Unsupervised learning on cooperative data

Alon Luboshitz - 312115090, Orit Hason Weiss - 322757048

May 29, 2025

1 Abstract

Our world's economy is largely driven by companies and corporations, making them of high interest. Understanding key features affecting one cooperative can improve economy, production, health, and far more. Here, we analyzed data from a single cooperative, focusing on various aspects of employee work, such as monthly working hours, remote work status, absences from work, and more. Using unsupervised methods such as dimensionality reduction, clustering, and feature enrichment, we identified key features differentiating between employee groups. For instance, we demonstrated that the data has a 3 main cluster structure, with one group exhibiting significantly higher absence rates than the others. This group is further characterized by low meeting attendance, lower salaries, and reduced productivity. This characterization enables cooperatives to manage employees more effectively and develop strategies to enhance productivity. We hope that our insights can shed light on better management. Our analysis can be found in the GitHub repository: `https://github.com/OritHason/Unsupervised-Learning-Project/tree/dev`.

2 Introduction

The global outbreak of COVID-19 in early 2020 [1] prompted an unprecedented transition to remote work for millions of employees worldwide. In Israel, as in many other countries, the pandemic served as a large-scale natural experiment, enabling researchers and policymakers to observe how remote work affects job performance and employee well-being. Several studies indicate that remote work during the pandemic did not compromise productivity—in some cases, it even improved it. For instance, a report by "TheMarker" found a 13% increase in employee performance while working from home, without any significant decline in work quality [2]. Additional surveys show that a substantial proportion of employees felt they were equally or more productive working remotely, compared to their pre-pandemic routines [3]. These studies collectively suggest that remote work can positively influence job satisfaction and work-life balance, though the outcomes may vary based on job roles, industry, and the level of organizational support provided. Our dataset contains 10,000 records of corporate employees across various departments, focusing on work hours, job satisfaction, and productivity performance [4].

3 Methods

3.1 Categorical feature processing and Multiple correspondence analysis

To handle categorical features for dimensionality reduction and model application, we used one-hot encoding followed by Multiple Correspondence Analysis (MCA) [5] to reduce the dimensionality of the categorical data to four components. After that, we combined the reduced categorical components with the numeric features for an additional round of dimensionality reduction.

3.2 Data standardization

To maintain stability and reduce the influence induced by feature scale, we standardize the data using sklearn StandartScaler function by removing the mean and scaling to unit variance.

3.3 dimensionality reduction and feature importance

Due to the high dimensions of the data, we performed principal component analysis (PCA) [6] and t-distributed stochastic neighbor embedding (t-SNE) [7] with two components. For the t-SNE algorithm, we used the sklearn library with the parameters: perplexity = 30, n_itter = 1000. For PCA feature importance, we analyzed the projection of features onto the first two principal components.

3.4 Guasine Mixture model and hierarchical clustering enrichment

We implemented the Gaussian Mixture Model (GMM)[8] with the sklearn library and hierarchical clustering with the scipy library. The Euclidean distance was calculated between points. We calculated the cluster enrichment by obtaining the cluster labels and measuring the proportion of the given label per cluster.

3.5 ANOVA testing

We applied an ANOVA statistical test[9] over 10 randomly initialized GMMs. Each model was fitted using the first two principal components from PCA, applied to both the full dataset and the numeric features subset. After obtaining the cluster labels of each data point, the standardized feature values were aligned, and the cluster labels grouped feature values. The feature differentiation values were calculated using ANOVA test, and the F-statistics[10] and p-value were averaged over the 10 models. The statistics for the numeric features only and whole data are supplied in (Table1, Table2) respectively.

4 Results

The data has a hierarchical structure of 3 clusters with 3 inner clusters based on *Remote_work*, *Work_life_balance*, and *Job_level*, respectively. We reduced the data dimensions using principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) to investigate the data structure and plotted the results over 2 dimensions. The PCA results demonstrated no separation between samples (Figure 1A). Interestingly, the t-SNE plot reveals three main clusters, each with a distinct internal structure (Figure 1B). To investigate the cause of this further, we plotted each feature value on top of the reduced dimensions. We observed 3 clusters based on the *Remote_work* feature (Figure 1C). Furthermore, we observed inner clusters within the *Remote_work* groups, driven by *Work_life_balance* and *Job_level* (Figure 1D–E). This hierarchical structure was validated in the t-SNE plot by removing *Remote_work*, revealing a 3×3 clustering pattern based on the remaining features (Figure 1F–G). When coloring the PCA results through the *Remote_Work* feature, we did not observe a specific clustering rather a split of the plot into areas (Figure 1H). To further validate this structure, we cross-validated the dimensionality reduction structure by fitting a hierarchical clustering (HC) to the data with 3 clusters. We ran 3 different HCs in which we kept all features and excluded *Remote_work* together with *Work_life_balance* or *Job_level.* Then, we calculated the enrichment of each label per cluster based on the HC (Methods 3.4). We demonstrated that the HC can capture the structure observed through the t-SNE algorithm. When using all the features in the data, the HC output matches the *Remote_work* labels up to 99% (Figure 1I). Furthermore, excluding *Remote_work* together with *Job_level* capture precisely the *Work_life_balance* (Figure1J). Lastly, excluding *Remote_work* together with *Work_life_balance* matches *Job_level* labels (Figure 1K). This demonstrates a good alignment between the HC and t-SNE outputs.

Next, we wanted to decipher other features contributing to this structure. First, we excluded *Remote_work*, *Work_life_balance*, and *Job_level* from the data and ran the same analysis as before. Sadly, removing the mentioned features disrupted the previous alignment between the labels and clusters obtained through the HC algorithm (Figure2A–C). We compared this analysis with the Guasine mixture model (GMM) and concluded the same result. To validate which clustering algorithm captures the structure, we evaluated the silhouette over 2-9 clusters with and without applying PCA and with and without categorical features. For the GMM algorithm, 3 clusters achieved a superior average silhouette score, strengthening the 3 main cluster structure observed through the dimensionality reduction. Moreover, we demonstrate peaks in the score for 6 and 9 clusters, strengthening the inner cluster structure observed in the t-SNE (Figure2G). Contrary, the HC algorithm silhouette score drops as the cluster number increases (Figure2G). Without applying PCA in advance, the score for 3 clusters was not superior, except for the HC algorithm with categorical features. We suggest that the high dimensions and feature interactions are misleading the model and thus preventing its accurate convergence (Figure2H). Furthermore, the silhouette scores of the GMM were superior to the HC, and after applying PCA, the scores were significantly superior up to 4 folds, thus we continue to evaluate the GMM clustering with 3 clusters and PCA (Figure2I).

Numeric features contributing to data differentiation. To decipher the features contributing to this clustering structure, we first excluded the *Remote_Work* feature and ran 10 randomly initialized GMM algorithms on the first 2 PCA components and conducted an ANOVA test between cluster groups to validate feature differentiation (Methods3.5). We demonstrated that without categorial features, most of the numeric features are differentiated between the clusters except *Years_at_Company* (Table1). Contraversy, when using the categorial features, the numeric feature *Absences_Per_Year* achieved superior f-statistics of more than 2-folds (Table2). Furthermore, in both cases, the different initiations did not change the ANOVA values, indicating on strong convergence, controversy to applying this analysis without PCA. Next, we plotted the feature distribution based on the GMM clustering with and without the categorical features. We first noticed that the numeric features differed similarly with and without the categorical features, indicating their effect on the clustering

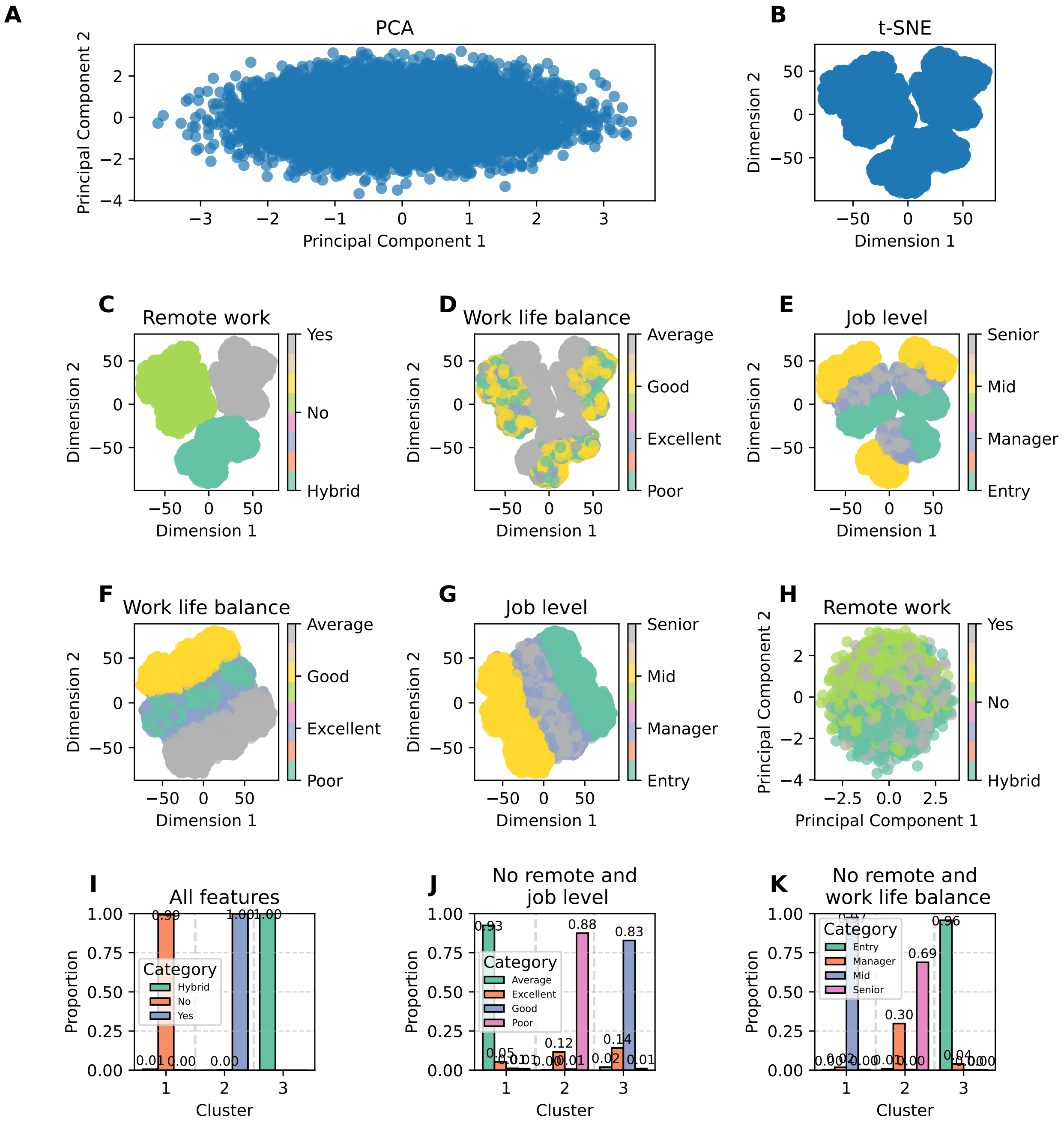

Figure 1: **Dimensionality reduction reveals hierarchical clustering structure.** (A) PCA reduction on the whole data. (B) t-SNE reduction on the whole data. (C–E) t-SNE reduction on the whole data colored by categorical features. (F–G) t-SNE reduction excluding the *Remote_Work* feature, colored by categorical features. (H) PCA reduction on the whole data colored by *Remote_work*. (I–K) Label enrichment per cluster based on the HC clustering. The titles correspond to features included in the data. Labels based on *Remote_work*, *Work_life_balance* and *Job_level* for (I–K) respectively.

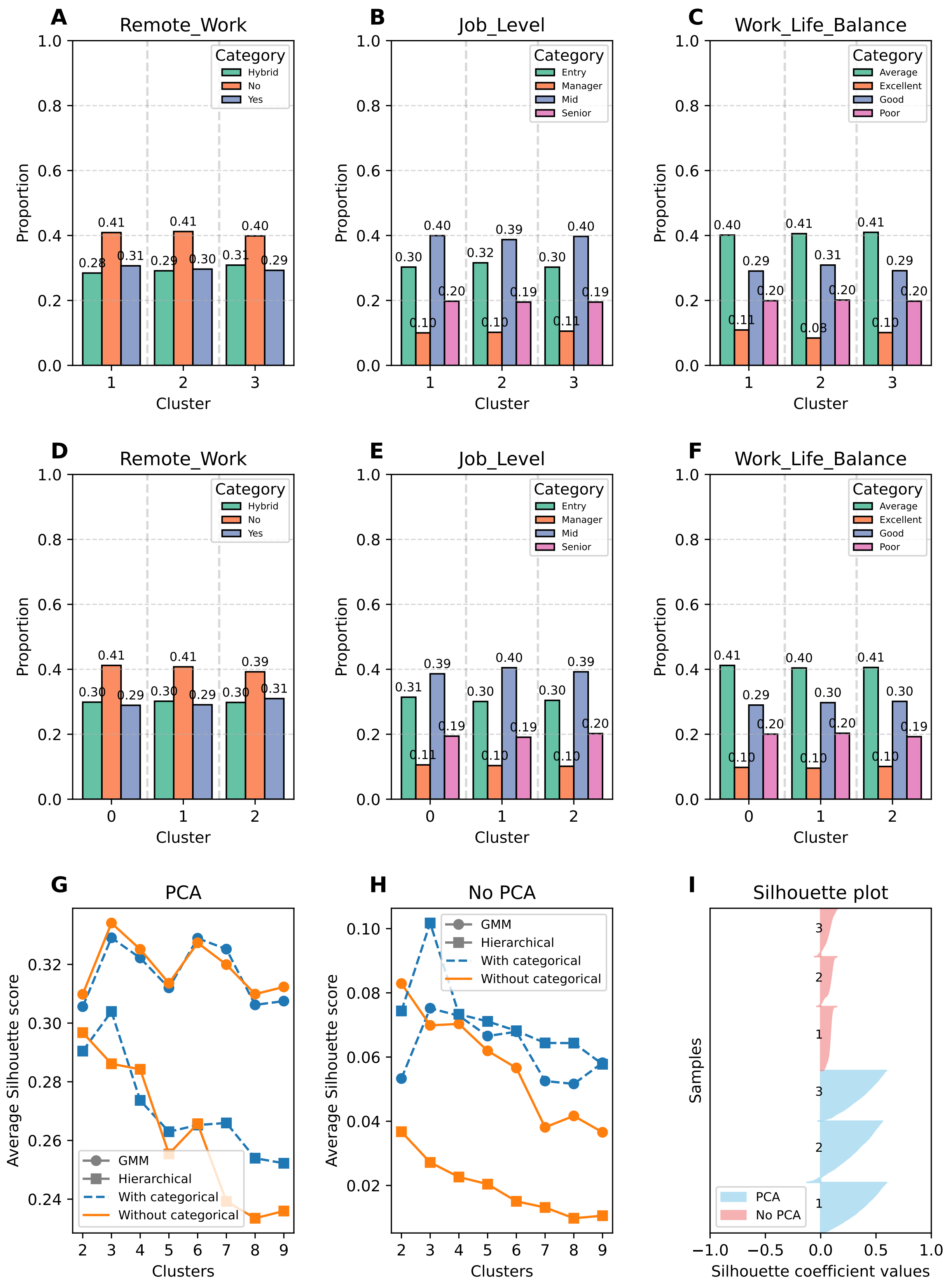

Figure 2: **Clustering reveals 3-cluster structure through latent space.** (A–F) Alignment of feature labels to cluster labels. The three mentioned features were removed prior to the clustering. Each title is the label aligned. (A–C) Alignment through Hierarchical clustering. (D–F) Alignment through GMM clustering. (G–H) Average silhouette score for 2-9 clusters based on Hierarchical clustering (squares) and GMM (circles). (G) PCA is applied to the data before the clustering. (H) No PCA is applied. (I) Silhouette scores per cluster with and without PCA based on GMM.

structure. We observed the same tendency among the following features: *Meetings_per_Week*, *Job_Satisfaction*, *Tasks_Completed_Per_Day*, revealing a strong connection between a high number of meetings, tasks completed and job satisfaction (cluster 1, Figure3A-C). Furthermore, the features: *Productivity_Score*, *Annual_Salary* indicated on concurrence between productivity and annual salary, one pushing the other (cluster 0, Figure3D-E). Surprisingly, the 2 sets of mentioned features are not visually correlated, where high productivity and salary don't go hand in hand with tasks completed, meetings, and job satisfaction as one would expect. Furthermore, cluster 2 differentiated by high values of *Absences_Per_Year* (Figure3F) is characterized by low monthly hours worked, meetings, tasks, satisfaction, and productivity, strengthening feature dependencies intuition. Surprisingly, we demonstrate an inverse trend between *Monthly_Hours_Worked* and *Overtime_Hours_Per_Week*. We conclude that workers who spend much time at work don't need to do overtime, whereas workers who don't show up often will stay longer (Figure3G-H).

The features *Overtime* and *Productivity* are correlated in clusters but not in the data. Lastly, to investigate the linear relations and dependencies among features, we calculated the Pearson correlation between numeric features for each cluster and plotted these values on heatmaps. We demonstrated that clusters 0,2 are similar compared to cluster 1 whereas the feature *Monthly_hours* is dominating in cluster 0,2 and not in cluster 1 (Figure4A–C). This emphasizes the relation between monthly hours worked and other features. Furthermore, *Overtime* and *Productivity* are positively correlated in clusters 0,2 and not in cluster 1, indicating a separation between the groups based on concurrence between these features (Figure4A–C). Moreover, to investigate features contribution in generall, we plotted the features values on top of the PCA components and followed the outlined vectors by magnitude and direction (Methods3.3). We demonstrate that conclusions from this analysis overlap with the clusters differentiation, for example, the following feature pairs: *Absences_Per_Year*, *Meetings_per_Week* and *Annual_Salary*, *Job_Satisfaction* are in opposite directions respectively, strengthening the trends in the observed box plots (Figure4D). Furthermore, in the whole data, the features *Overtime* and *Productivity* are in negative directions, whereas in the clusters they are correlated (Figure4A–D). This outlines the differentiation based on these features.

Feature	**F-stat Mean**	**F-stat Std**	**F-stat Median**	**p-val Mean**	**p-val Std**	**p-val Median**
Monthly_Hours_Worked	1138.85	1.61×10^{-13}	1138.85	0.0	0.0	0.0
Meetings_per_Week	1128.65	2.16×10^{-13}	1128.65	0.0	0.0	0.0
Absences_Per_Year	967.51	0.0	967.51	0.0	0.0	0.0
Job_Satisfaction	821.19	1.14×10^{-13}	821.19	0.0	0.0	0.0
Tasks_Completed_Per_Day	754.91	1.39×10^{-13}	754.91	4.59×10^{-306}	0.0	4.59×10^{-306}
Annual_Salary	702.10	9.51×10^{-14}	702.10	4.81×10^{-286}	0.0	4.81×10^{-286}
Productivity_Score	660.13	9.51×10^{-14}	660.13	5.28×10^{-270}	0.0	5.28×10^{-270}
Overtime_Hours_Per_Week	617.49	1.14×10^{-13}	617.49	1.38×10^{-253}	0.0	1.38×10^{-253}
Years_at_Company	12.54	1.49×10^{-15}	12.54	3.64×10^{-6}	0.0	3.64×10^{-6}

Table 1: ANOVA tests for GMM clustering on PCA without categorical features.

Feature	**F-Stats Mean**	**F-Stats Std**	**F-Stats Median**	**P-Vals Mean**	**P-Vals Std**	**P-Vals M**
Absences_Per_Year	2001.73	2.88×10^{-13}	2001.73	0.0000	0.0000	0
Tasks_Completed_Per_Day	886.53	5.08×10^{-14}	886.53	0.0000	0.0000	0
Meetings_per_Week	656.00	8.04×10^{-14}	656.00	2.02×10^{-268}	0.0000	2.02 × 1
Overtime_Hours_Per_Week	612.05	1.14×10^{-13}	612.05	1.77×10^{-251}	0.0000	1.77 × 1
Productivity_Score	602.71	6.23×10^{-14}	602.71	7.29×10^{-248}	0.0000	7.29 × 1
col_1	523.53	9.51×10^{-14}	523.53	5.87×10^{-217}	0.0000	5.87 × 1
Annual_Salary	503.77	2.54×10^{-14}	503.77	3.56×10^{-209}	0.0000	3.56 × 1
Job_Satisfaction	456.51	3.60×10^{-14}	456.51	1.90×10^{-190}	0.0000	1.90 × 1
col_2	215.44	3.60×10^{-14}	215.44	2.50×10^{-92}	3.41×10^{-108}	2.50 ×
Monthly_Hours_Worked	167.78	2.38×10^{-14}	167.78	2.15×10^{-72}	2.51×10^{-88}	2.15 ×
col_0	97.05	0.0000	97.05	1.81×10^{-42}	0.0000	1.81 ×
Years_at_Company	23.26	2.75×10^{-15}	23.26	8.33×10^{-11}	1.29×10^{-26}	8.33 ×
col_3	9.10	9.73×10^{-16}	9.10	0.0001	1.36×10^{-20}	0

Table 2: ANOVA tests for GMM clustering on PCA with categorical features.

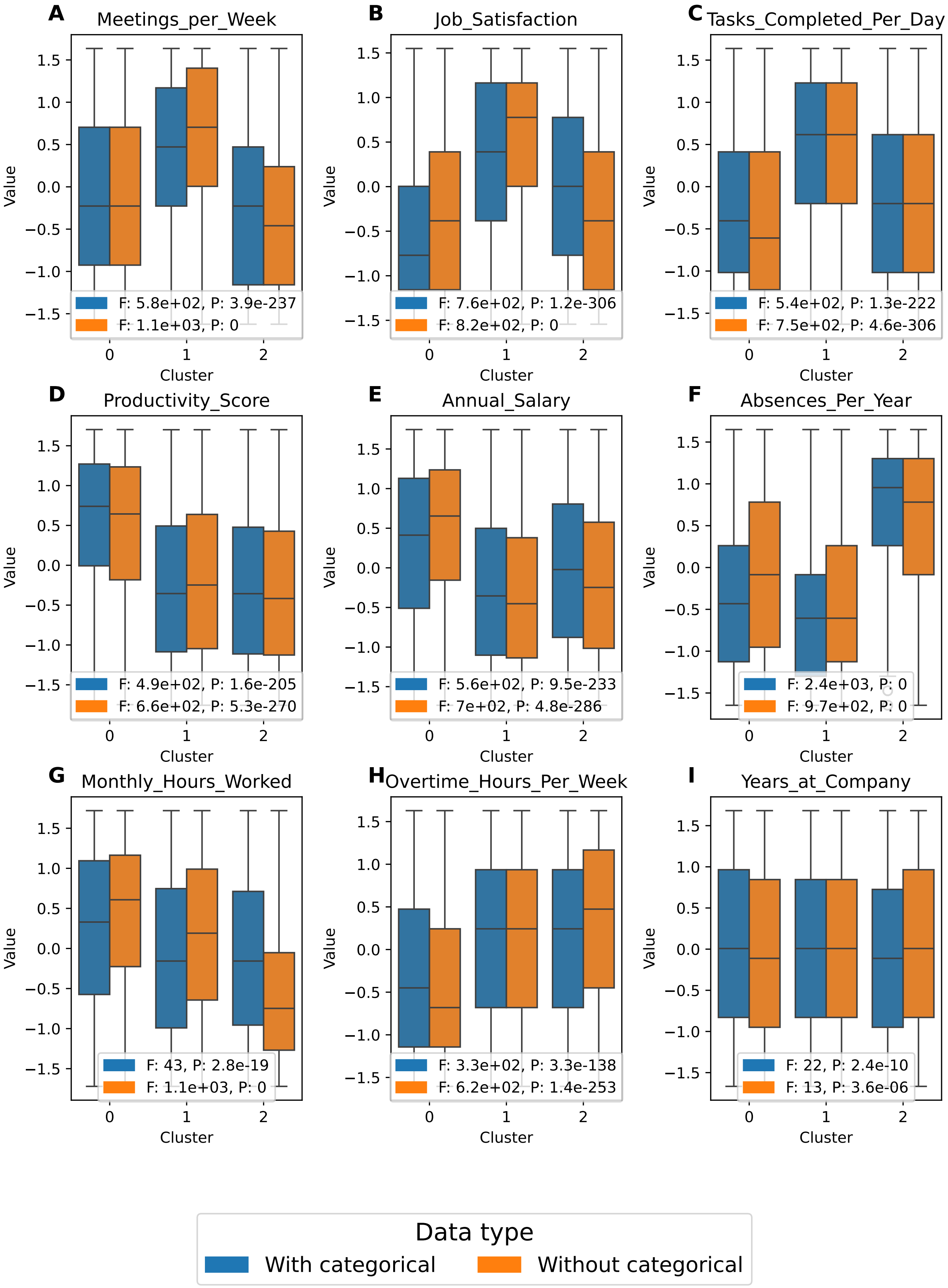

Figure 3: **GMM clustering reveals correlated and differentiated numeric features between clusters.** (A-I) Box plots of numeric features demonstrating differentiation between clusters. Cluster labels are outlined on the x-axis. The features were standardized and aligned to the cluster groups. F-statistics and p-values reported over 10 ANOVA tests (Methods3.5).

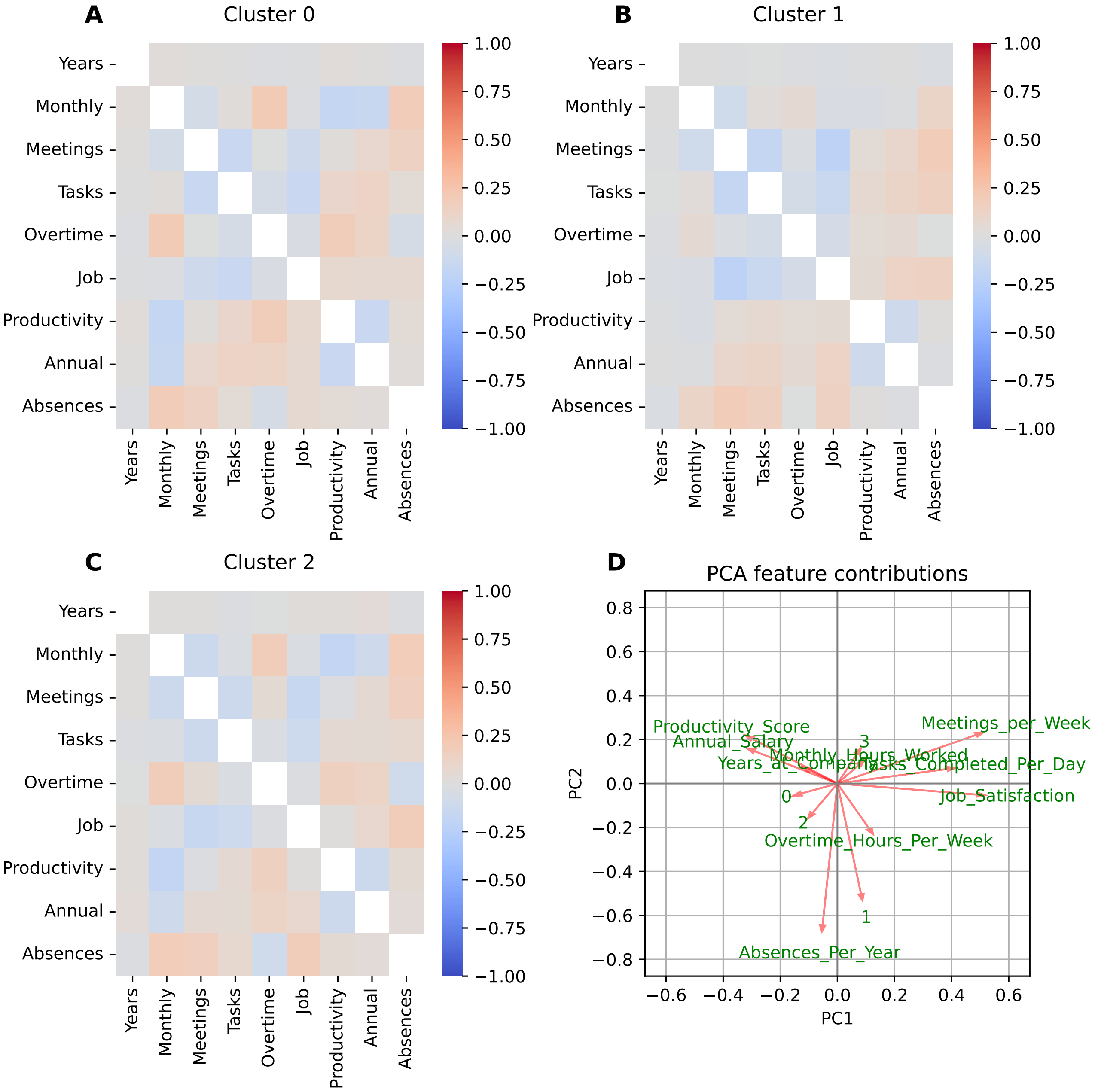

Figure 4: **GMM clustering reveals correlated numeric features among clusters.** (A–C) Heatmaps of Pearson correlation between features for each cluster obtained from the clustering algorithm. (D) Features contributing to the PCA components (Method 3.3). Each feature is projected to the first 2 PCs.

5 Discussion

In here we evaluated the following dataset of cooperative work in an unsupervised way. The data combines numeric and categorical features. We demonstrated that the data has a structure of 3 main clusters with 3 inner clusters of categorical features in each cluster. The main 3 clusters are due to remote-work status. We demonstrated that high number of meetings, tasks and satisfaction indicate one cluster, where high productivity and annual salary on another cluster. Moreover, monthly working hours are highly correlated with other features in some clusters. Furthermore, high number of absences indicates the last cluster. Sadly, these values are not well aligned with the remote-work status. Nevertheless, these features are of high importance due to the separation of the data through latent space. We suggest cooperatives to invest in these relations. We hypothesize that these features are connected indirectly to the remote-work status where remote work affects the mentioned features. For example, working from the office could enhance meetings and task loads, leading to better satisfaction among workers, whereas working remotely can lead to more absences and lower satisfaction. We suggest a deeper investigation of this latent structure and its relation to productivity using more advanced methods and cross-validating these insights through other datasets.

References

[1] Wikipedia contributors, "COVID-19 pandemic — Wikipedia, The Free Encyclopedia," 2025. [Accessed: 2025-04-29].

[2] TheMarker, "Working from home? you're probably more efficient – but your managers don't see it that way," 2020. Accessed: 2025-04-29.

[3] D. Keren, "What's up with the remote work trend? a situation review ahead of 2023," 2022. Accessed: 2025-04-29.

[4] S. Deepthi, "Corporate work hours & productivity," May 2022.

[5] K. Kamalja, S. D. S. S. R. Anjaneyulu, and V. S. R. Anjaneyulu, "Multiple correspondence analysis and its applications," *ResearchGate*, 2017. Accessed: 2025-04-29.

[6] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1, pp. 37–52, 1987. Proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists.

[7] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.

[8] L. P. Hansen, "Large sample properties of generalized method of moments estimators," *Econometrica*, vol. 50, no. 4, pp. 1029–1054, 1982.

[9] Wikipedia contributors, "Analysis of variance — Wikipedia, The Free Encyclopedia," 2025. [Accessed: 2025-04-29].

[10] Wikipedia contributors, "F-test — Wikipedia, The Free Encyclopedia," 2025. [Accessed: 2025-04-29].