Absenteeism Data Set
https://www.kaggle.com/datasets/HRAnalyticRepository/absenteeism-dataset/data

Data Summary

Data Source: It is an artificially generated HR Analytics data, generated by Lyndon Sundmark, who is a well-known author and data scientist in the field of HR Analytics. It is generated for the purpose of practice and learning, specifically for HR Analysts and for the purpose of learning to predict absence as an outcome. External Data.

Data Collection: Fictitious data

Data Content: This HR data set focuses on employee absence. It contains 8335 rows and 13 columns of data. The data set contains employee numbers and names, gender, city, job title, department, store location, business unit, division, age, length of service, and the number of hours absent.Absence is taken as the total annual absent hours per employee. The data is assumed for one year.

Why this data set?

I want to connect my previous professional experience as a recruiter to my data analysis studies and hopefully land a job

in HR Analytics. Therefore, this data set and this analysis would make a great addition to my portfolio.

Finding HR Analytics open datasets is not easy, since HR and personnel data is highly personal and confidential. It took me quite a long time to find a data set that matches the criteria given in the project brief, therefore I would really like to be able to use it.

Data Profile

6) Cleaning Your Data

1. Checked for duplicates, missing values and inconsistencies with data type. It all came back negative.
2. Filtered by age where it is <18 and >65 since it was not meaningful in this context and dropped these rows.
3. Dropped the columns EmployeeName, FirstName and LastName because of PIP data.

7) Basic Descriptive Statistics

	Age	LengthService	AbsentHours
count	8256.000000	8256.000000	8256.000000
mean	42.270237	4.786490	61.877815
std	9.615567	2.469612	48.901086
min	18.204720	0.053279	0.000000
25%	35.521573	3.578938	20.681011
50%	42.220543	4.599592	56.624944
75%	48.731601	5.624946	94.634461
max	77.938003	43.735239	272.530123

Most values appear reasonable in context. Will check for outliers later if needed

Data Limitations: Artificially generated data may not accurately represent real-world distributions and patterns. The data could miss nuances present in actual datasets. If the data generation process incorporates certain assumptions or rules, it may introduce biases that wouldn't exist in real-world data. While synthetic data avoids issues like privacy violations, the generation process might still embed sensitive or controversial patterns unintentionally.

Questions to Explore

- What is the overall absenteeism rate in the organization? How does it compare across different divisions or business units?
- How does absenteeism vary by gender? Are there significant differences in absenteeism rates between male and female employees?
- Is there a correlation between age and absenteeism rates? Do younger or older employees tend to have higher absenteeism?
- Which job titles or departments exhibit the highest absenteeism rates? What might contribute to these trends?
- Are there specific departments that consistently show lower absenteeism rates? What practices might they have in place?
- How does length of service affect absenteeism rates? Do newer employees tend to have higher absenteeism compared to those who have been with the organization longer?
- Is there a threshold of length of service beyond which absenteeism significantly decreases?
- How do absenteeism rates vary across different store locations? Are certain locations facing more challenges with absenteeism?