DAT-430 3-2:

Establishing a Baseline and Hypothesizing Outcomes

Chloe White

Professor Kotra

January 26, 2025

**Overview of the HR Training Dataset and Objectives**

The HR department is interested in evaluating the impact of training on employee performance and job satisfaction. The goal of this analysis is to establish a baseline for understanding how training hours influence these outcomes. The hypothesis for this analysis is that increasing training hours will lead to higher job satisfaction and better performance ratings among employees. To test this hypothesis, the HR Training dataset was analyzed using Python, focusing on regression modeling and visualization.

**Data Import and Exploration**

To begin, the necessary libraries for data analysis and modeling were imported. The HR Training dataset was loaded into a pandas DataFrame and displayed using the head() function to preview the first five rows. This dataset includes variables such as Age, BusinessTravel, Department, JobSatisfaction, PerformanceRating, and more.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

**Data Preparation**

A subset of the dataset was created containing only the columns needed for the analysis: training, JobSatisfaction, and PerformanceRating. These variables represent the number of training hours, employee satisfaction levels, and performance ratings.

After creating the sub-dataset, a check for missing values was performed using the .isnull().sum() function. The results confirmed that no missing values were present in the selected columns, ensuring that the data was complete and suitable for further analysis.

A screenshot of a computer

Description automatically generated

**Check for Data Integrity**

The dataset was summarized using the describe() function to ensure data integrity. The summary statistics showed that training hours ranged from 0 to 6, indicating variability in training participation across employees. The inclusion of 0 in the range for training hours indicates that some employees did not receive any training. Job satisfaction scores spanned from 1 to 5, aligning with the expected satisfaction scale, while performance ratings also ranged from 1 to 5, matching the anticipated rating scale. These results confirmed that the data was clean, fell within expected ranges, and showed no extreme outliers that could distort the analysis.

A screenshot of a computer

Description automatically generated

**Regression Analysis: Training and Job Satisfaction**

A linear regression model was applied to evaluate the relationship between training hours (X) and job satisfaction (Y). The model was implemented using the LinearRegression class from sklearn, and the coefficient of determination (R-squared value) was calculated (Stojiljkovic, 2019).

The R-squared value for this analysis was 0.86, indicating that 86% of the variation in job satisfaction can be explained by the number of training hours. This strong positive relationship suggests that increased training hours are likely to improve employee satisfaction. A scatter plot with a regression line overlay was generated to visualize the relationship. The plot clearly shows an upward trend, further supporting the hypothesis.

A screenshot of a computer program

Description automatically generated

A screen shot of a graph

Description automatically generated

**Regression Analysis: Training and Performance Rating**

The same method was applied to assess the relationship between training hours (X) and performance ratings (Y). The R-squared value for this analysis was 0.82, indicating that 82% of the variation in performance ratings can be explained by training hours. This demonstrates a strong positive relationship between training and employee performance. A scatter plot with a regression line was generated to visualize this relationship. The visualization shows a clear upward trend, consistent with the regression model’s results.

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screen shot of a graph

Description automatically generated

**Key Insights and Implications for HR**

The analysis demonstrates a strong positive relationship between training hours and both job satisfaction and performance ratings. The high R-squared values (0.86 and 0.82) suggest that increasing training hours is likely to enhance employee satisfaction and performance. These findings provide HR with a reliable baseline for predicting how additional training initiatives can benefit employees and guide future decision-making.

**References**

Cantillo, J. (2024, June 4). *Why Data Transformation Needs a Baseline* . MaterialPlus. https://www.materialplus.io/perspectives/why-data-transformation-needs-a-baseline

Stojiljkovic, M. (2019, April 15). *Linear Regression in Python*. Realpython.com; Real Python. https://realpython.com/linear-regression-in-python/