## Null values

No presence of Null values are reported in data

## Outliers

Degree of Financial Leverage (DFL) 22.041355
Interest Coverage Ratio (Interest expense to EBIT) 20.838833
Fixed Assets Turnover Frequency 20.794838
Current Asset Turnover Rate 20.516205
Total Asset Growth Rate 20.252236
...
Quick Asset Turnover Rate 0.000000
Cash Turnover Rate 0.000000
Operating Expense Rate 0.000000
Net Income Flag 0.000000
Inventory Turnover Rate (times) 0.000000
Length: 96, dtype: float64

the maximum outlier percentage is 22%

and the outliers is not a bad thing. They can explain the variance in our data which normally distributed values fail to explain.

In our case which is bankruptcy declaration, the presence of outliers is essential for considering different anomalies that can happen and different scenarios which influence bankruptcy
So right now we wont be dropping outliers we will create our model using them

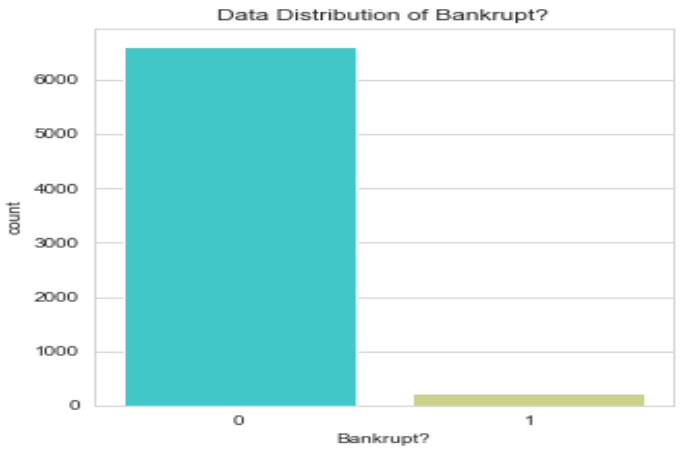
For example the column - Degree of Financial Levergae- which denotes how much a company relies on debt to finance its operations have maximum percentage of outliers. But it is not adding any ghost value to our model as for different companies depending on their revenue , he DFL may vary. DFL inturn depends upon other variable "
X1 Cost of Interest-bearing Debt:
The cost of debt directly impacts interest expenses, which are a component of the DFL calculation.

X6 Total Liability/Equity Ratio:
This ratio reflects the extent to which a company relies on debt (liabilities) versus equity. Higher leverage (more debt relative to equity) increases financial risk and affects DFL.

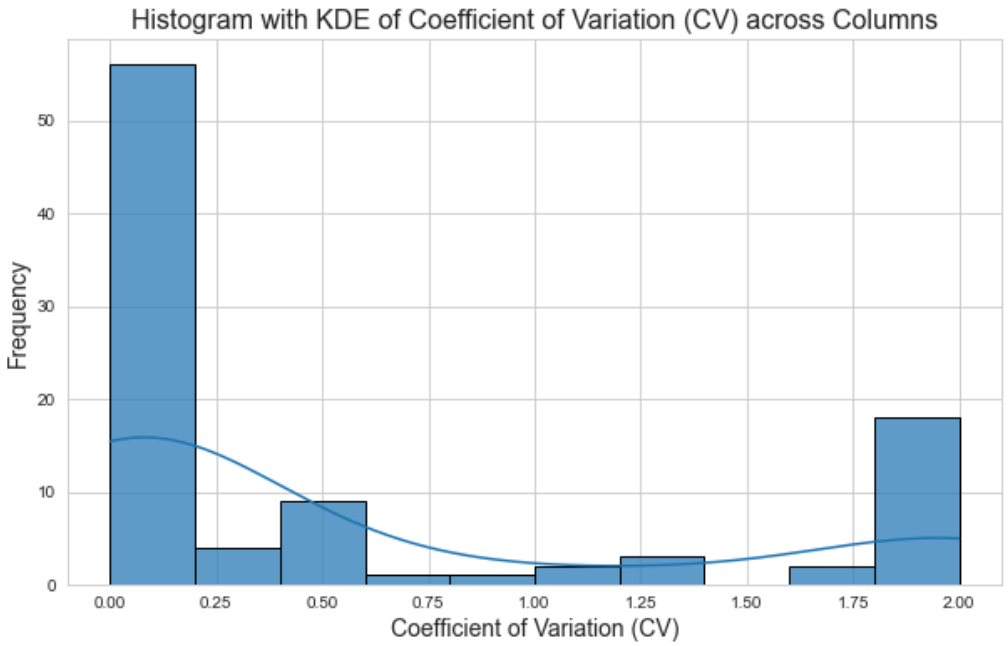
Hence, according to domain it is practical to not remove outliers.

## Balance of Data:

The data is imbalance
0 6599 - Not Bankrupt
1 220 - Bankrupt



**Histogram of Co-effiecient of  Variations of different Columns**




In short, a skewness value of 1.057 for the Coefficient of Variation (CV) histogram with KDE means that the distribution of
CV values across columns is moderately positively skewed. This suggests that there are more columns with higher coefficients of variation, indicating greater variability or dispersion in those datasets compared to columns with lower coefficients of variation.

**Columns with High CV**

| | Coefficient of Variation (CV) |
|---|---|
| Current Ratio | 1.999707 |
| Fixed Assets to Assets | 1.999706 |
| Net Value Growth Rate | 1.999413 |
| Revenue per person | 1.999413 |
| Quick Assets/Current Liability | 1.99912 |
| Revenue Per Share (Yuan Â¥) | 1.998533 |
| Liability-Assets Flag | 1.997654 |
| Total debt/Total net worth | 1.997654 |
| Quick Ratio | 1.99736 |
| Allocation rate per person | 1.99648 |

**Columns with Low CV**

| A | B | C | D |
|---|---|---|---|
| | Coefficient of Variation (CV) | | |
| Non-industry income and expenditure | 0.00173 | | |
| Continuous Net Profit Growth Rate | 0.001522 | | |
| Pre-tax net Interest Rate | 0.000871 | | |
| After-tax net Interest Rate | 0.000869 | | |
| Continuous interest rate (after tax) | 0.000827 | | |
| Operating Profit Growth Rate | 0.000613 | | |
| Operating Profit Rate | 0.00055 | | |
| Working capitcal Turnover Rate | 0.000479 | | |
| Cash Flow to Sales | 0.000387 | | |
| Net Income Flag | 0 | | |

**Presence of Heavily Skewed Columns and Negatively Skewed Columns**

'Heavily Positively Skewed Indicates that these columns have large number of value which are greater than the mean.

Possible Explanation:

1.The "Fixed Assets to Assets" ratio typically measures the proportion of a company's fixed assets (e.g., buildings, equipment) relative to its total assets.

Potential Reason for Skewness: Companies with large fixed asset bases relative to their total assets will have higher values for this ratio.

In financial datasets, especially in industries with significant capital investments (like manufacturing), there may be a few companies with disproportionately high fixed asset values compared to others, leading to a skewed distribution where the mean is pulled towards higher values.

Heavily Negatively Skewed Indicates that these columns have large number of value which are lesser than the mean.

Operating Profit Growth Rate:

Definition: This ratio measures the percentage change in operating profit over a specific period.

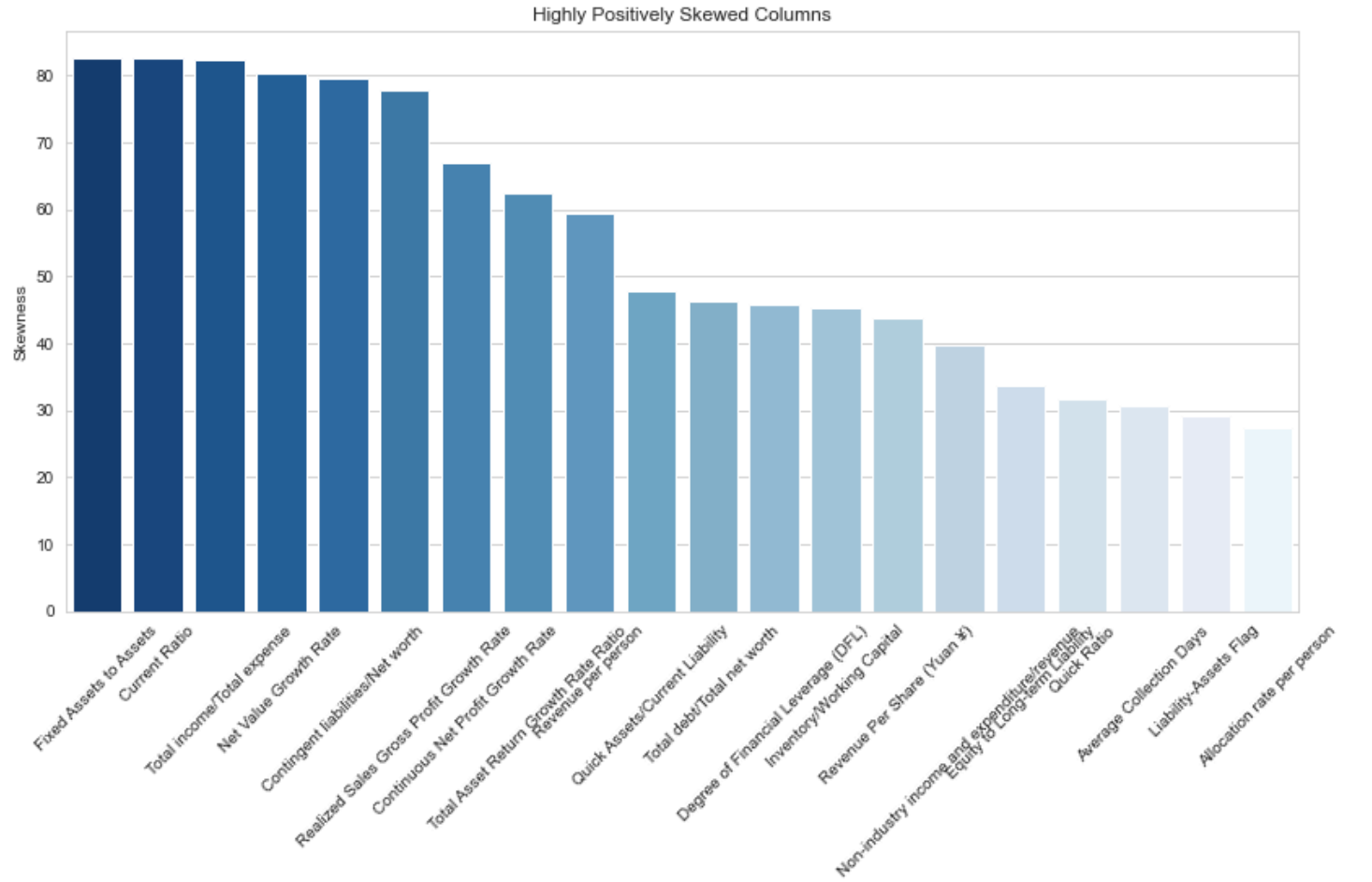
Potential Reason for Negative Skewness: Companies experiencing declines or negative growth in operating profits

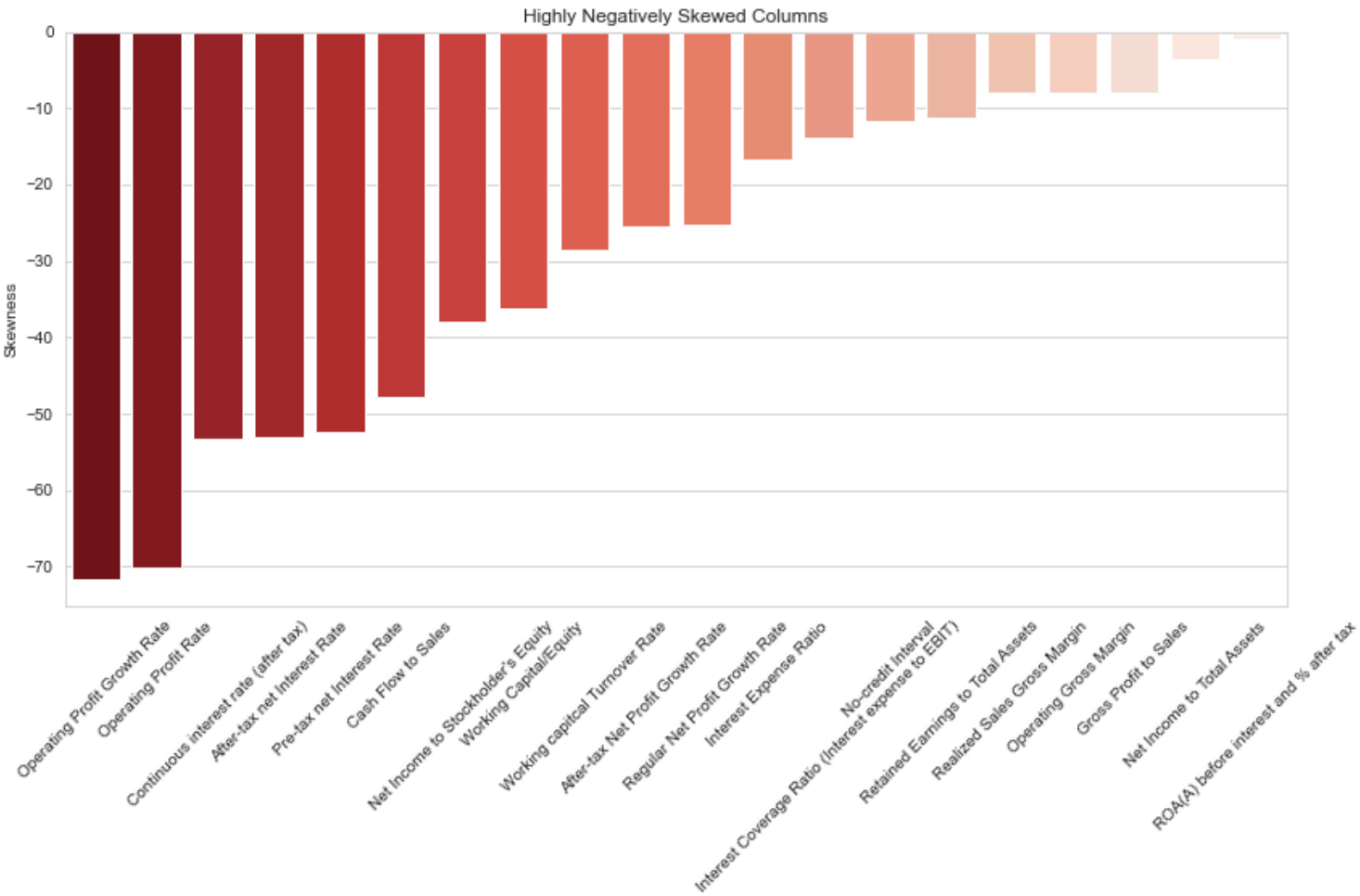
will have negative values for this ratio.

Economic downturns, operational challenges, or sector-specific issues can lead to a higher frequency of negative growth rates, resulting in a skewed distribution where more values are below the mean.

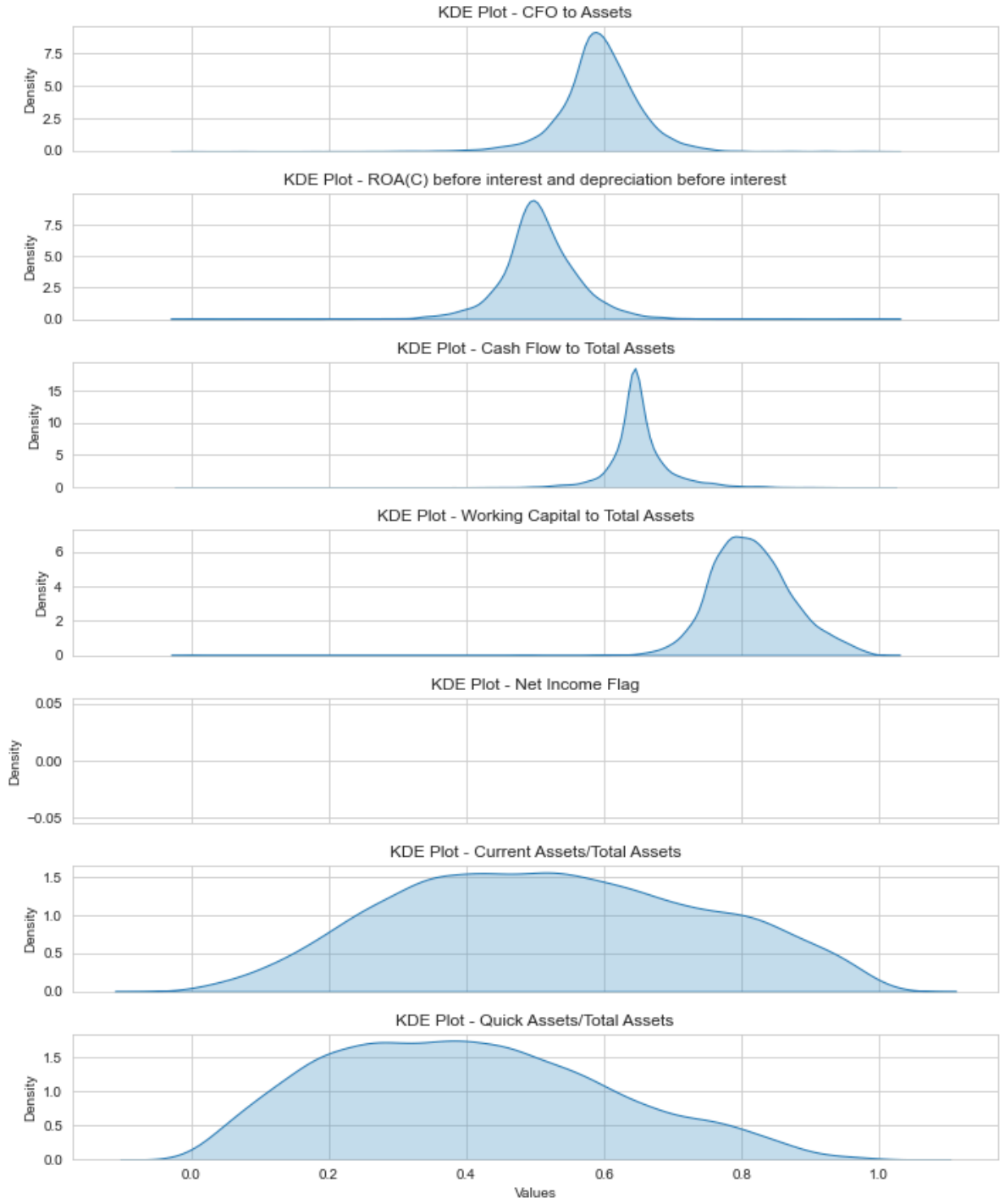
2. Operating Profit Rate:

Industries or companies with lower operating profit margins relative to revenue or assets will have lower values for this ratio. Sectors with high competition, cost pressures, or economic downturns may exhibit lower profitability ratio.


Highly Positively Skewed Columns
Skewness
0
10
20
30
40
50
60
70
80
Fixed Assets to Assets
Current Ratio
Total income/Total expense
Net Value Growth Rate
Contingent liabilities/Net worth
Realized Sales Gross Profit Growth Rate
Continuous Net Profit Growth Rate
Total Asset Return Growth Rate Ratio
Revenue per person
Quick Assets/Current Liability
Total debt/Total net worth
Degree of Financial Leverage (DFL)
Inventory/Working Capital
Revenue Per Share (Yuan ¥)
Non-industry income and expenditure/revenue
Equity to Long-term Liability
Quick Ratio
Average Collection Days
Liability-Assets Flag
Allocation rate per person


Highly Negatively Skewed Columns
Skewness
0
−10
−20
−30
−40
−50
−60
−70
Operating Profit Growth Rate
Operating Profit Rate
Continuous interest rate (after tax)
After-tax net Interest Rate
Pre-tax net Interest Rate
Cash Flow to Sales
Net Income to Stockholder's Equity
Working Capital/Equity
Working capital Turnover Rate
After-tax Net Profit Growth Rate
Regular Net Profit Growth Rate
Interest Expense Ratio
Interest Coverage Ratio (Interest expense to EBIT)
No-credit Interval
Retained Earnings to Total Assets
Realized Sales Gross Margin
Operating Gross Margin
Gross Profit to Sales
Net Income to Total Assets
ROA(A) before interest and % after tax

## Columns with almost Uniform  Distribution




## Top 10 columns showing very less p-values means highly co-related with Bankruptcy

| A | B | C |
|---|---|---|
| Column | P-value | |
| Research and development expense | 0.045399 | |
| Quick Ratio | 0.038528 | |
| Quick Asset Turnover Rate | 0.033042 | |
| Total assets to GNP price | 0.003742 | |
| Regular Net Profit Growth Rate | 0.002358 | |
| After-tax Net Profit Growth Rate | 0.001805 | |
| Revenue per person | 0.001036 | |
| Cash Flow to Liability | 0.000368 | |
| Total Asset Growth Rate | 0.000242 | |
| Current Assets/Total Assets | 0.000213 | |

## Hypothesis Testing

Null Hypothesis - Bankruptcy does not depend on research and development expense rate
p value<0..05 which is 0.048 therefore Ho is rejected
Hence we can conclude, Bankruptcy depends on research and development expense rate

Bankrupt and R&D Expense have negative co-efficient means when
R&D expense increases then bankruptcy will approch 0 means no bankrupt
As the R&D Expense increases, the probability of response variable decreases as there is -ve corelation

Conclusion: It is obvious a profitable company will spend more expenses on R and D

Null Hypothesis- Bankruptcy does not depend on Current Liability to Assets
p value<0.05 which is 0.000 therefore Ho is rejected
Hence we can conclude, Bankruptcy depends on Current Liability to Assets

As the value of Net Income to total assets increases by one unit, the log of
likelihood of bankruptcy=1 increases by 17.0380
Conclusion: It is obvious a profitable company will not have more liabilities than assets

Null Hypothesis - Bankruptcy does not depend on Net Income to Total Assets
p value<0.05 which is 0.000 therefore Ho is rejected
Hence we can conclude, Bankruptcy depends on Net Income to Total Assets

As the value of Net Income to total assets increases by one unit, the log of
likelihood of bankruptcy=1 decreases by -20

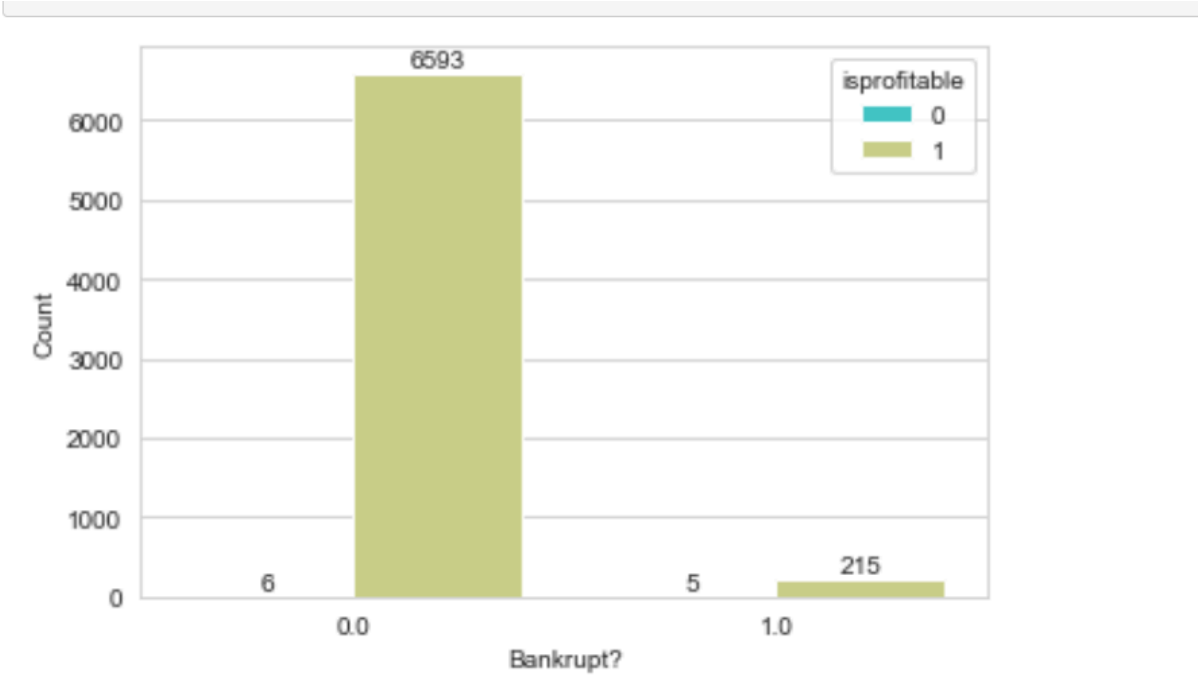
## Feature Engineering

Profit Indicators- value increases , profitability increases and bankruptcy decreases
Eg: Cash Flow Per Share,Total Asset Turnover
Loss Indicators- value increases , profitability decreases and bankruptcy increases
Eg:Debt ratio %,Degree of Financial Leverage (DFL)

```
# Calculate scores for profit and loss indicators
df['profit_score'] = df[Profit_Indicators].mean(axis=1)
df['loss_score'] = df[Loss_Indicators].mean(axis=1)

# Define a threshold or rule
threshold = 0.07  # Adjust this threshold as per your criteria

# Create isprofitable column based on the rule
df['isprofitable'] = (df['profit_score'] > df['loss_score'] + threshold).astype(int)
```




Null Hypothesis: IsProfitable does not influences Bankruptcy
Alternate Hypothesis: IsProfitable influences  Bankruptcy

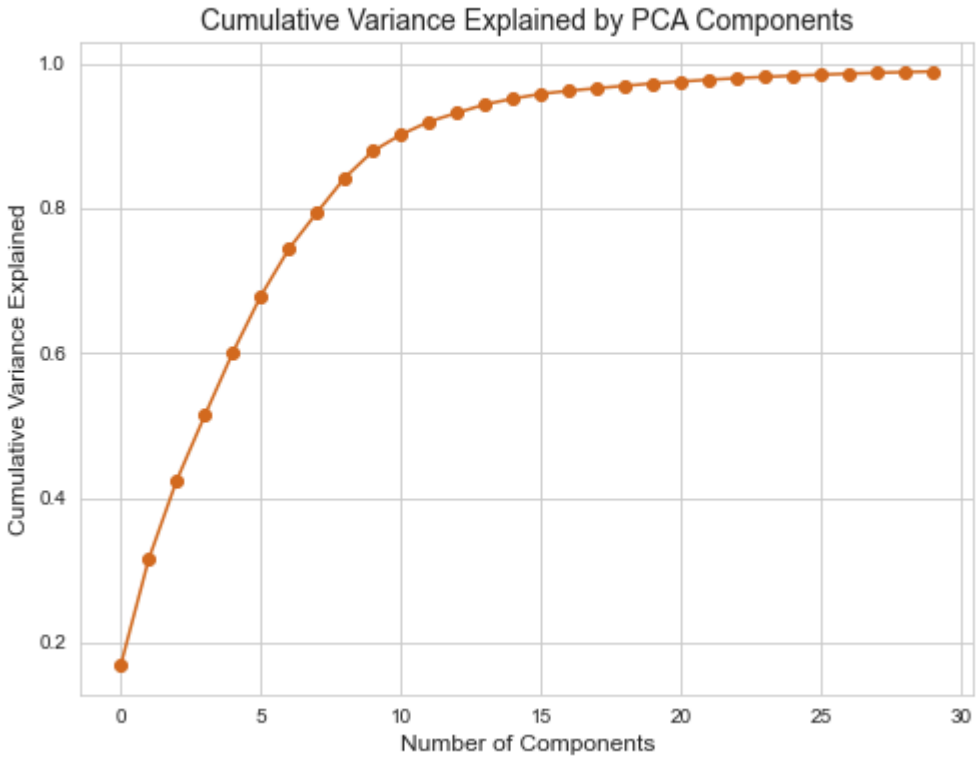
```
Bankrupt?       0.0  1.0
isprofitable
0                 6    5
1              6593  215
Chi-square statistic: 50.10960664365914
P-value: 1.4539344623108553e-12
```

As p-value =0.000000000001454 which is <0.05 we reject the Null Hypothesis
Hence concluded that isprofitable  influences Bankruptcy

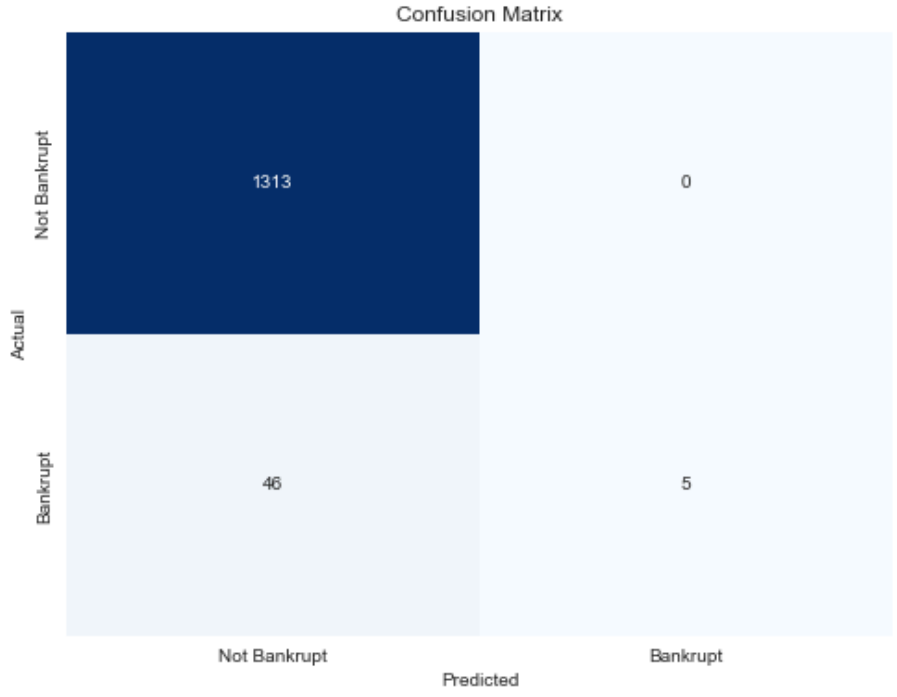
**Feature Selection Using PCA**

variance_percentage

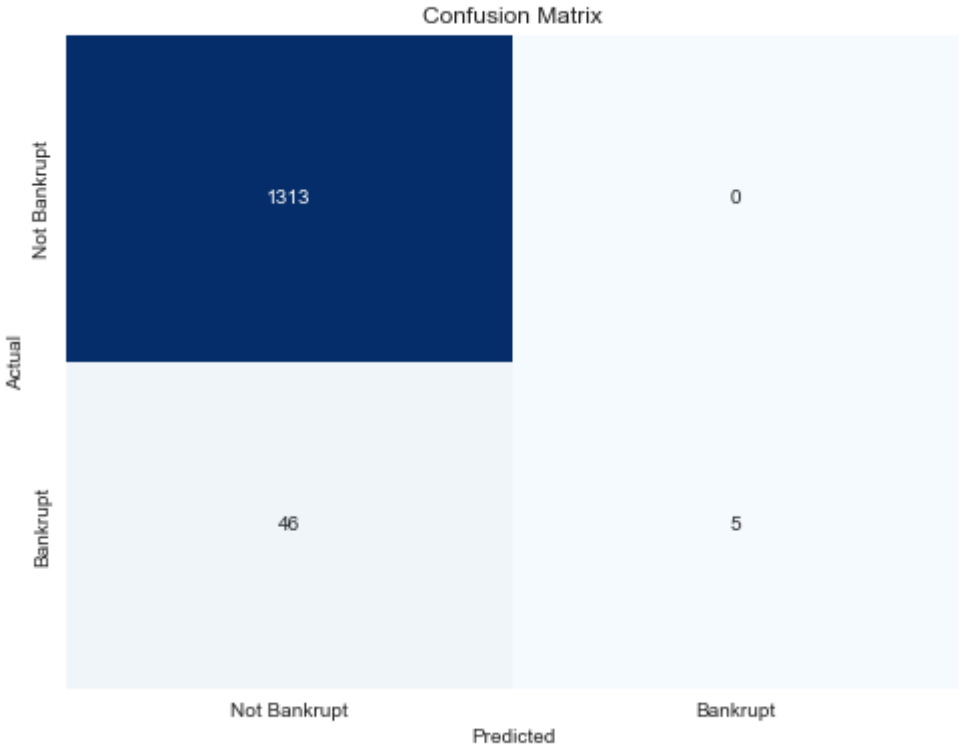
[16.87 14.66 10.9  8.95 8.68 7.82 6.49 5.01 4.83 3.67 2.27 1.79
 1.22 1.16 0.8  0.66 0.43 0.35 0.34 0.33 0.27 0.24 0.22 0.18
 0.16 0.15 0.14 0.12 0.1  0.09]
Total Asset Turnover is explaining maximum variance




Models:
1]Logistic Regression



2]Random Forest Classifier




In the context of financial risk management, the decision between prioritizing precision or recall depends on the specific consequences and goals related to managing risk.We should prioritize **recall** as the cost or impact of false negatives is high. This applies to scenarios where:

**Missing a Risk is Very Costly**: Failing to identify a true risk can lead to significant financial losses.

**Regulatory Compliance**: Some financial regulations may require that certain risks are always flagged and addressed, even at the cost of more false positives.

**Risk Aversion**: Organizations may prefer to catch as many potential risks as possible, even if it means more false positives, to avoid missing any significant threat.