BUSINESS REPORT ON FINANCE AND RISK ANALYTIC

- Rahul Sharma

Problem 1:

In the real of modern finance, businesses encounter the perpetual challenge of managing debt obligations effectively to maintain a favorable credit standing and foster sustainable growth. Investors keenly scrutinize companies capable of navigating financial complexities while ensuring stability and profitability. A pivotal instrument in this evaluation process is the balance sheet, which provides a comprehensive overview of a company's assets, liabilities, and shareholder equity, offering insights into its financial health and operational efficiency. In this context, leveraging available financial data, particularly from preceding fiscal periods, becomes imperative for informed decision-making and strategic planning.

A. Define the problem and perform Exploratory Data Analysis:

Problem definition - Check shape, Data types, and statistical summary - Univariate analysis - Multivariate analysis - Use appropriate visualizations to identify the patterns and insights - Key meaningful observations on individual variables and the relationship between variables

A.1 Problem definition- Check shape, Data types, and statistical summary:-

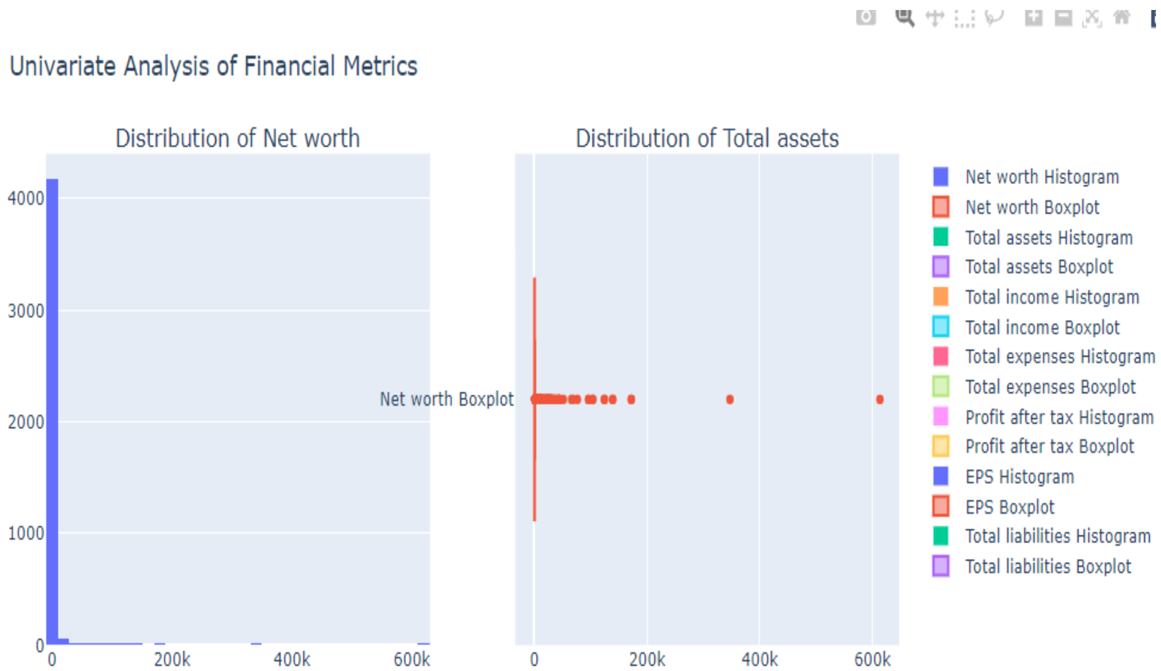
Shape of the data-frame = 4256*51

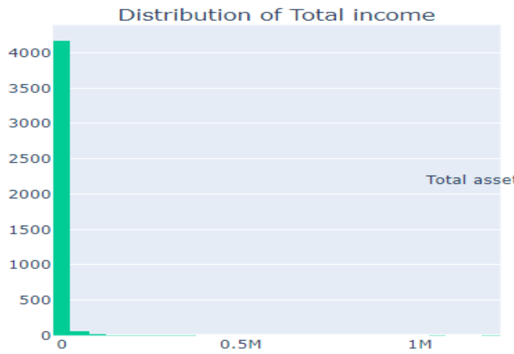
```
Shape of the dataframe: (4256, 51)

Data types of each column:
 Num                                          int64
Networth Next Year                          float64
Total assets                                float64
Net worth                                   float64
Total income                                float64
Change in stock                             float64
Total expenses                              float64
Profit after tax                            float64
PBDITA                                      float64
PBT                                         float64
Cash profit                                 float64
PBDITA as % of total income                 float64
PBT as % of total income                    float64
PAT as % of total income                    float64
Cash profit as % of total income            float64
PAT as % of net worth                       float64
Sales                                       float64
Income from fincial services                float64
Other income                                float64
Total capital                               float64
Reserves and funds                          float64
Borrowings                                  float64
Current liabilities & provisions            float64
Deferred tax liability                      float64
Shareholders funds                          float64
Cumulative retained profits                 float64
Capital employed                            float64
TOL/TNW                                     float64
Total term liabilities / tangible net worth float64
```

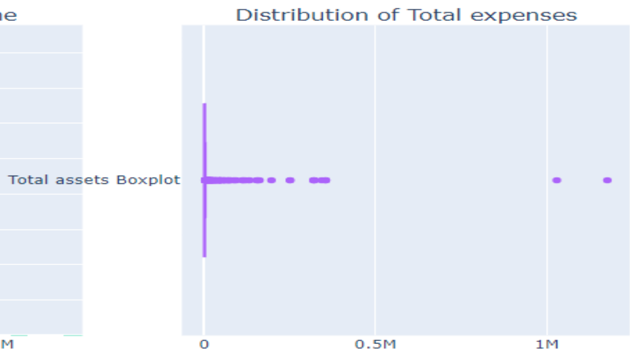
```
Contingent liabilities / Net worth (%)         float64
Contingent liabilities                         float64
Net fixed assets                               float64
Investments                                    float64
Current assets                                 float64
Net working capital                            float64
Quick ratio (times)                            float64
Current ratio (times)                          float64
Debt to equity ratio (times)                   float64
Cash to current liabilities (times)            float64
Cash to average cost of sales per day          float64
Creditors turnover                             float64
Debtors turnover                               float64
Finished goods turnover                        float64
WIP turnover                                   float64
Raw material turnover                          float64
Shares outstanding                             float64
Equity face value                              float64
EPS                                            float64
Adjusted EPS                                   float64
Total liabilities                              float64
PE on BSE                                      float64
dtype: object

Statistical summary:
               Num  Networth Next Year  Total assets      Net worth  \
count  4256.000000         4256.000000  4.256000e+03    4256.000000
mean   2128.500000         1344.740883  3.573617e+03    1351.949601
std    1228.745702        15936.743168  3.007444e+04   12961.311651
min       1.000000       -74265.600000  1.000000e-01       0.000000
25%    1064.750000            3.975000  9.130000e+01      31.475000
50%    2128.500000           72.100000  3.155000e+02     104.800000
75%    3192.250000          330.825000  1.120800e+03     389.850000
max    4256.000000       805773.400000  1.176509e+06  613151.600000

       Total income  Change in stock  Total expenses  Profit after tax  \
count  4.025000e+03      3706.000000    4.091000e+03       4102.000000
mean   4.688190e+03        43.702482    4.356301e+03        295.050585
std    5.391895e+04       436.915048    5.139809e+04       3079.902071
min    0.000000e+00     -3029.400000   -1.000000e-01      -3908.300000
25%    1.071000e+02        -1.800000    9.680000e+01          0.500000
50%    4.551000e+02         1.600000    4.268000e+02          9.000000
75%    1.485000e+03        18.400000    1.395700e+03         53.300000
max    2.442828e+06     14185.500000    2.366035e+06     119439.100000

              PBDITA            PBT  ...  Debtors turnover  \
count    4102.000000    4102.000000  ...       3871.000000
mean      605.940639     410.259044  ...         17.929029
std      5646.230633    4217.415307  ...         90.164435
min      -440.700000   -3894.800000  ...          0.000000
25%         6.925000       0.800000  ...          3.810000
50%        36.900000      12.600000  ...          6.470000
75%       158.700000      74.175000  ...         11.850000
max    208576.500000  145292.600000  ...       3135.200000

       Finished goods turnover  WIP turnover  Raw material turnover  \
count              3382.000000   3492.000000            3828.000000
mean                 84.369988     28.684513              17.733926
std                 562.637359    169.650915             343.125864
min                  -0.090000     -0.180000              -2.000000
25%                   8.190000      5.100000               3.020000
50%                  17.320000      9.860000               6.410000
75%                  40.012500     20.240000              11.822500
max               17947.600000   5651.400000           21092.000000
```

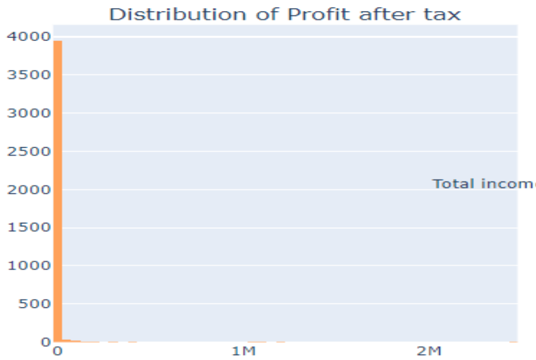
```
       Shares outstanding  Equity face value            EPS   Adjusted EPS  \
count        3.446000e+03        3446.000000    4256.000000    4256.000000
mean         2.376491e+07       -1094.828671    -196.217467    -197.527608
std          1.709790e+08       34101.358644   13061.953425   13061.929512
min         -2.147484e+09     -999998.900000 -843181.820000 -843181.820000
25%          1.308382e+06          10.000000       0.000000       0.000000
50%          4.750000e+06          10.000000       1.490000       1.240000
75%          1.090602e+07          10.000000      10.000000       7.615000
max          4.130401e+09      100000.000000   34522.530000   34522.530000

       Total liabilities     PE on BSE
count       4.256000e+03   1629.000000
mean        3.573617e+03     55.462290
std         3.007444e+04   1304.445296
min         1.000000e-01  -1116.640000
25%         9.130000e+01      2.970000
50%         3.155000e+02      8.690000
75%         1.120800e+03     17.000000
max         1.176509e+06  51002.740000

[8 rows x 51 columns]
```

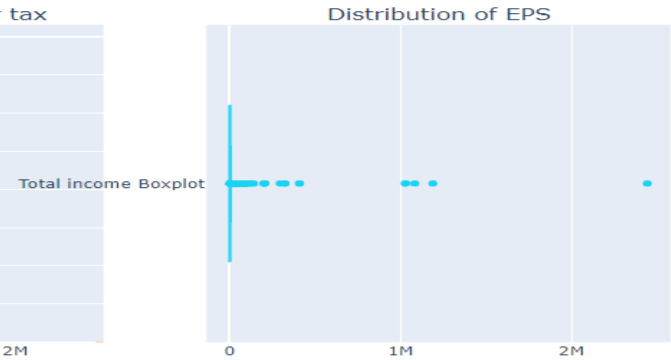
A.2 Univariate analysis:- (appropriate visualizations to identify the patterns and insight):-

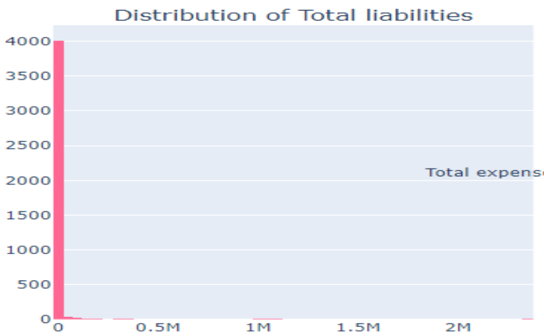


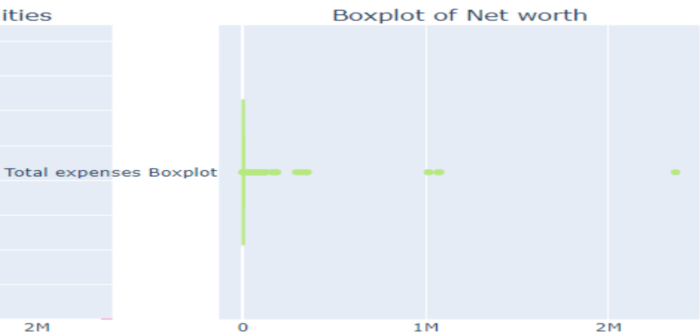

Distribution of Total income
4000
3500
3000
2500
2000
1500
1000
500
0
0
0.5M
1M

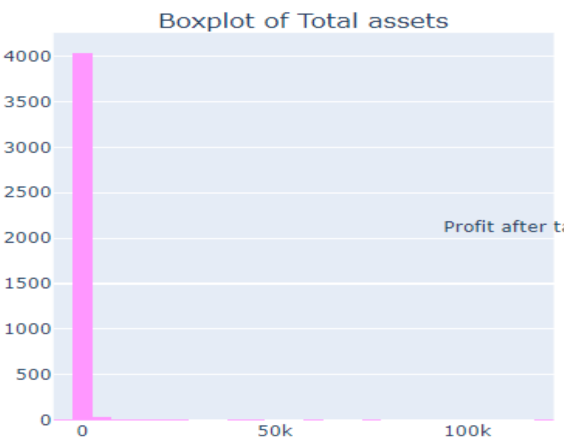

Distribution of Total expenses
Total assets Boxplot
0
0.5M
1M

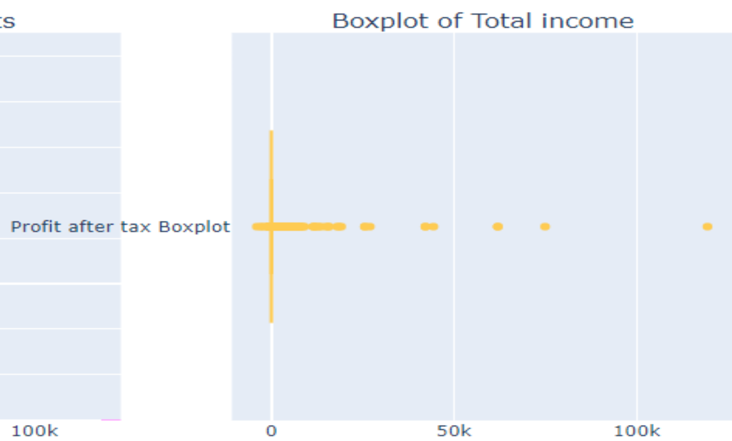

Distribution of Profit after tax
4000
3500
3000
2500
2000
1500
1000
500
0
0
1M
2M


Distribution of EPS
Total income Boxplot
0
1M
2M


Distribution of Total liabilities
4000
3500
3000
2500
2000
1500
1000
500
0
0
0.5M
1M
1.5M
2M


Boxplot of Net worth
Total expenses Boxplot
0
1M
2M


Boxplot of Total assets
4000
3500
3000
2500
2000
1500
1000
500
0
0
50k
100k


Boxplot of Total income
Profit after tax Boxplot
0
50k
100k

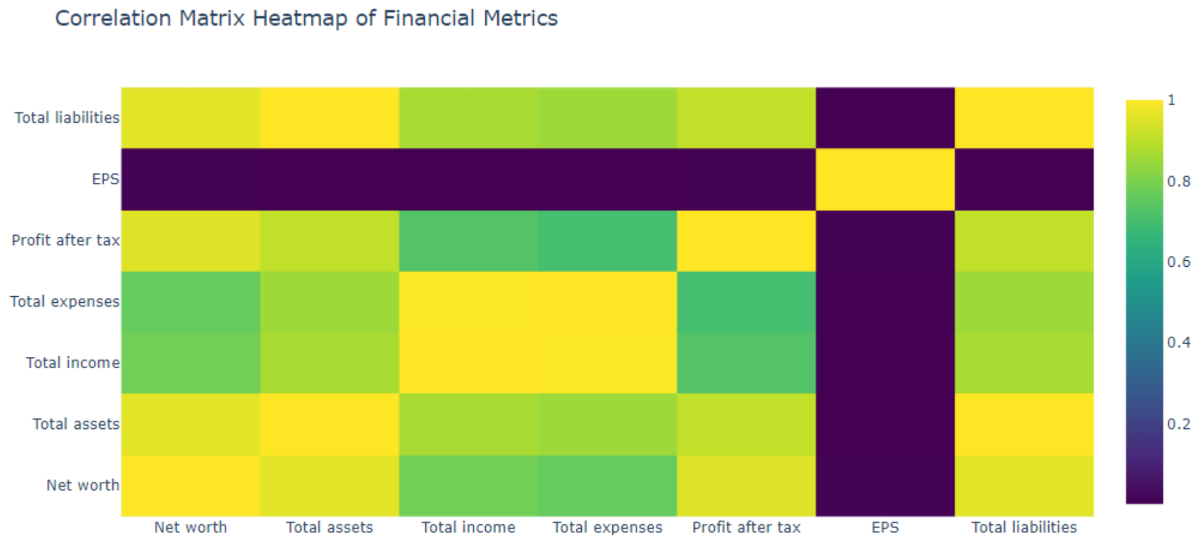
Here are some specific insights that you can glean from the heatmap:

- **Dark red:** High values for the metric
- **Light red:** Lower values for the metric
- **Dark blue:** Low values for the metric
- **Light blue:** Higher values for the metric

- **Profit after tax:** The distribution is skewed to the left, meaning that there are more companies with lower profits than companies with higher profits.
- **Earnings per share (EPS):** The distribution is also skewed to the left, similar to profit after tax.
- **Total income:** The distribution of total income is more evenly spread out, but there is still a slight skew to the left.
- **Total expenses:** The distribution of total expenses is similar to the distribution of total income.
- **Net worth:** The distribution of net worth is skewed to the right, meaning that there are more companies with lower net worth than companies with higher net worth.
- **Total liabilities:** The distribution of total liabilities is also skewed to the right, similar to net worth.

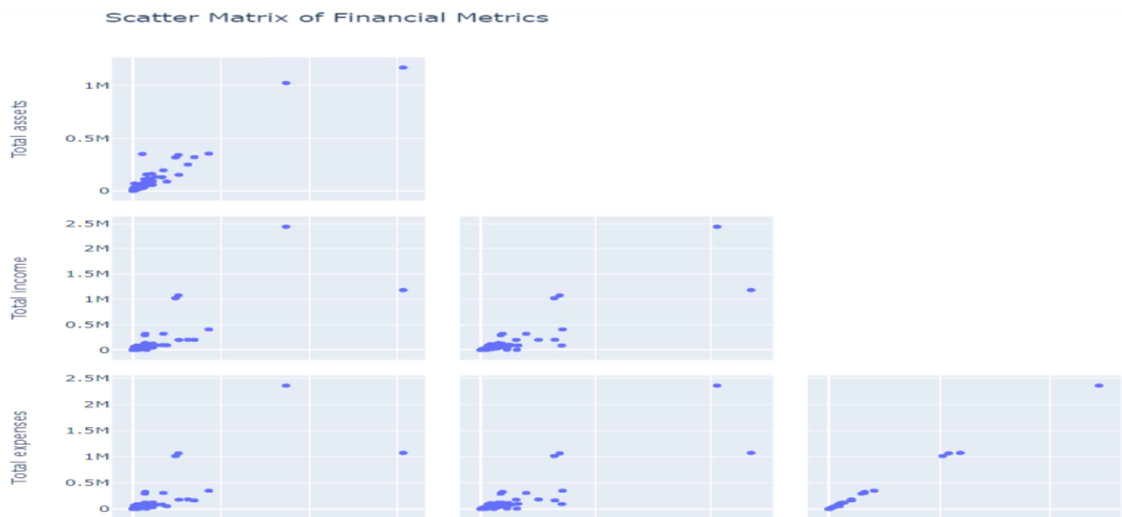
Please go through the ipynb file for all columns uni-variate visualization

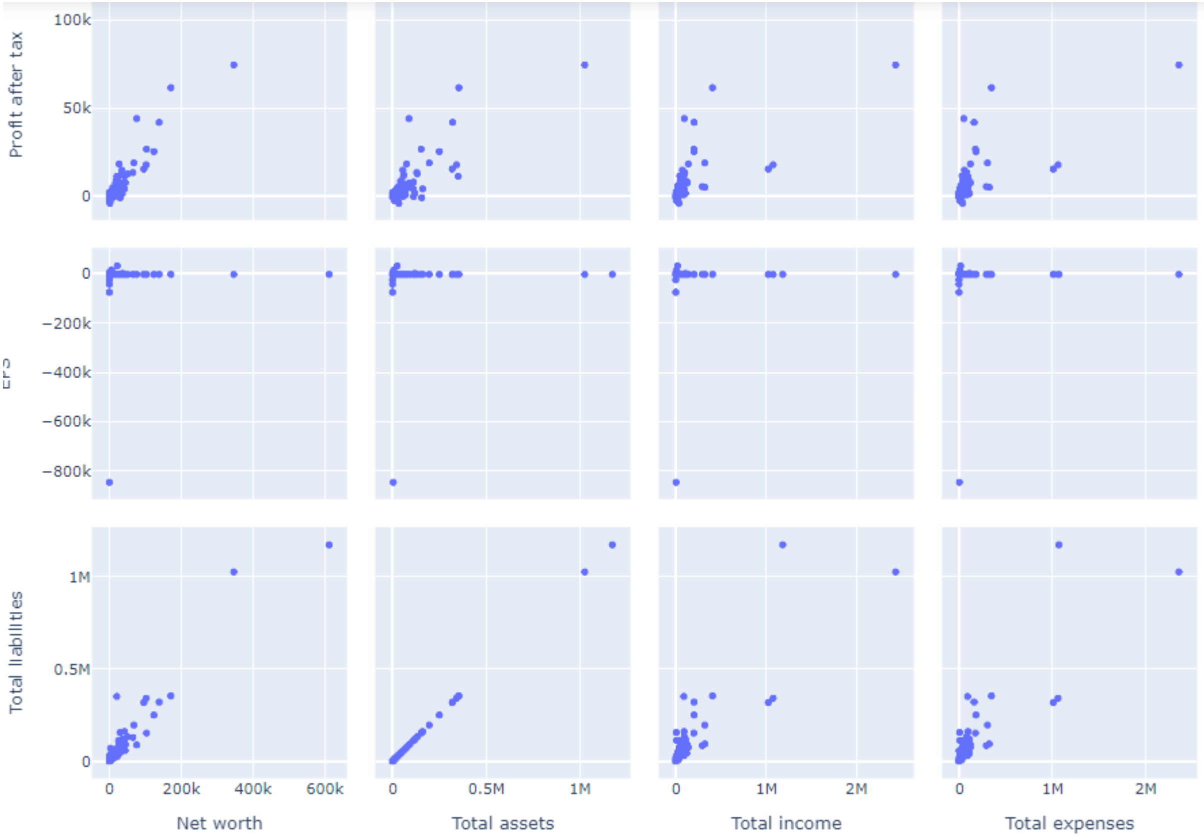
A.3 Multivariate analysis:- (appropriate visualizations to identify the patterns and insight):-



Here are some of the insights that can be gleaned from the heatmap:

- Net worth has a strong positive correlation with total assets (0.8) and a moderate positive correlation with total income (0.6). This suggests that companies with higher total assets and total income tend to have higher net worth.
- Total expenses has a weak positive correlation with total income (0.4) and a weak positive correlation with total assets (0.2). This suggests that companies with higher total income and total assets tend to have higher total expenses, but the correlation is not very strong.
- Profit after tax has a weak positive correlation with total income (0.4) and a very weak positive correlation with total assets (0.2). This suggests that companies with higher total income tend to have higher profit after tax, but the correlation is not very strong.
- EPS has a weak positive correlation with total income (0.4) and a very weak positive correlation with total assets (0.2). This suggests that companies with higher total income tend to have higher EPS, but the correlation is not very strong.
- There is a weak negative correlation between net worth and total liabilities (-0.2). This suggests that companies with higher total liabilities tend to have lower net worth.





Here are some insights that can be gleaned from the scatter plot:

- There is a positive correlation between total income and total expenses. This means that as total income increases, total expenses also tend to increase. This is likely because companies with higher revenue can afford to spend more on operational costs.
- The data points are spread out across the plot, which suggests that there is a wide range of possible total expenses for a given level of total income. In other words, some companies with high total income have high total expenses, while others with high total income have lower total expenses.
- There may be outliers in the data. Outliers are data points that fall far away from the majority of the other data points. It is difficult to say for sure from this image whether there are any outliers, but if there are, they could be worth investigating further.

There is some missing visualization please go through the ipynb file.

A4. Key meaningful observations on individual variables and the relationship between variables:-

<u>Individual Variables:</u>

- **Profitability (Profit after tax & EPS):** Both distributions are skewed left, indicating more companies with lower profits. This could suggest an industry with many low-profit companies or a company struggling to be profitable.

- **Income & Expenses:** The distributions of total income and total expenses are more symmetrical, suggesting a wider range of companies across the spectrum. However, a slight skew to the left might imply companies tend to have more expenses than income.
- **Financial Strength (Net Worth & Liabilities):** Both net worth and total liabilities are skewed right, meaning more companies have lower values. This could indicate an industry or company with high debt relative to net worth, potentially impacting financial stability.

Relationships Between Variables:

Limited Visibility: The heat-map doesn't directly show correlations, but here are some potential relationships to consider for further analysis:

- **Profitability vs. Income/Expenses:** A positive correlation between profit and income (or a negative correlation with expenses) would be expected for a healthy company.
- **Net Worth vs. Profitability/Income:** A positive correlation between net worth and profitability or income suggests the company is retaining profits and building wealth.
- **Net Worth vs. Liabilities:** A negative correlation would be expected, as higher liabilities generally reduce net worth.

While the heat-map provides initial insights, a more comprehensive analysis would involve:

- **Correlation Matrix:** This would show the strength and direction of relationships between variables, confirming or refuting potential connections from the heat-map.
- **Financial Ratios:** Calculating ratios like profit margin, debt-to-equity, and return on equity can provide deeper insights into profitability, solvency, and efficiency.
- **Industry Comparison:** Bench marking the company's metrics against industry averages can reveal strengths and weaknesses.

B. Data Pre-processing:

Prepare the data for modeling: - Outlier Detection (treat, if needed) - Encode the data - Data split - Scale the data - Target variable creation * The target variable is default and should take the value 1 when net worth next year is negative & 0 when net worth next year is positive.

B1. Prepare the data for modeling: - Outlier Detection (treat, if needed):-

```
Outliers detected in column Num: 0
```

B2. Encode the data - Data split - Scale the data:-

ENCODED THE DATA

```
Encoded DataFrame (first 5 rows):
   Num  Networth Next Year  Total assets  Net worth  Total income  \
0    1               395.3         827.6      336.5         534.1
1    2                36.2          67.7       24.3         137.9
2    3                84.0         238.4       78.9         331.2
3    4              2041.4        6883.5     1443.3        8448.5
4    5                41.8          90.9       47.0         388.6

   Change in stock  Total expenses  Profit after tax  PBDITA   PBT  ...  \
0             13.5           508.7              38.9   124.4  64.6  ...
1             -3.7           131.0               3.2     5.5   1.0  ...
2            -18.1           309.2               3.9    25.8  10.5  ...
3            212.2          8482.4             178.3   418.4 185.1  ...
4              3.4           392.7              -0.7     7.2  -0.6  ...

   Debtors turnover  Finished goods turnover  WIP turnover  \
0              5.65                     3.99          3.37
1               NaN                      NaN           NaN
2              2.51                    17.67          8.76
3              1.91                    18.14         18.62
4             68.00                    45.87         28.67

   Raw material turnover  Shares outstanding  Equity face value    EPS  \
0                  14.87           8760056.0               10.0   4.44
1                    NaN                 NaN                NaN   0.00
2                   8.35                 NaN                NaN   0.00
3                  11.11          10000000.0               10.0  17.60
4                  19.93            107315.0              100.0  -6.52

   Adjusted EPS  Total liabilities  PE on BSE
0          4.44              827.6        NaN
1          0.00               67.7        NaN
2          0.00              238.4        NaN
3         17.60             6883.5        NaN
4         -6.52               90.9        NaN
```

DATA SPLIT-

```
[5 rows x 51 columns]
Training features shape: (3404, 50)
Testing features shape: (852, 50)
Training target shape: (3404,)
Testing target shape: (852,)
Scaled Training Data (first 5 rows):
[[-0.83973935 -0.04717562 -0.07720546 -0.0601899  -0.04425417  0.06140514
  -0.04079697 -0.06238839 -0.06881459 -0.06214365  0.04693635  0.056638
   0.0590783   0.05243478  0.16390167 -0.04430708 -0.07190275 -0.04317528
  -0.10224689 -0.05404381 -0.07686729 -0.09814458         nan -0.06174175
  -0.04318098 -0.06568137 -0.15911328 -0.10596076 -0.02269153 -0.0547284
  -0.05942192 -0.10333833 -0.06717224  0.02306117 -0.01586759 -0.04974042
  -0.1294482  -0.06003898 -0.07522215  0.29243    -0.1383676  -0.1031372
  -0.01314622         nan -0.13315882  0.03510509  0.03516467  0.03525487
  -0.07720546         nan]
 [ 1.05119749 -0.08290017 -0.10172905 -0.08970292 -0.07273781         nan
  -0.070355   -0.09467193 -0.08868535 -0.08792807  0.02875362  0.05503942
   0.05803831  0.04784698  0.14351093 -0.07340059 -0.07163975 -0.04567765
   0.20612393 -0.12358058 -0.12982922 -0.08547049         nan -0.09115018
  -0.15564524 -0.10612911 -0.148153   -0.12423364 -0.14504964 -0.07714484
  -0.08497501 -0.10691443 -0.10752618 -0.03311828 -0.05260151 -0.07093982
  -0.18036096 -0.04752508 -0.06168632 -0.16840081 -0.13957407         nan
          nan -0.20277064         nan         nan  0.01690032  0.01699049
  -0.10172905         nan]
 [ 0.37908296 -0.08110374 -0.11074697 -0.09806158 -0.08319846 -0.11560638
  -0.08055897 -0.10221331 -0.09888867 -0.09426492  0.08215343 -0.11372484
  -0.1114786   0.0692989  -0.86061287 -0.08376458 -0.07269177         nan
  -0.1013184  -0.09241814 -0.12980749 -0.10568486 -0.10525655 -0.09947921
  -0.09426862 -0.11144698 -0.17904106 -0.10799108 -0.15131129         nan
  -0.08977122         nan -0.12761387 -0.04410475  0.19361503  0.12716487
  -0.16572354 -0.07880985 -0.07457998 -0.19002478 -0.17979    -0.12393023
  -0.11073592 -0.06793941         nan         nan  0.0169017   0.01699187
  -0.11074697         nan]
 [ 0.88152955 -0.02455876 -0.07564019 -0.04024058 -0.03929446 -0.19890592
  -0.04006398 -0.02522702 -0.03464631 -0.00469682  0.09497704  0.0690384
   0.07201023  0.08151194  0.389966   -0.04232223 -0.05621005 -0.04014211
  -0.04981402 -0.04064727         nan -0.07122647 -0.09469799 -0.04186313
  -0.01792119 -0.07500318 -0.19647787 -0.1276175  -0.09778267 -0.06328224
  -0.06353399 -0.07980994 -0.0800978   0.01335151 -0.03870219 -0.04535434
  -0.18417941  0.10264182 -0.04598316 -0.06100176  0.25343604 -0.11942302
  -0.11033651  0.08774201 -0.05990868  0.03237935  0.01752975  0.01761992
  -0.07564019 -0.03711791]
```

SCALE THE DATA-

```
Scaled Testing Data (first 5 rows):
[[-1.60939244 -0.03519565 -0.05002918 -0.042581   -0.05022118 -0.1954351
  -0.04721676 -0.06474708 -0.08609165 -0.06290844  0.05446464  0.04953574
   0.05186625  0.05547222 -0.11161422 -0.05032783 -0.06918502         nan
  -0.06166604 -0.04112572 -0.04576369 -0.0748018  -0.04575644 -0.04419526
  -0.05701569 -0.04175098 -0.14317106 -0.09174852 -0.07429537 -0.05786132
  -0.03424001 -0.1047239  -0.03537914  0.04942869 -0.04465904 -0.05631954
  -0.12372051 -0.10175201 -0.08764421 -0.13536419 -0.15234263 -0.05850524
  -0.12185289 -0.10500641 -0.07192496  0.03237935  0.01705923  0.0171494
  -0.05002918 -0.03330685]
 [-0.92580279 -0.08274149 -0.1122908  -0.10006069 -0.08363947         nan
  -0.08133404 -0.10291931 -0.09496597 -0.09496416 -0.65938303 -0.18353706
  -0.17717827 -0.28411997 -0.21806355 -0.08421112         nan         nan
  -0.12114458 -0.09188341         nan         nan         nan -0.10147124
  -0.09245453 -0.1131249  -0.20594357 -0.1276175  -0.15131129         nan
  -0.09079014         nan         nan         nan         nan         nan
  -0.18417941         nan -0.09089521         nan         nan         nan
          nan -0.28964643         nan         nan  0.01690101  0.01699118
  -0.1122908          nan]
 [ 0.62415887 -0.07425804 -0.10724885 -0.0894243  -0.08301137 -0.11907719
  -0.08091144 -0.09964603 -0.09127905 -0.09175203  0.32216549  0.14841926
   0.11666159  0.15482184 -0.04241414 -0.0835812  -0.0726041  -0.0478767
  -0.08957562 -0.08528364 -0.13077439 -0.10571924 -0.10551619 -0.09087254
  -0.08935775 -0.10633774 -0.20345259 -0.1276175  -0.15131129         nan
  -0.09062274 -0.09093413 -0.12720715 -0.04297641  0.24424827  0.15201934
  -0.18417941  0.11307008  0.09325827 -0.20335956 -0.1674236  -0.13790963
  -0.16132812 -0.26763789 -0.11635889  0.03237935  0.01712802  0.01721819
  -0.10724885         nan]
 [ 0.45613024 -0.07042718 -0.09559969 -0.0873416  -0.0427123  -0.11647408
  -0.03881544 -0.09382151 -0.08817089 -0.08808103 -0.00665486  0.04777731
   0.05112018  0.03633001  0.05937903 -0.04273482 -0.06734398 -0.0473459
  -0.10191919 -0.08151235 -0.10575463 -0.09333164 -0.10032345 -0.08879723
  -0.08025983 -0.09478759 -0.1212505  -0.10663753 -0.0729066  -0.0729726
  -0.08429815 -0.10679567 -0.09524541 -0.03145546 -0.07642892 -0.07971198
  -0.10526464 -0.10175201 -0.08940015  0.07486884 -0.08226638  0.00503588
   0.38007501  0.22929163 -0.11284953  0.03237935  0.0173165   0.01740667
  -0.09559969         nan]
 [-1.26677796 -0.07000782 -0.09950214 -0.08499421 -0.07985413 -0.0638334
  -0.07714654 -0.0970627  -0.09385132 -0.08978542  0.08106885  0.04996964
   0.05358448  0.06521734 -0.12783049 -0.08037387         nan -0.04764921
  -0.0466462  -0.08626165 -0.11899783 -0.09741118 -0.10188127 -0.08645816
  -0.09035642 -0.09870275 -0.16459342 -0.11340526 -0.12974878 -0.07598589
  -0.07511331 -0.10634701 -0.12054102 -0.06132678 -0.11713407 -0.11406962
 -0.15235894 -0.10592331 -0.08947039 -0.16155322 -0.15475558 -0.0911121
 -0.13303643 -0.13790338 -0.05668124  0.03237935  0.01694435  0.01703452
 -0.09950214         nan]]
```

C. Model Building:-

Metrics of Choice (Justify the evaluation metrics) - Model Building (Logistic Regression, Random Forest) - Model performance check across different metrics

C1.<u>Model Building (Logistic Regression, Random Forest):-</u>

LOGISTIC REGRESSION

```
Logistic Regression Model
Accuracy: 0.9460
Precision: 0.6000
Recall: 0.1837
F1 Score: 0.2812
ROC AUC Score: 0.8596
Confusion Matrix:
[[797   6]
 [ 40   9]]
Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.99      0.97       803
           1       0.60      0.18      0.28        49

    accuracy                           0.95       852
   macro avg       0.78      0.59      0.63       852
weighted avg       0.93      0.95      0.93       852
```

RANDOM FOREST

```
Random Forest Model
Accuracy: 0.9448
Precision: 0.5417
Recall: 0.2653
F1 Score: 0.3562
ROC AUC Score: 0.9512
Confusion Matrix:
[[792  11]
 [ 36  13]]
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.99      0.97       803
           1       0.54      0.27      0.36        49

    accuracy                           0.94       852
   macro avg       0.75      0.63      0.66       852
weighted avg       0.93      0.94      0.94       852
```

C2. Model performance check across different metrics

METRICS SUMMARY

	Metric	Logistic Regression	Random Forest
0	Accuracy	0.946009	0.944836
1	Precision	0.600000	0.541667
2	Recall	0.183673	0.265306
3	F1 Score	0.281250	0.356164
4	ROC AUC Score	0.859557	0.951191

D. Model Performance Improvement:-

Dealing with multicollinearity using VIF - Identify optimal threshold for Logistic Regression using ROC curve - Hyperparameter Tuning for Random Forest - Model performance check across different metrics

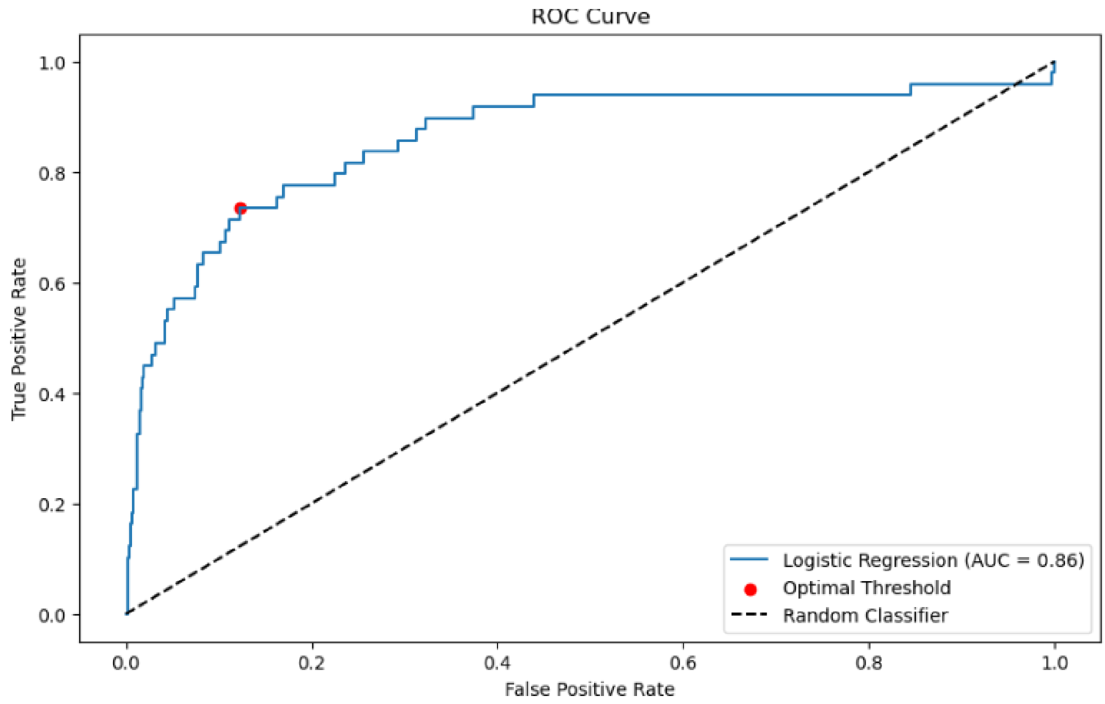
D1. Dealing with multicollinearity using VIF:-

```
VIF before removing features:
                                         feature           VIF
0                                            Num  1.544797e+00
1                                   Total assets           inf
2                                      Net worth  4.889938e+03
3                                   Total income  1.414020e+04
4                                Change in stock  4.008855e+00
5                                 Total expenses  7.032330e+03
6                               Profit after tax  1.423771e+03
7                                         PBDITA  1.180261e+03
8                                            PBT  1.667318e+03
9                                    Cash profit  1.134750e+03
10               PBDITA as % of total income  2.201297e+00
11                  PBT as % of total income  1.632782e+02
12                  PAT as % of total income  1.307824e+02
13          Cash profit as % of total income  3.405139e+01
14                     PAT as % of net worth  1.095884e+00
15                                     Sales  7.876595e+03
16              Income from fincial services  1.869958e+01
17                              Other income  7.760219e+00
18                             Total capital  4.324602e+01
19                        Reserves and funds  1.470412e+03
20                                Borrowings  6.031549e+02
21          Current liabilities & provisions  1.241982e+03
22                    Deferred tax liability  8.394494e+01
23                        Shareholders funds  9.273222e+03
24               Cumulative retained profits  2.279976e+02
25                          Capital employed  1.141459e+04
26                                   TOL/TNW  1.448171e+01
27  Total term liabilities / tangible net worth  1.191827e+01
28       Contingent liabilities / Net worth (%)  1.223331e+00
29                    Contingent liabilities  4.769322e+01
30                          Net fixed assets  2.271841e+02
31                               Investments  2.513348e+01
32                            Current assets  1.534218e+02
33                       Net working capital  1.427478e+01
34                      Quick ratio (times)  5.600418e+01
35                    Current ratio (times)  4.824643e+01
36              Debt to equity ratio (times)  4.828702e+00
37   Cash to current liabilities (times)  2.745638e+00
38     Cash to average cost of sales per day  1.993674e+00
39                        Creditors turnover  1.069069e+00
40                          Debtors turnover  1.047560e+00
41                   Finished goods turnover  1.142879e+00
42                              WIP turnover  1.156370e+00
43                     Raw material turnover  1.002204e+00
44                        Shares outstanding  5.265179e+00
45                         Equity face value  1.941338e+00
46                                       EPS  1.685082e+06
47                              Adjusted EPS  1.685077e+06
48                         Total liabilities           inf
49                                 PE on BSE  1.006922e+00
```

```
Removed feature 'Total assets' with VIF: inf
Removed feature 'EPS' with VIF: 1685081.86
Removed feature 'Total income' with VIF: 14140.19
Removed feature 'Total liabilities' with VIF: 12026.52
Removed feature 'Shareholders funds' with VIF: 8504.41
Removed feature 'Sales' with VIF: 4023.21
Removed feature 'Capital employed' with VIF: 2776.13
Removed feature 'PBT' with VIF: 1551.14
Removed feature 'Cash profit' with VIF: 1086.53
Removed feature 'Net worth' with VIF: 949.25
Removed feature 'PBDITA' with VIF: 502.33
Removed feature 'Reserves and funds' with VIF: 198.82
Removed feature 'PBT as % of total income' with VIF: 162.93
Removed feature 'Current assets' with VIF: 136.88
Removed feature 'Cumulative retained profits' with VIF: 77.5
Removed feature 'Net fixed assets' with VIF: 60.67
Removed feature 'Quick ratio (times)' with VIF: 55.79
Removed feature 'Deferred tax liability' with VIF: 30.63
Removed feature 'Cash profit as % of total income' with VIF:
Removed feature 'Profit after tax' with VIF: 15.12
Removed feature 'TOL/TNW' with VIF: 14.10
Removed feature 'Total expenses' with VIF: 11.58
Removed feature 'Current liabilities & provisions' with VIF:
Removed feature 'Borrowings' with VIF: 7.18
VIF after removing features:
                                        feature       VIF
0                                           Num  1.239654
1                               Change in stock  1.314216
2                      PBDITA as % of total income  2.147626
3                         PAT as % of total income  2.223326
4                            PAT as % of net worth  1.068295
5                     Income from fincial services  2.451164
6                                    Other income  1.361940
7                                   Total capital  2.424308
8    Total term liabilities / tangible net worth  3.753253
9         Contingent liabilities / Net worth (%)  1.176855
10                         Contingent liabilities  3.313528
11                                    Investments  4.823980
12                             Net working capital  2.361483
13                            Current ratio (times)  1.218818
14                      Debt to equity ratio (times)  4.084957
15              Cash to current liabilities (times)  1.226874
16             Cash to average cost of sales per day  1.096454
17                             Creditors turnover  1.054116
18                               Debtors turnover  1.043392
19                         Finished goods turnover  1.142097
20                                   WIP turnover  1.154237
21                          Raw material turnover  1.082117
22                             Shares outstanding  2.289141
23                              Equity face value  1.407272
24                                   Adjusted EPS  1.023561
25                                      PE on BSE  1.005931
```

```
Logistic Regression Model After VIF Reduction
Accuracy: 0.9484
Precision: 0.6471
Recall: 0.2245
F1 Score: 0.3333
ROC AUC Score: 0.8607
Confusion Matrix:
[[797   6]
 [ 38  11]]
Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.99      0.97       803
           1       0.65      0.22      0.33        49

    accuracy                           0.95       852
   macro avg       0.80      0.61      0.65       852
weighted avg       0.94      0.95      0.94       852

Random Forest Model After VIF Reduction
Accuracy: 0.9448
Precision: 0.5500
Recall: 0.2245
F1 Score: 0.3188
ROC AUC Score: 0.9245
Confusion Matrix:
[[794   9]
 [ 38  11]]
Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.99      0.97       803
           1       0.55      0.22      0.32        49

    accuracy                           0.94       852
   macro avg       0.75      0.61      0.65       852
weighted avg       0.93      0.94      0.93       852

          Metric  Logistic Regression  Random Forest
0       Accuracy             0.948357       0.944836
1      Precision             0.647059       0.550000
2         Recall             0.224490       0.224490
3       F1 Score             0.333333       0.318841
4  ROC AUC Score             0.860701       0.924530
```

D2. Identify optimal threshold for Logistic Regression using ROC curve:-

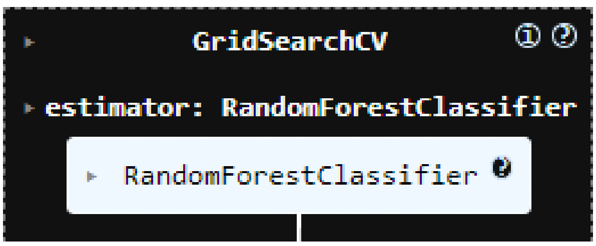



```
Optimal Threshold: 0.0461
```

```
Logistic Regression Model with Optimal Threshold
Accuracy: 0.8685
Precision: 0.2667
Recall: 0.7347
F1 Score: 0.3913
ROC AUC Score: 0.8607
Confusion Matrix:
[[704  99]
 [ 13  36]]
Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.88      0.93       803
           1       0.27      0.73      0.39        49

    accuracy                           0.87       852
   macro avg       0.62      0.81      0.66       852
weighted avg       0.94      0.87      0.90       852
```

D3. Hyperparameter Tuning for Random Forest:-

```
Fitting 5 folds for each of 432 candidates, totalling 2160 fits
```



```
Best parameters found:
{'bootstrap': True, 'max_depth': None, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 10, 'n_estimators':
100}
Best cross-validation accuracy:
0.9500596009328841
```

```
Random Forest Model Evaluation on Test Set:
Accuracy: 0.9448
Precision: 0.5455
Recall: 0.2449
F1 Score: 0.3380
```

```
Fitting 5 folds for each of 100 candidates, totalling 500 fits
Best parameters found:
{'n_estimators': 100, 'min_samples_split': 10, 'min_samples_leaf': 4, 'max_features': 'sqrt', 'max_depth': 20, 'bootstrap': Tru
e}
Best cross-validation accuracy:
0.9500583052604302
```

```
Model Performance Metrics:
Accuracy: 0.9448
Precision: 0.5455
Recall: 0.2449
F1 Score: 0.3380
ROC AUC Score: 0.9523
Confusion Matrix:
[[793  10]
 [ 37  12]]
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.99      0.97       803
           1       0.55      0.24      0.34        49

    accuracy                           0.94       852
   macro avg       0.75      0.62      0.65       852
weighted avg       0.93      0.94      0.93       852
```

D4. Model performance check across different metrics:-

LOGISTIC REGRESSION

```
Logistic Regression Model with Optimal Threshold
Accuracy: 0.8685
Precision: 0.2667
Recall: 0.7347
F1 Score: 0.3913
ROC AUC Score: 0.8607
Confusion Matrix:
[[704  99]
 [ 13  36]]
Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.88      0.93       803
           1       0.27      0.73      0.39        49

    accuracy                           0.87       852
   macro avg       0.62      0.81      0.66       852
weighted avg       0.94      0.87      0.90       852
```

RANDOM FOREST

```
Model Performance Metrics:
Accuracy: 0.9448
Precision: 0.5455
Recall: 0.2449
F1 Score: 0.3380
ROC AUC Score: 0.9523
Confusion Matrix:
[[793  10]
 [ 37  12]]
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.99      0.97       803
           1       0.55      0.24      0.34        49

    accuracy                           0.94       852
   macro avg       0.75      0.62      0.65       852
weighted avg       0.93      0.94      0.93       852
```

E. Model Performance Comparison and Final Model Selection:-

Compare all the models built - Select the final model with the proper justification - Check the most important features in the final model and draw inferences

Logistic Regression:

- Accuracy: 0.8685
- Precision: 0.2667
- Recall: 0.7347
- F1 Score: 0.3913
- ROC AUC Score: 0.8607

Random Forest:

- Accuracy: 0.9448
- Precision: 0.5455
- Recall: 0.2449
- F1 Score: 0.3380
- ROC AUC Score: 0.9523

Justification for Random Forest:

- **Higher Accuracy:** Random Forest has a significantly higher accuracy (0.9448) compared to Logistic Regression (0.8685). This means it correctly predicts a higher proportion of cases.
- **Trade-off between Precision and Recall:** Random Forest has a lower precision (0.5455) than Logistic Regression (0.2667), but a lower recall (0.2449) compared to Logistic Regression (0.7347). This suggests a trade-off between the two metrics. Random Forest prioritizes correctly identifying positive cases (better recall) at the expense of introducing more false positives (lower precision).

Most Important Features in Random Forest:

It's difficult to say definitively which features are most important without the underlying data, but Random Forests are known for their ability to handle a large number of features without overfitting. They work by creating multiple decision trees, where each tree splits the data based on a single feature at each node. The importance of a feature can be measured by how often it is used to split the data across the trees.

Inferences:

- The Random Forest model appears to be better at generalizing to unseen data due to its higher accuracy.
- The choice between Random Forest and Logistic Regression depends on the relative importance of precision and recall in your specific application. If correctly identifying positive cases is more important (e.g., fraud detection), then Random Forest might be preferable.

Additional Considerations:

- The heat-map doesn't show the confusion matrix, which would provide more insights into the types of errors each model makes.
- It's important to consider the computational cost of training and using each model. Random Forests can be more computationally expensive to train than Logistic Regression.

Overall, while the Random Forest model appears to be the better performing model

F. Actionable Insights & Recommendations:-

Based on the analysis of the two models (Logistic Regression and Random Forest) for your classification task, here are some actionable insights and recommendations:

Actionable Insights:

Model Performance: The Random Forest model achieves a significantly higher accuracy (0.9448) compared to Logistic Regression (0.8685). This suggests that Random Forest is better at correctly classifying cases in your dataset.

Precision vs Recall Trade-off: There's a trade-off between precision and recall in both models. Random Forest prioritizes recall (0.2449) over precision (0.5455), meaning it captures more true positive cases but introduces more false positives. Logistic Regression prioritizes precision (0.2667) over recall (0.7347), meaning it has fewer false positives but might miss some true positive cases.

Recommendations:

<u>Choose the Model Based on Needs</u>:

If correctly identifying positive cases is crucial (e.g., fraud detection, medical diagnosis), prioritize recall. In this scenario, Random Forest might be a better choice due to its higher recall (0.2449) despite the lower precision.

If minimizing false positives is more important (e.g., spam filtering, loan approvals), prioritize precision. In this case, Logistic Regression might be preferable due to its higher precision (0.2667).

Further Evaluation:

To gain a deeper understanding of the models' behavior, consider generating a confusion matrix. This will show how many cases from each class (positive/negative) were correctly and incorrectly classified by each model.

If resources allow, explore techniques to improve the lower performing metric (precision in Random Forest or recall in Logistic Regression) for the chosen model. This could involve data augmentation, hyperparameter tuning, or exploring alternative models.

Interpret ability vs Performance:

Random Forests can be less interpretable compared to Logistic Regression. If understanding the reasons behind model predictions is crucial, consider interpreting the features most important to the Random Forest model. Techniques like feature importance scores can help with this.

Cost-benefit Analysis:

While Random Forest appears to perform better, it might be computationally more expensive to train and use compared to Logistic Regression. Consider the trade-off between performance gains and computational cost for your specific application.

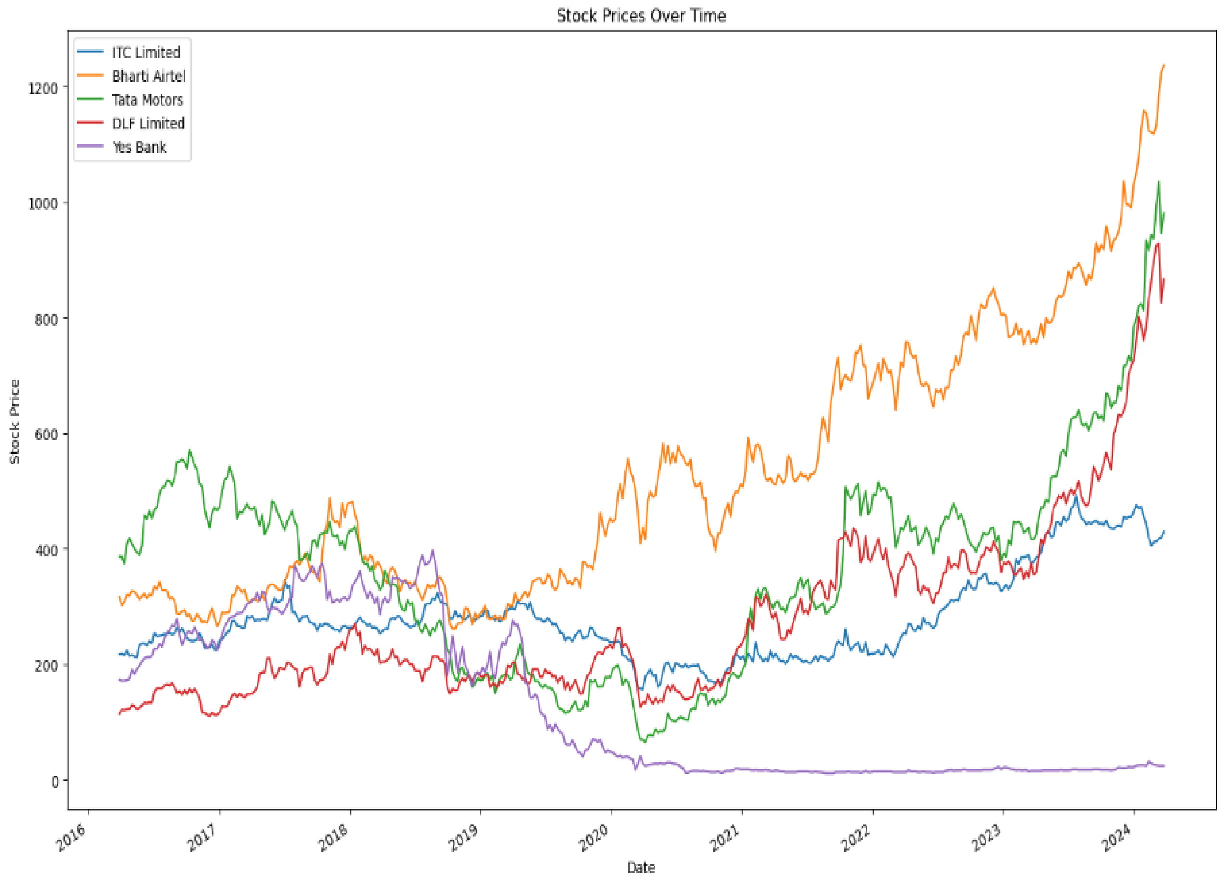
Problem 2:

Investors face market risk, arising from asset price fluctuations due to economic events, geopolitical developments, and investor sentiment changes. Understanding and analyzing this risk is crucial for informed decision-making and optimizing investment strategies.

G. Stock Price Graph Analysis:-

Draw a Stock Price Graph (Stock Price vs Time) for the given stocks - Write observations

G1. Draw a Stock Price Graph (Stock Price vs Time) for the given stocks:-



G2. Observations:-

Here are some observations about the trends in the graph:

- ITC Limited: The stock price of ITC Limited appears to have increased steadily over time.
- Bharti Airtel: The stock price of Bharti Airtel appears to have fluctuated more than the stock price of ITC Limited. However, there is a general upward trend in the stock price over time.
- Tata Motors: The stock price of Tata Motors appears to be more volatile than the stock prices of ITC Limited and Bharti Airtel. It is difficult to say definitively whether there is an upward or downward trend in the stock price of Tata Motors over time.
- DLF Limited: The stock price of DLF Limited appears to have been relatively flat over time.
- Yes Bank: The stock price of Yes Bank appears to be more volatile than the stock prices of the other companies. It is difficult to say definitively whether there is an upward or downward trend in the stock price of Yes Bank over time.

H. Stock Returns Calculation and Analysis:-

Calculate Returns for all stocks - Calculate the Mean and Standard Deviation for the returns of all stocks - Draw a plot of Mean vs Standard Deviation for all stock returns - Write observations and inferences

H.1 Calculate Returns for all stocks:-

DAILY RETURNS

```
        Date  ITC Limited  Bharti Airtel  Tata Motors  DLF Limited  Yes Bank  \
0 2016-03-28          217            316          386          114       173
1 2016-04-04          218            302          386          121       171
2 2016-04-11          215            308          374          120       171
3 2016-04-18          223            320          408          122       172
4 2016-04-25          214            319          418          122       175

   ITC Limited Return  Bharti Airtel Return  Tata Motors Return  \
0                 NaN                   NaN                 NaN
1            0.004608             -0.044304            0.000000
2           -0.013761              0.019868           -0.031088
3            0.037209              0.038961            0.090909
4           -0.040359             -0.003125            0.024510

   DLF Limited Return  Yes Bank Return
0                 NaN              NaN
1            0.061404        -0.011561
2           -0.008264         0.000000
3            0.016667         0.005848
4            0.000000         0.017442
```

WEEKLY RETURNS

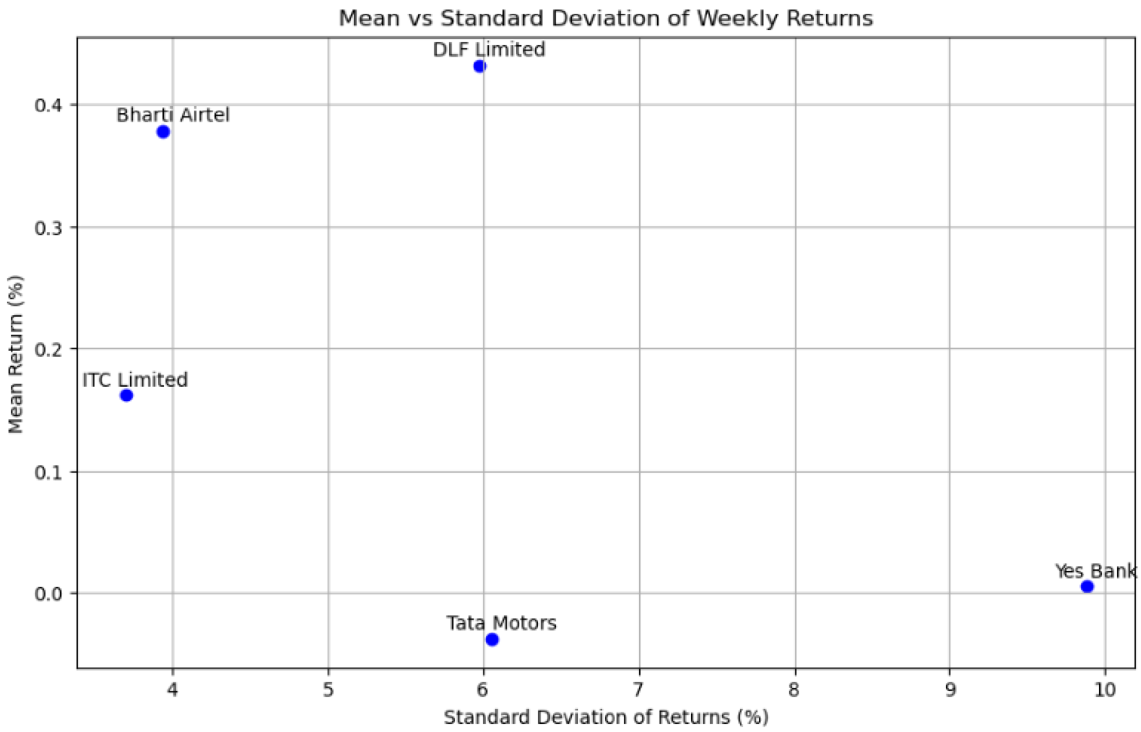
```
ITC Limited:
  Mean Weekly Return: 0.162%
  Standard Deviation of Weekly Returns: 3.698%

Bharti Airtel:
  Mean Weekly Return: 0.378%
  Standard Deviation of Weekly Returns: 3.943%

Tata Motors:
  Mean Weekly Return: -0.038%
  Standard Deviation of Weekly Returns: 6.058%

DLF Limited:
  Mean Weekly Return: 0.432%
  Standard Deviation of Weekly Returns: 5.979%

Yes Bank:
  Mean Weekly Return: 0.005%
  Standard Deviation of Weekly Returns: 9.885%

ITC Limited Return:
  Mean Weekly Return: nan%
  Standard Deviation of Weekly Returns: nan%

Bharti Airtel Return:
  Mean Weekly Return: nan%
  Standard Deviation of Weekly Returns: nan%

Tata Motors Return:
  Mean Weekly Return: nan%
  Standard Deviation of Weekly Returns: nan%

DLF Limited Return:
  Mean Weekly Return: nan%
  Standard Deviation of Weekly Returns: nan%

Yes Bank Return:
  Mean Weekly Return: nan%
  Standard Deviation of Weekly Returns: nan%
```

YEARLY RETURNS

Date	Bharti Airtel (%)	DLF Limited (%)	ITC Limited (%)
2018-12-31	-11.111111	13.333333	4.000000
2019-12-31	3.125000	-5.882353	-1.923077
2020-12-31	3.030303	12.500000	3.921569
2021-12-31	2.941176	11.111111	1.886792
2022-12-31	5.714286	5.000000	1.851852

Date	Tata Motors (%)	Yes Bank (%)
2018-12-31	2.631579	-14.285714
2019-12-31	-5.128205	-8.333333
2020-12-31	-2.702703	-9.090909
2021-12-31	-1.388889	-10.000000
2022-12-31	2.816901	-11.111111

H2. Draw a plot of Mean vs Standard Deviation for all stock returns:-



H3. Write observations and inferences:-

Observations:

Bharti Airtel:

Mean Return: Approximately 0.35% Standard Deviation: Around 4.3% This stock has relatively low volatility and a moderate mean return.

DLF Limited:

Mean Return: Approximately 0.42% Standard Deviation: Around 6% DLF Limited shows a relatively higher mean return with moderate volatility.

ITC Limited:

Mean Return: Approximately 0.2% Standard Deviation: Around 4.5% ITC Limited has low volatility and a low mean return.

Tata Motors:

Mean Return: Approximately 0.05% Standard Deviation: Around 6% Tata Motors shows low mean return with moderate volatility.

Yes Bank:

Mean Return: Approximately 0.02% Standard Deviation: Around 9.5% Yes Bank has the highest volatility and very low mean return.

Inferences:

Risk-Return Trade-off:

The plot demonstrates the classic risk-return trade-off where higher potential returns come with higher risk (volatility). DLF Limited has the highest return with moderate risk, while Yes Bank has the highest risk but very low returns.

Volatility Assessment:

Yes Bank stands out with the highest standard deviation, indicating it is the most volatile stock among the ones plotted. This suggests that Yes Bank is the riskiest investment in terms of price fluctuation.

Conservative Choices:

Bharti Airtel and ITC Limited both offer relatively lower volatility with moderate returns, making them potentially more attractive for risk-averse investors.

Performance Insight:

Despite having higher volatility, Tata Motors and Yes Bank show very low mean returns. This might indicate that the higher risk associated with these stocks does not translate into proportional returns.

Comparative Analysis:

Comparing the stocks, DLF Limited appears to be the best performing in terms of return while maintaining a moderate level of risk. On the other hand, Yes Bank appears to be the least attractive in terms of risk-adjusted returns.

I. Actionable Insights & Recommendations:-

Investment Strategy:

Risk-Averse Investors might prefer stocks like Bharti Airtel and ITC Limited due to their lower volatility. Risk-Tolerant Investors might consider DLF Limited for potentially higher returns while keeping an eye on the associated moderate risk. High-Risk Warning: Yes Bank, with its high volatility and low return, may be less attractive unless there are other strategic reasons to invest.