REMOTE WORK STRESS-DEPRESSION LEVELS DATA ANALYSIS

Darya Korbut

EPAM School of Digital Competencies, Savanorių pr. 28, Vilnius

Abstract

As remote work becomes increasingly common, it's essential to understand its effects on employees' well-being. This topic is relevant because while remote work offers flexibility, it can also come with its own set of challenges, especially when it comes to mental health. My research focuses on how working remotely influences stress levels, work-life balance, and mental health conditions across various industries and regions.

Keywords: remote work, stress levels, work-life balance, mental health, algorithm development

Introduction

It is necessary to pay attention, the main focus of the study is to develop an algorithm for the problem studying. The method is to choose and justify statistical tools for conducting the analysis. Therefore, the data set from the Kaggle educational resource was chosen. It preserves the data structure for developing the algorithm. And it allowed us to avoid the initial wrangling data.

With 5,000 records collected from employees worldwide, this dataset provides valuable insights into key areas like work location (remote, hybrid, onsite), stress levels, access to mental health resources, and job satisfaction. It's designed to help researchers, HR professionals, and businesses assess the growing influence of remote work on productivity and well-being.

Columns:

1. Employee_ID: Unique identifier for each employee.
2. Age: Age of the employee.
3. Gender: Gender of the employee.
4. Job_Role: Current role of the employee.
5. Industry: Industry they work in.
6. Work_Location: Whether they work remotely, hybrid, or onsite.
7. Stress_Level: Their self-reported level of stress.
8. Mental_Health_Condition: Any mental health condition reported (Anxiety, Depression, etc.).
9. Social_Isolation_Rating: A self-reported rating (1-5) on how isolated they feel.
10. Satisfaction_with_Remote_Work: How satisfied they are with remote work arrangements (Satisfied, Neutral, Unsatisfied).

This research was conducted in Google Colab resource using programming language Python and libraries for data manipulation, visualization and statistical analysis such as pandas, numpy, scipy, matplotlib, seaborn, tabulate, statsmodel and scikit-learn.

Moreover than that, the implementation of the algorithm step by step included three main stages such as:

1. Exploratory Data Analysis (EDA):
   - Examination of data types, shape, and basic descriptive statistics.
   - Count of unique values in categorical variables.
   - Filtering and identification of missing values and duplicates.
   - Visualization of feature distributions by using box, scatter, bar, pie plots and histograms.
2. Statistical Hypothesis Testing:
   - One-sample t-test.

- Two-sample (independent) t-test.
- Paired t-test.
- One-sample t-test with normally distributed data.
- One-sample z-test for population mean.
- Two-sample z-test for population means.
- One-sample z-test for population proportion.
- Two-sample z-test for population proportions.
- ANOVA for comparing group means.
- Chi-squared test for independence of categorical variables.

3. Multivariate Analysis:
- Principal Component Analysis (PCA) for dimensionality reduction.
- Cluster analysis (e.g., K-means) to detect group patterns and profiles.

Descriptive Statistics Analysis

The exploratory data analysis (EDA) provides a comprehensive overview of the dataset's composition, quality, and key patterns related to remote work and mental health.

Basic descriptive statistics showed that numeric fields fell within plausible ranges, and a summary of categorical fields indicated diverse entries. Only a small fraction of data was missing, which was addressed by removing those records, and no exact duplicate entries were found.

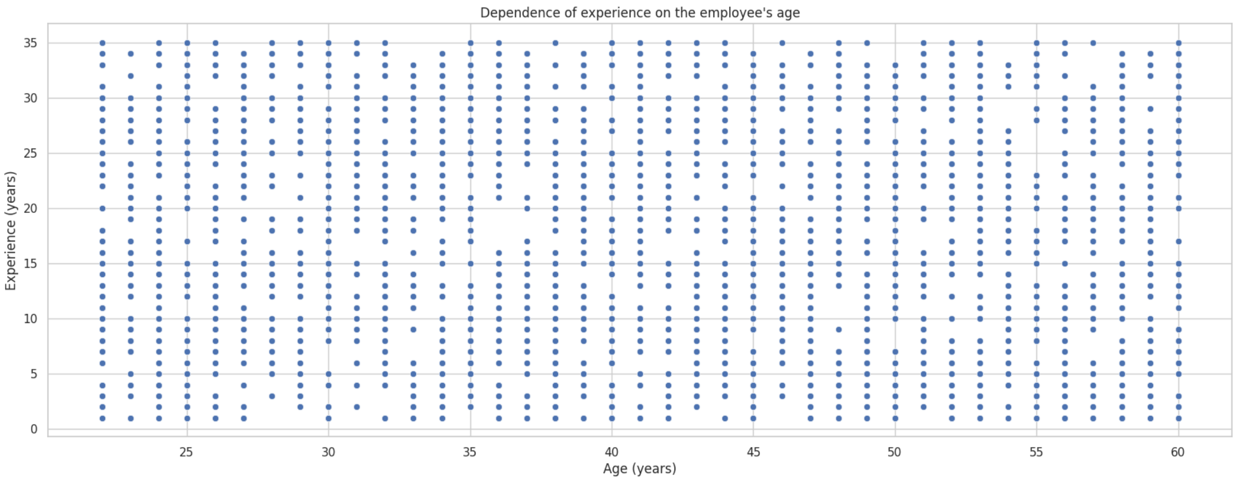

Figure 1 - Dependence of experience on the employee's age

Likewise, an important data integrity check uncovered a logical inconsistency between age and years of experience – for instance, some younger employees reported implausibly high experience. However, this anomaly was corrected by filtering out those entries, ensuring that age and experience values are consistent and the dataset remains reliable for further analysis.

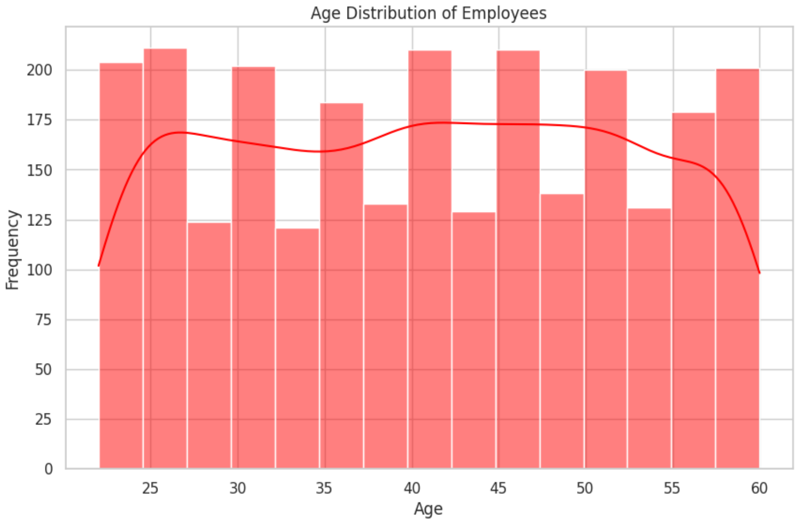

Figure 2 - Age distribution of Employees

Demographic distributions in the dataset appeared balanced and representative. The age of respondents ranged from about 22 to 60 years, with a fairly uniform distribution across this span.

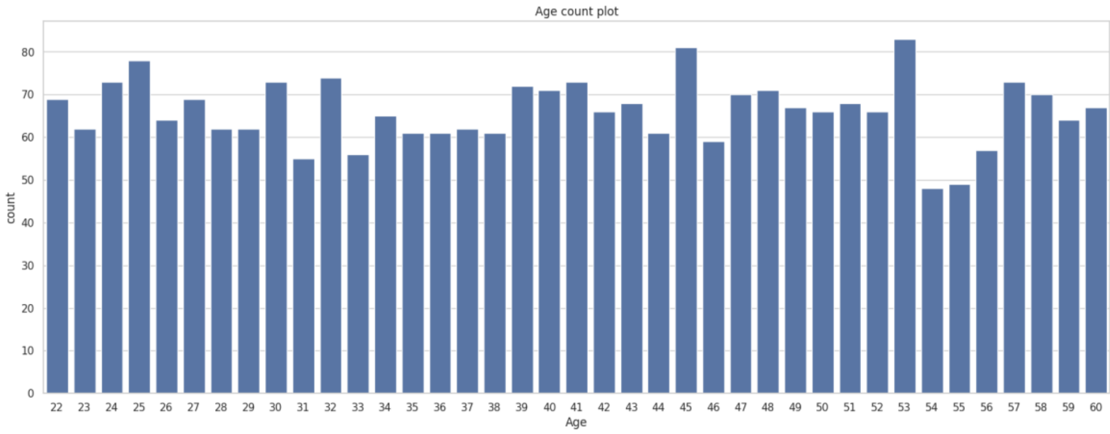

Figure 2 - Age count plot

Most age brackets had similar frequencies, aside from minor expected fluctuations (small peaks in the mid-20s, early 30s, and mid-40s, with slight dips around the late 40s and mid-50s). Overall, no age group within the 22–60 range was severely underrepresented.

The gender composition of the sample was well-balanced as well – no gender category dominated the dataset, indicating an equitable representation of different genders. These findings suggest that the dataset covers a broad cross-section of employees in terms of age, gender, and experience.

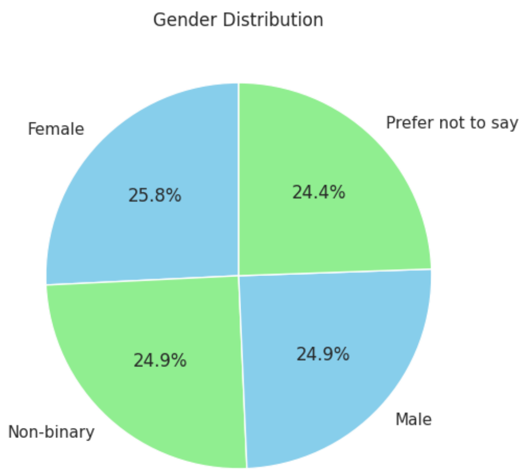

Figure 3 - Gender pie chart

The sample also spans a variety of job roles and industries, with no pronounced skew toward a single field.

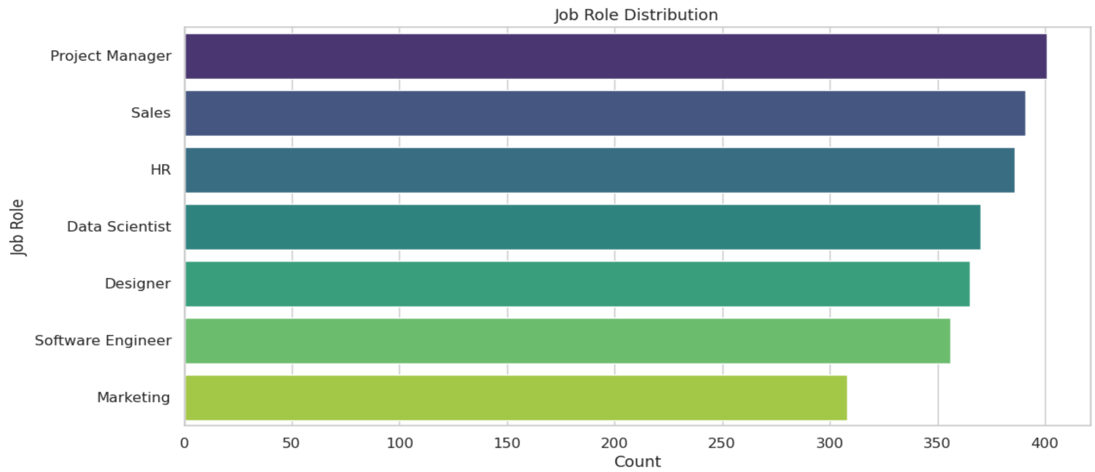

Figure 4 - Job Role distribution

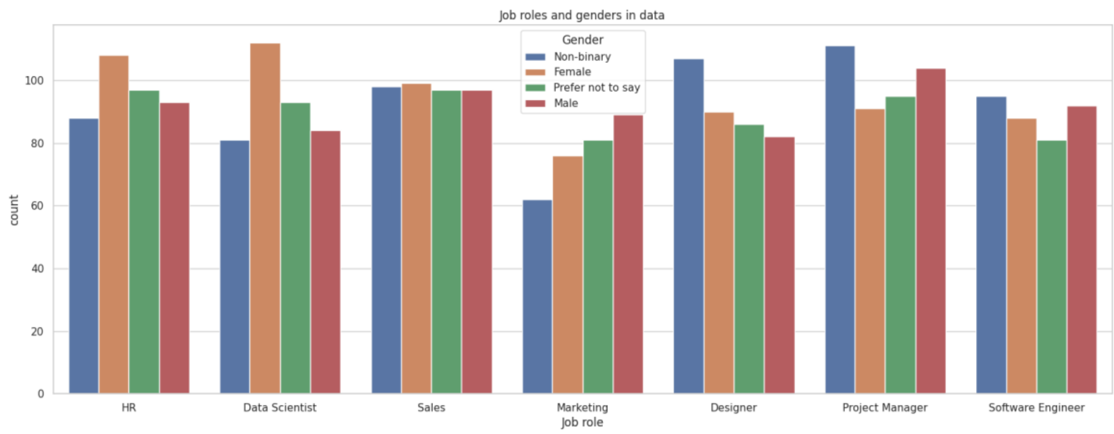

Figure 5 - Job roles and genders in data

A frequency analysis of job roles revealed that respondents work in many different positions, and importantly, there was no evidence of gender imbalance within those roles – each job role included a comparable proportion of different genders.

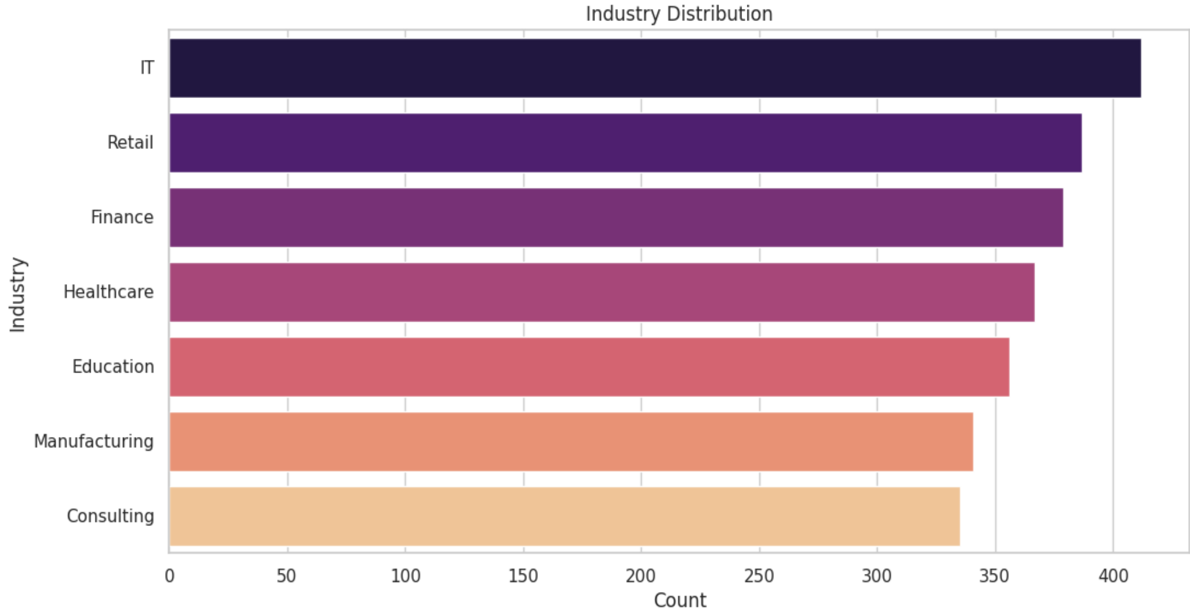

Figure 6 - Industry distribution

Similarly, the industry representation was diverse: while the IT industry had the highest number of respondents and fields like Consulting had the fewest, the overall counts across industries were of the same order of magnitude. This indicates that no industry was overwhelmingly dominant in the data.

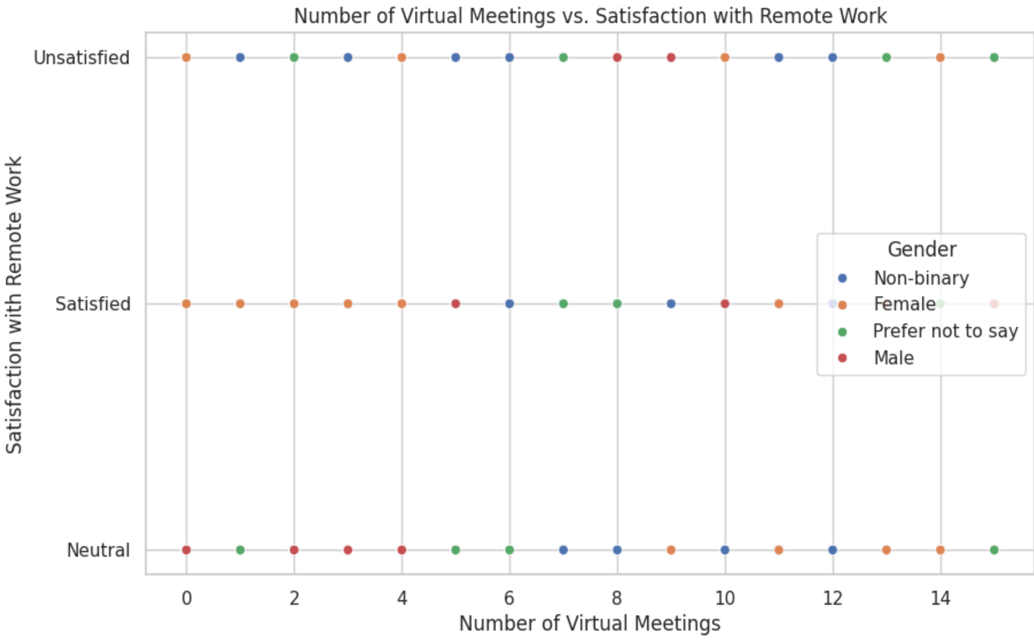

Figure 7 - Scatter plot of the number of virtual meetings vs. satisfaction with remote work

What is more, remote-work satisfaction was broadly distributed across "Satisfied," "Neutral," and "Unsatisfied" categories, with no implausible outliers.

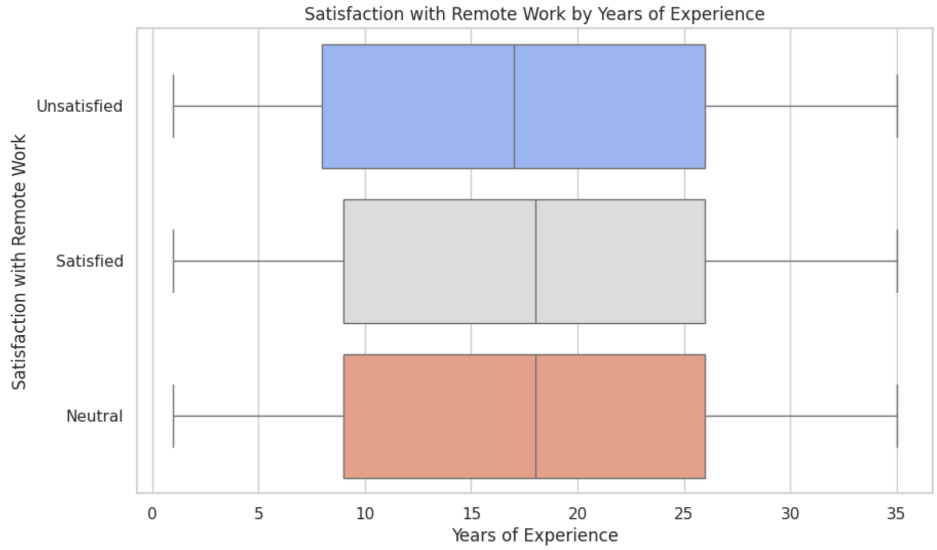

Figure 8 - Box plot between satisfaction with remote work and years of experience

Moreover, average age remained near 41 years in each group, and neither virtual meeting counts nor physical activity levels showed any clear influence on satisfaction. In sum, satisfaction levels were consistent across demographics and work patterns.

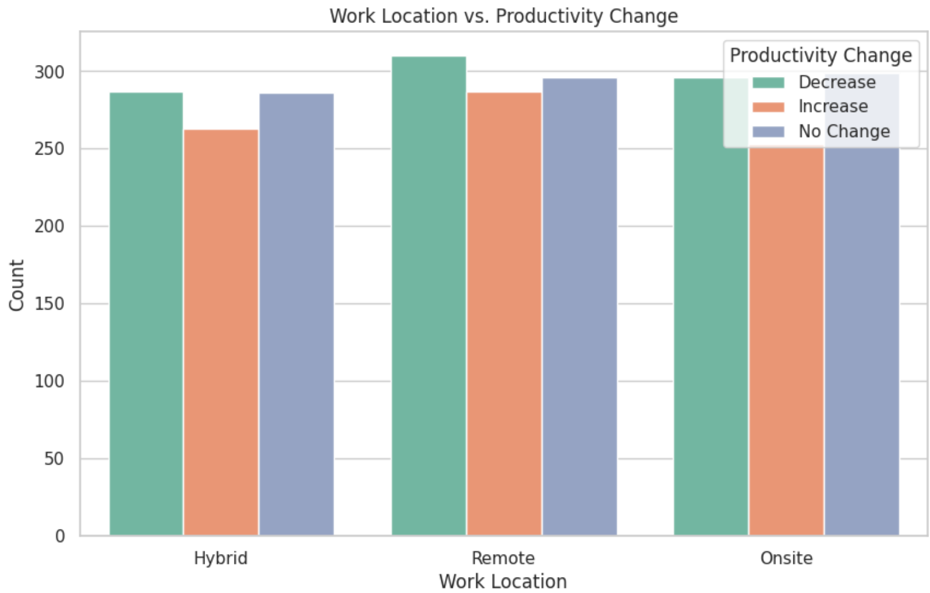

Figure 9 - Work-Life Balance Rating count plot

Work-location groups (remote, hybrid, onsite) reported similar productivity changes, with no group consistently outperforming the others. While most employees rated their work–life balance positively, about half still faced mental health challenges (e.g. anxiety, burnout) and only half had access to support resources, highlighting gaps in well-being provisions.

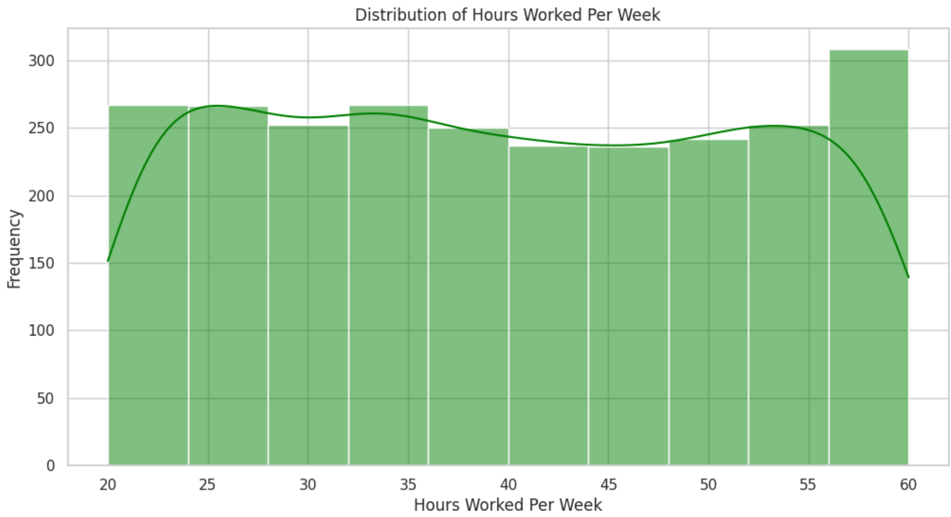

Figure 10 - Histogram of Hours Worked Per Week

Finally, weekly hours clustered between 20 and 60 with a modest uptick at 55–60 hours. Distributions were virtually identical across genders, with no major outliers.

Correlation Analysis Findings

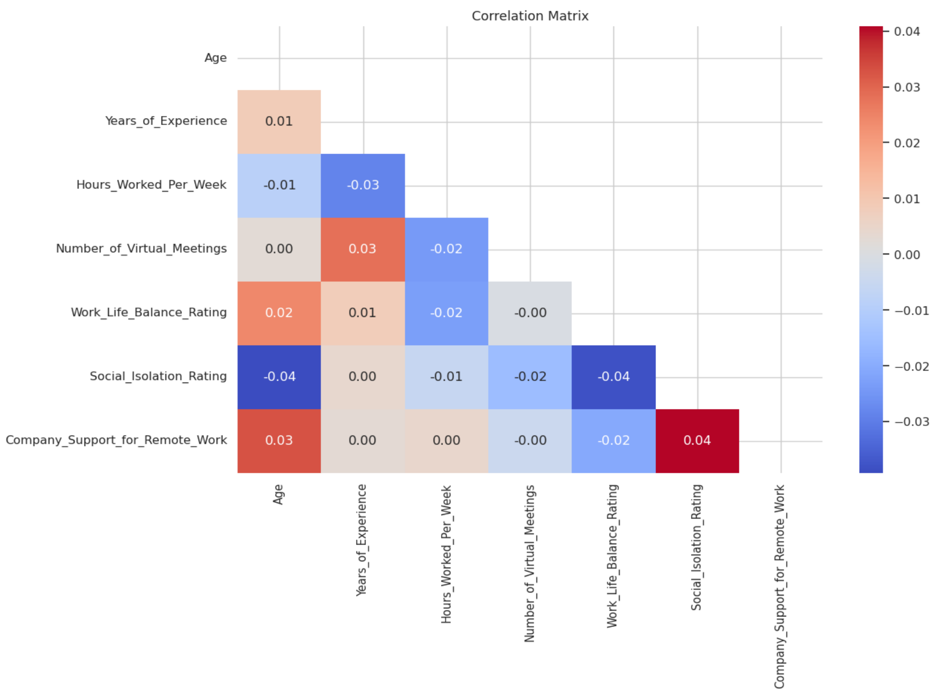

Figure 11 - Correlation matrix

Correlation analysis provides the following results:

1. Weak Positive Correlations:
   - Variables like Company_Support_for_Remote_Work and Social_Isolation_Rating show a weak positive correlation (around 0.04).
   - Age and Company_Support_for_Remote_Work also have a small positive correlation (0.03).
2. Weak Negative Correlations:

- Social_Isolation_Rating and Age have a weak negative correlation (-0.04)
- Other variables, such as Hours_Worked_Per_Week and Social_Isolation_Rating, show negligible or slightly negative correlations.

Hypothesis Analysis Findings

- With t = –1.5612 and p = 0.1186 (> 0.05), the one-sample t-test indicates no significant difference from the 40-hour benchmark.
- There are t = –0.6440 (small magnitude), p = 0.5197 (> 0.05), and confirmation of normality and equal variances, which means that the two-sample t-test shows no significant difference in average hours worked between remote and onsite workers.
- Despite non-normality indicated by Shapiro–Wilk (W = 0.9660, p < 0.05), the paired t-test (t = –0.4684, p = 0.6395) showed no significant change in work–life balance ratings before versus after remote work. Cohen's d (–0.0092) and its 95% confidence interval (–0.0966 to 0.0593) confirm a negligible effect.
- ANOVA results indicate no statistically significant difference in work–life balance ratings among remote, onsite, and hybrid workers, suggesting that work location does not impact these ratings and that other factors may be more influential.
- The chi-squared test found no significant association between gender and work location, indicating that gender does not influence whether employees work remotely, onsite, or in hybrid mode.

Multidimensional Analysis Results

PCA successfully reduces the dimensionality while retaining most of the variance.

Also, visualizations of the first two principal components help explore how well the target variables (Satisfaction_with_Remote_Work or Mental_Health_Condition) are distributed or correlated with the reduced-dimensional space.

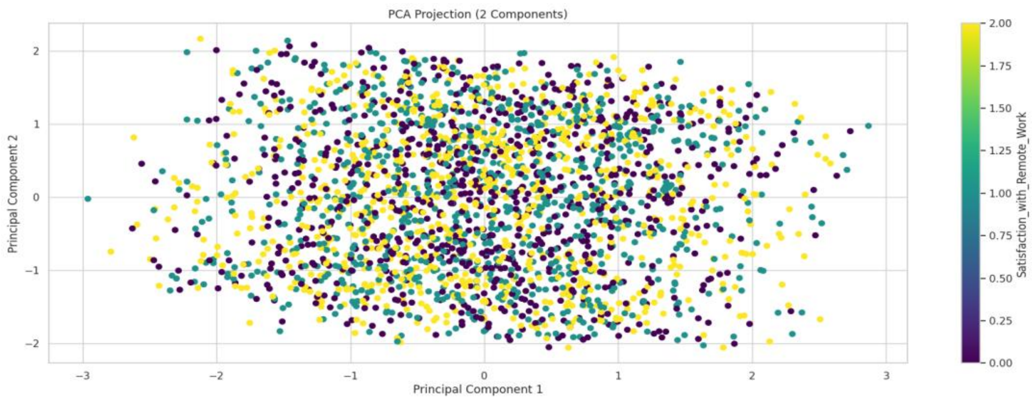

Figure 12 - PCA projection where target is Satisfaction_with_Remote_Work and features are Age, Years_of_Experience, Hours_Worked_Per_Week, Number_of_Virtual_Meetings

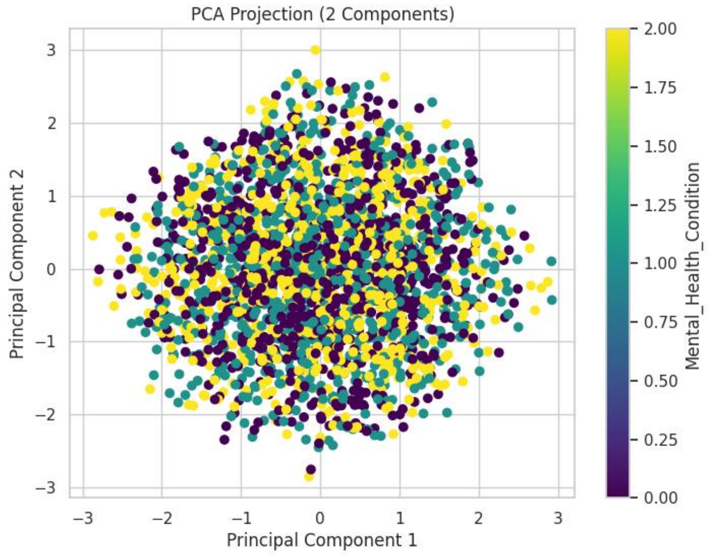

Figure 13 - PCA projection where target is Mental_Health_Condition and features are Age, Work_Life_Balance_Rating, Hours_Worked_Per_Week, Number_of_Virtual_Meetings

Conclusion

First of all, the EDA confirmed that data is synthetic because it was chosen from Kaggle educational tool. Key findings show uniformly distributed satisfaction, productivity changes, and hours worked across groups, alongside a notable minority experiencing mental health challenges and only half having access to support—insights that set the stage for subsequent hypothesis testing.

Moreover, about 50% of employees have access to mental health resources, but many still experience anxiety, burnout, and feelings of isolation. This suggests that while resources are available, they may not be sufficient to address the challenges of remote work.

Interestingly, there are some gender differences in job satisfaction related to virtual meetings. Women tend to be more satisfied with fewer virtual meetings, while men are more tolerant of a higher number of virtual meetings. However, both genders become frustrated when they have around 8 or more meetings a day.

Also, the work-life balance ratings showed that remote work does not significantly alter employees' overall work-life balance. However, some employees reported dissatisfaction due to excessive virtual meetings or lack of weekly physical activity.

Finally, the study shows that the proposed algorithm is effective in analyzing the impact of the form of organization of the organization's employees' work and their mental health. This avoids the problem of burnout at work.

Recommendations

While the data used in this research as an example is synthetic and well-balanced, further research using real-world data could help validate these findings and explore more targeted interventions. In other words, the proposed algorithm should be tested on data sets without preliminary processing in the future.

What is more, the data insights from this research can be valuable and helpful for HR professionals, businesses, and researchers who aim to better understand the complexities of remote work and how to support their employees effectively. For example, the analysis of the real data may suggest that companies may need to focus on improving access to mental health resources, adjusting meeting schedules, and encouraging physical activity among employees.

References

1. Şentürk E, Sağaltıcı E, Geniş B, Günday Toker Ö: Predictors of depression, anxiety and stress among remote workers during the COVID-19 pandemic, Work, 70 (1), pp. 41–51 (2021). doi: 10.3233/WOR-210082.
2. www.wikipedia.com (30 Sept 2011).
3. Bolten S: *Pharmaceutical Statistics: Practical and Clinical Application*, second edition, Drugs and Pharmaceutical Science, volume 40, New York: Dekker, 646 (1990).
4. Leon Lachman and Herbet A. Liberman, *Theory and Practice of Industrial Pharmacy*, CBS Publishers and Distributers, Special Asian, pp. 243–289 (2009).
5. Doe JR: Fundamentals of Data Science, 2nd ed., Athena Science Press, New York, pp. 1–32 (2020). https://www.academia.edu/43911785/Mathematical_Statistics_and_Data_Analysis?auto=download&email_work_card=download-paper.
6. Editorial Team: The Impact of Remote Working on Mental Health: Pros and Cons, Horton International Insights. https://hortoninternational.com/the-impact-of-remote-working-on-mental-health-pros-and-cons/ (2025).