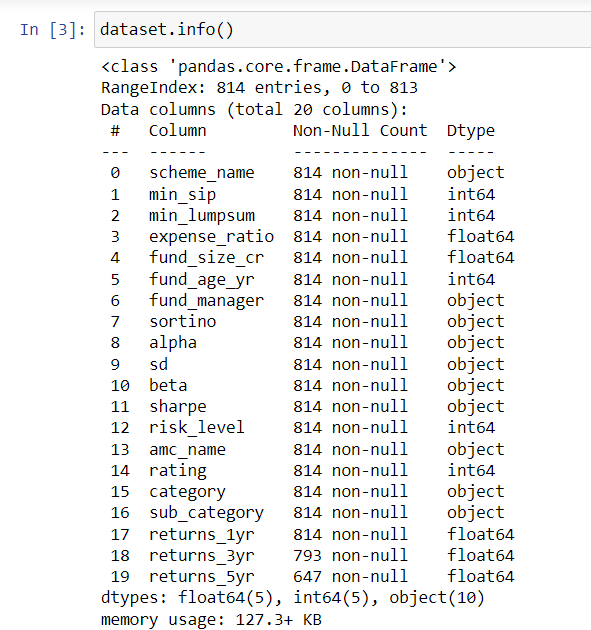
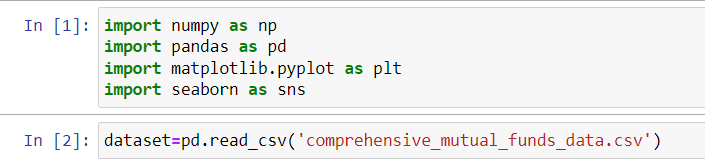
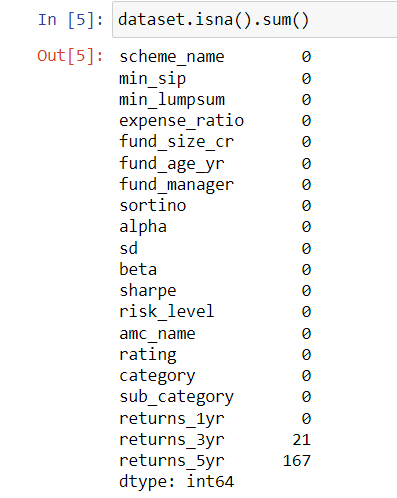
**Description Of Data:**

We have taken the dataset from kaggle. The dataset comprises of 20 columns and 814 rows. In the dataset there are 5 qualitative columns and 15 quantitative columns. Out of 15 quantitative columns, 9 of the columns have discrete quantitative values and 6 of the columns have continuous quantitative values. Following is the description of the required fields in the project.

1. **Scheme Name**: Name of the mutual fund scheme.
2. **Expense ratio**: It is a ratio of total annual expenses to average assets under management and expressed as a percentage. Higher expense ratios mean that a larger portion of the fund's returns will be used to cover operating expenses, potentially reducing the returns that investors receive. Lower expense ratios, on the other hand, leave more of the fund's returns for investors.
3. **Sortino** : It is a ratio of excess return to downside deviation. A higher Sortino ratio is generally preferred because it implies that the investment has provided a better return for the level of downside risk taken.
4. **Alpha:** It is the difference between actual return and expected return. If the alpha is positive, it suggests that the investment has outperformed its expected return, while a negative alpha indicates underperformance.
5. **Standard deviation**: A standard deviation is a number that can be used to show how much the returns of a mutual fund scheme are likely to deviate from its average annual returns. A higher standard deviation suggests higher risk because it means the fund's returns have been more variable. Conversely, a lower standard deviation suggests lower risk because the returns have been relatively stable and consistent.
6. **Beta:** It is a ratio of covariance between fund returns and benchmark returns to the variance of benchmark returns. A beta greater than 1 indicates that the mutual fund is more volatile or riskier than the benchmark. This means the fund tends to have larger price swings, both up and down, compared to the broader market.A beta less than 1 suggests that the mutual fund is less volatile or less risky than the benchmark. It means the fund tends to have smaller price fluctuations than the broader market.A beta of exactly 1 means the mutual fund's returns move in line with the benchmark. It indicates that the fund exhibits the same level of risk as the market.
7. **Sharpe**: It is the ratio of excess return to standard deviation. A higher Sharpe ratio indicates that the investment has provided a better risk-adjusted return because it has generated more excess return per unit of risk.
8. **Return\_1yr (%)**: The return percentage of the mutual fund scheme over 1 year.
9. **Return\_3yr (%)**: The return percentage of the mutual fund scheme over 3 years.
10. **Return\_5yr (%):** The return percentage of the mutual fund scheme over 5 years.

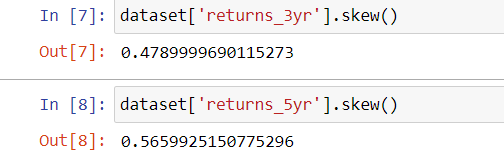
**Exploratory Data Analysis:**

1. **First we will see what our data comprises of:**
2. **Now we will find if there are any null values:**

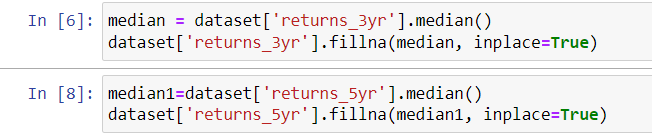
****

So we have null values in two columns namely ‘retruns\_3yr’ and ‘return\_5yr’.

Now we will see if our data is skewed or not and then accordingly fill the null values with mean or median.



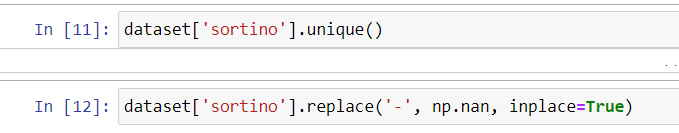
The above image shows us that our data is skewed, hence we will fill null values with median.

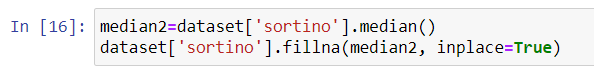


1. **Now we will remove certain characters like ‘-’ from the data:**

We came to know there are some unwanted characters in the dataset by using the ‘unique()’ function in python.

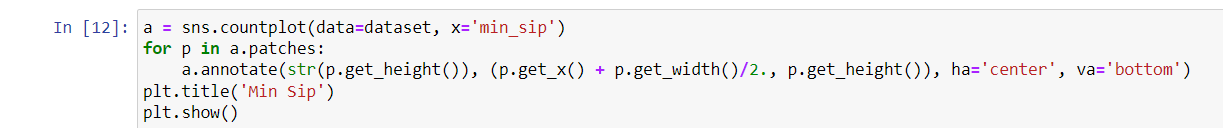
So we replaced these characters with null values and later filled the null values with the median value of the respected columns.

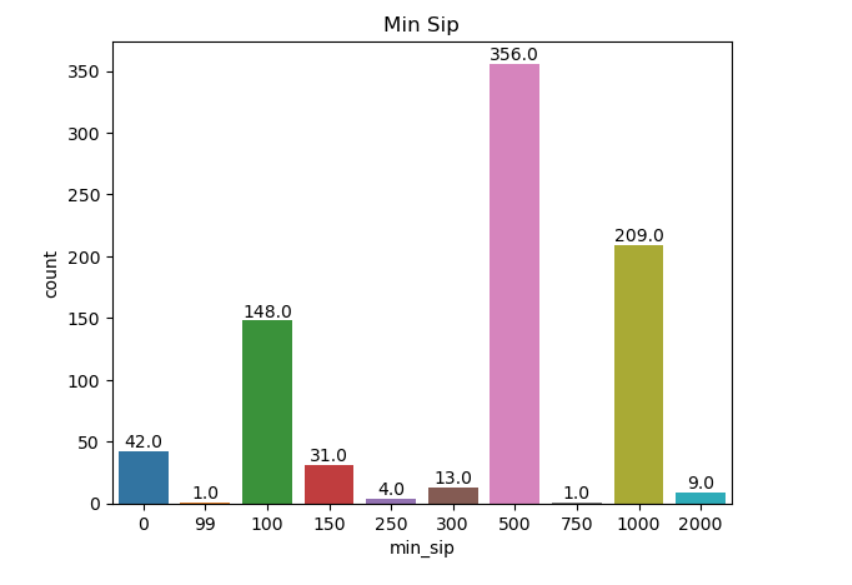


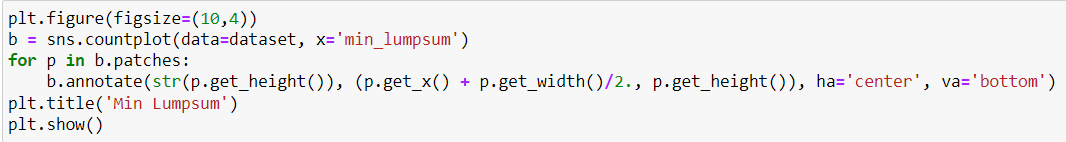


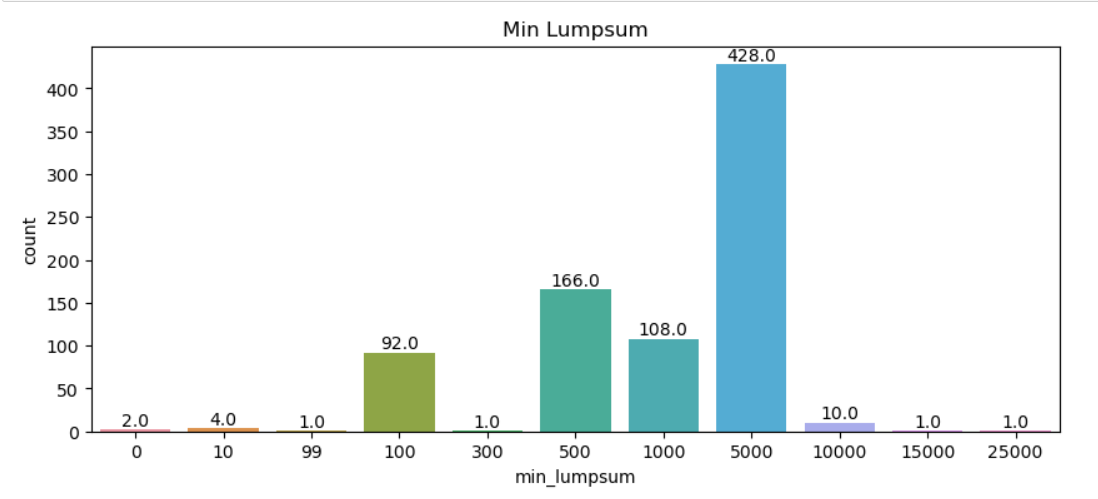
Similar steps were performed on the columns ‘alpha’, ‘beta’, ‘sd’, ‘sharpe’.

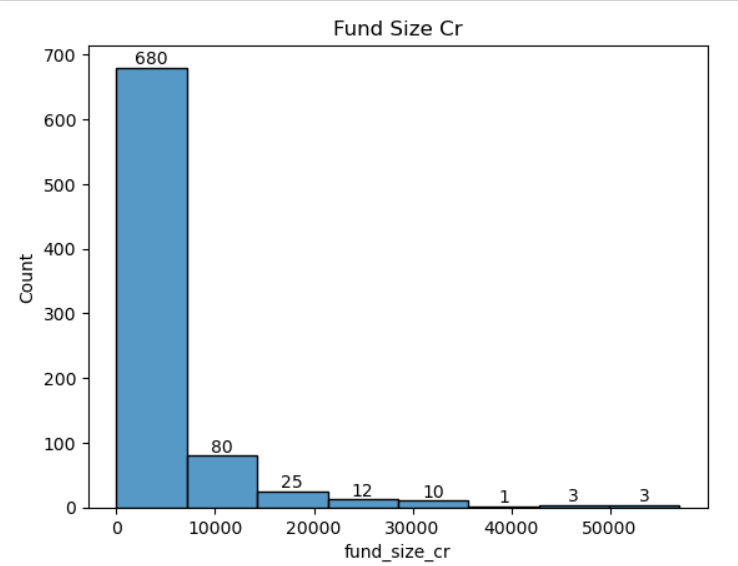
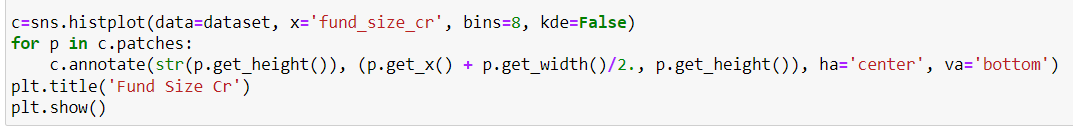
1. **Now we will try to gain some insights about our dataset by plotting various plots.**

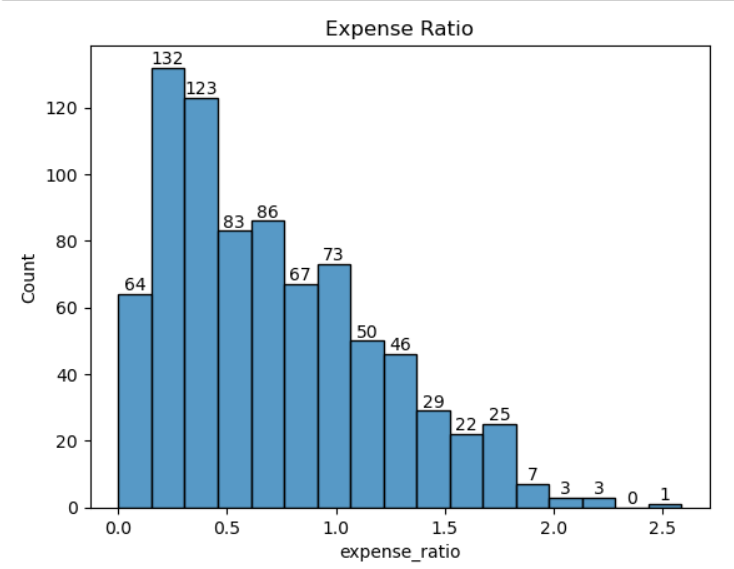
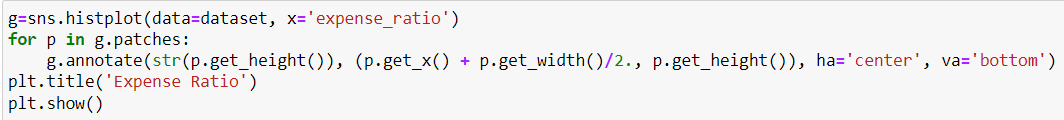
****

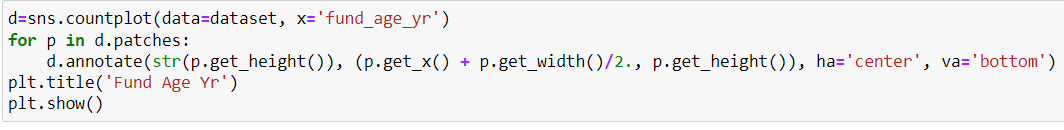
****

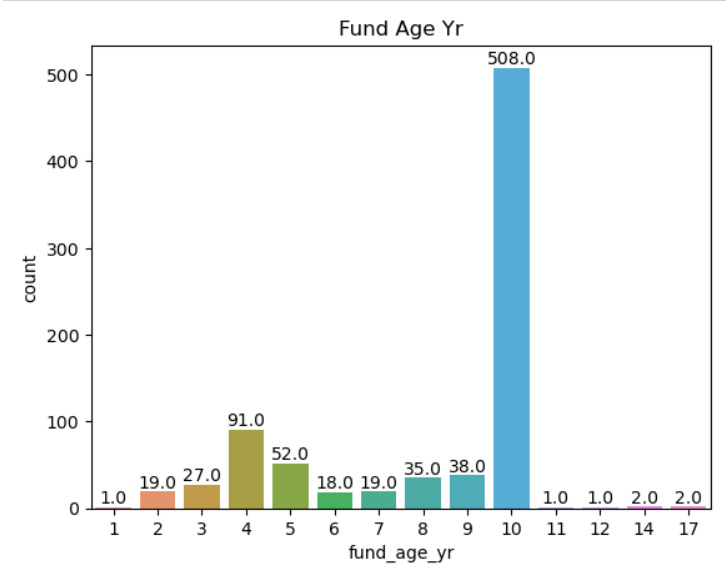
****

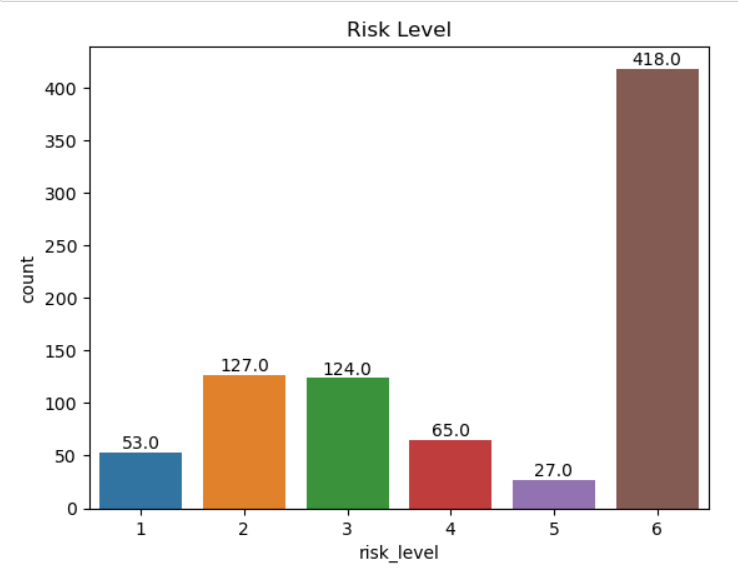
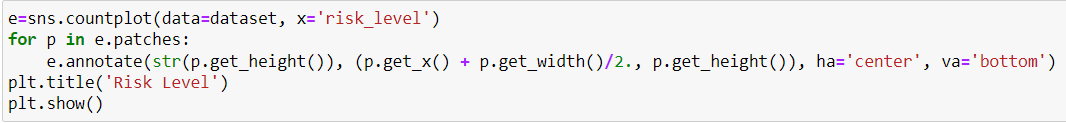
****

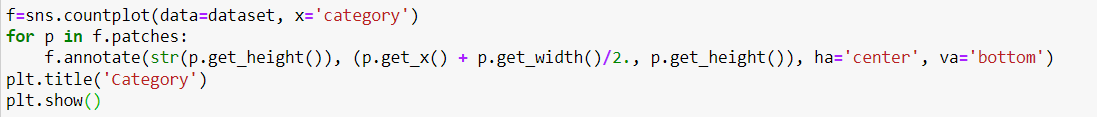
****

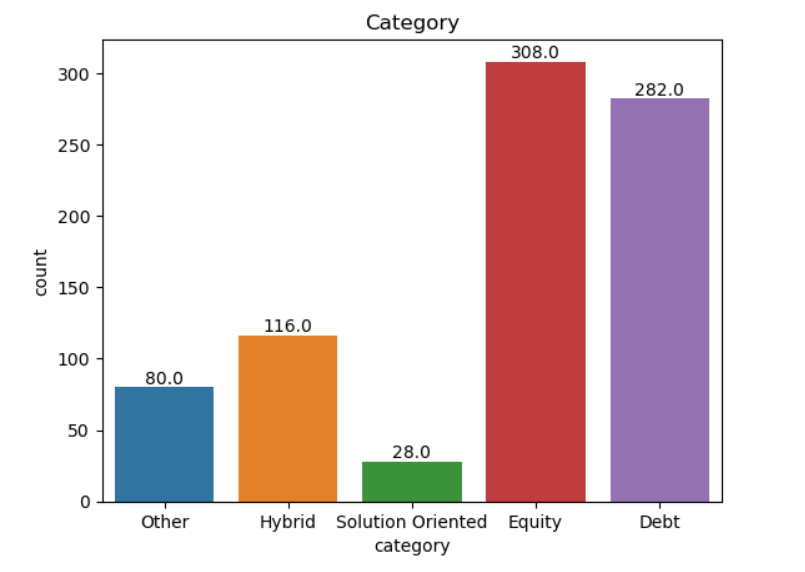


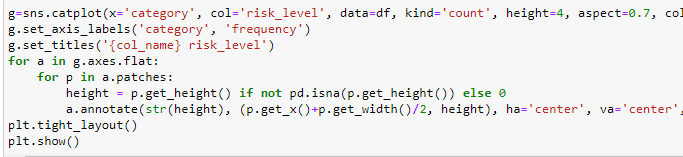


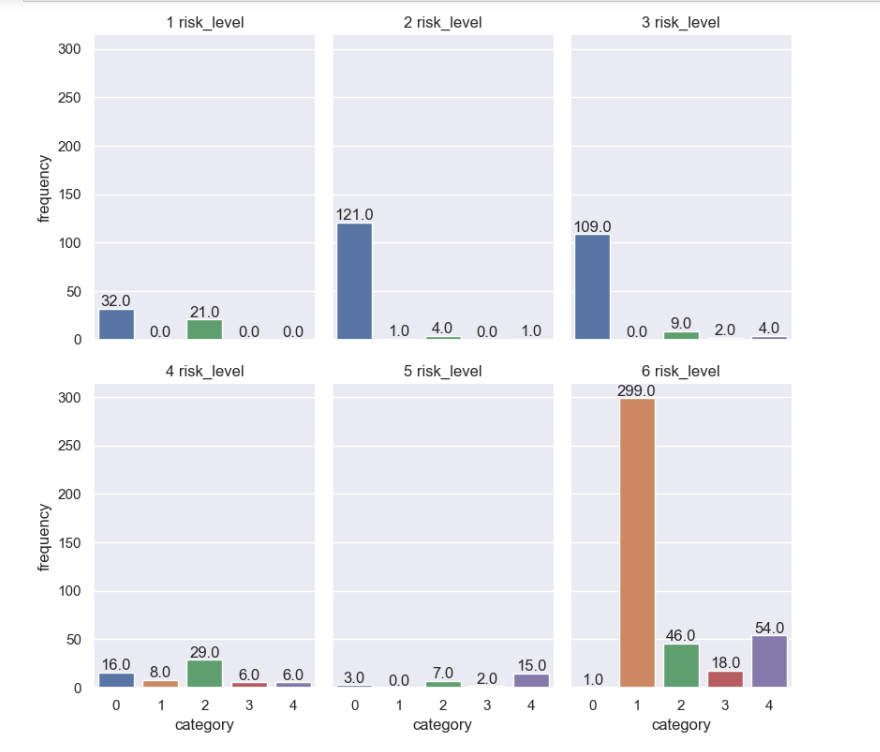


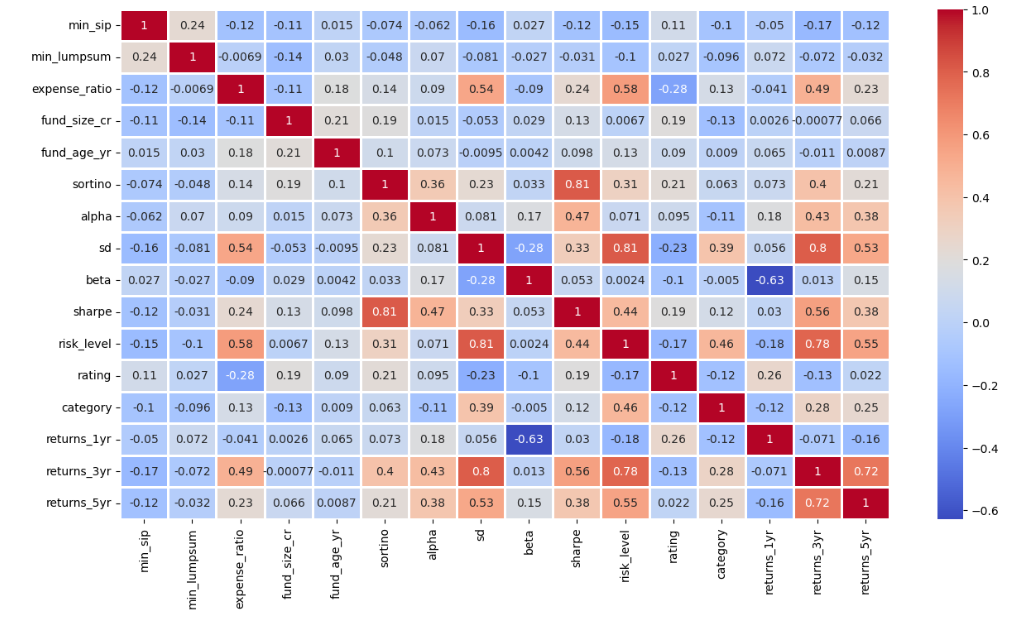




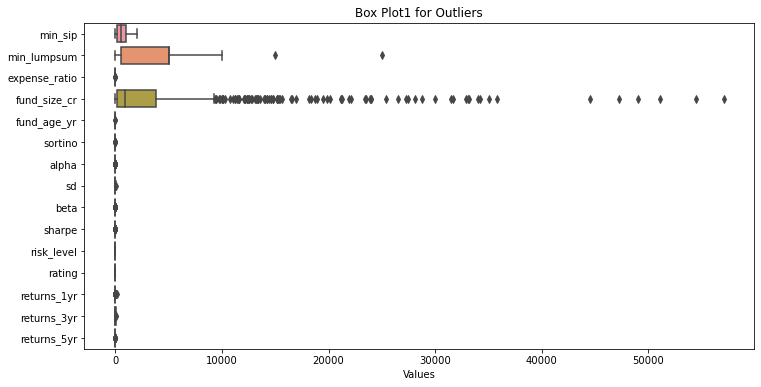




Note 0, 1, 2, 3, 4 represents categories Debt, Equity, Hybrid, Solution Oriented, Other respectively.

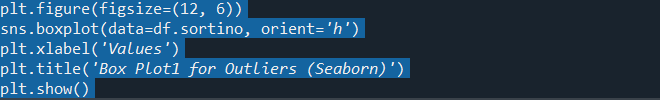


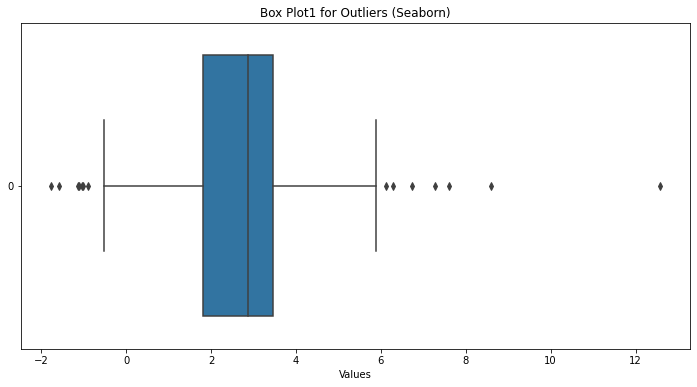
1. **Now we will see if we have any outliers**

****

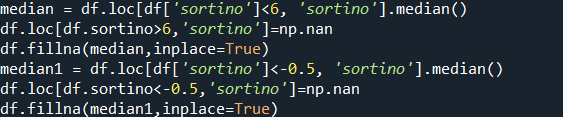
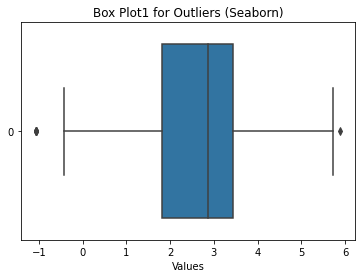
As you can see we have outliers, so we will take one column at a t time and treat the outliers by taking median

As you can see we have taken a single column sortino and finding its individual outlier

****



Treating the outliers with median

****

As you can see the outliers have been treated and filled with median

Similar steps were performed on the columns ‘alpha’, ‘beta’, ‘sd’, ‘sharpe’.

**Conclusion of EDA:**

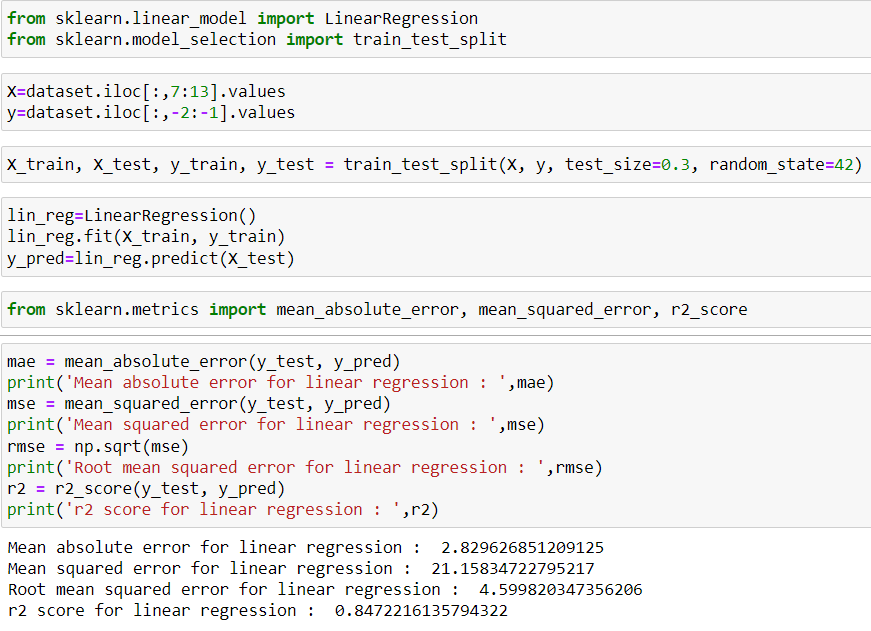
1. 44% of the mutual funds in the dataset have min\_sip Rs. 500
2. 59% of the mutual funds in the dataset have min\_lumpsum Rs. 5000.
3. 62% of the mutual funds in the dataset have fund\_age\_yr 10 years.
4. 51% of the mutual funds in the dataset fall under the category of very high risk (ie having risk level of 6).
5. Mutual funds of the category Equity are in majority in our dataset.
6. One of the plots shows that mutual funds having risk level 1 are mostly belong to the category of Debt mutual funds, mutual funds having risk level 2 are mostly belong to the category of Debt mutual funds, mutual funds having risk level 3 are mostly belong to the category of Debt mutual funds, mutual funds having risk level 4 are mostly belong to the category of Hybrid mutual funds, mutual funds having risk level 5 are mostly belong to the category of Other mutual funds and mutual funds having risk level 6 are mostly belong to the category of Equity mutual funds.
7. From the heat map we can see that there is a considerable correlation between the variables ‘returns\_3yr’,’expense\_ratio’, ‘sortino’, ‘alpha’, ‘beta’, ‘sd’, ‘sharpe’, ‘risk\_level’. So we will build our model using these variables.
8. Although we had detected outliers in various columns of our dataset but as they were very few we hadn’t removed them from the dataset as that would have been not good for our model building.

**Model Building:**

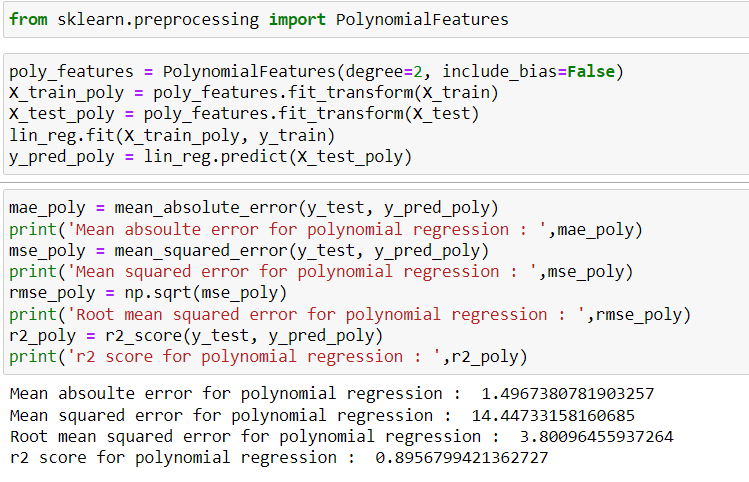
While building the model we considered 7 independent variables namely ‘expense\_raio’, ‘sortino’, ‘alpha’, ‘beta’, ‘sd’, ‘sharpe’, ‘risk level’ one target variable i.e. 'returns’.

We applied multiple models and found polynomial regression model with the lowest error scores.

**Linear Regression:**

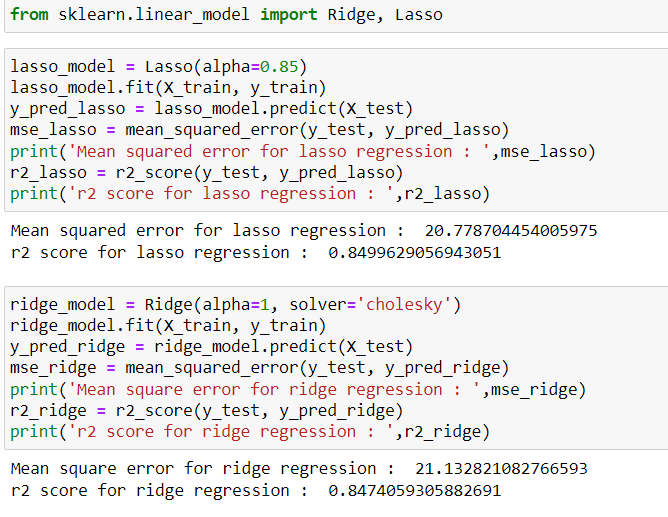


**Polynomial Regression:**

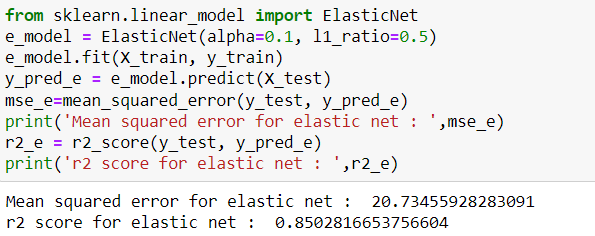
****

**Note:** While performing polynomial regression we changed the degree of the polynomial to 3 & 4 and observed that the error was increasing. So polynomial of degree 2 is best fitted in this problem.

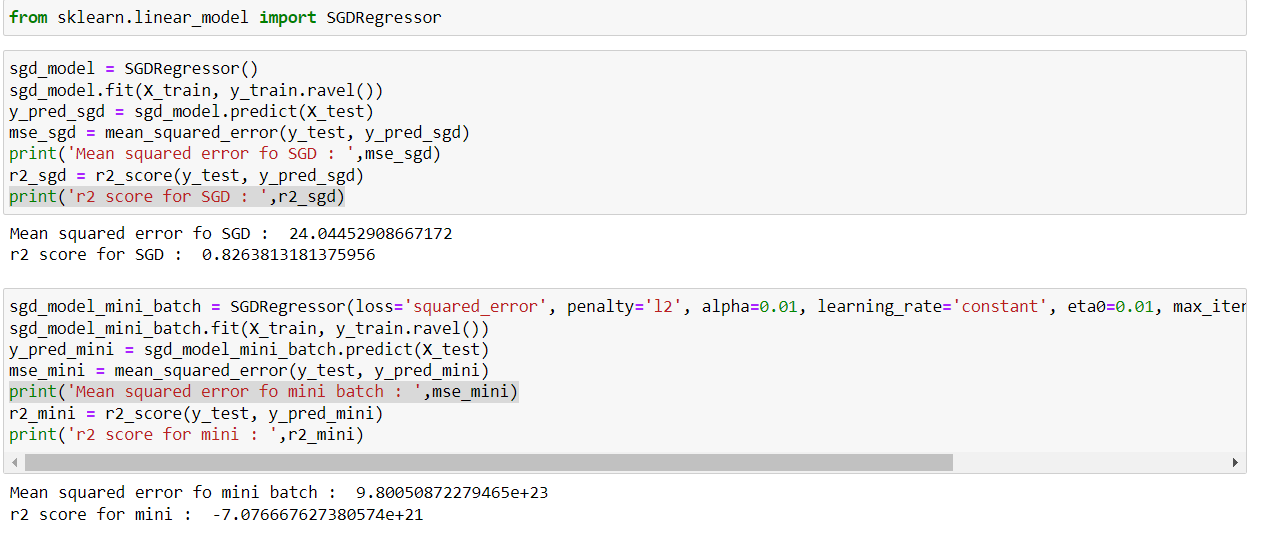
**Ridge and Lasso Regression:**

****

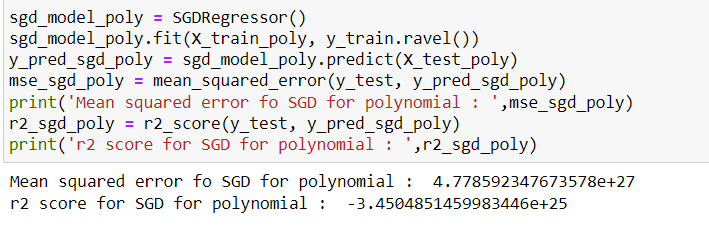
**Elastic Net Regression:**

****

**Gradient Descent on Linear Model:**



**Gradient Descent on Polynomial Model:**



After selecting polynomial regression as our model we created a dataset named ‘dataset1’ and entered 15 mutual funds in the dataset. After that using polynomial regression we predicted the returns of the mutual funds and sorted the ‘dataset1’ in descending order. This will give us the top 5 mutual funds with maximum returns.