Flourishing in the Workplace – An exploration *

Ayoub Chamakhi[1]

[0]emlyon business school
[1]Professor Dr. Franck JAOTOMBO

October 13, 2024

Abstract

This study examines the work-life balance and well-being of employees within a corporate setting by analyzing a dataset of 248 individuals, focusing on both personal and professional flourishing. Using Exploratory Data Analysis (EDA), we investigate categorical and numerical variables, exploring univariate and bivariate relationships among factors such as age, education, family status, mental flourishing in professional and private contexts, and positivity ratio. The analysis reveals significant associations between education level and emotional well-being, as well as strong correlations among professional flourishing, private flourishing, and flow. Bivariate analysis highlights the importance of family status and professional flourishing on personal well-being, while numerical correlations show that age, professional flourishing, and private flourishing are significant predictors of flow. Based on the EDA results, supervised modeling is employed to predict flow and positivity using multiple linear regression. The findings underscore the interconnectedness of professional and personal well-being factors and provide insights into predictors of employee flourishing.

Keywords: *Work-life Balance, Exploratory Data Analysis, Data Science, Machine Learning*

Link to Notebook and Dataset: Google Drive

[*] Individual project for the course Introduction to Machine Learning, under the supervision of Professor Dr. Franck JAOTOMBO.

Contents

1 Data Overview

A company would like to assess the work-life balance of its employees and has collected data from 248 individuals to gain insight into their well-being in both the personal and professional dimensions. The flourishing dataset contains 248 observations and 10 **variables**. The **ID** variable uniquely identifies each respondent, while **sex** denotes gender, with 1 representing males and 2 representing females. **famstatus** indicates the respondent's family status, categorized as Single, Separated/Divorced, or in a couple. The **education** variable reflects educational level, ranging from no high school degree up to five or more years of post-high school education. Additionally, **age** records the respondent's age. The data set also includes variables related to flourishing: **prof_cat** describes mental health in a professional context, with respondents classified as Languishing, Moderately Mentally Healthy, or Flourishing. Similarly, **priv_cat** reflects mental health in a private context using the same three categories. **positivity** measures the positivity ratio, calculated from the average positive and negative emotions, with categories from Depressed to Emotionally Flourishing based on the ratio. Furthermore, **prof_quant** and **priv_quant** provide quantitative scores for total flourishing levels in professional and private contexts, respectively. Lastly, the **flow** variable indicates the respondent's flow score, which assesses their sense of performance, mastery, and focus in various activities.

Among these 248 observations, we ensure that there are no missing values or incorrectly entered answers. We then proceed to encode each variable into its correct category, creating a numerical version of **education** since it has more than five modalities. We will retain the categorical version for further analysis. This results in five numerical variables: **age**, **education_num**, **pro_quant**, **priv_quant**, and **flow**. There are also five categorical variables that need to be transformed from numerical to categorical: **positivity**, **sex**, **pro_cat**,

priv_cat, **famstatus**, and **education_cat**. We also note that **pro_quant** and **priv_quant** have their versions in categorical form, which can lead to redundant information and cause multicollinearity.

2 Descriptive Statistics

We will start by performing an Exploratory Data Analysis to summarize the main characteristics of the data and uncover existing patterns.

2.1 Univariate Analysis

2.1.1 Categorical Variables

Categorical variables are analyzed using summary tables, pie charts, and bar charts. The **education_cat** variable shows that 44.35% of respondents have completed five or more years of post-secondary education, with lower percentages in the remaining categories, indicating a relatively high level of educational attainment. In the **sex** variable, 60.89% of the respondents are female, suggesting a slight gender imbalance in the sample. For **famstatus**, 69.76% of respondents are in a couple, indicating stable family structures which might influence the private life well-being. The **pro_cat** variable reveals that 64.11% are moderately mentally healthy, while 20.56% are flourishing and 15.32% are languishing. In **priv_cat**, the distribution is similar, with 66.13% moderately mentally healthy, indicating overall positive mental health. However, the **positivity** variable shows 59.68% of respondents are languishing, highlighting a concerning level of emotional distress among the population. The high level of education may correlate with better job productivity and mental health outcomes. The sample is slightly female-presented, potentially influencing responses to other variables. The majority of respondents are in stable relationships, providing emotional support. While a large proportion of individuals are classified as "Moderately Mentally Healthy," there is a notable presence of those who are languishing or struggling emotionally, which can help us understand the sources of well-being through the nuances (variations) compared to those who are healthier.

2.1.2 Numerical Variables

Numerical variables are analyzed using frequency tables, histograms, and box plots. The frequency analysis of the **age**, **pro_quant**, **priv_quant**, and **flow** variables indicates generally normal distributions. The **age** distribution shows a concentration of respondents in their 30s and 40s, with 14 respondents aged 37 (8.06%) and 18 respondents aged 33 (7.26%). The **pro_quant** variable has the highest frequency at 43, accounting for 4.44%, while the **priv_quant** variable also shows its peak frequency at 61, representing 4.44%. Both variables demonstrate a stable level of professional and private engagement among participants, with values primarily clustering around the mid-range. The **flow** variable exhibits a normal distribution as well, with the most common responses reported for 37 (8.06%) and 33 (7.26%) flow experiences. These distributions suggest a balanced representation across different levels of engagement, which may enhance the reliability of findings related to well-being and mental health outcomes. We note that the symmetrically distributed **prof_quant** and **priv_quant** have similar distribution as their categorical counterparts.

2.2 Bivariate Analysis

2.2.1 Mixed Categorical and Numerical

The mixed categorial and numerical variables are analyzed using ANOVA, difference in group means along with d-cohen for statistical significance in those differences effect size, and grouped boxplots. These analyses revealed significant differences among several variables. **famstatus** significantly influenced **priv_quant**, with couples scoring 57.26, separated/divorced individuals at 54.26, and singles at 52.00. The **pro_cat** also exhibited significant relationships, particularly with **age** (averages: Flourishing at 45.49, Moderately Mentally Healthy at 40.92, and Languishing at 39.76), and **pro_quant** (scores: Flourishing at 69.24, Languishing at 31.16, Moderately Mentally Healthy at 49.29). **pro_cat** further impacted **priv_quant** (scores: Flourishing at 66.63, Moderately Mentally Healthy at 55.16, and Languishing at 45.16) and **flow** (averages: Flourishing at 38.10, Moderately Mentally Healthy at 32.23, and Languishing at 29.08). The **positivity** variable had significant effects on **age** (averages: emotionally flourishing at 48.60, moderately emotionally healthy at 43.96, languishing at 40.25, and depressed at 40.37), **pro_quant** (scores: flourishing at 63.90, moderately emotionally healthy at 60.66, languishing at 48.20, depressed at 36.93), **priv_quant** (scores: flourishing at 67.70, moderately emotionally healthy at 64.48, languishing at 54.15, and depressed at 43.10), and **flow** (averages: flourishing at 38.25, moderately emotionally healthy at 36.32, languishing at 32.24, and depressed at 27.30). Being in a relationship significantly influences the **priv_quant** score, suggesting it may enhance overall well-being. Increased maturity, indicated by age, correlates with higher scores in both **pro_quant** and **priv_quant**, potentially reflecting the stability that comes with relationships and work as one ages. Additionally, there is a clear interconnection among the happiness indicators: **pro_quant**, **priv_quant**, **flow**, and **positivity**, all of which significantly impact one another.

2.2.2 Both Categorical

In analyzing both categorical variables, we use cotingency tables, cramer's V to measure the effect size of such independence, finally we draw side-by-side bar charts and stacked one. The analysis reveals significant associations among various categorical variables related to mental well-being. Firstly, there is a notable link between **education_cat** and **positivity** ($\chi^2 = 29.09, p - value = 0.0166$), suggesting that higher educational attainment positively influences emotional well-being. Furthermore, the relationship between **famstatus** and **priv_cat** is significant ($\chi^2 = 11.14, p - value = 0.0251$), confirming that family dynamics play a crucial role in determining private flourishing scores. Additionally, a strong association exists between **pro_cat** and **priv_cat** ($\chi^2 = 70.96, p-value < 0.0001$), highlighting the interconnectedness of professional and private mental health statuses. The relationship between **pro_cat** and **positivity** is also significant ($\chi^2 = 90.22, p - value < 0.0001$), demonstrating that professional flourishing directly impacts emotional flourishing. Lastly, there is a significant association between **priv_cat** and **positivity** ($\chi^2 = 71.63, p - value < 0.0001$), reinforcing the idea that private well-being is closely linked to overall emotional states.

2.2.3 Both Numerical

Last but not least, both numerical variables are analyzed using correlation tables, Pearson r-values, as normality was observed in these variables, and scatterplots to visualize the strength and type of relationships. The analysis reveals several significant Pearson correlations between key variables. Age is positively correlated with **pro_quant** ($r = 0.23, p - value < 0.001$), **priv_quant** ($r = 0.15, p - value < 0.02$), and **flow** ($r = 0.37, p - value < 0.0001$), indicating that older individuals tend to score higher in both professional and private flourishing, as well as in flow. The **pro_quant** and **priv_quant** scores are strongly correlated ($r = 0.63, p - value < 0.0001$), showing that professional and private flourishing are closely related. Additionally, **pro_quant** is significantly associated with **flow** ($r = 0.53, p - value < 0.0001$), and **priv_quant** is also positively correlated with **flow** ($r = 0.41, p - value < 0.0001$), suggesting that both professional and private flourishing contribute to higher flow scores.

2.3 EDA results

Based on the previous results, we will focus on modeling **positivity** as the categorical outcome and **flow** as the numerical outcome, given that they are constructed from other well-being indicators and demographic factors. For the predictors, **priv_quant** and **pro_quant** will be used instead of their categorical counterparts (**priv_cat** and **pro_cat**), as the quantitative variables introduce more variability and show high Variance Inflation Factor (VIF) scores, confirming their multicollinearity. Furthermore, the analysis using Cohen's d confirms significant differences in **flow** scores among the various **positivity** categories. Specifically, the results indicate strong effect sizes between depressed and emotionally flourishing individuals, as well as between depressed and moderately emotionally healthy individuals, highlighting a clear distinction in their **flow** experiences. The findings suggest that those who are emotionally **flourishing** and moderately emotionally healthy experience notably higher **flow** scores compared to those who are depressed.

3 Modeling

3.1 Supervised Models

We will start with supervised models, assuming a relationship between the various predictors we've identified and a single output variable, which may be either numerical or categorical. This approach will enable us to predict outcomes based on the specified input features, allowing us to understand how changes in the predictors can influence the target variable.

3.1.1 Multiple Linear Regression

We start by looking at the significant relationship between the flow variable and others identified in the previous Exploratory Data Analysis (EDA). Given the results of the prior correlation, we will use **age** ($r = 0.367$), **pro_quant** ($r = 0.53$), **positivity** (p-value < 0.0001), and **priv_quant** ($r = 0.4$) as predictors for modeling flow. We employed an iterative approach to determine the optimal predictors for our model, ultimately removing the positivity variables identified as relevant through the Cohen's d analysis, as their inclusion resulted in a reduced Adjusted R$^{\mathbf{2}}$. The code implements a structured method to model the relationship between selected predictors **age**, **pro_quant**, and **priv_quant** and the output variable **flow** using linear regression. The dataset is split into training (80%) and test (20%) sets to validate the model's performance on unseen data. Feature scaling is applied using StandardScaler to ensure that all predictors contribute equally to the model. A Linear Regression model is then instantiated and fitted on the scaled training data, allowing the algorithm to learn the relationship between predictors and the output.

After fitting the model, predictions are generated on the test set, and the model's performance is evaluated using two key metrics: Adjusted R$^{\mathbf{2}}$ and normalized RMSE. The results indicate an Adjusted R$^{\mathbf{2}}$ of 0.3121, suggesting that approximately 31.21% of the variability in flow can be explained by the selected predictors. The normalized RMSE value of 0.1584 indicates that the model's predictions are, on average, about 15.84% off from the mean flow values, which suggests moderate accuracy.

The regression analysis results indicate that among the predictors, **priv_quant** has the most substantial effect on flow, with a coefficient of 2.8614, suggesting that for each unit increase in **priv_quant**, flow increases by approximately 2.86 units. Following closely is **pro_quant**, which has a coefficient of 1.6813, indicating a flow increase of about 1.68 units for each unit increase in **pro_quant**. Both **priv_quant** and **pro_quant** are statistically significant, with p-values less than 0.001. In contrast, **age** has a coefficient of 0.5430, representing a smaller increase in flow of approximately 0.54 units for each unit increase in **age**; however, this effect is not statistically significant ($p = 0.262$). The constant term is 32.7879, which indicates the expected flow value when all predictors are zero. Overall, while the model captures some of the relationships, there is room for improvement, potentially through exploring alternative modeling techniques.

3.1.2 KNN Regression

We implemented the K-Nearest Neighbors (KNN) regression model to predict the variable **flow** based on the predictors **age**, **pro_quant**, and **priv_quant**. The model achieved an Adjusted R$^{\mathbf{2}}$ of 0.2197, indicating that approximately 21.97% of the variability in flow can be explained by the selected predictors. The normalized RMSE value of 0.1687 suggests that the model's predictions deviate from the mean flow values by an average of about 16.87%.

As a result, while both models have their merits, the Linear Regression model demonstrates a better fit and predictive performance for this dataset compared to the KNN model, warranting further exploration into additional predictors or modeling techniques to enhance prediction accuracy.

3.1.3 Logistic Regression

The model applies multinomial logistic regression to predict employee positivity levels based on variables identified as significant in the previous EDA. We selected **age**, **pros_quant**, **priv_quant**, **flow**, and **education_cat** as predictors because the EDA revealed strong relationships between positivity and these variables. The same steps used in Multiple Linear Regression were applied, but with One-Hot Encoding on categorical variables, creating a reference category to enable the use of Logistic Regression. After training the model, predictions were made on the test set, with an accuracy of 74%. The classification report reveals that while the model performs well overall, it struggles with smaller or less frequent categories, such as ”depressed” and ”emotionally flourishing,” where precision and recall are lower. This suggests that while the model captures major patterns, improvements are needed for better classification of underrepresented groups.

3.1.4 KNN Classifier

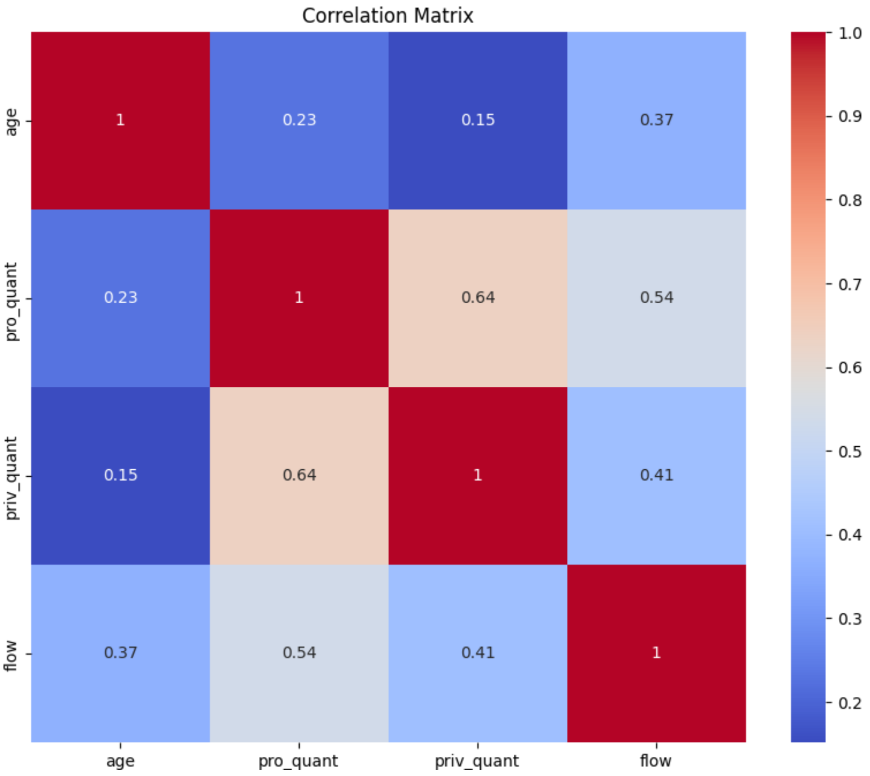
We apply the same steps of Logistic Regression, while adding a loop over values of n_neighbors from 1 to 20 to find the optimal number of neighbors for the KNN classifier. In each iteration, the model is trained on the training set, predictions are made on the test set, and accuracy is calculated. The best n_neighbors value (n = 5), corresponds to the highest accuracy, is stored and used on the model. While KNN and logistic regression show similar overall accuracy (74%), logistic regression provides a more balanced performance across all classes, whereas KNN struggles with minority classes like ”depressed” and ”emotionally flourishing,” overfitting the majority class (”languishing”).

3.2 Unsupervised Models

Following this, we will implement unsupervised models, which operate under the assumption that there are no predefined labels for the output. Instead, these models will help us identify patterns, groupings, or structures within the data, uncovering hidden relationships among the predictors without the need for specific target values.

3.2.1 Principal Component Analysis

We will start by applying Factor Analysis to reduce the dimensionality of the data used in the K-Means Clustering in the next subsection. Several assumptions must be met: the variables should be measured on an interval or ratio scale (**age**, **education_num**, **pro_quant**, **priv_quant**, **flow**), and the variables should be approximately normally distributed, as confirmed in the Exploratory Data Analysis. As a rough guideline, there should be five times as many observations (sample size) as there are variables ($10 variables * 5 = 50 < 248$).

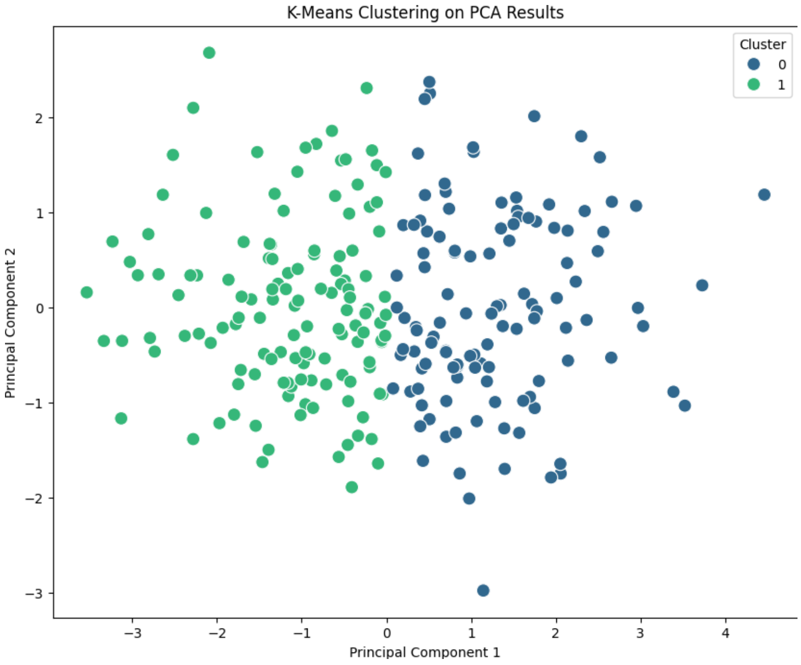


We begin by running a correlation matrix between these variables, eliminating education_num due to its very low correlation. The Bartlett test of sphericity rejects the null hypothesis (H0) with a ($\chi^2 = 248.67, p - value < 0.0001$) which allows us to proceed with factor analysis. Additionally, the KMO measure of sampling adequacy is 0.67, which is close to 1. The scree plot indicates that the variance explained starts to drop sharply after two factors, making this our optimal choice. There are two important groups that explain 78.5% of the variation in the data. 55.28% is explained by the first group, whereas, 23.23% is explained by the second.

Finally, the rotated PCA loadings, using Varimax rotation, reveal two distinct groups. PC1, that we will name well-being group, has a high loading on **pro_quant** and **priv_quant**. PC2, that we will name experienced group, has a high loading on **age** and **flow**.

3.2.2 K-Means Clustering

The K-Means clustering on the PCA-transformed data reveals two distinct groups.



Cluster 0 (blue) comprises individuals with higher well-being scores, indicating greater professional and personal flourishing, potentially linked with higher experience levels. Cluster 1 (green) includes individuals with lower well-being scores, suggesting they might face challenges in these areas, though they vary in experience. This clustering aligns with our PCA results, which showed two main components: well-being and experience, confirming a clear segmentation based on these factors and providing a basis for tailored interventions to address the specific needs of each group.

4 Conclusion

The analysis shows that professional and private flourishing are closely intertwined, with a strong association between positive private life experiences and better performance and well-being at work. Furthermore, the positive impact of higher education on mental and emotional health suggests that ongoing professional development and educational programs may benefit employees' well-being, especially for those lacking experience. Companies should consider implementing relationship support programs and encouraging work-life balance practices, as these can significantly contribute to private and professional flourishing. Additionally, targeting emotionally languishing employees with tailored mental health resources and well-being initiatives could help elevate their professional performance and overall flow experiences. Creating a supportive workplace culture that emphasizes holistic well-being could lead to improved productivity and greater employee satisfaction.

5 Bibliographie

References

M. Csikszentmihalyi. *Flow: The psychology of optimal experience*. Harper Perennial Modern Classics. Harper [and] Row, nachdr. edition, 2009. ISBN 978-0-06-133920-2.

B. Fredrickson. *Positivity: top-notch research reveals the 3-to-1 ratio that will change your life*. Three Rivers Press, 1st pbk. ed edition, 2009. ISBN 978-0-307-39374-6. OCLC: 419801686.

B. L. Fredrickson. The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions. 56(3):218–226, 2001-03. ISSN 1935-990X, 0003-066X. doi: 10.1037/0003-066X.56.3.218. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/0003-066X.56.3.218.

C. L. M. Keyes. The mental health continuum: From languishing to flourishing in life. 43(2):207, 2002-06. ISSN 00221465. doi: 10.2307/3090197. URL http://www.jstor.org/stable/3090197?origin=crossref.

J. Nakamura and M. Csikszentmihalyi. *The Concept of Flow*, pages 239–263. Springer Netherlands, 2014. ISBN 978-94-017-9087-1 978-94-017-9088-8. doi: 10.1007/978-94-017-9088-8_16. URL https://link.springer.com/10.1007/978-94-017-9088-8_16.

OpenAI. Chatgpt: A conversational ai model, 2024. URL https://www.openai.com/chatgpt. Accessed: 2024-10-12.