**DATA CLEANUP REPORT**

**A green and black logo

Description automatically generated**

**People Analytics at Seagate: A Strategic Approach**

**to Predict Voluntary Churn and Optimize the Hiring Process**

**Master of Science in Business Analytics**

**Course: Experiential Projects**

**S24 – 004 – Group 4 – Project 2**

**Prof. Noah Zikmund**

**Group Members**

**Anubhav Dubey** (anubhav.dubey@colorado.edu)

**Amandeep Bhardwaj**

(amandeep.bhardwaj@colorado.edu)

**Sandeep Reddy Modugu**

(sandeep.modugu@colorado.edu)

**Sai Arvind Atluri** (saiarvind.atluri@colorado.edu)

Table of Contents

[Data Overview 3](#_Toc162292052)

[Data Cleaning and Preparation Process 4](#_Toc162292053)

[Data Cleaning Phases 5](#_Toc162292054)

[Removing Unnecessary Columns 5](#_Toc162292055)

[Removing Unnecessary Rows 5](#_Toc162292056)

[Feature Engineering 6](#_Toc162292057)

[End Results 7](#_Toc162292058)

[Descriptive Statistics 8](#_Toc162292059)

[Data Visualizations 10](#_Toc162292060)

[Outliers vs. Non-Outliers of Numerical Variables 10](#_Toc162292061)

[Histograms 11](#_Toc162292062)

[**Employee Tenure** 11](#_Toc162292063)

[**Compa Ratio** 12](#_Toc162292064)

[**Base Pay** 13](#_Toc162292065)

[Scatter Plots 14](#_Toc162292066)

[**Tenure vs. Base Pay** 14](#_Toc162292067)

[**Compa Ratio vs. Base Pay** 15](#_Toc162292068)

[Correlation Heat Map 16](#_Toc162292069)

[Implications of observed changes 16](#_Toc162292070)

[Conclusion 17](#_Toc162292071)

# **Data Overview**

Navigating through the integrity and accuracy of data is a fundamental aspect of any project that aims to extract meaningful insights from available information. While this process is time-consuming, but it is essential for sifting through the diverse and often complex datasets, typically sourced from various systems. These datasets frequently exhibit a range of issues, including errors, inconsistencies, and missing values, which can significantly distort the analytical outcomes. The iterative nature of data cleaning requires continuous refinement to meet the analysis-specific demands and achieve the necessary data quality. Hence, a comprehensive data cleaning regimen is crucial for the project team to derive reliable insights and make well-informed business decisions.

As part of the Seagate HR data analysis project, we have undertaken meticulous steps to structure the data in a manner conducive to insightful analysis and decision-making. This section delineates the sophisticated data preparation and cleaning methodologies we employed, while grounded in our collective domain expertise and extensive research.

Initiating the process, we evaluated the dataset to identify and eliminate irrelevant rows and columns. This pruning was based on detailed assessment of the data's relevance on analytical goals and the feasibility of imputation for missing information. Through leveraging our analytical acumen, we were able to tailor the dataset to align precisely with the project’s objectives, thereby notably enhancing the dataset’s quality and relevance.

Addressing missing values was a pivotal aspect of our data preparation, wherein we employed feature engineering to judiciously estimate values. This methodical approach ensured that each data element not only enriched our analysis but also solidified the conclusions of our study. Whether dealing with missing values in numerical fields, particularly those related to financial metrics, or confronting columns with extensive missing data, we consistently applied feature engineering to derive accurate estimates. By imputing these missing values using both categorical and numerical features, we preserved the dataset’s analytical integrity and financial depth while ensuring that no decision to remove data compromised the overall analysis. Consequently, this strategy fortified our dataset, enhancing its ability to contribute significantly to the robustness and comprehensiveness of our analysis. This approach corroborated that each piece of data principally contributed to our analysis, strengthening the validity and breadth of our conclusions.

Moreover, we labeled columns with strikingly missing yet analytically valuable data as 'Not Specified / Not Terminated,' enhancing the clarity and interpretability of our dataset. This labeling not only preserved the structural integrity of the dataset but also facilitated a more nuanced and contextual analysis.

Consequently, our data cleaning and preparation regimen, marked by strategic eliminations, exhaustive management of missing values, and targeted feature engineering, has established a robust foundation for the next stages of the Seagate HR data analysis project. Subsequent sections will outline the precise procedures we implemented, highlighting our dedication to a rigorous and data-driven analytical approach.

## **Data Cleaning and Preparation Process**

**Data Profiling and Preliminary Exploration:** To identify any issues with the quality of the data, we examined the structure of the Seagate HR dataset and the findings from this preliminary investigation were meticulously recorded, providing a foundational background for the subsequent cleaning processes.

**Addressing Missing Data:** Our examination revealed missing data across several columns. To address these gaps, we employed a combination of statistical techniques and domain knowledge, varying from simple deletion of inconsequential data points to complex imputation of values with potential analytical significance.

**Duplicate Identification and Resolution:** Duplicate records, which could skew the analysis were identified, and removed to ensure the uniqueness of each data point. The process and its implications were transparently documented to maintain integrity.

**Normalization and Data Transformation:** We transformed the data, encoding categorical variables and normalizing skewed distributions, to ensure comparability and readiness for advanced analytical techniques.

**Feature Engineering and Enrichment:** Through feature engineering, we expanded our dataset and introduced new variables, specifically in relation to financial metrics and employee involvement, to enhance the analysis and provide deeper insights.

**Outlier Detection and Treatment:** Outliers and anomalies were identified and addressed to ensure the accuracy of our data. Methods such as thresholding and data curation were used to manage these abnormalities.

**Ensuring Consistency and Reliability:** A rigorous quality assurance process was implemented. This phase ensured that the data retained its original integrity and remained a dependable and consistent foundation for our post-cleaning analyses.

**Documentation of the Final Dataset:** To conclude the data cleaning phase, we diligently documented the final state of the dataset, including the number of variables and observations retained or removed, as well as any new variables added during the cleaning process.

**Data Cleaning Validation:** Final validation checks were conducted to verify the detailed completion of all data cleaning processes and to confirm the alignment of the cleaned data with the project’s objectives and standards.

# **Data Cleaning Phases**

## **Removing Unnecessary Columns**

We removed any irrelevant or unnecessary columns from the Seagate HR dataset to make it more streamlined and insightful for further research. Columns such as Job Code and Anonymized ID were eliminated, as they did not provide any analytical value for predictive modeling or aggregate assessment. The Job Group was also omitted due to its data largely overlapping with other job-related characteristics. Furthermore, Work Region was deemed unnecessary because Work Location and Work Country adequately represent the geographic information required for our research. Given the more detailed Tenure data, Tenure Bucket was considered redundant for our modeling endeavors, which favored comprehensive continuous data over its categorized summary.

A special case was the Work Structure column, which contained a significant number of null entries. Although we could have imputed these as Not Specified, it was determined that removing this column would not substantially impact the analytical results or the robustness of our predictive models. This decision was made with the objective of preserving data integrity and modeling efficiency.

A thorough review of each column's contribution to our analysis, guided by our domain expertise, enabled this strategic pruning of the dataset, ensuring the remaining information was ideally structured for our analytical goals and free of unnecessary complexity.

## **Removing Unnecessary Rows**

A close-up of a data

Description automatically generated

A screenshot of a computer

Description automatically generated

During the data cleaning stage of our analysis, we carefully pruned the Seagate HR dataset to remove rows that were not necessary for our analysis. This process involved the elimination of rows based on logical consistency and the presence of null values. Specifically, we removed a single record that lacked a value in the 'Generation' column, as this feature is crucial to our analysis. Additionally, we discarded five entries, where the Pay Level data was missing, assessing that this would not remarkably impact our overall model.

Upon a detailed review of the Compa Ratio statistic, which compares the compensation level to the market median, we identified 120 entries with a ratio exceeding 2, materially above the expected range. Considering that such high figures are infrequent and could indicate extreme outliers or data entry errors, we opted to exclude these records to maintain the coherence of our analysis. Additionally, we identified 12 rows in the dataset where the base pay for the Russian Federation was recorded as zero. Given that these entries constituted the entire data set for this work country, and it isn’t practicable to accurately predict the values, we decided to remove them.

Furthermore, we implemented a logical check to identify any records where an employee's termination was noted without an accompanying reason. These entries were extracted from our dataset and removed. Since, it was a single entry, no effect on the analysis was noticed.

These actions resulted in a dataset that is both logically consistent and analytically robust. Our dataset now provides a well-organized and reliable foundation for subsequent phases of our project, as verified by the updated descriptive statistics reflecting these changes. This tailored dataset is poised to deliver a more accurate and focused analysis.

## **Feature Engineering**

To address the missing values in our Seagate HR dataset, we employed a consistent set of features to impute both Base Pay and Compa Ratio. These features included Tenure, Cost to Replace Employee Multiplier, Job Title, Job Function, Job Category, Pay Level, Work Location, Work Country, Gender, Employee Status, and Generation. By consistently using these features, we ensured that the predictive models for the two financial metrics were built on the same underlying data dimensions, allowing for a reasoned and uniform approach to data imputation across the dataset.

This common set of features enhanced the feature engineering phase and maintained consistency in the analytical process, enabling us to leverage the correlations and patterns within the data to accurately predict and rectify missing values. Through a systematic approach, we enhanced the robustness of our imputation process and derived accurate and reliable values for Base Pay and Compa Ratio, thereby improving the dataset's analytical utility and completeness.

We developed a Python machine learning pipeline that used a Random Forest Regressor for predictive imputation and preprocessing transformations to operationalize this approach. To tackle the heterogeneous nature of our dataset, the preprocessor distinguished between numerical and categorical data and treated each appropriately, applying a One-Hot Encoder to categorical data and sending numerical data through. To anticipate missing Base Pay data, the model was trained on observations with non-missing target values. Ultimately, the dataset's zero values were substituted with these projections, maintaining the dataset's overall coherence, and setting it up for insightful analysis. This methodical, technologically advanced outlook supports our belief that the dataset is prepared for the next phases of our study.

A screenshot of a presentation

Description automatically generated

**Note:** The same set of 1,448 records contain null values in both 'Compa Ratio' and 'Base Pay'.

Additionally, to enhance the comparability of financial data across geographical locations, we created a new column in the Seagate HR dataset that leverages the exchange rates as on date to convert Base Pay values across different country denominations to USD. This step ensures that all base pay figures are standardized, allowing for a more accurate and fair analysis of compensation levels irrespective of the employee's work location. By doing so, we can perform cross-regional comparisons and analyses more effectively, contributing to a more rational understanding of global compensation trends within the organization.

## **End Results**

The table below provides a snapshot of the data cleaning steps undertaken for each column in our dataset. Identifiers and non-essential fields were removed, missing values were addressed through deletion or imputation, and incorrect entries were corrected streamlining our data for precise and focused analysis.



# **Descriptive Statistics**

A table comparing the original and cleaned datasets for Base Pay by Work Country is displayed in the image below. It provides insights into the distribution and central tendency of compensation numbers both before and after data cleaning. It does this by breaking down counts, means, standard deviations, and the range of base pay across percentiles for each country. By allowing an evaluation of the impact of cleaning on the statistical features of the data, this comparative perspective ensures that the cleaned dataset is suitable for further analysis.

A screenshot of a table

Description automatically generated

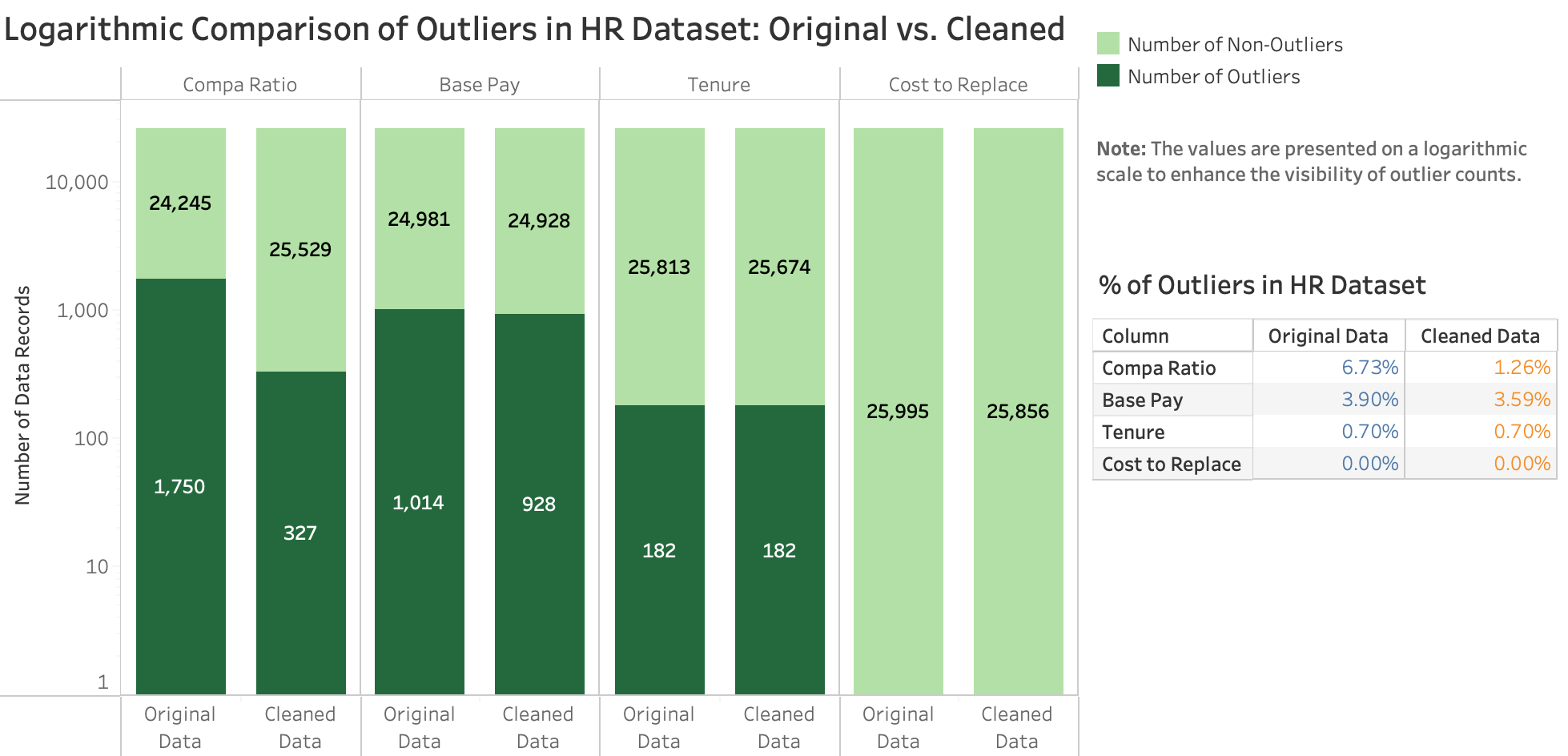
A screenshot of a data

Description automatically generated

Three metrics Compa Ratio, Cost to Replace, and Tenure are compared statistically between the original and cleaned datasets in the above table. Notably, data cleaning has raised the minimum value indicating the removal of extreme outliers and lowered the standard deviation in the Compa Ratio suggesting a decreased spread of data around the mean. As we bypassed the data cleaning owing to low percentage of outliers, the average costs to replace an employee and their tenure stayed pretty much the same. The spread of the numbers from smallest to the largest, didn't really change either. Indicative of our approach, we effectively eliminated extreme or unusual cases, maintaining the overall data's equilibrium and integrity. Moreover, these statistics demonstrate the effectiveness of the cleaning procedure, particularly in terms of normalizing the Compa Ratio variable which is essential for supplementary analysis.

# **Data Visualizations**

## **Outliers vs. Non-Outliers of Numerical Variables**



The above graph emphasizes on outlier and non-outlier values across numerical columns that considerably deviate, being much higher or lower than the rest. It demonstrates that Compa Ratio and Base Pay fields portray an apparent decrease in outliers after cleaning, indicating that data points which could have been errors or extreme values, have been addressed. Conversely, the Tenure and Cost to Replace columns exhibit minimal change in the number of outliers, suggesting that these areas either lacked predominant anomalies or that the cleaning criteria do not apply to these metrics.

When working with large datasets, employing a logarithmic scale is particularly beneficial as it aids in comparing counts across a broad range of values, thereby simplifying the visualization of outlier reduction after data cleaning. Notably, the outlier percentage for 'Compa Ratio' decreased from 6.73% to 1.26%, reflecting a more normalized dataset that is better suited for nuanced analysis.

## **Histograms**

### **Employee Tenure**

The histograms below illustrate the distribution of employee tenure prior and subsequent to data cleaning process. The striking similarities in their distributions and shapes suggest that the tenure data was already of high quality, with few errors or outliers to address. The minimal changes observed post-cleaning lends further support to this, indicating that no additional refinement measures were necessary.

A graph of blue and white lines

Description automatically generated

In the pre-cleaning histogram, there are notable number of employees with shorter tenures that gradually reduce as tenure length increases. This pattern is typical in workforce data, reflecting the dynamics of employee turnover and the higher influx of new hires compared to long-term employees.

A graph of green and black lines

Description automatically generated with medium confidence

The post-cleaning histogram displays a similar pattern, reinforcing the notion that the data cleaning was judicious, targeting only the most critical data points for removal or correction without disturbing the general trends within the Tenure column.

### **Compa Ratio**

A white background with numbers and text

Description automatically generated

The pre-cleaned histogram reveals that a significant number of both active and terminated employees had a Compa Ratio below 1, indicating that their pay was below the market median. The presence of a large initial bar at 0.50 denotes that a considerable portion of the workforce received significantly less than the market rate. As the histogram progresses to the right, the bars diminish in height, showing that fewer employees have higher Compa Ratios.

A graph of green and black bars

Description automatically generated

The post-cleaned histogram exhibits a remarkable transformation. The distribution of Compa Ratios is more uniform with a central peak. This pattern is characteristic of a normal distribution, which often emerges naturally in a variety of real-world scenarios and is generally preferred for analytical modeling. The data cleaning process appears to have moderated the extreme values that were skewing the original distribution.

### **Base Pay**

A graph of data in blue and white

Description automatically generated with medium confidence

The pre-cleaning histogram reveals the range of base salaries across the workforce. The bulk of employees, both active and terminated earned between $0 and $50,000. As base pay increases, fewer employees fall into each successive bracket indicating a substantial skew towards the lower end of the base pay range.

A graph of green and black bars

Description automatically generated with medium confidence

In the post-cleaning histogram, although there is still a peak at the lower end, the distribution across different base pay ranges is now more uniform. This indicates that the cleaning process has addressed some of the extreme values in base pay, but the overall trend of most employees earning lower base pays persist.

## **Scatter Plots**

The scatter plots below provide a granular view by illustrating how the tenure and compa ratio correlate with actual base pay across the organization and how data cleaning can influence these relationships.

## 

### **Tenure vs. Base Pay**

A screen shot of a graph

Description automatically generated

In the pre-cleaned plot, the data points are widely dispersed indicating variances in base pay across different countries and tenure levels. The spread suggests that base pay can vary predominantly for similar tenure lengths, possibly due to company specific pay scales or country-specific economic factors.

A screen shot of a graph

Description automatically generated

The data points in the post-cleaned plot seem more closely grouped around certain tenure ranges, indicating a consistent trend. This stipulates that the data cleaning process has removed or imputed some of the outliers, inaccurate entries, and null values in base pay column.

### **Compa Ratio vs. Base Pay**

A graph of a number of countries/regions

Description automatically generated with medium confidence

The pre-cleaned plot above, displays striking variation in base pay across all compa ratio levels suggesting that a substantial number of employees were compensated at rates below or above the market average.

A graph of colored circles

Description automatically generated with medium confidence

In the post-cleaned plot, the range of base pay across various compa ratios appears to be less variable and more clustered around the market rate i.e. compa ratio of 1. The consolidation of these data points demonstrate that the data cleaning process has eliminated some of the more extreme discrepancies in base pay and compa ratio columns.

## **Correlation Heat Map**

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

Concerning the pre-cleaned heat map, the compa ratio portrays minimal correlations with all other variables, indicating an absence of linear relationships. Tenure shows a slight positive correlation with the Base Pay and Cost to Replace metrics, suggesting a marginal positive relation. The strongest correlation though still nominal is between Base Pay and Cost to Replace columns, hinting that as base pay increases, the cost to replace an employee may rise proportionately.

The post-cleaned heatmap reveals a slightly more pronounced weak positive correlation between Compa Ratio and Tenure. The increased negative correlations of Compa Ratio with Base Pay and Cost to Replace columns suggest that a higher Compa Ratio may correlate with lower values in these variables. Furthermore, the weak positive correlation between Base Pay and Cost to Replace fields intensified marginally, indicating that data cleaning process may have unveiled a more pronounced association between these metrics.

# **Implications of observed changes**

The rigorous data cleaning process has notably enhanced the quality of the Seagate HR dataset, carrying substantial implications for the reliability and validity of subsequent analyses. Below is a detailed examination of these changes:

**Improved Data Quality:** Outliers, particularly in the Compa Ratio and Base Pay variables, have been eliminated to the most extent. This reduction in noise helps prevent the distortion of trend analyses and predictive modeling. The Compa Ratio now clusters more tightly around the market rate, manifesting that the data more accurately reflects the company's compensation landscape.

**Enhanced Analytical Precision:** By judiciously removing irrelevant rows and columns, the dataset has become more streamlined, potentially leading to more precise machine learning models. This refined dataset helps to improve the model’s generalizability while reducing the risk of overfitting.

**Increased Representativeness:** With the imputation of missing values through utilizing a consistent set of features, the dataset now presents a fuller picture. This is crucial for cross-country comparisons, where incomplete data could lead to biased interpretations of compensation trends.

**Better Decision Making:** The enhanced Seagate HR dataset fosters superior decision-making by providing a reliable, accurate foundation for supplemental analyses. This refinement ensures that decisions align with market trends and data integrity, leading to better alignment with project’s goals.

# **Conclusion**

To encapsulate, the data cleaning process applied to the Seagate dataset exemplifies a conscientious approach that balances the need for advanced analytics while maintaining the original data's integrity. The resulting dataset is not only cleaner but also primed to yield more insightful analyses.

The salient outcomes below reflect the effectiveness of the data cleaning process:

**Predictive Analysis Readiness:** The dataset is now better suited for advanced analytical techniques, such as hiring optimization and predictive modeling for voluntary turnover. The models constructed will be underpinned by the cleanest data possible, enhancing precision.

**Foundation for Strategic Decisions:** With more transparent data, we can make more informed strategic decisions. The refined dataset offers a dependable foundation for critical initiatives such as developing retention strategies.

The overall data cleaning process underscored the importance of data quality and its direct impact on analytical conclusions. The Seagate HR dataset now stands as a testament to the power of thorough data cleaning. The implications of the refined data resonate through every aspect of the analysis that follows, ensuring that any conclusions drawn are based on the most sound and robust information available.