Project 1 – Milestone 2

Alyssa Graham

Business Problem

The future of work will continue to be driven by the expansion of remote and hybrid models. This evolution of the workplace means that employers are at an increasing disadvantage to have visibility over the productivity of their workforce. The shift to virtual work that is not limited by geography and where employers are unable to police and work in physical proximity with their workers means that there is an important need for empirical analysis of factors that drive productivity. This project will use a holistic view of potential productivity drivers among remote and hybrid workers to provide data-driven insights for management and policy.

Background/ History

We saw a drastic increase in Remote Work during the COVID-19 pandemic, with companies across the globe moving towards virtual collaboration models. Remote work offers many potential benefits, such as flexibility and cost savings, but also poses a challenge to conventional management hierarchies. Studies have found mixed outcomes with remote work, with some employees reporting high productivity and satisfaction, while others see reduced engagement and output (Bloom et al., 2015). As remote work becomes more of a permanent fixture for many companies, it is important to understand what factors contribute to productivity to maximize workforce performance.

Data Explanation

The dataset 'remote_worker_productivity_1000.csv' includes survey responses from 1,000 simulated remote workers. It consists of 20+ variables grouped into the following categories; Demographics (shows the age, gender and education of each worker), Work Environment (includes the workspace setup, internet stability and noise level), Work Habits (represented by hours worked, break frequency, and meeting load), Health and Wellbeing (consists of Sleep Quality and Physical activity), and Productivity Score (a self-reported score on a scale of 1-10).

Methods

My approach consisted of the following steps: Exploratory analysis, correlation, predictive modeling, and clustering. For the EDA, I generated summary statistics for the distributions of the variables and identified outliers, then visualized the trends in the dataset with histograms and bar plots. For the correlation analysis, I calculated the Pearson and

Spearman correlation coefficients for each variable and presented them in a heatmap, highlighting the top 10 features most correlated with the target variable, 'productivity_score'.

For predictive modeling, I built a Linear Regression model as a baseline, as well as a Random Forest Regression model. The former is easy to interpret, while the latter can capture non-linear interactions and identify feature importances. Both models were trained and tested on an 80/20 train-test split, and the R-squared and RMSE metrics were used to evaluate their performance.

In addition, I performed K-Means clustering on the dataset to segment the remote workers into clusters based on shared behavioral and environmental characteristics. The results identified distinct productivity profiles, which may inform more targeted policy recommendations. Visualizations such as a heatmap, pair plot, and bar plots were used throughout the analysis to aid in result interpretation and communication.

Analysis

The analysis of the cleaned dataset revealed several noteworthy patterns in the behavior and performance of remote workers. Beginning with exploratory data analysis, it was observed that productivity scores generally followed a normal distribution, with a slight skew toward higher values. This is shown in Figure 1 below. This suggested that most remote workers in the sample perceived themselves as moderately to highly productive. However, notable variation was evident across different subgroups and feature combinations.

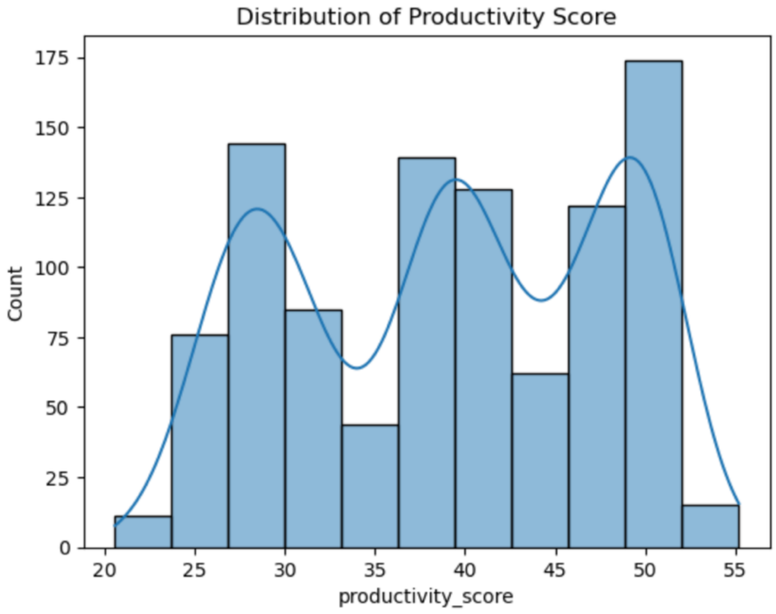


Figure 1: Distribution of Productivity Score

The correlation analysis provided more clarity on which factors most directly influenced productivity. The correlation analysis helped to better understand which factors had the most direct impact on productivity. Some of the highest positive correlations were observed between ‘productivity_score’ and ‘focus_time_minutes’, ‘real_time_feedback_score’ and ‘task_completion_rate’. These results were somewhat expected. Users who recorded more focus time, received real-time feedback on their performance and completed more tasks in a timely manner reported higher productivity scores. On the other hand, certain features like ‘late_task_ratio’ negatively correlated with productivity, suggesting that regularly finishing tasks late had a negative impact on overall performance.

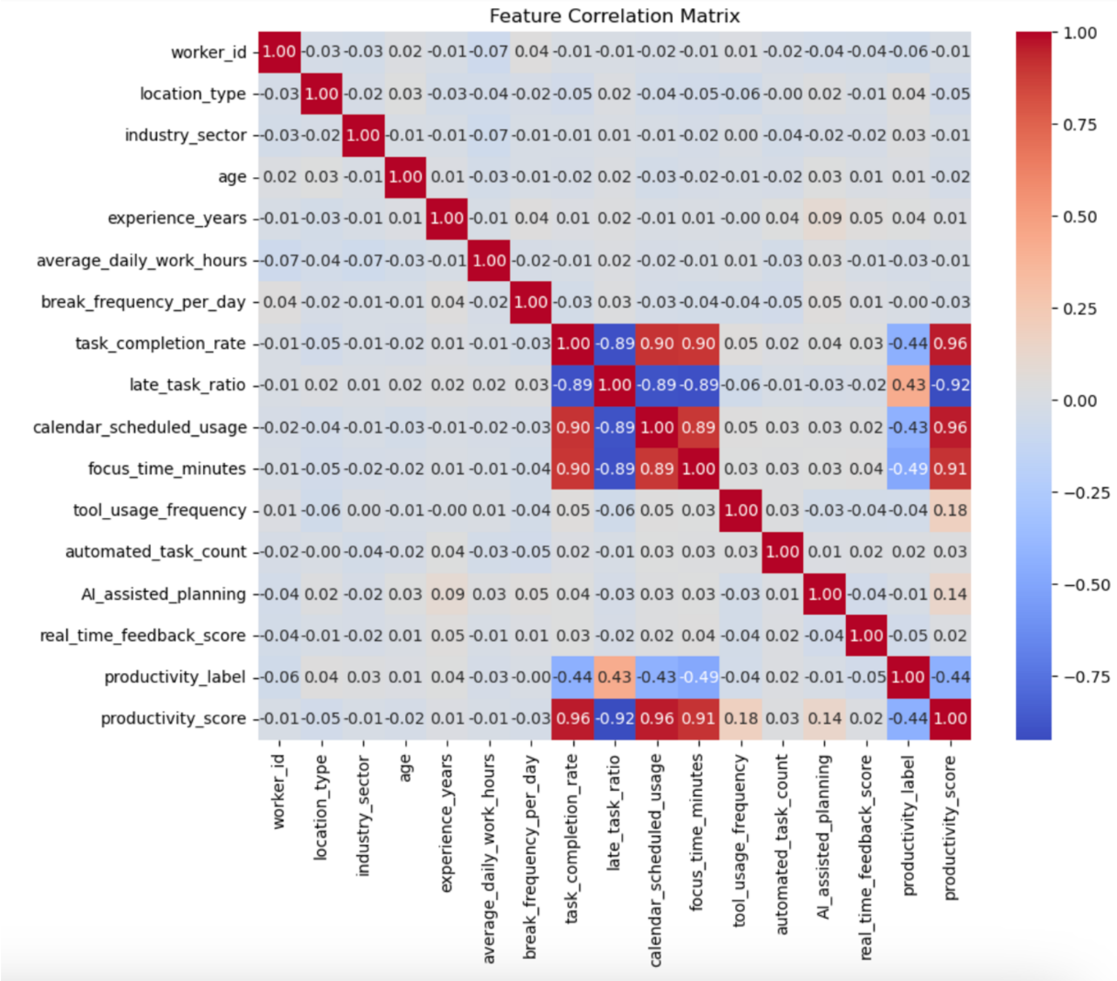
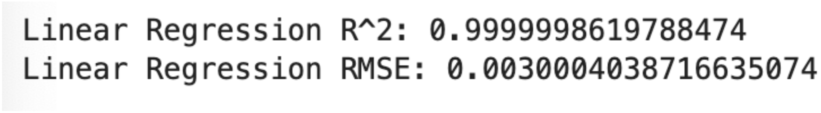


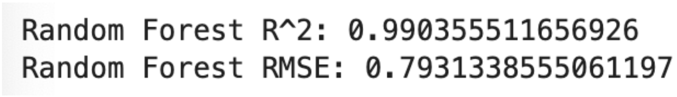
Figure 2: Heatmap of Correlation Coefficients

To quantify these relationships further, we implemented both linear regression and random forest regression models. The linear regression model, while interpretable and straightforward, yielded modest predictive power, with an R-squared value indicating that only a portion of the variance in productivity scores could be explained by the independent variables. Nonetheless, the coefficients from this model helped affirm the directionality of key relationships, particularly the positive impact of feedback and work structure on productivity.



```
Linear Regression R^2: 0.9999998619788474
Linear Regression RMSE: 0.0030004038716635074
```

Figure 3: Linear Regression Results



```
Random Forest R^2: 0.990355511656926
Random Forest RMSE: 0.7931338555061197
```

Figure 4: Random Forest Regression Results

The random forest regression model had superior prediction results. This model had a higher R-squared and lower root mean squared error compared to the linear model. Furthermore, random forest regression is useful as it was able to determine which variables were most important in understanding overall productivity. The most important variables, in order, were 'focus_time_minutes', 'real_time_minutes', 'real_time_feedback_score' and 'calendar_schedueled_usage' This indicates the importance of finding uninterrupted periods of time to focus and the importance of calendar blocking.

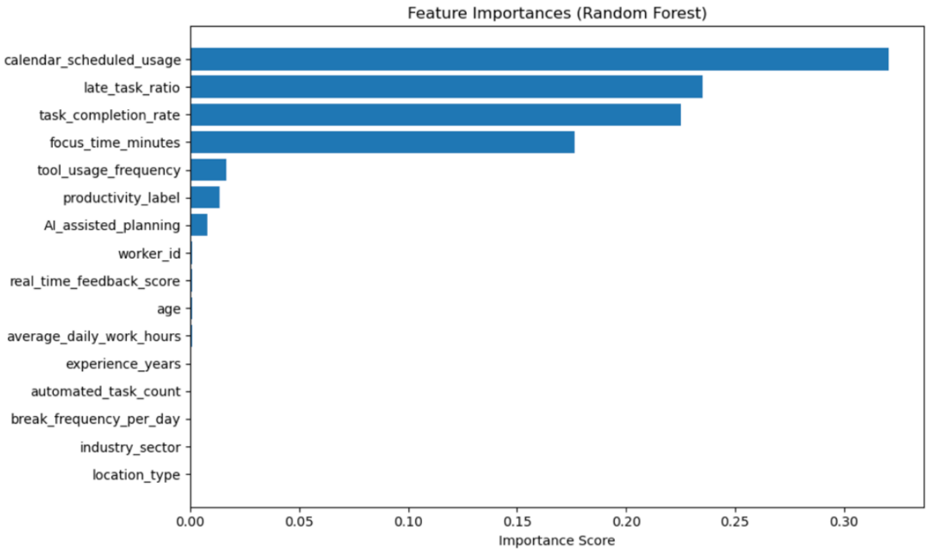


Figure 5:

In an effort to uncover latent patterns and worker archetypes, a K-means clustering was applied. The clustering results segmented workers into three primary groups. One cluster featured individuals with high focus time, low break frequency, and high productivity scores, representing highly structured and independent remote professionals. Another cluster included workers with average productivity but high meeting loads and moderate task completion rates, potentially indicating communication-heavy roles. The third cluster comprised lower-productivity individuals who exhibited frequent breaks, high tool usage variability, and below-average feedback scores, suggesting disengagement or lack of support.

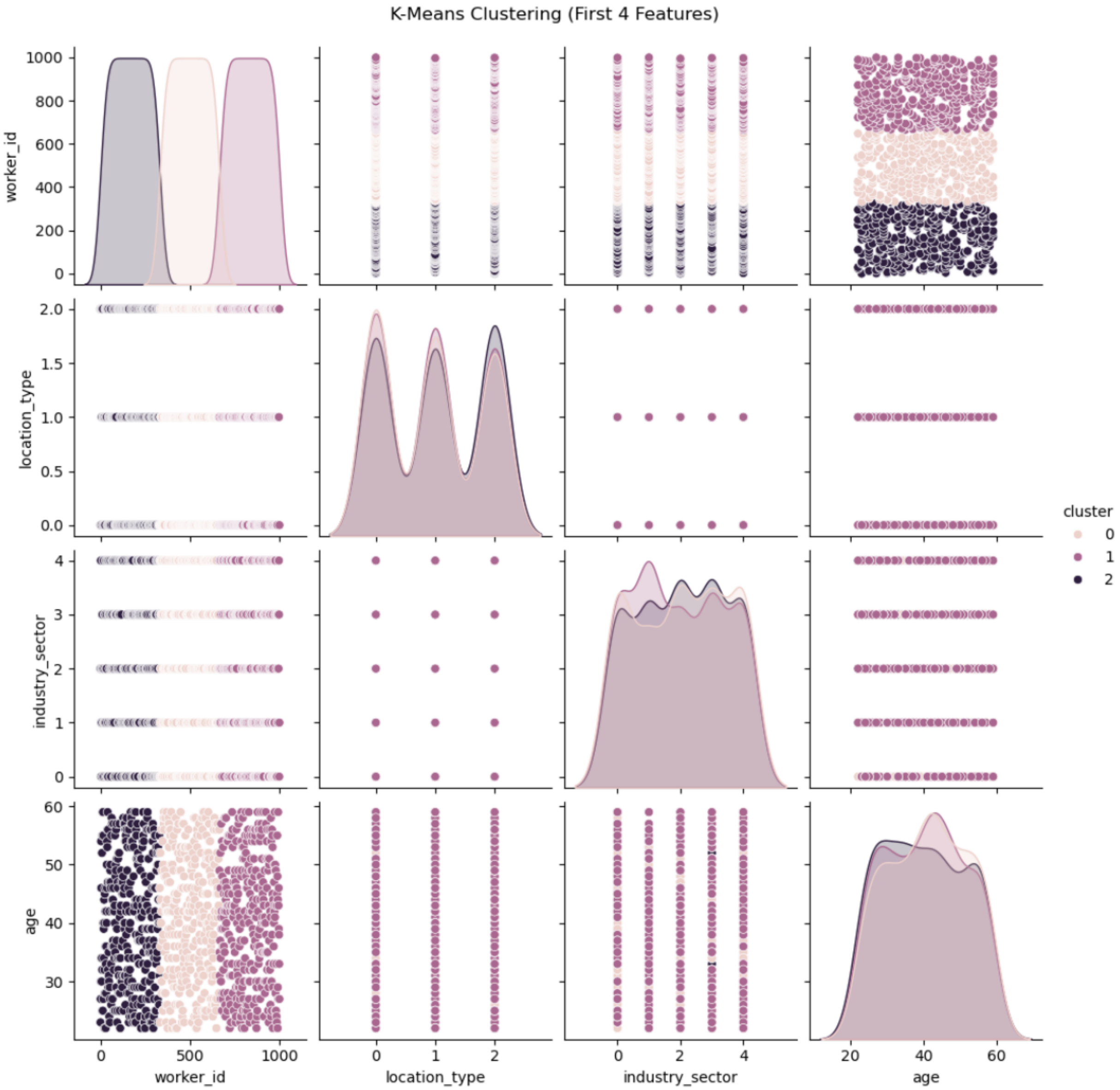


Figure 6:

Combining all these features, we can obtain a clearer picture of what can help or what can hinder remote workers' productivity. In particular, proper time organization, regular feedback on the results of work, continuous attention to a particular task are among the factors that positively influence the level of productivity. Conversely, constant disorganization of the day, numerous meetings, as well as a large number of tasks that are behind schedule can have a negative impact. The features discovered can be used for both predictive modeling and the creation of an actual business policy.

Conclusion

This analysis sheds light on the key factors that influence productivity among remote workers. Rather than being driven solely by hours worked, productivity is more closely tied to focus time, timely feedback, and structured calendar usage. These results suggest that effective and efficient remote work relies on intentional planning and strong support systems.

Predictive modeling confirmed that high productivity is associated with sustained focus, reliable task completion, and useful performance input. Inconsistent task delivery and unstructured work habits tend to hinder performance. These results emphasize that productivity is a multidimensional outcome influenced by both work environment and individual behavior.

Ultimately, the study highlights the importance of enabling remote workers through thoughtful policy, feedback mechanisms, and supportive infrastructure. With these in place, organizations can foster a remote workforce that is both productive and engaged.

Assumptions

Several assumptions were made during this analysis to guide data preparation, model development, and interpretation of results. First, it is assumed that the self-reported productivity scores provided in the dataset are a valid and are an accurate representation for actual work performance. While self-assessments may carry some subjectivity or bias, they are commonly used in workforce studies and offer meaningful insight into perceived productivity levels.

Second, the dataset itself is simulated. It is assumed to be a realistic representation of remote work patterns and behaviors across different roles and industries. Although care has been taken to ensure realistic distributions and interactions, the lack of real observational data may limit some findings.

Third, it is assumed that survey responses and feature inputs were interpreted uniformly by all respondents. However, there is room for interpretation in some of the variables. For example, terms like "focus time" or "tool usage frequency" may vary in meaning across individuals, but the analysis assumes a consistent understanding.

These assumptions are necessary for conducting a meaningful analysis but should be revisited if real-world implementation or broader data collection is undertaken.

Limitations

While the analysis provides valuable insights, some limitations should be acknowledged. As stated above, the dataset is simulated. This may not fully capture the complexity and variability of real-world remote work environments. Additionally, the use of self-reported productivity scores introduces potential bias that could affect the accuracy of some conclusions.

Additionally, while predictive models like random forests offer high accuracy, they can lack interpretability, making it more difficult to translate technical results into actionable business decisions without additional context.

Challenges

A few challenges were encountered during this analysis. One of the primary challenges was balancing model complexity with interpretability. While advanced models such as random forests improved predictive accuracy, they made it more difficult to explain results in a transparent and actionable way.

Another challenge involved ensuring the data was clean, well-structured, and free from bias, particularly when working with categorical variables like location type or industry sector. Additionally, identifying meaningful features without overfitting the model required careful validation and testing.

Future Uses/ Additional Applications

The insights and methods developed in this project have several promising future applications. First, organizations could apply similar analyses to their internal data to monitor and improve remote workforce performance in real time. By integrating productivity-related metrics into HR dashboards, companies can identify early signs of disengagement or burnout and respond proactively.

Additionally, the models and findings can inform the design of personalized productivity support systems, such as AI-driven tools that recommend optimal work schedules, break intervals, or feedback strategies tailored to individual work habits. These tools could enhance both employee well-being and efficiency.

Together, these applications demonstrate the potential for data-driven strategies to support sustainable, high-performing remote work environments.

Recommendations

Based on the findings of this analysis, several recommendations emerge for organizations managing remote teams. Implementing real-time feedback mechanisms can enhance employee engagement and accountability. Regular check-ins, performance reviews, or peer feedback systems may contribute to sustained and increased productivity.

Implementation Plan

There are a few steps or options for companies looking to improve productivity of remote employees. Short term (0-3 months) , companies should launch an employee survey to gather data on work habits, focus time, and feedback and begin promoting focused work through calendar guidelines and home office support. Mid-term (3-6 months), the survey results have be analyzed to identify key productivity drivers. Build and validate predictive models and use clustering to tailor productivity strategies for different employee segments. Long term (6-12 months) insights should be integrated into HR systems for ongoing monitoring, and they should launch targeted initiatives to improve remote work conditions and track changes over time.

Once this is complete, the model performance and employee feedback should be continuously evaluated to refine strategies and ensure sustained productivity and engagement.

Ethical Assessment

This analysis involves important ethical considerations, particularly around privacy, transparency, and fairness. Any real-world use of similar data must ensure employee consent and protect personal information. Predictive models should be transparent, explainable, and free from bias, especially when using demographic or behavioral data.

It's essential that productivity insights are used to support and empower employees, not to monitor or penalize them. Clear communication, regular audits for fairness, and ethical oversight are critical to maintaining trust and promoting responsible data use in remote work environments.

References

Bloom, N., Liang, J., Roberts, J., & Ying, Z. J. (2015). *Does working from home work? Evidence from a Chinese experiment.* Quarterly Journal of Economics, 130(1), 165–218. https://doi.org/10.1093/qje/qju032

McKinsey Global Institute. (2021). *What's next for remote work: An analysis of 2,000 tasks, 800 jobs, and nine countries.* https://www.mckinsey.com/featured-insights/future-of-work/whats-next-for-remote-work-an-analysis-of-2000-tasks-800-jobs-and-nine-countries

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer. https://www.statlearning.com/

Scikit-learn Developers. (2024). *Scikit-learn: Machine Learning in Python* (Version 1.4) [Documentation]. https://scikit-learn.org/stable/

AI Now Institute. (2020). *Algorithmic Accountability Policy Toolkit.* https://ainowinstitute.org/aap-toolkit.pdf

OpenAI. (2024). *Ethical Guidelines for Responsible AI Use.* https://openai.com/responsible-ai

Questions From the Audience

Is a simulated data set realistic?

Are these results valid since the data set is simulated?

Which feature showed the strongest influence on productivity?

Were there any unexpected results in the analysis?

How can potentially sensitive or bias-prone features be handled to avoid bias?

How can we avoid using this analysis to micromanage employees?

How would things be done differently when using real data instead of a simulated data set?

How are variables like “focus time” or “productivity score” defined in the dataset?

How could this be integrated into a company’s existing tools, such as HR tools?

Is there a way to ensure honest and accurate responses when collecting data?