Project Report

# GitHub URL

<https://github.com/mariemcsweeney/UCDPA_mariemcsweeney>

# Abstract

To summarise the project, I searched the web to source a dataset that reflected a large company with a lot of employees and attempted to provide some insights around employee attrition and performance to senior leadership and to HR. I wanted to also be able to look at this company in more of a macro context so I searched the web for open-source data both on a particular company (IBM [sample employee data]) and aggregated global data around employee satisfaction, based on a study from Gallup. I had to clean and pre-process both sets of data to use them and merge them, then I had to analyse this data. Then I had to come up with visualisations around the data and derive insights. As a result of insights around groups that are vulnerable (ages 45-64) around job satisfaction, work/life balance, and how training initiatives are deployed. I therefore wish to suggest some particular departmental focus – i.e. training in the R&D Department and performance concerns in HR Dept. Finally, I set up a GitHub project repository.

# Introduction

I have both a personal and professional interest in workplace culture and team dynamics. As technological advancements have brought about new opportunities, so has the intensity of work on the average employee. Some statistics I have encountered:

* (World Economic Forum) only 33% of the 27,000 employees surveyed in 2021 felt as though they were thriving in their work.
* (America Bureau of Labor Statistics) 44% of Western European workers feel stressed at or about their work.
* (European Agency for Safety and Health at Work) 65% of the working population in the US are actively seeking a new role.

The ability to derive insights on the most valuable resource to a company – people – is often underexplored and underappreciated. I believe People analytics is key to bolstering strategic business partnerships and businesses who undertake the activity are proven to financially outperform those that do not.

# Dataset

For this assignment, I used one main report and brought in a second report. For the main dataset, I searched Kaggle.com – I was confident that it is reputable enough that downloading a dataset wouldn’t have any concerns around academic or commercial licensing. Kaggle is a huge community of online Data Scientists and one I would like to get further involved with as they do competitions and hackathons. Kaggle is open source and free.

Gallup data is also free, highly reputable, and continually surveys citizens in 160 countries, representing more than 98% of the world’s adult population.

# Implementation Process

* I imported all the packages I knew I would initially need – pandas, NumPy, requests
* I went onto Kaggle, saved to my local drive and brought into Jupyter
* So I wouldn’t have to constantly write down the local location, I named my DataFrame ‘employee\_data’.
* To check it, I applied an index on ‘Gender’ and viewed the first 4 rows to make sure it imported as expected and did a reverse sort on years at company to make sure I brought in all the rows
* I used the .info function to get a concise summary of the DataFrame and I didn’t know what the index would be so I set it at zero, one, etc
* To see the data all listed out to see the data to appreciate it, I was able to use a looped function to print out the column names and iterate over the rows
* Employee data -> Employee # 2068 is a duplicate, removed this to prepare the data, line 1472. This code is not specific to employee #2068 as it simply sorts by the employee number and then drops the duplicate and in this way, is reusable code.
* Employee data -> Column V missing some entries of ‘Over 18’, as column A has the employee ages, we can interpolate data of ‘Y’ for these, could use a fill statement but as it is conditional, I use a ‘loc’ function.
* I categorized the ages into bands, for useful comparison with the Global Study and rolled up those employees who rated their job satisfaction as 3 or 4 as ‘enjoying their job’. Covered Conditional statements by the way we added a new age category, for ‘if’, ‘then’ rules, we assigned a category based on the value of age,
* Groupby – we pivoted up the age categories and grouped together the higher elements of the job satisfaction scores to broadly suggest if the employee enjoyed their job.
* I then scraped the web to bring in the Global Wellbeing Initative dataset – I had to use urlretireve function to do so and the resulting .xlsx file had two tabs – I just wanted to see ‘Global’ and name the url wellness\_df
* I utilized the pivot table function to compare the job satisfaction and expressed the results as a percentage and compared to the Global Wellness Index and could see for every age group, the employees of IBM score their job satisfaction less than that of those surveyed by Gallup

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

# Results

There is a linear relationship between employee retention and the higher their job satisfaction, how they perceive their work/life balance and their commutes:

A graph showing a line of a plane

Description automatically generated with medium confidence

A blue line on a white background

Description automatically generated

A blue triangle with a line

Description automatically generated with medium confidence

A graph showing a line of blue lines

Description automatically generated with medium confidence

Then, I wanted to look bring in the job level into the analysis to reflect on any nuances from those employees with different responsibilities:

A graph of different colored lines

Description automatically generated

Next, I wanted to look at performance and look at training and overtime, and focus in on different department to see if their approach was different.

A graph of performance rating

Description automatically generatedA graph with a line

Description automatically generated with medium confidence

A graph showing the results of a job

Description automatically generated with medium confidence

Finally, I wanted to combine the insights between performance and attrition to see if there was any actionable insights to be derived.

A graph of different colored lines

Description automatically generated

# Insights

**Engagement and Attrition analysis**

Comparing to a benchmarked Global Wellness Index, the employees are well below the job satisfaction levels of the global population, in all age groups, but in particular 45-64 Age Group, with 59% reporting a 3 or 4 in job satisfaction vs 86.1% of the general population reporting job satisfaction. In IBM, there is actually an inverse impact of job satisfaction vs general population – the older an employee gets, the less satisfied they are with their role. As a result, I would make a recommendation to Senior Management to target initiatives towards all groups around engagement, starting with those particularly in mid-career in particular that might need their needs at their particular age and stage of life. It is important not to let any biases around their attitudes to work cloud any grasps from the data.

It is intuitive to believe that an employee’s job satisfaction has a linear relationship with likelihood of attrition. In particular, IBM’s management group should take note of an employee’s satisfaction rating of 2.47 or below as this is a high predictor of the employee leaving the business and as a general rule, People Managers and/or HR should focus in on an individual employee. One suggestion would be to review the telecommuting policies and work-from-home flexibility – the distance the employees have to travel is an even more meaningful predictor of an employee leaving than their own perception of work/life balance. It is hard to state from the data the commuting time impact on work/life balance due to the variability (so suggesting a number of other factors have an influence on the employees perception of his/her work/life balance). That being said, the commute does appear to have a significant impact for some, with a high standard deviation.

Broadly looking at a three-way visualization between attrition, job satisfaction and job level, the more senior an employee is, the less likely they are to turnover. An interesting exception to this is in the mid-level job level. This insight is predictive in nature and may identify some key personnel at risk due to low job satisfaction – it may potentially signify mid-level manager burnout – is there enough support being put in place so they in turn, can develop and support others? This is a recommendation for HR for further investigation.

**Performance Analysis**

The relationship between the amount of times training was implemented last year and the performance rating is a negative correlation in the aggregate. This would potentially indicate that mangers at IBM spend a lot of their training resources/budget etc augmenting the knowledge of underperformers to get them to a basic level of understanding or fill in particular gaps but then no more. This finding appears to be unique to those who work in Research & Development department.

Conversely, however, those who got the highest performance ratings, also were in receipt of the most training and this effect was magnified in Sales roles.

Suggesting to HR that Learning & Development department should market themselves as training for continuous growth mindset, instead of just filling in the gaps. Some cross-functional learning may also be relevant as any sales training appears to be an investment.

Overtime in the aggregate does not have any effect on an employee’s performance. The visualisation indicates that it may do as much harm as good as there are those doing overtime that have low performance and high performance and the opposite. This may signify to IBM that in knowledge-type roles, working longer hours doesn’t actually produce higher results and HR could do analysis on employee’s overtime to understand the factors driving that and put supports in place to reduce it as the cost of overtime does not appear to be recouping many or any benefits.

Such as the analysis I discovered while looking through the data to identify concerns around attrition, I see a similar pattern when appreciating data around performance. Broadly speaking, there is a slight upswing in performance ratings the more senior an employee but again, we see a dip in performance in the mid-level job role. On deeper analysis, this appears to be very prevalent within the HR department. Suggesting therefore Senior leadership within HR, open communication lines with mid-level employe

# Machine Learning Questions

**Describe what kind of prediction you could perform in future using machine learning and/or deep learning?**

In future, we could use the IM data in aggregate from e.g. MS Teams or Sack in an unsupervised way to conduct Sentiment Analysis and thus be able to identify a disengaged employee. We would need to test the model over time by continuing to put managerial/performance/wellness support in place and then use the feedback received to give feedback to the data.

**Would you use classification or regression methods?**

In this case, we would want to use a classification method – i.e. we want to be able to predict the performance of an employee or assess the likelihood that an employee will leave the business.

# References

**There** Datacamp – Podcast ‘DataFramed’ episode -Kaggle and the future of datascience

UCD Class Notes

[How people analytics is transforming the HR landscape | McKinsey](https://www.mckinsey.com/capabilities/people-and-organizational-performance/our-insights/how-to-be-great-at-people-analytics)

* (World Economic Forum) only 33% of the 27,000 employees surveyed in 2021 felt as though they were thriving in their work;
* (America Bureau of Labor Statistics) 44% of Western European workers feel stressed at or about their work;
* (European Agency for Safety and Health at Work) 65% of the working population in the US are actively seeking a new role.

[Using people analytics to drive business performance: A case study | McKinsey](https://www.mckinsey.com/capabilities/quantumblack/our-insights/using-people-analytics-to-drive-business-performance-a-case-study)

[Organizational Health Index | McKinsey & Company](https://www.mckinsey.com/solutions/orgsolutions/overview/organizational-health-index)

Kaggle.com – level up with the largest AI and Machine Learning Community = https://www.datacamp.com/blog/what-is-kaggle

[HR Analytics Case Study | Kaggle](https://www.kaggle.com/datasets/vjchoudhary7/hr-analytics-case-study)

[5 Ways Technology Can Help Build a Strong Company Culture - Spiceworks](https://www.spiceworks.com/hr/engagement-retention/articles/5-ways-technology-can-help-build-a-strong-company-culture/)

Podcast – the path to becoming a Kaggle Grandmaster – ‘DataFramed’ the podcast by DataCamp

[IBM HR Analytics Employee Attrition & Performance | Kaggle](https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset)

[Collecting HR Data With Web Scraping (Why It Matters) | Scraping Robot](https://scrapingrobot.com/blog/hr-data/)

<https://dummyapis.com/dummy/employee>

note:data is always random]

<https://www.gallup.com/file/education/468923/Global%20Wellbeing%20Initiative%20Dataset.xlsx>

<https://www.gallup.com/467702/indicator-employee-retention-attraction.aspx>

Chat GPT – attempting to express a pivot table as a %

https://en.wikipedia.org/wiki/Gallup,\_Inc.