Data Cleanup Report

Seagate Employee Churn Project

|  |  |
| --- | --- |
| *Submitted by:* | |
| Charles Biddle Porter  [charles.biddleporter@colorado.edu](mailto:charles.biddleporter@colorado.edu) | Duc Hoang  [duc.hoang@colorado.edu](mailto:duc.hoang@colorado.edu) |
| Harshit Gole  [harshit.gole@colorado.edu](mailto:harshit.gole@colorado.edu) | Kirti Vatsh  [kirti.vatsh@colorado.edu](mailto:kirti.vatsh@colorado.edu) |
| Manan Dhanteja  [Manan.dhanteja@colorado.edu](mailto:Manan.dhanteja@colorado.edu) |  |

Table of Contents

[Report Objective 2](#_Toc161773337)

[Data Description 3](#_Toc161773338)

[conclusion 12](#_Toc161773339)

# **Data Cleaning overview**

Data cleaning is a critical step in the data analysis process, ensuring the integrity and usability of the dataset. It involves scrutinizing the data for inaccuracies, inconsistencies, and redundancies. The process typically begins with removing duplicate records to avoid skewed results. Next, we address null or missing values, which may involve imputation techniques or, in some cases, removing affected records if they're not critical to the analysis. We also perform quick checks to validate the logical consistency of the data, such as ensuring that active employees do not have a termination date. Another critical task is normalizing values, particularly with variables like salary, which may need to be standardized to a common currency.

# data cleaning pipeline

# **Data cleaning phases**

1. removing unnecessary rows

Our team took targeted steps during the data cleaning to enhance the dataset's predictiveness and accuracy. To address the unique and unpredictable nature of Position Elimination events, we removed rows featuring "Position Elimination/ RIF Involuntary" or "Position Elimination/ RIF Voluntary" in the Termination Reasons column. These events are contingent on the company's strategic decisions and occur sporadically, making them unsuitable for inclusion in a model to forecast future trends.

Additionally, we scrutinized the Compa Ratio Column, removing any entries with Compa Ratio values equal to or greater than two and those with a value of 0. Such extreme values are atypical and could heavily skew the dataset. By eliminating these anomalies, we aim to stabilize the dataset's skewness, creating a more balanced and representative data collection for subsequent analysis. This step is crucial in ensuring that outliers do not distort our predictive models and can accurately reflect the typical compensation structure within the company.

1. removing null values

In the data cleaning phase, we tackled the issue of null values with precision, specifically within the Pay Level column. We recognized that only four rows contain null values in the Pay Level column—approximately 0.015% of our total observations. We then decided that filtering out these rows would be the most effective. This deletion technique removes irrelevant data without significantly impacting the overall dataset. The scarcity of these null instances means their removal will not compromise the statistical integrity of our analyses. Ensuring that each entry has a corresponding Pay Level is vital for any subsequent compensation analysis, as it forms a core metric for our project's objectives.

1. normalizing values

In the normalization section of our data cleaning phase, we note that the Base Pay salaries still need to be standardized due to the absence of currency conversion rates, which the stakeholders still need to provide. Once we can normalize Base Pay values to a common currency, it will likely result in eliminating current outliers. This normalization will bring the mean Base Pay closer to the median value, which we approximate around $140,000, and considerably reduce the standard deviation. This adjustment is expected to mitigate the effect of currency fluctuations and regional economic disparities, yielding a dataset that more accurately reflects the compensation levels across the global workforce. The normalization process is crucial as it will enhance the comparability and relevance of the Base Pay data, making it a more reliable metric for analysis and decision-making.

1. feature engineering

The variables sampled here were used to generate a correlation heatmap, and the methodology to compare categorical and numerical variables objectively is described below.

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Variable Type** | **Details** |
| Job Title | Categorical | Polytomous and Nominal |
| Compa Ratio | Numerical | - |
| Work Country | Categorical | Polytomous and Nominal |
| Gender | Categorical | Polytomous and Nominal |
| Tenure Bucket | Categorical | Polytomous and Ordinal |
| Generation | Categorical | Polytomous and Nominal |
| Base Pay Midpoint Annualized USD | Numerical | - |
| Cost to Replace Employee Multiplier | Numerical | - |

*Table 1: Variables description*

At first, the data was separated into categorical and numerical data. The categorical data is further represented as polytomous (meaning it has three or more categories), with its subtype being nominal, meaning unordered, or ordinal, meaning ordered data. Usually, the correlation indices would be computed with different methods based on the relationships between the different data types; however, using one-hot encoding, the categorical variables were represented by binary indicator variables, effectively making them numerical. This approach makes all further comparative data calculations in machine learning and predictive analytics quantitative.

A graph with a line and a chart

Description automatically generated with medium confidence

*Figure 2: Correlation Heatmap*

1. data cleaning visualization

Comparing the histograms of Compa Ratio values before and after data cleaning provides a clear visual representation of the impact of our cleaning procedures. Initially, the histogram reveals a distribution with a long tail, indicating the presence of outliers with Compa Ratio values that extend far beyond the typical range. These outliers create a highly skewed distribution. After cleaning, the histogram exhibits a more bell-shaped distribution, indicative of a normal distribution. This transformation results from removing extreme Compa Ratio values, including those equal to or greater than two and those at 0. By eliminating these outliers, we have realigned the dataset better to reflect the central tendency of the Compa Ratios.

*A screenshot of a computer screen

Description automatically generated*

*Figure 3: Termination type before and after data cleaning*

The scatter plot comparisons before and after data cleaning show a pronounced difference in the distribution and range of the Compa Ratio values. Prior to cleaning, the Compa Ratio values are spread out across a broad range, extending from 0 to beyond 15, with a concentration of points at the lower end, indicating a skewed distribution with some extreme values. In contrast, the post-cleaning scatter plot presents a much more concentrated range of Compa Ratio values, mostly clustered between 0.5 and 1.5. This consolidation suggests the removal of outliers and extreme values from the data set. The Cost to Replace Employee metric appears to remain consistently distributed in both scatterplots, which implies that the cleaning process primarily affected the Compa Ratio variable. The presence of points across a similar range on the y-axis in both plots indicates that the cleaning did not significantly impact the variability in the cost to replace an employee.

*A comparison of a graph

Description automatically generated*

*Figure 4: Scatterplot of Compa ratio vs Cost before and after data cleaning*

# **conclusion**

In conclusion, data cleaning is fundamental to the analytical process. It enhances data quality and accuracy by correcting errors, reconciling inconsistencies, and filling in missing values. This report presents a comprehensive data-cleaning pipeline, offering a clear guideline for the process, enhancing transparency in the analysis, and articulating the rationale behind each method.

The report has detailed the critical steps undertaken for data cleaning, including deleting extraneous rows, outlier removal, handling of null values, and feature engineering. Decisions regarding outliers and specific value removal were tailored to the project's aim of forecasting future trends. The planned normalization of the data is also expected to improve data reliability, serving as a more dependable metric for analysis and decision-making. Moreover, feature engineering has been employed to convert categorical data into a binary format, enabling their inclusion in regression analysis and enhancing the model's predictive power.

Having completed the data cleaning phase, we now have a dataset that bolsters decision-making, yields more precise insights during the analysis phase, and sets the stage for a more accurate machine learning model. This cleaner data positions us to identify trends and insights critical to informed business decisions more precisely.

Overall, the report offers a complete overview of the data cleaning and preprocessing steps taken, making the data valuable and significant for statistical analysis, aiming to enhance the quality of insights and the ability to predict future trends.