An Analysis of What Factors Correlate With Last Years Employee Performance Rating

Google Data Analytics Professional Certificate Capstone Project

The purpose of this analysis is to explore the this Human Resources dataset. I was interested in seeing what (if any) factors correlated with an employees previous year performance rating. The following report will include the steps I took to reach my conclusion. I will share foot notes when necessary to talk about any problems I had at any particular stage, and how I worked through it. Below is a list of skills and programs I used for this project:

1. Excel - Data cleaning
2. SQL (Postgresql) - Querying and data aggregation
3. R (R Studio) - Regression analysis and visualisations
4. Tableau - Additional visualisations

Ask

The business tasks is to explore what factor(s) correlated with the previous years performance scores of the employees in the dataset. The following variables will be of particular interest to help answer the business task: age, tenure, average training scores and the number of training sessions attended. I'll be using these dependent variables and analysing them against the independent variable (which will be the previous years performance rating).

The insights provided in this report will allow the business to clearly see where improvements can be made. For example, if

Prepare

The data allows me to answer the business questions because it has a numeric rating for their employee performance scores. Plus it has other information such as gender, age and department. Which allows one to deep dive into what variables correlated with an employees performance rating.

The dataset is updated annually on Kaggle, with the last update being in July 2023. This means that I am confident in the data as the information will not be outdated. Which would be an issue for the business if this wasn't the case, as the insights from this report will contribute to the wider business strategy.

The licence for the dataset falls under the Open Data Commons - Public Domain Dedication and Licence. Essentially, this allows the dataset to be used and modified freely. In this case, it is for education and practice purposes as part of my course.

These are the limitations with the data:

1. It is only one dataset, even though the initial spreadsheet has over 17,000 entries.
2. It is subject to human error with it's collection method.
3. The categories are very broad, e.g. one category is if over 80% of KPI's have been met, the dataset doesn't say what the KPI's were.

Process

In order to process the data, I have chosen to use Microsoft Excel. Not only am I familiar with the program, but this project will allow me to implement the skills I have learnt in the earlier modules of the Google Data Analytics course. From there, I did the following in Excel:

1. I duplicated the data then start cleaning the duplicated sheet.
2. Removed any columns I did not need for my analysis.
3. Check for any duplicates via VLOOKUP and using filters on all the columns.
4. Used the =LEN function to make sure cells in certain columns contained the correct length. E.g. 'Previous Year Rating' column was scored from 1-5, and no entries were in double digits.
5. Used the =COUNTIF function for the 'KPI's Met More Than 80' column. The results from this column followed the Boolean logic. Therefore, there should be no double digits as it's either 1 or 0.

6. Used the =PROPER function to change the ‘m/f’ results under the ‘Gender’ column to ‘Male/Female’.
7. Used the =TRIM function to get rid of excess spaces. I was particularly wary of any entries that would have an extra ‘space’ at the end, that could cause problems later on.
8. Deleted rows with blank entries by using ‘Find & Select’.
9. The original, uncleaned dataset had 17417 rows of data. After cleaning, it had 15424 rows of data.

Footnote: there are inefficiencies in a lot of the functions and steps I performed. However, I took the long route because I wanted to familiarise myself more with different Excel functions.

Analyse

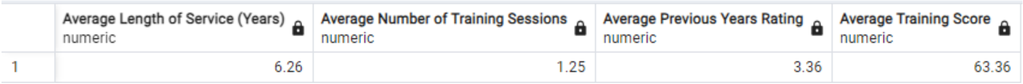
This section starts of with some general demographic information relating to the employees using mostly SQL but also R. Then a regression was performed in R with the selected dependent variables against the independent variable. Most of the results were predictable but there was one or two that was very surprising but sparks more questions.

SQL

Using Postgresql as my preferred application for SQL, I imported the cleaned CSV file.

I created a table to show the averages of certain columns which showed some good insights. The average performance rating was 3.36 or 67%; which is interesting because the average training score was 63%.

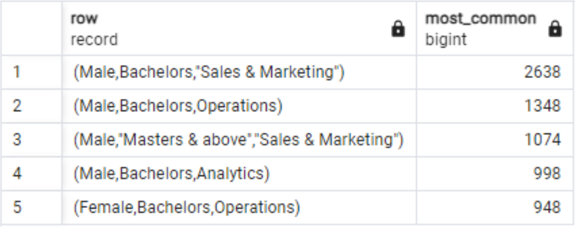
```
SELECT
ROUND(AVG(length_of_service), 2) AS "Average Length of Service (Years)",
ROUND(AVG(no_of_trainings), 2) AS "Average Number of Training Sessions",
ROUND(AVG(previous_year_rating), 2) AS "Average Previous Years Rating",
ROUND(AVG(avg_training_score), 2) AS "Average Training Score"
FROM
hr_analytics;
```



	Average Length of Service (Years) numeric	Average Number of Training Sessions numeric	Average Previous Years Rating numeric	Average Training Score numeric
1	6.26	1.25	3.36	63.36

With a clean dataset of over 14,000 entries, I was curious to know what the most common types of demographic. The query below produced a table that showed that the most common worker was a male, with a bachelor’s degree who worked in sales and marketing. Which was 1290 more than the second most common category. Another thing to note (as shown in the dashboard), is that the sales and marketing team collectively have the lowest performance score by department. Coming in at 3.11.

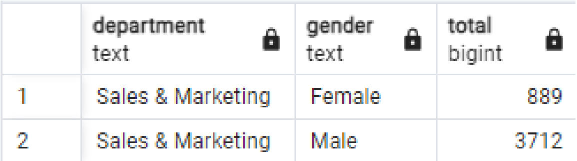
```
SELECT (gender, education, department), COUNT (*) AS most_common
FROM hr_analytics
GROUP BY (gender, education, department)
ORDER BY most_common DESC
LIMIT 5
```



	row record	most_common bigint
1	(Male,Bachelors,"Sales & Marketing")	2638
2	(Male,Bachelors,Operations)	1348
3	(Male,"Masters & above","Sales & Marketing")	1074
4	(Male,Bachelors,Analytics)	998
5	(Female,Bachelors,Operations)	948

I also wanted to find out the make up of the sales and marketing team. It is possible that they are more harshly marked than everyone else because the sales department has a lot of influence in how much money a company is making. If sales for a particular year or quarter are low, then the sales and marketing team may be the first department/factor that’s scrutinised.

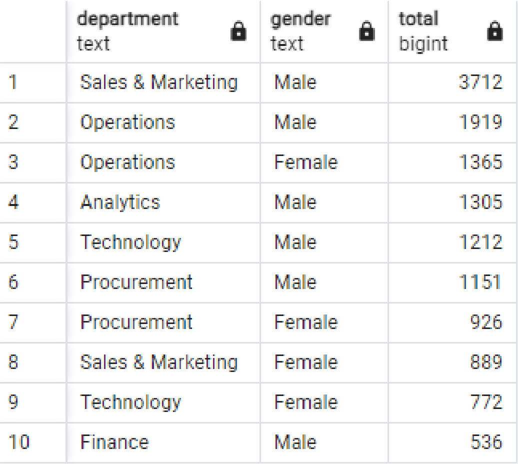
```
SELECT
department,
gender,
COUNT(*) AS total
FROM
hr_analytics
WHERE
department IN ('Sales & Marketing')
GROUP BY
department, gender;
```



	department text	gender text	total bigint
1	Sales & Marketing	Female	889
2	Sales & Marketing	Male	3712

The above table made me interested to see the overall gender distribution for all the departments. The operations and procurement team seem to have a better male-to-female ratio. However, visuals later in the report will validate that a gender balance or imbalance may not affect performance scores for a department. This is due to varying departmental performance scores: the operations team average rating was 3.66 (2nd highest). However, the procurement teams average was 3.24 (3rd worst).

```
SELECT
department,
gender,
COUNT(*) AS total
FROM
hr_analytics
GROUP BY
department, gender
ORDER BY
total DESC
LIMIT 10;
```



	department text	gender text	total bigint
1	Sales & Marketing	Male	3712
2	Operations	Male	1919
3	Operations	Female	1365
4	Analytics	Male	1305
5	Technology	Male	1212
6	Procurement	Male	1151
7	Procurement	Female	926
8	Sales & Marketing	Female	889
9	Technology	Female	772
10	Finance	Male	536

R

Using R Studio as my preferred application for R, I imported the cleaned CSV file.

I installed and loaded the following packages that'll be best suited to doing statistical analysis.

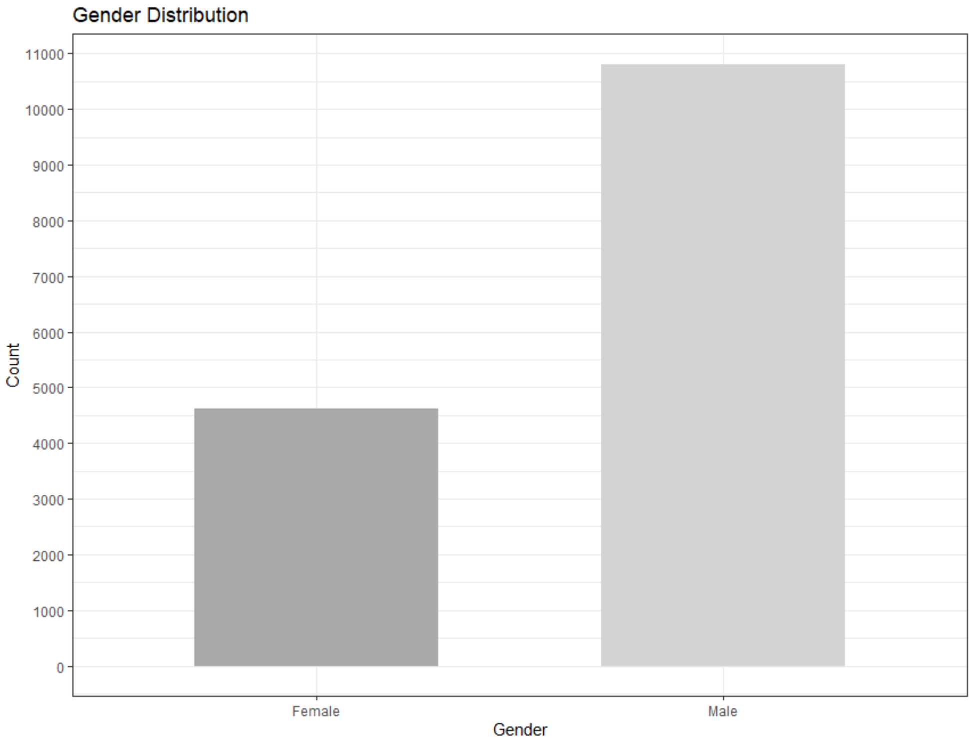
```
install.packages('ggplot2')
install.packages('tidyr')
library('ggplot2')
library('tidyr')
library('tidyverse')
```

I first wanted to see the gender totals that were in the dataset.

```
ggplot(data = Employees_dataset, aes(x = gender, fill = gender)) +
geom_bar(fill = custom_colors, width = 0.6) +
labs(title = "Gender Distribution",
x = "Gender",
y = "Count") +
scale_fill_manual(values = custom_colors) +
scale_y_continuous(breaks = seq(0, 12000, by = 1000))
```

I ran the below custom colours code to get the right colours for the bar chart.

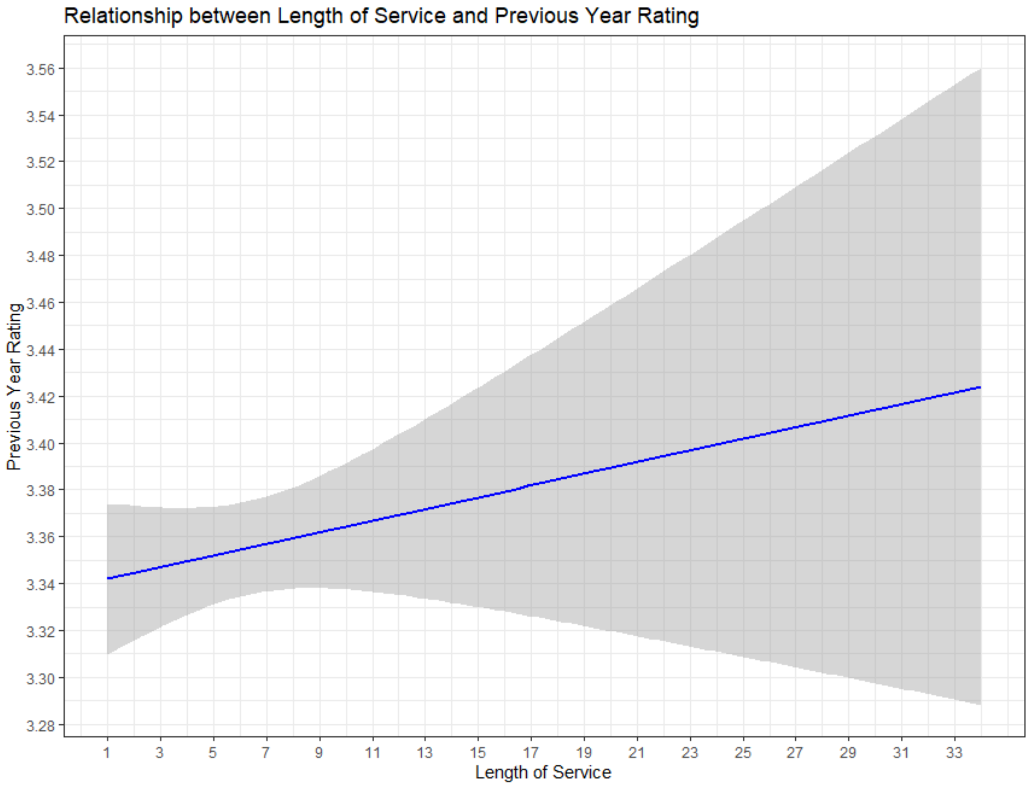
```
custom_colors <- c("Male" = "darkgrey", "Female" = "lightgrey")
```



With the basic demographic noted, it was time to move on to the analysis of the dependent variables. Below is a linear regression for the length of service and the previous year rating. The graph shows that there is a positive correlation between the two variables. It starts of very narrow for the first decade of service and then it widens by a lot as length of service continues.

The widening of performance ratings based on service after 10 years or so is compelling. Especially for the ratings that dip below 3.31, which is the lowest rating for staff with a only a year of service. This threshold is reached around 24+ years of service. This was another surprise in the dataset and it could be down to a few things: (1) an employee could have had a bad year, even if they have over 24 years of experience. (2) it is possible that there isn't enough of a distinction with staff ratings within their first few years of working.

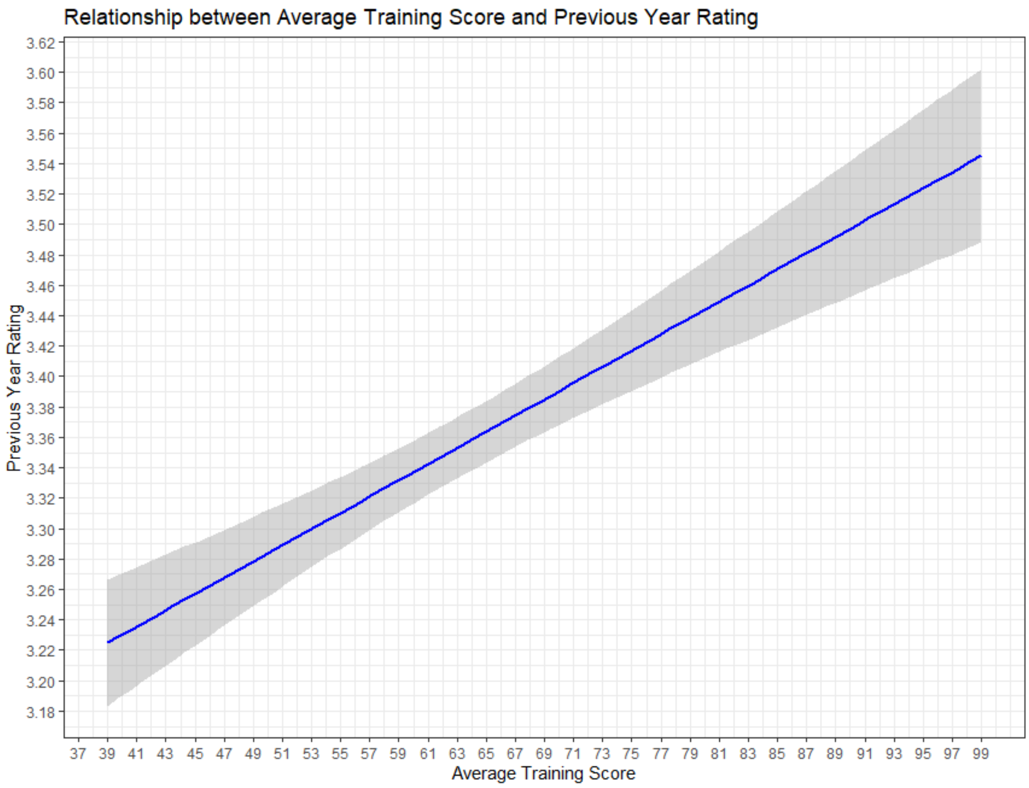
```
ggplot(data = Employees_dataset, aes(x = length_of_service, y = previous_year_rating)) +
geom_smooth(method = "lm", formula = y ~ x, color = "blue") +
labs(title = "Relationship between Length of Service and Previous Year Rating",
x = "Length of Service",
y = "Previous Year Rating") +
scale_x_continuous(breaks = seq(1, max(Employees_dataset$length_of_service), by = 2)) +
scale_y_continuous(breaks = seq(3.1, 5, by = 0.02))
```



The regression for average training score and previous years rating showed an extremely strong correlation. This can be a positive and negative that will need to be investigated further. On the positive side, this looks like the company has been able to align the training courses offered with an employees work. Resulting in better performance. However, (and this depends on many things including culture), a manager could look at an employee more favourably if they know they scored high during training. Resulting in an inflated performance rating that may not be correlated to the work they produced that year.

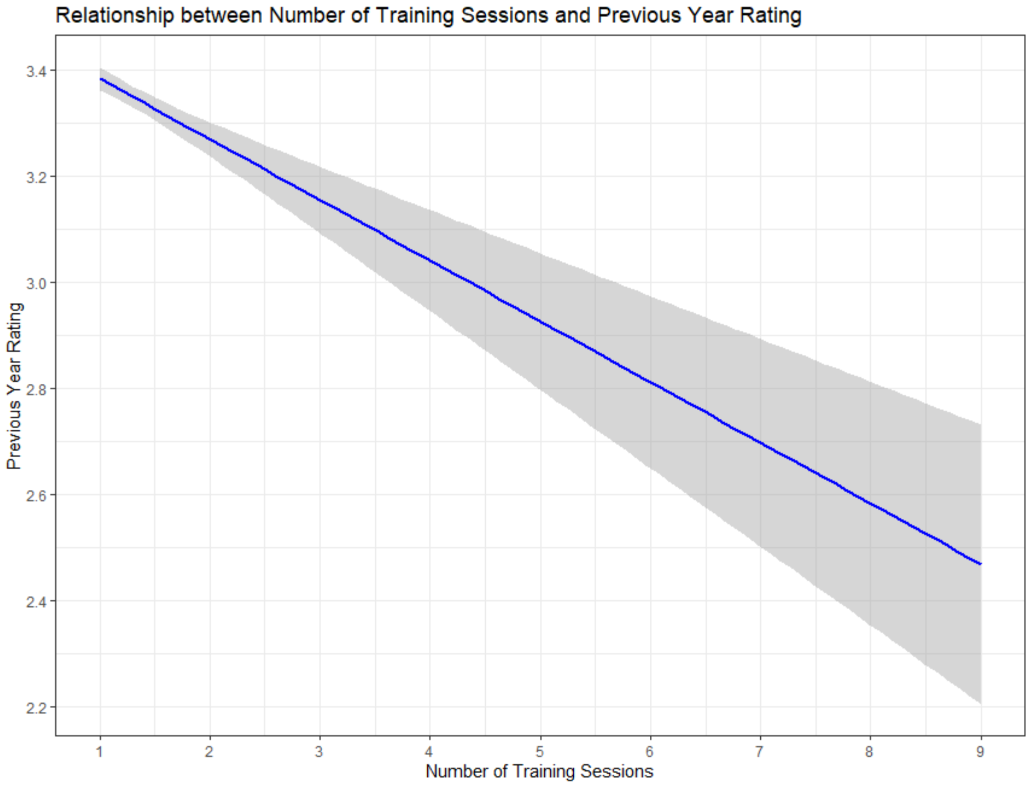
The graph shows a narrow range but that range widens at the extremes and is narrower in the middle. Which is around the 60's on the x-axis and 3.36 on the y-axis; highlighting the averages that were calculated using SQL.

```
ggplot(data = Employees_dataset, aes(x = avg_training_score, y = previous_year_rating)) +
geom_smooth(method = "lm", formula = y ~ x, color = "blue") +
labs(title = "Relationship between Average Training Score and Previous Year Rating",
x = "Average Training Score",
y = "Previous Year Rating") +
scale_x_continuous(breaks = seq(1, max(Employees_dataset$avg_training_score), by = 2)) +
scale_y_continuous(breaks = seq(3.1, 5, by = 0.02))
```



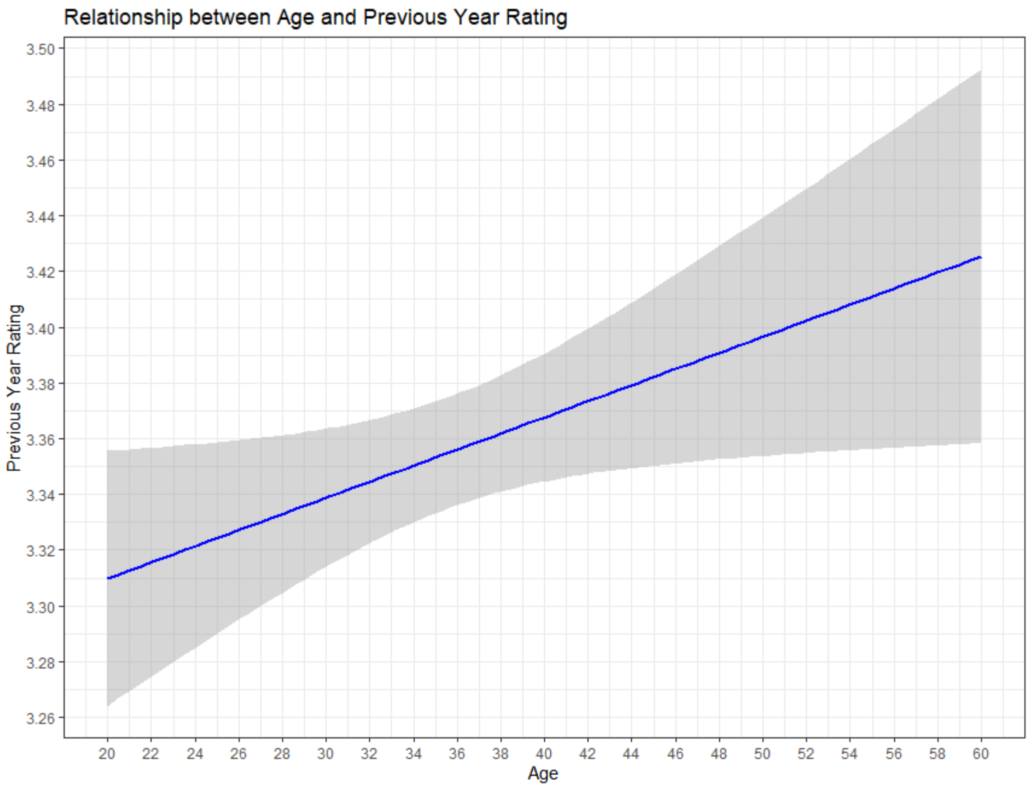
Upon first inspection of the analysis, this one stood out the most. It shows a negative correlation between number of training sessions and performance rating. This makes sense as training sessions take employees away from their work. So, if they fall behind because they've attended numerous sessions, it makes sense that they'll likely receive a negative rating.

```
ggplot(data = Employees_dataset, aes(x = no_of_trainings, y = previous_year_rating)) +
geom_smooth(method = "lm", formula = y ~ x, color = "blue") +
labs(title = "Relationship between Number of Training Sessions and Previous Year Rating",
x = "Number of Training Sessions",
y = "Previous Year Rating") +
scale_x_continuous(breaks = seq(1, max(Employees_dataset$no_of_trainings), by = 1)) +
scale_y_continuous(breaks = seq(2, 3.5, by = 0.2))
```



Lastly, we have the correlation between age and previous years rating. Which shows a positive correlation between the variables. This makes sense as with age (hopefully) comes experience. And not just with work but in social situations which can influence the bias one has towards a person.

```
ggplot(data = Employees_dataset, aes(x = age, y = previous_year_rating)) +
geom_smooth(method = "lm", formula = y ~ x, color = "blue") +
labs(title = "Relationship between Age and Previous Year Rating",
x = "Age",
y = "Previous Year Rating") +
scale_x_continuous(breaks = seq(20, max(Employees_dataset$age), by = 2)) +
scale_y_continuous(breaks = seq(2.8, 3.5, by = 0.02))
```



Footnote: there is overlap with how I used Excel, SQL and R. Again, I wanted to familiarise myself more with these programs. So, I used a little bit of all of them for this project, and kept to the strengths of all the applications as best as possible.

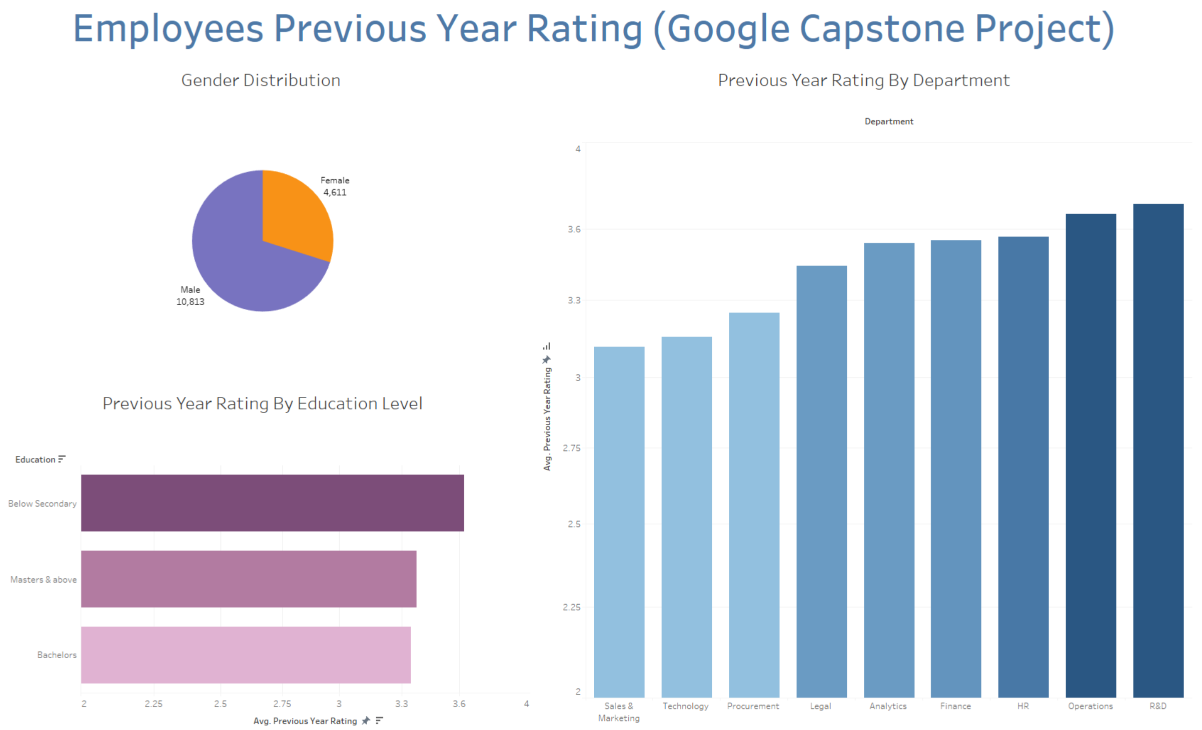
Share

One of the benefits of R is the visualisation aspect of it. So, I would be showing those graphs (and the SQL tables) alongside the Tableau graphs.

The business question was answered but that spurred on more questions in general. For example, in the right-hand chart below, the operations and R&D departments were the only departments to average over 3.6 on their previous year rating. The question would now be - why? This will then allow one to explore how they are managed and organised. This can then be compared to other departments within the business. It'll be compelling to see multi-year data with the same metrics. Then one could really see how well departments are being managed.

One vital story that my data tells is that taking training sessions seriously is a good strategy to get ahead in the business. Trying to achieve the best mark possible may pay dividends with ones performance rating.   Even though this is for the executive team, it is available for all staff to see. As it may inspire them to improve on certain aspects of their work.

Below is the dashboard I created on Tableau with additional information such as the average performance rating for various education levels.



Act

My final conclusion is that there needs to be more exploration in to certain variables before a strategy can be created to improve performance ratings. For example, the different types of training sessions offered needs to be collected. That data will then be transposed in Excel during the ‘Process’ phase to see which employees went to which session(s). Statistical analysis on that would yield much better insight than the contemporary one.

However, that is in a perfect world. Here are my suggestions based on the current data:

1. Training sessions should be limited to a maximum of 3 per year. As after that the negative correlation starts to become quite noticeable.
2. The most important next step stakeholders should take is to implement a mentoring scheme between young professionals (not just graduates), and seasoned workers. It’ll be very beneficial for older workers to show younger workers how they manoeuvre in the professional realm. I believe they’ll be able to provide exemplary mentorship, especially on the social aspect of things.