|  |
| --- |
| DU Data Analytics Bootcamp |
| Remote Work and Employee Mental Health |
| An exploration of a dataset |

|  |
| --- |
| Mélisa Teglas, Tamya Louis, Don Ross  11-4-2024 |

Introduction

With the chosen dataset, this project was to identify patterns in the mental health and stress levels of employees working on-site vs. hybrid vs. remote. We scrutinized the relationships between survey responses from the employees and their demographic characteristics, features of work habit trends, and other associations drawn from the data.

Specifically, the following questions were posed:

1. Demographics trends
2. What is the relationship between work location and work life balance ratings?
3. Are there regions where there is more stress than others? What about by work location?
4. What is the relationship between work location and stress level? What about by access to mental health?
5. What is the relationship between physical activity and work location? What about by region?
6. What is the relationship between work location and social isolation ratings? What about by mental health condition?
7. Do employees endorse certain mental health conditions more when working hybrid vs. onsite vs. remote?

Data

The dataset titled “Remote Work & Mental Health” was identified from Kaggle.com and was collected, organized, analyzed and posted by Waqar Ali. According to the author, this dataset draws from a variety of sources and was collected via a structured survey that included both quantitative and qualitative questions to explore how working remotely affects stress levels, work-life balance, and mental health conditions across various industries, job roles, and world regions. The questionnaire was distributed digitally across social media platforms, professional networks, and online forums, and the responses were anonymized to ensure confidentiality.

The data itself is contained within a downloadable CSV file, which was stored in the GitHub repository for the project and contains 5000 rows and 20 columns of data. Some data are categorized using numeric values, such as ratings ranging between 1 and 5. Other data use objects, such as strings, to describe the respondents’ ratings.

Limitations of the Dataset

There are a few limitations of this dataset, including:

* Job Roles:
  + Although the job roles are the same within each industry, it seems possible, if not likely, that these roles may vary based on the industry they represent.
* Years of Experience:
  + The column that represents the number of years of work experience appears inaccurate, as some of the reported durations of experience surpass the respondents’ reported ages.
* Work Location:
  + “Hybrid” is not clearly defined in this section. For example, does it refer specifically to an employee splitting their time between the office and working remotely 50% of the time? Or 25%/75%? Or 5%/95%? Also, it is a little unclear whether the employee’s work location was their choice. For example, it would be expected that an employee who works at their desired location might be more satisfied than one who is forced to work at a non-preferred location.
* Number of Virtual Meetings:
  + The data reported here are not defined clearly, as the duration of each virtual meeting could be a standard one hour or could be a 15-minute check-in or a 4-hour work session.
* Mental Health Conditions:
  + In the mental health conditions column, it appears there were only three options from which the respondents could choose, and there was no option for “other.” Additionally, it is unclear whether these conditions are self-diagnosed or whether they were clinically diagnosed by a healthcare professional. Further, there are some missing values in that it appears not all of the respondents answered this question.
* Access to Mental Health Resources:
  + This is not specifically defined, other than to report that this refers to whether the employee had access to mental health resources. However, it is unclear what type of resources were included in this definition. For example, did the employees have access to employer-supported mental health treatment, such as psychotherapy and/or medication management? Or did the employees only have access to recommended fact sheets and self-care apps, for example?
* Physical Activity:
  + The only options for the categorized data are “weekly,” “daily,” and “none.” What if a person exercises less frequently than weekly but more frequently than none?
* Values:
  + Certain values were collected using discrete objects (strings). For example, the respondents’ stress levels were recorded as “low,” “medium,” and “high.” Other values were collected using numeric data. For example, respondents rated their work-life balance according to a scale ranging from 1 to 5. Furthermore, for the items that were scored using an integer, there is no explanation as to whether 1 refers to the lower end or upper end of the scale.

Methods

* Placed csv file in GitHub repository
* Read file
* Printed basic info about data frame: range, data types, number of columns, etc.
* Removed Employee\_ID column
* Assigned integer values to:
  + Stress level
  + Productivity change
  + Satisfaction w/ remote work
  + Physical activity
    - Filled in NaN values with “none”
    - Then, assigned values to “weekly,” “daily”, and “none”
  + Sleep quality
* Filled in NaN values for mental health conditions with “not reported”
* Sorted the data according to the reported MH conditions and then reset the index
* Plotted various data to explore trends and themes

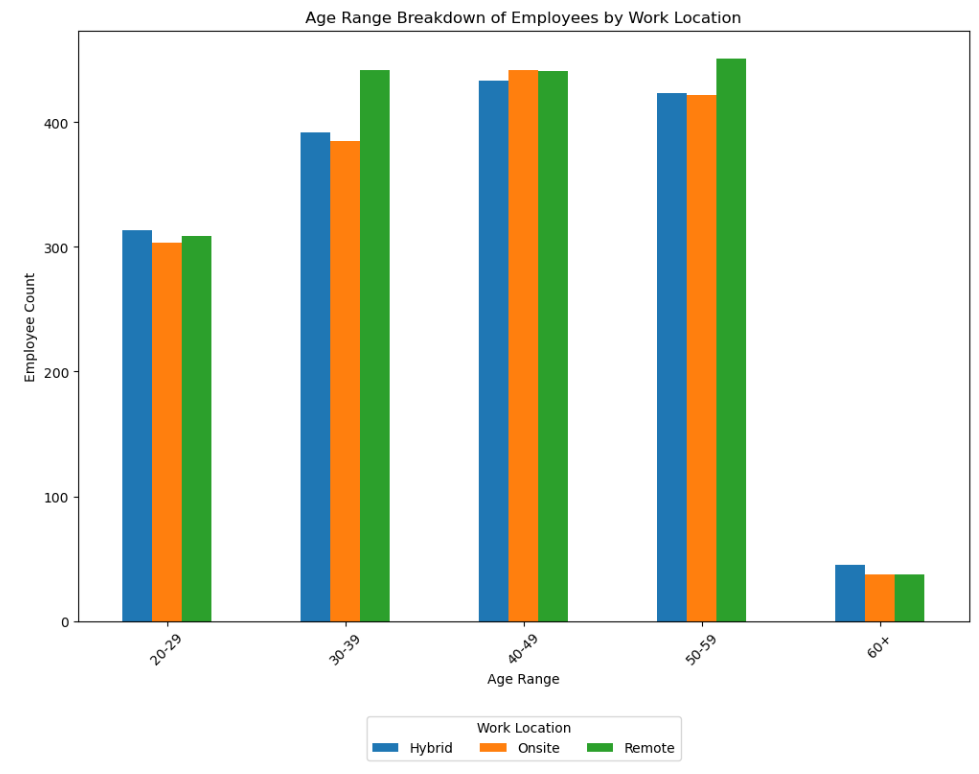
Visualizations and Analysis

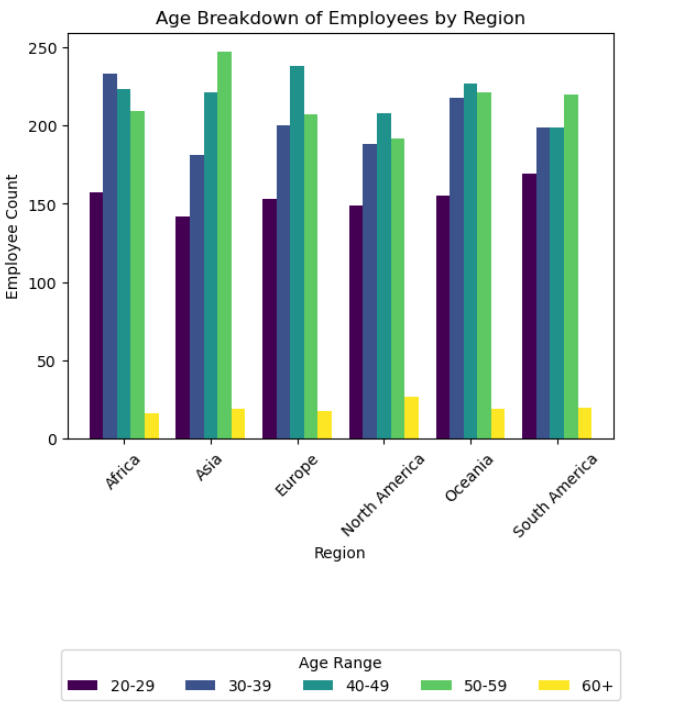
1. Demographic trends

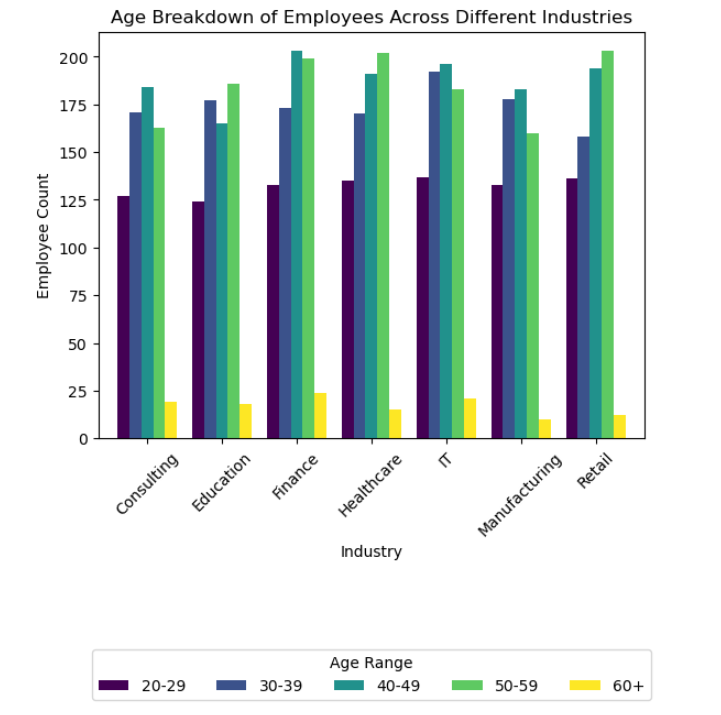
Mid-career employees (ages 30-59) make up most of the workforce across different work locations, regions, and industries. The youngest (20-29) and oldest (60+) age groups have lower representation, possibly due to career-entry timing, early retirement trends, or evolving career preferences.

Regional and industry-specific trends:

* Asia and South America have relatively younger workforces, with a higher concentration of employees in the 20-29 age range.
* Healthcare and IT attract a larger share of mid-career professionals, likely due to the demand for experienced employees in these fields.
* Education stands out with a higher proportion of employees in the 50-59 age range. This suggest that employees in Education tend to stay longer in their roles, possibly due to stability or commitment within the field.





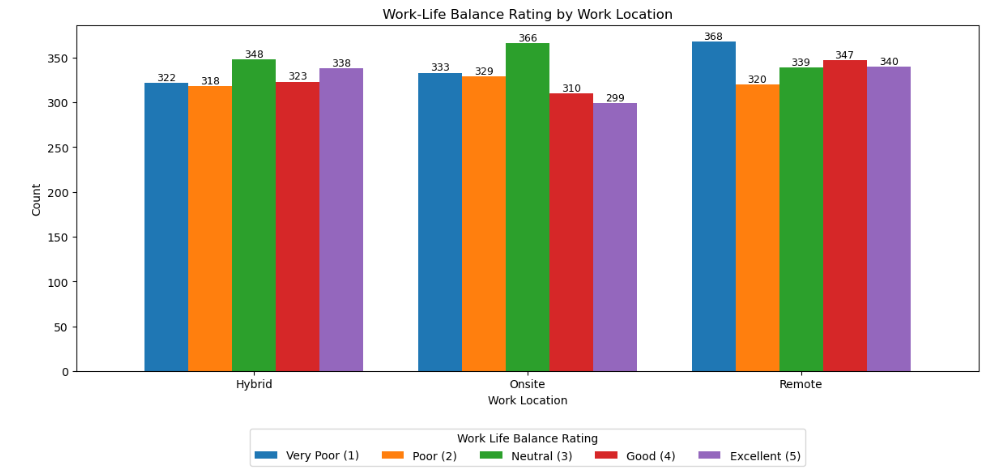


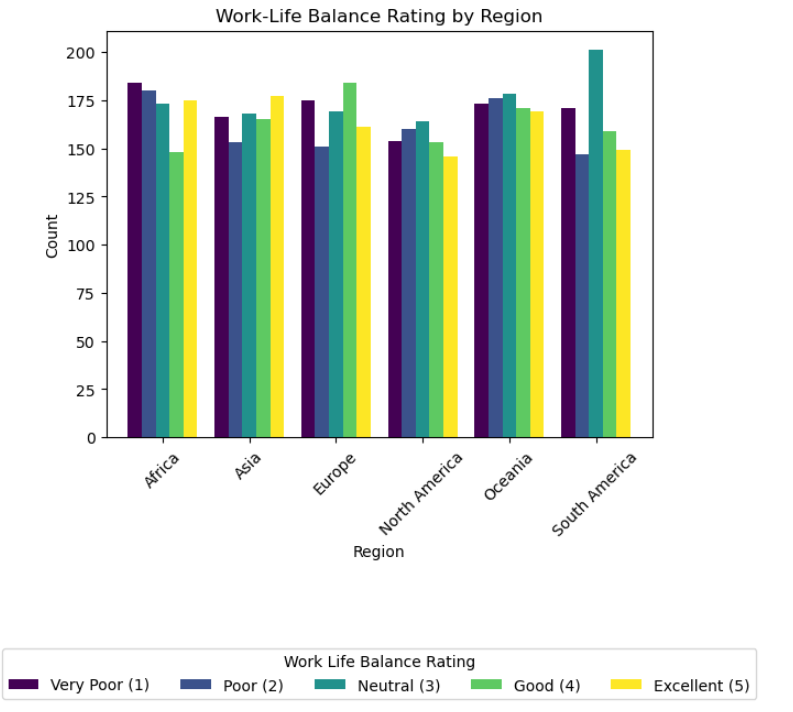
1. What is the relationship between work location and work life balance ratings?

Work-life balance satisfaction varies somewhat across regions. Europe and South America tend to have higher satisfaction, with a greater proportion of "Good" and "Excellent" ratings. In contrast, North America and Africa show a higher incidence of "Very Poor" and "Poor" ratings, indicating potential challenges with work-life balance in these areas.

Work location has a more distinct impact on work-life balance satisfaction:

* Remote work is associated with higher satisfaction, as shown by a higher proportion of positive ratings ("Good" and "Excellent").
* Onsite work is linked with lower satisfaction, with more "Neutral" to "Poor" ratings, possibly due to commuting demands and less flexibility.
* Hybrid work offers a balanced experience, with a relatively even distribution of satisfaction ratings, suggesting that it provides moderate flexibility without fully replicating the benefits of remote work.





1. Are there regions where there is more stress than others? What about by work location?

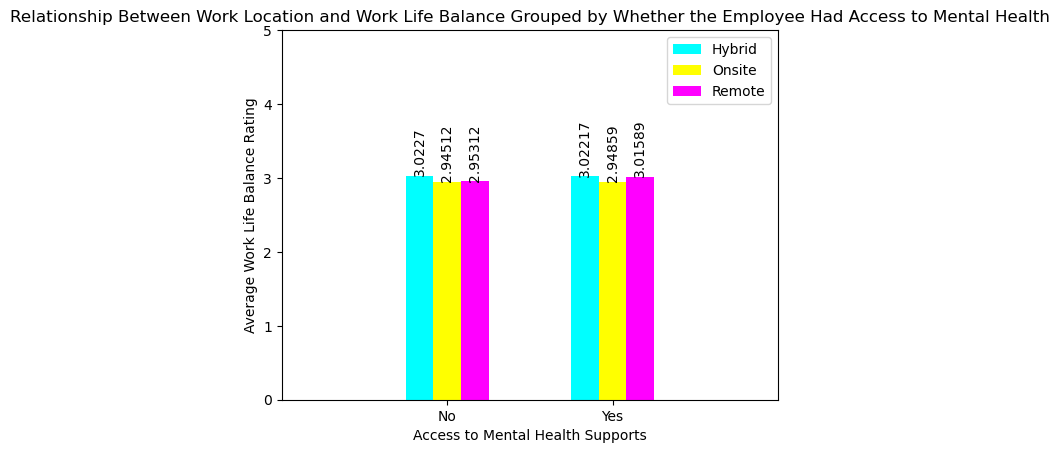
According to the grouped bar chart below, Europe and Africa appear to have the highest ratings for stress levels for workers working remotely. Interestingly, working onsite appears to be approximately less (or the same amount of) stressful than working remotely in some regions, including Africa, Europe, Northa America, Oceania, and South America. That being said, the differences between the average stress levels are very small.

A graph of stress levels

Description automatically generated

1. What is the relationship between work location and stress level? What about by access to mental health?

The chart below represents the relationship between work location and average work life balance ratings of workers. In general, it appears that there is a higher level of work life balance for hybrid workers, regardless of whether they had access to mental health supports. There also appears to be no difference between the two groups representing access to mental health supports.

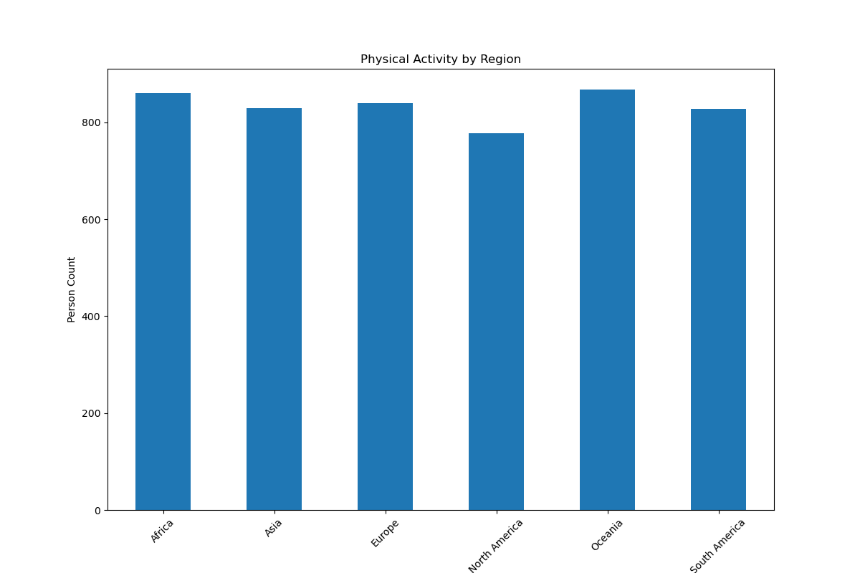


1. What is the relationship between physical activity and work location? What about by region?

Remote and Hybrid workers get more 'daily' and 'weekly' activity than onsite workers, while Onsite workers have a higher instance of 'none' for activity. Africa and Oceana have highest instances of 'daily' activity, with Asia, North American, and South America the lowest for 'daily' activity and fairly even. Asia has the highest instance of 'none', and North America has the lowest instance of 'weekly' activity.

Region Instances

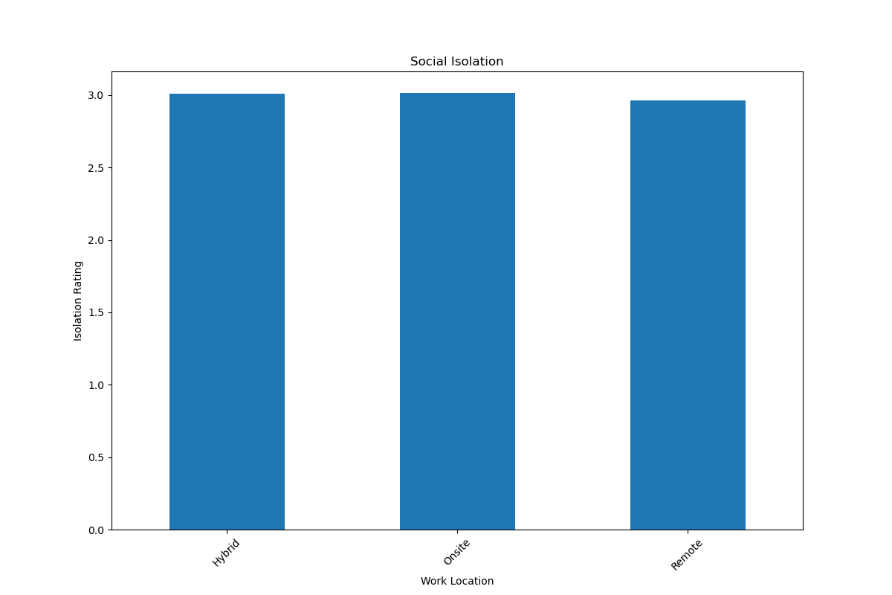
Africa 860  
Asia 829  
Europe 840  
North America 777  
Oceania 867  
South America 827



1. What is the relationship between work location and social isolation ratings? What about by mental health conditions?

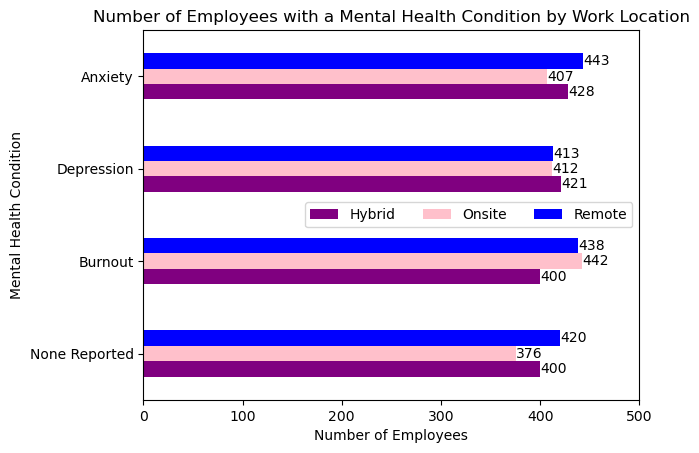
There is a higher count/instance of Remote workers in the study, while they show the lowest average social isolation rating. Onsite workers show a higher instance of Burnout, Hybrid workers show a higher instance of Anxiety, and Remote workers show a high instance of both Anxiety and Burnout, the instances on this are very similar (443 and 438 respectively).

Work Location Avg Social Isolation Rating  
Hybrid 3.008490  
Onsite 3.012828  
Remote 2.961494



1. Do employees endorse certain mental health conditions more when working hybrid vs. onsite vs. remote?

According to the chart below, it appears that anxiety is endorsed the most for remote and hybrid workers and burnout was identified as the most endorsed for onsite workers. However, there was a significant number of employees who chose not to answer this question. One hypothesis for why this might be is that only three discrete options for mental health conditions were offered on the survey, anxiety, depression, and burnout. There was no option for “other”, if the individual had a mental health condition that was not one of the three options. Additionally, there was no option for “none” on the survey. So, it’s possible that employees chose to avoid answering this question if relevant options were not available. Additionally, from this graph, it appears that the frequencies of the reported mental health conditions were similar across mental health conditions and across work locations.



Conclusion

The dataset that was used for this project appeared to have a relatively normal distribution, so there were no major significant differences between groups. That being said, there were several instances where we could carefully deduce relevant information from this dataset. However, caution should be used when reviewing the visualizations and analysis, as there were several limitations to this dataset.