

Analytics Capstone Project

Himani Dave

GMG Data Visualization

June 28th, 2019



# Introduction

GMG stands for Global Mining Guidelines group. It is “a network of like-minded mining companies driving change in the mining industry.” They develop “multi-disciplinary guidelines to manage risk, enhance performance, design solutions, inform planning and decision making, and save time and money”(“What we do for the Mining Industry,” n.d.). The company hosts various events for collaboration between these companies and has workgroups that work on different guidelines.

The primary method used by the company for storing this data is the Highrise CRM system. Highrise contains “Contacts” which includes companies as well as individuals. Each contact has multiple tags. Each tag represents an event or a workgroup as well as other information such as whether the individual or contact is a member of GMG or not, whether they have been a speaker in any of the events etc. Contacts can have multiple tags associated with them.

The key problems that the company faced was populating their scorecard. GMGs scorecard contains information about the various projects, the companies that have been working on those projects, the events held for those projects, and the number of employees attending the various events for a particular project. Populating this scorecard manually was a tedious task which is why they wanted a better solution to this problem. Another problem was to calculate engagement of the companies with GMG. There was no way for them to quantify the engagement levels.

The key questions that needed to be answered were as follows:

* Which companies have the maximum participation in the various events of GMG?
* What are the events that each company has attended?
* What workgroups are the various companies a part of?
* How many employees, and from which companies have participated in the various events of GMG?
* What is the change in the engagement level of the companies through the years 2017, 2018, and 2019?

# Methods

**Data collection:**

Data was mainly collected in the form of access to the Highrsie CRM system. GMG wanted a dynamic dashboard in PowerBI which is why it was required to import all the data in PowerBI from Highrise using a connector. CData Software provides data access and connectivity solutions(“CData Software - About Us,” n.d.) in the form of drivers that connect one system to another. A CData PowerBI Highrise connector was used for the purpose. There were a lot of technical issues with the connector which took up a lot of time. The company later realized that there was a glitch in the driver and fixed it during week 3 of the project.

Data was imported to PowerBI once the connector was setup. However, Highrise CRM does not export relationships, which came to light once the data was imported in PowerBI using the connector. Because of this, the tags associated with each contact were not visible in PowerBI. Since direct connection to PowerBI did not yield the desired data, contacts were directly exported from Highrise in the form of an Excel file.

**Data Cleaning:**

Initially, R programming was used to clean the data, but since I found it difficult, not having learned complicated data cleaning in R, I switched to Python. The raw data exported from Highrise looked like below:

|  |  |  |
| --- | --- | --- |
| Name | Company or Person | Tag1, Tag2, Tag3, Tag4, Tag5,… |
| Name 1 | Company | Tag1, tag2, tag3,… |
| Name 2 | Person | Tag1, tag2, tag3, tag4… |
| … |  |  |

In order to clean the data, two excel files were created from the exported contacts, one for data analysis and another for data visualization. The file for data analysis, “conatcts\_py” used for data analysis had the following columns:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Name | Company or Person | Tag1 | Tag2 | Tag3 | Tag4 | Tag5 | … |
| Name1 | Company | Tag1 | Tag2 | Tag3 | NaN | NaN | … |

The other file that was used for data analysis was “conatcts\_visual.py”. The snapshot of the file is as follows:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Company or Person | Company | Name | Tag1 | Tag2 | Tag3 | Tag4 | … |
| Person | XYZ Company | Name1 | Tag1 | Tag2 | NaN | … | … |

The above file contained information about individuals as well as companies, the companies that the individuals worked for, and the tags associated with them.

To clean this data, these files were imported into Jupyter notebooks using Python and were first converted into data frames. Pandas library was used for both data cleaning and analysis. The NaNs were removed from the files and replaced with spaces so as to make the analysis easier. Relevant data was then extracted to perform the analysis.

**Data Analysis:**

Data analysis was carried out using Python in Jupyter notebooks. The above data frames were transformed into other data frames as required to answer the key questions. The data frames were created based on the different types of events. They were further divided by years. These data frames were then exported in the form of Excel files.

One of these files that contained statistical information about the number of events attended by the company, the number of forums, the number of workshops all of which were further divided by years, was imported into PowerBI to create a report.

# Analysis

The data analysis resulted into a number of data frames that were later exported in the form of excel files. The details about the resultant data frames are as follows:

1. Companies2017, Companies2018, Companies2019: These data frames contained a list of companies that had attended a particular event. Three such data frames were created for 2017, 2018, and 2019. Similar data frames were created for individuals as well.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Tags | Company1 | Company2 | Company3 | Company4 | … |
| Tag1 |  |  |  |  |  |
| Tag2 |  |  |  |  |  |
| … |  |  |  |  |  |

1. Evenst2017, Events2018, Events2019: These data frames contained information about the events that different companies attended. This includes all type of events held by the company except workgroups. Similar data frames were created for individuals.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Companies | Event1 | Event2 | Event3 | Event4 | … |
| Company1 |  |  |  |  |  |
| Company2 |  |  |  |  |  |
| … |  |  |  |  |  |

1. Similar data frames were created for Forums and Workshops for both, companies as well as individuals.
2. Workgroups were not separated by years, which is why two data frames were created listing all the companies that were part of a workgroup.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Companies | WG1 | WG2 | WG3 | WG4 | … |
| Company1 |  |  |  |  |  |
| Company2 |  |  |  |  |  |
| … |  |  |  |  |  |

1. ContactStats – This data frame contains statistical information about various companies. These stats are available for individuals in a separate data frame.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Company | No of workgroups | No of Forums 2017 | No of Forums 2018 | No of Forums 2019 | No of Workshops 2017 | No of Workshops 2018 | No of Workshops 2019 |
| Company1 |  |  |  |  |  |  |  |
| Company2 |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Company | No of Events 2017 | No of Events 2018 | No of Events 2019 | Total forums | Total Workshops | Total workgroups | Total events |
| Company1 |  |  |  |  |  |  |  |
| Company2 |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Company | Percent of total workgroups | Percent of total forums | Percent of total workshops | Percent of total events |
| Company1 |  |  |  |  |
| Company2 |  |  |  |  |
| … |  |  |  |  |

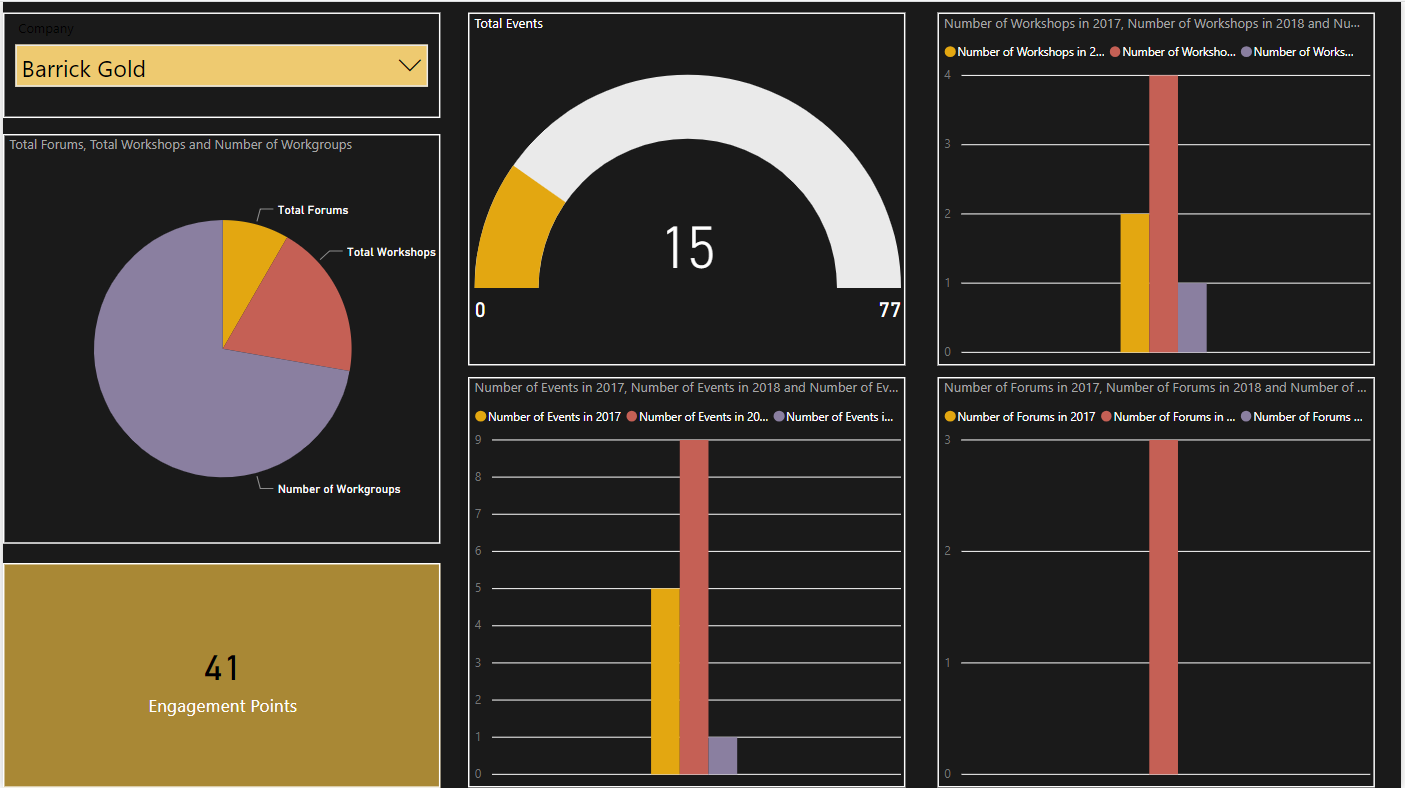
1. ContactAndCompanies – This data frame includes the name of the people, the company they are associated with and the events they have attended in 2019.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Name of the Person | Company | Tag1 | Tag2 | Tag3 | Tag4 | … |
| Person1 | Company |  |  |  |  |  |
| Person2 | Company |  |  |  |  |  |
| … |  |  |  |  |  |  |

1. NumberOfAttendees - This data frame contains the number of people attending a particular event by company. This only includes events of 2019.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Company | Tag1 | Tag2 | Tag3 | Tag4 | Tag5 | … |
| Company1 | No of people for Company1 |  |  |  |  |  |
| Company2 | No of people for Company2 |  |  |  |  |  |
| … |  |  |  |  |  |  |

Apart from these data frames, a PowerBI report was created to visualize the engagement of the companies. A snapshot of the report is as below:



In the report, the top left corner is a drop-down list of companies which controls all the charts. The pie chart shows the number of workgroups, the total number of forums, and the total number of workshops the company has participated in over the years. The engagement points, below the pie chart, are calculated by adding the number of workgroups to the total number of events attended by the company. The gauge shows the total number of events attended by a company out of the 77 events conducted so far by GMG. The three bar graphs show the total number of events, forums, and workshops in the years 2018, 2018, and 2019 that were attended by the selected company.

# Conclusion

Thus, all the above mentioned excel files, the PowerBI report, the data files, as well as the code were provided to the external partners. These provide the answer to all the key questions as was required by the external partners. Stackoverflow(“Stack Overflow - Where Developers Learn, Share, & Build Careers,” n.d.) has been my main source for getting help with code and troubleshooting the problems that I encountered while coding.

The project has been a learning curve for me which has helped me transition from a business analytics student to a true business analyst. It has helped me discover my potential and passion for data analysis and also given me an experience of working on real world data analysis projects. It has taught me how to deal with difficult and often long drawn situations and turn them around to find results.

# References

CData Software - About Us. (n.d.). Retrieved June 27, 2019, from CData Software website: https://www.cdata.com/company/default.aspx

Stack Overflow - Where Developers Learn, Share, & Build Careers. (n.d.). Retrieved June 27, 2019, from Stack Overflow website: https://stackoverflow.com/

What we do for the Mining Industry. (n.d.). Retrieved June 26, 2019, from Global Mining Guidelines Group website: https://gmggroup.org/about-us/gmg/