**Firm Level Data Analysis**

Name: Saurabh Dharmadhikari

PGP-DSBA Online January’ 21

Date: 18/10/2022

Table of Contents:

Executive Summary……………………...................................................................................................................................3

Introduction ...................................................................................................................................................................3

Data Description ............................................................................................................................................................3

Sample of the dataset ...................................................................................................................................................4

Exploratory Data Analysis..............................................................................................................................................4

Let us check the types of variables in the data frame. .................................................................................................4

Check for missing values in the dataset .......................................................................................................................5

Question 1.1 Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, data types, shape, EDA). Perform Univariate and Bivariate Analysis. .................................................................................7

Question 1.2 Impute null values if present? Do you think scaling is necessary in this case? ......................................15

Question 1.3 Encode the data (having string values) for Modelling. Data Split: Split the data into test and train (30:70). Apply Linear regression. Performance Metrics: Check the performance of Predictions on Train and Test sets using R-square, RMSE. …………………………………………………………………………………………………………………………...................15

Question 1.4 Inference: Based on these predictions, what are the business insights and recommendations. ..........16

Conclusion & Insights ..................................................................................................................................................16

THE END!......................................................................................................................................................................16

Executive Summary

You are a part of an investment firm and your work is to do research about these 759 firms. You

are provided with the dataset containing the sales and other attributes of these 759 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.

Introduction

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset using central tendency and other parameters. The data consists of 759 entries of different firms. We are provided with different attributes of these 759 firms mentioned below in the Data Description. Our mission is to build a model that can predict sales for these firms using a linear regression model.

Data Description

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 institutions.

Sample of the data set

| **Unnamed: 0** | | **sales** | | **capital** | | **patents** | | **randd** | | **employment** | | **sp500** | | **tobinq** | | **value** | | **institutions** | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 0 | | 826.995050 | | 161.603986 | | 10 | | 382.078247 | | 2.306000 | | no | | 11.049511 | | 1625.453755 | | 80.27 |
| **1** | 1 | | 407.753973 | | 122.101012 | | 2 | | 0.000000 | | 1.860000 | | no | | 0.844187 | | 243.117082 | | 59.02 |
| **2** | 2 | | 8407.845588 | | 6221.144614 | | 138 | | 3296.700439 | | 49.659005 | | yes | | 5.205257 | | 25865.233800 | | 47.70 |
| **3** | 3 | | 451.000010 | | 266.899987 | | 1 | | 83.540161 | | 3.071000 | | no | | 0.305221 | | 63.024630 | | 26.88 |
| **4** | 4 | | 174.927981 | | 140.124004 | | 2 | | 14.233637 | | 1.947000 | | no | | 1.063300 | | 67.406408 | | 49.46 |

Exploratory data analysis

Checking types of variables in the data set

RangeIndex: 759 entries, 0 to 758

Data columns (total 10 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Unnamed: 0 759 non-null int64

1 sales 759 non-null float64

2 capital 759 non-null float64

3 patents 759 non-null int64

4 randd 759 non-null float64

5 employment 759 non-null float64

6 sp500 759 non-null object

7 tobinq 738 non-null float64

8 value 759 non-null float64

9 institutions 759 non-null float64

dtypes: float64(7), int64(2), object(1)

We have a total of 759 rows and 9 columns. All these columns or features are either float or integer type except sp500.

Check for null values

sales 0

capital 0

patents 0

randd 0

employment 0

sp500 0

tobinq 21

value 0

institutions 0

dtype: int64

There are 21 null values in tobinq column. All the other features do not have any null value and thus we will have to replace the missing values in tobinq column. We will be using the mean value of the tobinq column to fill in the missing values. So this method will let us keep the rest of the data for the prediction and also mean value will not effect much standard deviation of tobinq feature.

|  | **sales** | **capital** | **patents** | **randd** | **employment** | **sp500** | **tobinq** | **value** | **institutions** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **14** | 26.309996 | 5.072000 | 2 | 9.321929 | 0.215000 | no | 2.79491 | 15.925752 | 24.99 |
| **22** | 26.478002 | 15.928000 | 0 | 3.829102 | 0.450000 | no | 2.79491 | 44.354573 | 42.66 |
| **38** | 50.858007 | 23.438000 | 1 | 1.159340 | 0.616000 | no | 2.79491 | 48.779925 | 15.96 |
| **39** | 11.182001 | 8.923000 | 1 | 6.957602 | 0.120000 | no | 2.79491 | 17.094293 | 10.76 |
| **45** | 666.319036 | 591.956894 | 44 | 409.005371 | 5.300000 | no | 2.79491 | 1163.897196 | 58.50 |
| **60** | 172.800995 | 43.720996 | 13 | 16.280581 | 1.185000 | no | 2.79491 | 160.127315 | 15.04 |
| **103** | 242.970015 | 114.746010 | 0 | 46.152855 | 1.800000 | no | 2.79491 | 269.288483 | 36.62 |
| **120** | 611.270017 | 318.793948 | 0 | 63.899895 | 4.440000 | no | 2.79491 | 541.281046 | 54.14 |
| **123** | 667.923892 | 279.040040 | 5 | 253.020279 | 4.060000 | no | 2.79491 | 1028.687964 | 36.23 |
| **263** | 177.270983 | 55.697997 | 0 | 7.769489 | 2.120000 | no | 2.79491 | 116.445783 | 46.53 |
| **288** | 211.847044 | 104.624999 | 6 | 123.462677 | 1.820000 | no | 2.79491 | 183.842582 | 33.88 |
| **312** | 322.601935 | 129.875982 | 25 | 63.278076 | 2.080000 | no | 2.79491 | 296.758388 | 71.82 |
| **390** | 722.400097 | 336.819017 | 0 | 25.273182 | 7.729999 | no | 2.79491 | 658.528551 | 23.14 |
| **435** | 1525.500220 | 1089.106119 | 30 | 425.467957 | 4.613000 | no | 2.79491 | 2416.172318 | 65.08 |
| **461** | 123.586022 | 41.197000 | 3 | 26.305023 | 0.958000 | no | 2.79491 | 156.287797 | 35.63 |
| **525** | 190.185998 | 85.633012 | 3 | 48.030548 | 1.123000 | no | 2.79491 | 326.913854 | 30.50 |
| **542** | 351.779041 | 288.367054 | 11 | 73.962776 | 2.012000 | no | 2.79491 | 471.130289 | 39.57 |
| **569** | 2131.053990 | 158.675033 | 28 | 248.000992 | 2.984000 | no | 2.79491 | 776.607351 | 73.42 |
| **636** | 722.120099 | 292.663017 | 5 | 45.872894 | 5.769000 | no | 2.79491 | 632.589951 | 72.49 |
| **702** | 34.568002 | 15.081000 | 2 | 2.741614 | 0.318000 | no | 2.79491 | 33.721811 | 8.18 |
| **755** | 171.821025 | 73.666008 | 1 | 0.037735 | 1.684000 | no | 2.79491 | 228.475701 | 46.41 |

We can see above that all the missing values are now filled with the mean value of tobinq.

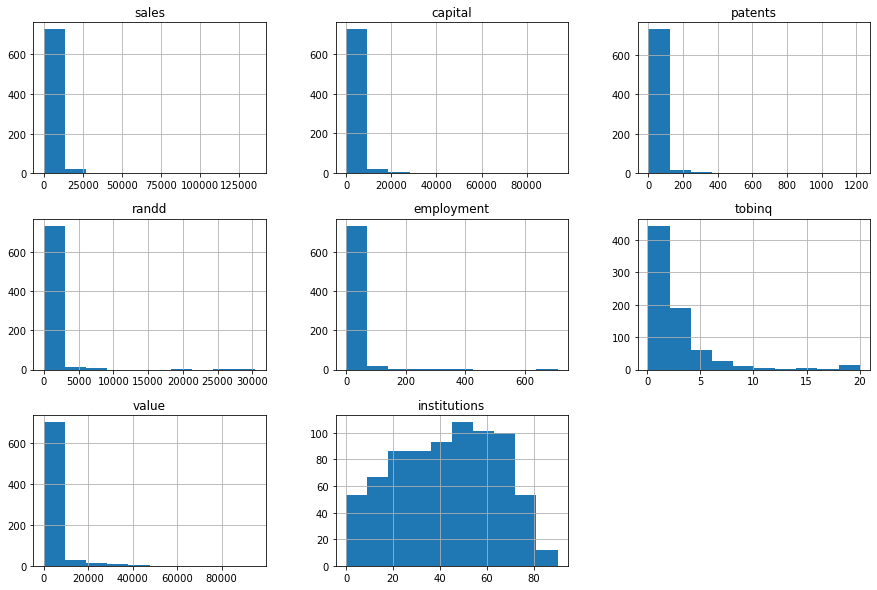
Question 1.1

* 1. Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, data types, shape, EDA). Perform Univariate and Bivariate Analysis.

|  | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **sales** | 759.0 | 2689.705158 | 8722.060124 | 0.138000 | 122.920000 | 448.577082 | 1822.547366 | 135696.788200 |
| **capital** | 759.0 | 1977.747498 | 6466.704896 | 0.057000 | 52.650501 | 202.179023 | 1075.790020 | 93625.200560 |
| **patents** | 759.0 | 25.831357 | 97.259577 | 0.000000 | 1.000000 | 3.000000 | 11.500000 | 1220.000000 |
| **randd** | 759.0 | 439.938074 | 2007.397588 | 0.000000 | 4.628262 | 36.864136 | 143.253403 | 30425.255860 |
| **employment** | 759.0 | 14.164519 | 43.321443 | 0.006000 | 0.927500 | 2.924000 | 10.050001 | 710.799925 |
| **tobinq** | 759.0 | 2.794910 | 3.319629 | 0.119001 | 1.036000 | 1.741800 | 3.082979 | 20.000000 |
| **value** | 759.0 | 2732.734750 | 7071.072362 | 1.971053 | 103.593946 | 410.793529 | 2054.160386 | 95191.591160 |
| **institutions** | 759.0 | 43.020540 | 21.685586 | 0.000000 | 25.395000 | 44.110000 | 60.510000 | 90.150000 |

In the above table we can see the counts, mean, standard deviation, minimum values and perncent of values under the percentage and maximum values of each column.

Mean sales value of a firm is about 2689 with a minimum of 0.13800 and a maximum of 135696. It has a standard deviation of 8722.



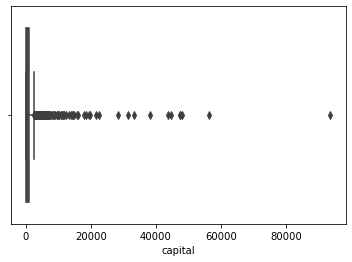
In the above histogram we observe that the features do not have a normal distribution but it is rather a skewed data.

Let us check for outliers in our data set.

Sales



Capital



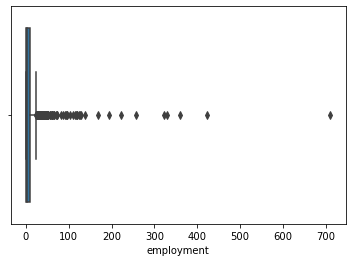
Patents



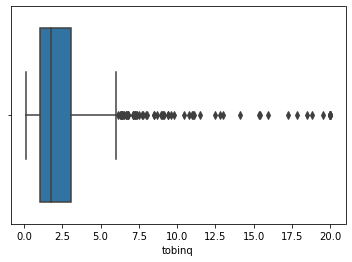
R & D



Employment



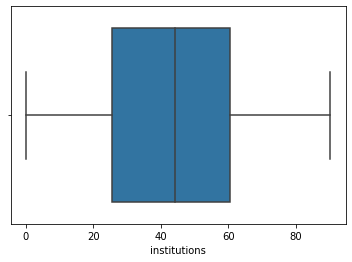
Tobinq



Value

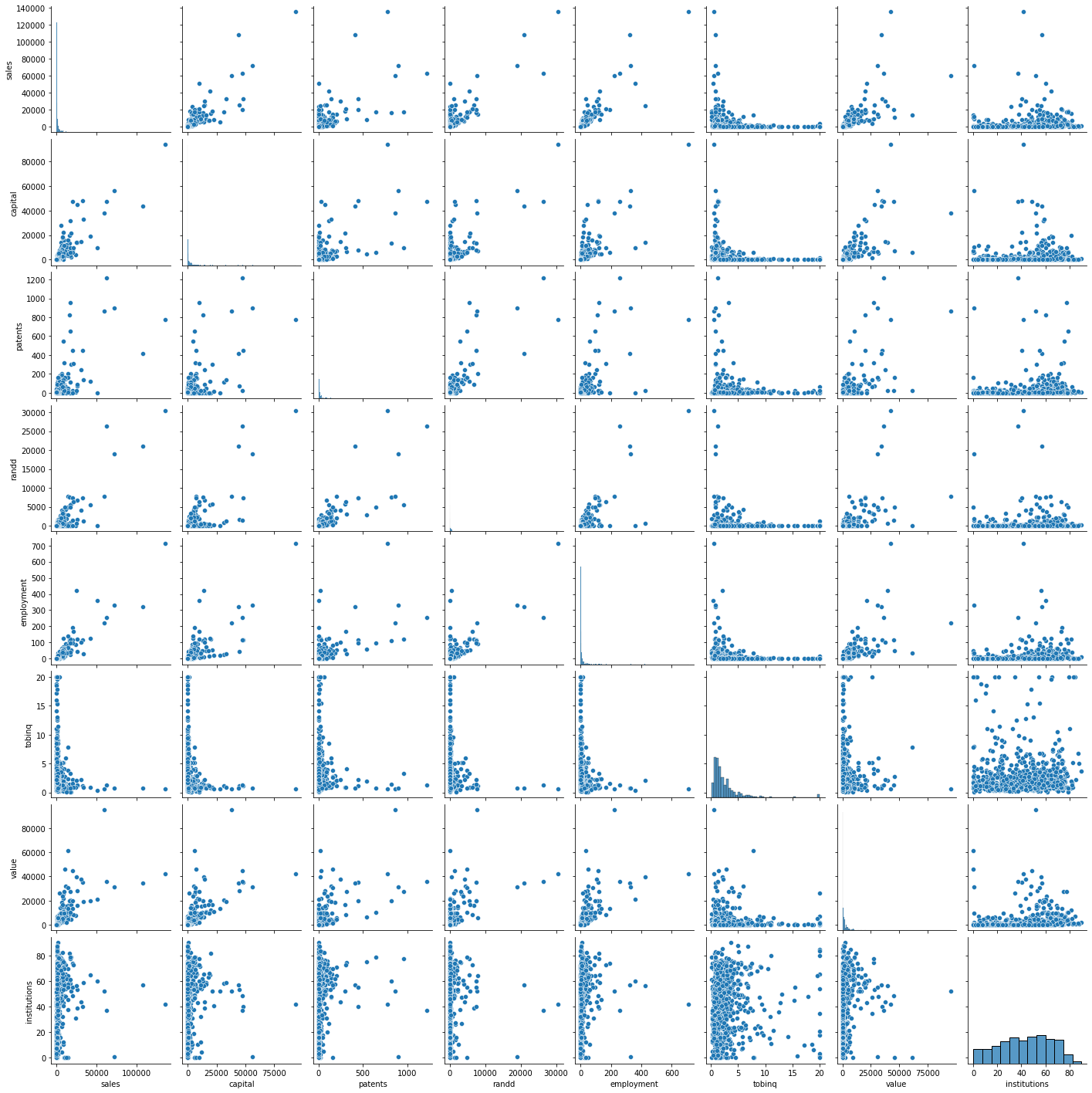
****

Institutions

****

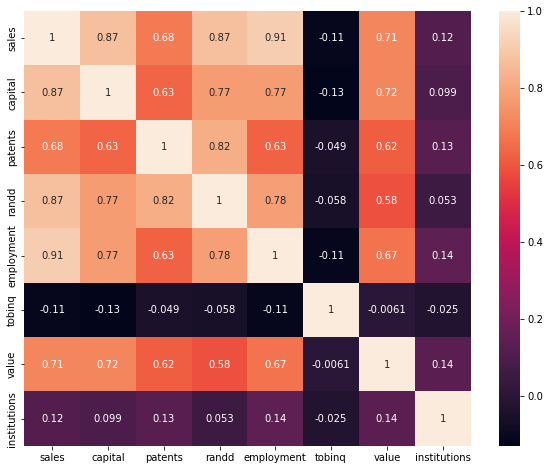
From the above box plot we can see that there are many outliers in our features. We particularly would not want to treat them as they are not one of case that could lead our predictions to stray but rather with so many outliers, we may be able to use this data to get better accuracy for our predictions. Hence, we will not be treating the outliers in this case.

Let us have a look at the pair plot for this data set and see.



We see that most features have a linear relationship with sales.

Let us also have a look at heatmap



As we studied earlier, we can see that there is a linear relationship between features and sales. But institutions do not have a strong linear relationship with sales.

We will still go ahead with our prediction model because we also observe from the scatter plot that probably we do not have enough data points to establish an effective relationship. If we had more data maybe we could establish a relationship between sales and institutions.

* 1. Impute null values if present? Do you think scaling is necessary in this case?

We have already imputed null values shown in page number 5.

Data is already in a comparative unit thus there is no need to scale the data.

1.3) Encode the data (having string values) for Modelling. Data Split: Split the data into test and

train (30:70). Apply Linear regression. Performance Metrics: Check the performance of

Predictions on Train and Test sets using R-square, RMSE.

We type of data for sp500 from object to integer, having values of 0s and 1s. Where 0 is no and 1 is yes.

This is how our sample datasheet looks now.

|  | **sales** | **capital** | **patents** | **randd** | **employment** | **sp500** | **tobinq** | **value** | **institutions** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 826.995050 | 161.603986 | 10 | 382.078247 | 2.306000 | 0 | 11.049511 | 1625.453755 | 80.27 |
| **1** | 407.753973 | 122.101012 | 2 | 0.000000 | 1.860000 | 0 | 0.844187 | 243.117082 | 59.02 |
| **2** | 8407.845588 | 6221.144614 | 138 | 3296.700439 | 49.659005 | 1 | 5.205257 | 25865.233800 | 47.70 |
| **3** | 451.000010 | 266.899987 | 1 | 83.540161 | 3.071000 | 0 | 0.305221 | 63.024630 | 26.88 |
| **4** | 174.927981 | 140.124004 | 2 | 14.233637 | 1.947000 | 0 | 1.063300 | 67.406408 | 49.46 |

X

Linear regression model was applied to get predictions. Thus, we have results now.

Let us check performance with R-square.

R-square score for train set = 0.9363769281190529

R-square score for test set = 0.8927713978886768

Let us check performance with RMSE.

RMSE score for train set = 2164.5197630396633

RMSE score for test set = 2953.5253885719326

1.4) Inference: Based on these predictions, what are the business insights and

recommendations.

Coefficients and canstant

const 53.694023

capital 0.414244

patents -5.044542

randd 1.026063

employment 83.958224

sp500 -102.006251

tobinq -31.296414

value 0.126675

institutions 1.062729

dtype: float64

Summary

OLS Regression Results

==============================================================================

Dep. Variable: sales R-squared: 0.936

Model: OLS Adj. R-squared: 0.935

Method: Least Squares F-statistic: 960.3

Date: Wed, 19 Oct 2022 Prob (F-statistic): 1.38e-306

Time: 01:23:09 Log-Likelihood: -4831.5

No. Observations: 531 AIC: 9681.

Df Residuals: 522 BIC: 9719.

Df Model: 8

Covariance Type: nonrobust

================================================================================

coef std err t P>|t| [0.025 0.975]

--------------------------------------------------------------------------------

const 53.6940 233.554 0.230 0.818 -405.128 512.516

capital 0.4142 0.027 15.563 0.000 0.362 0.467

patents -5.0445 2.407 -2.096 0.037 -9.773 -0.316

randd 1.0261 0.127 8.052 0.000 0.776 1.276

employment 83.9582 3.629 23.135 0.000 76.829 91.087

sp500 -102.0063 267.962 -0.381 0.704 -628.423 424.410

tobinq -31.2964 30.297 -1.033 0.302 -90.816 28.223

value 0.1267 0.022 5.885 0.000 0.084 0.169

institutions 1.0627 4.964 0.214 0.831 -8.690 10.816

==============================================================================

Omnibus: 231.611 Durbin-Watson: 1.932

Prob(Omnibus): 0.000 Jarque-Bera (JB): 31505.675

Skew: 0.809 Prob(JB): 0.00

Kurtosis: 40.701 Cond. No. 2.90e+04

==============================================================================

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The condition number is large, 2.9e+04. This might indicate that there are

strong multicollinearity or other numerical problems.

R-square is about 89% for our test results. It means 89% of the variance is explained by the model we built.

We can see that coefficient of sp500 is in negative. So it has as a very decremental effect on sales.

Toinq has a decremental effect on sales. Every time toinq increases by 1 unit sales decrease by 31.

Also, we see employment has an incremental effect on a firm’s sales.

Value and capital has the least effect on sales of a firm.

Other features also have effect on sales but lesser than compared to sp500, toinq and employment.