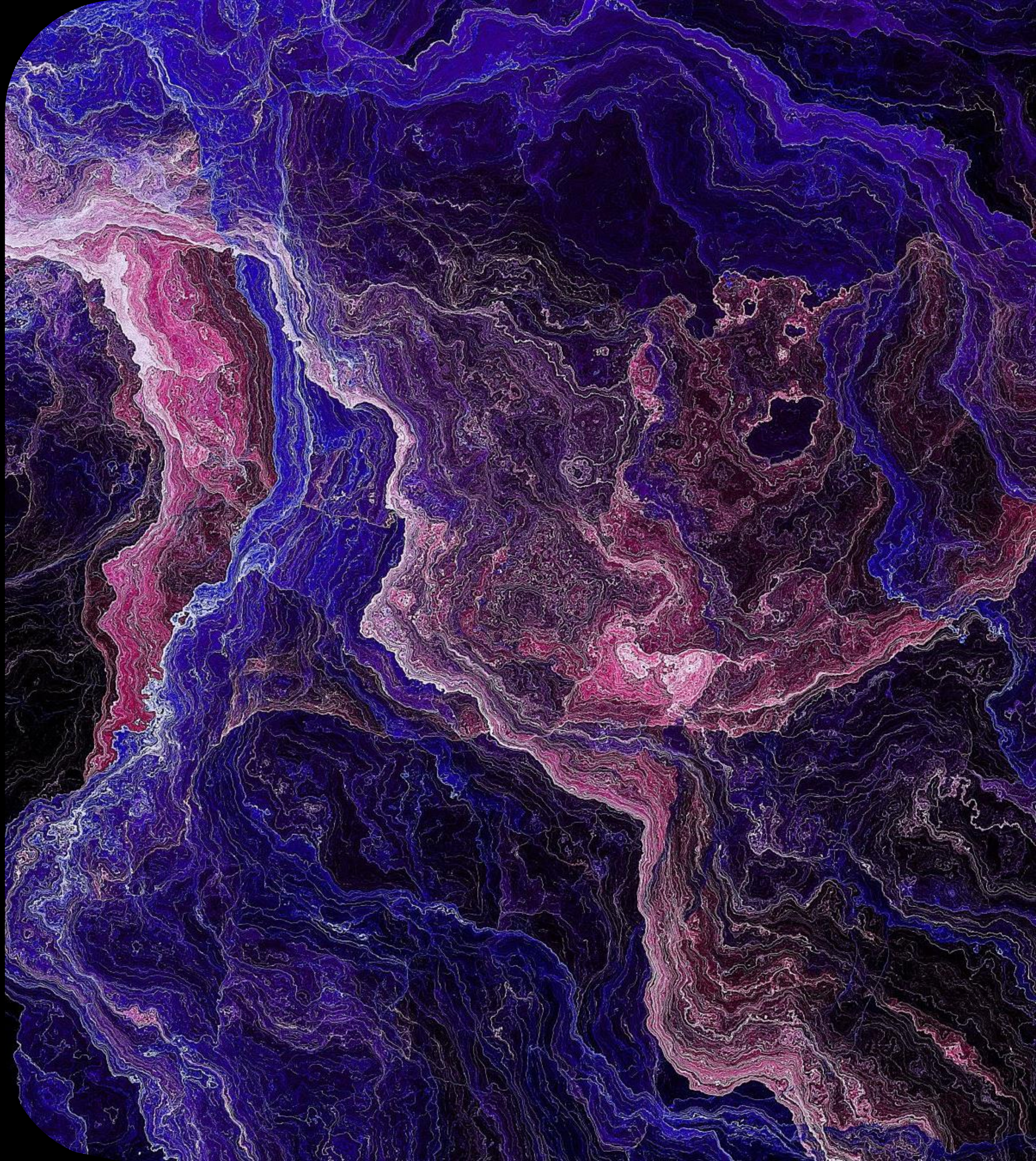
Binary Logistic Regression on Employee Data

Predicting Employee Burnout

Western Governors University

D606 Task 3: Data Science Capstone Presentation

Student: Nichole Gonzales



About me...

- Born and raised in Galveston County, Texas.
- Bachelor of Science in Criminal Justice from Sam Houston State University.
- Data Analyst experience for executive level decision making at an organization.
- I enjoy spending time with my family, watching movies, and going to the gym.

Project Overview

- Study Objective – Problem and Hypothesis
- Data Analysis Process
- Findings
- Understanding Limitations of Techniques and Tools
- Proposed Action
- Benefits of the Study

Study Objective: Problem and Hypothesis

- Problem with Employee Burnout
  - + In 2024, half (52%) of employees reported feeling symptoms of Burnout (NAMI, 2024).
  - + In 2025, Forbes reported that two-thirds (66%) of American employees experienced Burnout.
- Research Question
  - + Can a logistic regression model classify employees as high or low risk of Burnout using workplace data?
- Hypothesis
  - + Null Hypothesis: Burnout risk cannot be predicted using this data
  - + Alternative Hypothesis: Burnout risk can be predicted with accuracy greater than 70%

Data Analysis Process Summary

- Data Collection
- Data Preparation
  - Exploratory Data Analysis
  - Data Cleaning
  - Variable Optimization – Backward Stepwise Elimination
- Final Analysis – Binary Logistic Regression

Data Analysis Process: Data Collection

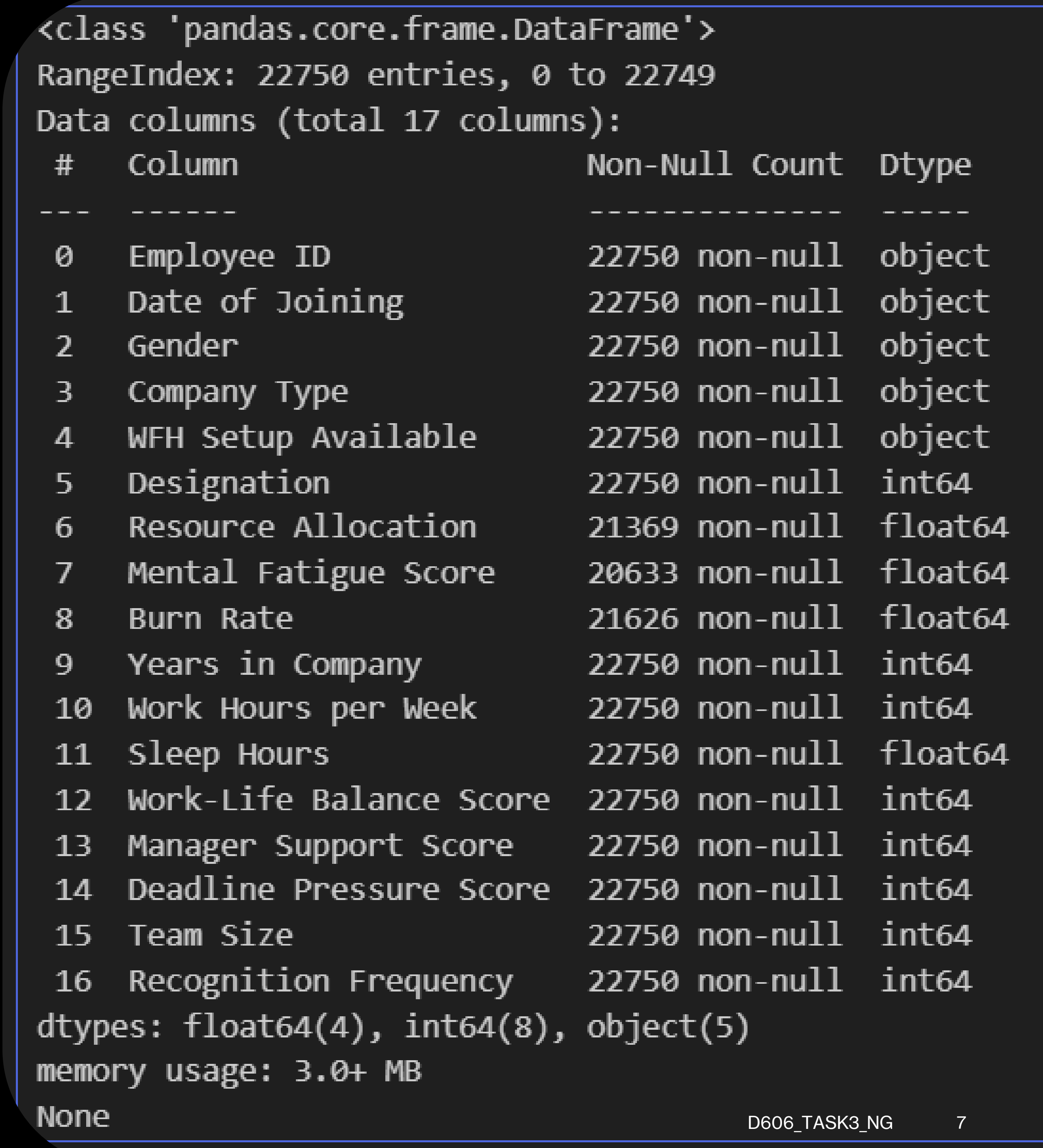
- Harvard Dataverse dataset "Burnout Among Corporate Employees" (Devkar, 2025).

- Number of Observations: 22,750
- Number of Variables: 17
  - Includes numeric and categorical data types

- Structure supports the application of statistical classification techniques.

Data Analysis Process: Exploratory Data Analysis

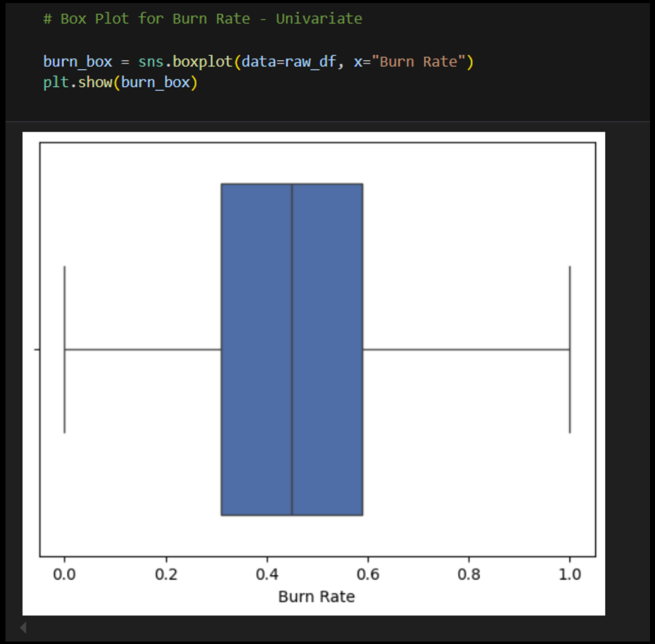
- **Understanding Data Structure, Data Types, and Completeness**
  - How many records?
  - Data Types
  - Variable Names
  - How many gaps (nulls) in dataset?
  - Descriptive Statistics on Numeric Data
    - Count, Mean, Min, Max, Each Quartile



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 22750 entries, 0 to 22749
Data columns (total 17 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Employee ID               22750 non-null  object
 1   Date of Joining           22750 non-null  object
 2   Gender                    22750 non-null  object
 3   Company Type              22750 non-null  object
 4   WFH Setup Available       22750 non-null  object
 5   Designation               22750 non-null  int64
 6   Resource Allocation       21369 non-null  float64
 7   Mental Fatigue Score      20633 non-null  float64
 8   Burn Rate                 21626 non-null  float64
 9   Years in Company          22750 non-null  int64
 10  Work Hours per Week       22750 non-null  int64
 11  Sleep Hours               22750 non-null  float64
 12  Work-Life Balance Score   22750 non-null  int64
 13  Manager Support Score     22750 non-null  int64
 14  Deadline Pressure Score   22750 non-null  int64
 15  Team Size                 22750 non-null  int64
 16  Recognition Frequency     22750 non-null  int64
dtypes: float64(4), int64(8), object(5)
memory usage: 3.0+ MB
None
```

Data Analysis Process: Clean and Prepare the Dataset

- Treat missing values: Drop or Impute
  + Dropping records with null “Burn Rate”
  + Impute Median for “Resource Allocation” and “Mental Fatigue Score”
- Convert categorical variables to numeric form
- Create and Define Target Variable to identify Employees at High Risk of Burnout
  + Use top quartile (>= 0.59) Burn Rate for High Risk, Remaining will be Low Risk



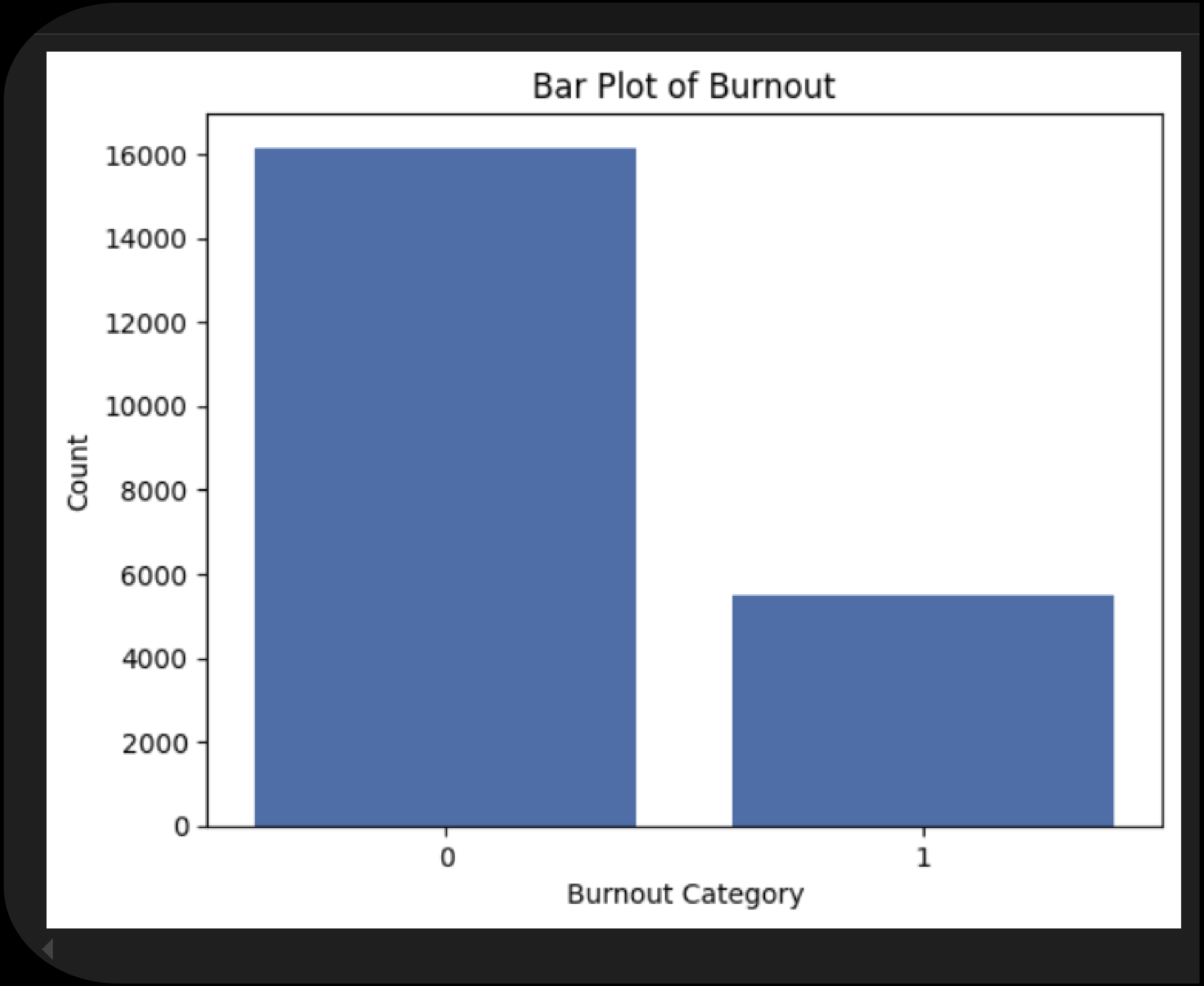
```
# Box Plot for Burn Rate - Univariate

burn_box = sns.boxplot(data=raw_df, x="Burn Rate")
plt.show(burn_box)
```

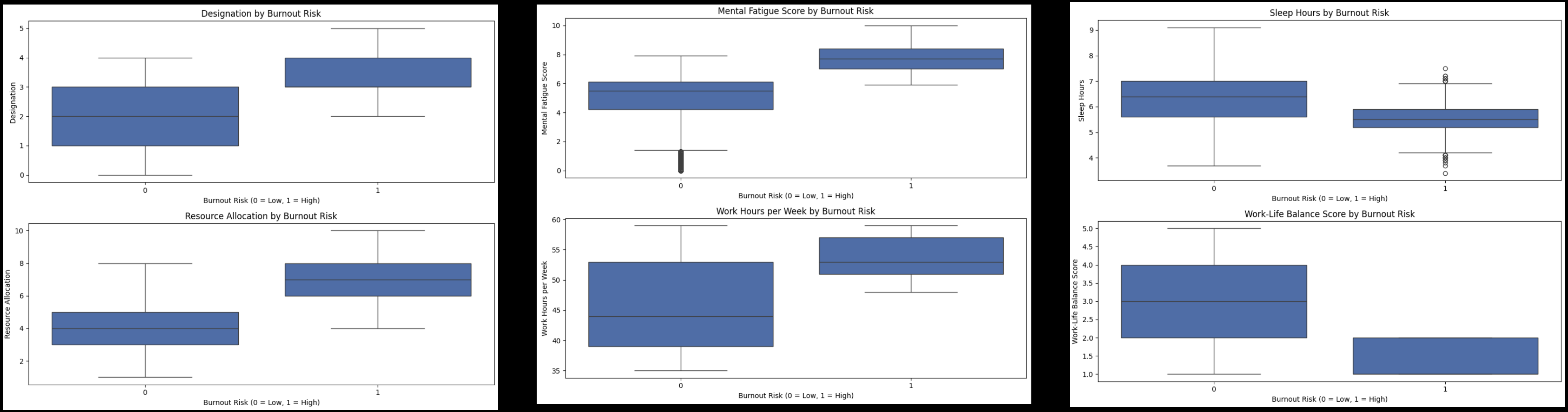
Image of Burn Rate boxplot from analysis.

Data Analysis Process: Understanding Target Variable

- Variable: Burnout_Risk
- Number of Employees Identified as High-Risk Burnout: 5,480
- Number of Employees Identified as Low-Risk Burnout: 16,146

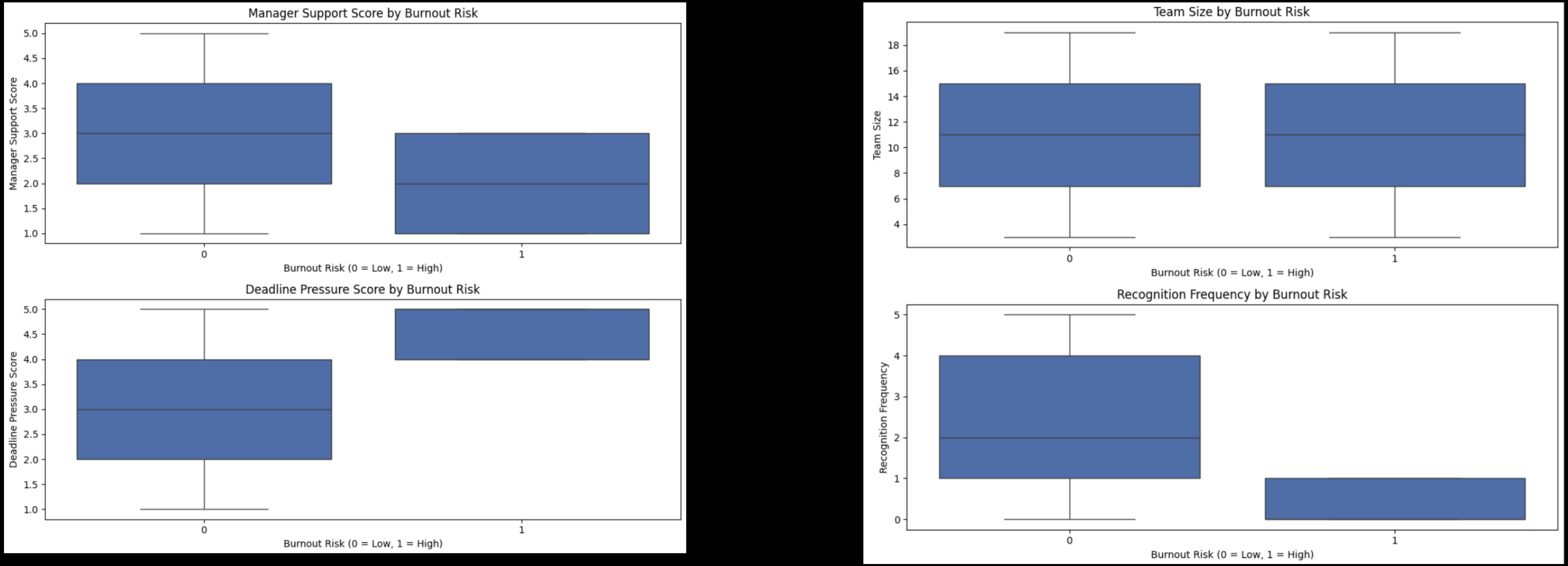


Data Analysis Process: Understanding Target Variable “Burnout Risk”



Boxplot of Variables Split on Burnout Risk Category

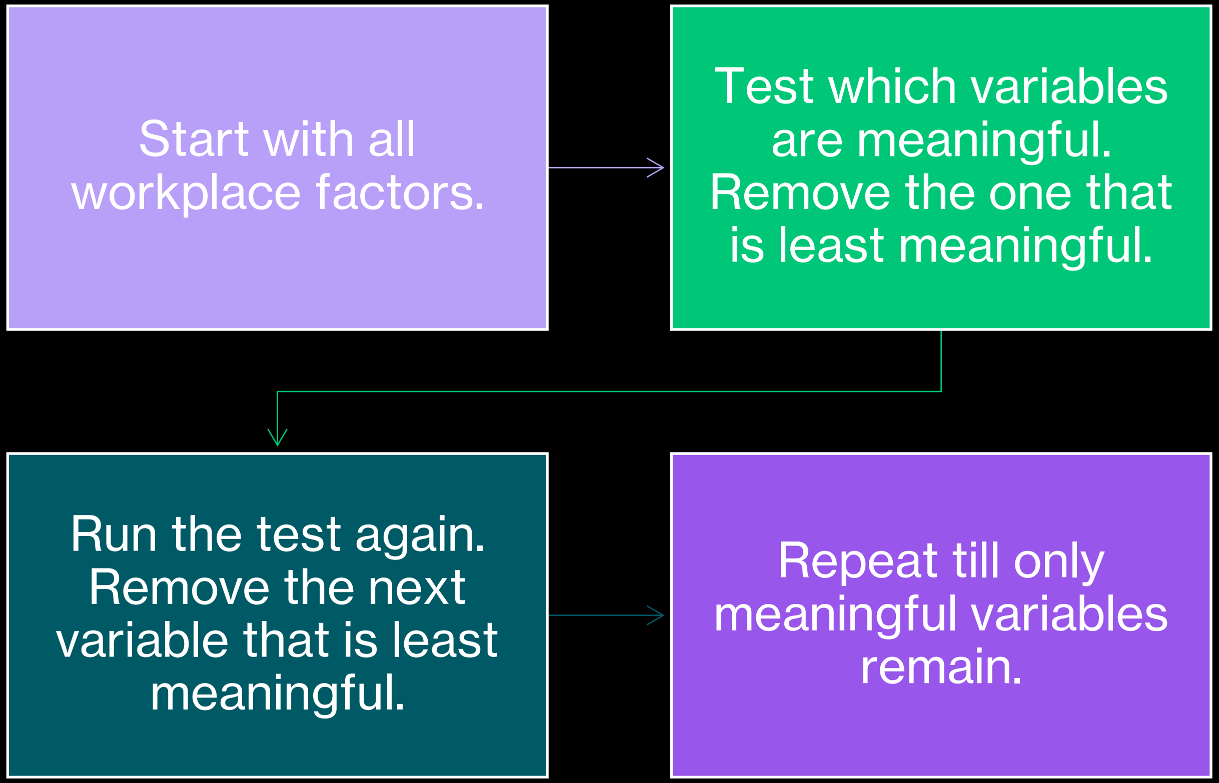
Data Analysis Process:
Understanding Target Variable “Burnout Risk”



Boxplot of Variables Split on Burnout Risk Category

Data Analysis Process: Variable Optimization

Backward Stepwise Elimination



Start with all workplace factors.

Test which variables are meaningful. Remove the one that is least meaningful.

Run the test again. Remove the next variable that is least meaningful.

Repeat till only meaningful variables remain.

Retained Variables to Use in Model

Before Optimization
Gender
Company Type
WFH Setup Available
Designation
Resource Allocation
Mental Fatigue Score
Years in Company
Work Hours per Week
Sleep Hours
Work-Life Balance Score
Manager Support Score
Deadline Pressure Score
Team Size
Recognition Frequency

After Optimization
Gender
~~Company Type~~
WFH Setup Available
Designation
Resource Allocation
Mental Fatigue Score
~~Years in Company~~
Work Hours per Week
~~Sleep Hours~~
Work-Life Balance Score
~~Manager Support Score~~
~~Deadline Pressure Score~~
~~Team Size~~
Recognition Frequency

Logistic Regression

- Used for Classification
- A method that uses data patterns to estimate the probability that something can be put into one of two categories, such as High or Low Burnout risk.

- Steps:
  - + Split dataset into training and testing data
  - + Perform Logistic Regression
  - + Measure Accuracy of Model

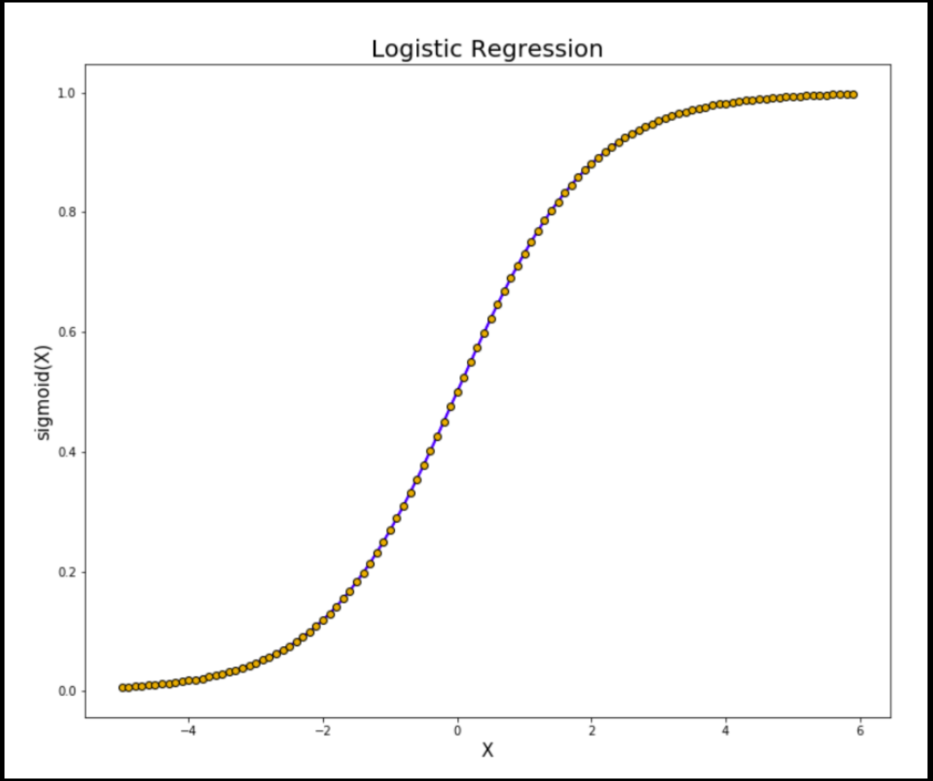


Image by Logistic Regression Explained (Thorn, 2020).

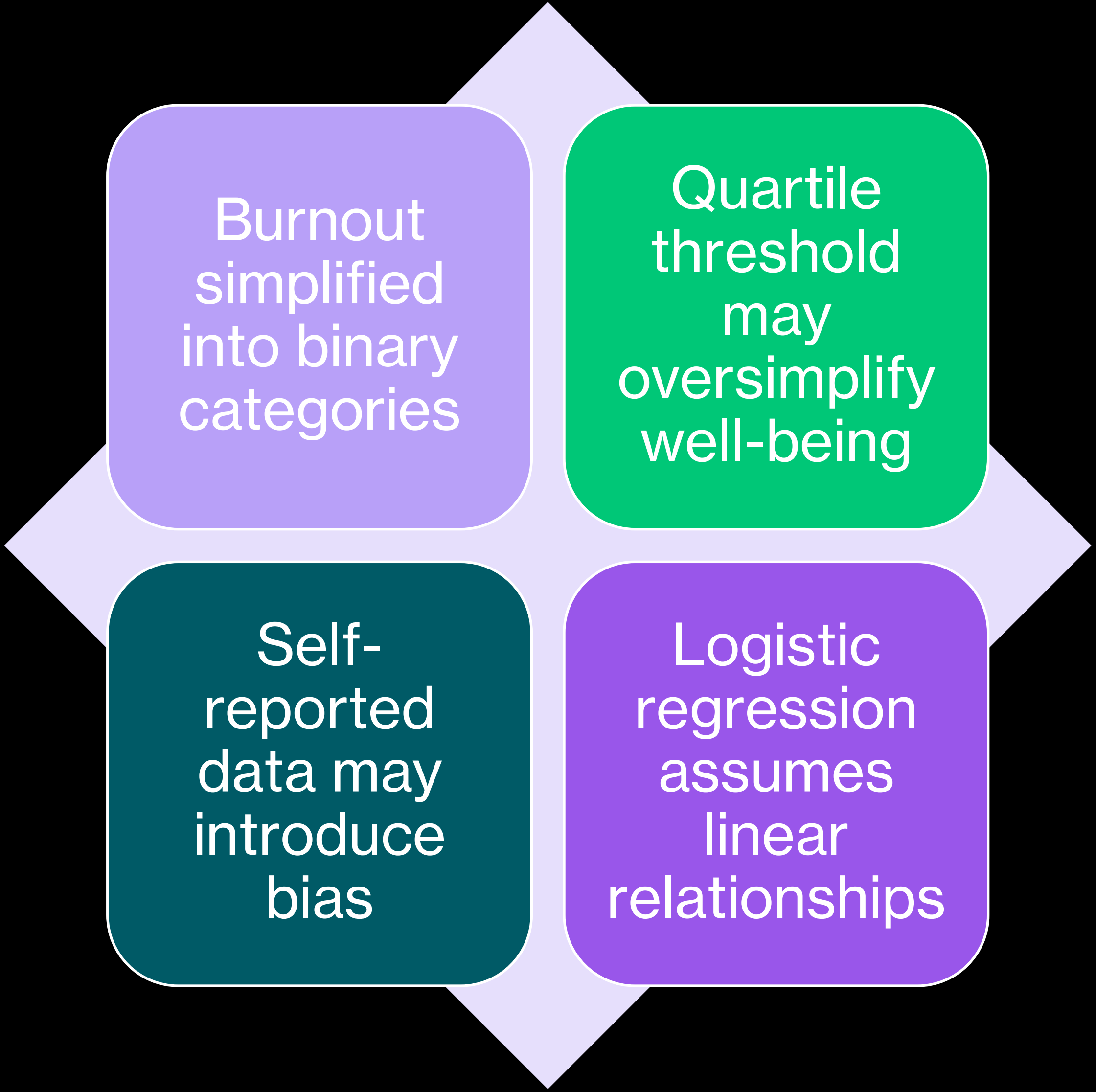
Analysis Findings

Model Accuracy: 89.65%

Accept Alternative Hypothesis: Burnout risk can be predicted with accuracy greater than 70%.

Mental Fatigue Score was the most dominant influencer of Burnout.

Limitations of Techniques and Tools



Burnout simplified into binary categories

Quartile threshold may oversimplify well-being

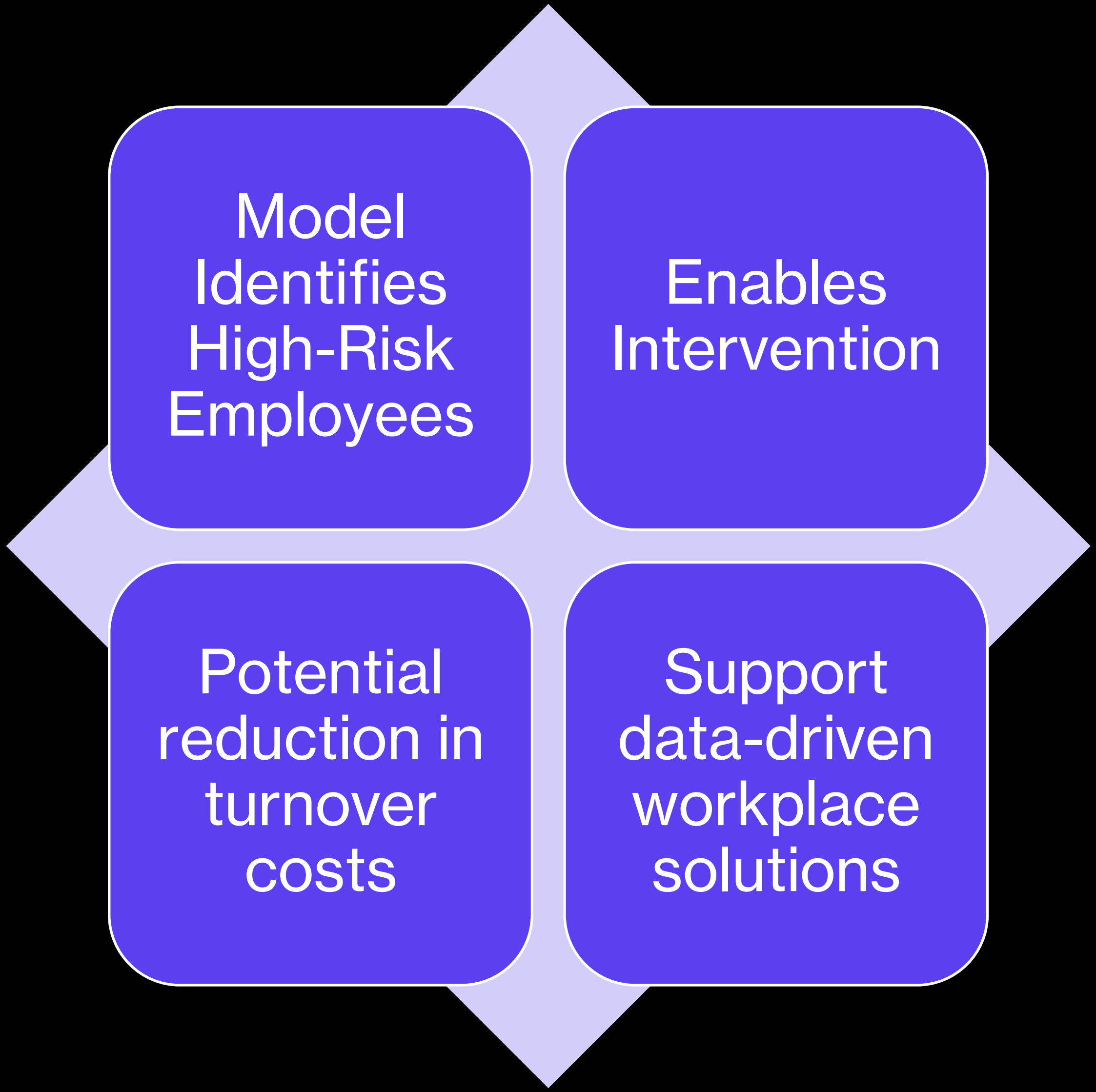
Self-reported data may introduce bias

Logistic regression assumes linear relationships

Proposed Action

- Identify employees at high-risk of burnout.
- Implement a form of intervention.
- Focus on reducing mental fatigue.
- Support proactive wellness strategies.

Expected Benefits



Model Identifies High-Risk Employees

Enables Intervention

Potential reduction in turnover costs

Support data-driven workplace solutions

References

Devkar, P. (2025). *Burnout Among Corporate Employees*. Harvard Dataverse. Dataset retrieved December 10, 2025, from https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/VG6KQD

National Alliance on Mental Illness (NAMI). (2024). *The 2024 NAMI Workplace Mental Health Poll*. NAMI. https://www.nami.org/support-education/publications-reports/survey-reports/the-2024-nami-workplace-mental-health-poll/

Robinson, B. (2025). *Job Burnout at 66% in 2025, New Study Shows.* Forbes. https://www.forbes.com/sites/bryanrobinson/2025/02/08/job-burnout-at-66-in-2025-new-study-shows/

Thorn, J. (2020). *Logistic Regression Explained.* Towards Data Science. https://towardsdatascience.com/logistic-regression-explained-9ee73cede081/

Thank You!