Phase 3 Project: CRISP-DM

Predicting loan defaults for

Asset loan applications of vehicles with

Logistic Regression, Decision Trees and Random Forest

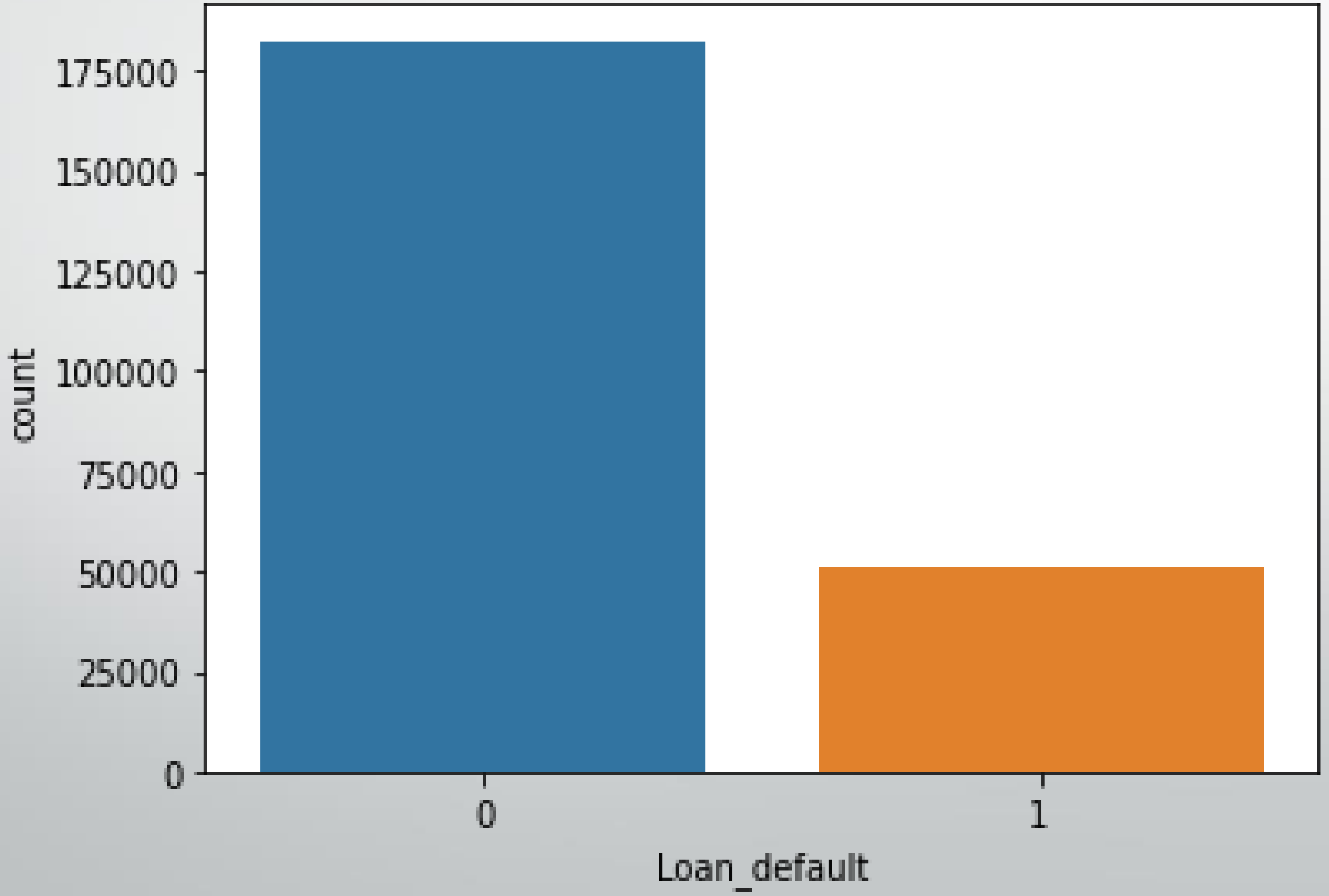
Business Understanding:

- An asset based credit lending company, that offers vehicle loans has contracted us, to predict the rate of loan default, based on their existing customer history. They need to create a better credit risk scoring model, to minimize the default rate to less than 10% of their portfolio.
- The dataset is of loans given for motor vehicles. We have been tasked with building an asset loan default prediction, for motor vehicles based on the given dataset. We will be predicting the rate of default using the Logistic Regression, Decision Trees and Random Forest predictive models. Our target variable, rate of loan default, is based on KYC(Know Your Customer) data collected on the loan borrowers. This data includes the borrower's personal information, employment status and credit status.
- The 5C's are often used when accessing credit worthiness, which are Capital, Capacity, Collateral, Character and Conditions.

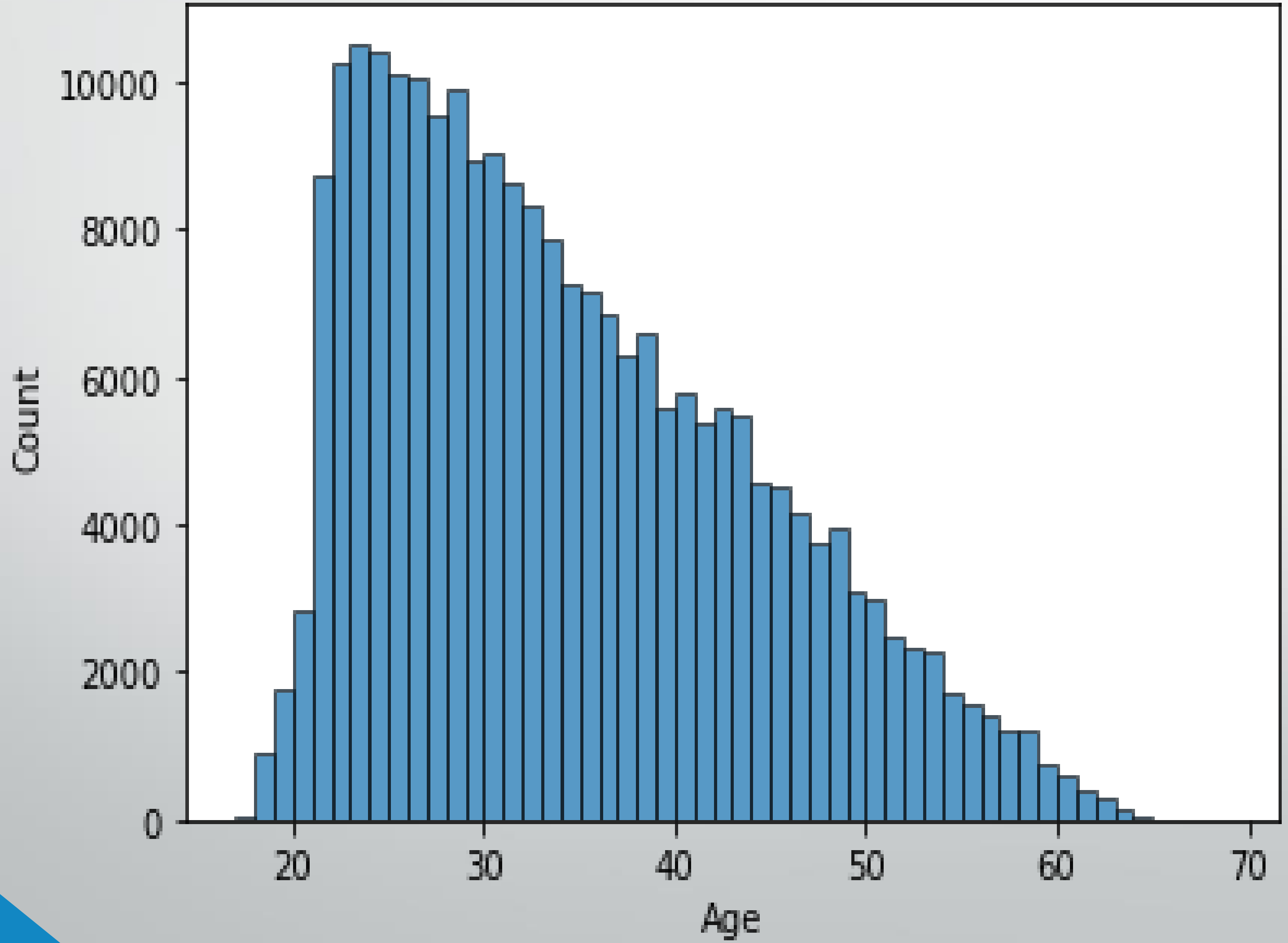
Data Understanding:

- *The company will provide us with their historic data, so we may make accurate predictive models*
  - *Our target variable is Loan default*
  - *Our predictor variables will be chosen from the remaining 39 columns, and other predictor columns that we will create*
  - *The most significant columns based on market research and domain information, center on the SC's of Credit .In this project, we have identified them as :*
    - *Disbursed Amount*
    - *Disbursal date*
    - *Asset Cost*
    - *LTV*
    - *State ID*
    - *Employment type*
    - *Credit history length*
    - *Age*
    - *Perform_cns_score*
    - *Manufacturer_id*

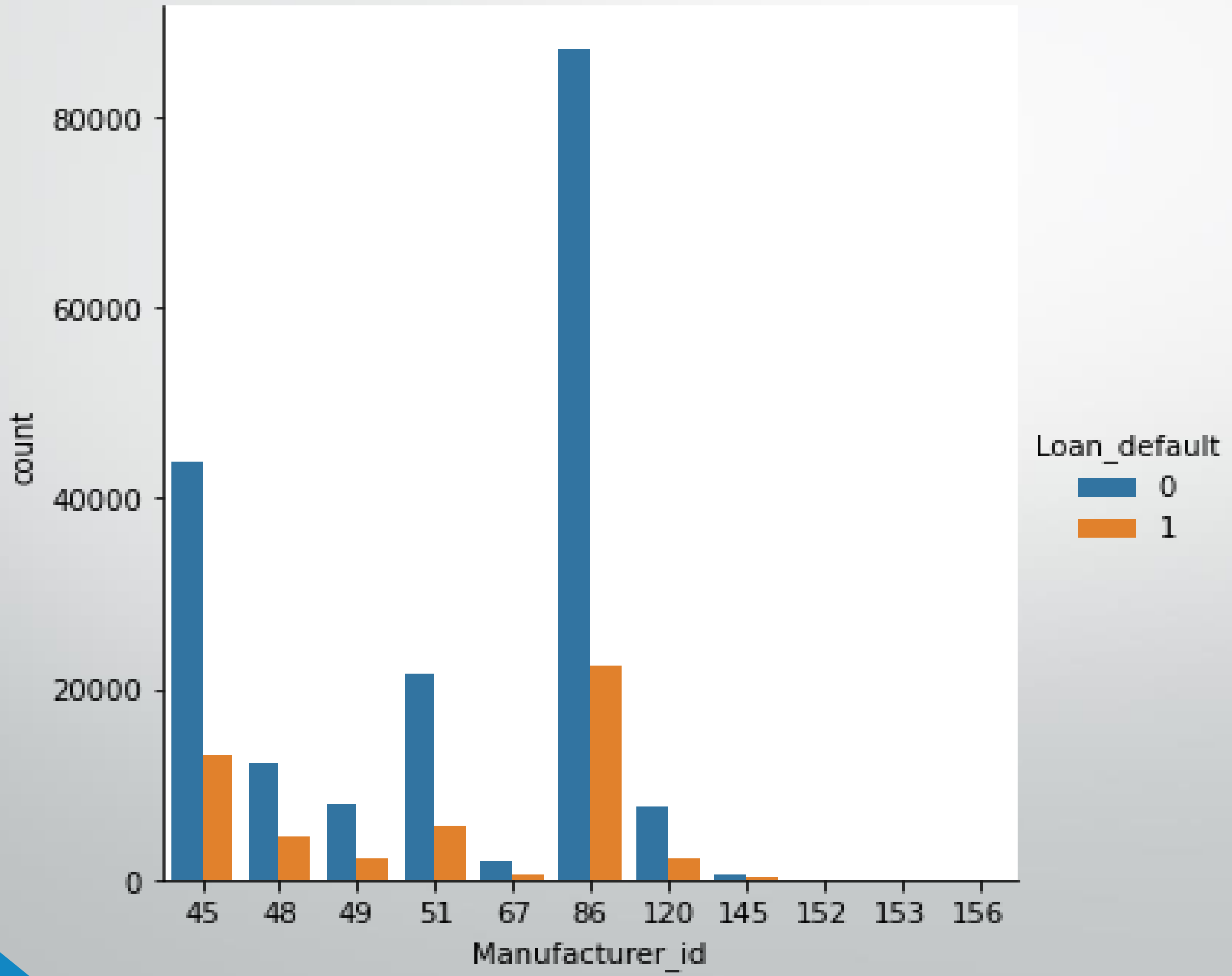
- According to the historical data, the current loan default rate is 21%. Our goal is to create a predictive model that will decrease this rate to 10%.
- No default = 182,543
- Defaulted = 50,611



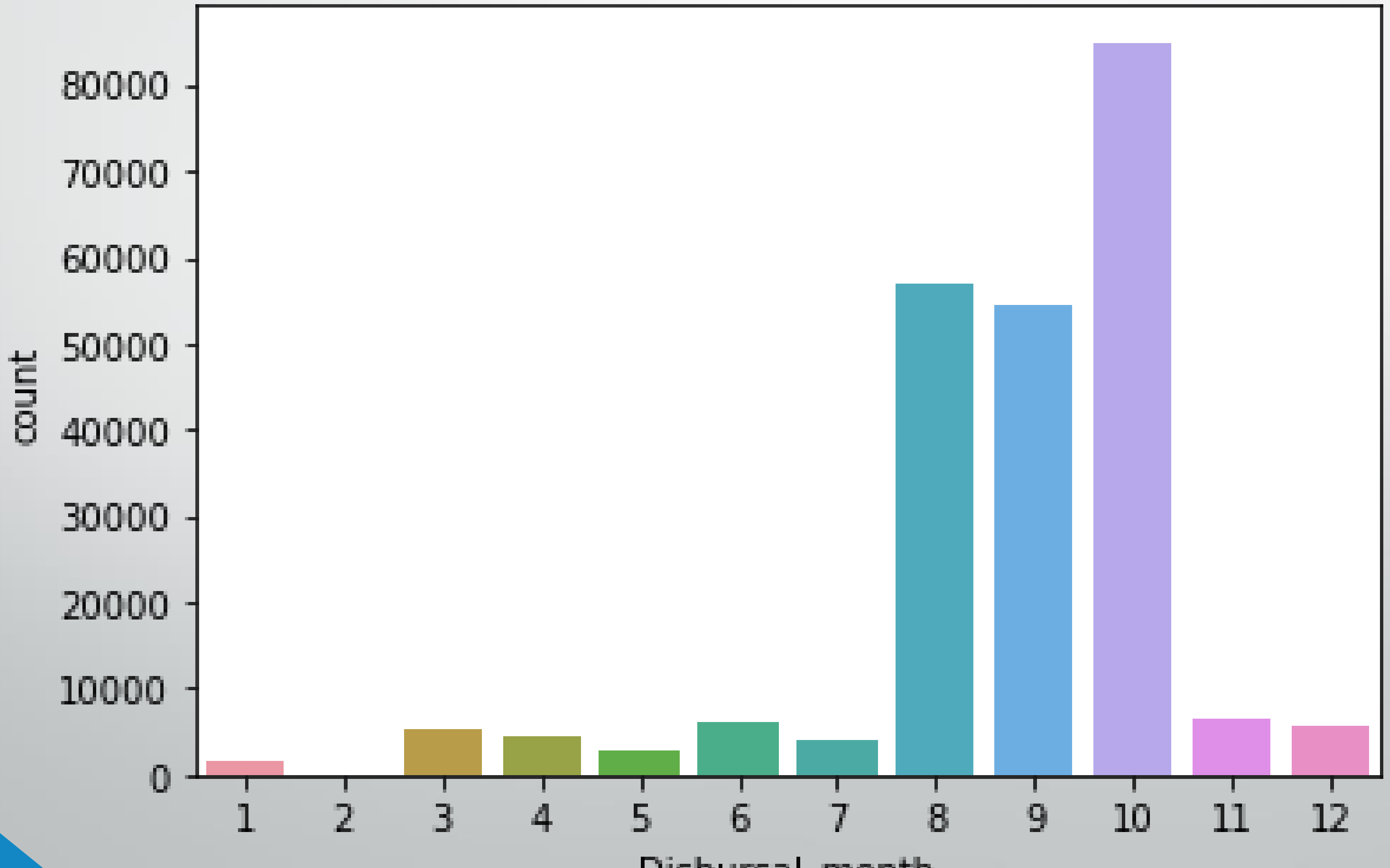
- A visualization showing the age distribution of the borrowers.
- The median age Is 32 years.



- Visualization showing the default rate based on the vehicle manufacturer.
- Manufacturer 86, had the highest default rate, asset cost value, and vehicle count



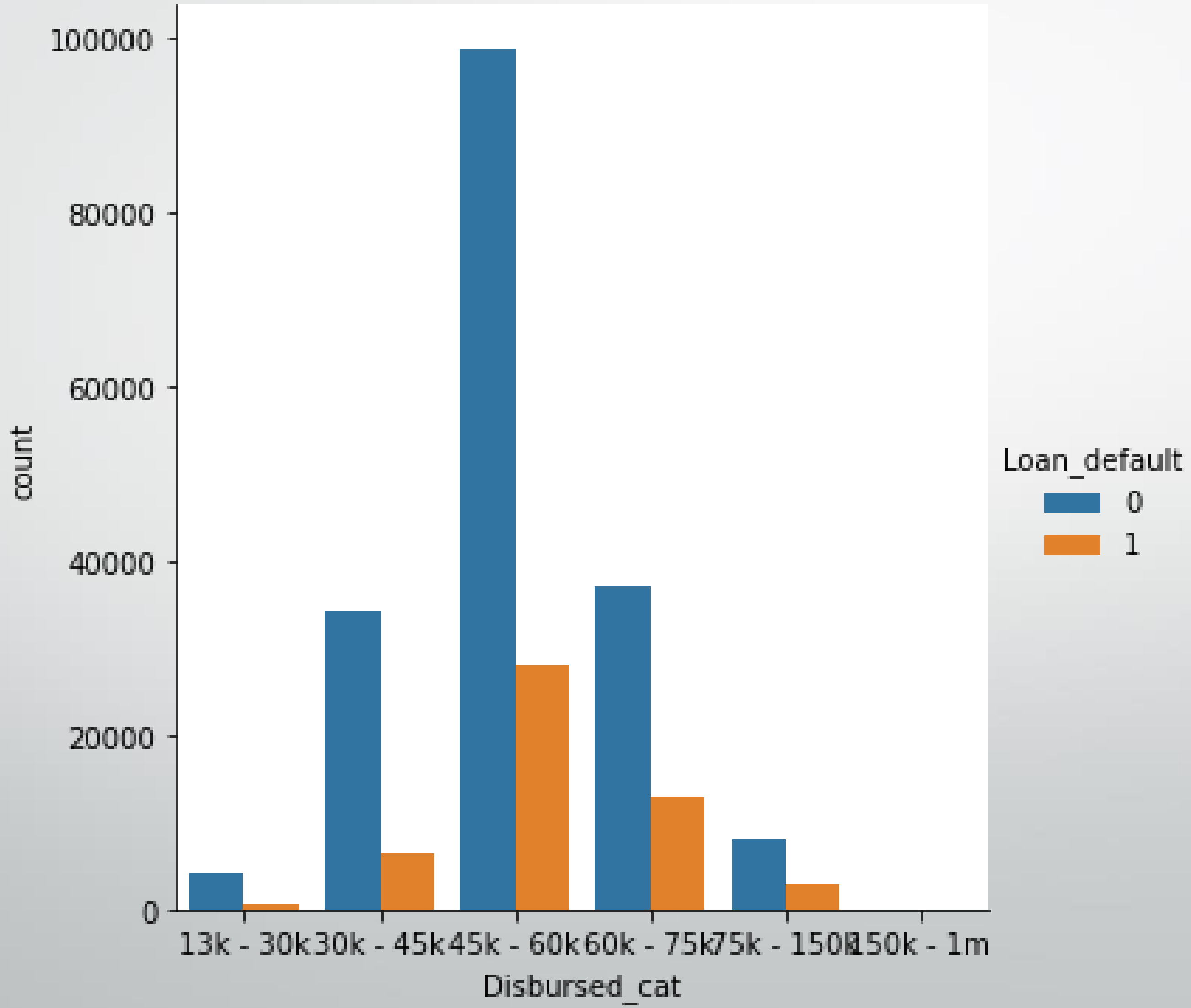
Visualization showing Disbursal Amount per month



Data Preparation

- *We imputed missing values in employment type column, which was the only column with null values.*
- *We scaled the data using the MinMax Scaler, due to the high variance in values of the columns.*
- *We binned the Disbursed Amount, to create a categorical column for easier evaluation*
- *We one hot encoded the categorical data, in preparation for modelling.*

- The highest loan defaults are in the 45-60K bin
- Loan default rate increased as disbursed amount increased as well, they are positively correlated.



Modeling

- *This is a classification problem, where we are trying to predict if a customer will default on a loan or not.*
- *We will use:*
  - *Logistic Regression*
  - *Decision Tree Classifier*
  - *An ensemble model, Random Forest*

Model Evaluation

- The evaluation metrics used in this project are:
  - Accuracy
  - Precision
  - Recall
  - F1 score
  - AUC

Logistic Regression Metrics

- Accuracy: 0.7826926101190902
- Precision: 0.2
- Recall: 0.00019748535316963992
- F1: 0.0003945810864132579
- AUC: 0.6195024611468474

- The model has not done a bad job at predicting the loan default based on the accuracy assumption

- f1 score of ~0.0003 should prove beyond doubt that our model is not reliable despite the 78% accuracy

Decision Tree Metrics

We performed 3 iterative models, while tuning hyper parameters, and got the following results:

Evaluation Metric	Model 1	Model 2	Model 3
Accuracy:	67%	67%	78%
Precision:	26%	25%	0%
Recall	28%	26%	0%
F1	27%	26%	0%
AUC	53%	52%	55%

Random Forest Metrics

- Accuracy: 0.6345993008942549
- Precision: 0.3028237585199611
- Recall: 0.5206815048256844
- F1: 0.3829355738239235
- AUC: 0.6383335366609525

Conclusion

The best model seems to be the logistic regression model, based on the higher accuracy, and the AUC.
The model will require more feature engineering to make better predictive analysis, of the target variable.