DAE Project Two

Sayfullah Jumoorty (2430888)

Muhammed Muaaz Dawood (2425639)

Mujammil Mohsin Gulam Sakhidas (2436109)

May 2024

Question 1

The dataset provided contains information on data scientists who filled out employee satisfaction surveys, along with whether they remained at the company three months later. The variables included in the dataset are: *Stay*, indicating whether or not the data scientist is still with the company three months later; *Pay*, the monthly salary in dollars; *Estimated Happiness*, a score derived from employees' reported happiness and their comments on what they'd like to see changed; and *Performance*, the results of managers' performance reviews.

First, we checked for any missing values in the dataset and found none, ensuring that we have a complete dataset for analysis. Next, we computed summary statistics to understand the central tendency and spread of the numerical variables, including mean, median, standard deviation, and range. This step helps us to get an initial understanding of the data distribution.

We visualized the distribution of the *Pay*, *Estimated Happiness*, and *Performance* variables using histograms. The histogram for *Pay* revealed a skewed distribution, indicating that most salaries are clustered around lower values with a few outliers. The *Estimated Happiness* score showed a multimodal distribution, and the *Performance* scores were roughly normally distributed.

To understand the relationships between the variables, we computed the correlation matrix. Interestingly, the correlation between *Performance* and *Stay* was slightly negative. This finding might seem counterintuitive since we would expect higher performance to correlate with higher retention. This anomaly could be intentionally included in the dataset to provoke critical thinking and deeper analysis.

To further investigate, we visualized the relationship between *Performance* and *Stay* using a box plot. The box plot showed that the *Performance* scores are similar for those who stayed and those who left, with a slight negative trend. Additionally, there was an unexpected third category in *Stay*, which we identified as an error. We corrected this by replacing the value 2 with 1 in the *Stay* column and rechecked the correlations.

After this correction, the correlation matrix was recalculated, showing a more intuitive relationship between variables. The data exploration revealed key insights and potential anomalies in the dataset. The negative correlation between *Performance* and *Stay* suggests that high-performing employees might still leave the company, possibly due to other factors not captured in the dataset, such as job satisfaction, work-life balance, or career opportunities. This finding emphasizes the importance of considering multiple dimensions when analyzing employee retention.