Cleaning data

--Here I will check for columns that contain NULL values, which can affect your analysis and may need to be addressed.

SELECT

COUNT(\*) AS total\_rows,

SUM(CASE WHEN credit\_policy IS NULL THEN 1 ELSE 0 END) AS missing\_credit\_policy,

SUM(CASE WHEN purpose IS NULL THEN 1 ELSE 0 END) AS missing\_purpose,

SUM(CASE WHEN int\_rate IS NULL THEN 1 ELSE 0 END) AS missing\_int\_rate,

SUM(CASE WHEN installment IS NULL THEN 1 ELSE 0 END) AS missing\_installment,

SUM(CASE WHEN log\_annual\_inc IS NULL THEN 1 ELSE 0 END) AS missing\_log\_annual\_inc,

SUM(CASE WHEN dti IS NULL THEN 1 ELSE 0 END) AS missing\_dti,

SUM(CASE WHEN fico IS NULL THEN 1 ELSE 0 END) AS missing\_fico,

SUM(CASE WHEN revol\_bal IS NULL THEN 1 ELSE 0 END) AS missing\_revol\_bal,

SUM(CASE WHEN revol\_util IS NULL THEN 1 ELSE 0 END) AS missing\_revol\_util,

SUM(CASE WHEN inq\_last\_6mths IS NULL THEN 1 ELSE 0 END) AS missing\_inq\_last\_6mths,

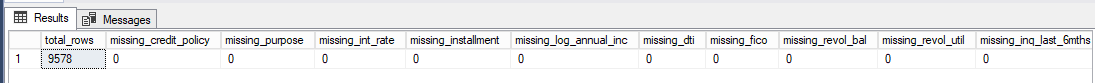
SUM(CASE WHEN delinq\_2yrs IS NULL THEN 1 ELSE 0 END) AS missing\_delinq\_2yrs,

SUM(CASE WHEN pub\_rec IS NULL THEN 1 ELSE 0 END) AS missing\_pub\_rec,

SUM(CASE WHEN not\_fully\_paid IS NULL THEN 1 ELSE 0 END) AS missing\_not\_fully\_paid

FROM

dbo.loandata;



--Here, I will look for data points that don't make sense, such as a FICO score that's too low to be realistic: FICO scores below a reasonable threshold (500) or Interest rates that are negative or too high.

SELECT

COUNT(\*) AS anomalous\_fico\_count

FROM

dbo.loandata

WHERE

fico < 500;

SELECT

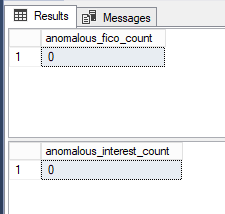
COUNT(\*) AS anomalous\_interest\_count

FROM

dbo.loandata

WHERE

int\_rate < 0 OR int\_rate > 1;



--Here, I want to ensure that the data is consistent. I want to make sure that loans marked as not fully paid have a non-zero balance.

SELECT

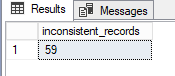
COUNT(\*) AS inconsistent\_records

FROM

dbo.loandata

WHERE

not\_fully\_paid = 1 AND revol\_bal = 0;



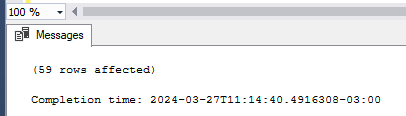
It appears that Ihave identified 59 records in the dbo.loandata table where the not\_fully\_paid column has a value of 1 (which I assume indicates loans that are not fully paid), but at the same time, the revol\_bal (presumably revolving balance) is 0. This could be considered inconsistent since you would expect a non-zero balance for loans that have not been fully paid.

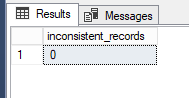
–Assuming I have determined that the inconsistency is indeed an error and the revol\_bal should be greater than 0, I will set it to a NULL.

UPDATE dbo.loandata

SET revol\_bal = NULL

WHERE not\_fully\_paid = 1 AND revol\_bal = 0;





–Duplicate records can skew my results, so I wanted to identify and remove any duplicates from my dataset.

SELECT

credit\_policy, purpose, int\_rate, installment, log\_annual\_inc, dti, fico, days\_with\_cr\_line,

revol\_bal, revol\_util, inq\_last\_6mths, delinq\_2yrs, pub\_rec, not\_fully\_paid, COUNT(\*) AS DuplicateCount

FROM

dbo.loandata

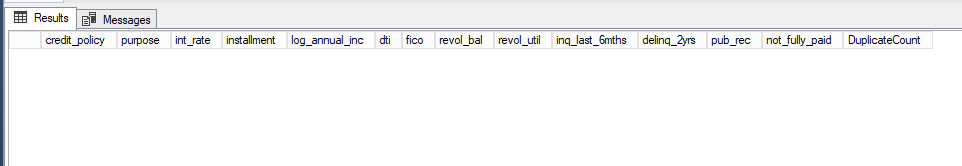
GROUP BY

credit\_policy, purpose, int\_rate, installment, log\_annual\_inc, dti, fico, days\_with\_cr\_line,

revol\_bal, revol\_util, inq\_last\_6mths, delinq\_2yrs, pub\_rec, not\_fully\_paid

HAVING

COUNT(\*) > 1;



–Lets look at the table structure:

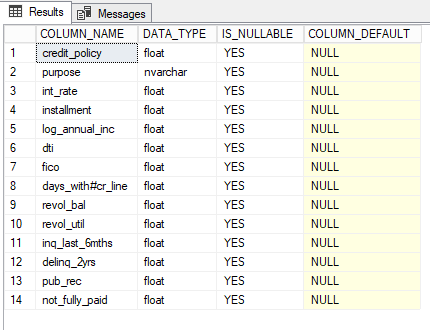
SELECT COLUMN\_NAME, DATA\_TYPE, IS\_NULLABLE, COLUMN\_DEFAULT

FROM INFORMATION\_SCHEMA.COLUMNS

WHERE TABLE\_CATALOG = 'loan\_data'

AND TABLE\_SCHEMA = 'dbo'

AND TABLE\_NAME = 'loandata';



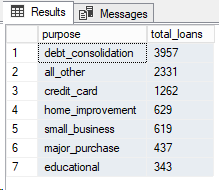
--Lets look into the most common reasons borrowers are seeking loans.

SELECT purpose, COUNT(\*) as total\_loans

FROM dbo.loandata

GROUP BY purpose

ORDER BY total\_loans DESC;



The results from this query indicate that the most common reason borrowers are seeking loans is for debt consolidation, followed by 'all\_other' which likely includes a variety of other reasons not listed separately. Credit card payments, home improvement, and small business are also significant categories, while major purchases and educational expenses are less common reasons for seeking loans. These results give an initial indication of the borrowing patterns and could be the basis for further analysis into the risk profiles associated with each loan purpose, among other potential insights.

–Average Interest Rates by FICO Score

Let’s look at the range of FICO scores, by grouping them into categories (e.g., 650-699, 700-749) to see how average interest rates vary across these ranges.

SELECT

CASE

WHEN fico >= 650 AND fico < 700 THEN '650-699'

WHEN fico >= 700 AND fico < 750 THEN '700-749'

WHEN fico >= 750 AND fico < 800 THEN '750-799'

ELSE '800+'

END AS ficorange,

AVG(int\_rate) as average\_interest\_rate

FROM dbo.loandata

GROUP BY

CASE

WHEN fico >= 650 AND fico < 700 THEN '650-699'

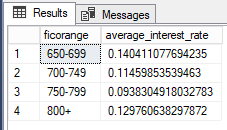
WHEN fico >= 700 AND fico < 750 THEN '700-749'

WHEN fico >= 750 AND fico < 800 THEN '750-799'

ELSE '800+'

END

ORDER BY ficorange;



The results from your query show the average interest rates for loans across different FICO score ranges. Borrowers with FICO scores in the 750-799 range enjoy the lowest average interest rate, while the rates are higher for the other ranges, with the 800+ range having a surprisingly higher average rate than the 700-749 range. This could indicate a non-linear relationship between FICO scores and interest rates, or it might reflect a smaller sample size or different loan characteristics in the 800+ range.

–This query calculates the number of loans not fully paid and the total number of loans for each purpose, along with the proportion of loans not fully paid, offering insights into which loan purposes have higher default rates.

SELECT purpose,

SUM(CASE WHEN not\_fully\_paid = 1 THEN 1 ELSE 0 END) as not\_fully\_paid,

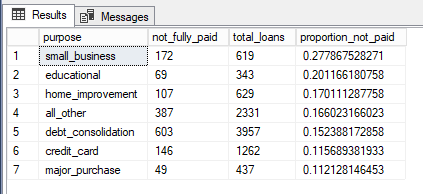
COUNT(\*) as total\_loans,

SUM(CASE WHEN not\_fully\_paid = 1 THEN 1 ELSE 0 END) / COUNT(\*) as proportion\_not\_paid

FROM dbo.loandata

GROUP BY purpose

ORDER BY proportion\_not\_paid DESC;



From the results, it's evident that loans taken out for small business purposes have the highest proportion of not being fully paid, followed by educational and home improvement loans. These insights can inform risk assessments for different loan purposes. The analysis suggests that small business loans might be riskier than those for other purposes, which could be due to various factors, such as the inherent risks in starting or maintaining a small business.

–This query provides a high-level view of loan repayment across the entire dataset, showing how many loans were fully paid versus those that were not.

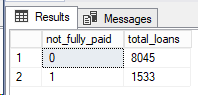
SELECT

not\_fully\_paid,

COUNT(\*) as total\_loans

FROM dbo.loandata

GROUP BY not\_fully\_paid;



This provides a high-level view of the loan performance in your dataset, indicating that roughly 16% of the loans (1533 out of 9578 total loans) are not fully paid.

--Lets get a preliminary understanding of the distribution and central tendencies of our data by running some summary statistics.

SELECT

AVG(fico) AS average\_fico\_score

FROM

Dbo.loandata;



A FICO score of around 710 is generally considered to be a good credit score. This suggests that, on average, borrowers in our dataset are likely to be seen as creditworthy by lenders.

SELECT

MIN(dti) AS min\_dti\_ratio,

MAX(dti) AS max\_dti\_ratio

FROM

Dbo.loandata;



The maximum DTI ratio of about 30% is typically considered acceptable by lenders, though it may be approaching the upper limit of what is considered affordable for borrowers. It’s a common benchmark in the lending industry where lenders prefer borrowers to have a DTI ratio under 36%.

SELECT

AVG(int\_rate) AS average\_interest\_rate

FROM

Dbo.loandata;



An average interest rate of 12.26% gives me a general idea of the cost of borrowing. This rate is above what you would expect for secured lending, like mortgages, but might be typical for unsecured personal loans or credit card borrowing.

SELECT

AVG(revol\_bal) AS average\_loan\_amount,

MIN(revol\_bal) AS min\_loan\_amount,

MAX(revol\_bal) AS max\_loan\_amount

FROM

Dbo.loandata;



The average loan amount gives an indication of the typical size of loans being taken out. In this case, the average is around $17,000, which could suggest that the dataset might include personal loans or smaller business loans rather than large loans like mortgages.

A minimum revolving balance of $0 is normal and expected. It simply indicates that some borrowers have not utilized their revolving credit or have paid off their balances in full.

The maximum revolving balance of $120,735.90 suggests that at least one borrower has a high amount of revolving credit utilized. This could be a high-utilization borrower, possibly with significant credit card debt or a large line of credit.

SELECT

SUM(CASE WHEN not\_fully\_paid = 0 THEN 1 ELSE 0 END) \* 100.0 / COUNT(\*) AS percentage\_fully\_paid,

SUM(CASE WHEN not\_fully\_paid = 1 THEN 1 ELSE 0 END) \* 100.0 / COUNT(\*) AS percentage\_not\_fully\_paid

FROM

Dbo.loandata;



A 16.05% rate of non-repayment is significant and could indicate a moderate level of credit risk within the loan portfolio. Understanding the factors that contribute to this rate could help in developing better credit models or strategies to reduce defaults.

–I am interested to see the relationship between fico score and performance.

SELECT

AVG(fico) AS avg\_fico\_score\_fully\_paid,

COUNT(\*) AS count\_fully\_paid

FROM

dbo.loandata

WHERE

not\_fully\_paid = 0;

SELECT

AVG(fico) AS avg\_fico\_score\_not\_fully\_paid,

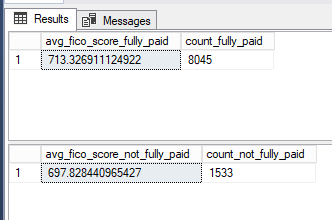
COUNT(\*) AS count\_not\_fully\_paid

FROM

dbo.loandata

WHERE

not\_fully\_paid = 1;



–I wanted to see the relationship between DTI and loan repayment:

SELECT

not\_fully\_paid,

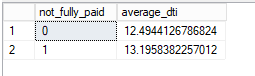
AVG(dti) AS average\_dti

FROM

dbo.loandata

GROUP BY

not\_fully\_paid;



The higher average DTI among loans that have been fully paid supports the general notion that a higher DTI is a positive predictor of loan repayment.

–Now, let analyze interest rate across different fico ranges.

SELECT

CASE

WHEN fico >= 650 AND fico < 700 THEN '650-699'

WHEN fico >= 700 AND fico < 750 THEN '700-749'

WHEN fico >= 750 AND fico < 800 THEN '750-799'

ELSE '800+'

END AS fico\_range,

AVG(int\_rate) AS average\_interest\_rate

FROM

dbo.loandata

GROUP BY

CASE

WHEN fico >= 650 AND fico < 700 THEN '650-699'

WHEN fico >= 700 AND fico < 750 THEN '700-749'

WHEN fico >= 750 AND fico < 800 THEN '750-799'

ELSE '800+'

END

ORDER BY

CASE

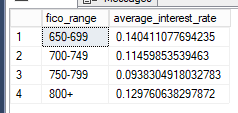
WHEN fico >= 650 AND fico < 700 THEN '650-699'

WHEN fico >= 700 AND fico < 750 THEN '700-749'

WHEN fico >= 750 AND fico < 800 THEN '750-799'

ELSE '800+'

END;



The general trend indicates that those with higher FICO scores tend to receive lower interest rates, which aligns with standard lending practices—better credit scores usually qualify for better loan terms. However, the uptick in interest rate for the 800+ category is an interesting anomaly. This could be due to various factors, such as a smaller sample size for that group, the inclusion of different loan types, or perhaps these borrowers are taking out larger loans which might carry slightly higher rates.

–Comparing Loan Purposes and Their Impact on Repayment Rates.

SELECT

purpose,

SUM(CASE WHEN not\_fully\_paid = 0 THEN 1 ELSE 0 END) \* 100.0 / COUNT(\*) AS percent\_fully\_paid,

SUM(CASE WHEN not\_fully\_paid = 1 THEN 1 ELSE 0 END) \* 100.0 / COUNT(\*) AS percent\_not\_fully\_paid

FROM

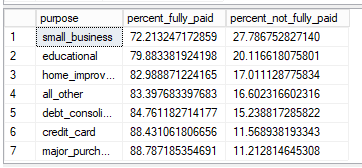
dbo.loandata

GROUP BY

purpose

ORDER BY

percent\_not\_fully\_paid DESC;



These findings are valuable for lenders as they can help in risk assessment and portfolio management. For instance, lenders might consider tightening lending criteria for small business and educational loans or adjusting interest rates to mitigate the higher risks associated with these loan purposes.

— Let's confirm hypothesis #3.

SELECT

purpose,

COUNT(\*) AS total\_loans,

SUM(CASE WHEN not\_fully\_paid = 0 THEN 1 ELSE 0 END) AS fully\_paid,

SUM(CASE WHEN not\_fully\_paid = 1 THEN 1 ELSE 0 END) AS not\_fully\_paid,

SUM(CASE WHEN not\_fully\_paid = 0 THEN 1 ELSE 0 END) \* 100.0 / COUNT(\*) AS percent\_fully\_paid,

SUM(CASE WHEN not\_fully\_paid = 1 THEN 1 ELSE 0 END) \* 100.0 / COUNT(\*) AS percent\_not\_fully\_paid

FROM

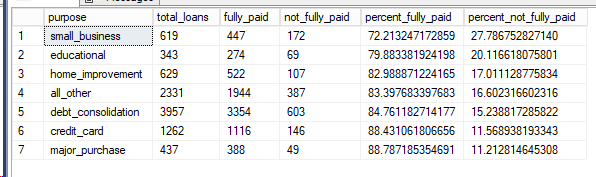
dbo.loandata

GROUP BY

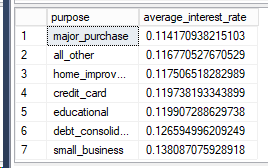
purpose

ORDER BY

percent\_not\_fully\_paid DESC;



The data provided shows repayment rates for different loan purposes. With small business loans having a higher rate of non-repayment and major purchase loans showing a lower rate, there appears to be a pattern in loan repayment based on the purpose. This suggests that loan purpose is indeed a significant factor in the likelihood of a loan being fully repaid, supporting hypothesis #3.



The additional data shows the average interest rates for loans categorized by their purpose. The interest rates vary by loan purpose, with the lowest average rate for major purchases and the highest for small business loans. This variation in interest rates by loan purpose aligns with hypothesis #3, suggesting that loan conditions, such as interest rates, are indeed influenced by the purpose of the loan.

— Let's confirm hypothesis #3.

SELECT

not\_fully\_paid,

AVG(dti) AS average\_dti,

MIN(dti) AS min\_dti,

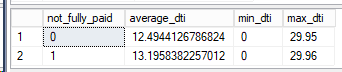
MAX(dti) AS max\_dti

FROM

dbo.loandata

GROUP BY

not\_fully\_paid;



The data shows that loans not fully paid have a slightly higher average DTI ratio compared to those that are fully paid, which could suggest that higher DTI is associated with a higher likelihood of not fully repaying a loan, supporting hypothesis #4. However, the difference in average DTI between the two groups is relatively small, indicating that while there may be a relationship, DTI alone may not be a strong predictor of repayment without considering other factors.