Analysis of how workplace factors impact mental health: conclusions from employee research

Maja Placek
Applied Computer Science
Wrocław University of Science and Technology
266538@student.pwr.edu.pl

I. Introduction

With the growing awareness of the impact of mental health state on an employee and their productivity, the topic of how workplace influences mental well-being seems to be gaining increasing interest. In recent years, factors like company size, remote work-style, mental health benefits have drawn particular attention.

In this report I am going to analyze correlations between workplace-related factors and outcomes concerning employees mental health state. To achieve this, I will utilize the results of a survey available at the following website: [1]

This way, I will attempt to identify means in which employers could promote mental health of their employees and consequently improve overall atmosphere in the company.

II. Data set and its processing

A. Data set

The data set was obtained from the website [1]. It includes results from a survey that measures attitudes towards mental health and frequency of mental health disorders in the workplace.

The survey included 27 questions:

1) Timestamp
2) Age
3) Gender
4) Country
5) state: If you live in the United States, which state or territory do you live in?
6) self_employed: Are you self-employed?
7) family_history: Do you have a family history of mental illness?
8) treatment: Have you sought treatment for a mental health condition?
9) work_interfere: If you have a mental health condition, do you feel that it interferes with your work?
10) no_employees: How many employees does your company or organization have?
11) remote_work: Do you work remotely (outside of an office) at least 50% of the time?
12) tech_company: Is your employer primarily a tech company/organization?
13) benefits: Does your employer provide mental health benefits?
14) care_options: Do you know the options for mental health care your employer provides?
15) wellness_program: Has your employer ever discussed mental health as part of an employee wellness program?
16) seek_help: Does your employer provide resources to learn more about mental health issues and how to seek help?
17) anonymity: Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources?
18) leave: How easy is it for you to take medical leave for a mental health condition?
19) mental_health_consequence: Do you think that discussing a mental health issue with your employer would have negative consequences?
20) phys_health_consequence: Do you think that discussing a physical health issue with your employer would have negative consequences?
21) coworkers: Would you be willing to discuss a mental health issue with your coworkers?
22) supervisor: Would you be willing to discuss a mental health issue with your direct supervisor(s)?
23) mental_health_interview: Would you bring up a mental health issue with a potential employer in an interview?
24) phys_health_interview: Would you bring up a physical health issue with a potential employer in an interview?
25) mental_vs_physical: Do you feel that your employer takes mental health as seriously as physical health?
26) obs_consequence: Have you heard of or observed negative consequences for coworkers with mental health conditions in your workplace?
27) comments: Any additional notes or comments

B. Pre-processing

1) Data cleaning:

- Column names were changed into lower case.
- Age: Entries with invalid values (-29, 5, 329 etc.) were discarded. The data set was filtered to include only individuals within the age range of 18 to 70 years.
- Gender: The gender variable values were narrowed down to [Female, Male, Other].

- Timestamp: entries were filtered to include only those from year 2014 (due to very low variance).

2) Dropped columns: I have decided to drop the following columns due to specified problems:

- timestamp: low variance
- country: low variance
- state: correlated with country
- comments: unusable data

3) Dealing with missing data:

- self_employed: Nan entries were dropped because the number of them was negligible and it would be difficult to interpret missing values.
- work_interfere: Empty entries were substituted with 'Unknown'.

4) Data encoding: The encoding technique used in the analysis is a combination of label encoding, manual mapping and one hot encoding. The variables 'work_interfere', 'no_employees', and 'leave' are encoded using label encoding based on a predefined order or ranking. Each unique category within these variables was assigned a numerical value based on a predefined order or ranking. For example, in the 'work_interfere' variable, the categories 'Unknown' and 'Never' were encoded as 0, 'Rarely' as 1, 'Sometimes' as 2, and 'Often' as 3. Similarly, the 'no_employees' variable was encoded with numerical values ranging from 0 to 5, representing different employee count ranges. One hot encoding technique was used to encode gender variable.

C. *Exploratory Data Analysis*

	raw data	after pre-processing
number of entries	1259	1163
number of columns	27	25

TABLE I: Data before and after pre-processing

	0	1	2
timestamp	2014-08-27 11:29:31	2014-08-27 11:29:37	2014-08-27 11:29:44
age	37	44	32
gender	Female	M	Male
country	United States	United States	Canada
state	IL	IN	NaN
self_employed	NaN	NaN	NaN
family_history	No	No	No
treatment	Yes	No	No
work_interfere	Often	Rarely	Rarely
no_employees	6-25	More than 1000	6-25
remote_work	No	No	No
tech_company	Yes	No	Yes
benefits	Yes	Don't know	No
care_options	Not sure	No	No
wellness_program	No	Don't know	No
seek_help	Yes	Don't know	No
anonymity	Yes	Don't know	Don't know
leave	Somewhat easy	Don't know	Somewhat difficult
mental_health_consequence	No	Maybe	No
phys_health_consequence	No	No	No
coworkers	Some of them	No	Yes
supervisor	Yes	No	Yes
mental_health_interview	No	No	Yes
phys_health_interview	Maybe	No	Yes
mental_vs_physical	Yes	Don't know	No
obs_consequence	No	No	No
comments	NaN	NaN	NaN

TABLE II: Representative data sample

Due to the low variance in 'Country' variable I have decided to discard this column. Consequently 'state' variable was also dropped.

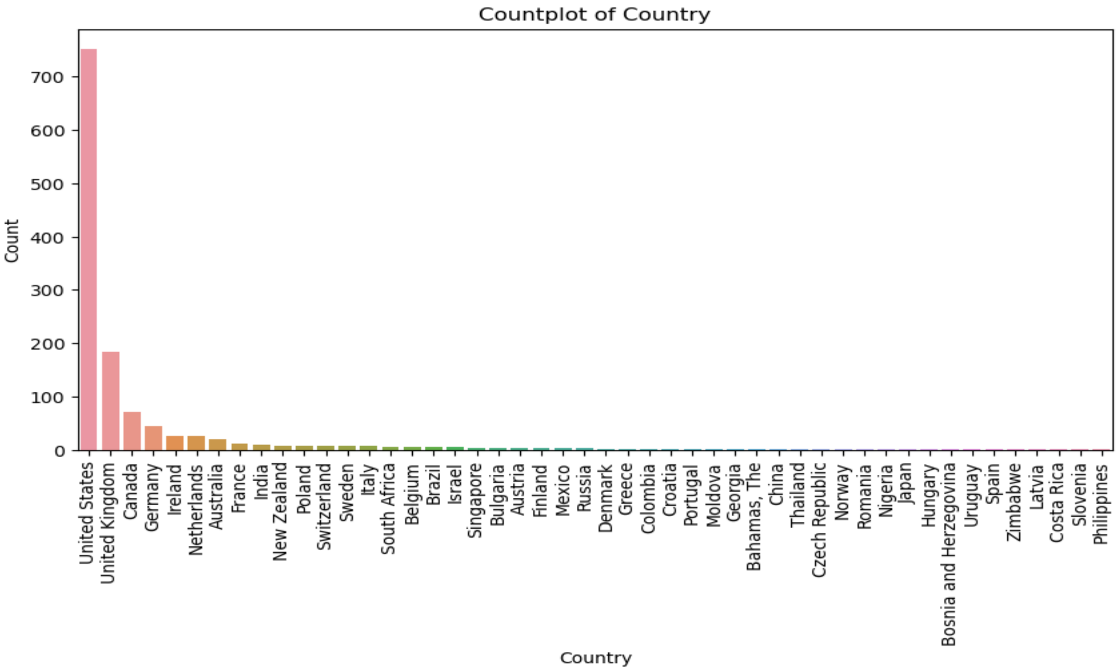

Fig. 1: Count plot of Country

In order to provide better human-readability, summary statistics will be shown in their primary form (before encoding). Detailed statistics for every feature are included in the appendix section.

I decided to divide the features into following categories:

*1) **Personal info**:* The rounded average age is equal 32.03.

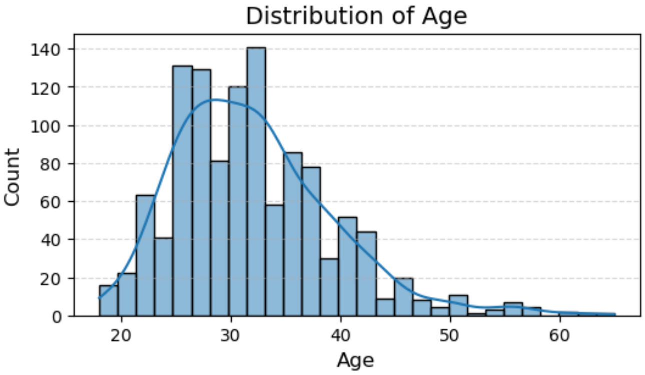

Fig. 2: Age distribution

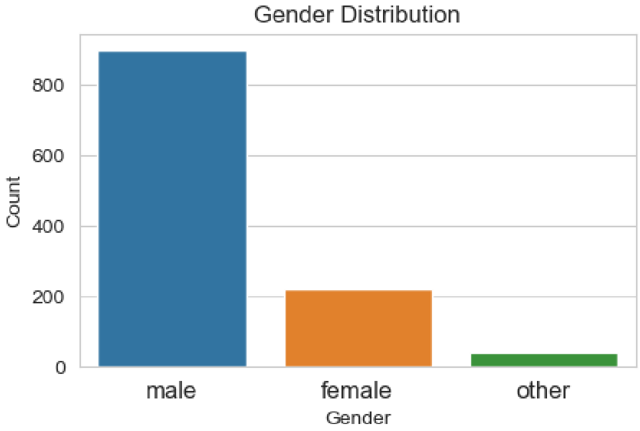

Fig. 3: Gender distribution

The survey findings reveal a higher representation of males, mirroring the greater male presence observed in tech companies.

*2) **Workplace environment**:* Most of the respondents weren't self-employed. Predominantly they didn't work remotely and their company belonged to tech industry.

*3) **Care options**:* The main conclusion from this part of the survey is that generally people are unaware about their care options situation at work. Moreover employers tend to pay little attention to providing resources to learn more. Described stigma may contribute to detrimental effects on ones mental well-being.

*4) **Attitudes towards mental health**:* Generally people tend to be more concerned about their mental than physical health. Surprisingly, respondents prefer to discuss their psychological problem with their supervisors, rather than with coworkers.

*5) **Mental health state**:* From chart below we can conclude that feature 'treatment' and 'work_interfere' are strongly correlated. People who seek treatment tend to be more influenced by mental health conditions at work. Moreover, individuals who did not seek treatment for mental health issues predominantly had values of 'Never' or 'Unknown' in the 'work_interfere' column. Therefore 'Unknown' and 'Never' were encoded with the same value - 0. In further analysis, a new output variable will be introduced: 'treatment_work_interfere_product'.

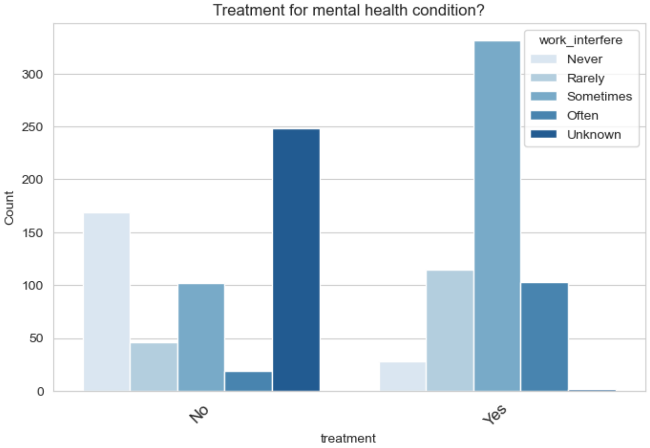

Fig. 4: 'Work_interfere' and 'treatment' chart

III. Experiments

To discern the pivotal factors influencing the output variables: 'treatment', 'work_interfere' and 'treatment_work_interfere_product' ('treatment' * 'work_interfere), I utilized the following techniques:

1) Correlation matrix analysis
2) Principal component analysis (PCA)
3) Linear regression

Analyzed models were created based on estimators available in scikit-learn.

A. Correlation matrix analysis

To enhance readability, only the highest correlations with outcome variables (excluding correlations with themselves) are presented in tables III, IV, V:

Feature	Importance
family_history	0.334235
mental_health_consequence	0.196258
obs_consequence	0.177886
care_options	0.144964
leave	0.129800

TABLE IV: Top correlations with 'work_interfere'

Feature	Importance
family_history	0.384161
care_options	0.235402
gender_female	0.185874
obs_consequence	0.159074
benefits	0.138294

TABLE III: Top correlations with ’treatment’

Feature	Importance
family_history	0.356483
care_options	0.203670
obs_consequence	0.182370
gender_female	0.163761
mental_health_consequence	0.140258

TABLE V: Top correlations with ’treatment_work_interfere_product’

B. Principal component analysis (PCA)

The second method used in the analysis was the PCA. After conducting multiple iterations of this method with different numbers of components and evaluating their results, I discovered that PCA1 and PCA2 exhibited the highest accuracy in terms of explained variance. As a result, in order to provide better clarity in the report only those components will be presented.

Explained variance for PC1 = 0.51373783.
Explained variance for PC2 = 0.45604862.

Feature	Importance
treatment	0.480958
family_history	0.376527
gender_female	0.305439
benefits	0.268926
care_options	0.267652

TABLE VI: Most important features for PC1

Feature	Importance
mental_health_consequence	0.378170
leave	0.191418
phys_health_consequence	0.169851
work_interfere	0.115581
obs_consequence	0.114568

TABLE VII: Most important features for PC2

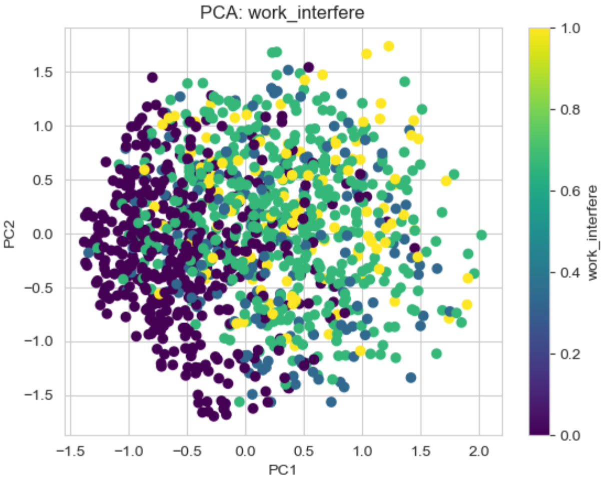

Fig. 5: PCA chart: work_interfere

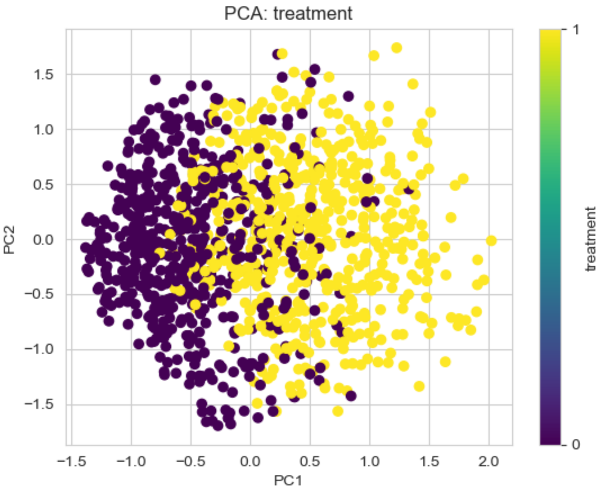

Fig. 6: PCA chart: treatment

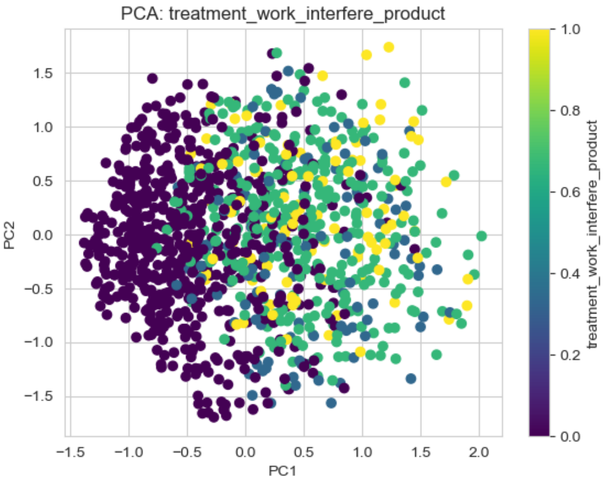

Fig. 7: PCA chart: treatment_work_interfere_product

C. Linear regression analysis

All data used for linear regression is divided into training and testing data in a 75:25 ratio, with a random seed set to 13 to ensure consistent data sampling.
As a measure of model accuracy, r-squared score will be presented in table VIII.

Feature	**R-squared score**
work_interfere	0.1961795217360064
treatment	0.16670596803757243
treatment_work_interfere_product	0.15178530861513928

TABLE VIII: R-squared score for every model

*1) **work_interfere**:* Most important features for determining work_interfere in this model are: family_history, mental_health_consequence, coworkers, care_options and obs_consequence.

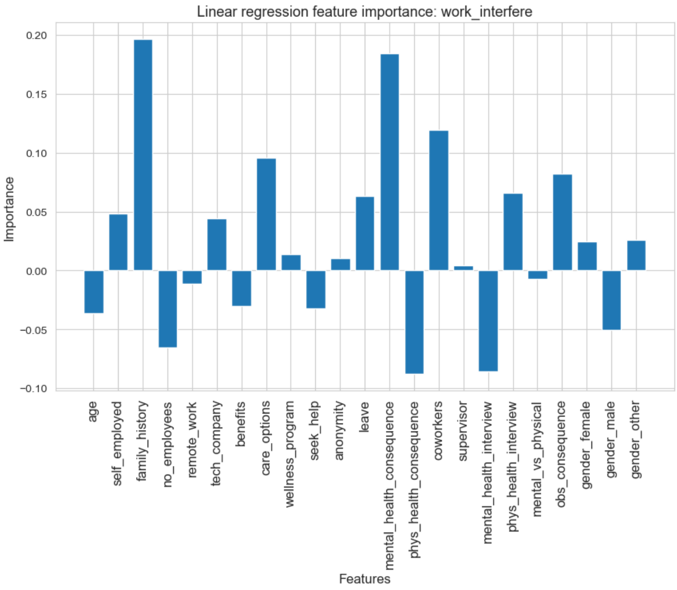

Fig. 8: Linear regression feature importance: work_interfere

*2) **treatment**:* Most important features for determining treatment in this model are: family_history, coworkers, age, care_options and mental_health_consequence.

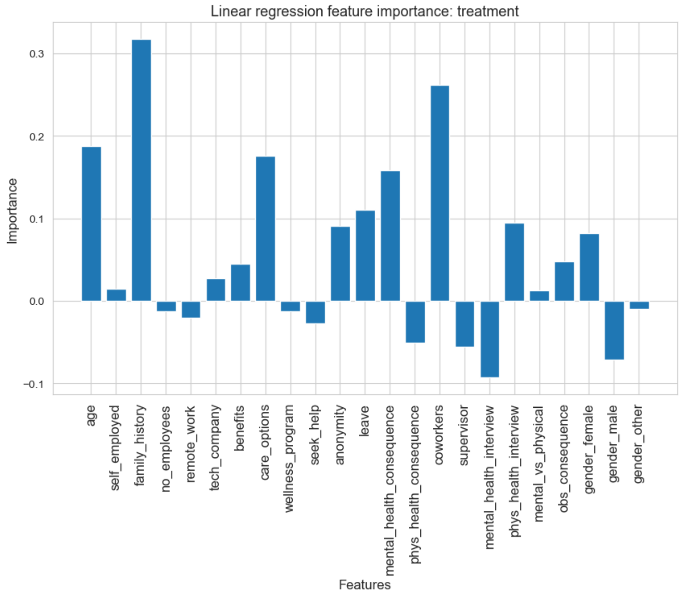

Fig. 9: Linear regression feature importance: treatment

*3) **work_interfere**:* Most important features for determining treatment_work_interfere_product in this model are: family_history, coworkers, mental_health_consequence, care_options and phys_health_interview.

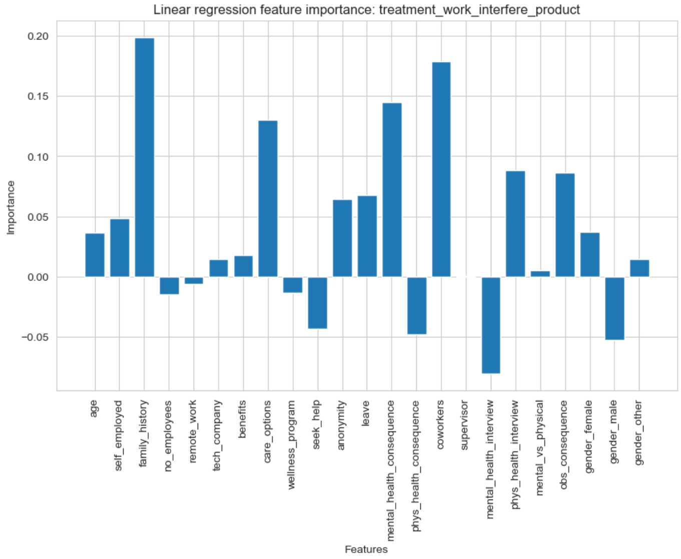

Fig. 10: Linear regression feature importance: treatment_work_interfere_product

IV. Conclusions

Despite the fact that experiments accuracy was quite underwhelming (relatively low R-squared score in linear regression model (about 20%) and medium explained variance in PCA (about 50%)), the outcomes were notably similar across all of them.

For every conducted experiment the strongest predictor for every output feature was family history of mental illness. This implies a strong genetic correlation or the impact family environment has on ones psychological state.

Secondly, the feature that appeared in every top 5 most important features was the awareness of care options provided by the employer. It suggests that when employees are well informed about available mental health resources they are more likely to seek treatment when needed. Without a doubt, this can be applied only when an appropriate program is already led. Therefore employers should emphasize this topic more often.

From the regression model, it can be concluded that employees who suffer from psychological problems are more prone to discuss it with their coworkers rather than with their supervisors.

Surprisingly, women generally tended to seek help more willingly than men.

Additionally, the feature indicating potential negative consequences of discussing a mental health issue with one's employer exhibits a strong positive correlation with the feature representing the extent to which an individual's mental state interferes with their work (work_interfere).

Lastly, features describing workplace environment did not play a significant role in determining respondents mental state.

Summing up, it is vital for company's management to encourage employees to seek treatment for mental health problems and reduce stigma associated with it. Increasing awareness concerning that topic is also essential. Finally, employers could improve communication with their employees in order to reduce the amount of detrimental impact work has on their mental state.

References

[1] "Mental health survey." [Online]. Available: https://www.kaggle.com/datasets/osmi/mental-health-in-tech-survey

V. Appendix

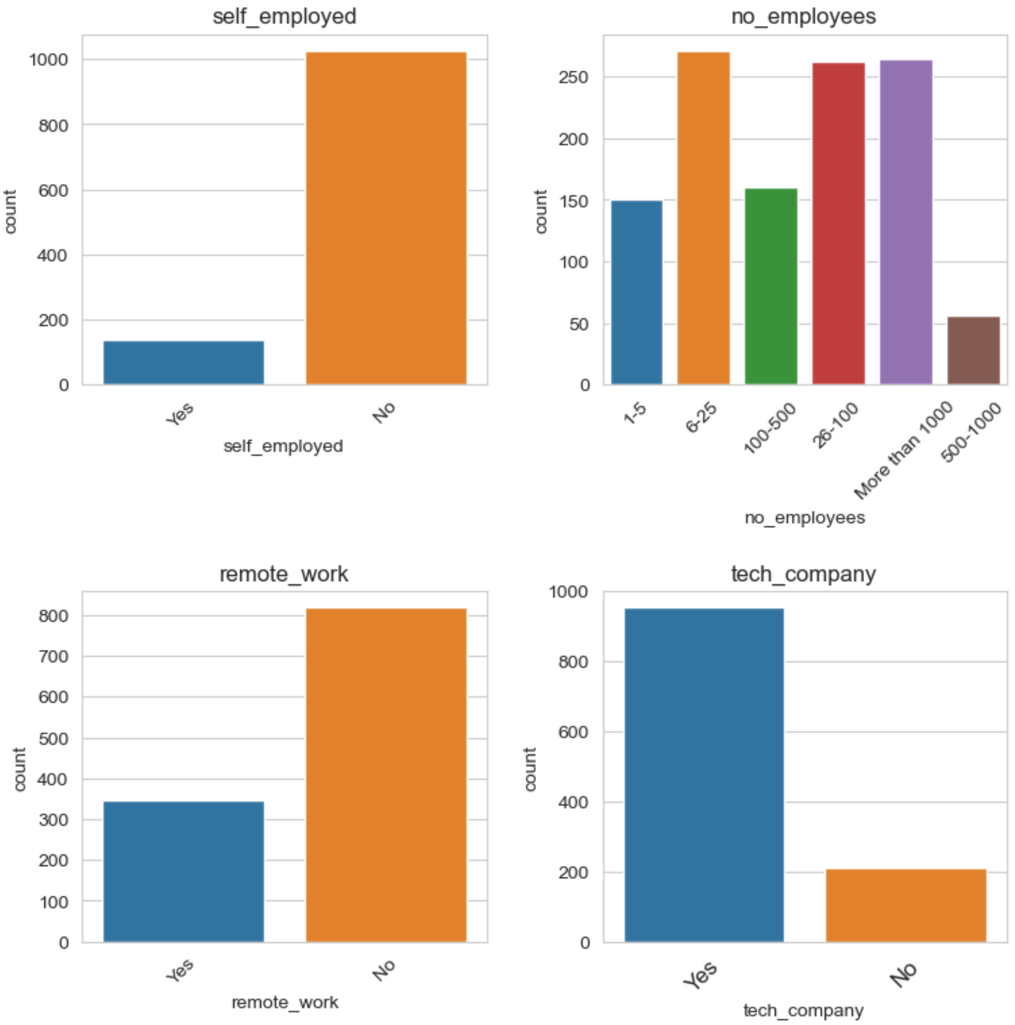

Fig. 11: Basic statistics for worplace environment features

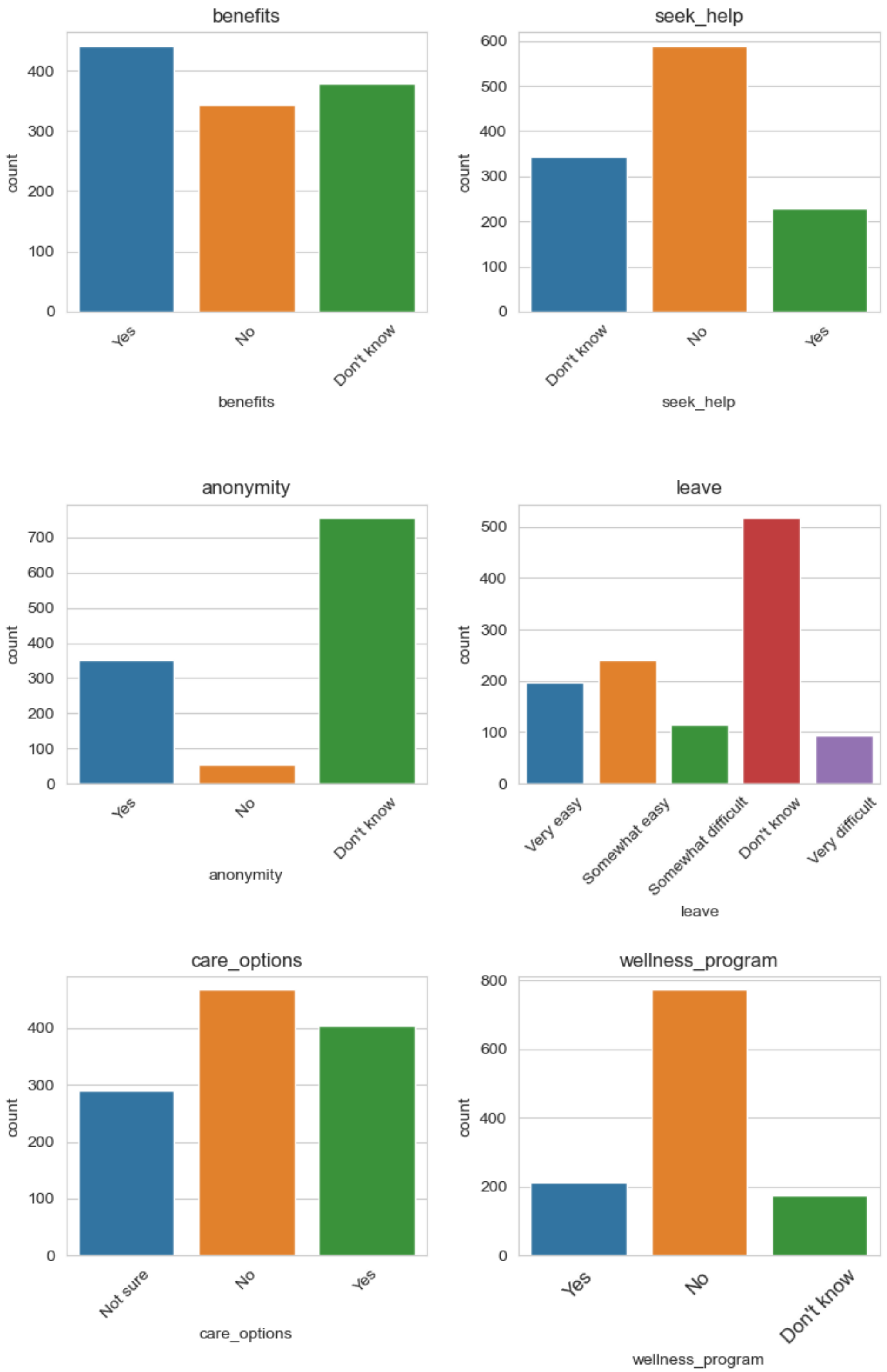

Fig. 12: Basic statistics for care options features

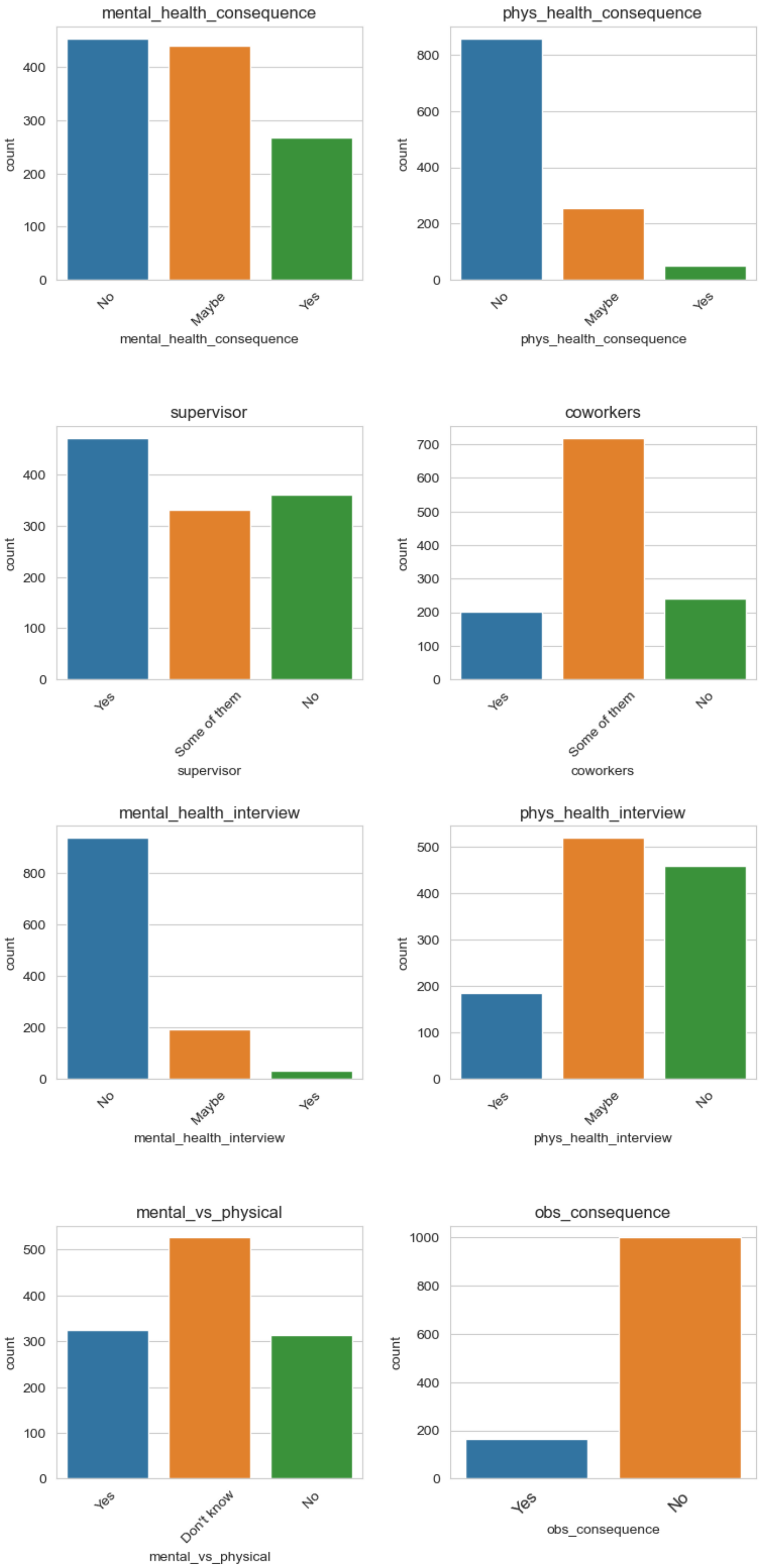

Fig. 13: Basic statistics for attitude features

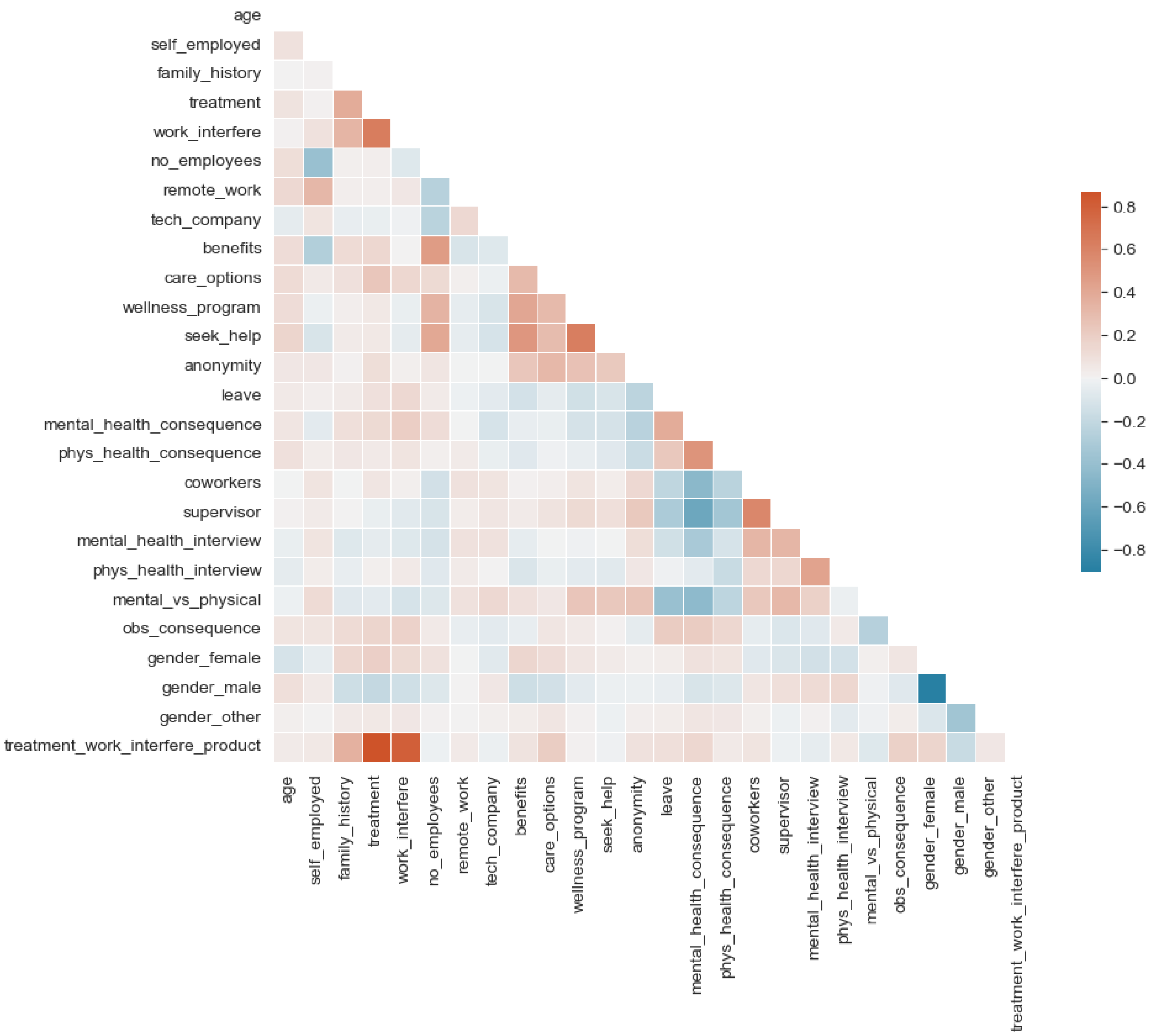

Fig. 14: Correlation matrix