PREDICTIVE MODEL

Linear Regression :

You are a part of an investment firm and your work is to do research about these 738 firms.
You are provided with the dataset containing the sales and other attributes of these 738 firms. Predict the sales of these firms on the bases of the details given in the dataset so as to
help your company in investing consciously. Also, provide them with 5 attributes that are most important

Data Dictionary :

1. Sales: Sales (in millions of dollars).
2. Capital: Net stock of property, plant, and equipment.
3. Patents: Granted patents.
4. Randd: R&D stock (in millions of dollars).
5. Employment: Employment (in 1000s).
6. sp500: Membership of firms in the S&P 500 index. S&P is a stock market index that measures the stock performance of 500 large companies listed on stock exchanges in the United States
7. Tobinq: Tobin's q (also known as q ratio and Kaldor's v) is the ratio between a physical asset's market value and its replacement value.
8. Value: Stock market value.
9. Institutions: Proportion of stock owned by institution

Problem 1.1:

Exploratory Data Analysis- Problem definition - Data background and contents - Univariate analysis - Bivariate analysis - Insights based on EDA

A) Rows And Column Views : (Rows - 738,9)

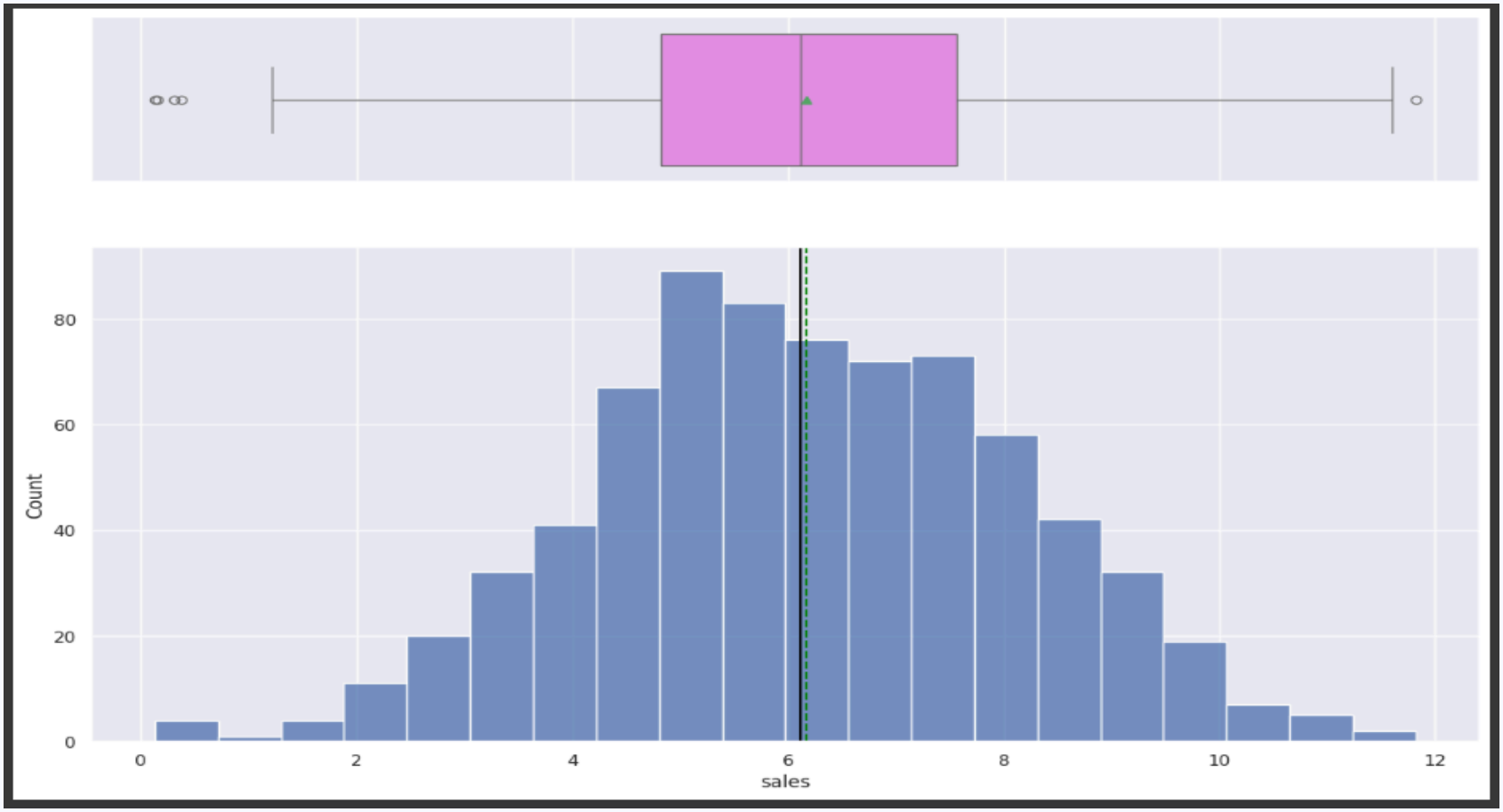
	sales	capital	patents	randd	employment	sp500	tobinq	value	institutions
0	6.719007	161.603986	10	382.078247	2.306000	no	11.049511	1625.453755	80.27
1	6.013113	122.101012	2	0.000000	1.860000	no	0.844187	243.117082	59.02
2	9.037039	6221.144614	138	3296.700439	49.659005	yes	5.205257	25865.233800	47.70
3	6.113682	266.899987	1	83.540161	3.071000	no	0.305221	63.024630	26.88
4	5.170075	140.124004	2	14.233637	1.947000	no	1.063300	67.406408	49.46

Statistical summary of the dataset :

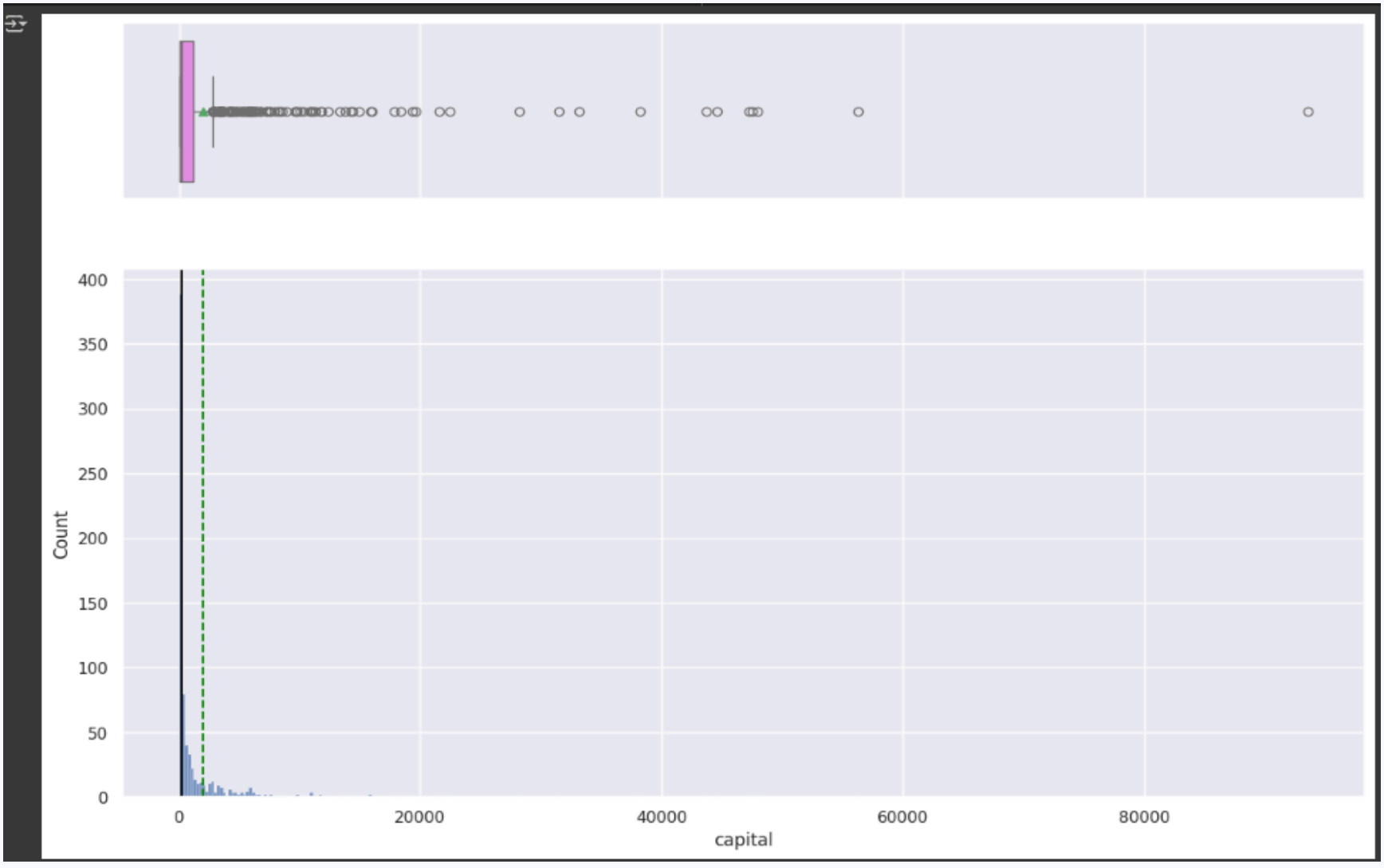
	sales	capital	patents	randd	employment	tobinq	value	institutions
count	738.000000	738.000000	738.000000	738.000000	738.000000	738.000000	738.000000	738.000000
mean	6.167375	2028.505862	26.322493	449.882342	14.497931	2.794910	2797.505101	43.104390
std	1.959889	6550.941548	98.569850	2034.803742	43.887131	3.366591	7159.919660	21.732167
min	0.129272	0.057000	0.000000	0.000000	0.006000	0.119001	1.971053	0.000000
25%	4.819261	52.832747	1.000000	4.621146	0.926250	1.018783	102.982570	25.430000
50%	6.114605	205.811964	3.000000	36.824968	3.035500	1.680303	418.519363	44.430000
75%	7.557078	1147.033133	11.750000	145.604980	10.358501	3.139309	2091.235844	60.547500
max	11.818186	93625.200560	1220.000000	30425.255860	710.799925	20.000000	95191.591160	90.150000

Univariate Analysis

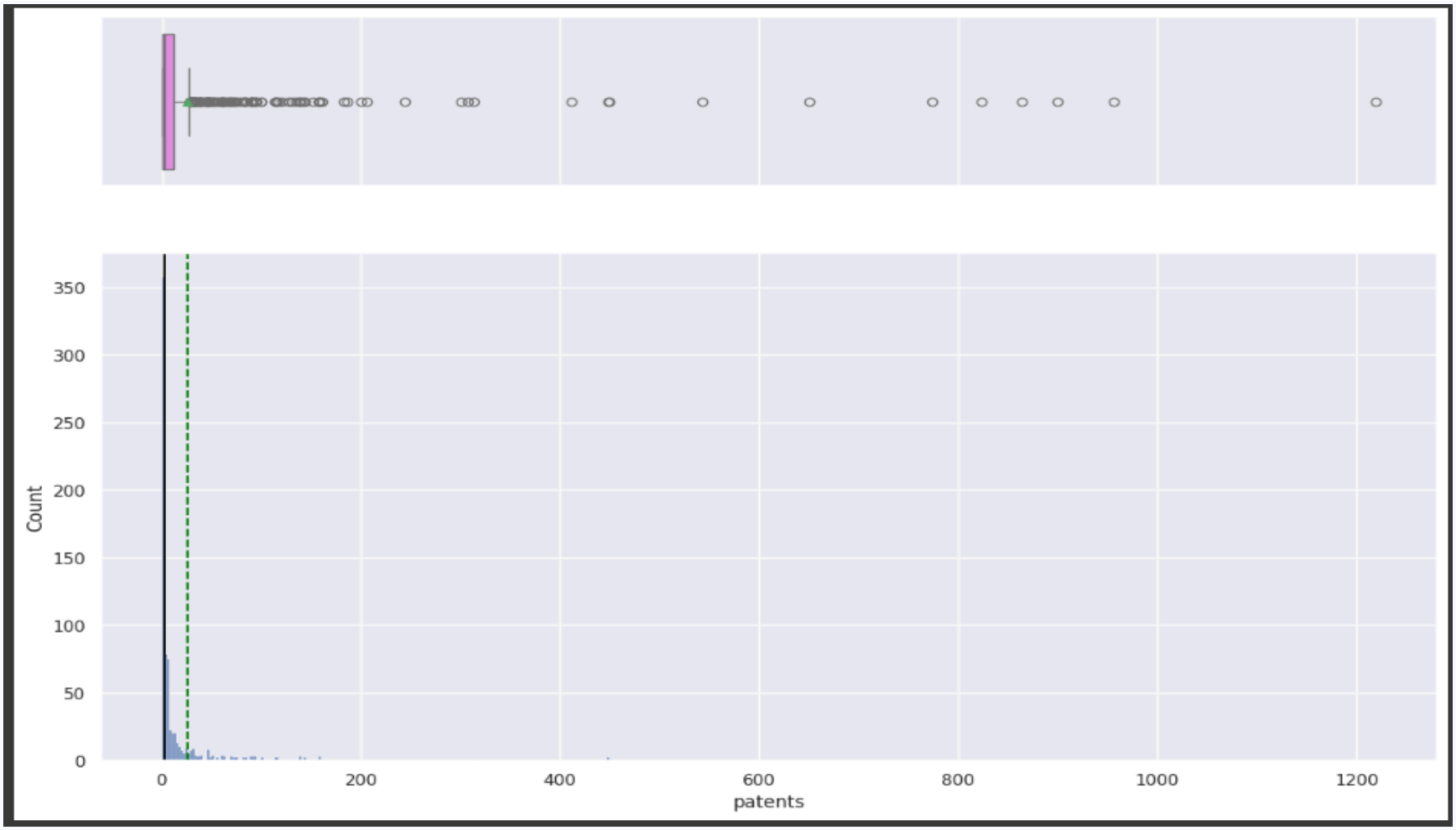
Histplot (count vs sales) :



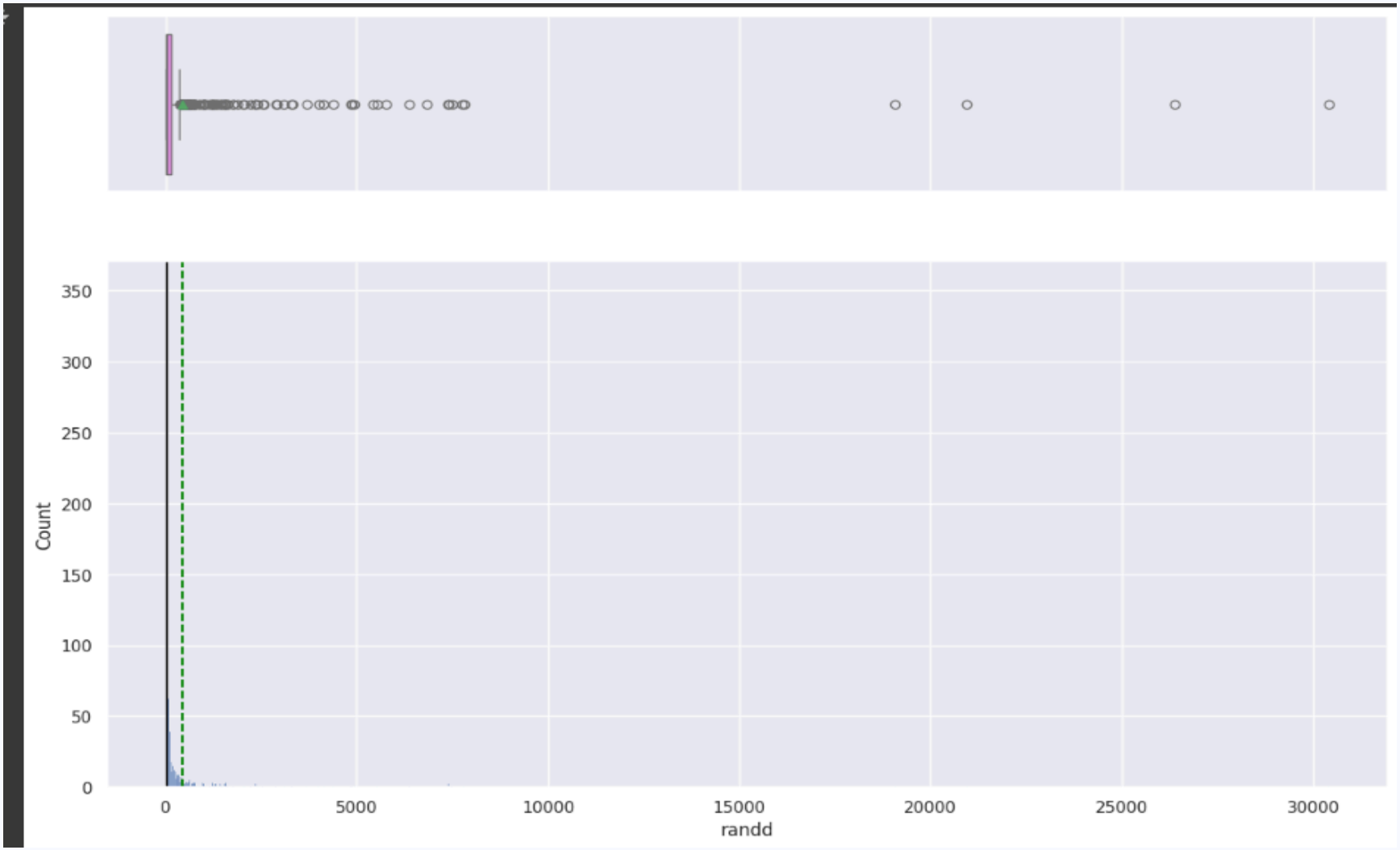
Histplot (count vs capital) :



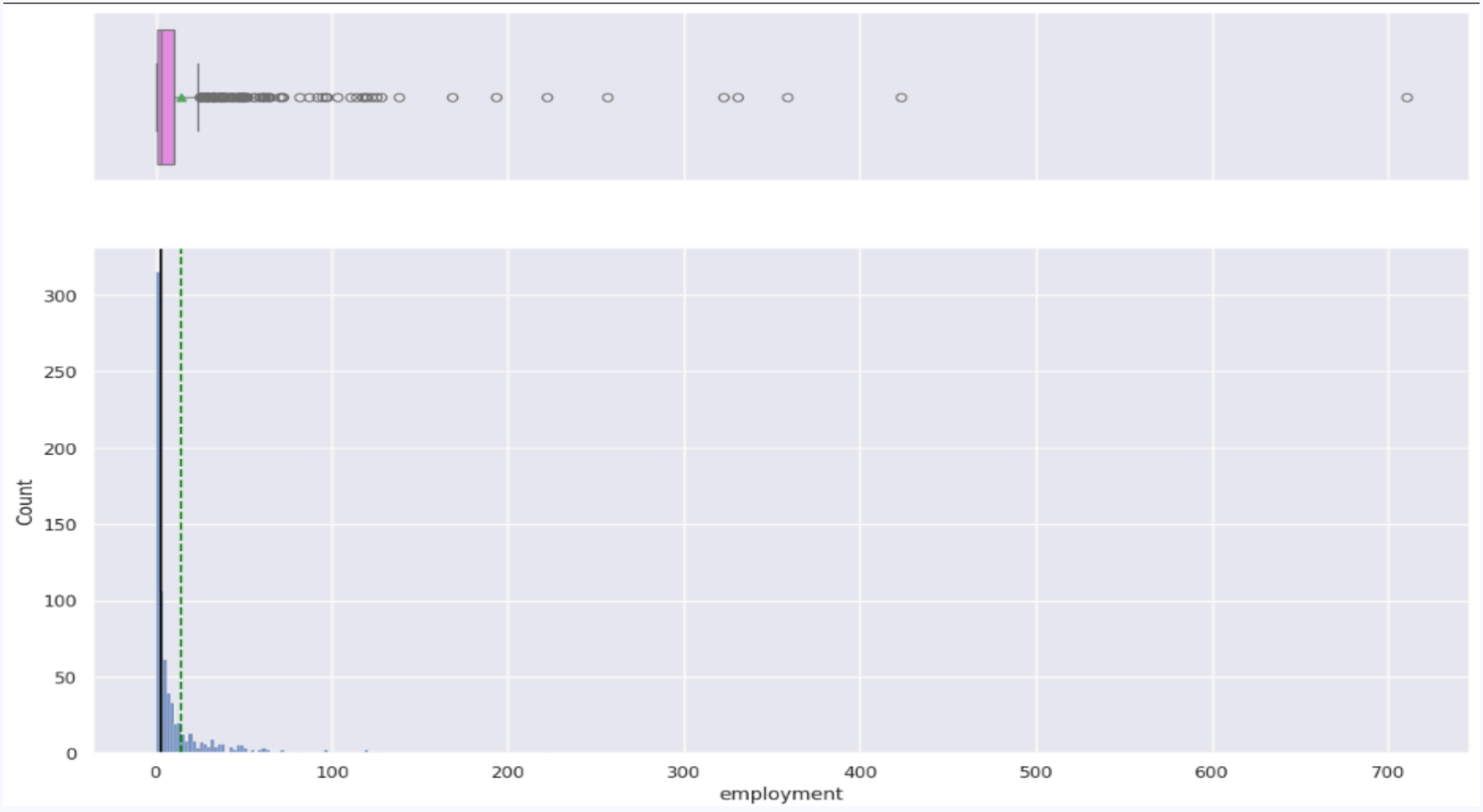
Histplot (count vs patents) :



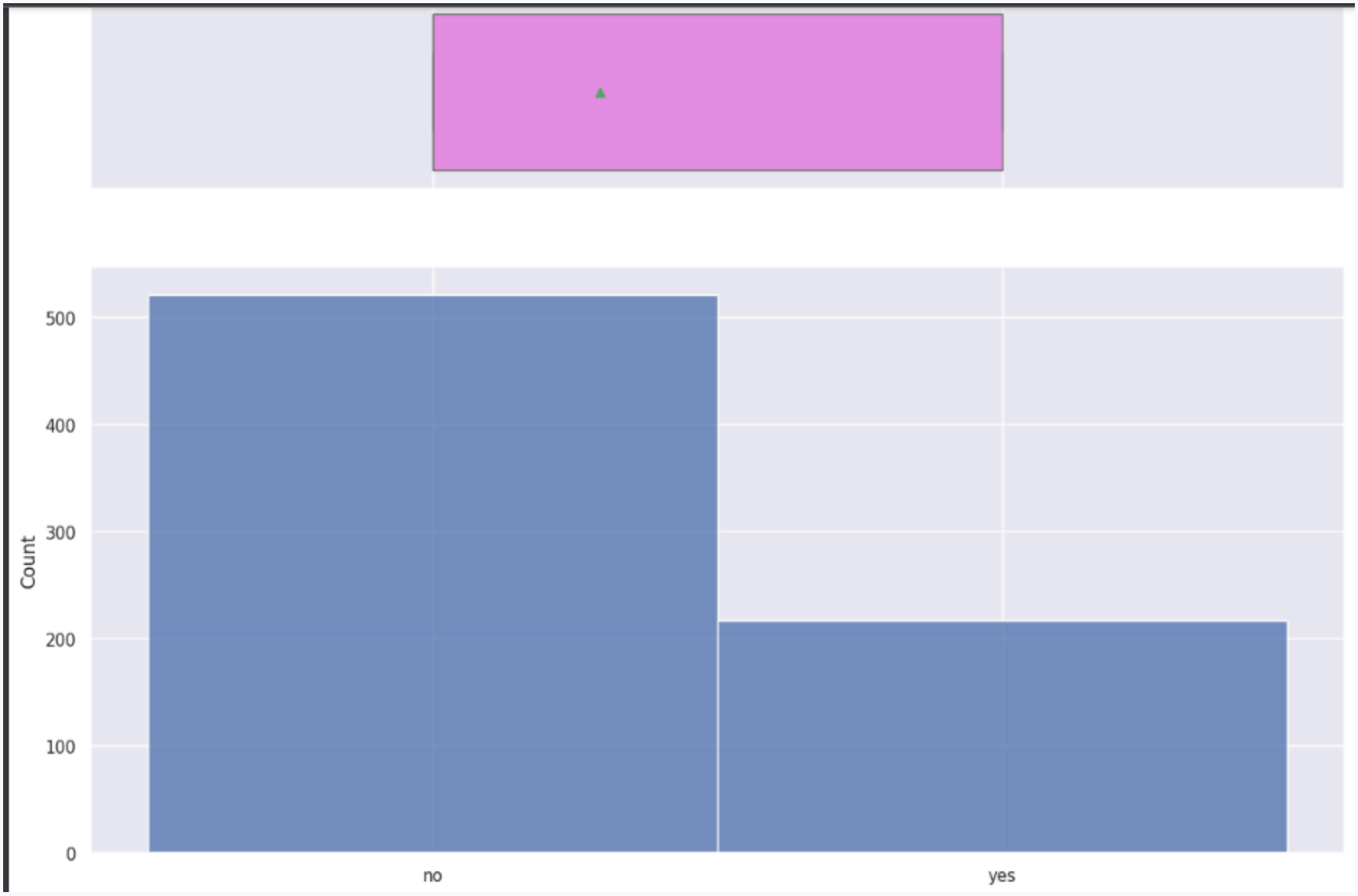
Histplot (count vs randd) :



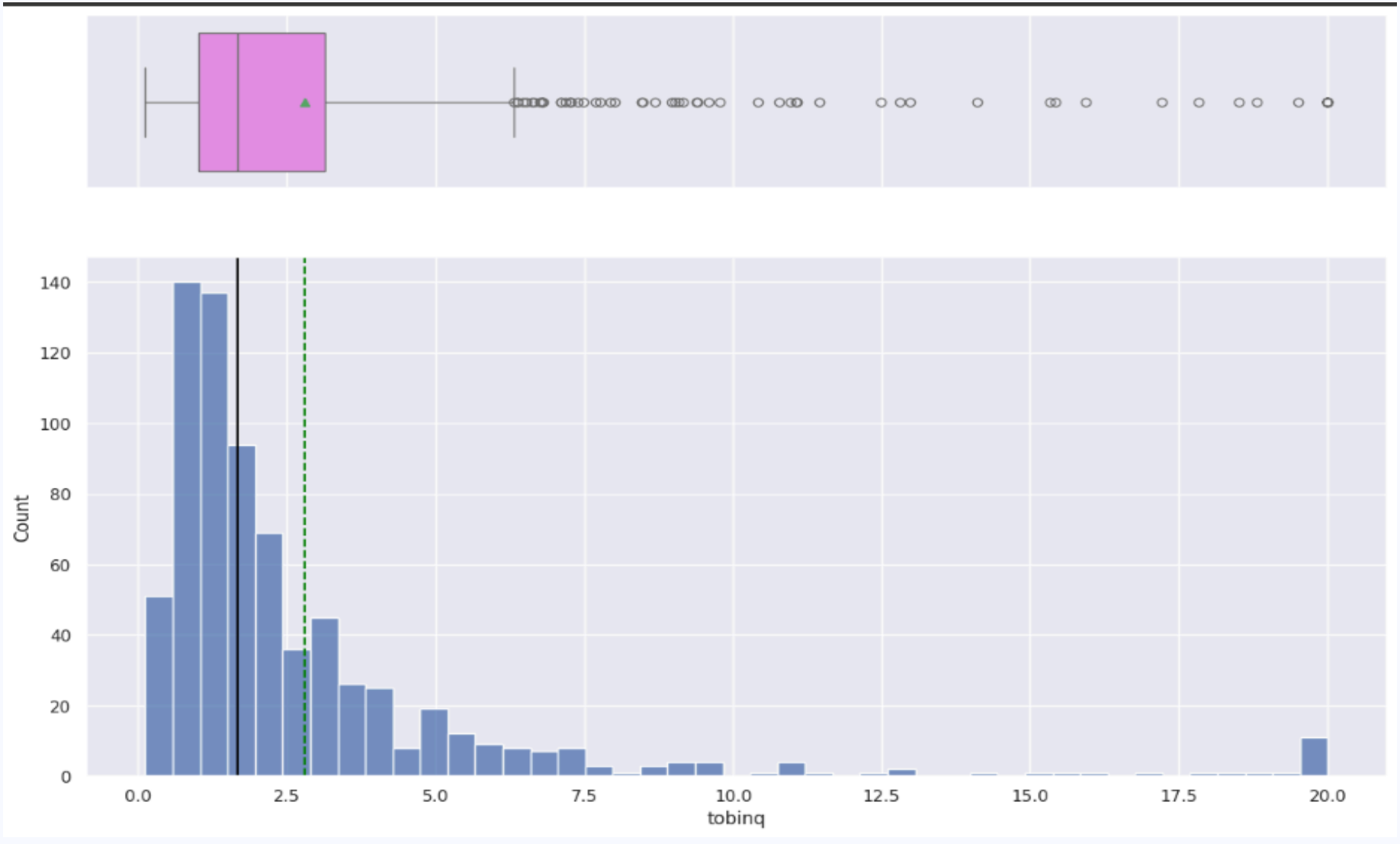
Histplot (count vs employment) :



Histplot (count Vs sp500) :



Histplot (count vs tobinq) :



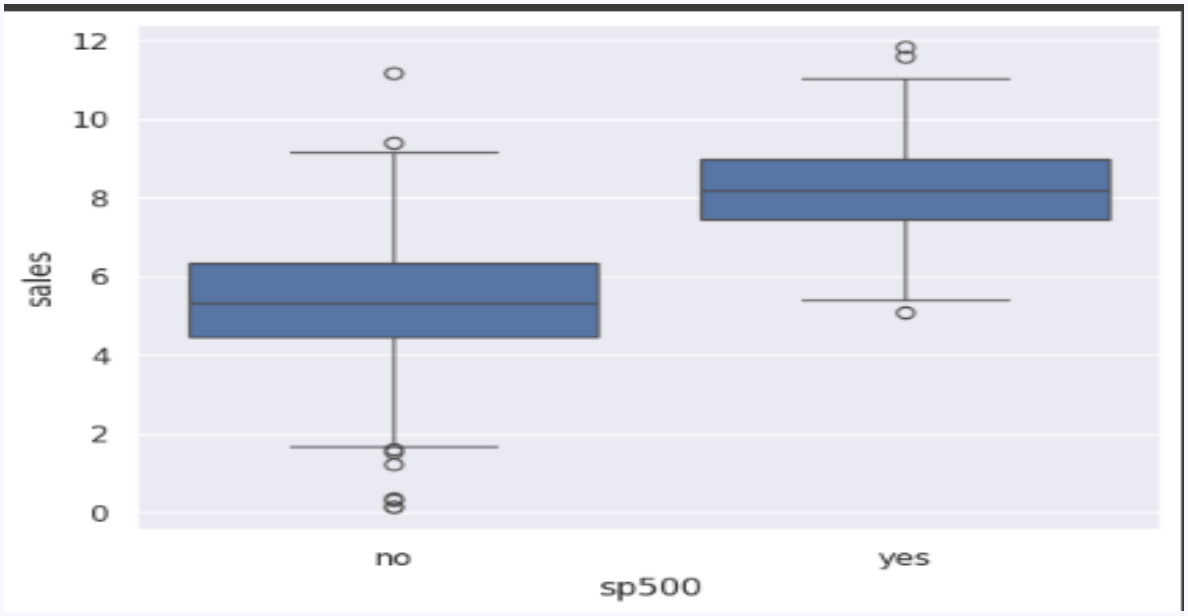
Histplot (count vs value) :

Insights using univariate analysis **1)** Mostly the sales of the firms is around 10000.

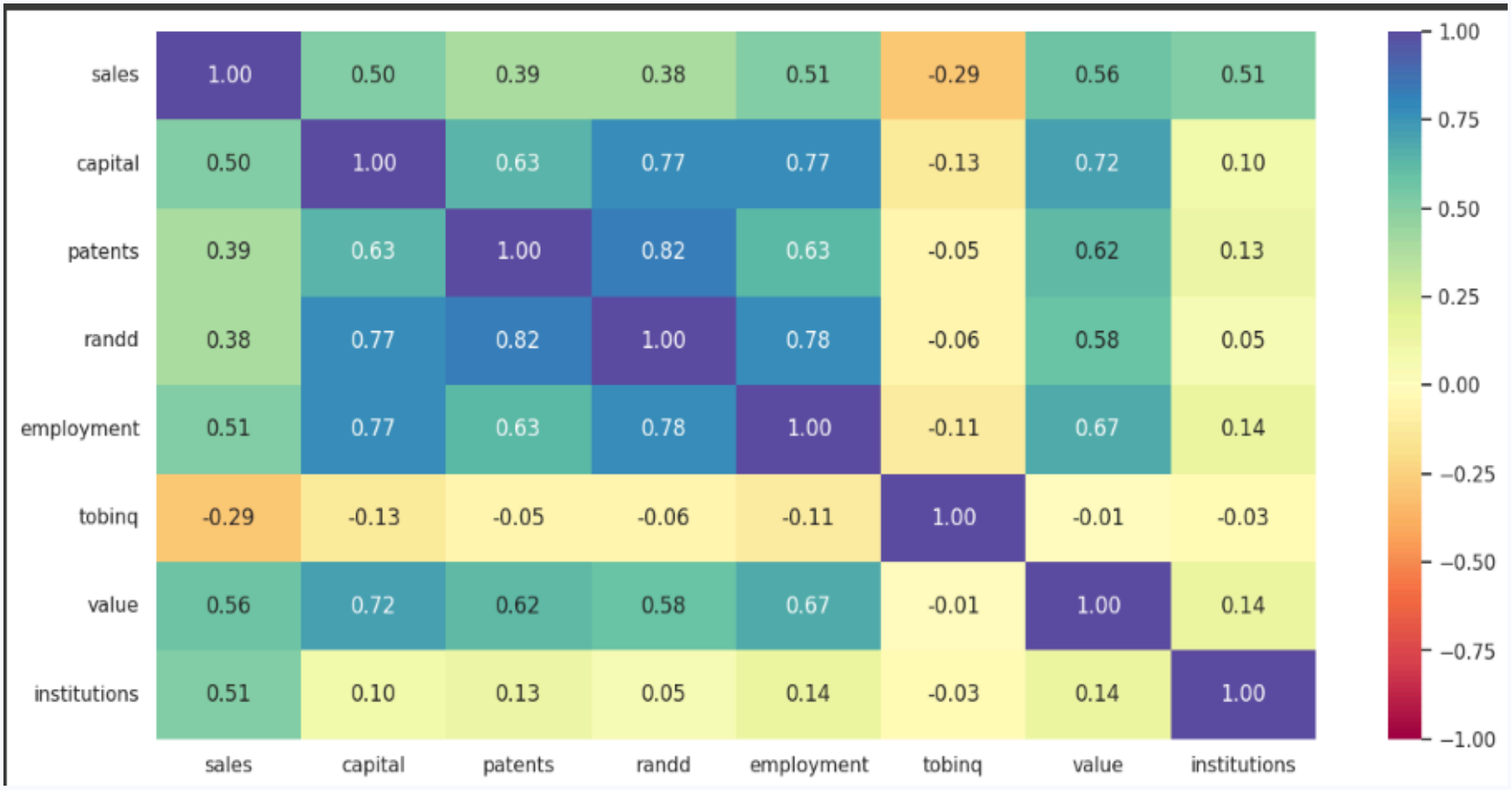
2)Most of the firms are having capital less than 25000 and the patents acquired by them were also in small quantities. So, this indicates that most of the firms are smaller or mid-sized ones.However, there are certain outlier populations indicating the firms belonging to larger sizes.

BIVARIATE ANALYSIS :

Boxplot : Sales Vs sp500



Heat Map :



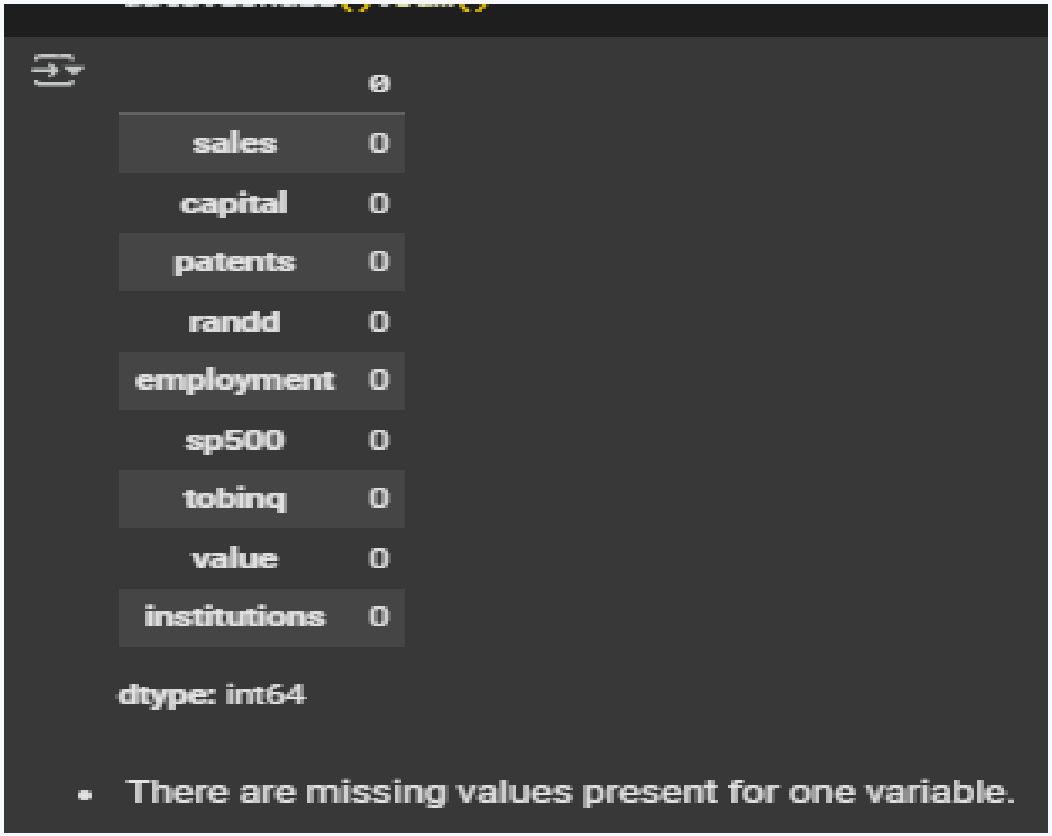
When the capital is High the Patents are also high. This indicates that firms with higher capital are holding many patents.

Problem 1.2: Data Preprocessing

- Duplicate value check - Missing value check and treatment - Outlier check (treatment if needed) - Feature engineering (if needed) - Data preparation for modeling

There is no duplicate value in the dataset

Missing Value Check :



	0
sales	0
capital	0
patents	0
randd	0
employment	0
sp500	0
tobinq	0
value	0
institutions	0

dtype: int64

- There are missing values present for one variable.

Data Preparation for Modeling :

Defining the dependent and independent variables :

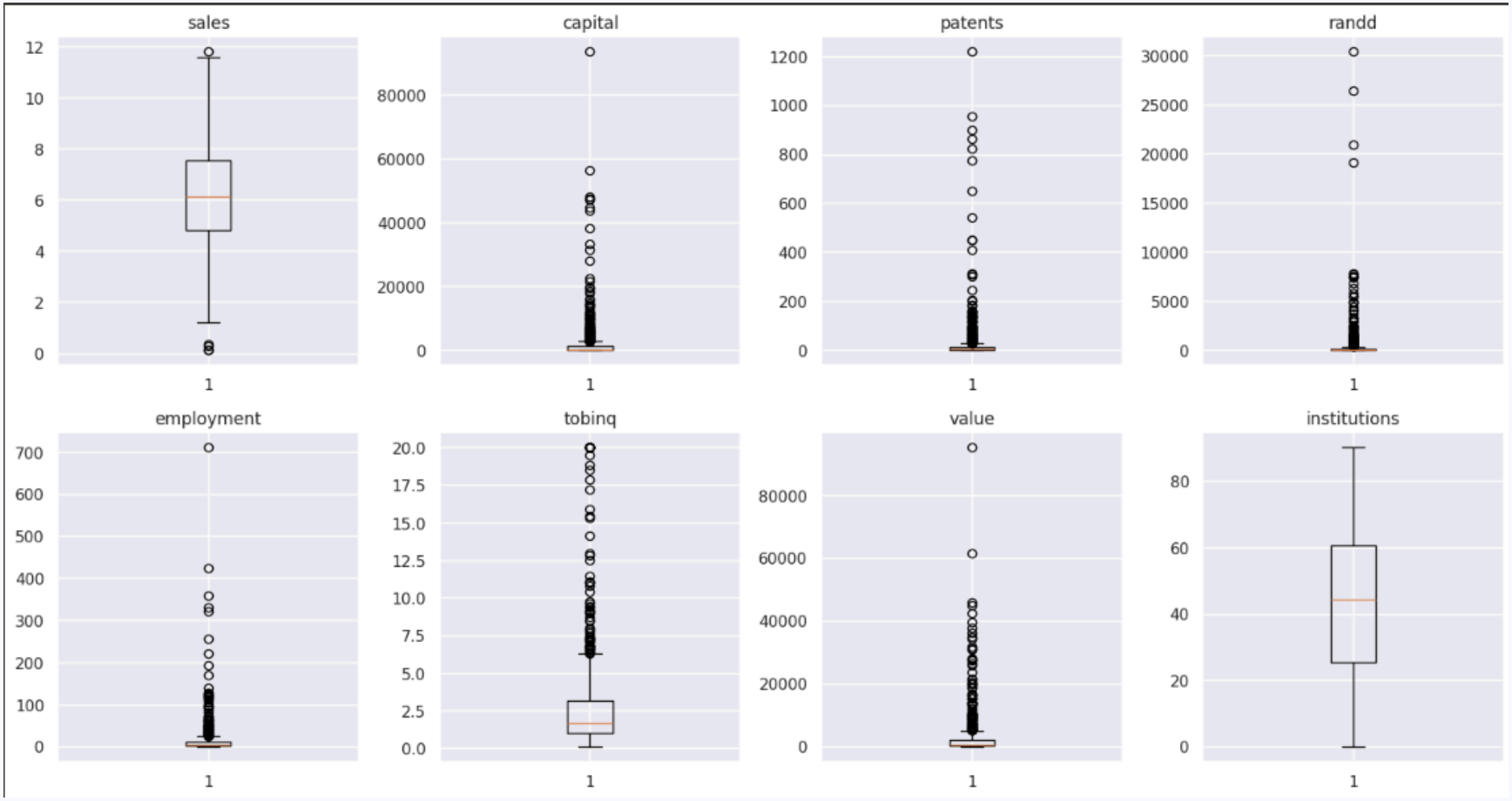
```
0    6.719007
1    6.013113
2    9.037039
3    6.113682
4    5.170075
Name: sales, dtype: float64
       capital  patents        randd  employment sp500     tobinq  \
0   161.603986       10   382.078247    2.306000    no  11.049511
1   122.101012        2     0.000000    1.860000    no   0.844187
2  6221.144614      138  3296.700439   49.659005   yes   5.205257
3   266.899987        1    83.540161    3.071000    no   0.305221
4   140.124004        2    14.233637    1.947000    no   1.063300

           value  institutions
0   1625.453755         80.27
1    243.117082         59.02
2  25865.233800         47.70
3     63.024630         26.88
4     67.406408         49.46
```

Creating dummy variables :

	capital	patents	randd	employment	tobinq	value	institutions	sp500_yes
0	161.603986	10.0	382.078247	2.306000	11.049511	1625.453755	80.27	0.0
1	122.101012	2.0	0.000000	1.860000	0.844187	243.117082	59.02	0.0
2	6221.144614	138.0	3296.700439	49.659005	5.205257	25865.233800	47.70	1.0
3	266.899987	1.0	83.540161	3.071000	0.305221	63.024630	26.88	0.0
4	140.124004	2.0	14.233637	1.947000	1.063300	67.406408	49.46	0.0

```
Number of rows in train data = 590
Number of rows in test data = 148
```

Outlier Check :



The black circles represent the outliers and it is present in all the columns except Institutions. The majority of the variables are highly skewed towards the right.

Problem 1.3:Model Building - Linear Regression

- Build the model and comment on the model statistics - Display model coefficients with column names

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  sales   R-squared:                       0.669
Model:                            OLS   Adj. R-squared:                  0.664
Method:                 Least Squares   F-statistic:                     146.7
Date:                Sun, 23 Feb 2025   Prob (F-statistic):          4.33e-134
Time:                        03:39:30   Log-Likelihood:                -907.85
No. Observations:                 590   AIC:                             1834.
Df Residuals:                     581   BIC:                             1873.
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          4.7634      0.116     41.158      0.000       4.536       4.991
capital     1.893e-05   1.38e-05      1.370      0.171   -8.21e-06    4.61e-05
patents       -0.0007      0.001     -0.886      0.376      -0.002       0.001
randd       3.853e-05   5.56e-05      0.693      0.489   -7.07e-05       0.000
employment     0.0032      0.002      1.699      0.090      -0.000       0.007
tobinq        -0.1384      0.015     -9.285      0.000      -0.168      -0.109
value       7.105e-05   1.04e-05      6.806      0.000    5.05e-05    9.16e-05
institutions   0.0257      0.003     10.220      0.000       0.021       0.031
sp500_yes      1.4808      0.129     11.499      0.000       1.228       1.734
==============================================================================
Omnibus:                       22.073   Durbin-Watson:                   1.980
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               55.245
Skew:                          -0.023   Prob(JB):                     1.01e-12
Kurtosis:                       4.498   Cond. No.                     2.87e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.87e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Problem - 1.4 Model Performance Evaluation and evaluate the model on different performance metrics

Training Performance

	RMSE	MAE	MAPE
0	1.127269	0.844911	26.95745

Test Performance

	RMSE	MAE	MAPE
0	1.030785	0.81383	23.978281

Problem 1.5)Now we'll check the rest of the assumptions on olsmod2.

Linearity of variables

Independence of error terms

Normality of error terms

No Heteroscedasticity

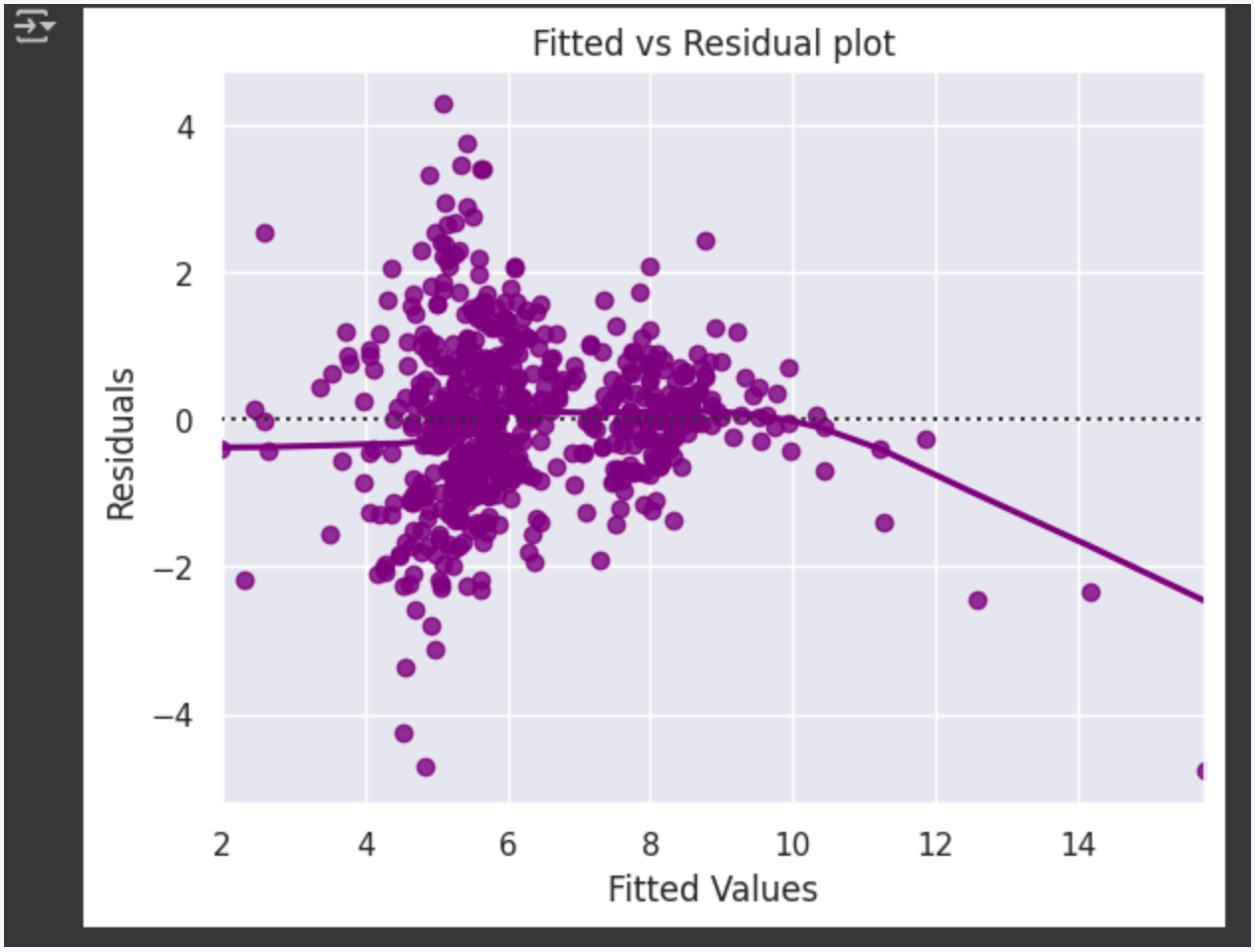
TEST FOR LINEARITY AND INDEPENDENCE

We will test for linearity and independence by making a plot of fitted values vs residuals and checking for patterns.

If there is no pattern, then we say the model is linear and residuals are independent.

Otherwise, the model is showing signs of non-linearity and residuals are not independent.

	Actual Values	Fitted Values	Residuals
652	5.772882	5.774955	-0.002073
366	6.340426	5.396227	0.944200
447	9.259054	9.546073	-0.287019
618	6.229126	5.588692	0.640434
610	5.455543	5.333378	0.122165

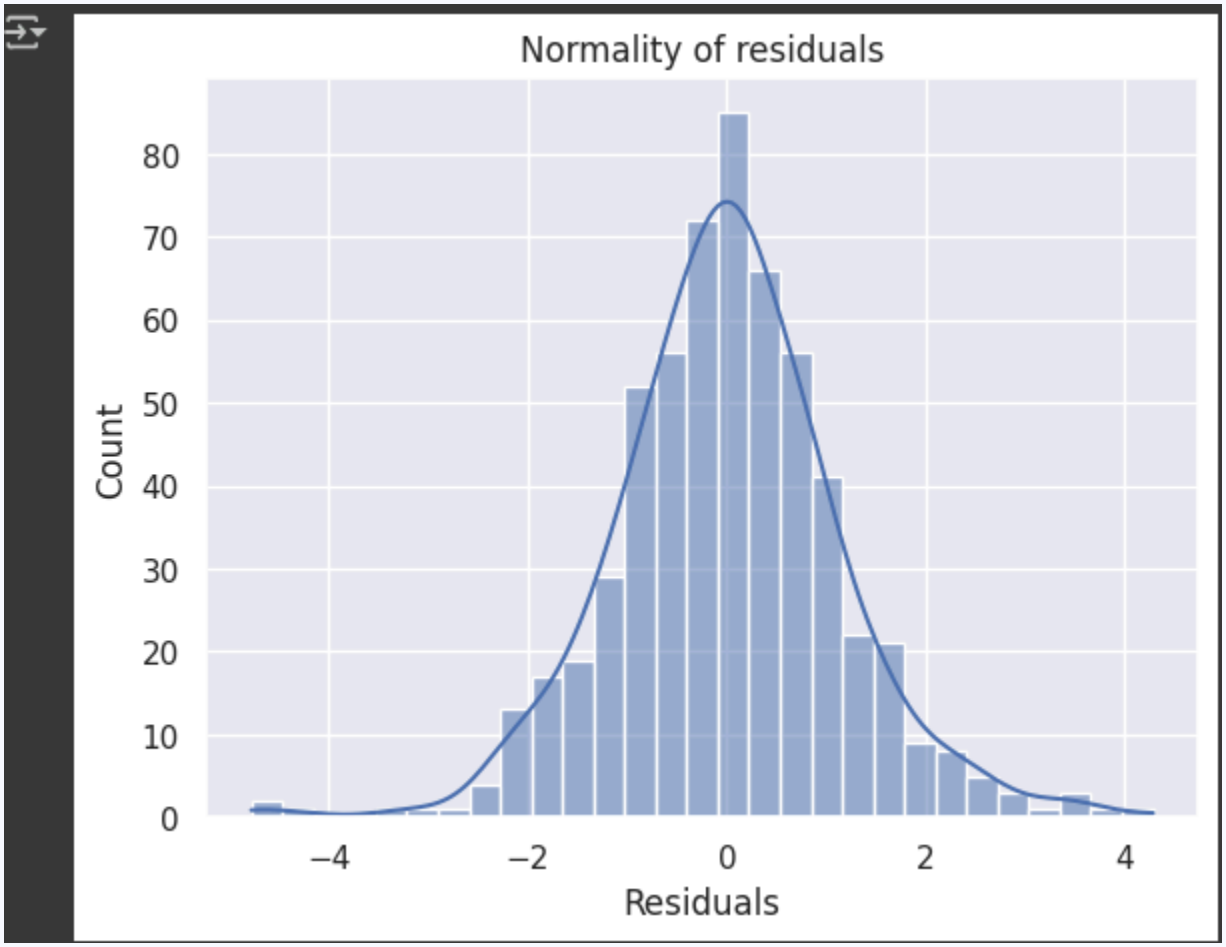


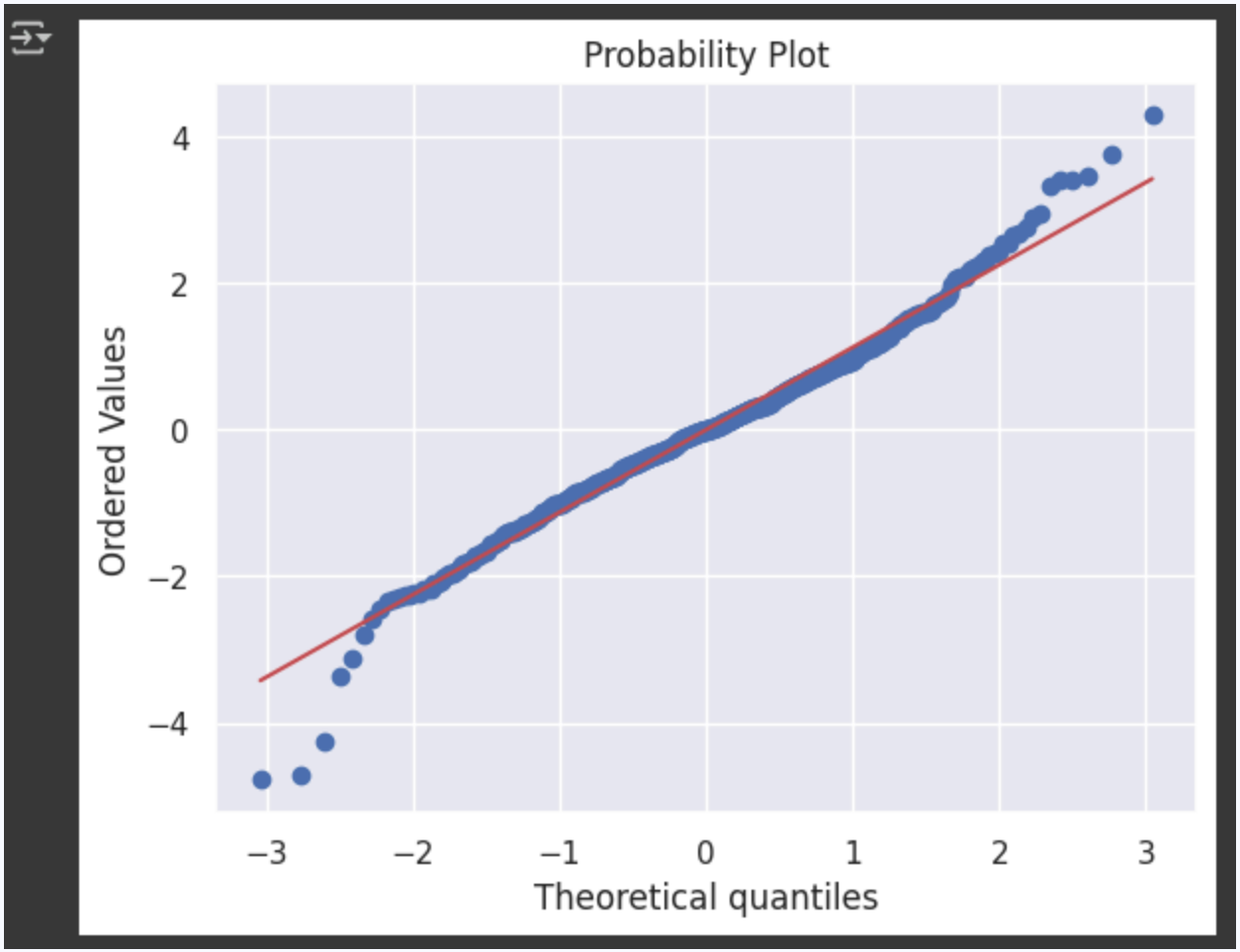
TEST FOR NORMALITY

We will test for normality by checking the distribution of residuals, by checking the Q-Q plot of residuals, and by using the Shapiro-Wilk test.

If the residuals follow a normal distribution, they will make a straight line plot, otherwise not.

If the p-value of the Shapiro-Wilk test is greater than 0.05, we can say the residuals are normally distributed.






```
ShapiroResult(statistic=0.9822825883697885, pvalue=1.4029046526110507e-06)
```

TEST FOR HOMOSCEDASTICITY

We will test for homoscedasticity by using the goldfeldquandt test.

If we get a p-value greater than 0.05, we can say that the residuals are homoscedastic. Otherwise, they are heteroscedastic.

```
[('F statistic', 1.0339466399164483), ('p-value', 0.3883948729469446)]
```

Final Model Summary

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  sales   R-squared:                       0.667
Model:                            OLS   Adj. R-squared:                  0.664
Method:                 Least Squares   F-statistic:                     233.5
Date:                Sun, 23 Feb 2025   Prob (F-statistic):          1.09e-136
Time:                        03:44:35   Log-Likelihood:                -909.96
No. Observations:                 590   AIC:                             1832.
Df Residuals:                     584   BIC:                             1858.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          4.7867      0.115     41.533      0.000       4.560       5.013
employment     0.0053      0.001      3.947      0.000       0.003       0.008
tobinq        -0.1406      0.015     -9.555      0.000      -0.170      -0.112
value       7.475e-05   8.81e-06      8.488      0.000    5.74e-05     9.2e-05
institutions   0.0251      0.002     10.122      0.000       0.020       0.030
sp500_yes      1.4786      0.129     11.487      0.000       1.226       1.731
==============================================================================
Omnibus:                       25.118   Durbin-Watson:                   1.983
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               68.680
Skew:                          -0.020   Prob(JB):                     1.22e-15
Kurtosis:                       4.671   Cond. No.                     2.28e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.28e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Training Performance

	RMSE	MAE	MAPE
0	1.131306	0.843946	26.941502

Test Performance

	RMSE	MAE	MAPE
0	1.030857	0.812045	23.962577

1.6)Actionable Insights and Recommendations:

The investment criteria for any new investor is mainly based on the capital invested in the company by the promoters and investors are vying on the firms where the capital investment is good as also reflecting in the scatter plot. To generate capital the company should have the combination of the following attributes such as value, employment, sales and patents. The highest contributing attribute is employment followed by patents. When the Employment increases by 1 Unit the Sales increase by 80.33 units, by keeping all the predictors constant, When the Capital increases by 1 Unit the Sales increase by 0.42 units by keeping all the predictors constant.