Technical Assessment Report - Data Scientist - 006191

Candidate: Eirini Zygoura

Phase 1: *Data cleaning*

The output of this phase will be a data table of all the firms containing valid records for the most significant economic factors that will be examined in Phase 2 to identify the firms that worth the most attention. These factors are:

- Gross Written Premium (GWP)
- Total Assets (TA)
- Gross claims incurred (GCI)
- Net Written Premium (NWP)
- SCR coverage ratio (SCRcov)

As a first step cleaning the data, records with false data, such as invalid negative values of GWP and TA, are excluded. Also, bankrupt companies should be excluded, so all firms with zero GWP or TA for the last year are excluded. Furthermore, non-zero values in between zeros in the timeseries of all of the above factors are not acceptable, as they can easily indicate recording errors.

Phase 2a: *Identification of the biggest firms*

In this phase the biggest firms are categorised, based on the indicators of Total Assets and Gross Written Premium. Their values per year are calculated. To account for the case that a firm was established during this period, calculations are made from the years with no zero values.

Equal weights on GWP per year and TA per year are used for sorting the firms according to their size. Figure 1 presents the resulting 15 biggest firms and their associated economic factors per year.

Firms 10, 105 and 120 in the figure are identified with extremely large TA per year. This firms should be handled with caution, since values like these could imply recording errors.
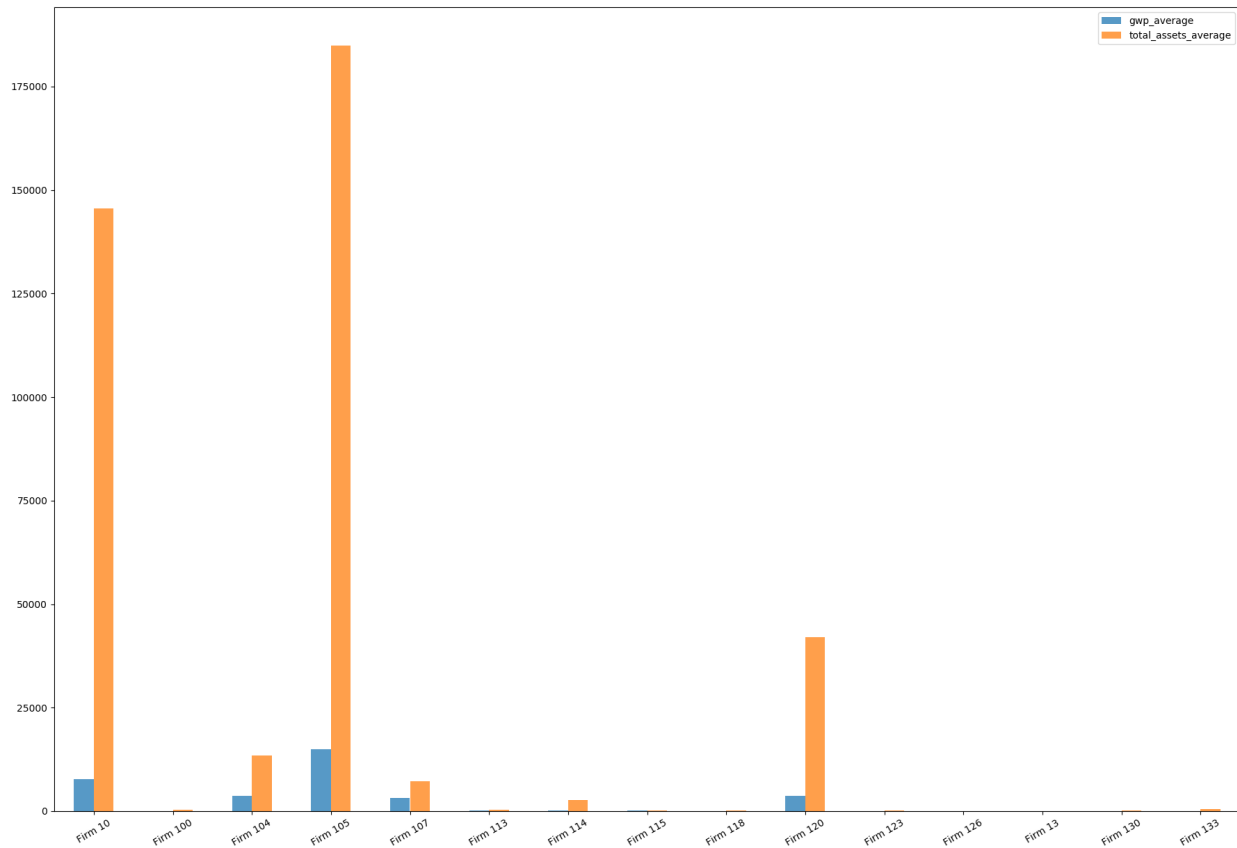
Figure1: GWP and TA per year of the biggest firms

Phase 2b: _Identification of the business profile changing firms_

From the indicators that reveal changes in a company, Gross claims incurred and Gross Written Premium were selected and evaluated. For each factor data are normalised so that values for each firm are between 0 and 1. The mean absolute deviation is a measure of the distance of the observed values from the mean value of the factor. Biggest deviations indicate the most changing firms.

The firms are then sorted by profile changes using equal weights on the standard deviations of the GCI and GWP. Figure 2 presents the mean absolute deviations of GCI and GWP for the 15 firms that changed business profile a lot over the 5 years or records.

Firm 107 shows a big difference in the deviations of GCI and GWP. This difference could indicate false records or an anomaly (such as a special event) that may have happened during the 5 year period. Further insight on the rest of the economical factors may reveal the case.
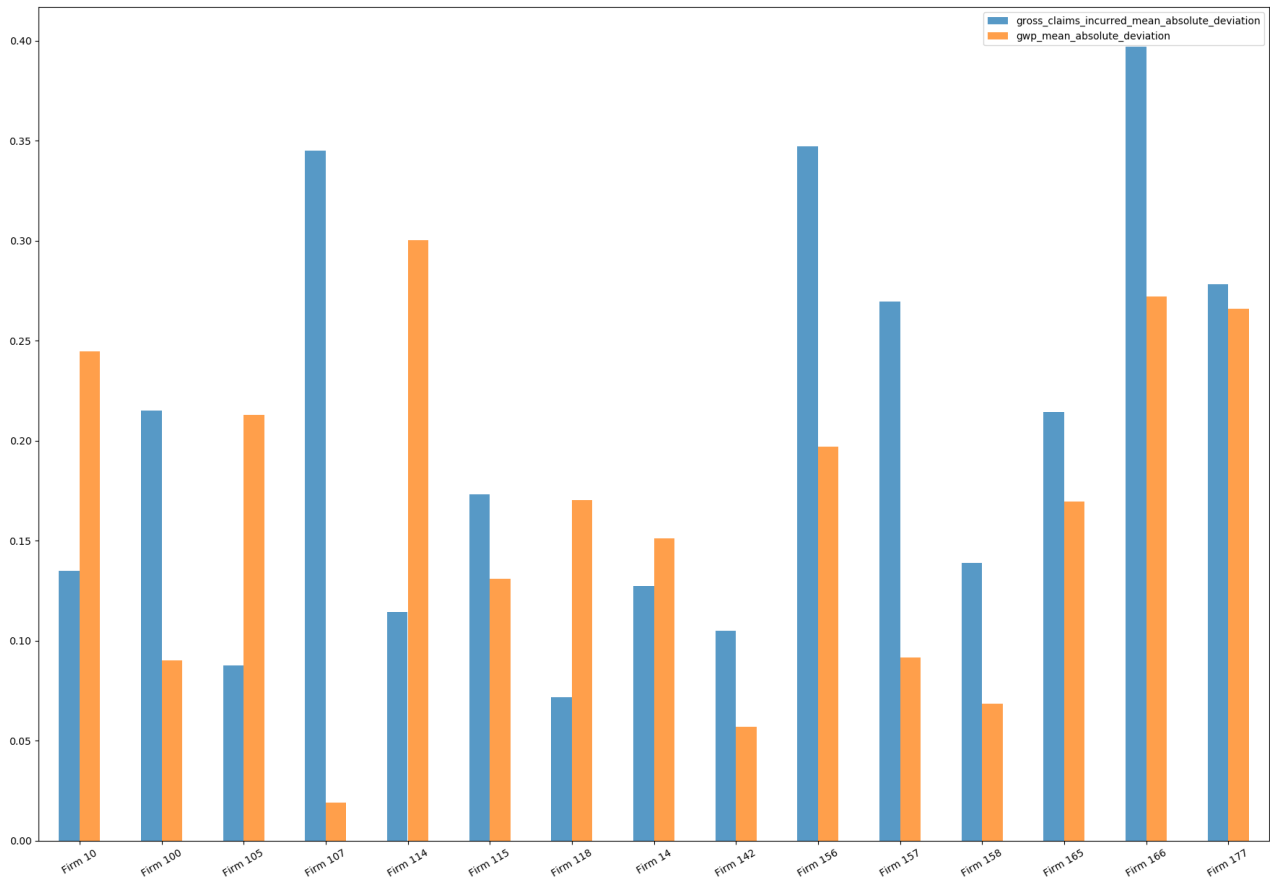
Figure 2: Mean absolute deviations for GCI and GWP of the most business profile changing firms

Phase 2c: *Outliers detection*

Outliers are detected based on some of the basic factors that characterise a company, such as GWP, NWP and SCRcov. Analysis is done separately for each year. To present the separation of the outlier firms in the feature space of these 3 factors, 3D scatter plots are presented for each year in Figures 3, 4, 5, 6 and 7 respectively. The outlier firms are marked in blue colour.

Furthermore, the firms are sorted in descending order for the number of the years that were characterised as outliers. This can be helpful to identify the excess to which a firm indeed differs from the others or there may have been an error in recording or a special event through a year. Firms with higher number of the years being outliers are first in the order. Table 1 presents the order or the outliers from 5 to 1 year being outlier.
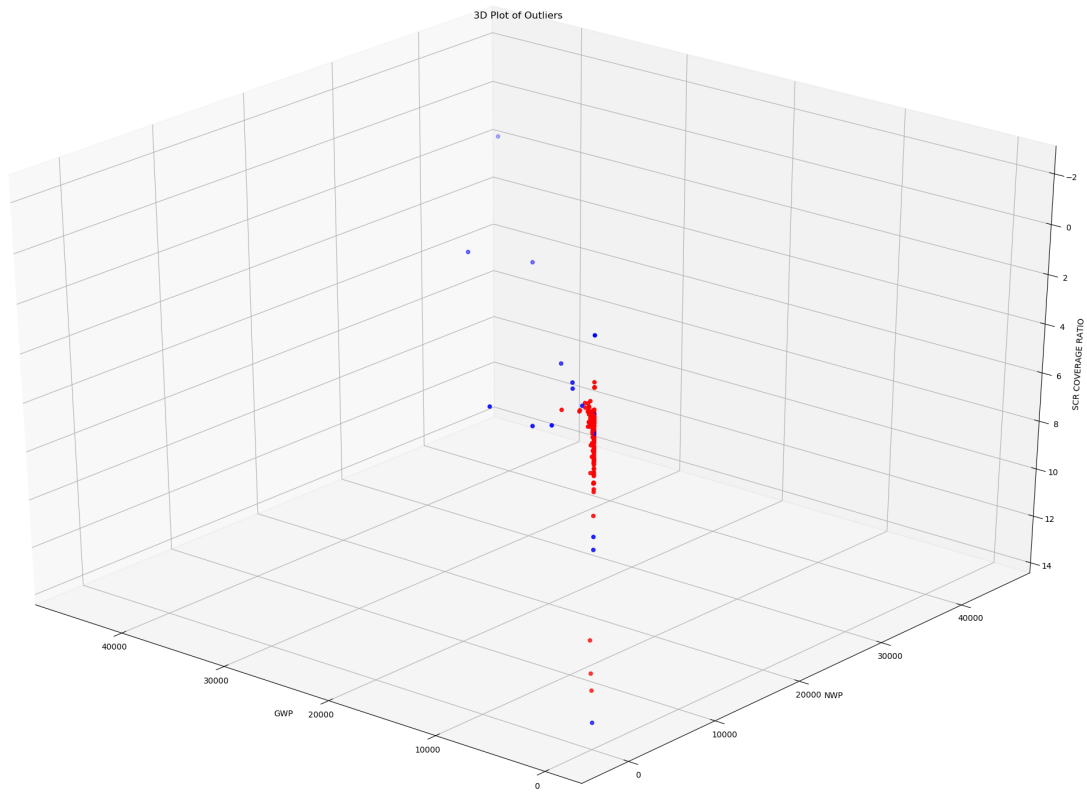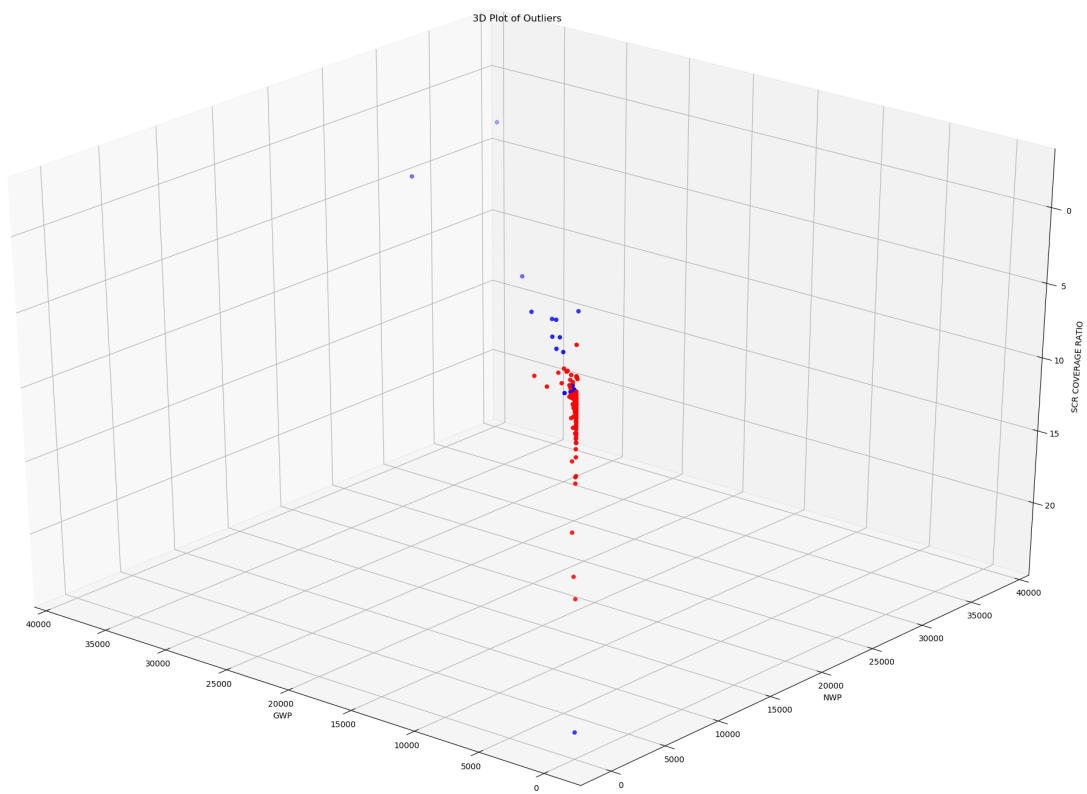
Figure 3: Scatter plot of outliers for the year 2016
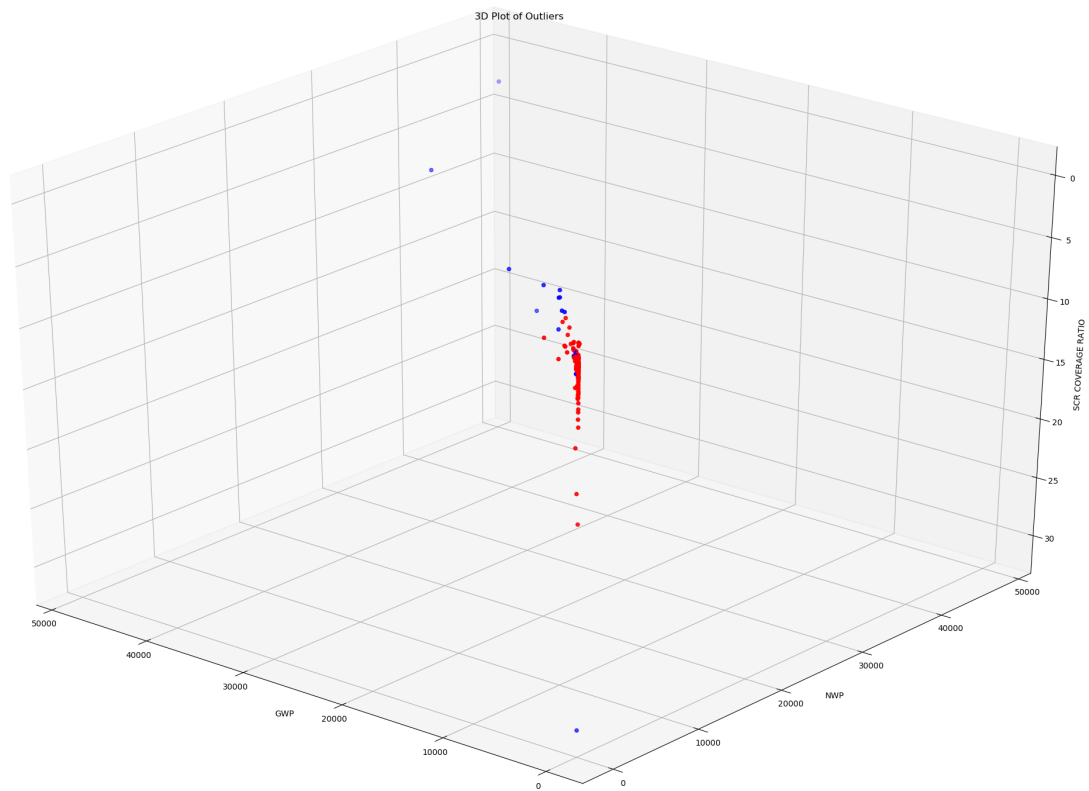

Figure 4: Scatter plot of outliers for the year 2017
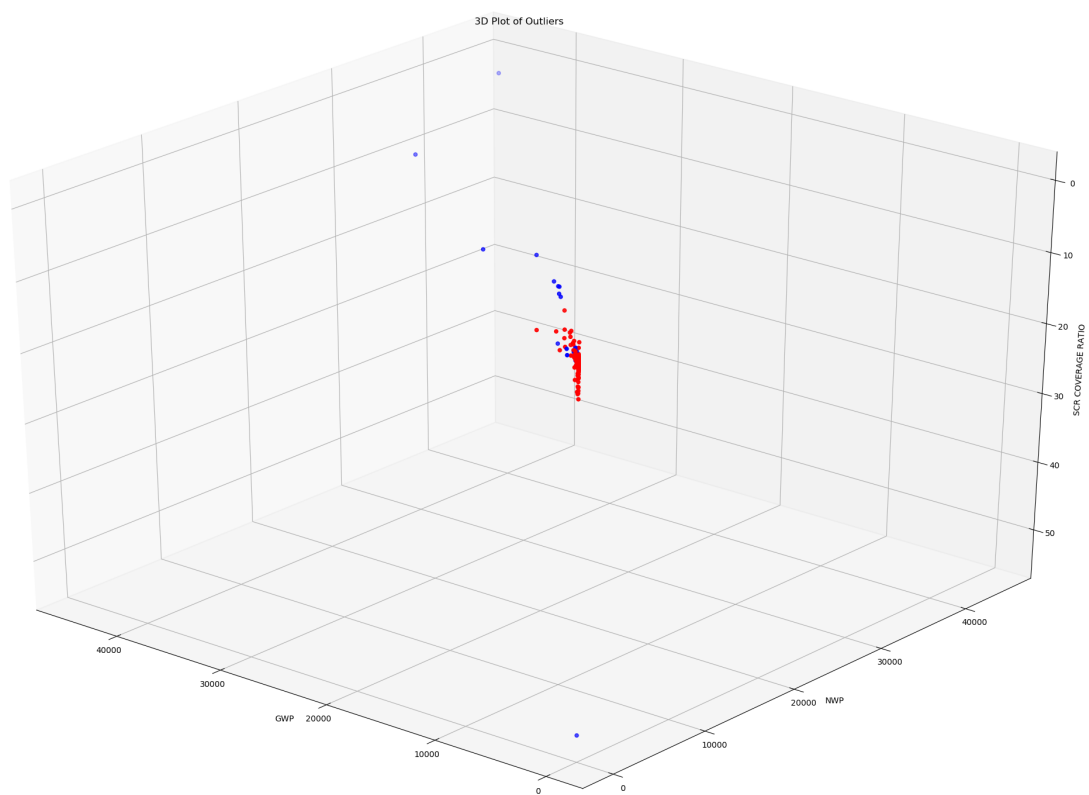
Figure 5: Scatter plot of outliers for the year 2018

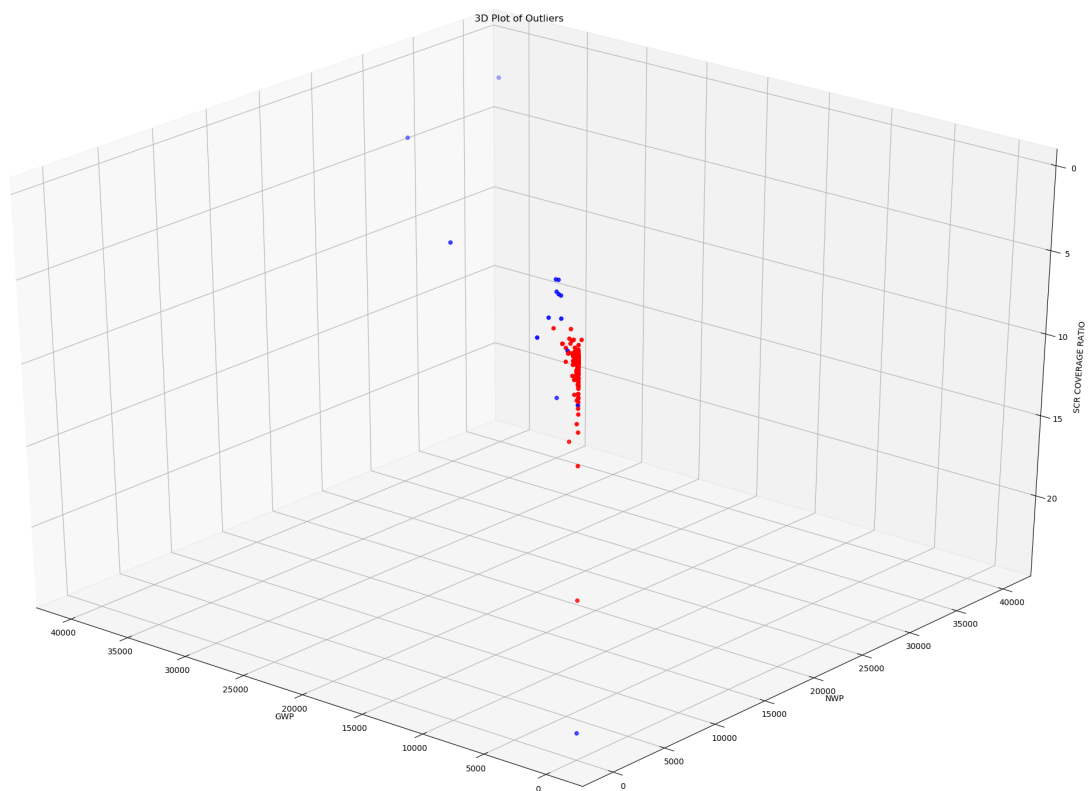*Figure 6: Scatter plot of outliers for the year 2019*

Figure 7: Scatter plot of outliers for the year 2020

Order	Outlier Firms	Order	Outlier Firms
1	Firm 10	18	Firm 247
2	Firm 104	19	Firm 25
3	Firm 120	20	Firm 26
4	Firm151	21	Firm 266
5	Firm 170	22	Firm 276
6	Firm 177	23	Firm 281
7	Firm 183	24	Firm 30
8	Firm 196	25	Firm 301
9	Firm 199	26	Firm 302
10	Firm 200	27	Firm 311
11	Firm 203	28	Firm 33

Order	Outlier Firms	Order	Outlier Firms
12	Firm 209	29	Firm 37
13	Firm 210	30	Firm 4
14	Firm 221	31	Firm 40
15	Firm 23	32	Firm 69
16	Firm 235	33	Firm 7
17	Firm 240	34	Firm 73

Table 1 :Outlier firms ordered by descending number of years being outlier