# Predicting Absenteeism in the Workplace

By: Myra Rust

## Executive Summary

< To Do: Write abstract with short description of project objective, methods, and results>

## Introduction/Background

Workplace absenteeism negatively impacts businesses in a number of ways. It reduces overall productivity, it negatively impacts employee morale, and it cuts down on the profit margin of the business. < Insert statistics on how much, in dollars or percentage of profits, absenteeism costs employers on average > This project will use data collected from a business on employee absenteeism over a period of three years to conduct research using data science methodologies and techniques to identify possible causes and patterns in absenteeism.

The objective of this project is to answer the following research questions:

Can absenteeism be predicted?

What are the main causes of high absenteeism?

Are there patterns in absence rates by season or day of the week?

Are there groups of contributing factors that lead to high absenteeism?

Providing employers with information on when higher absenteeism will occur and reasons surrounding the absences will provide them with information that can be used to assist in absence management and ultimately lead to better productivity, a more positive employee experience, and higher profit margins.

## Preliminary Analysis

Data Source: The dataset contains 740 observations with 21 features detailing reasons for employee absences and personal/demographic information of the employee. This dataset was recorded by

a courier company in Brazil between July 2007 to July 2010.

Link to dataset: https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work

Data Cleaning: The dataset consisted mostly of numbers and required very little cleaning. Column names were changed to be more user friendly, weight column was converted from kilograms to pounds, and height column was converted from centimeters to inches.

List of Features:

1. Individual identification (id)
2. Reason for absence (reason) – Includes 21 categories linked to International Code of Diseases (ICD) medical reasons and 7 categories for other extenuating factors.
    a. (1) Certain infectious and parasitic diseases, (2) Neoplasms, (3) Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism, (4) Endocrine, nutritional and metabolic diseases, (5) Mental and behavioral disorders, (6) Diseases of the nervous system, (7) Diseases of the eye and adnexa, (8) Diseases of the ear and mastoid process, (9) Diseases of the circulatory system, (10) Diseases of the respiratory system, (11) Diseases of the

digestive system, (12) Diseases of the skin and subcutaneous tissue, (13) Diseases of the musculoskeletal system and connective tissue, (14) Diseases of the genitourinary system, (15) Pregnancy, childbirth and the puerperium, (16) Certain conditions originating in the perinatal period, (17) Congenital malformations, deformations and chromosomal abnormalities, (18) Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified, (19) Injury, poisoning and certain other consequences of external causes, (20) External causes of morbidity and mortality, (21) Factors influencing health status and contact with health services, (22) Patient follow-up, (23) Medical consultation, (24) Blood donation, (25) Laboratory examination, (26) Unjustified absence, (27) Physiotherapy, (28) Dental consultation.

3. Month of absence (month)
4. Day of the week (weekday)
   a. (2) Monday, (3) Tuesday, (4) Wednesday, (5) Thursday, (6) Friday
5. Seasons (season)
6. Transportation expense (trans_expense)
7. Distance from Residence to Work in Kilometers (distance)
8. Service time (service_time)
9. Age (age)
10. Daily Work load Average (workload)
11. Hit target (hit_target)
12. Disciplinary failure (discipline_fail)

a. (0) No, (1) Yes

13. Education (education)

    a. (1) High school, (2) Graduate, (3) Postgraduate, (4) Master and doctor

14. Number of Children (children)

15. Social drinker (drinker)

    a. (0) No, (1) Yes

16. Social smoker (smoker)

    a. (0) No, (1) Yes

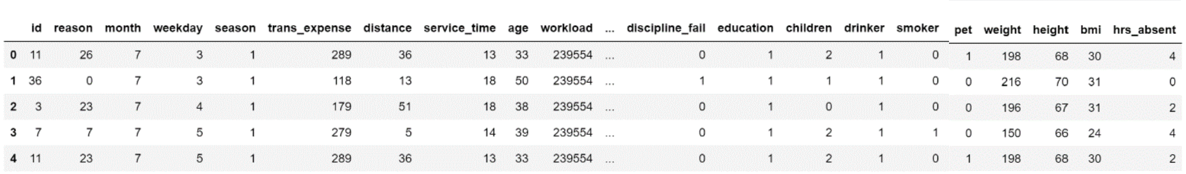
17. Number of Pets (pet)

18. Weight in pounds (weight)

19. Height in inches (height)

20. Body mass index (bmi)

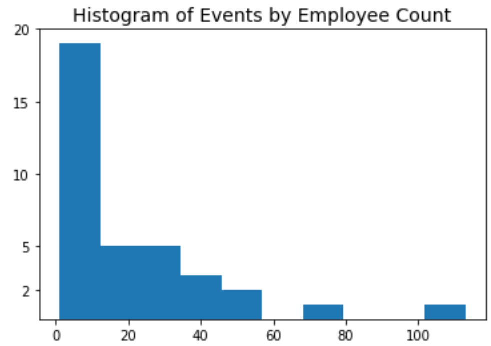
21. Absenteeism time in hours (hrs_absent)

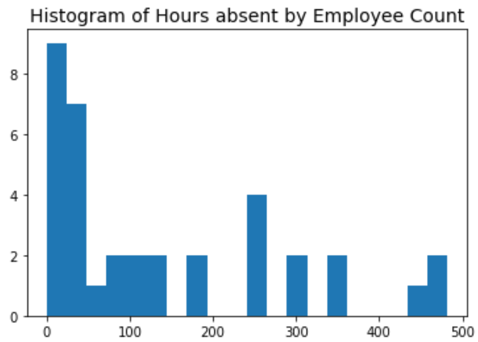


| | id | reason | month | weekday | season | trans_expense | distance | service_time | age | workload | ... | discipline_fail | education | children | drinker | smoker | pet | weight | height | bmi | hrs_absent |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 11 | 26 | 7 | 3 | 1 | 289 | 36 | 13 | 33 | 239554 | ... | 0 | 1 | 2 | 1 | 0 | 1 | 198 | 68 | 30 | 4 |
| **1** | 36 | 0 | 7 | 3 | 1 | 118 | 13 | 18 | 50 | 239554 | ... | 1 | 1 | 1 | 1 | 0 | 0 | 216 | 70 | 31 | 0 |
| **2** | 3 | 23 | 7 | 4 | 1 | 179 | 51 | 18 | 38 | 239554 | ... | 0 | 1 | 0 | 1 | 0 | 0 | 196 | 67 | 31 | 2 |
| **3** | 7 | 7 | 7 | 5 | 1 | 279 | 5 | 14 | 39 | 239554 | ... | 0 | 1 | 2 | 1 | 1 | 0 | 150 | 66 | 24 | 4 |
| **4** | 11 | 23 | 7 | 5 | 1 | 289 | 36 | 13 | 33 | 239554 | ... | 0 | 1 | 2 | 1 | 0 | 1 | 198 | 68 | 30 | 2 |

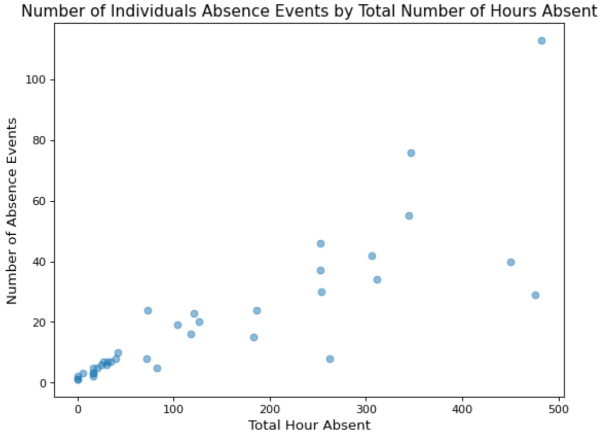
Preliminary Observations:

Workplace absences occur for many different reasons but there are individuals that incur more absences than others. This dataset tracked absences from 36 employees over three years. The following histograms depict employee counts for number of absence events and number of hours absent.



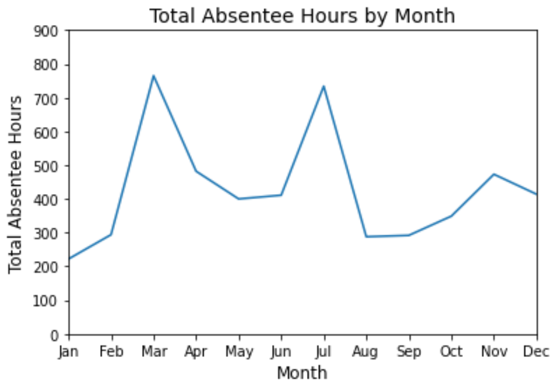


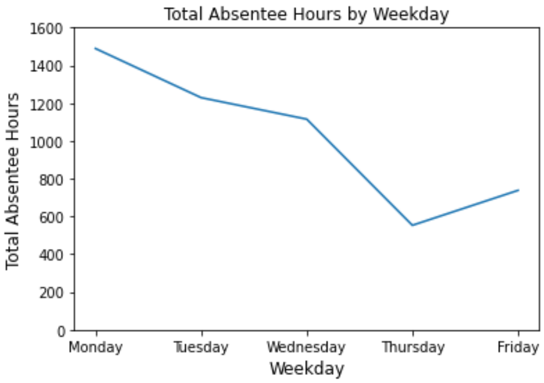
As you can see most employees had between 0-10 absence events, but there were some individuals that had more, including one employee that had 113 absence events. As far as number of hours absent, we see the same trend. Most employees fell into the range of either 0-25 hours absent or 25-50 hours absent, but a few employees had over 400 absentee hours. This coordinates with what is shown in the below scatterplot. As the number of absence events increases, so does the number of hours absent, with the majority of employees falling in the lower left range and a small number of employees incurring the majority of absence events and hours absent.



< Possibly enter counts of events/hours by reason. Provide Table printout. Might be too much though…maybe put this in appendix? >

The season appears to have little connection to absentee hours, with the number of total hours absent fluctuating only ~300hrs, between seasons. However, when looking at monthly and day of the week absentee hours, we can see obvious trends. March and July have significant spikes in absentee hours and most absentee hours occur on Mondays, decreasing thereafter until Thursday with a slight rise again on Fridays.


Total Absentee Hours by Month


Total Absentee Hours by Weekday

No significant correlation between features was discovered. Also, worthwhile to note is that most features had very little correlation with the target feature (hrs_absent) as well. The correlation heatmap and correlation to target feature visualization can be found in Appendix A.

< To Do: Figure out how to create an actual appendix and link it to the above text >

Data Transformation: Once all data cleaning was complete categorical variable were transformed using one-hot encoding, the data was scaled using the MinMaxScaler function and then divided into a 70/30 train-test split.

## Feature Selection

<still working>

Right now, I am trying to figure out the best way to conduct feature selection. I have tried correlation feature selection, mutual information feature selection, and next will try decision tree feature selection, which hopefully will give me better results than the other two, because initial modeling results are not looking so good.

## Model Selection and Evaluation

<still working>

Linear Regression to predict number of hours an employee will be absent based off the employee's demographic information.

If I have time, I would like to try clustering to see if I can find groups of features that correlate with high rates of absenteeism.

## Results

<still working>

## Conclusion

<still working>

## Future Work

<still working>

## Acknowledgements

I would like to thank Professor Catherine Williams and all other professors that are part of Bellevue University's data science program. Without their direction, support, and guidance, this paper would not have been possible. I would also like to thank my daughter for sacrificing her time with me to allow me to pursue this degree for the betterment of myself.

## References

Cushard, B. (2021). The Impact of Absenteeism. ADP, LLC. Retrieved from https://www.adp.com/spark/articles/2017/01/the-impact-of-absenteeism.aspx.

bountiXP Team. (June 18, 2020). The Real Effects of Absenteeism in the Workplace. bountiXP. Retrieved from https://www.insights.bountixp.com/blog/effects-of-absenteeism-in-the-workplace.

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Investopedia Contributor. (n.d.). The Causes and Costs of Absenteeism in the Workplace. Forbes. Retrieved from https://www.forbes.com/sites/investopedia/2013/07/10/the-causes-and-costs-of-absenteeism-in-the-workplace/?sh=23df84e53eb6.

Martiniano, A., Ferreira, R. P., Sassi, R. J., & Affonso, C. (2012). Application of neuro fuzzy network in prediction of absenteeism at work. In Information Systems and Technologies (CISTI), 7th Iberian Conference on (pp.1-4). IEEE.

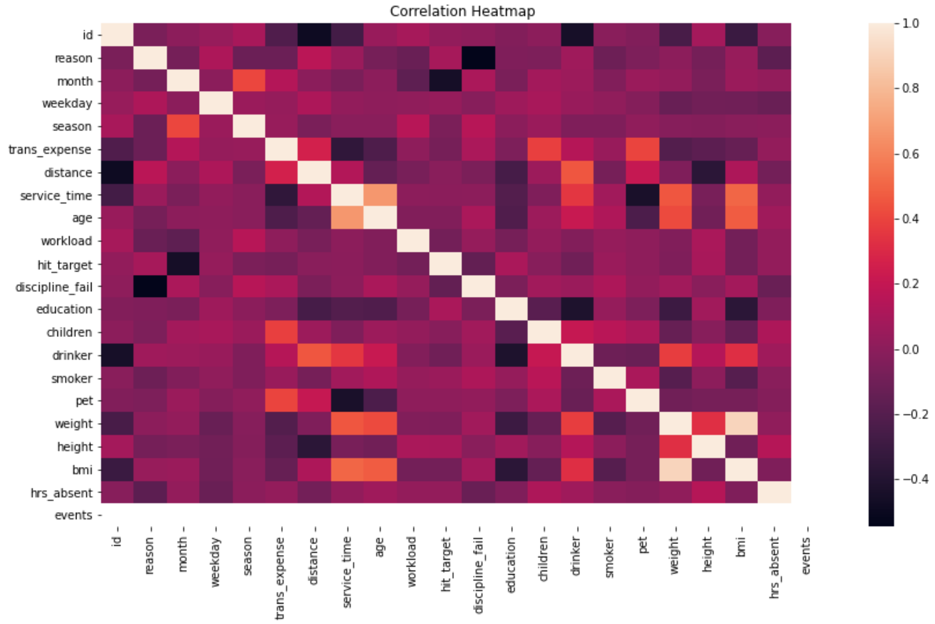
Sharma, Shiv. (n.d.). What is Absence Management? 6 Ways to Reduce Absenteeism at Work. TaskWorld. Retrieved from https://taskworld.com/blog/what-is-absence-management-6-ways-to-reduce-absenteeism-at-work/.

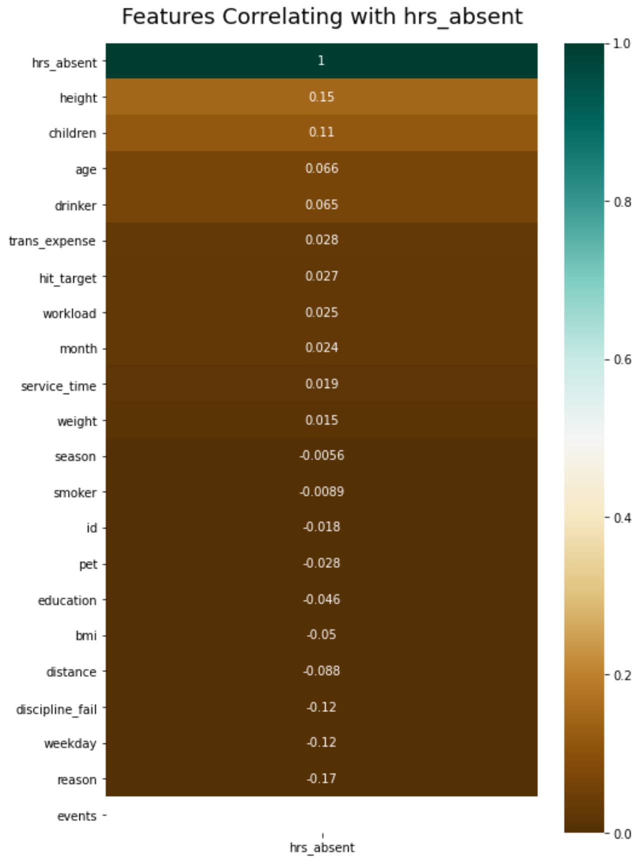
Society for Human Resource Management (SHRM). (December 15, 2014). Employee Absences Have Consequences for Productivity and Revenue, SHRM Research Shows. SHRM. Retrieved form https://www.shrm.org/about-shrm/press-room/press-releases/pages/employeeabsencessurvey.aspx.

U.S. Bureau of Labor Statistics. (February 23, 2021). Data on Absences from Work. U.S. Bureau of Labor Statistics. Retrieved from https://www.bls.gov/cps/absences.htm.

Van Vulven, Erik. (n.d.). Absenteeism in the Workplace: A Full Guide. AIHR Academy. Retrieved from https://www.aihr.com/blog/absenteeism/.

# Appendix A.


Correlation Heatmap
id
reason
month
weekday
season
trans_expense
distance
service_time
age
workload
hit_target
discipline_fail
education
children
drinker
smoker
pet
weight
height
bmi
hrs_absent
events
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4


Features Correlating with hrs_absent
hrs_absent 1
height 0.15
children 0.11
age 0.066
drinker 0.065
trans_expense 0.028
hit_target 0.027
workload 0.025
month 0.024
service_time 0.019
weight 0.015
season -0.0056
smoker -0.0089
id -0.018
pet -0.028
education -0.046
bmi -0.05
distance -0.088
discipline_fail -0.12
weekday -0.12
reason -0.17
events
hrs_absent
1.0
0.8
0.6
0.4
0.2
0.0