What Employee Characteristics Are Associated with Differences in Productivity and Job Satisfaction Among U.S. Workers?

Persephone G. Moran

College of Information, University of Maryland

INST414 (0202) Data Science Techniques

Professor Kem Nguyen-Le

May 18, 202

Abstract

This project investigates how observable employee characteristics—such as department, position, gender, and tenure—are associated with differences in productivity and job satisfaction among U.S. workers. Using a realistic but synthetic HR dataset of 200 employees, multiple linear regression models were applied to assess the statistical relationship between employee features and two key outcomes: self-reported productivity and satisfaction. While a few predictors—such as being male or working in IT—were modestly associated with higher productivity, and interns and junior developers reported greater satisfaction, most variables were not statistically significant. Both models explained less than 10% of the outcome variance, underscoring the limited insight offered by traditional HR metrics alone. These results support broader theoretical insights from *How Women Rise* and *David and Goliath*, which suggest that performance and fulfillment are shaped as much by perception, power structures, and behavioral norms as by measurable traits. The findings call for deeper, more structural approaches to understanding workplace outcomes—ones that consider not just what can be measured, but what remains unseen.

Introduction

Understanding the factors that shape employee productivity and job satisfaction remains a critical issue for organizations striving to foster performance, engagement, and retention. While this project originally set out to explore how work modalities—remote, hybrid, and in-person—affect these outcomes, limitations in available data required a refinement of the research question. Instead, this study investigates how observable employee characteristics—such as department, position, gender, and tenure—are associated with differences in productivity and job satisfaction among U.S. workers.

This project adopts a structural lens alongside its statistical approach, recognizing that workplace dynamics often reflect deeper behavioral patterns and systemic biases. Helgesen and Goldsmith (2018) identify persistent barriers women face in professional advancement—not from lack of capability, but from socially conditioned habits like overvaluing expertise, reluctance to self-promote, or prioritizing others' comfort. These patterns, while rarely captured in HR datasets, significantly shape how performance is both enacted and evaluated.

In parallel, this project draws on Gladwellian research principles from *David and Goliath* (2013), which question conventional notions of advantage. Gladwell argues that perceived strengths—like confidence or seniority—can conceal structural vulnerabilities, while disadvantages may foster hidden resilience. In the context of employee outcomes, these frameworks remind us that statistical models may capture correlations, but not the power dynamics or psychological forces that underlie them.

To explore these dynamics, this project uses a fictional but realistic HR dataset and applies quantitative methods including exploratory data analysis and ordinary least squares

(OLS) regression. The analysis aims to detect which demographic or job-related variables predict productivity and satisfaction—and to consider what remains unseen when models rely solely on observable characteristics.

Methodology

This project uses a synthetic dataset titled *Employee Productivity and Satisfaction HR Data*, downloaded from Kaggle on April 27, 2025. The dataset contains 200 employee records with variables including age, gender, department, position, salary, number of projects completed, feedback score, and self-reported productivity and satisfaction rates (both as percentages). Though artificial, the data reflects plausible organizational conditions and is well-suited for statistical exploration.

All preprocessing and analysis were performed in Python using pandas, statsmodels, and Jupyter Notebook. Textual identifiers like employee names were dropped, and the *Joining Date* field was converted into *Tenure (Years)* to serve as a proxy for experience. Categorical variables—*Gender*, *Department*, and *Position*—were encoded as binary dummy variables, with the first category of each set dropped to prevent multicollinearity. Boolean values were cast to integers to ensure full compatibility with regression modeling.

The two dependent variables—*Productivity (%)* and *Satisfaction Rate (%)*—were analyzed separately using multiple linear regression via Ordinary Least Squares (OLS). Predictor variables included age, tenure, salary, feedback score, number of projects completed, and the encoded categorical variables. OLS was chosen for its interpretability and capacity to quantify linear relationships between employee characteristics and outcomes of interest.

This methodology allows for a straightforward assessment of whether and how observable employee features relate to reported productivity and satisfaction, while also highlighting the constraints of HR data when isolated from deeper behavioral or structural context.

Results and Analysis

To assess whether observable employee characteristics are associated with differences in productivity and job satisfaction, two multiple linear regression models were constructed using Ordinary Least Squares (OLS). The first model used self-reported productivity as the dependent variable, and the second used satisfaction rate. Both models included the same set of predictors: age, tenure, salary, feedback score, number of projects completed, and dummy variables for gender, department, and position.

Productivity Model

The productivity model explained roughly 8% of the variance in productivity scores ($R^2 = 0.080$), and the overall model was not statistically significant ($F = 1.065$, $p = 0.392$). However, two individual predictors reached or approached statistical significance. Being male was marginally associated with higher productivity ($\beta = 8.30$, $p = 0.051$), suggesting a potential gender effect in either performance or its reporting. Working in the IT department was significantly associated with higher productivity ($\beta = 14.15$, $p = 0.031$), likely reflecting job-role alignment with independent, digital workflows. Other variables—including tenure, feedback score, salary, and most position types—did not show significant relationships with productivity.

Satisfaction Model

The satisfaction model accounted for 7% of the variation in satisfaction rates ($R^2 = 0.070$) and also lacked global statistical significance ($F = 0.926$, $p = 0.537$). Two position variables, however, were individually significant. Interns reported the highest average satisfaction ($\beta = 34.07$, $p = 0.047$), followed closely by junior developers ($\beta = 22.88$, $p = 0.025$). These results may reflect onboarding support or early-career optimism. No other predictors, including gender, department, tenure, or salary, were statistically significant.

Interpretation

Taken together, the regression models suggest that while a few job-related and demographic variables are modestly associated with productivity or satisfaction, the majority of variation in these outcomes remains unexplained. The low R^2 values in both models indicate that unobserved factors—such as workplace culture, interpersonal dynamics, or psychological safety—likely play a larger role than observable features alone can capture.

These results support the project's broader argument: statistical models can reveal surface-level trends, but often fail to account for deeper behavioral patterns and systemic inequities that shape employee outcomes. Observed differences may reflect not just real performance gaps, but also differences in perception, opportunity, and reward structures embedded within workplace systems.

Discussion

This project's regression results show that while some employee characteristics—such as gender and department—are modestly associated with productivity, and others—like position—are associated with satisfaction, the majority of variance remains unexplained by observable features. These findings affirm the need to interpret data not just statistically, but structurally and behaviorally.

Gender, Department, and the Challenge of Observable Patterns

The productivity model revealed that being male ($\beta = 8.30$, $p = 0.051$) and working in IT ($\beta = 14.15$, $p = 0.031$) were among the only statistically or marginally significant predictors of productivity. These findings superficially resemble those in *Zhuang & Pan (2022)*, who found that technical departments often report higher satisfaction and performance outcomes—but their results emphasized that such patterns were mediated by support structures like peer networks and perceived fairness. This model, in contrast, lacks these latent variables, underscoring the limits of quantitative HR datasets in capturing the conditions that shape how performance is evaluated.

Moreover, the gender finding aligns with insights from *How Women Rise* (Helgesen & Goldsmith), where women's reluctance to self-promote—or social penalties for doing so—can distort how productivity is reported or perceived. Our results may reflect not actual differences in output, but in how performance is socially validated—a caution echoed in *Elanwer (2021)*, who emphasized the risk of over-interpreting surface-level patterns in HR analytics without deeper behavioral modeling.

Job Satisfaction: Optimism, Role Identity, or Structural Naïveté?

Interns and junior developers were the only groups with statistically significant associations with satisfaction. Zhuang & Pan (2022) suggest that early-career employees tend to report higher satisfaction when they feel supported or have clear growth paths. This may explain our findings—these employees could be experiencing a "honeymoon period" supported by structured onboarding and low expectation-pressure, which fades over time.

However, Nagpal et al. (2024) note that early satisfaction is not a reliable predictor of retention without sustained mentorship and development—a risk not observable in our dataset but highly relevant for HR policy.

The near-significance of tenure (p = 0.199) may also mirror *Zhuang & Pan's* findings that satisfaction increases slightly with experience, but only when roles include perceived growth and autonomy. Since our dataset does not include performance reviews or psychological safety metrics, we can't observe whether experienced employees feel more secure or just more resigned.

Methodological Parallels and Divergences

All three scholarly sources used more advanced modeling than basic OLS—*Elanwer* applied decision trees and neural networks, *Nagpal et al.* implemented gradient boosting, and *Zhuang & Pan* explored quantile regression. These methods were more capable of detecting nonlinear patterns and interaction effects. In comparison, our linear model explained only 7–8% of the variance in outcomes (R^2 = 0.070–0.080). This confirms a key critique raised by *Elanwer*

(2021): basic statistical models are often insufficient for capturing the behavioral nuance embedded in HR data.

Moreover, both *Elanwer* and *Nagpal et al.* stress the importance of model explainability—either through tools like LIME or stakeholder-focused visualization—to ensure HR professionals understand how predictions are generated. This project echoes that goal by integrating interpretability as a future direction, especially given the risk of overtrusting black-box models without context.

A Structural Interpretation: Beyond What the Model Can See

Ultimately, this project supports the view that HR data alone rarely tells the full story. As Helgesen & Goldsmith argue, “just working hard won’t get you where you want to go”—particularly for women and marginalized workers. And as Gladwell reminds us, “giants are not what we think they are”: statistical significance may cloak deeper structural inequity or social bias.

The limited explanatory power of our regressions doesn’t mean that employee characteristics are irrelevant. Instead, it signals that core determinants of productivity and satisfaction—like mentorship, cultural fit, or psychological safety—are absent from standard HR datasets. As *Elanwer* notes, quantitative models must be situated within real-world decision-making processes, not mistaken for them.

Limitations

This project is subject to several limitations, both methodological and structural, that impact the interpretability and generalizability of the findings.

Data Quality and Generalizability

The dataset used for analysis is fictional, designed for instructional purposes. While its structure reflects plausible HR data, its synthetic nature means that results cannot be generalized to real-world organizations. Furthermore, the data lacks several dimensions that are critical for understanding workplace dynamics—such as employee-manager relationships, psychological safety, peer support, and performance evaluations. These missing variables likely account for much of the unexplained variance in both regression models ($R^2 = 0.070–0.080$).

Missing Work Modality Data

The project's original research question focused on remote versus in-person work. However, because the dataset did not include a variable identifying work modality, the scope was necessarily refined to focus on demographic and job-role characteristics instead. This limits the study's ability to contribute directly to contemporary conversations about post-pandemic work models.

Modeling Constraints

This project used multiple linear regression (OLS) for its transparency and interpretability. However, as demonstrated by Elanwer (2021) and Nagpal et al. (2024), more advanced models such as decision trees or gradient boosting are often better suited for uncovering nonlinear or interaction effects in HR data. Future iterations may benefit from these techniques, particularly when paired with explainability tools.

Multicollinearity and Feature Encoding

Some predictor variables, such as salary and number of projects completed, exhibited high correlation in earlier exploratory analysis. While steps were taken to mitigate multicollinearity (e.g., dummy-variable encoding with reference categories), it may still subtly impact the model's estimates. In addition, all categorical variables were encoded as dummies, which treats categories as distinct without accounting for ordinal or relational meaning.

Omitted Variable Bias and Confounding Factors

Key drivers of productivity and satisfaction—such as motivation, workplace discrimination, leadership quality, and organizational culture—are not represented in the dataset. Their omission introduces potential confounding bias. For example, a perceived gender-productivity gap might be less about output and more about how work is rewarded or recognized—especially in contexts where assertiveness is interpreted differently across gender lines (Helgesen & Goldsmith, 2018).

Temporal Limitations

The dataset includes a snapshot of employee attributes and self-reported outcomes at a single time point. Without longitudinal data, it is impossible to assess how satisfaction or productivity change over time or in response to internal changes like promotion, turnover, or burnout.

Conclusion

This project examined whether observable employee characteristics—like gender, department, tenure, and position—can explain differences in productivity and job satisfaction. Using multiple linear regression on a synthetic HR dataset, it found only limited evidence of statistically significant relationships. Being male and working in IT were modestly associated with higher productivity, while interns and junior developers reported higher satisfaction. However, both models explained less than 10% of the variance, suggesting that most drivers of workplace outcomes lie outside the scope of conventional HR data.

These results support the broader argument that performance and satisfaction are shaped not just by individual traits, but by structural dynamics and social perception. As Helgesen and Goldsmith (2018) argue in *How Women Rise*, women's workplace outcomes are often influenced by unspoken behavioral expectations—like self-silencing or over-preparing—that go unmeasured but heavily affect evaluation. Similarly, Gladwell's *David and Goliath* reminds us that perceived strengths—like assertiveness or visibility—can mask deeper inequities, while undervalued traits may reflect hidden resilience.

Future work should expand the modeling approach and the dataset itself. Nonlinear methods like decision trees or quantile regression, as demonstrated in Elanwer (2021) and Zhuang & Pan (2022), may better capture complex dynamics. Just as importantly, richer data—including metrics on psychological safety, mentorship, and perceived fairness—could offer a clearer picture of what truly drives workplace outcomes. Until then, this project urges caution: what organizations choose to measure often defines what they value—but not always what matters most.

References

Elanwer, R. (2021). *HR Analytics in practice: Using machine learning to predict employee absenteeism* (Master's thesis,The Pennsylvania State University). https://etda.libraries.psu.edu/files/final_submissions/23839

Gladwell, M. (2015). *David and Goliath: Underdogs, misfits, and the Art of Battling Giants*. Back Bay Books / Little, Brown and Company.

Helgesen, S., & Goldsmith, M. (2019). *How women rise: Break the 12 habits holding you back*. Random House Business Books.

Nagpal, P., Pawar, A., & Sanjay, M. (2024). Predicting employee attrition through HR analytics: A machine learning approach," In *Proceedings of the 2024 4th International Conference on Innovative Practices in Technology and Management (ICIPTM)* (pp. 1-4). IEEE. https://doi.org/10.1109/ICIPTM59628.2024.10563285

Zhuang, M.-E., & Pan, W.-T. (2022). Data modelling in human resource management: Influencing factors of employees job satisfaction. *Discrete Dynamics in Nature and Society*, 2022, 1-19. https://doi.org/10.1155/2022/3588822