Absenteeism in the Workplace

Bradley Chang, Elijah Colwill, Haleigh Gronwold, Yeling Jiang, Akshaya Kumar, Alex Liu, Shashi Pingolia, Rachel Zhang, Coby Zorn

PURDUE UNIVERSITY

Introduction

Absenteeism is a large and growing issue in the manufacturing industry. Based on market research:

- The average manufacturing absence rate was 3% in 2020. [1]
- Over 2% of total work time was lost due to absenteeism. [1]
- A shift worker in the U.S. costs a company ≈ \$3,600 in annual absenteeism costs. [2]
- Direct costs include payroll and overtime; indirect costs include increased workload and reduced quality of work. [2]

Our team has estimated, based on average industry absence rates and costs per shift worker, that the annual cost of absenteeism for large manufacturing companies is between 4 and 5 million dollars.

Objective

Identify existing patterns of absenteeism ❯ Investigate reasons behind rates of absenteeism in departments ❯ Create and test models, and interpret their findings

Research Methodologies

Analysis
- Exploratory Analysis
- Correlation Testing
- Significance Testing
- Data Visualization
- Datasets: Absenteeism, Quality, Engagement

Modeling
- ANOVA
- Time Series
- Regression
- Toolset: R, Python, GitHub, Jupyter Notebook

Deliverables
- Feature-Engineered Datasets
- Data-Driven Insights
- Forecasting Models

Data Overview

We began by pre-processing the data (i.e. data cleaning, data type conversion, and eliminating missing values) and by feature engineering our three datasets.

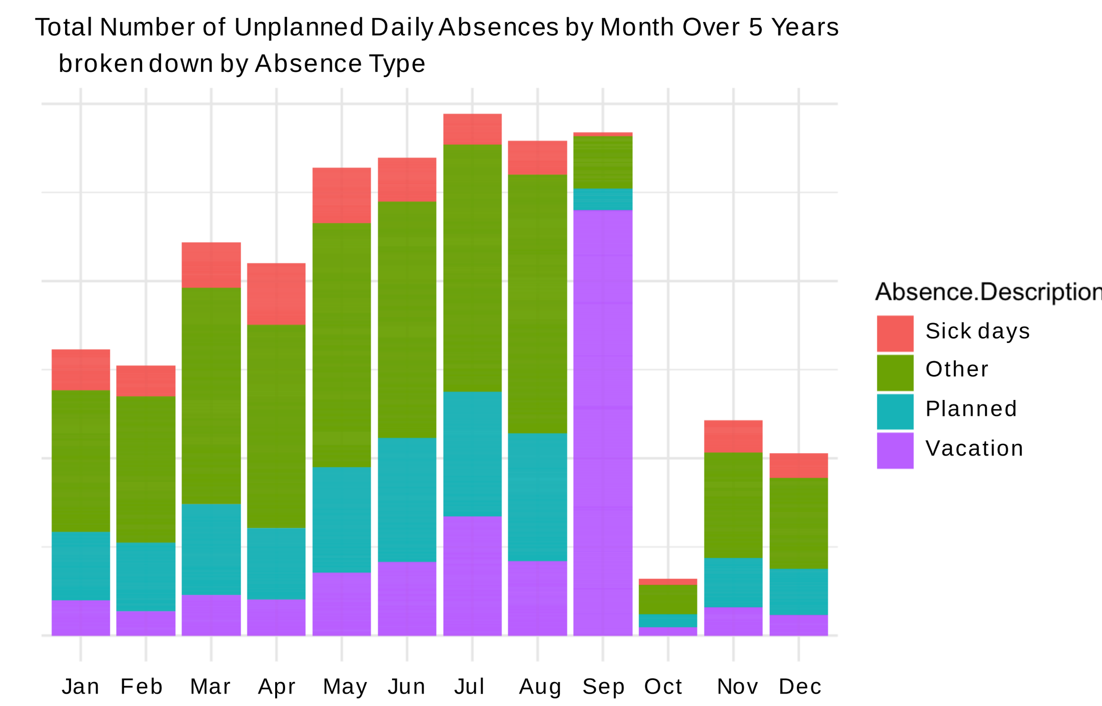
Engagement Dataset
└ Survey Items, Participation Rates, Engagement Score, Perception of Survey Responses, etc.

Quality Score Dataset
└ Start Date of Week, Work Department, Quality Score
- + Validation Set
  └ data not in Absenteeism dataset (i.e., FY21)

Absenteeism Dataset
└ Work Department, Shift, Assigned Employee ID, Absence Type, Absence Description, etc.
- + Day of Week
- + Fiscal Year
- + Department Size
  └ by aggregating unique employee IDs per department per fiscal year
- + Federal Holiday Flag
- + Unplanned Leave
  └ by removing federal holidays and inventory adjustment days

We used these three files to generate a master dataset

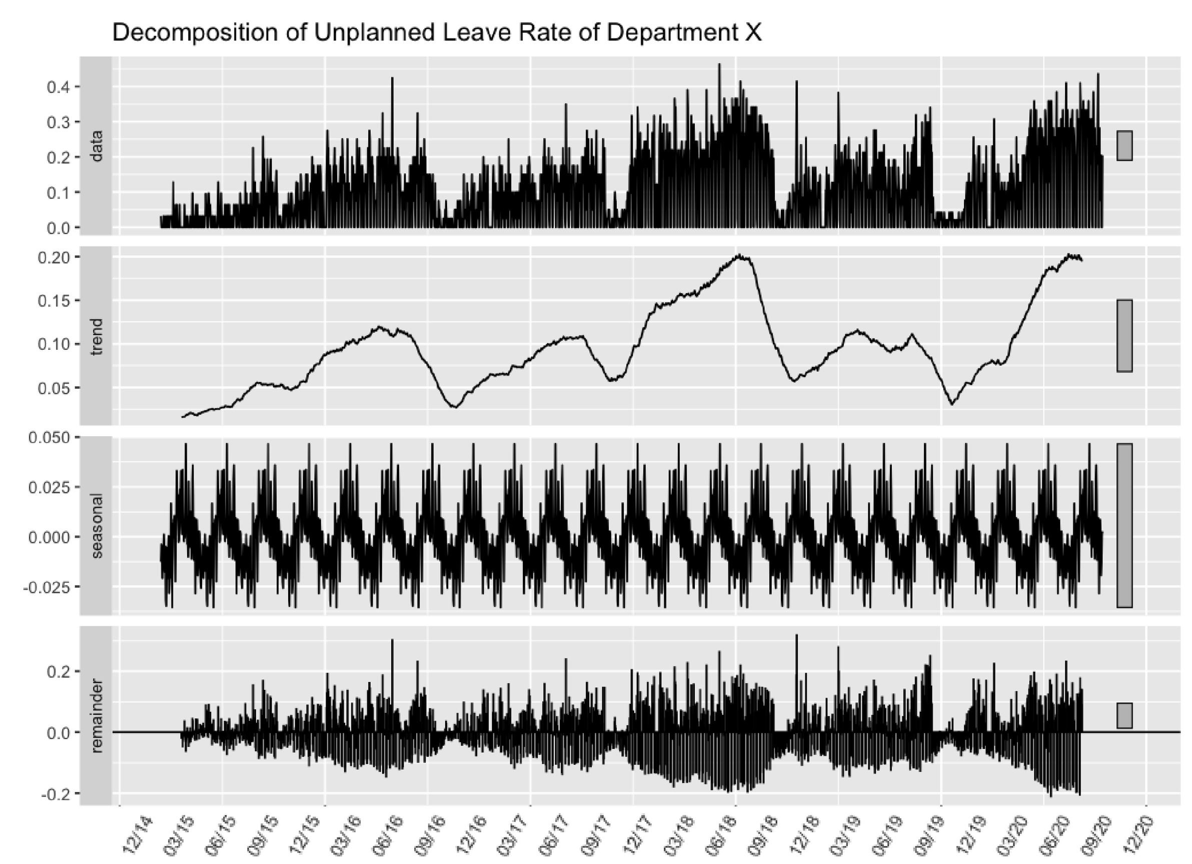
Through exploratory analysis, we found that the reason for absence each month differs significantly.



Time Series for Trend & Seasonality

By using the Auto Regressive Integrated Moving Average (ARIMA) model, we decomposed the unplanned leave rate series into trend, seasonality, and residuals. We found that:

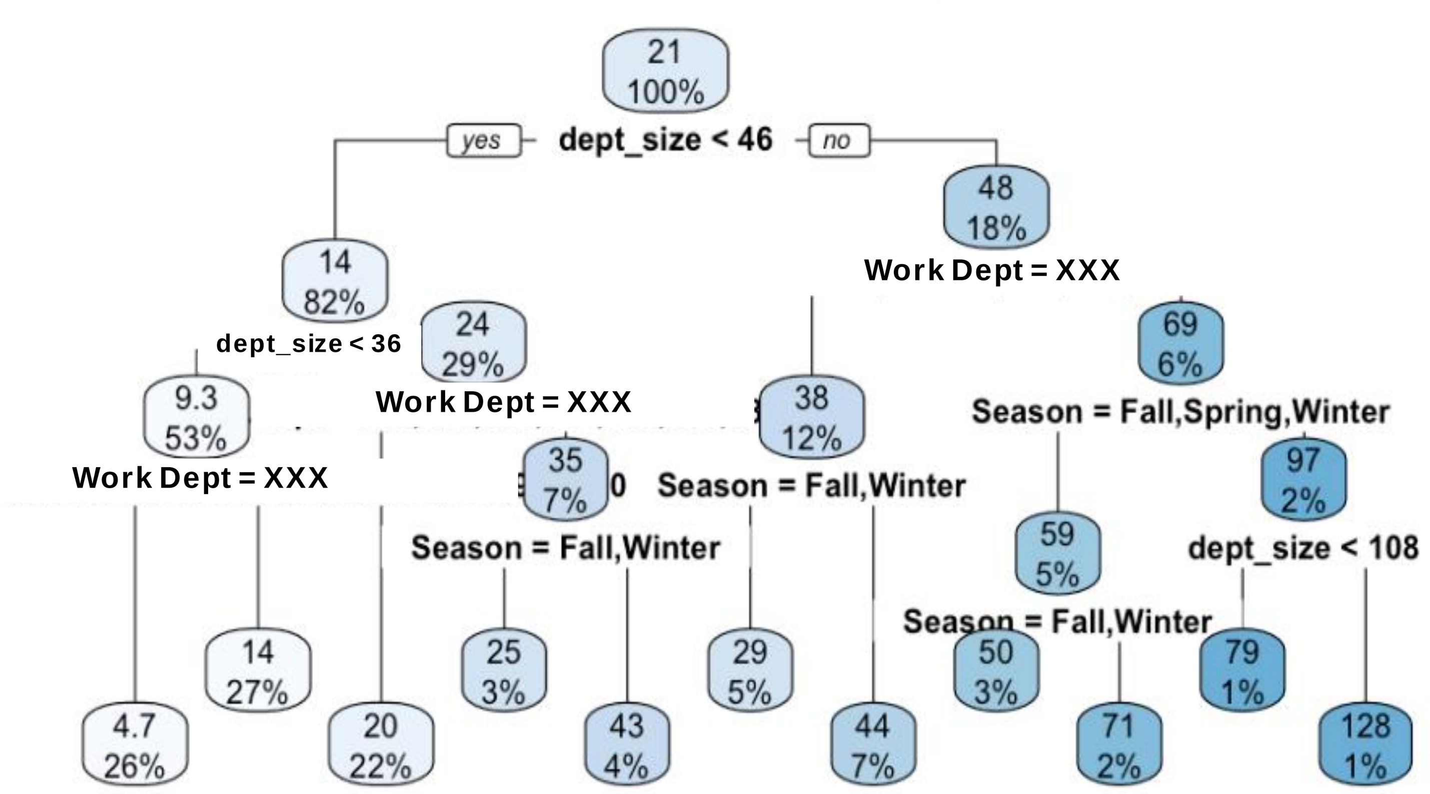
- Absence rate fluctuates with quarterly seasonality.
- Around half of the departments have strong timely patterns and are perfect for an ARIMA model.
- September tends to have higher unplanned leave rates. This might be because of hunting season and planned factory shutdowns, according to HR and Operations Managers.



Model Results & Evaluation

We used multiple linear regressions (MLR) with double-log, quadratic, and interaction variations and a decision tree to forecast weekly unplanned leave and product quality.

$$\text{Weekly Unplanned Leave} = -0.39 + \sum_{i=1}^{48} \beta_i\, Dept\ i + \sum_{j=1}^{52} \beta_j\, Week\ j + 0.74\ Dept\ size - 0.31 Quality + \sum_{k=1}^{4} \beta_k\, Season\ k$$



We evaluated model performances by adjusted R^2, mean absolute error (MAE), and mean absolute percentage error (MAPE).

Model	Adjusted R^2	MAE	MAPE
Times series ARIMA*	-	2.65	0.50
MLR	0.675	7.89	0.80
Log-log MLR	0.686	8.83	0.82
MLR with Quality^2	0.62	8.74	0.88
MLR with Interactions	0.674	7.93	0.81
Decision Tree	-	7.66	0.78

* ARIMA model tests on department-level data, while the rest test on factory-level data.

Conclusions

- Our forecasting models can predict weekly unplanned absences with high accuracy. The decision tree and time series models have the highest accuracy followed by the multi-linear regression model.

ANOVA Results

- More absences occur on Mondays & Fridays.
- Occupational injury is higher in October and November. Sick day absences are significant in Winter.

Regression Results

- No significant relationship exists between the number of absences and production quality.
- The absenteeism rate is negatively related to the engagement score.

Next Steps

- ***Expanded Data*** – Data from other factories without seasonal bias, differing production types, etc. for further correlation analysis.
- ***Individual Data*** – Access to individual level engagement data for more granular insights.
- ***Fiscal Impact*** – Quantifiable measurement of fiscal impact of absenteeism to target monetary impact instead of rate.
- ***Cross-validation of models*** – Train the models with more datasets using cross-validation to avoid overfitting and improve model performances.

References

[1] "Absences from Work of Employed Full-Time Wage and Salary Workers by Occupation and Industry." *U.S. Bureau of Labor Statistics*, 22 Jan. 2021, www.bls.gov/cps/cpsaat47.htm.

[2] Folger, Jean. "The Causes and Costs of Absenteeism." *Investopedia*, Investopedia, 29 Aug. 2020, www.investopedia.com/articles/personal-finance/070513/causes-and-costs-absenteeism.asp.

Acknowledgements

Thank you to our corporate mentors, Emma A. and Dave L., and our Purdue mentors, Nathan DeLay and the Data Mine staff, for their guidance and support!