EMPLOYEE ABSENTEEISM ANALYSIS

EXPLORING EMPLOYEE ABSENTEEISM THROUGH DASHBOARDS & MACHINE LEARNING

In this project, I explored a cleaned absenteeism dataset using data science tools in Python and visual analytics in Tableau. The goal was to understand factors that drive employee absenteeism and use predictive models to estimate absentee hours. I followed a structured workflow from data cleaning and transformation through modeling and dashboard creation to generate actionable insights.

📊 **Tableau Dashboard Link:**

https://public.tableau.com/views/EmployeeAbsenteeismAnalysisDashboard/Dashboard1?:language=en-GB&:sid=&:redirect=auth&:display_count=n&:origin=viz_share_link

1. INTRODUCTION AND PROBLEM DEFINITION

This project was completed as part of the *Google Data Analytics Professional Certificate* and focuses on analyzing a dataset related to employee absenteeism. The goal was to explore what factors influence how many hours employees are absent from work, and to use both data visualization and machine learning to identify trends and make predictions. Absenteeism can be costly for companies, and understanding the patterns behind it can help HR teams make better decisions about resource planning, employee wellness, and productivity.

2. DATASET AND PREPROCESSING

The dataset was provided as part of the course on Coursera. It contains HR records of employees from a courier company, with information about the reasons for their absence, the time missed, and several demographic and work-related features. The original data included 28 detailed medical reason codes based on ICD-10 categories, which were difficult to analyze directly.

To make the analysis more meaningful and easier to interpret, I grouped these 28 reasons into 4 main categories:

- Group 1: Medical conditions (e.g., infections, neoplasms, mental disorders)
- Group 2: Pregnancy and related
- Group 3: Musculoskeletal issues and trauma
- Group 4: Routine check-ups, dental visits, and minor administrative reasons

I created four binary columns to represent each group and later combined them into one simplified Reason_Group column. I also checked for missing values and duplicate entries, but the dataset was already cleaned. Additional columns like "ID" or others not needed for modeling were dropped. Finally, I applied a **log transformation** on the *Absenteeism Time in Hours* column, because it was highly skewed; most people missed very few hours, but a few had extremely high values.

	Reason_1	Reason_2	Reason_3	Reason_4	Month Value	Day of the Week	Transportation Expense	Distance to Work	Age	Daily Work Load Average	Body Mass Index	Education
count	700.000000	700.000000	700.000000	700.000000	700.000000	700.000000	700.000000	700.000000	700.000000	700.000000	700.000000	700.000000
mean	0.250000	0.008571	0.090000	0.597143	6.897143	2.410000	222.347143	29.892857	36.417143	271.801774	26.737143	0.167143
std	0.433322	0.092250	0.286386	0.490823	3.342319	1.761669	66.312960	14.804446	6.379083	40.021804	4.254701	0.373370
min	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	118.000000	5.000000	27.000000	205.917000	19.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	4.000000	1.000000	179.000000	16.000000	31.000000	241.476000	24.000000	0.000000
50%	0.000000	0.000000	0.000000	1.000000	7.000000	2.000000	225.000000	26.000000	37.000000	264.249000	25.000000	0.000000
75%	0.250000	0.000000	0.000000	1.000000	10.000000	4.000000	260.000000	50.000000	40.000000	294.217000	31.000000	0.000000
max	1.000000	1.000000	1.000000	1.000000	12.000000	6.000000	388.000000	52.000000	58.000000	378.884000	38.000000	1.000000

Figure 1: Sample of the dataset after preprocessing. Reason codes have been grouped, missing values checked, and a log transformation applied to the target variable.

3. EXPLORATORY DATA ANALYSIS

To begin understanding the data, I created visualizations and summary statistics using Python (mainly pandas, seaborn, and matplotlib). The distribution of absenteeism hours showed that most employees were absent for 1–4 hours, with a few being absent for up to 120 hours. This justified using a log transformation for modeling.

As shown in Figure 2 below, absenteeism hours are heavily skewed toward shorter durations.

```
# Plotting the distribution of absenteeism time
plt.figure(figsize=(10, 6))
sns.histplot(df['Absenteeism Time in Hours'], bins=30, kde=True)
plt.title("Distribution of Absenteeism Time in Hours")
plt.xlabel("Hours Absent")
plt.ylabel("Number of Employees")
plt.show()
```

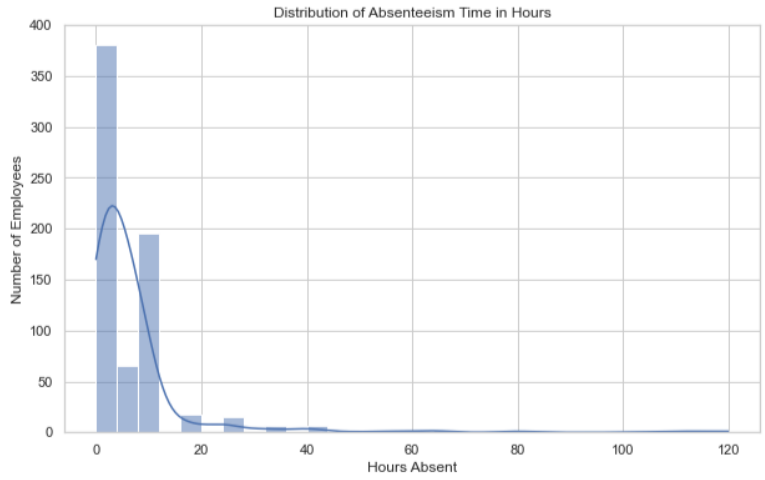

Figure 2: Distribution of absenteeism hours. Most employees are absent for fewer than 10 hours, but a few outliers exist, up to 120 hours.

I then explored how absenteeism varied across different features. Reason Group 1 had the highest average absence hours, which makes sense since it includes many medical conditions. March and November appeared to be the months with the highest total absenteeism. When looking at demographics, I noticed that younger and mid-aged employees (around 30–45 years) were absent more often. The number of children and pets had a small but visible impact — employees with dependents had slightly more hours absent. Education level seemed to have a reverse trend: employees with higher education levels had fewer hours missed on average.

A correlation heatmap confirmed that most features did not have a strong linear relationship with absenteeism time, which suggested that non-linear models like decision trees or ensemble methods might perform better.

```
# Correlation heatmap for all numerical columns
plt.figure(figsize=(12, 8))
sns.heatmap(df.corr(), annot=True, cmap='coolwarm', fmt='.2f')
plt.title("Correlation Heatmap")
plt.show()
```

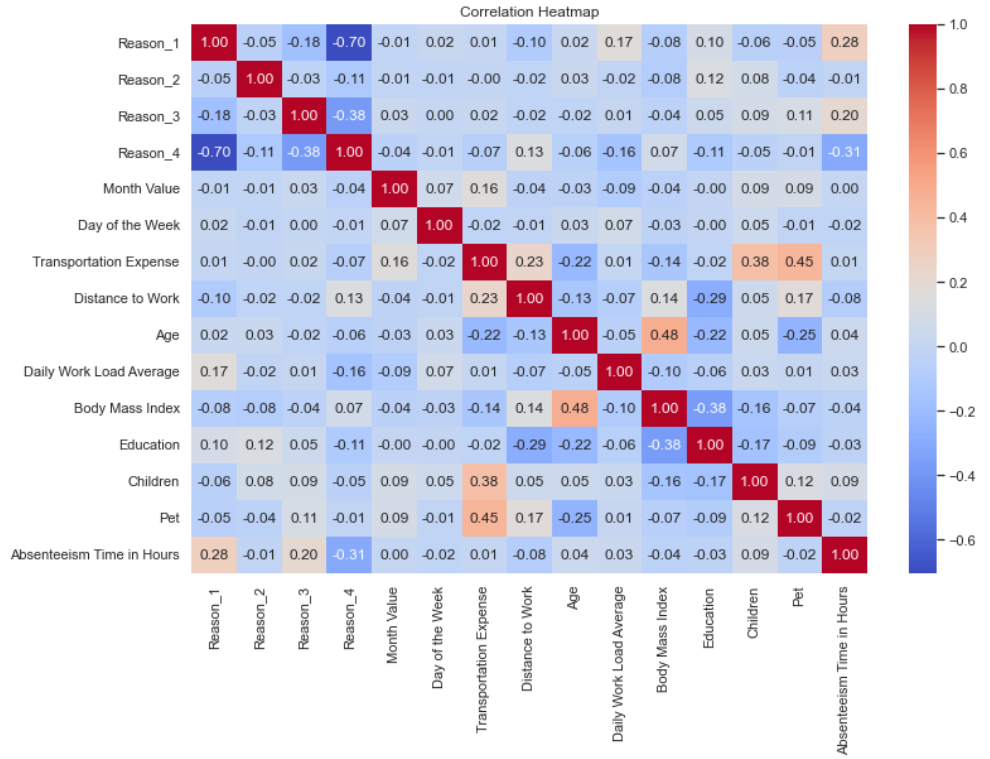

Figure 3: Correlation heatmap between numerical features. No strong linear relationships are observed between predictors and absenteeism time.

The heatmap in Figure 3 above highlights the weak linear correlation among most features, suggesting the use of non-linear models.

4. REGRESSION MODELLING

After completing the exploratory analysis, I built three models to predict absenteeism hours: Linear Regression, Random Forest, and XGBoost.

Linear Regression served as a baseline but performed poorly, explaining only 6 percent of the variance ($R^2 = 0.06$) with an RMSE of about 12.1 hours. This confirmed that absenteeism patterns are not easily captured by simple linear methods.

The Random Forest model, applied to the log-transformed target, performed much better with an R^2 of 0.46 and RMSE around 11.2 hours. XGBoost delivered similar results ($R^2 = 0.42$, RMSE $\approx$ 11.8), slightly behind Random Forest but still effective.

Both models identified the same top predictors: Reason Group, Month Value, Daily Work Load Average, and Education; all of which significantly influenced absence duration.

```
Random Forest R²: 0.46
Random Forest RMSE (log scale): 0.60
```

```
y_pred_original = np.expm1(y_pred_log)
y_test_original = np.expm1(y_test_log)

rmse_original = np.sqrt(mean_squared_error(y_test_original, y_pred_original))
print(f"Random Forest RMSE (original hours): {rmse_original:.2f}")
```

```
Random Forest RMSE (original hours): 11.24
```

```
feature_importance = pd.Series(rf.feature_importances_, index=X.columns)
feature_importance = feature_importance.sort_values(ascending=True)

plt.figure(figsize=(8, 6))
sns.barplot(x=feature_importance, y=feature_importance.index)
plt.title("Feature Importance (Random Forest)")
plt.xlabel("Importance Score")
plt.ylabel("Feature")
plt.show()
```

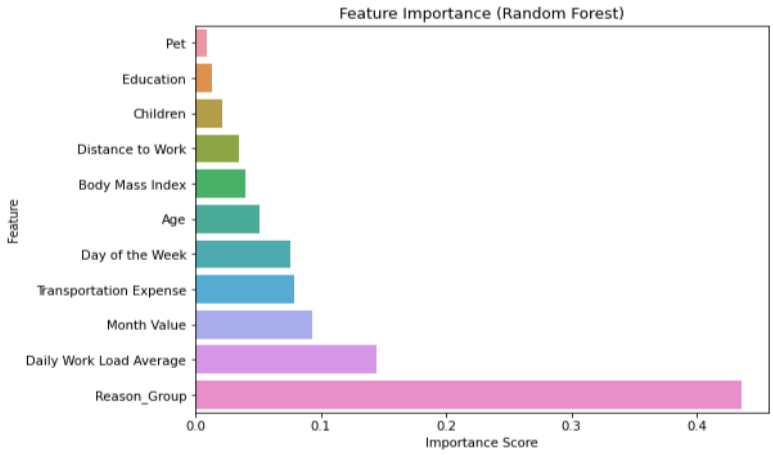

Figure 4: Feature importance from the Random Forest model. Reason Group, Month Value, Work Load, and Education are the most predictive features.

5. IMPLEMENTATION IN TABLEAU

To make the project interactive and visually appealing, I designed a dashboard using Tableau Public. I exported a cleaned version of the dataset (Absenteeism_for_Tableau.csv) and used it to build the dashboard entirely online using Tableau Web Authoring.

The dashboard includes three main views:

1. **Bar chart by Reason Group** — shows which categories cause the most absenteeism.
2. **Line chart by Month** — displays seasonal trends in absenteeism.
3. **Bar chart by Age Group** — compares average hours missed across age brackets.

I also added filters so users can view trends based on education level, number of children, and number of pets. These filters allow HR staff or other users to explore the data based on real-world segments.

The dashboard is styled with clean colors, visible legends, and hover tooltips to make it easy to understand without needing to look at the raw data.

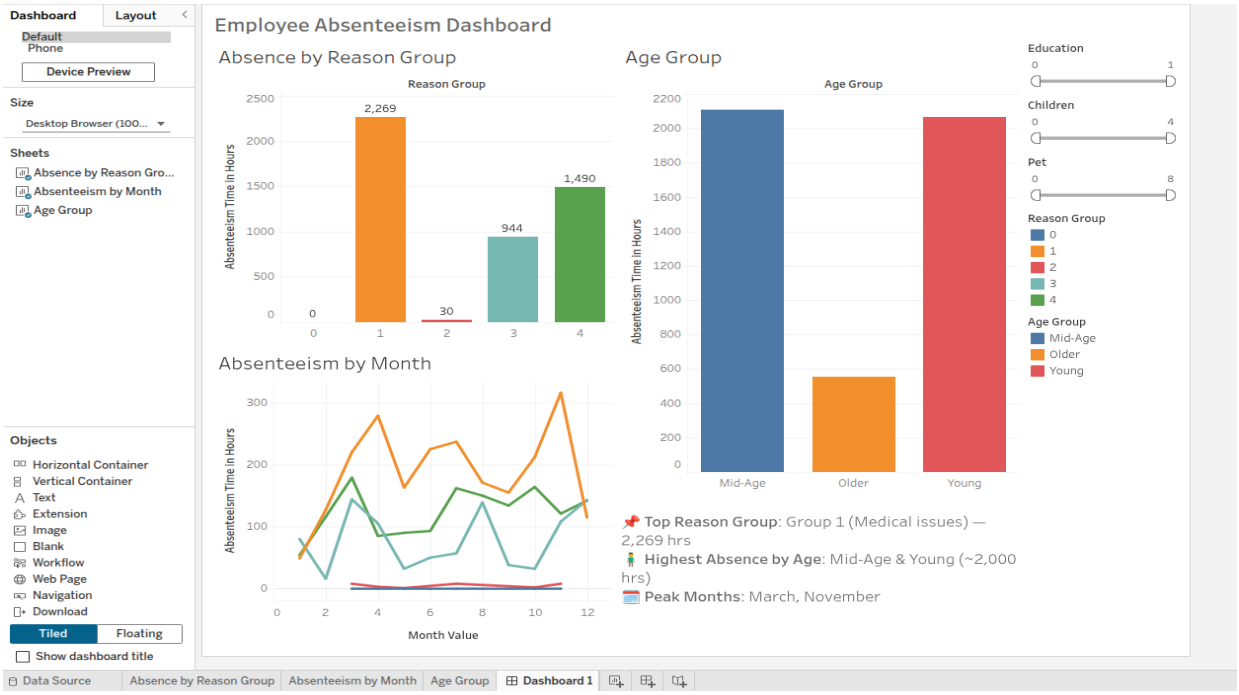

Figure 5: Tableau dashboard showing absenteeism trends by reason group, month, and age. Filters allow exploration by demographic features like education, pets, and children

6. KEY FINDINGS

This project revealed several important insights about absenteeism in the company:

- **Medical-related issues (Reason Group 1)** were the most common cause of absenteeism.
- **Absences were higher in March and November**, suggesting possible seasonal health trends.
- **Employees aged 30–45** tended to miss more hours than others.
- **Random Forest outperformed Linear Regression** and was the best model for this dataset.
- **Workload and education** were useful predictors — those with heavier workloads and lower education levels were more likely to miss more hours.

These insights can help HR departments design targeted interventions, such as health programs during peak absence months or support for employees in high-risk groups.

7. REFLECTION

This project helped me apply the skills I learned in the Google Data Analytics Certificate in a real world context. I practiced data preprocessing, visualization, feature engineering, and predictive modeling, all within a structured workflow. I also gained valuable experience building an interactive dashboard with Tableau, which was a new tool for me at the time.

One of the most important lessons was understanding how different types of models behave when applied to real business data. I discovered that even if a model is statistically correct, it might not perform well without proper data transformation or thoughtful feature selection.

Most importantly, I now feel more confident in my ability to communicate insights through clear visuals and written analysis. This project showed me the value of not just building models, but also turning them into tools that decision makers can understand and act on.

8. CONCLUSION

This project successfully demonstrates how data analytics and machine learning can be applied to a real-world business problem: employee absenteeism. From exploring patterns in HR records to predicting future absenteeism hours using multiple regression models, the process gave me hands on experience with tools like pandas, seaborn, scikit learn, and XGBoost. In addition to coding and analysis, building an interactive Tableau dashboard taught me how to turn numbers into clear visuals that stakeholders can use to make informed decisions.

The biggest takeaway was understanding how different models behave on real world data. While Linear Regression was simple, it performed poorly, which led me to explore and appreciate more powerful models like Random Forest and XGBoost. I also learned how to engineer features, choose the right transformations, and clean data in a way that improves model performance.

Beyond the technical skills, this project strengthened my ability to communicate insights through storytelling, both visually and in writing. I now feel more confident designing complete analytics projects and presenting findings to business audiences. I look forward to continuing this journey and applying these skills in professional settings.