# DSCI 550 Project Report:
# Exploring the Impact of Remote Work on Employee Performance and Satisfaction

**Team Members:** Cheryl Khau, Mandy Chen, Silvia Choi

## 1. Problem Definition and Background

As companies transition to post-pandemic work models [3], understanding the impact of remote work on employee performance, satisfaction, and retention has become increasingly critical. Flexible work arrangements, which gained prominence during the pandemic, have been shown to influence employee productivity, well-being, and overall job satisfaction [4]. To address this need, our project focuses on three key research questions:

1. **Do high-performing employees report higher satisfaction levels with increased remote work [1]?**
2. **What percentage of low performers primarily work remotely, and does remote work impact their productivity?**
3. **How does remote work frequency vary across performance categories (high, mid, low) [2], and how does this influence retention?**

These questions aim to uncover actionable insights for optimizing remote work policies and improving workforce outcomes.

## 2. Description of Dataset

The dataset used for this project was sourced from Kaggle and contains comprehensive corporate employee performance data. It consists of 100,000 rows and 18 columns, providing a robust dataset for statistical analysis and predictive modeling. The dataset includes a diverse range of variables spanning demographics such as age, gender, and education, job-related details like department, salary, and job title, and work habits including remote work frequency and overtime hours. Additionally, performance indicators such as performance scores, employee satisfaction scores, and resignation data are included. These features make the dataset highly relevant to our research questions, as it enables an examination of the relationships between remote work, employee performance, and retention. Its size and diversity ensure that it is suitable for uncovering actionable insights through both statistical analysis and machine learning techniques.

## 3. Methods

### 3.1 Data Cleaning and Preprocessing

For our dataset preparation, several data cleaning and preprocessing steps were undertaken. Missing data was addressed by removing rows with null values to maintain data integrity. Categorical variables were transformed into numerical representations through one-hot encoding to facilitate analysis. Additionally, feature engineering was performed to enhance the dataset's structure: employees were grouped into Low, Mid, and High performers based on predefined performance score thresholds, and remote work frequencies were categorized into None, Low, and High groups. These steps ensured the dataset was clean, consistent, and ready for statistical and predictive modeling.

### 3.2 Statistical Analysis

Statistical analyses were conducted to explore the relationships outlined in our research questions. For the first question, an ANOVA test assessed differences in satisfaction levels among high performers across various *remote work categories*, supplemented by regression models to examine the relationship between *performance levels* and *satisfaction scores*. In addressing the second question, the Kruskal-Wallis test evaluated productivity metrics—specifically, projects handled and overtime hours—for low performers across *remote work categories*. Additionally, one-way ANOVA and pairwise T-tests were performed to identify any significant differences between groups. These statistical methods provided foundational insights into the data, revealing patterns and trends essential for further analysis.

### 3.3 Predictive Modeling

To analyze how *remote work frequency* and *performance categories* influence employee retention, predictive modeling techniques were applied. Logistic regression was initially used to predict resignation probabilities; however, due to class imbalance in the target variable, the Synthetic Minority Oversampling Technique (SMOTE) was implemented to balance the dataset. This adjustment improved the model's ability to detect patterns associated with employee resignation. Advanced machine learning models, including Random Forest and Gradient Boosting Classifiers, were then employed to enhance prediction accuracy. These models not only improved classification performance but also facilitated feature importance analysis, highlighting the significant impact of employee satisfaction on retention, while indicating that remote work frequency and performance levels had lesser effects.

## 4. Experiment Setup and Results

### 4.1 Research Question 1: Satisfaction and Remote Work for High Performers

The ANOVA analysis for high performers revealed no significant differences in satisfaction scores across remote work categories (test statistic: 0.86, p-value: 0.42). The accompanying boxplot (Figure 1) supports this finding, showing no visible variation in remote work frequency distribution across performance groups, with consistent medians and variability across categories. Further insights from the heatmap (Figure 2) highlight mean satisfaction scores for each combination of performance group and remote work category. Satisfaction scores are remarkably close across all categories, with slight variations observed. Low remote work (3.01) resulted in slightly higher satisfaction scores for both low and high performers, while high remote work (3.00) and no remote work (2.99) were marginally lower across all groups. These minor differences reinforce that remote work frequency does not significantly influence satisfaction for high-performing employees, suggesting other factors likely play a more substantial role.

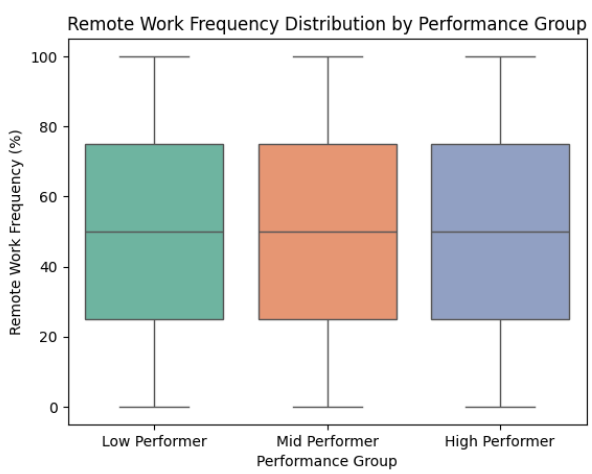


**Figure 1**

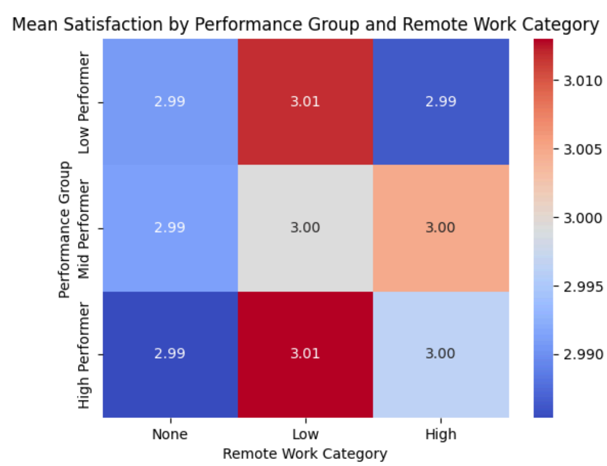


**Figure 2**

We tested two additional models for this analysis. The linear regression model (Figure 3) showed a coefficient of 0.0019, indicating that moving from mid to high performer groups increases satisfaction scores by only 0.0019. The intercept of 2.9976 represents the baseline satisfaction score when the low performer group is mapped to 0. The R-squared value of 1.54e^(-6) indicates the model explains virtually no variance in satisfaction scores. Linear regression was selected as a standard method to quantify relationships between the continuous dependent variable (satisfaction scores) and the categorical independent variable (performance groups). While the results showed no significant variance explained by performance groups, this approach aligns with methods commonly applied in workplace studies, such as [1]'s regression analysis of telecommuting and job satisfaction.

```
---- Linear Regression ----
Linear Regression Coefficient: 0.00191182274502061, Intercept: 2.9975628610406106
Mean Squared Error: 1.3241398789312875
R-squared: 1.5451924859632982e-06

---- Logistic Regression ----
Logistic Regression Accuracy: 0.50208
Confusion Matrix:
[[50208     0]
 [49792     0]]
Classification Report:
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:1531: Un
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:1531: Un
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:1531: Un
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
              precision    recall  f1-score   support

           0       0.50      1.00      0.67     50208
           1       0.00      0.00      0.00     49792

    accuracy                           0.50    100000
   macro avg       0.25      0.50      0.33    100000
weighted avg       0.25      0.50      0.34    100000


---- Random Forest ----
Random Forest Feature Importances: [0.47846129 0.52153871]
R-squared on test set: -0.00021003201256464266
```

**Figure 3**

The logistic regression model achieved an accuracy of 50.21%, struggling to predict satisfaction accurately due to class imbalance. The confusion matrix showed all employees predicted as "not satisfied" (satisfied = 0), highlighting the data's class imbalance issue and the model's limited ability to distinguish between satisfaction levels. Despite these limitations, logistic regression was chosen for its established use in analyzing binary relationships in organizational research.

### 4.2 Research Question 2: Remote Work and Productivity for Low Performers

The Kruskal-Wallis test was used to evaluate differences in two productivity metrics: projects handled and overtime hours across remote work categories (none, low, and high). The results showed no significant differences, with a test statistic of 0.80 and p-value of 0.67 for projects handled, and a test statistic of 3.16 and p-value of 0.21 for overtime hours. The boxplots (Figure 4) further confirmed these findings, showing consistent medians and similar ranges across all remote work categories for both metrics. This suggests that remote work does not significantly influence productivity for low performers in terms of either projects handled or overtime hours.

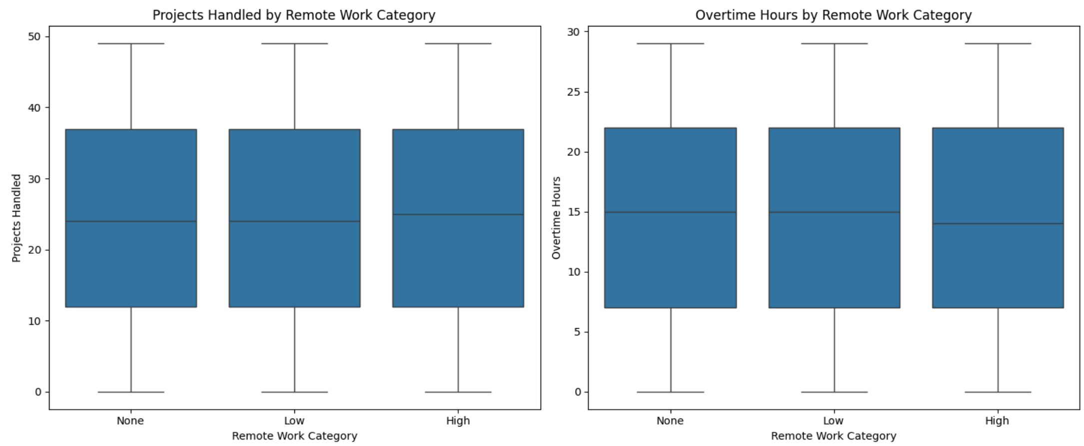


**Figure 4**

Additionally, A one-way ANOVA test corroborated these findings, showing no statistically significant difference in productivity metrics across remote work categories for low performers (Figure 4). The test statistics and p-values closely matched the Kruskal-Wallis results (projects handled: 0.79, p=0.67; overtime hours: 3.15, p=0.21), suggesting that remote work does not appear to have a measurable impact on low performers' productivity.

```
Projects_Handled - Kruskal-Wallis test statistic: 0.7958009219395946, p-value: 0.6717288875763976
Overtime_Hours - Kruskal-Wallis test statistic: 3.15885351084151, p-value: 0.20609320616252785
```

**Figure 5**

Finally, pairwise T-tests with Bonferroni correction further validated these findings. All comparisons for projects handled and overtime hours showed p-values greater than 0.05, with most corrected p-values at or above 1.0. Even the lowest uncorrected p-value (0.10 for none vs. high in overtime hours) remained above the significance threshold, consistently indicating that remote work frequency does not significantly influence productivity metrics for low performers. This method has been previously applied in organizational studies [4].

### 4.3 Research Question 3: Retention Analysis

When we tried to find out how remote work frequency and performance categories influence retention, logistic regression was initially applied. Due to severe class imbalance, the model achieved 90% accuracy but failed to predict resignations effectively, with an F1-score of 0.0.

After balancing the dataset using SMOTE, the balanced logistic regression improved prediction performance slightly, achieving 51% accuracy with moderate precision and recall.

We then applied a Random Forest Classifier due to its ability to handle non-linear relationships and capture feature interactions, a model widely recognized in employee retention studies [5]. The Random Forest classifier (Figure 6) produced an overall accuracy of 78%, with F1-scores of 0.80 and 0.77 for non-resigned and resigned employees respectively. The model achieved balanced precision (0.75 for False and 0.83 for True) and recall (0.85 for False and 0.72 for True). Feature importance analysis revealed that employee satisfaction score was the most significant predictor of resignation (99.3%), while remote work frequency (0.51%) and performance categories had minimal influence.

```
Classification Report:
              precision    recall  f1-score   support

       False       0.75      0.85      0.80     26904
        True       0.83      0.72      0.77     27090

    accuracy                           0.78     53994
   macro avg       0.79      0.78      0.78     53994
weighted avg       0.79      0.78      0.78     53994


Confusion Matrix:
[[22764  4140]
 [ 7507 19583]]

Feature Importances:
                            Feature  Importance
1       Employee_Satisfaction_Score    0.992988
0             Remote_Work_Frequency    0.005091
2   Performance_Group_Mid Performer    0.001085
3  Performance_Group_High Performer    0.000837
```

**Figure 6**

Moreover, we decided to discover deeper insights into the influence of retention. Therefore, we had applied the Gradient Boosting Classifier for our analysis. Referring to Figure 7 below, the model achieved an accuracy of 55%, with balanced precision and recall for both classes, yielding F1-scores of 0.53 and 0.58, respectively. While employee satisfaction score was the most influential feature (99.3%), remote work frequency (0.51%) and performance categories (below 0.2%) had minimal influence. These results align with previous analyses, confirming that resignation is strongly driven by satisfaction rather than other variables. We think the Gradient Boosting Classifier is a good model because of its ability to capture complex, non-linear relationships, making it suitable for analyzing multifaceted factors like retention. This model has been effectively applied in similar studies, such as Yu et al. (2019), where Gradient Boosting was used to predict resignation patterns in corporate settings, demonstrating its ability to identify subtle predictors in imbalanced datasets. Uncovering the dominant role of satisfaction while minimizing the impact of remote work and performance categories, Gradient Boosting provides an understanding of resignation dynamics.

```
Classification Report:
              precision    recall  f1-score   support

       False       0.56      0.51      0.53     26904
        True       0.55      0.60      0.58     27090

    accuracy                           0.55     53994
   macro avg       0.55      0.55      0.55     53994
weighted avg       0.55      0.55      0.55     53994


Confusion Matrix:
[[13616 13288]
 [10770 16320]]
```

| | Feature | Importance |
|---|---|---|
| 1 | Employee_Satisfaction_Score | 0.992988 |
| 0 | Remote_Work_Frequency | 0.005091 |
| 2 | Performance_Group_Mid Performer | 0.001085 |
| 3 | Performance_Group_High Performer | 0.000837 |

**Figure 7**

## 5. Observations and Conclusions

### 5.1 Key Findings

In conclusion, our project provided valuable insights into the relationships between remote work frequency, employee performance, satisfaction, and retention. For high performers, statistical analysis revealed that remote work frequency does not significantly impact satisfaction levels, as demonstrated by the ANOVA test and consistent results across remote work categories. Similarly, for low performers, remote work frequency showed no significant influence on productivity metrics such as projects handled and overtime hours, as confirmed by the Kruskal-Wallis test, ANOVA, and T-tests. In analyzing retention, machine learning models like Random Forest and Gradient Boosting Classifiers highlighted employee satisfaction scores as the most influential factor, while remote work frequency and performance categories had minimal direct impact. These findings suggest that while remote work frequency has limited influence on employee satisfaction and productivity, fostering employee satisfaction remains critical for improving retention across all performance categories.

### 5.2 Limitations and Future Work

Despite the robustness of our analyses, we faced certain limitations. First, the dataset was imbalanced, particularly in the resignation variable, which impacted the initial logistic regression model's performance. Although SMOTE helped address this issue, the models still struggled to predict resignations effectively, indicating the need for more sophisticated approaches. Additionally, the dataset lacked detailed contextual variables such as employee preferences, specific job roles, or organizational policies, which could further explain variations in satisfaction and retention. Future work could focus on integrating additional data sources, such as survey feedback or qualitative insights, to capture these nuances. Moreover, exploring more advanced models like XGBoost or neural networks may enhance prediction accuracy and uncover deeper patterns, offering a more comprehensive understanding of the dynamics between remote work and workforce outcomes.

## 6. References

[1] Bloom, Nicholas. "How working from home works out." *Stanford Institute for Economic Policy Research*, June 2020, https://siepr.stanford.edu/publications/policy-brief/how-working-home-works-out

[2] Luc, Kelly, and Jackson, Fresia. "Improving employee performance: Categorizing high, mid, and low performers." *Culture Amp*, 9 November 2022, https://www.cultureamp.com/blog/improving-employee-performance

[3] Ray, Tapas K, and Pana-Cryan, Regina. "Work flexibility and worker well-being." *Centers for Disease Control and Prevention*, 11 March 2022, https://blogs.cdc.gov/niosh-science-blog/2022/03/11/work-flexibility/

[4] Peckham, Trevor. "Evaluating Employment Quality as a Determinant of Health in a Changing Labor Market." *National Library of Medicine*, 5 September 2019, https://pmc.ncbi.nlm.nih.gov/articles/PMC6756794/

[5] Strobl, C., Boulesteix, AL., Zeileis, A. et al. "Random forest variable importance measures: Illustrations, sources and a solution." *BMC Bioinformatics*, 25 January 2007, https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-25