Technical Assessment (all areas) – Data Scientist - 006191

Problem statement –

Identify the firms to which scarce resources need to be allocated.

Data –

5 years data for various firms were presented in two excel sheets.

EDA –

The distribution of data indictes the data is having kurtosis and is highly skewed. All of the columns contains outliers as seen in the box plots.
An eg. Box plot for Total assets variable is shown in fig.1

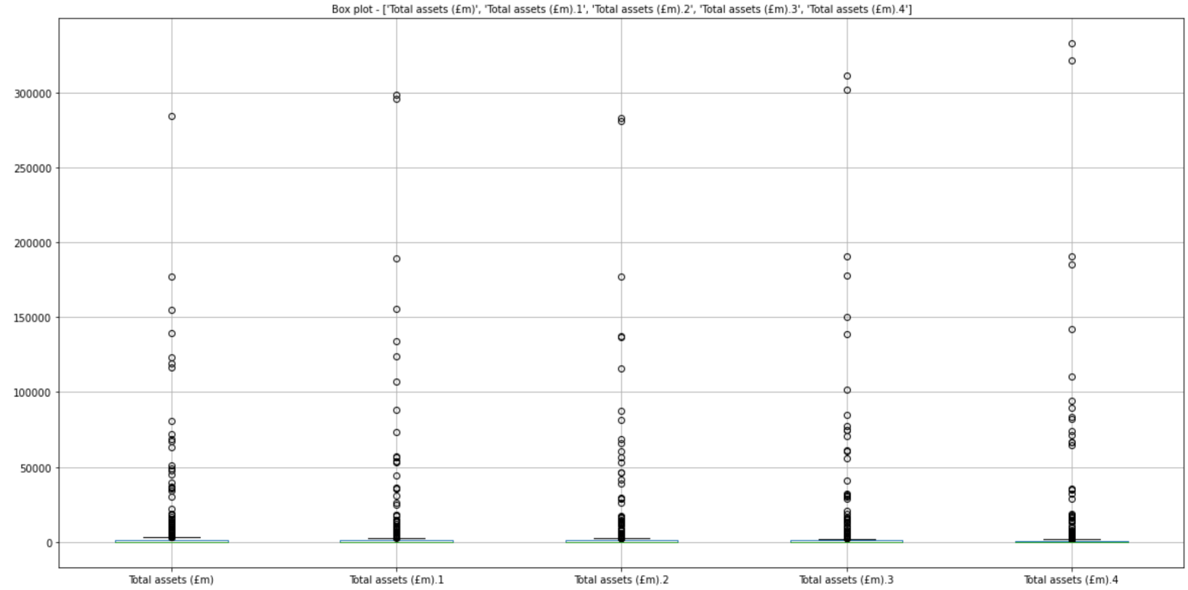
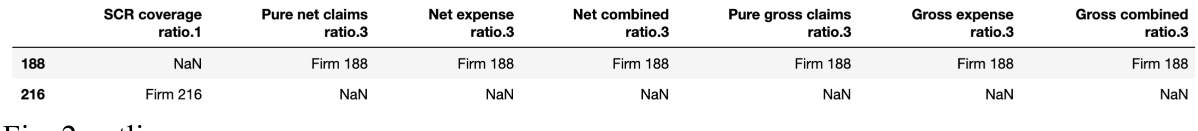


Fig.1 Total Assets Box plot

However, in columns 'Pure net claims ratio.3', 'Net expense ratio.3', 'Net combined ratio.3', 'Pure gross claims ratio.3', 'Gross expense ratio.3', 'Gross combined ratio.3' are significantly affected by outliers. Intrestingly these outliers for all of these 6 columns are present in the year 2019 and occurs from the same firm Firm 188, as shown in fig.2. Besides, point outliers exist for 'SCR coverage ratio.1' in the year 2017 for the firm Firm 216. These outliers are very extreme values and can be considered as errors.



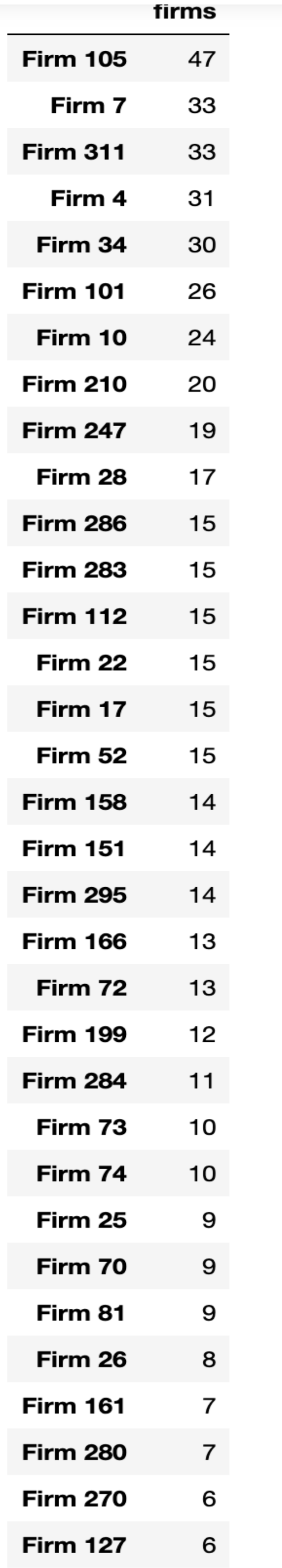
	SCR coverage ratio.1	Pure net claims ratio.3	Net expense ratio.3	Net combined ratio.3	Pure gross claims ratio.3	Gross expense ratio.3	Gross combined ratio.3
188	NaN	Firm 188	Firm 188	Firm 188	Firm 188	Firm 188	Firm 188
216	Firm 216	NaN	NaN	NaN	NaN	NaN	NaN

Fig. 2 outliers
After removing the positive outliers, this time the data is plagued with negative outlier for the same 5 columns. However, this time the outlier exists in the year 2017 for the firm 99. further the outliers were identified with the help of distribution of data and standard deviation around the mean. Any data point present outside +/-3 will be identified for further analysis.

Measuring the deviation around the mean –

fig. 3 shows the frequency with which the data for each firm deviates the +/-3 standard deviation rule across all columns. Data for firms with frequency of outliers greater than 5 will be removed and the ones with a frequency less than 5 will be treated cautiously.

Since the a single firm is having multiple instances of outliers its might not be an error in data.



	firms
Firm 105	47
Firm 7	33
Firm 311	33
Firm 4	31
Firm 34	30
Firm 101	26
Firm 10	24
Firm 210	20
Firm 247	19
Firm 28	17
Firm 286	15
Firm 283	15
Firm 112	15
Firm 22	15
Firm 17	15
Firm 52	15
Firm 158	14
Firm 151	14
Firm 295	14
Firm 166	13
Firm 72	13
Firm 199	12
Firm 284	11
Firm 73	10
Firm 74	10
Firm 25	9
Firm 70	9
Firm 81	9
Firm 26	8
Firm 161	7
Firm 280	7
Firm 270	6
Firm 127	6

Fig.3 frequency of outliers per firm across all variables

Year over year changes –

most of the significant year over year changes recorded for 'Pure net claims ratio'. Firms 29 and 50 recorded significant change in 2020. firms 29 and 64 recorded significant year over changes in 2020 under the 'Gross claims incurred (£m).4' field. firms 29 and 308 recorded significant year over changes in 2020 under the 'Pure net claims ratio.4' field. most of the significant year over year change was recorded for firm 47. However all of these changes happened during 2018. where as most of the significant year over changes happened for firm 29 in 2020. for these two firms, 47 and 29 significant year over year changes happening in a single year indicates outliers.

Big Firms – Total Assets –

Top 30 big firms are shown in fig.4

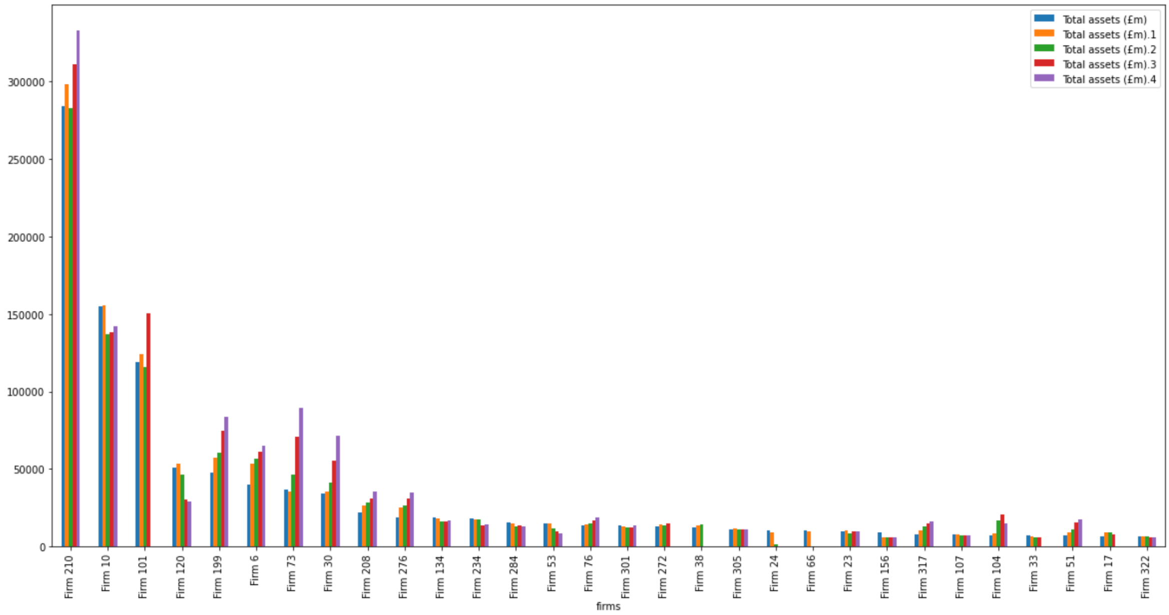
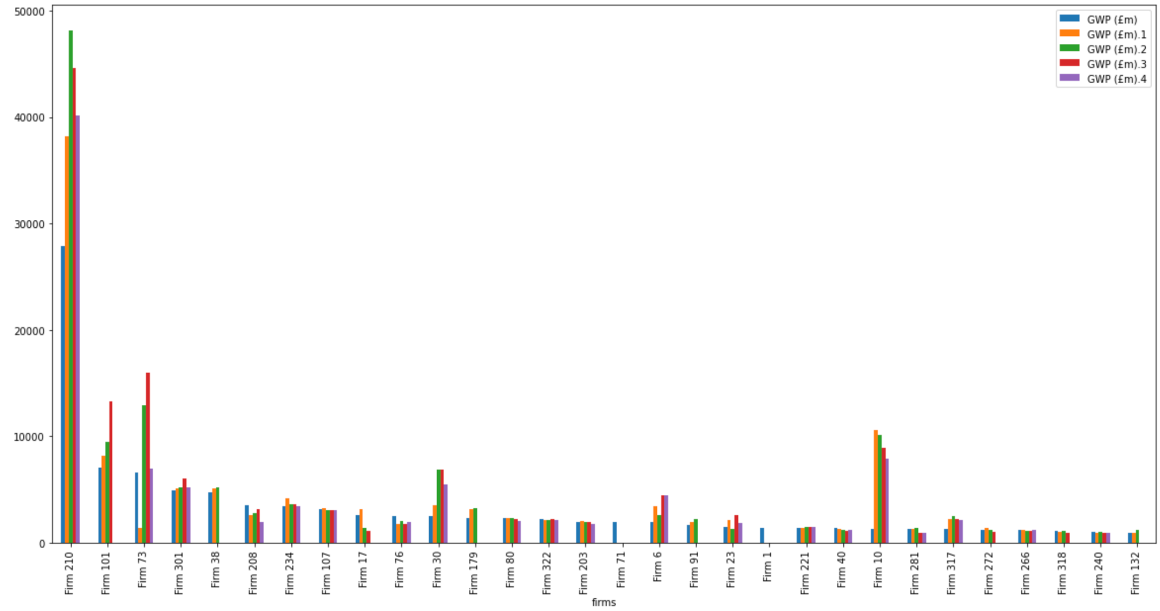


Fig.4 top 30 firms.

for firm 1, 67 and 131 the 2016 numbers might be an anamoly because the later four year figures are following the similar trend. for firm 101 and 272 the 2020 figures might be an error because the latest figures differ significantly from its earlier numbers

GPW –

Top GPW are shown in fig.5



It has been observed that most of the values in GWP are lower values. Among such lower values firm 210 appears to be an outlier(error entry).

for firm 1, and 71 the 2016 numbers might be an outlier because the later four year figures are foloowing the similar trend. on the other hand 17, 101, 272, 318 might be an outlier because the first four year figures following the similar trend

NWP –
Top NWP are shown in fig.6

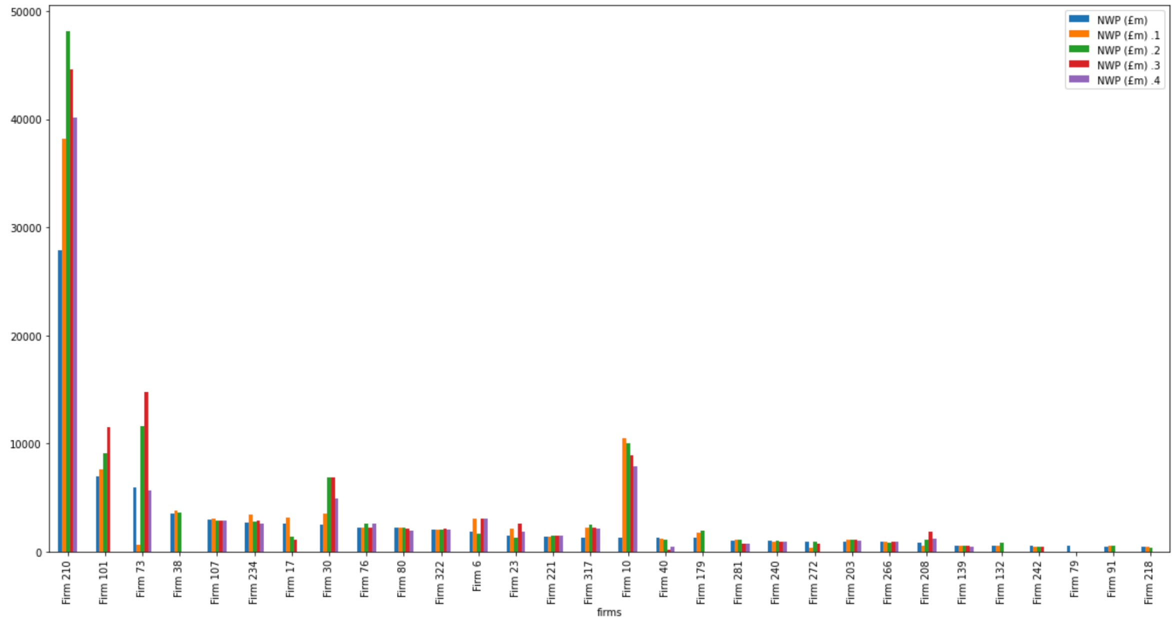


Fig.6 NWP

nwp and gwp have similar number of value counts per grouping.

It has been observed that most of the values in GWP are lower values. Among such lower values firm 210 appears to be an outlier(error entry).

for firm 79, 242 and 272 the 2016 numbers might be an outlier because the later four year figures are following the similar trend. on the other hand 17 might be an outlier because the first four year figures following the similar trend

SCR Coverage Ratio
SCR Coverage Ratio Fig.7

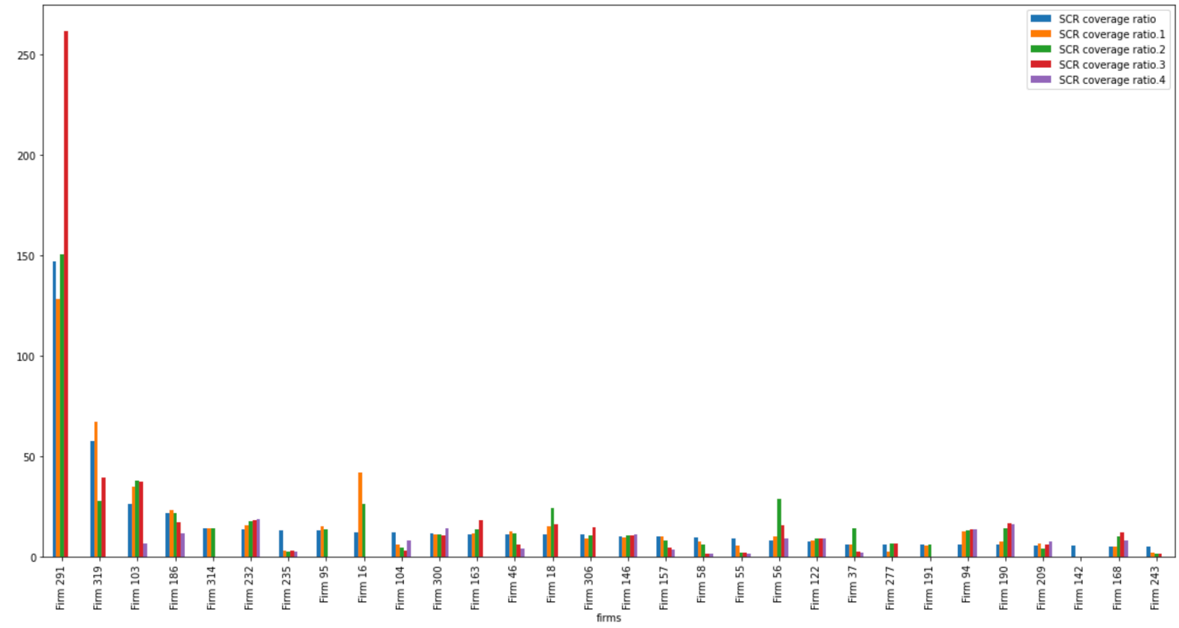


Fig.7 SCR Coverage Ratio
firm 320 for the year 2016 recorded an error input

for firm 142 the 2016 numbers are be an outlier because the later four year figures are following the similar trend. on the other hand 18, 163, 277, 291, 306, 319 might be an outlier because the first four year figures following the similar trend

Gross Claims Incurred –

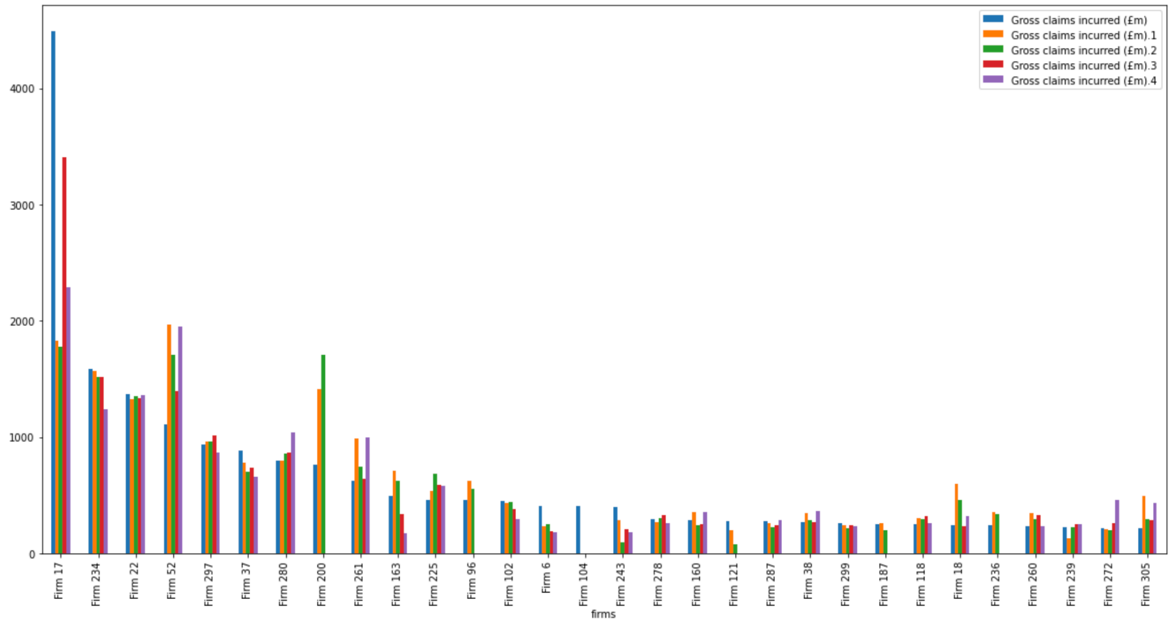


Fig.8 Gross Claims incurred

Net Combined Ratio –

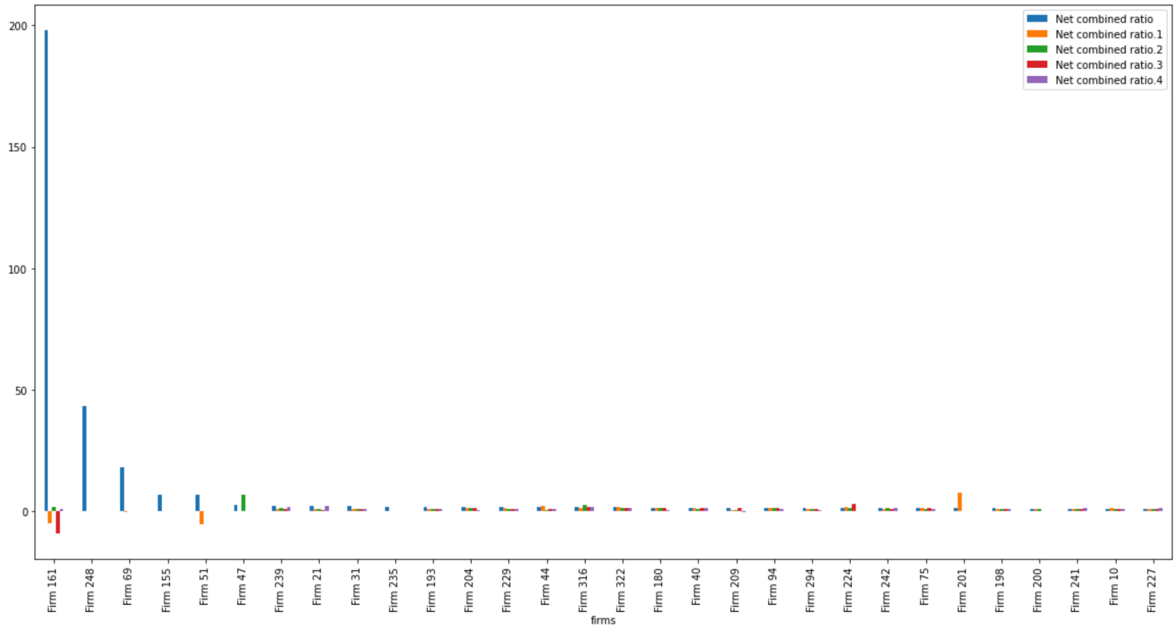


Fig.9 Net Combined Ratio
firms 161, 248, 69, 155, 47 and 201 require the most attention among net combined ratio as these firms are loss making.

Correlation –

there's no linear relation ship between scr coverage ratio and total assets, gross claims incurred and total assets, net combined ratio and total assets as shown in fig.10

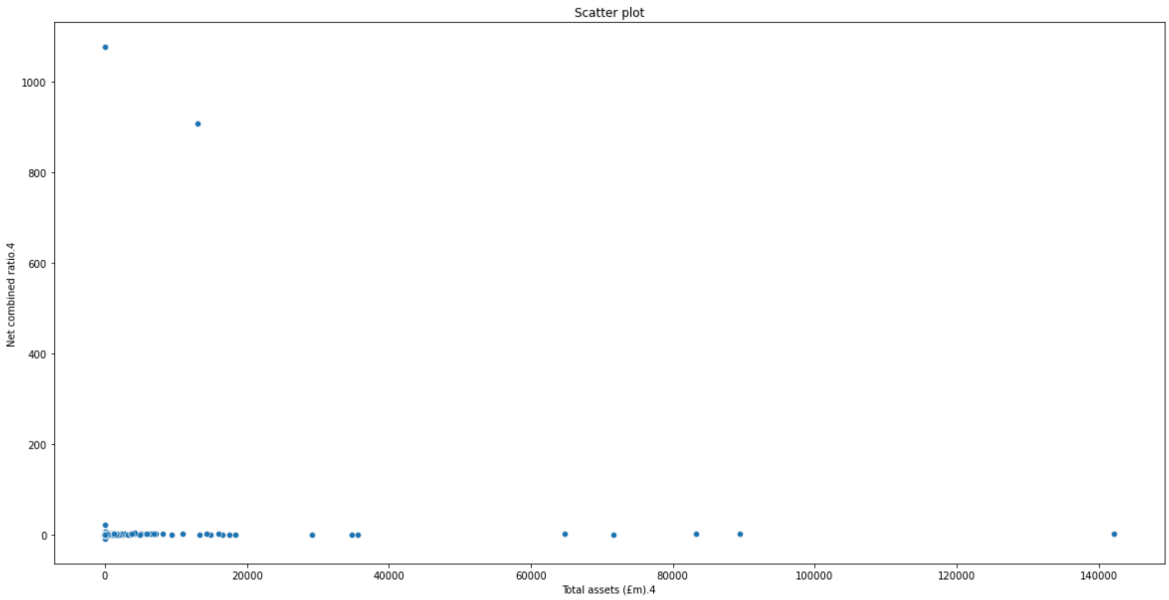


Fig.10 scatterplot between total assets and net combined ratio

There's a linear relation between total assets and gwp, total assets and nwp and gwp and nwp, total assets and total liabilities as ahown in fig.11

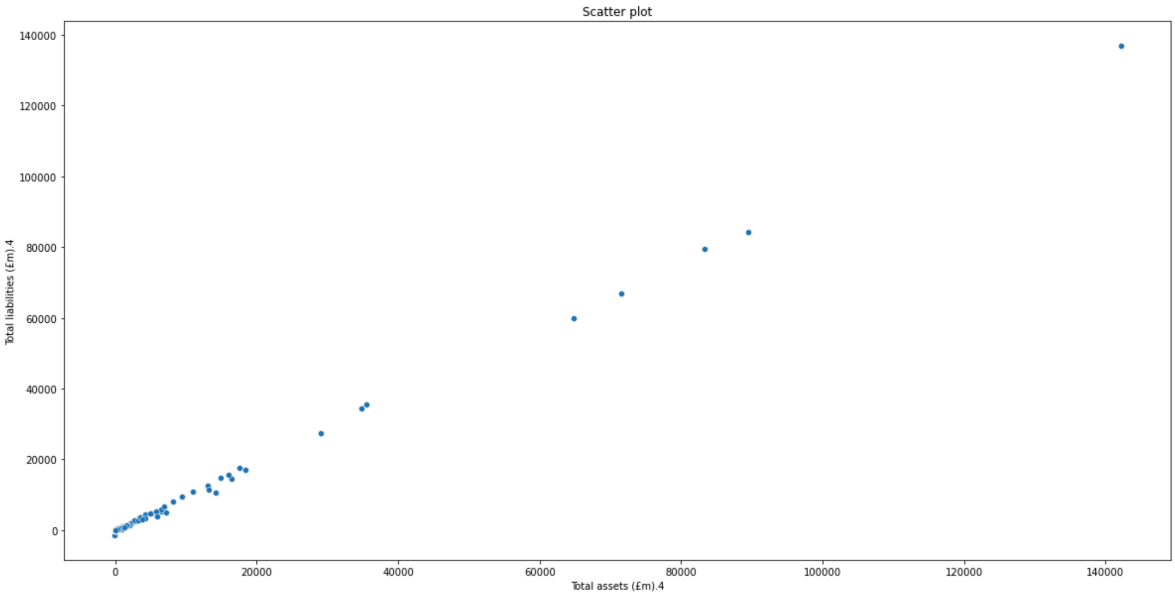
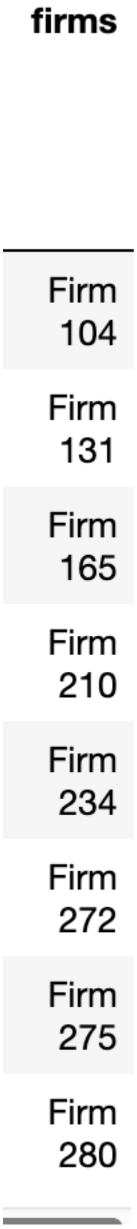


Fig.11 Scatter plot between total assets and total liabilities.

Firms that require the scare resources –

Firms that require scare resources are as shown in fig.12



firms
Firm 104
Firm 131
Firm 165
Firm 210
Firm 234
Firm 272
Firm 275
Firm 280

Fig.12 firms that require scarce resources.

Conclusion –

Data was handled with caution to look after for potential outliers and verify for possible errors across multiple columns. Correlations were drawn between variables. Discretization of the continuous variables was done to reduce the dimensionality and to look deep into the data to identify the firms which require the most resources. Finally the firms requiring the scarce resources was identified and submitted. At the end machine learning use cases in handling outliers such as winsorization, log transformation and unsupervised k means clustering was performed.

Annex

the data set is skewed and contains lots of outliers hence would like to start with normalizing the data normal and outlier treatment.

For outlier treatmenet, started with Winsorization, winsorizing is treating the outliers but the effectiveness is limited.
hence trying to treat the outliers with log transformations.

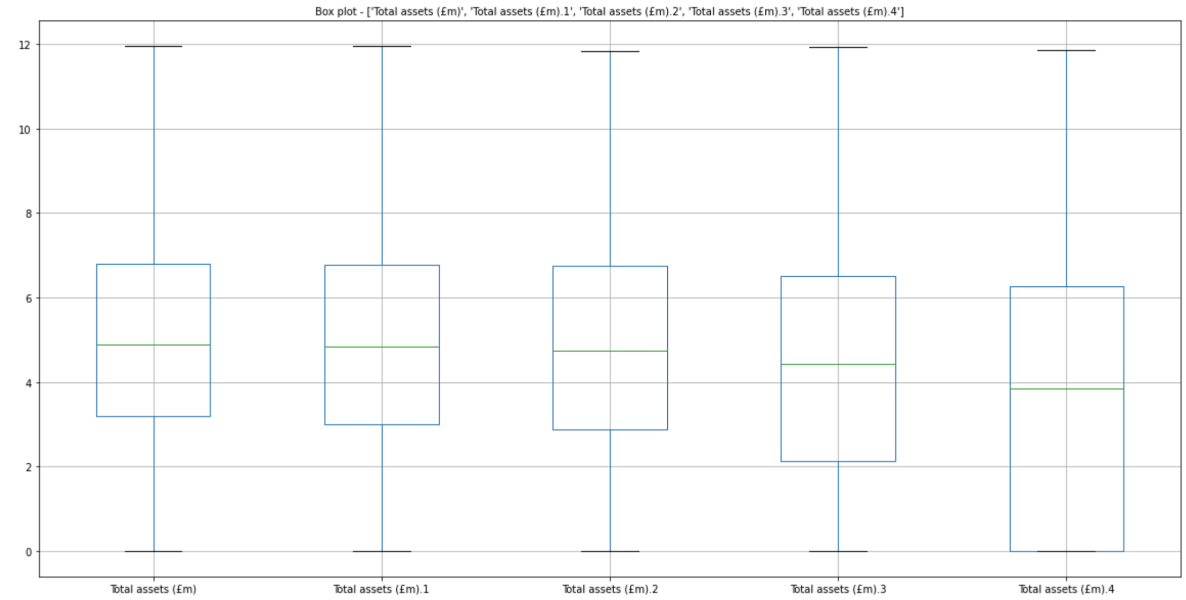


fig.13 total assets box plots after log transformed.
 log transformations helps in normalising the data, reducing skewness and kurtosis and most importantly it eliminates outliers as shown in fig.13

finally performed k means clustering to identify outliers/anomalies unsupervised as shown in fig.14

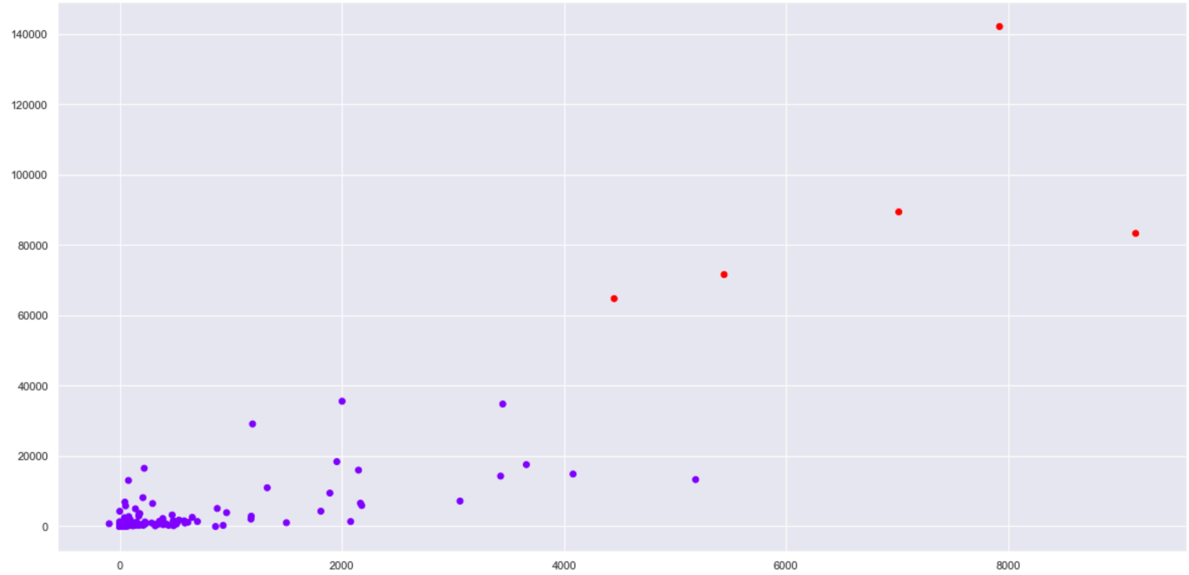


fig.14 clustering of GWP and total assets.