# Technical Assessment (all areas) – Data Scientist – 006191 – Fan Sun

## Task 1. Data Exploratory Analysis

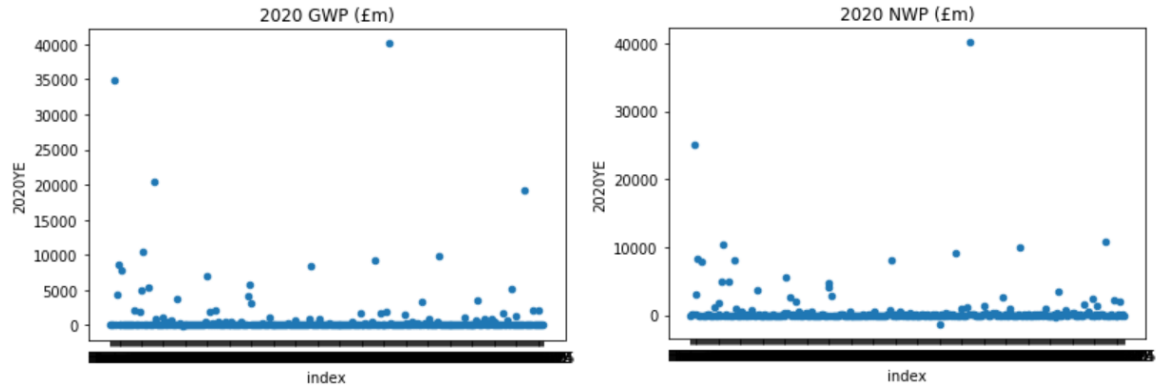
In this part, we going to use the given dataset to do exploratory analysis from three angels by using the five metrics provided and giving insights to the team on which firms we should focus. The reason these methodologies are employed is to limit the number of businesses being watched to a reasonable number, i.e. we don't want to watch all of the metrics all the time.

### 1.1 Big sized firms

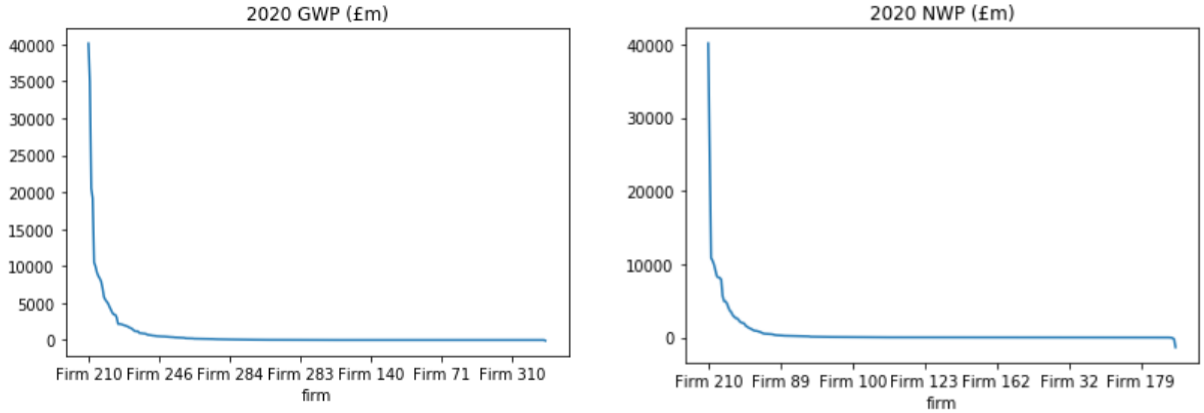
The latest (2020 year) NWP and GWP metrics are chosen to identify the firm size, as

- NWP and GWP are both money figures indicating a company's size
- they are static data implying the company's present performance
- yearly data is long enough to discount the unexpected situations or seasonal patterns.

To be specific, scatter plot of 2020 GWP and 2020 NWP are plotted to spot the extreme high values, which are the big firms.




It's fairly easy to spot the top two or three extreme values by looking at these above figures, but not easy to make the big firms list longer. So, another two scatter plots sorted by value are produced to figure out this problem. This also helps avoid the problem that randomly picking a number - i.e. top 10 - is insufficient.



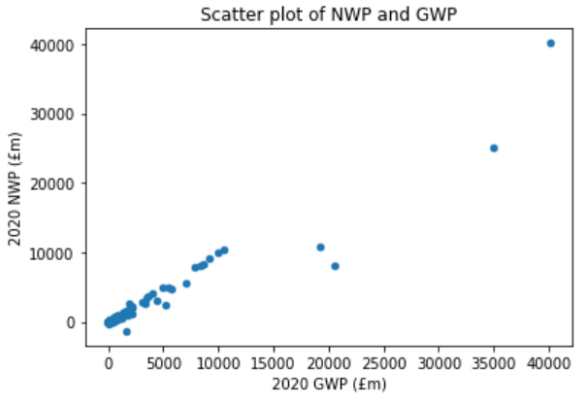

Both figures contain an inflection point: there are few companies with big values on the left of the inflection point and loads of small companies on the right. We extract company name and their values in the below tables.

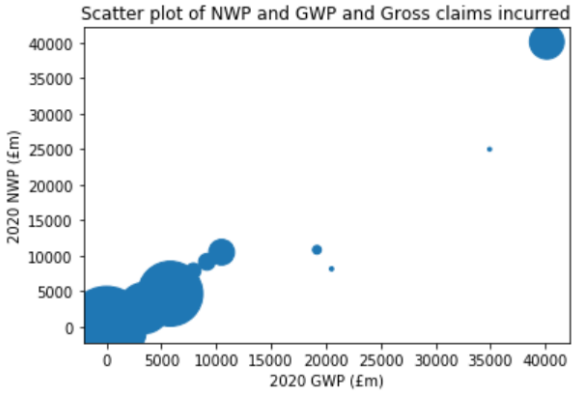
| Firm name | 2020 GWP (£m) |
|---|---|
| Firm 210 | 40135.69 |
| Firm 4 | 34922.70 |
| Firm 34 | 20510.75 |
| Firm 311 | 19180.02 |
| Firm 26 | 10489.25 |
| Firm 247 | 9961.52 |
| Firm 199 | 9149.58 |
| Firm 7 | 8652.95 |
| Firm 151 | 8341.64 |
| Firm 10 | 7923.37 |
| Firm 73 | 7016.19 |
| Firm 105 | 5811.66 |
| Firm 30 | 5442.88 |
| Firm 301 | 5186.51 |

| Firm name | 2020 NWP (£m) |
|---|---|
| Firm 210 | 40135.69 |
| Firm 4 | 24996.02 |
| Firm 311 | 10830.97 |
| Firm 26 | 10489.25 |
| Firm 247 | 9961.52 |
| Firm 199 | 9134.28 |
| Firm 7 | 8359.91 |
| Firm 151 | 8180.39 |
| Firm 34 | 8145.62 |
| Firm 10 | 7893.06 |

Firm names marked in red means they show up in both of the top GWP and NWP rank, which is where our team needs to focus.

Then we include both metrics on the same chart. The slope represents NWP/GWP, and the shallower slope, the more the risk is being passed on to reinsurers. There are three points below the most companies' slope, which means that they have transferred the risk to the reinsurer, they are firm 4, firm 311, and firm 34. All of them are in the red list of big companies with big NWP and GWP. However, if we add the gross claims incurred as the size in the scatter, then we find these three companies do not have huge gross claims incurred but they still transfer the risk to reinsurer, as such, they might need more attention.



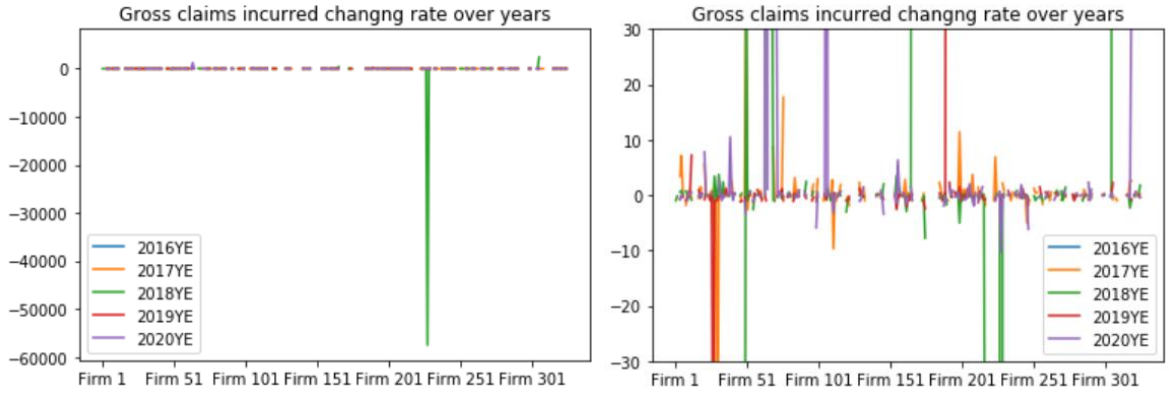

### 1.2 Changing business profile

We are considering the changing rate of gross claims incurred over five years in this part as:

- changing rate is the measurement discount company size or profitability
- gross claims incurred is a large cost to an insurer, monitoring how these change over time for a firm is vital
- it isn't affected by the missing values

We plotted the changing rate over years, and from the left plot below we can spot the first outlier. Firm 228 has a big trough in green, which means it's decreased dramatically from 2017 to 2018, and the decrease rate is too huge to be reliable. We rescale the y axis to see other firms'

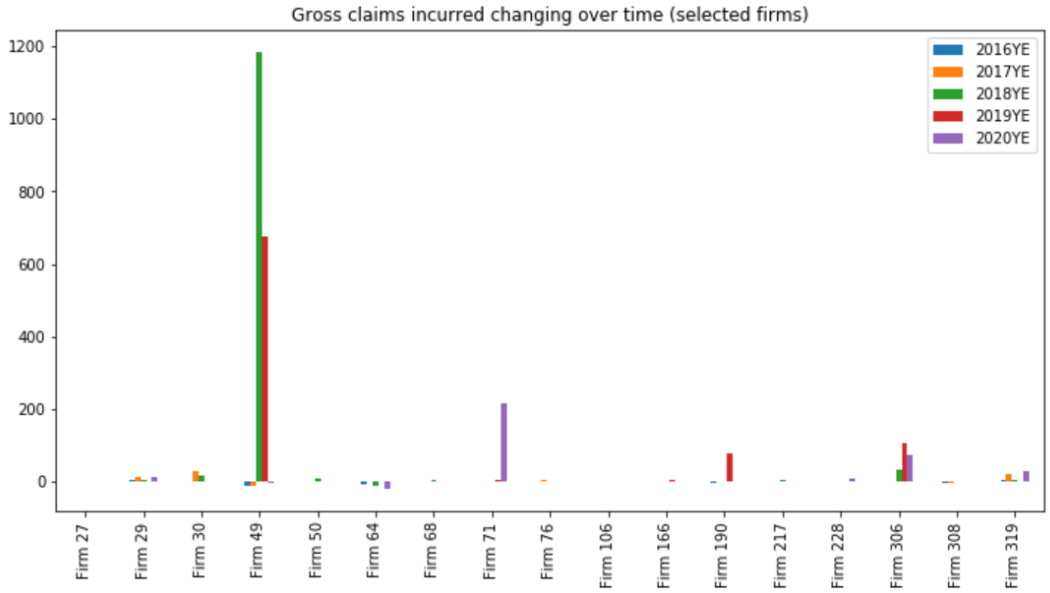
patterns, any company – or any legend colour – with obvious peaks or troughs experienced substantial change year on year.




We retrieved the firm with abnormal change rate in the table below. However, for some start-up companies, a small increase in claims can lead to huge growth rates, so it will be more reasonable to analyse these listing company with their gross claims incurred money as well.

| Firm name | 2016-2017 gross claims incurred changing rate | 2017-2018 gross claims incurred changing rate | 2018-2019 gross claims incurred changing rate | 2019-2020 gross claims incurred changing rate |
|---|---|---|---|---|
| Firm 27 | | | -57.55 | |
| Firm 29 | | | | 5137.30 |
| Firm 30 | -72.55 | | | |
| Firm 49 | | -113.46 | | |
| Firm 50 | 33.04 | 41.70 | | |
| Firm 64 | | 183.18 | | 1255.91 |
| Firm 68 | | 57.12 | | |
| Firm 71 | | | | 48.01 |
| Firm 76 | 17.67 | | | |
| Firm 106 | | | | 163.05 |
| Firm 166 | | 409.53 | | |
| Firm 190 | | | 209.74 | |
| Firm 217 | | -84.91 | | |
| Firm 228 | | -57530.08 | | -10.43 |
| Firm 306 | | 2378.98 | | |
| Firm 308 | | | | 90.19 |
| Firm 319 | | | | 41.48 |

From the bar chart below, we could spot that some companies have a huge changing rate. Many of these may be because of a comparison between a small starting amount. We narrow our search based on this to focus on those companies with both huge gross claims incurred and the huge changing rate such as Firm 49, 71, 190, and 360.



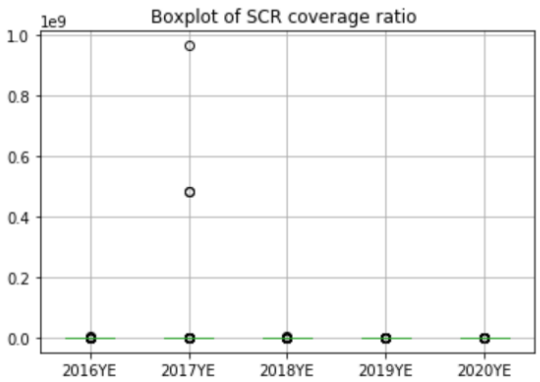

### 1.3 Outliers

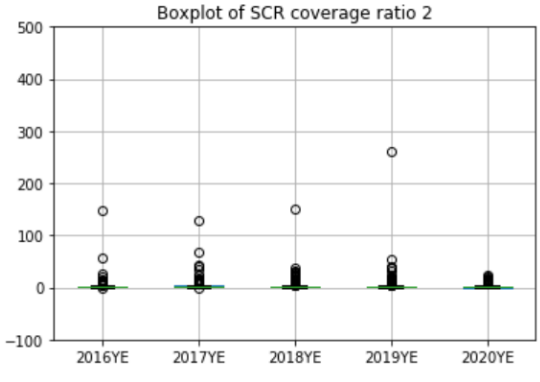
In this section, we are going to compare firms' SCR coverage ratio and net combined ratio with average value for every single reporting period as

- SCR coverage ratio greater than 100% means the firm is holding enough capital to meet the requirement, and the size of the buffer (i.e. surplus over 100%) can be important
- Net combined ratio indicates the profitability of a firm, and if less than 100% it indicates a profit.

We firstly check the average ratios of each year, finding out some numbers are way more than normal magnitude (100% ish). Then we plot all the data into boxplots to see if there are extreme outliers.

| | 2016 | 2017 | 2018 | 2019 | 2020 |
|---|---|---|---|---|---|
| **Avg SCR coverage ratio** | 12869.47 | 5930313.00 | 12467.41 | 533.02 | 514.22 |
| **Avg net combined ratio** | 1.35 | -133.05 | 5.99 | 7477.47 | 9.85 |

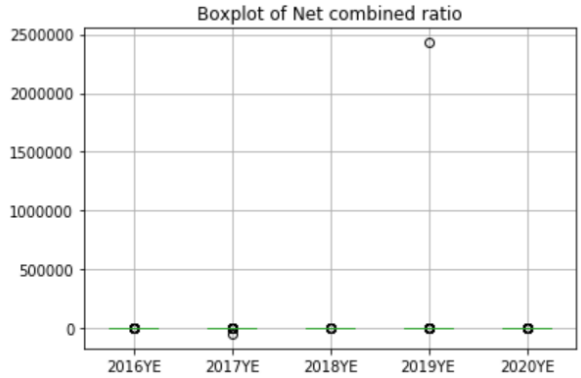


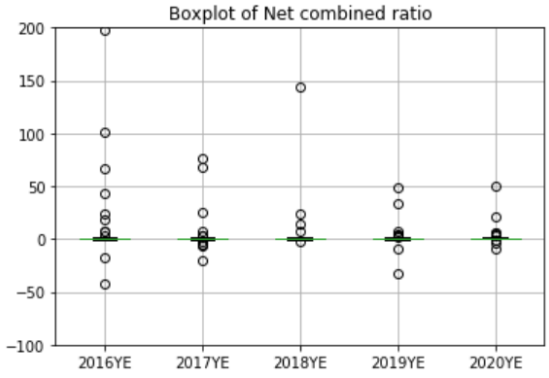




The left above figure is about SCR ratio, the outliers are in the magnitude in $10^9$, which are obviously wrong outliers. After picking out those wrong outliers listed in the following table, another boxplot with rescaled y axis is plotted and the outliers in the right above figure are genuine outliers.

| Firm name | 2016 SCR coverage ratio | 2017 SCR coverage ratio | 2018 SCR coverage ratio | 2019 SCR coverage ratio | 2020 SCR coverage ratio |
|---|---|---|---|---|---|
| Firm 1 | | 481792000.00 | | | |
| Firm 66 | | | 3856018.00 | | |
| Firm 127 | | 182412.20 | 194753.80 | 171974.69 | 166394.58 |
| Firm 131 | | 481792000.00 | | | |
| Firm 216 | | 963584000.00 | | | |
| Firm 320 | 4181573.00 | | | | |

We do the same for net combined ratio, the below left boxplot shows the wrong outliers, while the below right shows the genuine outliers. The only difference will be there are lots of negative net combined ratio and huge net combined ratio. The negative ratio may come from one of the negative claims, expenses, and premium earned (if these companies do record profit as negative expenses like this), but the large extreme magnitude is still obviously wrong as listed in the below table.




| Firm name | 2016 net combined | 2017 net combined ratio | 2018 net combined ratio | 2019 net combined ratio | 2020 net combined ratio |
|---|---|---|---|---|---|
| Firm 28 | -124.29 | 821.47 | 1578.85 | | |
| Firm 70 | | 1738.30 | | | |
| Firm 99 | | -46116.70 | | | |
| Firm 166 | 101.72 | -2.44 | | -248.63 | 989.16 |
| Firm 188 | | | | 2430023.00 | |
| Firm 228 | | | | | 1076.16 |
| Firm 284 | | | | 435.58 | 906.31 |

## Task 2. Clustering Analysis

Regarding the further insights from detecting outliers, we could employ ML techniques. Unsupervised learning is applied here because the data set is not labelled. Clustering can be used by grouping the variables in such a way that objects in the same group are more similar to each other than to those in other groups.

One common clustering algorithm is K means analysis. In the K-means based outlier detection technique the firms are partitioned into K groups by assigning them to the closest cluster centres. Once assigned we can compute the distance to its cluster centre, and then calculate the difference between the average cluster distance to it, and pick those with largest difference as outliers.

Because the five metric we picked includes value and ratio, which the magnitudes are very different, so an initial normalization is required.

We fit several K means models by choosing K from 3 to 13 with 2 as steps and to check the evaluation score (BC/WC) first, where BC is overall between cluster score and WC is overall within cluster score. The goal is trying to minimise the WC and maximize BC, which is looking for larger values of the overall clustering score. The associated results are in the below table.

| K | 3 | 5 | 7 | 9 | 11 | 13 |
|---|---|---|---|---|---|---|
| BC | 1.89 | 9.69 | 22.74 | 31.62 | 46.23 | 69.80 |
| WC | 5.90 | 2.45 | 1.03 | 0.58 | 0.40 | 0.21 |
| Score (BC/WC) | 0.32 | 3.96 | 22.14 | 54.37 | 116.76 | 325.04 |

The score increases as K become larger, but when the K moved from 5 to 7, the score improves almost 7 times, indicating K=7 is a good start.

We fit the data into K means model and assign the cluster name, calculate the distance to the cluster centre, and compare the difference with the average cluster distance shown in the table below.

| Cluster name | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| Avg distance | 21.52 | 13.71 | 22.85 | 13.92 | 16.78 | 18.85 | 15.32 |

Then we sort the difference to spot the outliers as shown in the below table.

| Firm name | GWP | NWP | Gross claims incurred | SCR coverage ratio | Net combined ratio | Cluster | distance | avg_dis | diff |
|---|---|---|---|---|---|---|---|---|---|
| Firm 112 | 0.03 | 0.00 | 1.00 | 0.00 | 0.01 | 4.00 | 7.16 | 5.89 | -1.28 |
| Firm 210 | 1.00 | 1.00 | 0.19 | 0.00 | 0.01 | 1.00 | 8.93 | 8.07 | -0.86 |
| Firm 34 | 0.23 | 0.51 | 0.01 | 0.00 | 0.01 | 5.00 | 5.26 | 4.60 | -0.66 |
| Firm 311 | 0.29 | 0.48 | 0.02 | 0.00 | 0.01 | 5.00 | 5.18 | 4.60 | -0.58 |
| Firm 228 | 0.03 | 0.00 | 0.01 | 0.00 | 1.00 | 3.00 | 7.38 | 6.96 | -0.42 |
| Firm 283 | 0.03 | 0.00 | 0.49 | 0.00 | 0.01 | 2.00 | 4.86 | 4.52 | -0.34 |
| Firm 74 | 0.04 | 0.01 | 0.48 | 0.00 | 0.01 | 2.00 | 4.81 | 4.52 | -0.29 |
| Firm 22 | 0.03 | 0.00 | 0.45 | 0.00 | 0.01 | 2.00 | 4.73 | 4.52 | -0.21 |
| Firm 304 | 0.03 | 0.00 | 0.43 | 0.00 | 0.01 | 2.00 | 4.67 | 4.52 | -0.15 |

To make the results more reliable, we repeat the process setting K as 5 and 9, because their BC/WC score are good and avoid an overfitting problem. The below table shows the firm with largest difference distance when K =5, 7, and 9. The name marked in red means they are detected as outlier by K means.

| K=5 | |
|---|---|
| Firm name | diff |
| Firm 112 | -1.37 |
| Firm 17 | -0.78 |
| Firm 210 | -0.73 |
| Firm 52 | -0.57 |
| Firm 105 | -0.55 |
| Firm 34 | -0.12 |
| Firm 283 | -0.12 |
| Firm 74 | -0.11 |
| Firm 311 | -0.10 |

| K=7 | |
|---|---|
| Firm name | diff |
| Firm 112 | -1.28 |
| Firm 210 | -0.86 |
| Firm 34 | -0.66 |
| Firm 311 | -0.58 |
| Firm 228 | -0.42 |
| Firm 283 | -0.34 |
| Firm 74 | -0.29 |
| Firm 22 | -0.21 |
| Firm 304 | -0.15 |

| K=9 | |
|---|---|
| Firm name | diff |
| Firm 34 | -0.68 |
| Firm 311 | -0.65 |
| Firm 210 | -0.51 |
| Firm 112 | -0.32 |
| Firm 228 | -0.28 |
| Firm 283 | -0.14 |
| Firm 74 | -0.11 |
| Firm 301 | -0.10 |