# Data Quality Tests and Summary Statistics

To begin with the data quality assessment, I divided the dataset into company and individual accounts since they needed to be modeled independently. I conducted separate data quality checks for each dataset.

## *Company Accounts*

**Initial Observations:**

- The company accounts dataset contained several variables: Account_ID, Loan_Date, Company/Individual_Flag, Cumulative_Months_on_Book, Remaining_Months_on_Book, Days_in_Arrears, Loan_to_Value, Credit_Score, Account_Management_Score, County_Court_Judgement, Number_of_Searches_in_Last_Month, Public_Information_Sources, Consumer_Indebtedness_Index, and Stage.
- The dataset initially had 1,307 records with missing values in key variables such as Credit_Score (70 missing), Account_Management_Score (144 missing), County_Court_Judgement (144 missing), Number_of_Searches_in_Last_Month (144 missing), Public_Information_Sources (144 missing), and Consumer_Indebtedness_Index (144 missing).

**Handling Staging Issues:**

- I noticed that the dataset included incorrect stages like Stage 10 and Stage 11. Upon examining these cases, I found that the performance metrics (Days_in_Arrears, Loan_to_Value, etc.) did not indicate defaults. Therefore, I assumed these were data input errors and reclassified them into Stage 1.

**Missing Data Analysis and Handling:**

- I conducted a detailed analysis of missing data by stages:
  - **Stage 1:** Most missing values (136) were found in this stage. Given that Stage 1 is considered non-default and generally low-risk, I chose to drop the records with missing values in this stage due to the abundance of data points.
  - **Stage 2:** There were 19 cases with missing data. I could not identify a clear pattern in these missing values, so I opted to drop these records to avoid introducing bias by imputing.

    - **Stage 3:** Out of 18 Stage 3 records, four had missing values. One of these records had extreme anomalies (e.g., Remaining_Months_on_Book being negative), which led me to drop it. For the remaining three cases, I imputed missing Credit_Score values using the mean Credit_Score of Stage 3 records.

**Descriptive Statistics and Data Quality Issues:**

- I performed a descriptive statistics analysis on the cleaned dataset:
    - **Remaining_Months_on_Book**: I found negative values, which indicated data errors or extended loan durations. I set these to a minimum of 2 to reflect a high chance of default.
    - **Loan_to_Value**: A maximum value of 6.8 was recorded, which was unrealistic as this ratio should not exceed 1. I assumed this was a percentage rather than a ratio, so I divided the value by 100.
    - **Credit_Score**: One case had a Credit_Score of -16, which seemed incorrect. I replaced it with the mean Credit_Score of Stage 1 accounts, considering the account's other metrics indicated good performance.
    - **Account_Management_Score**: With 161 cases having a value of -999, I assumed this represented special cases, likely indicating no information available. I left these values as they were, allowing the model to capture this information.
    - **County_Court_Judgement** and **Public_Information_Sources**: Some values were extremely high or negative, which may indicate special cases. I chose not to alter these as their distribution could provide meaningful insights for the model.
    - **Consumer_Indebtedness_Index**: Since this index usually ranges from 0 to 99, I assumed that negative values indicated no information was available. I replaced these negative values with the mean Consumer_Indebtedness_Index for each stage.

**Final Dataset for Company Accounts:**

- After cleaning, the final company accounts dataset consisted of 1,106 records:
    - **Stage 1:** 961 records
    - **Stage 2:** 128 records
    - **Stage 3:** 17 records

## *Individual Accounts*

- **Initial Data Exploration:**

  - The individual accounts dataset had a distinct issue—Credit_Score had missing values in 682 out of 692 records, which is over 90% of the data. Given this high level of missingness, I decided to drop the Credit_Score column entirely.
  - Other variables like Account_Management_Score, County_Court_Judgement, Number_of_Searches_in_Last_Month, Public_Information_Sources, and Consumer_Indebtedness_Index also had missing values, but these were more manageable.

- **Stage-wise Missing Data Analysis:**
  - **Stage 1:** Despite dropping the Credit_Score column, there were still 133 missing values for other variables. Since Stage 1 had a larger dataset (601 cases), I decided to drop the records with missing values to maintain data quality.
  - **Stages 2 and 3:** Similar to the company accounts, I inspected the missing data in these stages. Where missing values seemed to affect less than 10% of the total cases, I chose to drop them instead of imputation. This reduced the data size but maintained integrity.
- **Data Quality Observations:**
  - **Negative Values:** Some individual accounts also had negative values for variables like Remaining_Months_on_Book and Consumer_Indebtedness_Index. I applied similar corrections as with the company accounts, setting negative Remaining_Months_on_Book to 0 and imputing reasonable values for Consumer_Indebtedness_Index.
- **Final Dataset for Individual Accounts:**
  - After cleaning, the dataset had a final structure ready for modeling with only necessary and complete data points retained.
  - **Summary Statistics:**
    - Remaining_Months_on_Book: Adjusted to ensure no negative values.
    - Loan_to_Value: Adjusted to fall between 0 and 1.
    - Other variables showed appropriate ranges following data cleaning.

**2.)** Below is a overview of the modelling process, please look into the pdf of the notebook to get more information about the model.

# Company and Individual Accounts PD Model

## *Company Accounts PD Model*

**Objective:** To predict both the 12-month and lifetime Probability of Default (PD) for company accounts, focusing on Stage 3 accounts as indicating default.

**Process Followed:**

1. **Data Preparation:**
   - **Default Labeling:** Created a new binary column default, marking defaults (default = 1) for accounts in Stage 3 and non-defaults (default = 0) for Stage 1 and Stage 2.
   - **Feature Selection:** Selected key features such as Days_in_Arrears, Account_Management_Score, County_Court_Judgement, Number_of_Searches_in_Last_Month, Public_Information_Sources, Credit_Score, Consumer_Indebtedness_Index, Loan_to_Value, Months_on_Book, and Remaining_Months_on_Book.
2. **Multicollinearity Check (VIF):**
   - Calculated the Variance Inflation Factor (VIF) to check for multicollinearity among the features.
   - **Initial VIF Results:** Found high VIF values for some features, particularly Months_on_Book.
   - **Adjusted VIF:** Removed Months_on_Book and checked again. All remaining features had VIF values less than 10, indicating no significant multicollinearity.
   - **Correlation Analysis:** Used a heatmap to visualize correlations between the features, ensuring that there were no major correlations causing multicollinearity issues.
3. **12-Month PD Model:**
   - **Scaling Features:** Scaled the selected features using standardization.
   - **Logistic Regression:** Built a logistic regression model to predict the 12-month PD.
   - **Model Coefficients:** Examined the model's coefficients to ensure they aligned with expectations regarding the impact of each feature on the default prediction.
   - **12-Month PD Calculation:** Used the fitted model to predict the 12-month PD for each company account, storing the probabilities in a new column PD_12_month.

4. **Lifetime PD Calculation for Stage 2 and 3:**
    - **Identifying Stage 2 and 3 Accounts:** Isolated accounts in Stage 2 and Stage 3 for lifetime PD calculation.
    - **Remaining Lifetime:** Calculated Remaining_Lifetime_Years by dividing Remaining_Months_on_Book by 12.
    - **Cumulative Survival Probability:** Iterated through the remaining lifetime of each loan to calculate cumulative survival probability:
        - Adjusted the survival probability annually using the 12-month PD.
    - **Lifetime PD:** Computed the lifetime PD as *1−survival probability1 - \text{survival probability}*1−survival probability for each account and added it to the PD_lifetime column.
5. **Visualization:**
    - Plotted the distribution of the 12-month PD and lifetime PD to analyze risk patterns across the stages.
    - Used box plots to compare the 12-month PDs across different stages, highlighting how risk varies between Stage 1, Stage 2, and Stage 3.

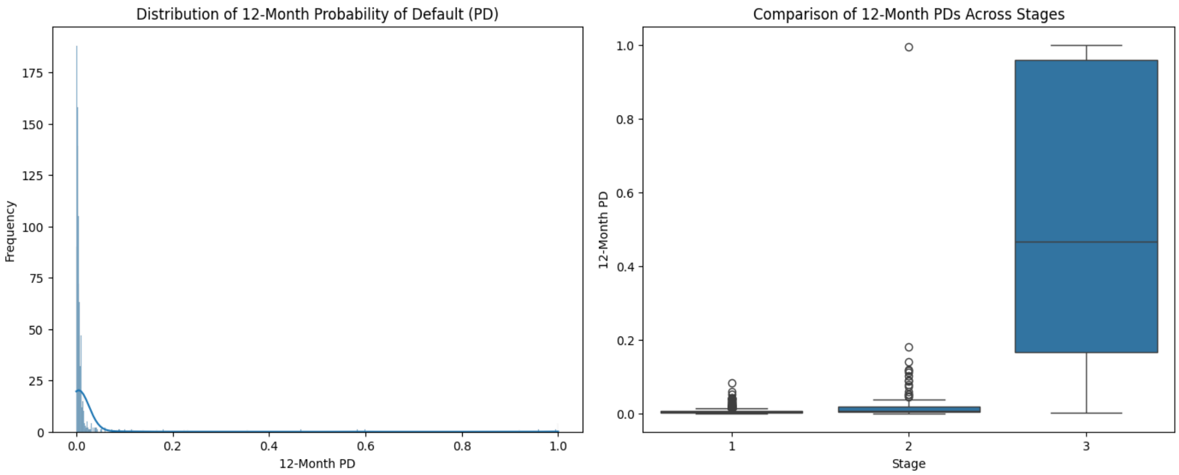
## Coefficients of the 12-month PD model for Companies and some visualizations.:

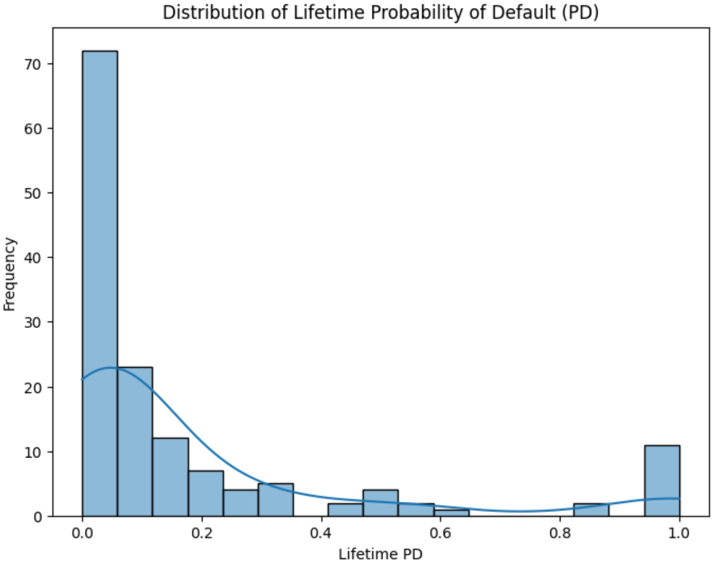
Coefficients of the 12-month:

| | |
|---|---|
| Intercept: | [-5.38993394] |
| Days_in_Arrears: | 1.78774017 |
| Account_Management_Score: | 0.426407697 |
| County_Court_Judgement: | 0.243764812 |
| Number_of_Searches_in_Last_Month | -0.06672998 |
| Public_Information_Sources: | -0.372463426 |
| Credit_Score: | -0.588580628 |
| Consumer_Indebtedness_Index | 0.072511587 |
| Loan_to_Value: | -0.387787676 |
| Remaining_Months_on_Book: | -0.366374885 |

All the coefficients make sense apart from Account_Management_Score haiving a positive impact on PD while it should be negative and Public_Information_Sources being negative, while according to what i interpret of it, it feels like the more public_information about losses or delinquencies you increase in PD. And Loan_to_value should also have a positive impact on the PD but it is negative.

## Result charts





## *Individual Accounts PD Model*

**Objective:** To predict the 12-month and lifetime PD for individual accounts, with the same focus on treating Stage 3 as default.

**Process Followed:**

6. **Data Preparation:**
    - **Default Labeling:** Created a binary default column for individual accounts, setting default = 1 for Stage 3 accounts and default = 0 for others.
    - **Feature Selection:** Selected relevant features including Days_in_Arrears, Account_Management_Score, County_Court_Judgement, Number_of_Searches_in_Last_Month, Public_Information_Sources,

Consumer_Indebtedness_Index, Loan_to_Value, and Remaining_Months_on_Book.

7. **Multicollinearity Check (VIF):**
    - **Initial VIF Check:** Conducted VIF analysis on the selected features.
    - **Feature Adjustment:** Removed Months_on_Book due to higher VIF values.
    - **Final VIF Analysis:** Confirmed that the adjusted feature set had acceptable VIF values, indicating no significant multicollinearity. Loan-to-value was on the borderline but was retained.
8. **12-Month PD Model:**
    - **Scaling Features:** Scaled the features using standardization.
    - **Logistic Regression:** Built a logistic regression model to predict the 12-month PD for individual accounts.
    - **Model Coefficients:** Assessed the coefficients to ensure they were consistent with expectations.
    - **12-Month PD Calculation:** Used the model to calculate the 12-month PD for each individual account, adding the probabilities to the PD_12_month column.
9. **Lifetime PD Calculation for Stage 2 and 3:**
    - **Identifying Stage 2 and 3 Accounts:** Filtered individual accounts in Stage 2 and Stage 3 for the lifetime PD calculations.
    - **Remaining Lifetime:** Computed the Remaining_Lifetime_Years by converting Remaining_Months_on_Book into years.
    - **Cumulative Survival Probability:** For each account, the cumulative survival probability was calculated annually:
        - Survival probability was adjusted iteratively based on the 12-month PD.
    - **Lifetime PD:** Calculated the lifetime PD as *1−survival probability1 - \text{survival probability}*1−survival probability and stored it in the PD_lifetime column.
10. **Visualization:**
    - Visualized the distribution of 12-month PD and lifetime PD for individual accounts.
    - Used box plots to compare the 12-month PDs across different stages, offering insights into the risk differentiation among stages.

The models for both company and individual accounts used logistic regression to predict 12-month PDs and survival analysis to calculate lifetime PDs. Key steps included checking for multicollinearity, scaling features, fitting the model, and then interpreting the coefficients. The analysis and visualization of 12-month and lifetime PDs provided a comprehensive view of the risk associated with different stages, aligning with the IFRS9 framework.

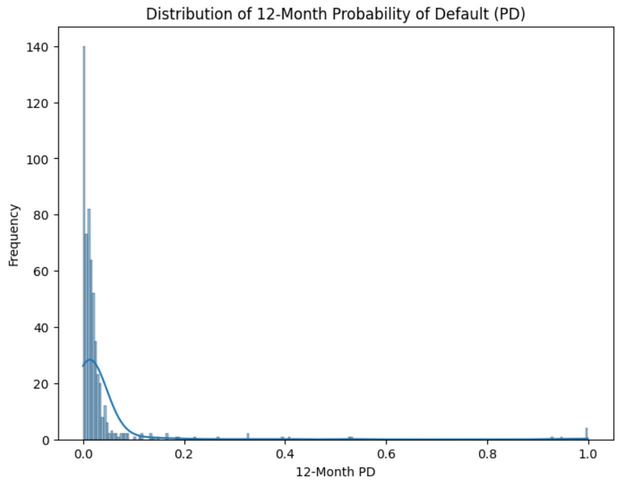
**Coefficients of the 12-month PD model for Companies and some visualizations.:**

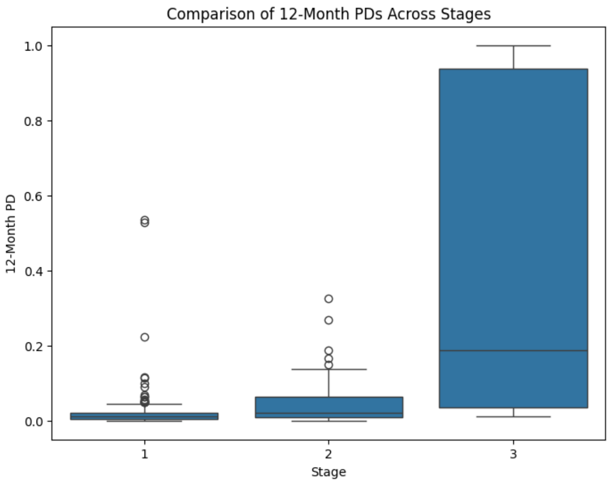
## Cofficeints:

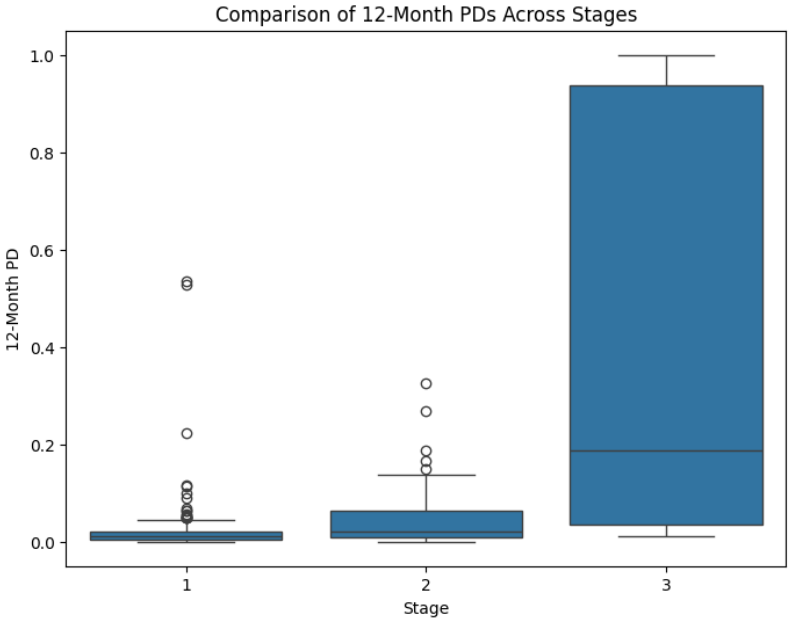
| | |
|---|---|
| Intercept: | [-4.52224379] |
| Days_in_Arrears: | 0.820714569 |
| Account_Management_Score: | -0.067592133 |
| County_Court_Judgement: | 0.394538054 |
| Number_of_Searches_in_Last_Month | -0.486307528 |
| Public_Information_Sources: | -0.243088864 |
| Consumer_Indebtedness_Index | 0.469167275 |
| Loan_to_Value: | -0.274306624 |
| Remaining_Months_on_Book: | -1.180227764 |

All the coefficients make sense apart from Public_Information_Sources being negative, while according to what i interpret of it, it feels like the more public_information about losses or delinquencies you increase in PD.  And I was expecting Loan_to_value to have a positive impact on the PD but it is negative.

## **Results Charts:**







## 3.)

Considering the variables already used in the model (Days_in_Arrears, Account_Management_Score, County_Court_Judgement, Number_of_Searches_in_Last_Month, Public_Information_Sources, Credit_Score, Consumer_Indebtedness_Index, Loan_to_Value, Remaining_Months_on_Book), some additional variables that I think would enhance the model would be:

- **Annual Income:** Including the annual income of individual borrowers could provide insight into their ability to service debt. This would complement existing variables like Credit_Score and Consumer_Indebtedness_Index by adding another dimension of financial capacity.

- **Debt-to-Income Ratio:** This ratio indicates the proportion of a borrower's income used to service debt. A higher ratio suggests greater financial strain and a higher likelihood of default.
- **Installment Amount:** The monthly installment amount for each loan could indicate the borrower's repayment burden. Higher installment amounts relative to income can signal potential repayment difficulties.

- **Credit Score:** Credit Score data would have been very usefull in case of individual accounts as well. ( We had it in the company account level )

- **Total Number of Open Accounts:** The total number of open credit accounts (including loans, credit cards, etc.) can indicate the borrower's overall credit exposure. Higher exposure could signal a greater risk of overextension.

- **Number of Recent Delinquencies:** The count of recent delinquencies (e.g., late payments) in the borrower's credit history can directly indicate higher default risk.

- **Company Age:** The number of years since the company was established can help assess its stability. Newly established companies might pose a higher default risk compared to more mature companies. (this might also cause bias in the model)

- **Industry Sector:** Including the industry sector of the company can help identify sector-specific risks. Some industries are inherently riskier and more prone to economic downturns than others. (this might also cause bias in the model)

Also including macro-economic variables like the ones listed below can increase the models predictive power and allow us to make better PD predictions in a forward looking time span.

- **GDP Growth Rate:** Reflects the overall economic health. A slowing GDP growth rate could signal economic downturns that might affect borrowers' repayment capabilities.

- **Unemployment Rate:** Higher unemployment rates can increase the likelihood of defaults, especially for individual borrowers. This variable can indicate the economic stress that affects borrowers' income stability.

- **Interest Rates:** Higher interest rates increase borrowing costs, potentially affecting the borrowers’ ability to repay. Including variables like the central bank's interest rate or average market interest rates can provide insight into the financial burden on borrowers.

- **Inflation Rate:** Rising inflation can erode purchasing power, potentially leading to higher default rates as borrowers' real income decreases.