
        WHEN loan_count BETWEEN 1 AND 10 THEN '1-10 loans'
        WHEN loan_count BETWEEN 11 AND 20 THEN '11-20 loans'
        WHEN loan_count BETWEEN 21 AND 30 THEN '21-30 loans'
        WHEN loan_count BETWEEN 31 AND 40 THEN '31-40 loans'
        ELSE 'Over 40 loans'
    END AS number_of_loans,
    ROUND(AVG(loan_count), 2) AS avg_loan_count,
    COUNT(application.SK_ID_CURR) AS num_applicants,
    COUNT(CASE WHEN TARGET = 1 THEN 1 END) AS YES_payment_difficulty,
    COUNT(CASE WHEN TARGET = 0 THEN 1 END) AS NO_payment_difficulty,
    ROUND(COUNT(CASE WHEN TARGET = 1 THEN 1 END) * 100.0 / COUNT(*), 2) AS
YES_percent_pd,
    ROUND(COUNT(CASE WHEN TARGET = 0 THEN 1 END) * 100.0 / COUNT(*), 2) AS
NO_percent_pd
FROM loan_per_id
INNER JOIN application
ON application.SK_ID_CURR = loan_per_id.SK_ID_CURR
GROUP BY number_of_loans
```

Results 1

WITH loan_per_id AS (SELECT SK_ID_CURR, COI | Enter a SQL expression to filter results (use Ctrl+Space)

	num_of_loans	avg_loan_count	num_applicants	YES_payment_difficulty	NO_payment_difficulty	YES_percent_pd	NO_percent_pd
1	1-10 loans	4.31	231,214	17,727	213,487	7.67	92.33
2	11-20 loans	13.62	29,808	2,365	27,443	7.93	92.07
3	21-30 loans	23.58	2,215	247	1,968	11.15	88.85
4	31-40 loans	33.62	205	23	182	11.22	88.78
5	Over 40 loans	49.96	49	6	43	12.24	87.76

Task 8 FreeStyle

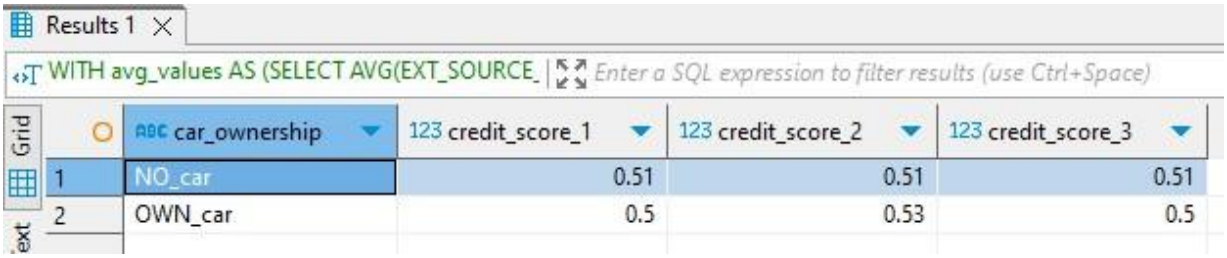
Now, conduct your own research and analysis to see what factors from the “application” and the “bureau” tables are affecting

- The Credit Scores

FLAG_OWN_CAR factor from “application” table affecting **CREDIT SCORES**

```sql
WITH avg_values AS
    (SELECT
        AVG(EXT_SOURCE_1) AS avg_ext_source_1,
        AVG(EXT_SOURCE_2) AS avg_ext_source_2,
        AVG(EXT_SOURCE_3) AS avg_ext_source_3
    FROM application
    WHERE EXT_SOURCE_1 IS NOT NULL AND EXT_SOURCE_2 IS NOT NULL AND EXT_SOURCE_3 IS
```

```
NOT NULL)
SELECT
    CASE
        WHEN FLAG_OWN_CAR = 'Y' THEN 'OWN_car'
        ELSE 'NO_car'
    END AS car_ownership,
    ROUND(AVG(COALESCE(EXT_SOURCE_1, avg_values.avg_ext_source_1)), 2) AS
credit_score_1,
    ROUND(AVG(COALESCE(EXT_SOURCE_2, avg_values.avg_ext_source_2)), 2) AS
credit_score_2,
    ROUND(AVG(COALESCE(EXT_SOURCE_3, avg_values.avg_ext_source_3)), 2) AS
credit_score_3
FROM
    application, avg_values
GROUP BY
    car_ownership
```



Explanation:

Result shows that those who own a car in EXT_SOURCE_1 and EXT_SOURCE_3 show very similar averages for both groups, suggesting that car ownership does not have a strong influence on these credit scores. EXT_SOURCE_2 shows a noticeable difference between non-car owners and car owners, with car owners having a higher average credit score. This suggests that in this category, owning a car may be linked to better credit scores.

Car owners showing slightly better performance in one of the credit scores (EXT_SOURCE_2) might indicate a slight trend in creditworthiness among car owners compared to non-owners.

AMT_CREDIT_SUM_OVERDUE factor from “bureau” table affecting **CREDIT SCORES**

```
WITH avg_values AS
    (SELECT
        AVG(EXT_SOURCE_1) AS avg_ext_source_1,
        AVG(EXT_SOURCE_2) AS avg_ext_source_2,
        AVG(EXT_SOURCE_3) AS avg_ext_source_3
    FROM application
    WHERE EXT_SOURCE_1 IS NOT NULL AND EXT_SOURCE_2 IS NOT NULL AND EXT_SOURCE_3 IS
NOT NULL)
SELECT
    CASE
```

```
        WHEN AMT_CREDIT_SUM_OVERDUE < 20000 THEN 'Under 20K'
        WHEN AMT_CREDIT_SUM_OVERDUE BETWEEN 20000 AND 49999 THEN '20K-49K'
        WHEN AMT_CREDIT_SUM_OVERDUE BETWEEN 50000 AND 79999 THEN '50K-79K'
        WHEN AMT_CREDIT_SUM_OVERDUE BETWEEN 80000 AND 109999 THEN '80K-109K'
        ELSE '110K+'
    END AS total_credit_overdue,
    ROUND(AVG(COALESCE(EXT_SOURCE_1, avg_values.avg_ext_source_1)), 2) AS
credit_score_1,
    ROUND(AVG(COALESCE(EXT_SOURCE_2, avg_values.avg_ext_source_2)), 2) AS
credit_score_2,
    ROUND(AVG(COALESCE(EXT_SOURCE_3, avg_values.avg_ext_source_3)), 2) AS
credit_score_3
FROM application
INNER JOIN bureau
ON application.SK_ID_CURR = bureau.SK_ID_CURR
CROSS JOIN avg_values
WHERE AMT_CREDIT_SUM_OVERDUE > 0
GROUP BY total_credit_overdue
ORDER BY MIN(AMT_CREDIT_SUM_OVERDUE)
```

Results 1

WITH avg_values AS (SELECT AVG(EX | Enter a SQL expression to filter results (use Ctrl+Space)

Grid | Text

	total_credit_overdue	credit_score_1	credit_score_2	credit_score_3
1	Under 20K	0.5	0.49	0.28
2	20K-49K	0.45	0.44	0.17
3	50K-79K	0.47	0.46	0.2
4	80K-109K	0.51	0.48	0.15
5	110K+	0.45	0.45	0.17

Explanation:

Since we're focusing on the one who currently have credit sum overdue, we included the WHERE condition which exclude those who do not have any credit overdue even though they makes up a big part of the data.

Result above shows that the average for all credit score 1, 2, 3 does not have any correlation with the total amount credit overdue since the higher the total amount credit overdue, the credit score average will either increase or decrease randomly.

The average credit score for 1 and 2 would likely be considered average to slightly below average. It may suggest the applicant is not highly risky, but not very safe either. They probably fall in the moderate risk category.

Credit score 3 in overall however have a very low average credit score which may suggest that the applicant is likely to have poor credit history or financial stability, or is otherwise considered risky by external sources. Such a low score can indicate that the applicant may struggle with repayment, and lenders would likely consider them high risk for defaulting on loans.

- The Payment Difficulty

AMT_INCOME factor from “application” table affecting **TARGET**

```
SELECT
    CASE
        WHEN AMT_INCOME_TOTAL < 20000 THEN 'Under 20K'
        WHEN AMT_INCOME_TOTAL BETWEEN 20000 AND 49999 THEN '20K-49K'
        WHEN AMT_INCOME_TOTAL BETWEEN 50000 AND 79999 THEN '50K-79K'
        WHEN AMT_INCOME_TOTAL BETWEEN 80000 AND 109999 THEN '80K-109K'
        ELSE '110K+'
    END AS total_income,
    COUNT(CASE WHEN TARGET = 1 THEN 1 END) AS YES_payment_difficulty,
    COUNT(CASE WHEN TARGET = 0 THEN 1 END) AS NO_payment_difficulty,
    ROUND(COUNT(CASE WHEN TARGET = 1 THEN 1 END) * 100.0 / COUNT(*), 2) AS
YES_percent_pd,
    ROUND(COUNT(CASE WHEN TARGET = 0 THEN 1 END) * 100.0 / COUNT(*), 2) AS
NO_percent_pd
FROM
    application
GROUP BY
    total_income
ORDER BY
    MIN(AMT_INCOME_TOTAL);
```

Results 1

SELECT CASE WHEN AMT_INCOME_TOTAL < 20(| Enter a SQL expression to filter results (use Ctrl+Space)

	total_income	YES_payment_difficulty	NO_payment_difficulty	YES_percent_pd	NO_percent_pd
1	20K-49K	343	4,174	7.59	92.41
2	50K-79K	1,826	21,020	7.99	92.01
3	80K-109K	3,542	38,518	8.42	91.58
4	110K+	19,114	218,974	8.03	91.97

Explanation:

Table shows that higher income groups (80K-109K and 110K+) have slightly higher percentages of payment difficulties compared to the lower income groups. Lower-income groups (20K-49K and 50K-79K) tend to have slightly lower percentages of payment difficulty, which might indicate that loan terms for lower-income groups are more conservative or that these applicants are more cautious with borrowing.

Overall, this suggests that income level alone may not be a strong predictor of payment difficulty, as the percentages of those with difficulties are rather close across income groups. However, those in the 80K-109K income bracket have a slightly higher percentage of difficulties, possibly indicating that they are more likely to take on larger loans or have more financial commitments accordance to their income.

</talentlabs>

CREDIT_DAY_OVERDUE factor from “bureau” table affecting **TARGET**

```
SELECT
    CASE
        WHEN CREDIT_DAY_OVERDUE BETWEEN 1 AND 50 THEN '1-50 days'
        WHEN CREDIT_DAY_OVERDUE BETWEEN 51 AND 100 THEN '51-100 days'
        WHEN CREDIT_DAY_OVERDUE BETWEEN 101 AND 150 THEN '101-150 days'
        WHEN CREDIT_DAY_OVERDUE BETWEEN 151 AND 200 THEN '151-200 days'
        ELSE 'Over 200 days'
    END AS credit_days_overdue,
    COUNT(*) AS num_applicants,
    COUNT(CASE WHEN TARGET = 1 THEN 1 END) AS YES_payment_difficulty,
    COUNT(CASE WHEN TARGET = 0 THEN 1 END) AS NO_payment_difficulty,
    ROUND(COUNT(CASE WHEN TARGET = 1 THEN 1 END) * 100.0 / COUNT(*), 2) AS
YES_percent_pd,
    ROUND(COUNT(CASE WHEN TARGET = 0 THEN 1 END) * 100.0 / COUNT(*), 2) AS
NO_percent_pd
FROM bureau
INNER JOIN application
ON application.SK_ID_CURR = bureau.SK_ID_CURR
WHERE CREDIT_DAY_OVERDUE > 0
GROUP BY credit_days_overdue
ORDER BY MIN(CREDIT_DAY_OVERDUE);
```

Results 1

SELECT CASE WHEN CREDIT_DAY_OV | Enter a SQL expression to filter results (use Ctrl+Space)

	credit_days_overdue	num_applicants	YES_payment_difficulty	NO_payment_difficulty	YES_percent_pd	NO_percent_pd
1	1-50 days	2,141	449	1,692	20.97	79.03
2	51-100 days	453	90	363	19.87	80.13
3	101-150 days	156	20	136	12.82	87.18
4	151-200 days	67	17	50	25.37	74.63
5	Over 200 days	889	99	790	11.14	88.86

Explanation:

As expected, result shows that the longer the overdue period, the higher the chance of encountering payment difficulties, particularly in the 151-200 days range. Interestingly, the over 200 days group does not have as high a percentage of difficulties as the 151-200 days group.

Overall, credit overdue days can be a strong indicator of potential payment issues, with different thresholds (especially around 151-200 days) needing closer observation.