**Technical Assessment – Analyst, Data Science - 010242**

**Overview:**

There are total 325 Firms that need be analysed based on the metrics provided. First, I have joined the data from two sheet i.e. ‘General and Underwriting’.

Then I have Filtered the data for recent year YE2020 its would become easier for data preparation and handling outliers.

**Data Preparation:**

First, I have created new column, NWP/GWP and filtered the data where this metrics is not zero, since it’s one of the important metrics to identify firms. Also, there are volume of incomplete data present in the dataset and we have removed the firms for which we don’t have enough data to analyse.

Then I have performed outlier treatment to deal with outliers using Standard deviation technique:

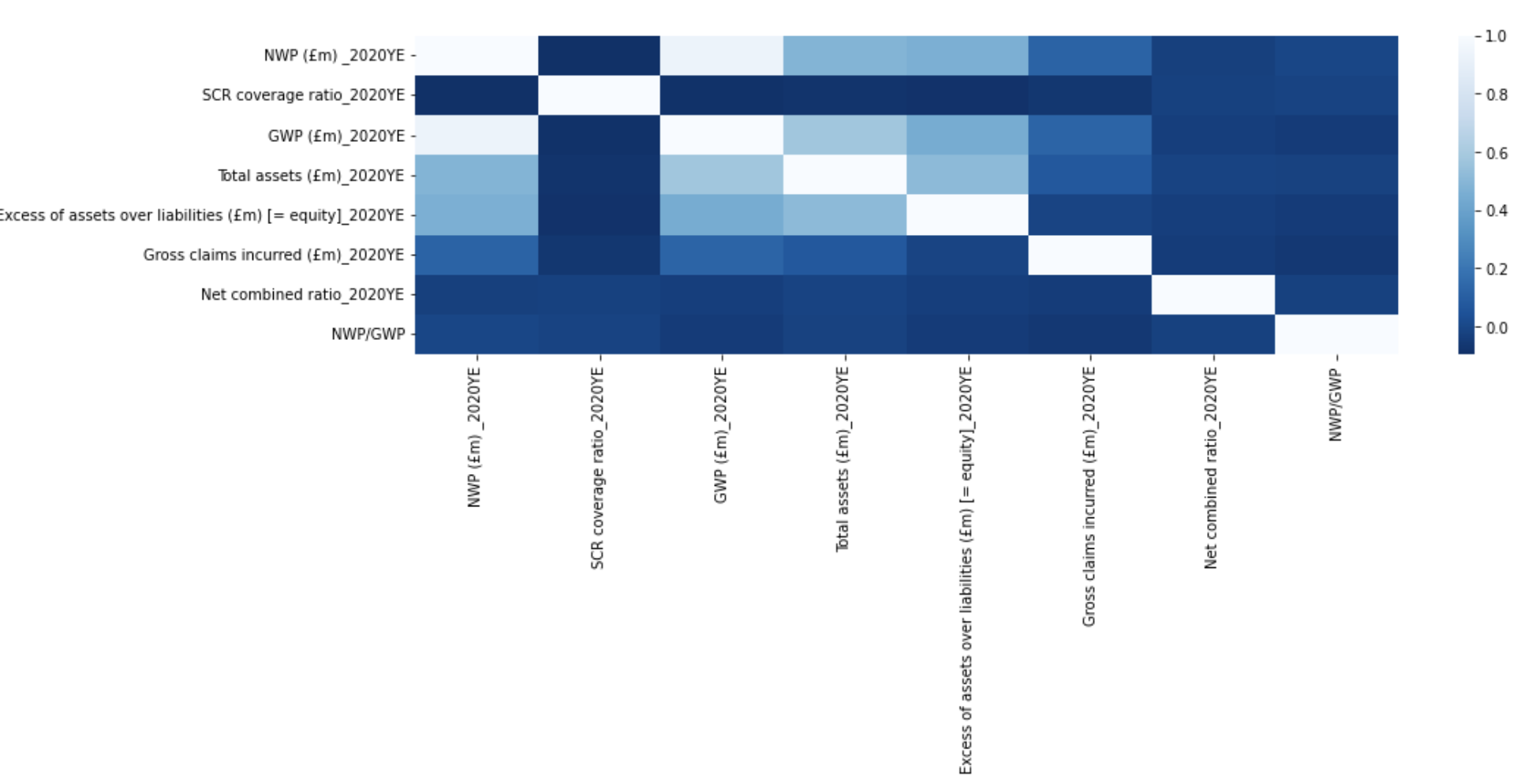
For Gross Claimed Incurred and Net combined ratio, threshold = 2 STD

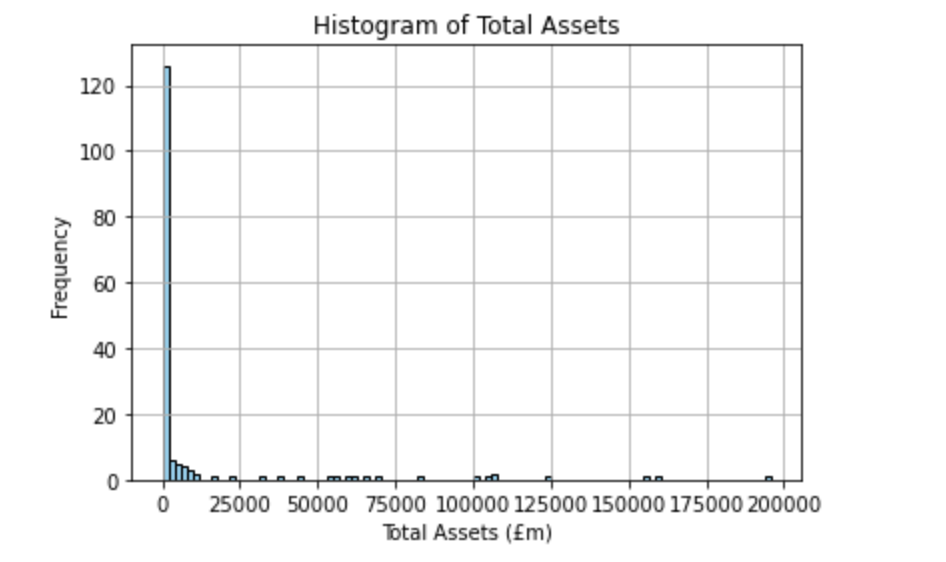
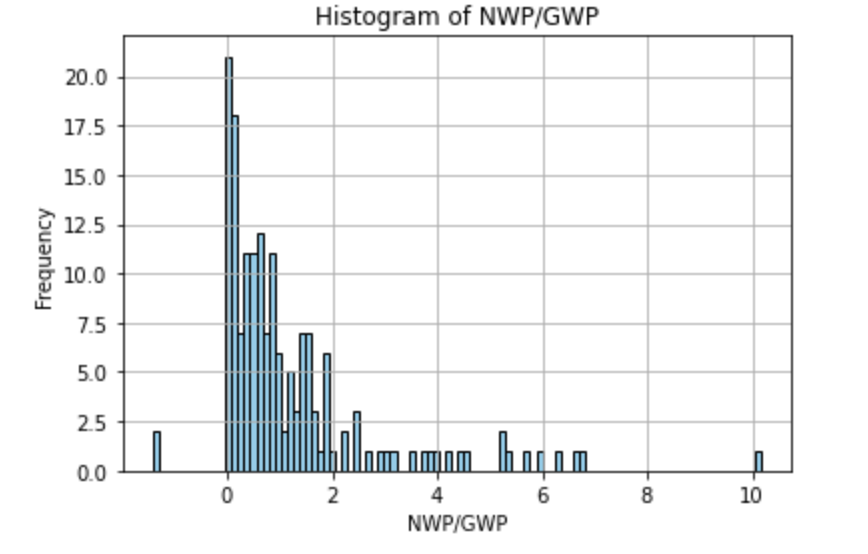
For SCR coverage ratio, threshold = 3 STD

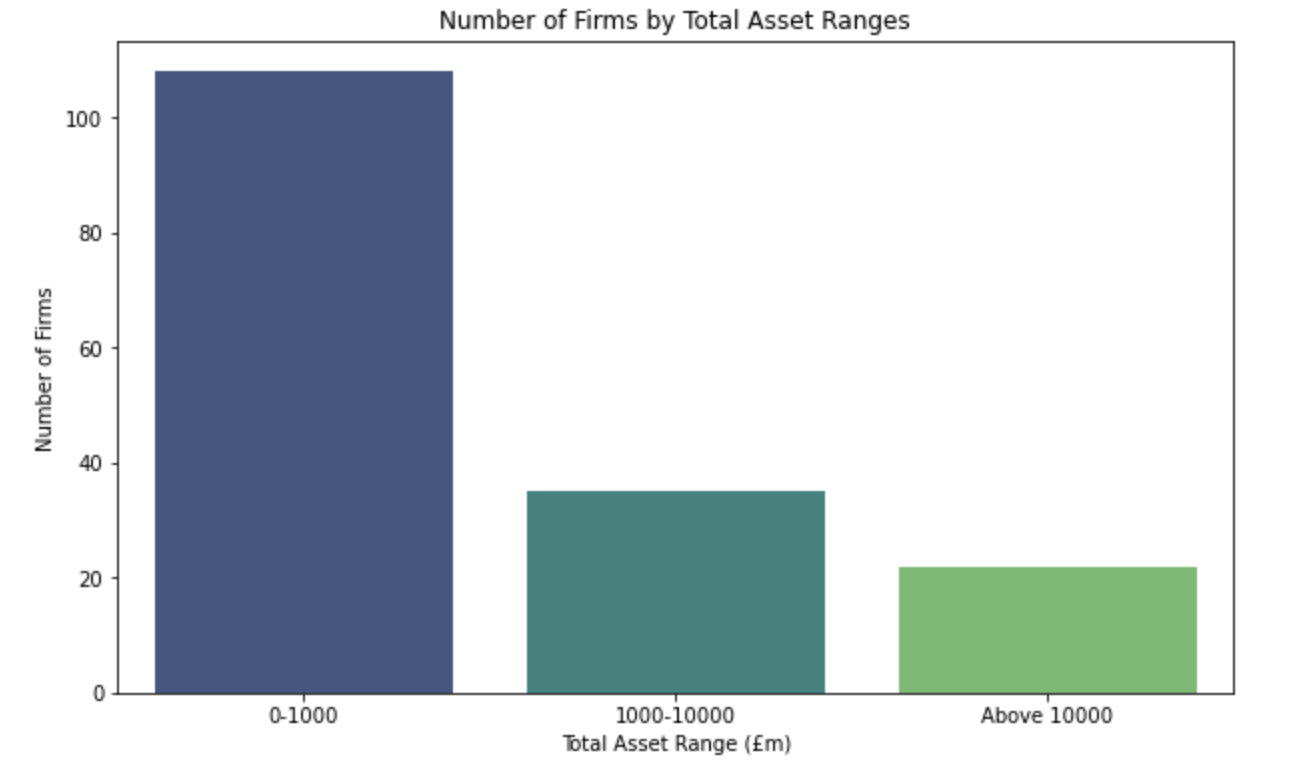
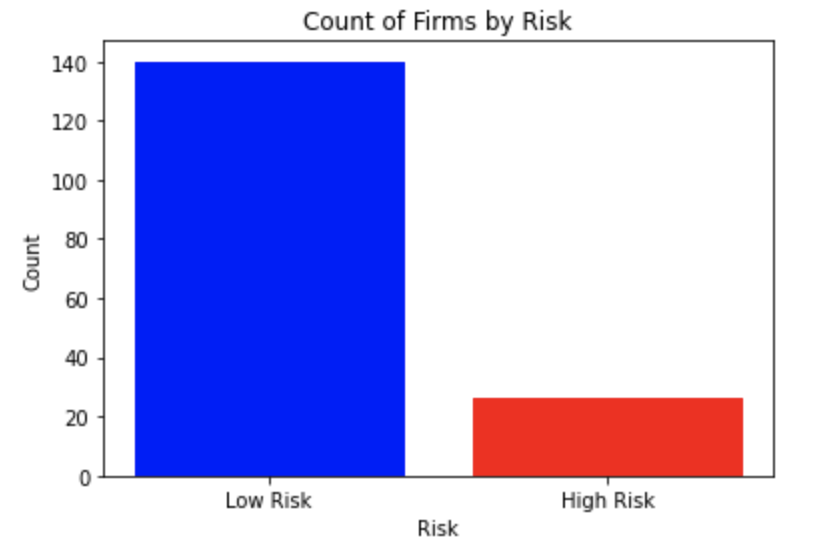
Finally, we got total 166 firms, I think data is consistent to perform advance analysis.

**Exploratory data Analysis:**

We can see that there’s high correlation between NWP and GWP, and calculating NWP/GWP metrics makes analysis more robust.



Using the histogram analysis, I have categorised the Firm Size based on Total Assets, and Firms associated Risk with NWP/GWP.

**Firm Classification:**

Firm Size:

0-1000 -Small

1000-10000-Medium

Above 10000 - Large

Firm Risk Associated:

NWP/GWP < 2, low risk

NWP/GWP > 2, high risk

Firm Capital requirement

SCR coverage ratio < 1, low

SCR coverage ratio 1-2, As per

SCR coverage ratio > 2, high

Firm Profit,

Net combined ratio < 1, yes else no

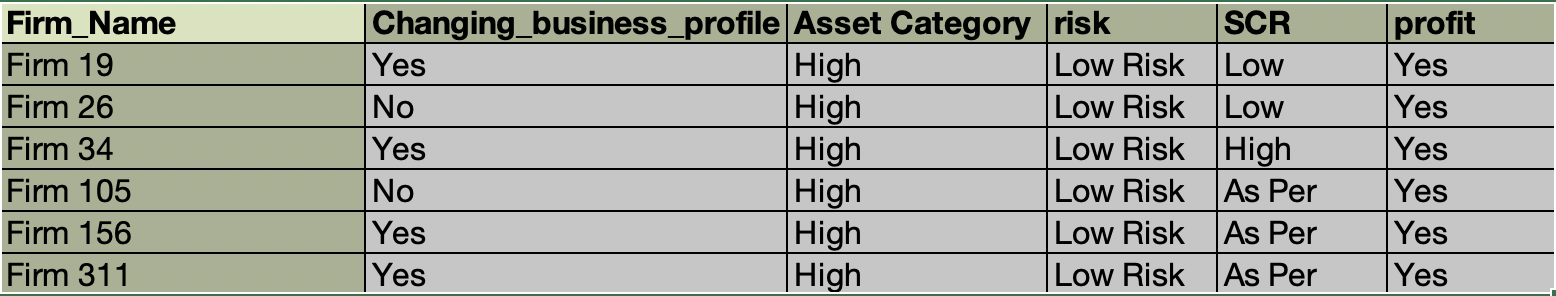
**Analysis of changing business profile:**

For this I have used, Gross Claimed incurred and analysed the year to year data for last 5 years and have categorised the Firms based on the threshold that if Year to year change have more than 500% then we will considered that business profile keeps changing. Of course, we can set the threshold as per the business requirement. Also, we can consider the other metrics for analysing the business profile change.

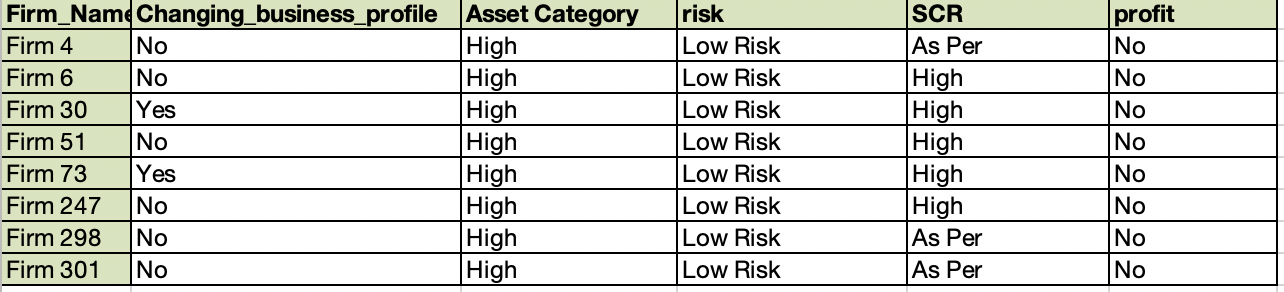


**Result: The Firms which needs most attention**

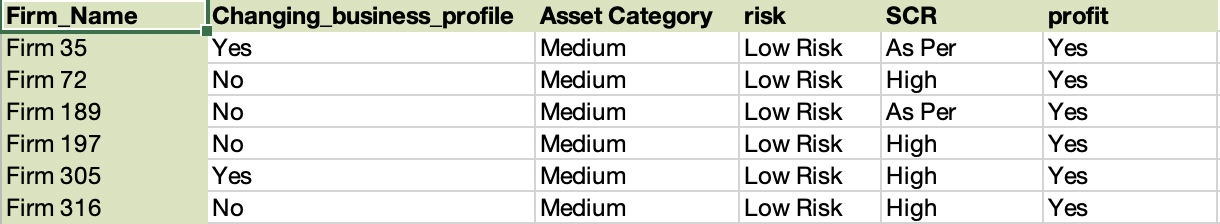
**Large Firms with Profit**

****

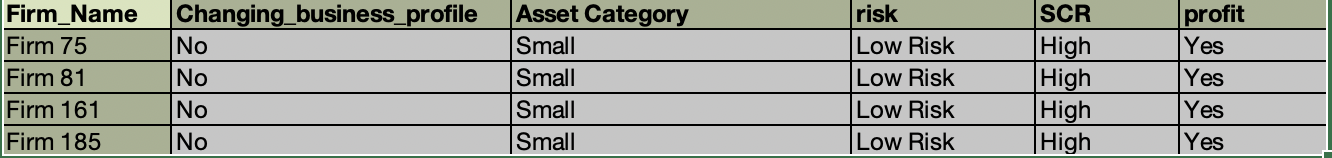
**Large Firms with No profit but low risk associated and meeting capital requirement**



**Medium Size Firms performing well in all metrics**

****

**Small Firms performing well and change in ‘Gross Claimed incurred’ below defined threshold meaning stable**

****

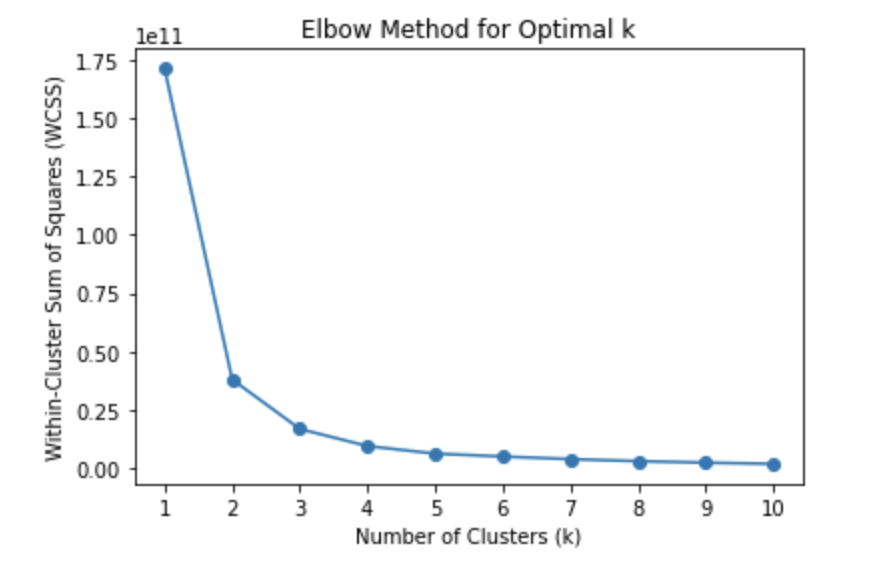
**Section 3**

While creating data pipelines and processing data into Microsoft SQL, I would consider tools such as Azure data factory, Azure Data lake gen 2, Azure Databrick and Power BI for advance visualisation.

1. Azure Data Factory: Using ADF, we can perform ETL transformation and extract the data from required resources and ingest into Azure storage. Also, we can implement data quality standard within ADF which includes, data exploration, data profiling and data validation. Once we make sure that data is as per standard quality, we can store this data into Azure data lake.
2. Azure Blob Storage: This can be utilised to store both structured and unstructured data for data analysis.
3. Azure Data Lake Storage: It is a data lake solution for big data analytics which can be used for both batch processing and real time workload. Since in this project data is batch processed daily, we will use Azure Data lake Storage.
4. Azure Databrick: Once the data is loaded is Azure data lake, I will use Azure data brick to work in collaborative environment with other data scientist and data engineer. I will use Apache Spark to query the data from Azure data lake storage and then utilise python notebook present within the Azure data brick to perform advance analytics building machine learning algorithm.
5. Azure Synapse Analytics: Then I will create dedicated SQL pools using Azure Synapse for running complex queries for advance analytics. Then I will integrate it with Azure Analysis Services to create semantic model so that BI capabilities can be planned further.
6. Data Visualisation and Reporting: Finally I will utilise Power BI platform to build BI dashboarding solution and automate the reporting process so that end user can have visual representation of the analysis and gain critical insights to make strategic changes.

**Annex:**

**Cluster Analysis:** Now I have performed cluster analysis on the total 166 Firms for we have consistent data. First, I have normalised the numerical data so that we can perform advance machine learning exercise. Elbow method shows that optimal clustering should be 3.



Using the Boxplot distribution plot, We can analyse the characteristics of each clusters, meaning what kind of firms belong to each of the cluster.

**Cluster 1** involves firms having total assets more than 100000m and most firms have low risk profile, profitable and enough buffer to meet capital requirements, but also have large cost to insurer. Further in depth analysis is required to identify low risk and high risk firms.

**Cluster 2** involves med size firms total assets ranging between 25000m to 75000m, involves low risk as well as high risk portfolio, low cost to insurer and are in profitable business.

**Cluster 3** involves small firms. This firms tends to have greater cost to insurer comparer to mid-size firm, also size of the buffer is good compared to mid-size firms. But few firms are also there which are not in profit.

**Note:** Further in depth analysis is required due to time constraint.

