INTERNSHIP @ MONETARY AUTHORITY OF SINGAPORE, INSURANCE DPT

CONVERTING RAW DATA TO CLEAN DATA, AND DISPLAYING IN TABLEAU DASHBOARD

DONE BY: CAI WENJING

SEPT 2019 - FEB 2020

Contents

1 Introduction

In Monetary Authority of Singapore, I was attached to the insurance department and worked under the Supervisory Analytics Division. During my 6 months of internship, I was exposed to many insurance and financial terms while doing ad hoc task related to research on impact of negative yield and low interest rate environment on global insurers and analyse survey on Risk Based Capital 2 Collective Investment Scheme Look-Through Approach. Furthermore, I was exposed into looking the insurance department dashboards and data as part of the project I am going to do.

About Company and Department

Company

The Monetary Authority of Singapore (MAS) is Singapore's central bank and financial regulatory authority. It administers the various statutes pertaining to money, banking, insurance, securities and the financial sector in general, as well as currency issuance.

As Singapore progressed, an increasingly complex banking and monetary environment required more dynamic and coherent monetary administration. Therefore, in 1970, the parliament of Singapore passed the Monetary Authority of Singapore Act leading to the formation of MAS on 1 January 1971. The act gives MAS the authority to regulate all elements of monetary policy, banking, and finance in Singapore.

Department

The Insurance Department supervises and regulates insurance companies and has as its primary objective the protection of policyholders' interests. The department adopts a risk-focused approach in the prudential and market conduct supervision of insurance companies. In its standards setting role, the department works closely with industry associations to promote the adoption of best practices by the industry.

2 Risk and Compliance Analytics Project

Project Introduction

Insurance department uses tableau dashboards to visualise financial information and risk metrics. As MAS uses single risk assessment system, called **Comprehensive Risk Assessment Framework and Techniques (CRAFT)** to assess the risks of a financial institution, the insurance department has been collecting CRAFT Ratings for years and they would want to have dashboards to visualise the ratings.

The scoresheet (diagram 1) used in the insurance department is an unstructured data. This data structure is easy for stakeholders to visualise when they fill up their CRAFT Ratings. However, due to its unstructured format, the machine cannot generate into a tableau dashboard. Thus, it is a must to generate a script to convert the unstructured data into a clean data. In clean data mode, it can do machine learning algorithms or do any kind of statistical analysis.

[Refer to Appendix A on CRAFT]

Project Objective

Design an end to end analysis of risk management from collecting raw data to visualizing of dashboards. From there on, dashboards are to be enhanced and maintained.

Project Process

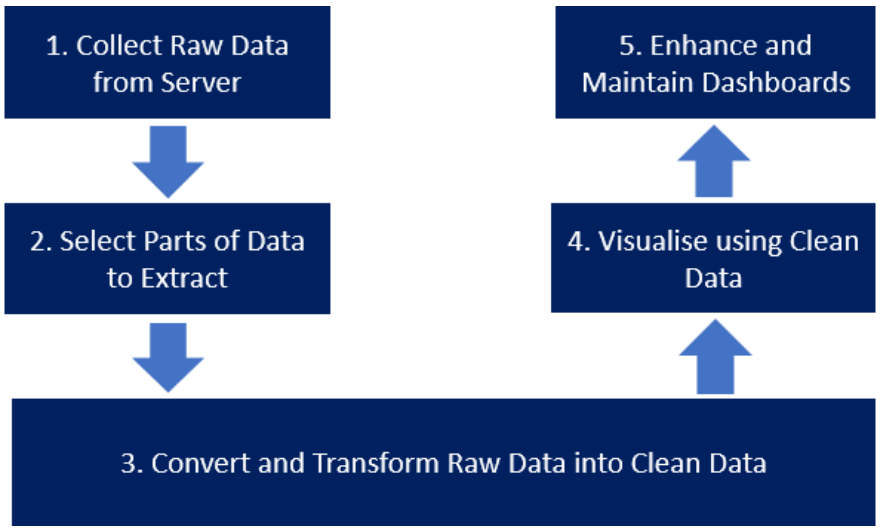


Diagram A

Initialisation of Project Process:

Step 1: Collecting raw data from server

This is the scoresheet, also known as the raw data, used by the insurance department to assess the CRAFT Ratings. After stakeholders are done filling up their CRAFT Ratings on the scoresheet, they will upload into a server. In total, I have collected about 100 scoresheets totalling 8000 rows which is about 64,000 data fields for year 2017 and 2018.

*As data in the scoresheet is confidential, only the layout of the scoresheet can be displayed in this report.

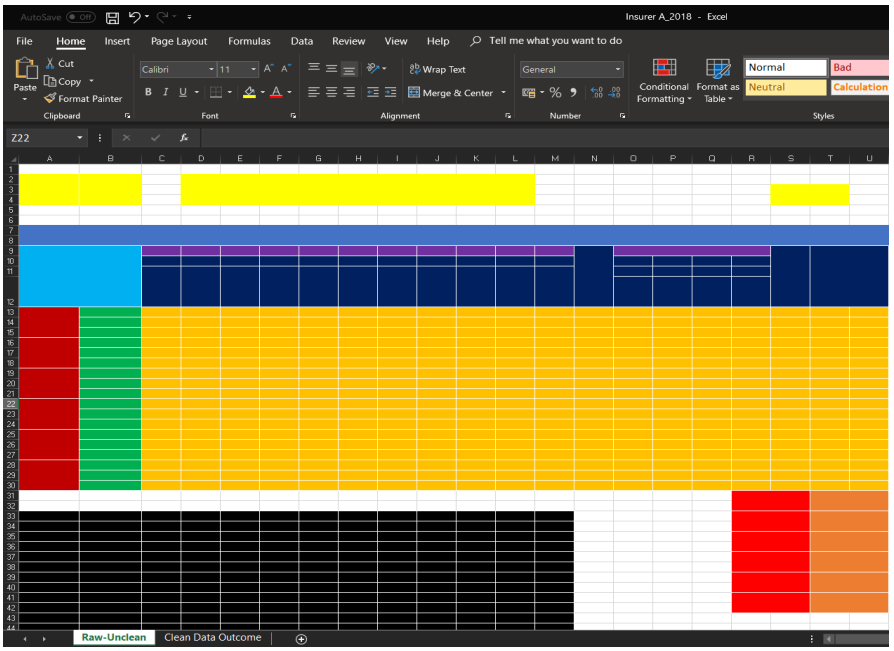


Diagram 1 - Scoresheet (Raw Data)

Step 2: Choosing parts of the data to be extracted

For this scoresheet, I will have to extract out different parts of the data due to business requirement.

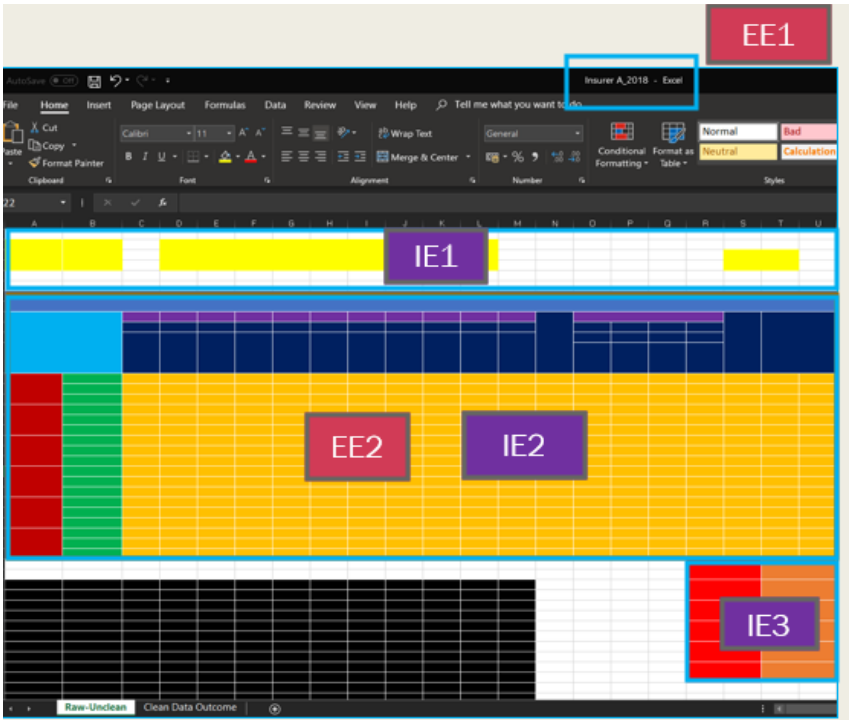


Diagram 2- Parts of the data to be extracted out

Classified into 2 types of Extraction:

Internal Extraction (IE): Extracts directly from the scoresheet

IE1: Extract cell value at a specific cell

IE2: Extract main data

IE3: Extract cell value while referring to a column

External Extraction (EE): Extracts indirectly from the scoresheet

EE1: Extract Year based on file name (InsurerName_Year.xlsx)

EE2: Extract Rank based on number of Significant Activities

Step 3: Converting and transforming raw data to clean data

Before converting from raw data to clean data, I will have to manually standardise the format and content of the scoresheet as the current scoresheet can be edited by the stakeholders. For example, they can add or remove rows and columns as well as change value of the cell. For empty cell, some are intended to be empty, and for those are not, I will have to input a value in it while referring to the CRAFT Framework.

For example:

Wrong cell value: (Affects some scoresheet)

Value edited in the scoresheet by Stakeholders	Standardised value in the scoresheet (Before the stakeholders fill up the ratings)
Institution Net Risk Score Institution Net Risk INSTITUTION NET RATINGS	INSTITUTION NET RISK SCORE
IR Ratings IR Rating IR RATINGS	IR RATING

Diagram 3A – Examples of wrong cell value

Data cleaning is the process of detecting and correcting inaccurate records from table and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the unclean data. – Extracted from Wikipedia on "What is Data Cleaning"

When converting raw data to clean data, I have used Jupyter Notebook which is an open source web as it allows me to visualise data and do data cleaning, data transforming and data extraction. Jupyter notebook supports python coding and python libraries. The libraries I have used were pandas, numpy, os and glob.

A data frame is a table or a two-dimensional array-like structure with columns and rows. It is identical to a SQL table. In jupyter notebook, I have declared a variable for this scoresheet called df, which is the short form for data-frame.

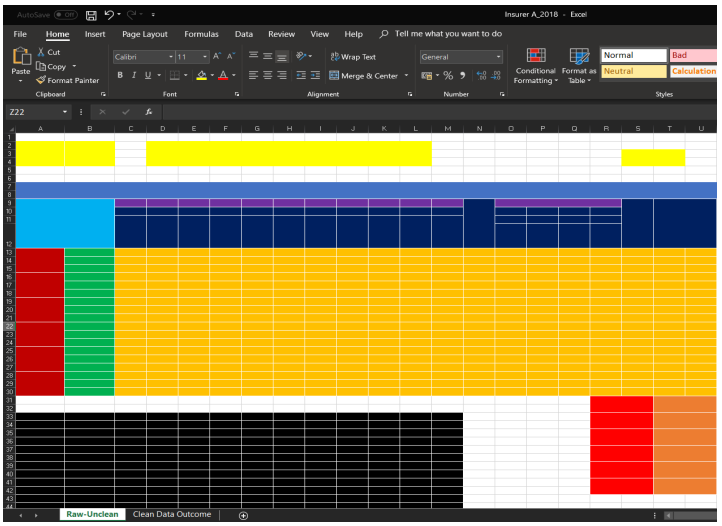
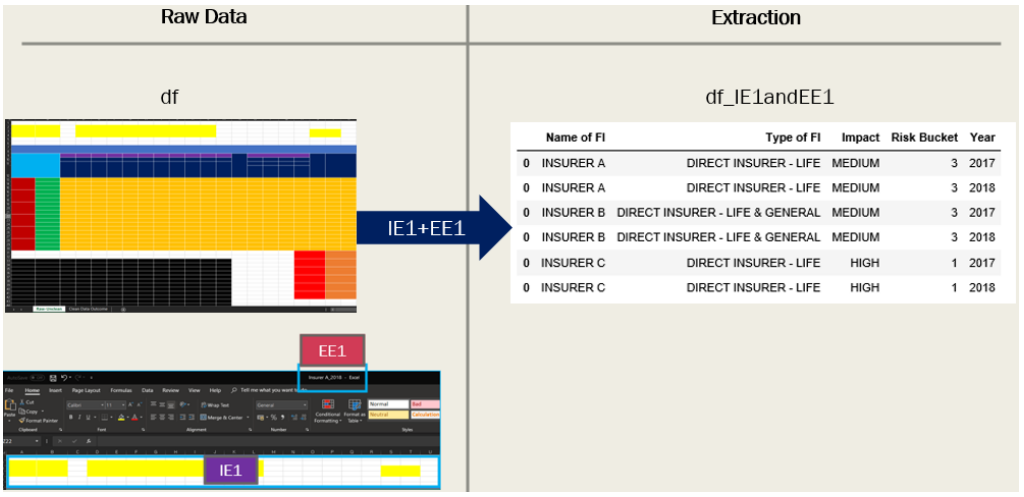


Diagram 3.1 - df

Using the df in diagram 3.1, IE1 and EE1 are extracted and it forms a working data frame.



	Name of FI	Type of FI	Impact	Risk Bucket	Year
0	INSURER A	DIRECT INSURER - LIFE	MEDIUM	3	2017
0	INSURER A	DIRECT INSURER - LIFE	MEDIUM	3	2018
0	INSURER B	DIRECT INSURER - LIFE & GENERAL	MEDIUM	3	2017
0	INSURER B	DIRECT INSURER - LIFE & GENERAL	MEDIUM	3	2018
0	INSURER C	DIRECT INSURER - LIFE	HIGH	1	2017
0	INSURER C	DIRECT INSURER - LIFE	HIGH	1	2018

Diagram 3.2 – IE1 and EE1 Data Extracted

In order to extract out the remaining data, I will have to remove IE1 because it is not in the correct column header and row to begin with. By using the raw data, I will truncate IE1 from df to form df_noIE1.

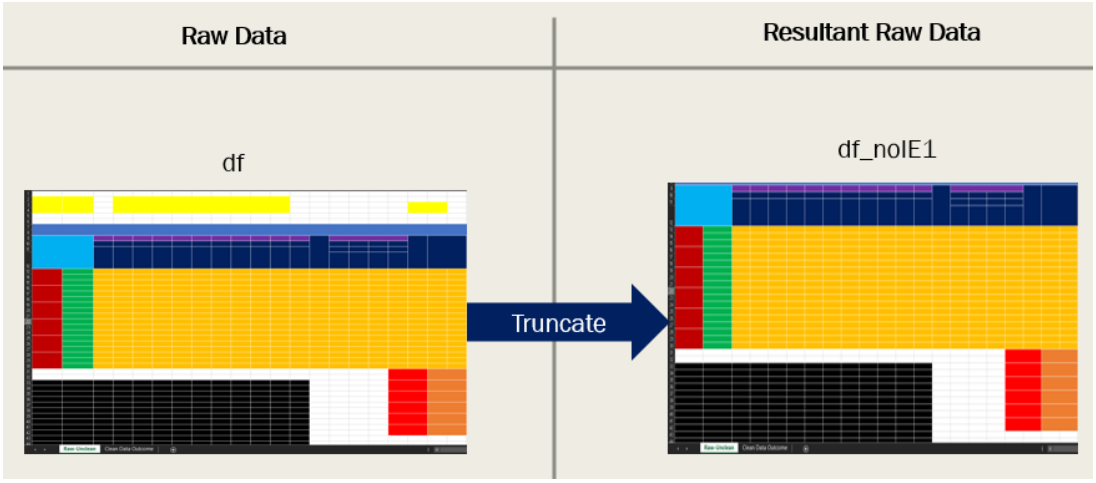


Diagram 3.3 – Remove IE1 to extract out remaining IEs and EE2

To begin extracting out the remaining data, I have developed 4 possible conditions. They serve the purpose to extract out the rest of the data accurately.

1. **FillSA** (known as Fill empty cell in Significant Activity Column) Replicates the value above as shown in the red circle.

   Due to data validation that empty cell must be dropped, I will have to fill up the empty cell under the Significant Activity column so that the CRAFT rating can be extracted.

Fill SA	
Initial	After
SA1	SA1
	SA1
INVESTMENT	INVESTMENT
	INVESTMENT
MOTOR	MOTOR
	MOTOR

Diagram 3.4(A) – Example of FillSA

2. **NoFillSA** □ Values will not get replicated.

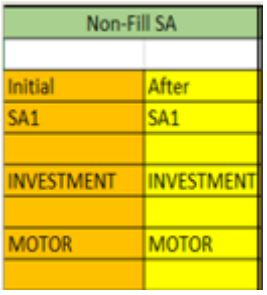


Diagram 3.4(B) – Example of NoFillSA

3. **Shift** (known as shift cell position) □ To shift cell value by 1. (Due to business requirement)

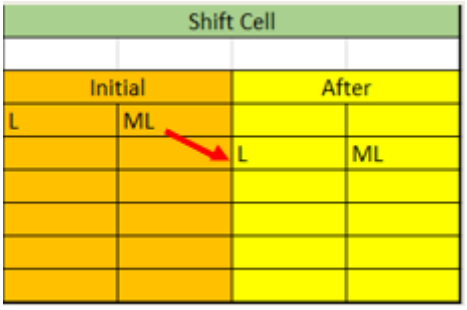


Diagram 3.4(C) – Example of Shift cell position

4. **NoShift** □ Data remain the same as it is.

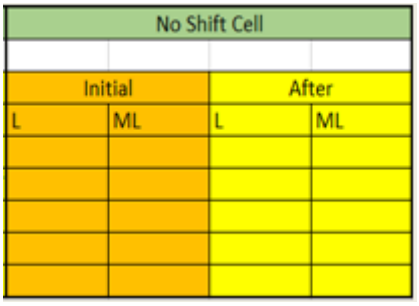


Diagram 3.4(D) – Example of No Shift cell

By using df_noIE1 in diagram 3.3, I have created 3 variables from the conditions provided in diagram 3.4(A), 3.4(B), 3.4(C), 3.4(D). They are df_FillSA_Shift, df_NoFillSA_Shift and df_NoFillSA_NoShift. These variables with conditions will overwrite the selected value in df_noIE1.

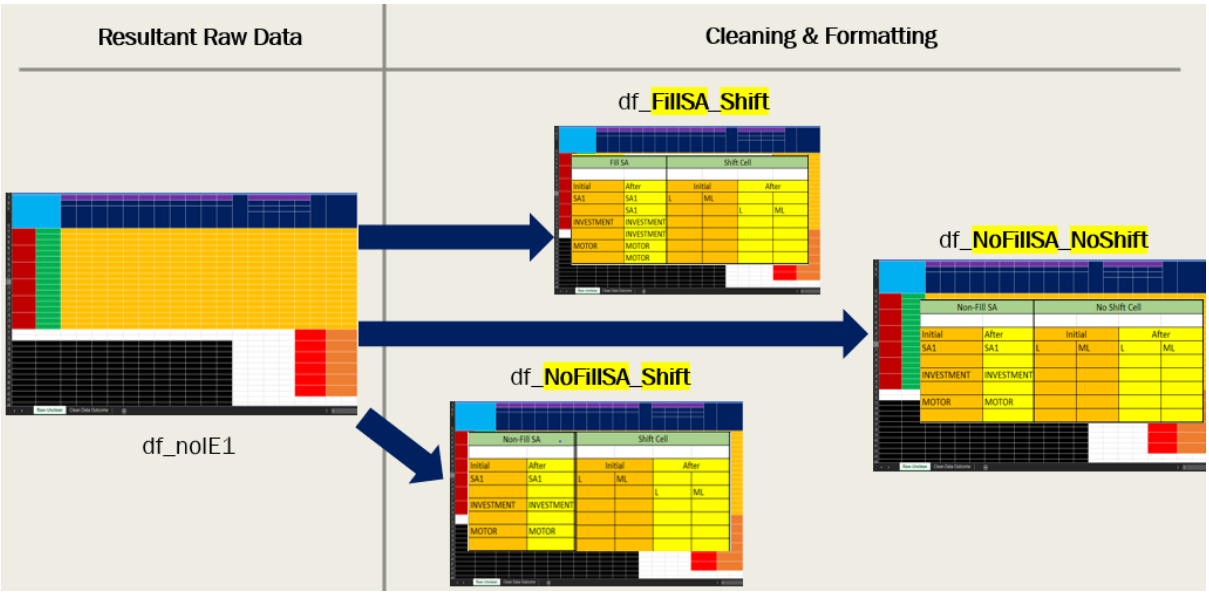


Diagram 3.4 – Create each variable for different conditions

Using **FillSA** and **Shift** (df_FillSA_Shift), IE2 and EE2 are extracted and it forms a working data frame.

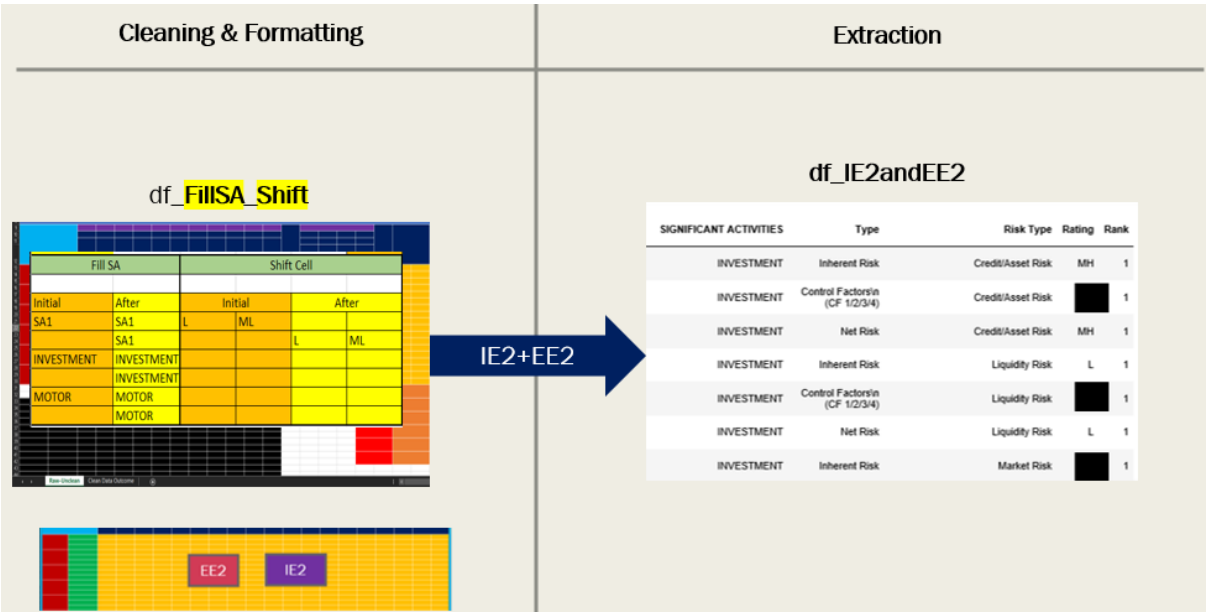


Diagram 3.5 – IE2 and EE2 Data Extracted

Using **NoFillSA** and **Shift** (df_NofillSA_Shift), a part of IE2 is extracted and it forms another working data frame.

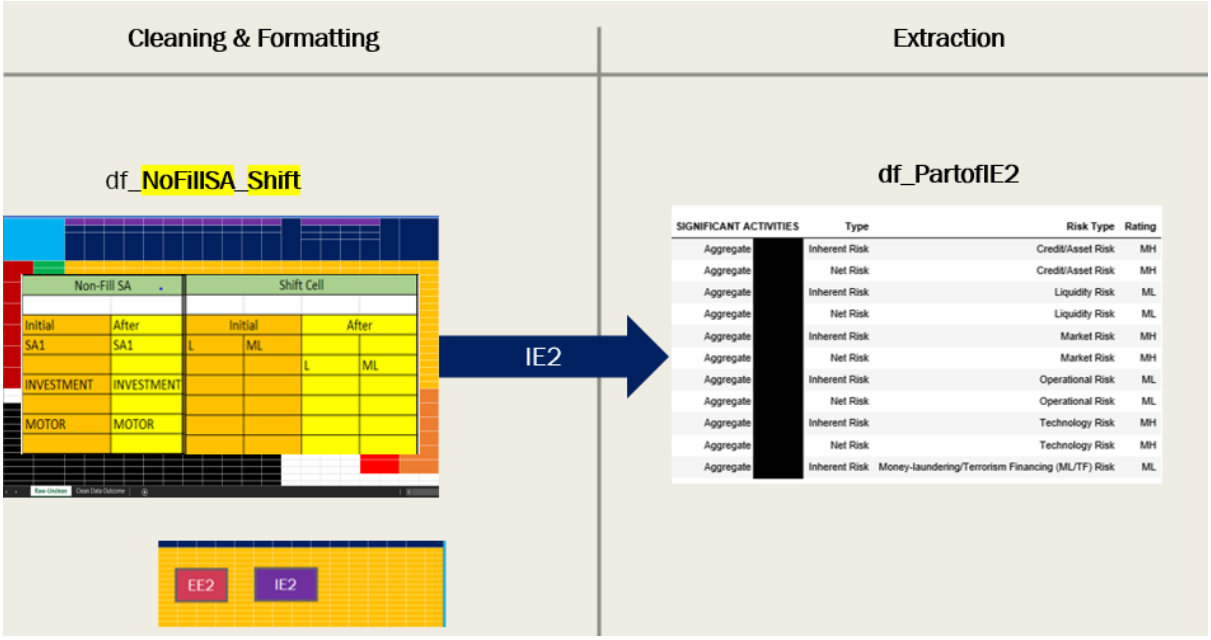


Diagram 3.5 – part of IE2 Data Extracted

Using **NoFillSA** and **NoShift** (df_NofillSA_NoShift), IE3 is extracted and it forms the last working data frame.

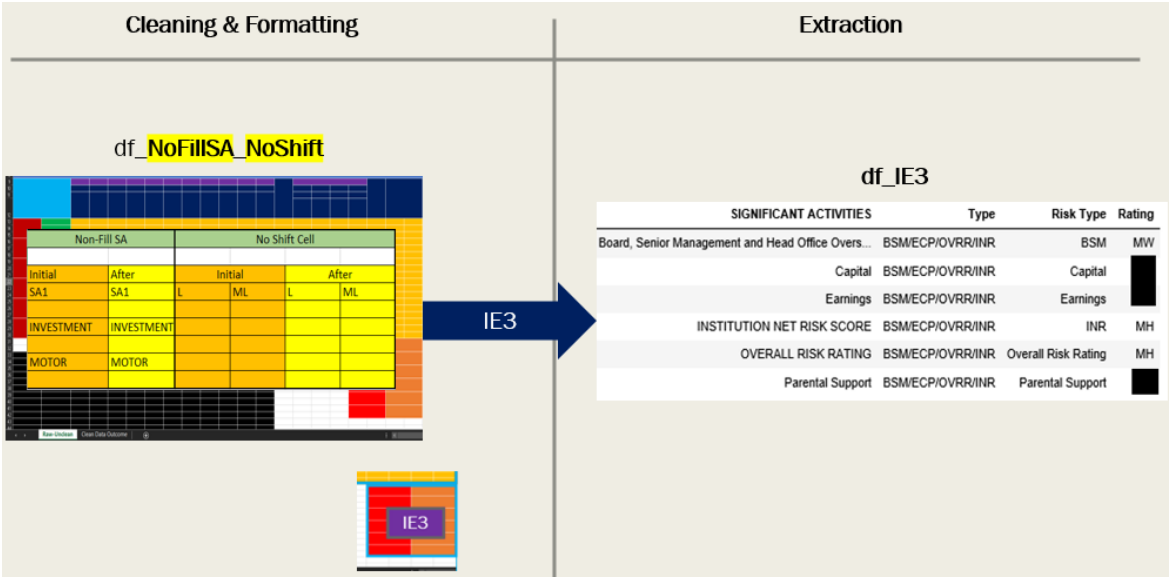


Diagram 3.6 – IE3 Data Extracted

After all the working data frames are developed, with careful insertion (python code), they are merged into data frame called df_cleandata which is in a clean data format.

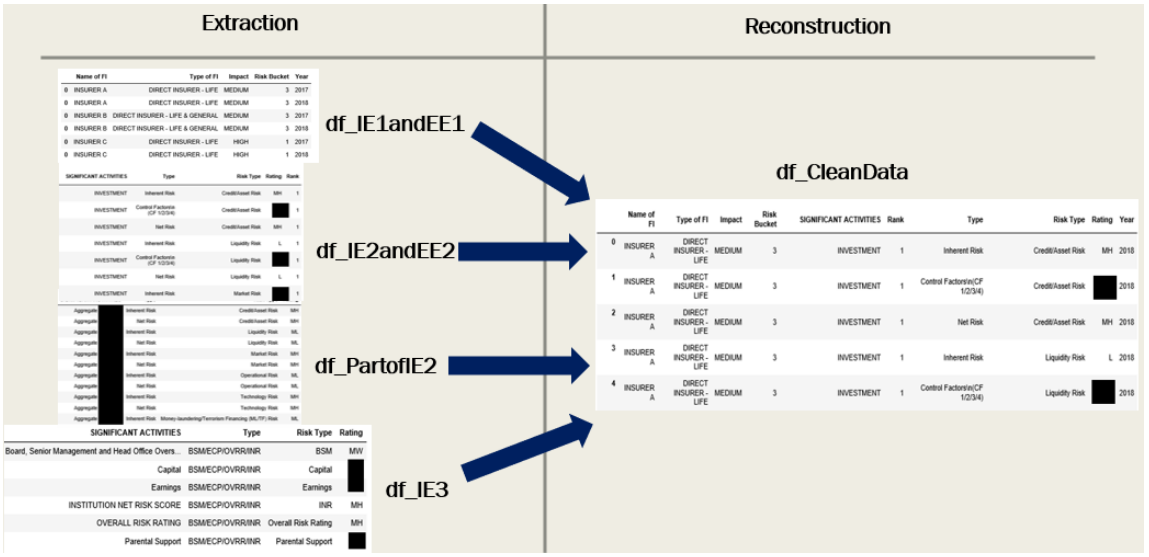


Diagram 3.7 – Clean Data is form

As df_cleandata in diagram 3.7 is stored as a memory, I will have to export out from jupyter notebook to an excel file containing 18000 data fields. By comparing the raw data which has 64000 data fields and clean data which has 18000 data fields, the clean data has the most accurate record.

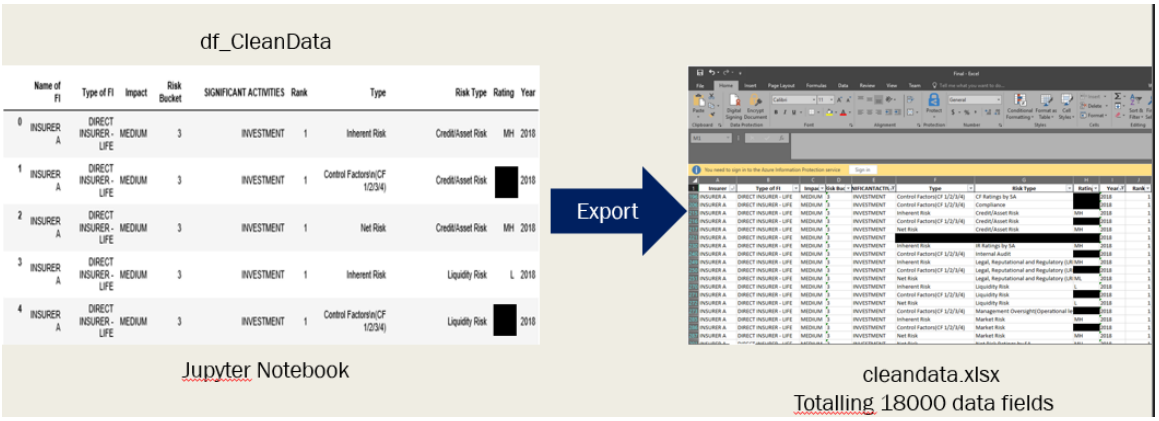
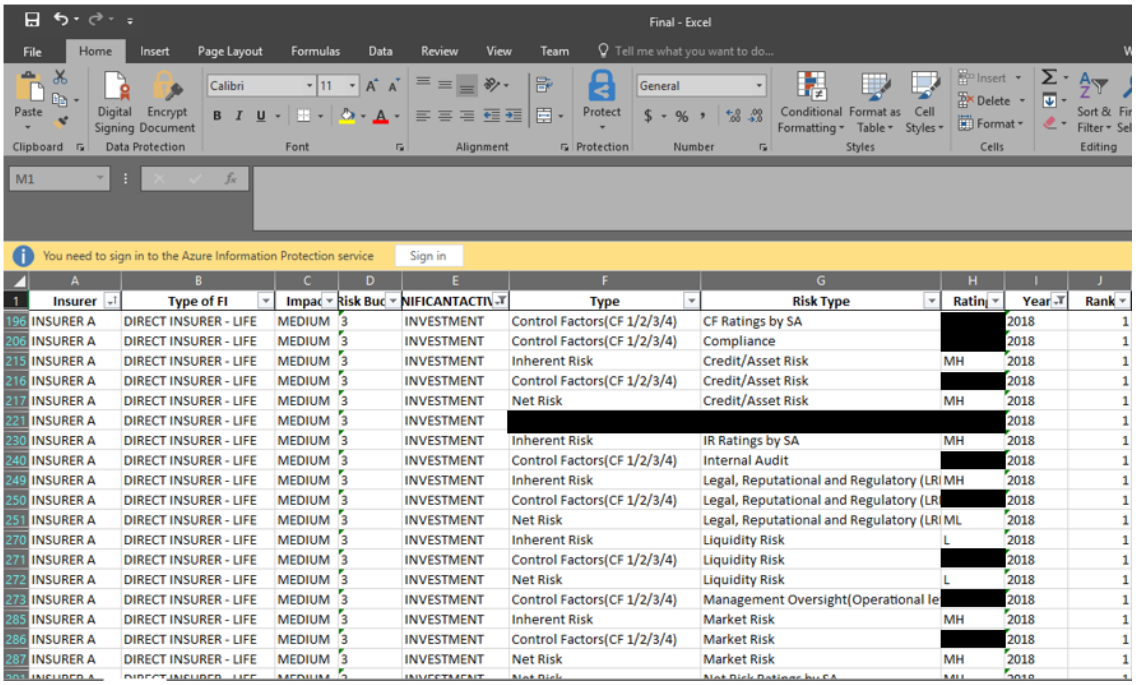


Diagram 3.8 – Export from jupyter notebook to cleandata.xlsx

Clean Data:



	A	B	C	D	E	F	G	H	I	J
1	Insurer	Type of FI	Impac	Risk Buc	NIFICANTACTI	Type	Risk Type	Ratin	Year	Rank
196	INSURER A	DIRECT INSURER - LIFE	MEDIUM	3	INVESTMENT	Control Factors(CF 1/2/3/4)	CF Ratings by SA	[redacted]	2018	1
206	INSURER A	DIRECT INSURER - LIFE	MEDIUM	3	INVESTMENT	Control Factors(CF 1/2/3/4)	Compliance	[redacted]	2018	1
215	INSURER A	DIRECT INSURER - LIFE	MEDIUM	3	INVESTMENT	Inherent Risk	Credit/Asset Risk	MH	2018	1
216	INSURER A	DIRECT INSURER - LIFE	MEDIUM	3	INVESTMENT	Control Factors(CF 1/2/3/4)	Credit/Asset Risk	[redacted]	2018	1
217	INSURER A	DIRECT INSURER - LIFE	MEDIUM	3	INVESTMENT	Net Risk	Credit/Asset Risk	MH	2018	1
221	INSURER A	DIRECT INSURER - LIFE	MEDIUM	3	INVESTMENT	[redacted]			2018	1
230	INSURER A	DIRECT INSURER - LIFE	MEDIUM	3	INVESTMENT	Inherent Risk	IR Ratings by SA	MH	2018	1
240	INSURER A	DIRECT INSURER - LIFE	MEDIUM	3	INVESTMENT	Control Factors(CF 1/2/3/4)	Internal Audit	[redacted]	2018	1
249	INSURER A	DIRECT INSURER - LIFE	MEDIUM	3	INVESTMENT	Inherent Risk	Legal, Reputational and Regulatory (LR	MH	2018	1
250	INSURER A	DIRECT INSURER - LIFE	MEDIUM	3	INVESTMENT	Control Factors(CF 1/2/3/4)	Legal, Reputational and Regulatory (LR	[redacted]	2018	1
251	INSURER A	DIRECT INSURER - LIFE	MEDIUM	3	INVESTMENT	Net Risk	Legal, Reputational and Regulatory (LR	ML	2018	1
270	INSURER A	DIRECT INSURER - LIFE	MEDIUM	3	INVESTMENT	Inherent Risk	Liquidity Risk	L	2018	1
271	INSURER A	DIRECT INSURER - LIFE	MEDIUM	3	INVESTMENT	Control Factors(CF 1/2/3/4)	Liquidity Risk	[redacted]	2018	1
272	INSURER A	DIRECT INSURER - LIFE	MEDIUM	3	INVESTMENT	Net Risk	Liquidity Risk	L	2018	1
273	INSURER A	DIRECT INSURER - LIFE	MEDIUM	3	INVESTMENT	Control Factors(CF 1/2/3/4)	Management Oversight(Operational le	[redacted]	2018	1
285	INSURER A	DIRECT INSURER - LIFE	MEDIUM	3	INVESTMENT	Inherent Risk	Market Risk	MH	2018	1
286	INSURER A	DIRECT INSURER - LIFE	MEDIUM	3	INVESTMENT	Control Factors(CF 1/2/3/4)	Market Risk	[redacted]	2018	1
287	INSURER A	DIRECT INSURER - LIFE	MEDIUM	3	INVESTMENT	Net Risk	Market Risk	MH	2018	1

Diagram 3.9 – Clean data format

Clean Data:

Using the excel clean data file (diagram 3.9), I proceeded to clean it for the 2nd time as there were issues with the selected data.

1. **Business Requirement:** (Affects all the data)

First extracted information from scoresheet to clean data:

Type	Risk Type
Inherent Risk	Compliance
Inherent Risk	Compliance
Inherent Risk	Compliance
Inherent Risk	Compliance
Inherent Risk	Compliance
Inherent Risk	Compliance

Diagram 3.10(A) – Extracted information from the scoresheet reflected in the clean data

By looking at the chart below, diagram 3.10(B). Compliance falls under control factors and not inherent risk.

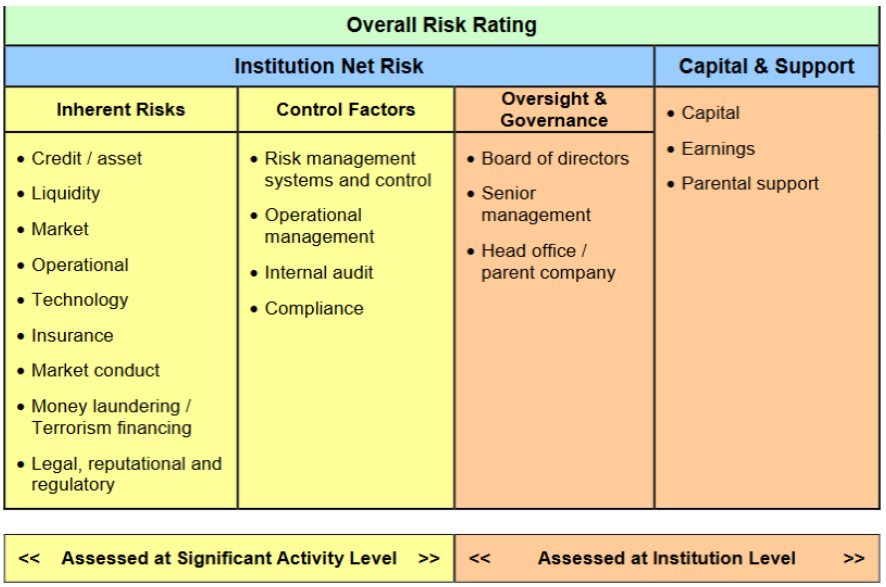


Diagram 3.10(B) – Risk Type Chart

Solution: As this affects all the data and this is considered a business requirement, I will have to solve it by code.

Type	Risk Type
Control Factor(CF 1/2/3/4)	Compliance
Control Factor(CF 1/2/3/4)	Compliance
Control Factor(CF 1/2/3/4)	Compliance
Control Factor(CF 1/2/3/4)	Compliance
Control Factor(CF 1/2/3/4)	Compliance
Control Factor(CF 1/2/3/4)	Compliance

Diagram 3.10(C) – Corrected version due to business requirement

2. **Entity Resolution** (Affects some of the data)

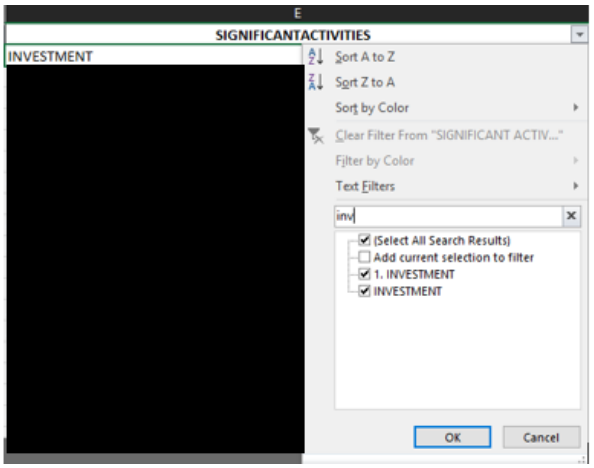


Diagram 3.11(A) – Multiple Entries of Significant Activities with the same entity

Stakeholders inputted value in Significant Activities column	Standardised Significant Activities
SA1. Investment INVESTMENT Investment investment	Investment
SA1. MOTOR Motor motor	Motor

Diagram 3.11(B) – Values inputted by the stakeholders for Significant Activity

Solution: Manually replace all the same Significant Activity with different entities into one entity.

Step 4: Visualising using tableau dashboard

Using the fully cleaned clean data in Diagram 3.10 and 3.11, I will import into the tableau database.

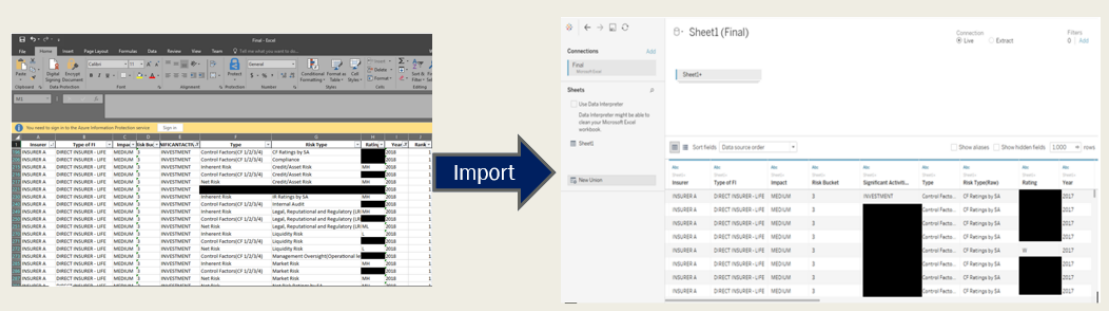


Diagram 4 – Importing into tableau database

With the data in the tableau database, I have crafted out 3 dashboards:

1. **Stacking Worksheet**

Stacking worksheet is used to benchmark and compare the risk rating among insurers. This view allows stakeholders to assess the risk for different types of insurers: Life Insurer, General Insurer and Reinsurer.

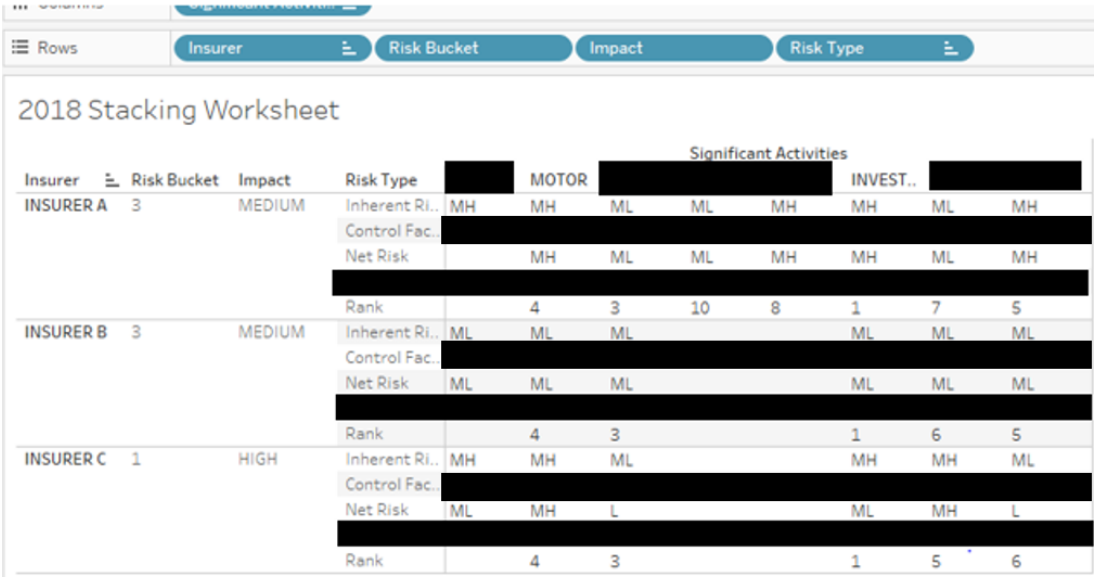


Diagram 4.1 - Stacking Worksheet

2. 100% Stacked Bar Chart

100% stacked bar chart is designed to show relative percentage of multiple data series in the stacked chart, where the total cumulative of each stacked bar is always 100%.

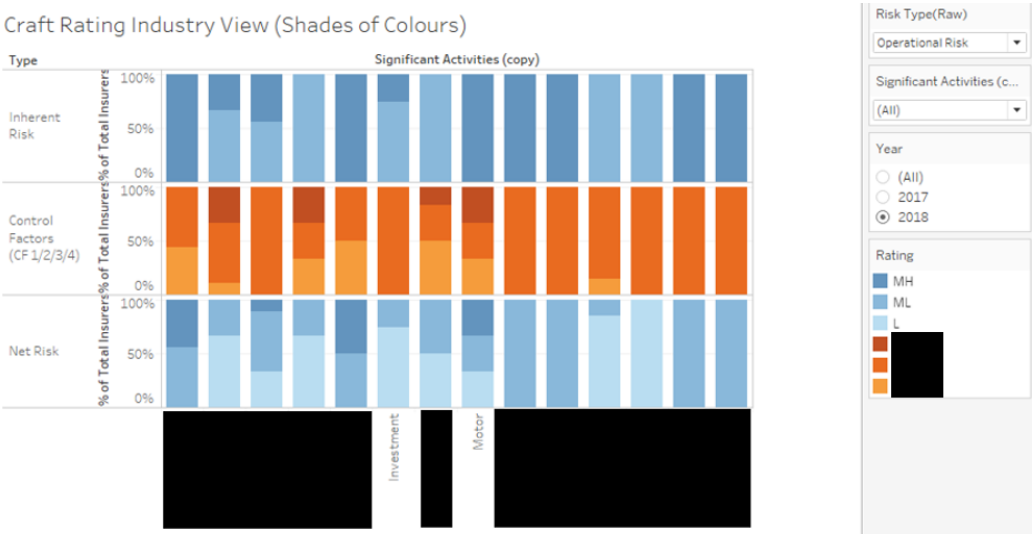


Diagram 4.2 - 100% Stacked Bar Chart

3. Pivot Table

A pivot table helps to summarise data in a quick and easy manner. This view compares all insurers against a significant activity.

Diagram 4.3 – Pivot Table Worksheet

Step 5: Maintaining and Enhancing of Dashboards

After developing the dashboards, we want to make the dashboards more user-friendly and intuitive.

For Stacking Worksheet:

1. Ordering for Insurance Companies

The ordering for insurance companies is for viewing the relevance and importance to today's market in the insurance industry. For example, if I want insurer B to be first, followed by insurer C and lastly insurer A. The simple way is to create an excel file to store the order for the insurer.

	A	B	C	D
1	Order	Insurer	Year	Life/Gen/Re Stackin
2	1	INSURER B	2018	LIFE
3	2	INSURER C	2018	LIFE
4	3	INSURER A	2018	LIFE
5				
6				

Diagram 5.1(A) – Excel file contains ordering for insurers

This is the outcome when the diagram 5.1(A) is imported into the tableau database:

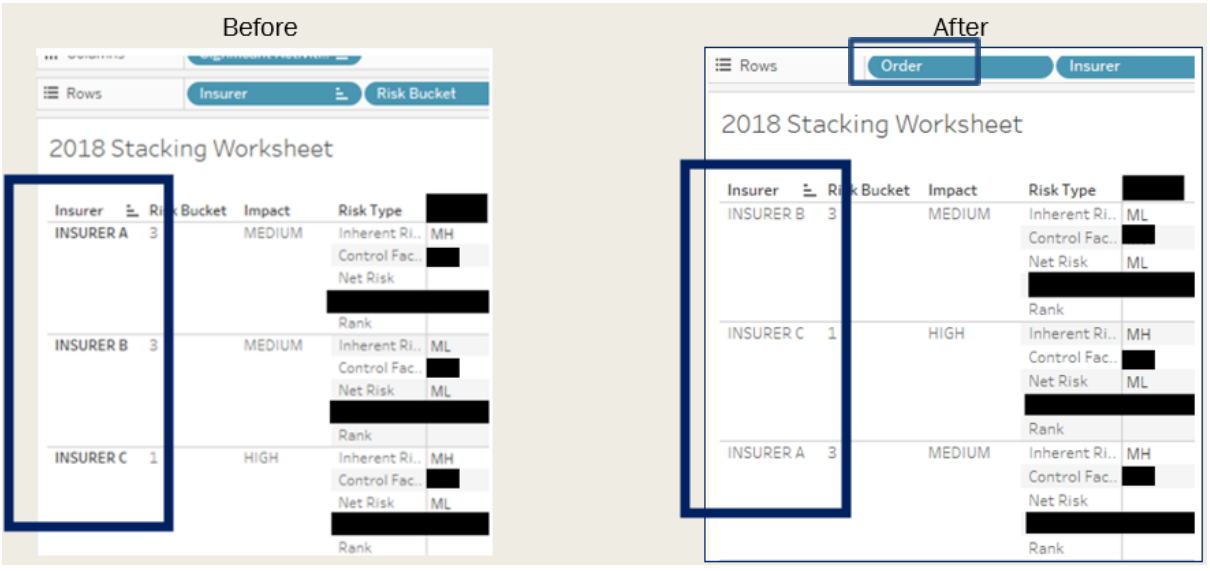


Diagram 5.1(B) – Ordering for insurer name

2. Year on Year (YoY) Rating Change

It is to see if the ratings will deteriorate, improve, or stay constant over years.

When the stakeholders want to view the changes for the rating over years, it is not intuitive for them to toggle between years repeatedly. Furthermore, the colour of the texts is the same, and it can be quite a hassle for them if they were to compare many ratings.

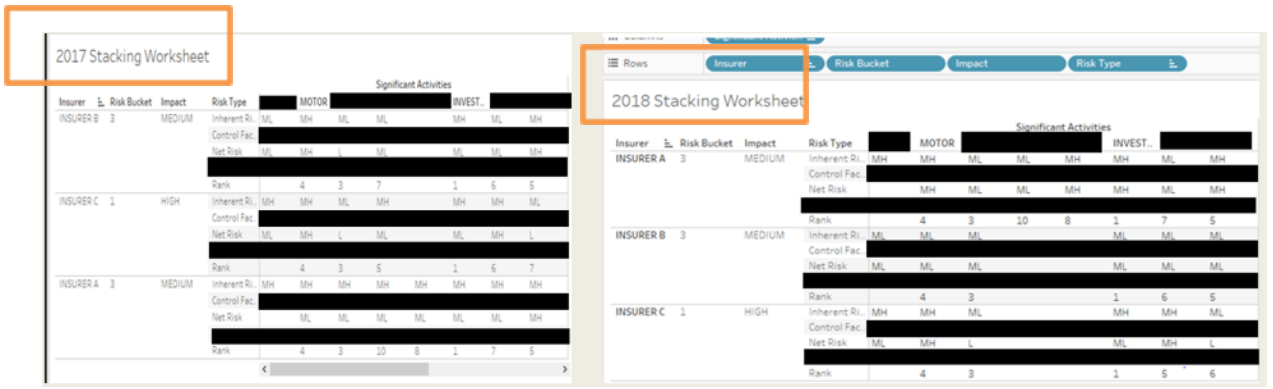


Diagram 5.2(A) – Toggle between year to see rating change

To make the stakeholders viewing easier, I have implemented a colour system on year rating change so that they would not need to toggle between years. This is comparing between 2017 and 2018 where ratings in 2018 will have the colour change taking place.

Red □ Deteriorate

Green □ Improve

Black □ No change/ Constant

Diagram 5.2(B) – Colour system to see rating change

For 100% Stacked Bar Chart:

1. Implement tooltip to show insurer name

When hover into one of the bars in the bar chart, there were no insurance companies name in the tooltip, and it can be difficult for the stakeholders to identify the insurance companies although there is number of insurers shown.

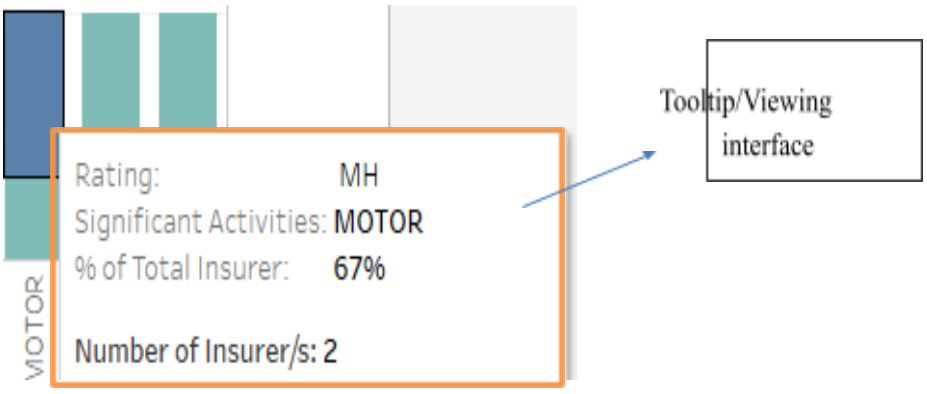


Diagram 5.3(A) – Current tooltip information

After all the codes are implemented, all the names of the insurance companies are listed out when hover over to the bar.

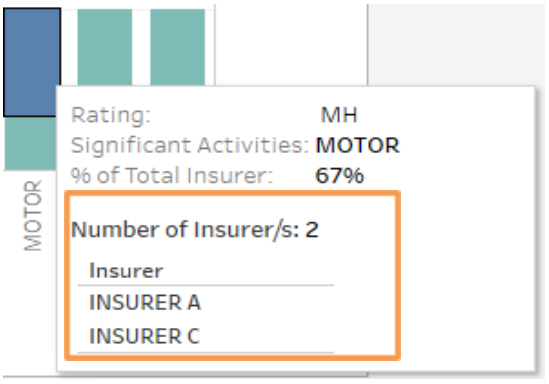
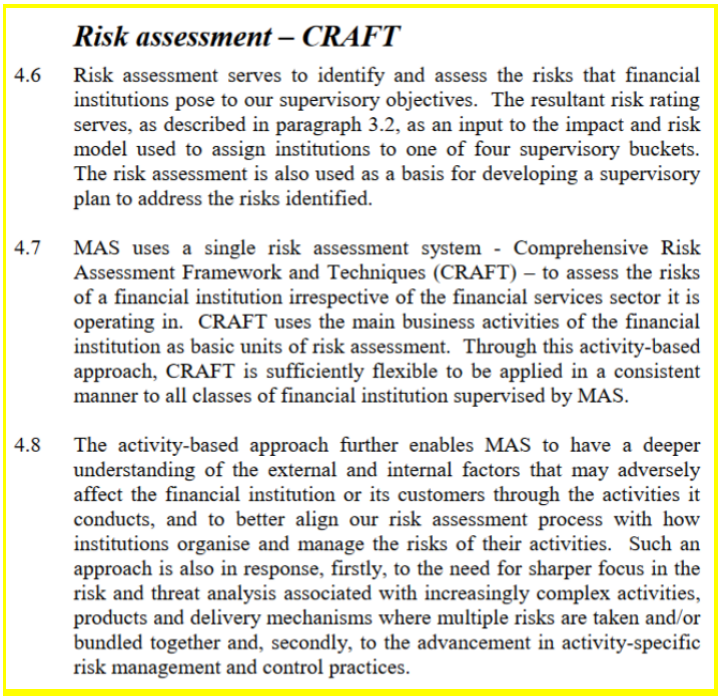


Diagram 5.3(B) – After implementing the tooltip on insurer name

Appendices

Appendix A – MAS CRAFT Framework



Risk assessment – CRAFT

4.6 Risk assessment serves to identify and assess the risks that financial institutions pose to our supervisory objectives. The resultant risk rating serves, as described in paragraph 3.2, as an input to the impact and risk model used to assign institutions to one of four supervisory buckets. The risk assessment is also used as a basis for developing a supervisory plan to address the risks identified.

4.7 MAS uses a single risk assessment system - Comprehensive Risk Assessment Framework and Techniques (CRAFT) – to assess the risks of a financial institution irrespective of the financial services sector it is operating in. CRAFT uses the main business activities of the financial institution as basic units of risk assessment. Through this activity-based approach, CRAFT is sufficiently flexible to be applied in a consistent manner to all classes of financial institution supervised by MAS.

4.8 The activity-based approach further enables MAS to have a deeper understanding of the external and internal factors that may adversely affect the financial institution or its customers through the activities it conducts, and to better align our risk assessment process with how institutions organise and manage the risks of their activities. Such an approach is also in response, firstly, to the need for sharper focus in the risk and threat analysis associated with increasingly complex activities, products and delivery mechanisms where multiple risks are taken and/or bundled together and, secondly, to the advancement in activity-specific risk management and control practices.

Extracted from:

https://www.mas.gov.sg/-/media/MAS/News-and-Publications/Monographs-and-Information-Papers/Monograph--MAS-Framework-for-Impact-and-Risk-Assessment.pdf