**PRELIMINARY DATA REPORT**

A close-up of a 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

[Metadata 3](#_Toc161783940)

[Data Analysis Tools/Software 5](#_Toc161783941)

[Data Exploration 5](#_Toc161783942)

[Ø Data Classification: Categorical vs. Numerical 5](#_Toc161783943)

[Ø Unique Value Count: Categorical 6](#_Toc161783944)

[Ø Unique Value Count: Numerical 6](#_Toc161783945)

[Ø Employee Demographic and Statistics 7](#_Toc161783946)

[Ø Histogram of Employee Tenure 8](#_Toc161783947)

[Ø Scatter Plot 9](#_Toc161783948)

[Ø Correlation Heat Map 11](#_Toc161783949)

[Ø Descriptive Statistics 12](#_Toc161783950)

[Ø Employee Termination Analysis 13](#_Toc161783951)

[Data Issues 14](#_Toc161783952)

[Ø Outliers 14](#_Toc161783953)

[Ø Null or Missing values 14](#_Toc161783954)

[Ø Duplicates 15](#_Toc161783955)

[Ø Sanity Checks 15](#_Toc161783956)

# **Metadata**

The HR dataset under analysis is a comprehensive resource for studying employee turnover among non-manufacturing specialists at Seagate while omitting executives and manufacturing specialists. It includes a wide array of variables related to employment outcomes, job characteristics, and employee demographics. Our primary objective is to utilize this data to predict and identify the main factors contributing to employee termination, understand its financial impact, and devise a recruitment strategy that effectively mitigates the incidence of turnover over the next two years.

**A close-up of a number

Description automatically generated**

Each row in the dataset corresponds to individual records or data points for an employee in the company, capturing details such as their tenure, job function, remuneration metrics, anonymized ID, and reasons for termination. This granular perspective of employee’s career trajectory within the organization is pivotal for understanding the dynamics of employee churn and its patterns.

Each column in the dataset acts as a descriptor for various facets of employee data, collectively offering a holistic view of workforce characteristics. These columns form a multi-dimensional matrix that provides essential variables required for an in-depth analysis of employee turnover. Below are the detailed descriptions for each column:

A screenshot of a computer

Description automatically generated

A close-up of a text

Description automatically generated

A group of white rectangular boxes with green text

Description automatically generated

A screenshot of a computer

Description automatically generated

# **Data Analysis Tools/Software**

**Microsoft Excel:** Initially, Excel provided a straightforward and user-friendly environment for early-stage exploratory analysis, facilitating quick data sorting, preliminary cleansing, and basic visual representation. This phase was essential for gaining an initial understanding of the data's structure and identifying any apparent trends or data quality issues.

**Python (Pandas):** As the analysis progressed to a more detailed phase, the Pandas library became crucial for performing complex data manipulations and generating coherent summary statistics. With its flexibility and robustness, Pandas enabled intricate transformations, exhaustive data cleaning, and preparation for detailed analysis. This preparation laid the groundwork for identifying factors contributing to employee turnover.

**Tableau**: In the visualization phase, Tableau played a pivotal role in offering advanced and interactive graphical representations of the data. Utilizing Tableau, histograms, scatter plots, and other visualization forms were created to explore the data's underlying patterns and relationships effectively. This helped in identifying key trends and anomalies that were not immediately apparent in tabular data. Tableau's dynamic and intuitive interface allowed for the quick adjustment of visual parameters, enabling deeper and more nuanced analysis of the HR dataset.

# **Data Exploration**

## **Data Classification: Categorical vs. Numerical**

The image below displays two categories of attributes: Numerical and Categorical. Under the Categorical heading, columns such as Anonymized ID, Job Title, Gender, and Termination Reason are included, which represent non-numeric data points that describe traits or characteristics. Meanwhile, the Numerical heading lists attributes like Compa Ratio, Tenure, Base Pay, and Cost to Replace. These are quantitative data values that measure variables. This classification assists in distinguishing between qualitative and quantitative elements for the employee-related analysis.

A diagram of a work site

Description automatically generated with medium confidence

## **Unique Value Count: Categorical**

A table with numbers and text

Description automatically generated

Anonymized ID denotes a unique identifier for each record in the data table, which displays a broad variety of unique values across categorical columns. Thousands of distinct items in the Job Code and Termination Date columns demonstrate the diversity of work roles and termination events. In contrast, the data for Gender, Work Structure, and Work Region show less variety and very few unique values. The moderate diversity displayed by mid-range values in columns such as Job Title and Work Location suggests an organizational structure that gradients from specific identifiers to general categorization.

## **Unique Value Count: Numerical**

A table with blue squares

Description automatically generated

With 8,841 unique values, the Tenure column has the most variety, indicating a wide range of employee service periods. The Compa Ratio and Base Pay columns have a moderate number of distinct values, 910 and 665, respectively, representing compensation variability. On the other hand, there are only two distinct multipliers for Cost to Replace, indicating standardization of replacement costs.

## **Employee Demographic and Statistics**

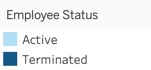
A chart of a group of people

Description automatically generated

## **Histogram of Employee Tenure**

A graph of blue and white bars

Description automatically generated



This graph depicts the distribution of employee tenure, contrasting active employees with those who have been terminated. Most employees, both active and terminated, have a tenure of less than 10 years, with the highest concentration seen in the latest tenure groups. The number of employees decreases as tenure increases, suggesting that fewer employees stay with the company for longer periods.

## **Scatter Plot**

**Analyzing the Diverse Spectrum of Tenure and Base Pay**

A graph showing a number of blue dots

Description automatically generated

The scatter plot illustrates the relationship between base pay and employee tenure across various locations. The majority of employees have fewer than ten years of tenure, as shown on the horizontal axis. There is a dense clustering of data points where both tenure and base pay levels are lower, with data becoming more sparse as these variables increase. The vertical spread of base pay at similar tenure levels suggest significant variations in remuneration, which could be due to differences in job functions, pay scales, or regional currency values. The figure reveals the wide range of pay scales, which may result from including multiple currencies and the diverse economic conditions of the regions represented, rather than a strong correlation between tenure and base salary. Outliers at higher pay levels may indicate executive compensation or specialized high-skill positions that diverge from typical pay scales.

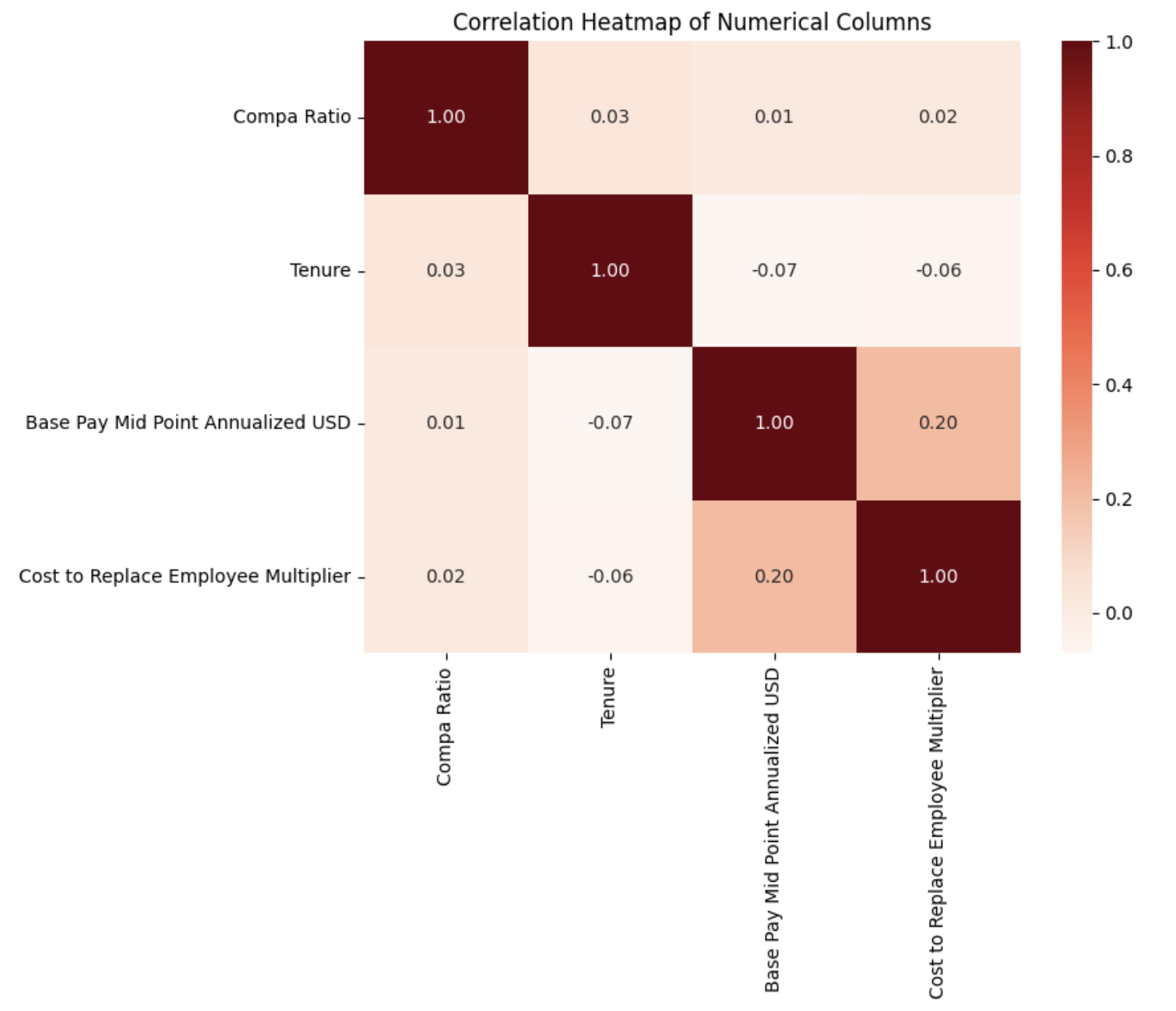
**Interpreting Base Pay Variability Across Compa Ratios**

A graph of a graph

Description automatically generated with medium confidence

Considering that Base Pay is presented in a variety of currencies and locations, this scatter plot depicts the relationship between the Compa Ratio and Base Pay. Most data points are clustered at the lower end of the Compa Ratio spectrum, indicating that most employees' pay aligns closely with the organization's established compensation structure. Fewer data points appear as the Compa Ratio increases, representing employees likely earning well above the average, which could be attributed to exceptional performance or roles requiring specialized skills. It is important to note that what appears to be high Base Pay figures could be a result of conversions from other currencies with higher nominal values compared to the USD. The graph shows a broad range that could be influenced by factors such as location-specific pay standards, market rates, and currency strengths, rather than a strong or consistent correlation between Compa Ratio and Base Pay.

## **Correlation Heat Map**



The relationship between various numerical variables is visually represented in this correlation heatmap. The strength of the correlation is indicated by the color intensity, where 1.00 denotes a perfect positive correlation that naturally occurs along the diagonal, which is the point where each variable intersects with itself. There is extremely little to no linear relationship between the variables, as indicated by the other values that are close to 0. For instance, values of -0.07 and -0.06 suggest a marginally negative association between Tenure and Base Pay Mid Point Annualized USD, and Cost to Replace Employee Multiplier, respectively. Conversely, a weak positive association (r = 0.20) between Base Pay Mid Point Annualized USD and Cost to Replace Employee Multiplier indicates that as base pay increases, so does the cost of replacing an employee. Overall, this heatmap suggests that none of these factors strongly predict the others, due to the lack of significant linear correlations between them.

## **Descriptive Statistics**



The data indicates that the Base Pay and Compa Ratio columns have significantly different means and standard deviations. Notably, the standard deviation of the Base Pay is substantially higher than its mean, which suggests a wide distribution of values around the mean. This condition typically signals the presence of outlier values that significantly diverge from the standard data range. Such outliers can drastically elevate the mean and inflate the standard deviation, demonstrating that the data points span around broad spectrum of values.

The analysis could be further complicated by null values. If not appropriately addressed, these can artificially influence the calculations of the mean and standard deviation. To ensure the most accurate results, null values must be diligently managed or imputed before proceeding with these statistical measures.

In summary, the presence of outliers and null values is likely responsible for the inflated mean and standard deviation values. In such cases, examining the median is often beneficial, as it is more robust to outlier influence. Unlike the mean, which can be heavily skewed by outliers, the median remains relatively unaffected and therefore provides a better sense of the data's central tendency in the presence of such anomalies.

## **Employee Termination Analysis**

A graph with numbers and colored squares

Description automatically generated with medium confidence

The bar chart shows that layoffs and downsizing, primarily through Position Elimination/RIF (Reduction in Force) Involuntary, is the most common reason for employee termination. Career Advancement is the second major cause, indicating that many employees leave to pursue better career opportunities. A significant number also leave to pursue a Change of Career Direction, seeking different professional experiences. A smaller group opts for partially voluntary separation, as indicated by Position Elimination/RIF Voluntary. The least common reason is Return Home/Care for Family, suggesting that personal obligations are a less frequent cause for departure. Addressing these primary issues, especially involuntary terminations, and career advancement opportunities, may be crucial for improving employee retention strategies.

# **Data Issues**

During a thorough data exploration process, several data quality issues were identified that require careful management and evaluation to support the validity and trustworthiness of the data for future analyses.

## **Outliers**

In the dataset, outliers were observed, indicating significant variations in critical metrics related to pay and length of employment. Specifically, 1,750 employees had Compa Ratio values that were anomalous, signifying large disparities in how wages compare to market prices. Additionally, 182 employees exhibited extraordinary employment lengths in the tenure column, compared to their peers. Moreover, abnormal base pay statistics were found for 3,416 employees, potentially highlighting instances of exceptionally high or low pay.

The image below shows the profile of an employee with a compa ratio of 0.0 who has been employed for more than 20 years. Since this number usually indicates how much a person makes in relation to the market median, a compa ratio of 0.0 for someone with such a long tenure is quite unusual and may be a sign of an anomaly that warrants further investigation.



## **Null or Missing values**

There were instances of significant issues with null or missing values occurred, especially concerning fired personnel. Notably, the Compa Ratio field for 1,465 employees contained null values, indicating a lack of information on remuneration records compared to market rates. Furthermore, five individuals had missing information in the Pay Level column, which is crucial for understanding their pay scale within the company. Most conspicuously, the employment arrangements such as preferred remote, onsite, or hybrid work for 8,947 employees were obscured, as the Work Structure column lacked entries for them. These data gaps are principally concerning because they pertain solely to terminated employees, potentially skewing the analysis of turnover trends and reasons. Therefore, addressing these null values through imputation or exclusion is essential, depending on their impact on the analysis and the proportion of missing data in each column.

The example below indicates instances where null or missing values occur in the Work Structure and Pay Level columns for certain terminated employees. These omissions underscore the necessity of conducting a comprehensive review and data cleansing to ensure the accuracy of any research related to pay and employee turnover.



## **Duplicates**

The examination of the dataset revealed no duplicate records, maintaining the dataset's integrity. The absence of duplicates ensures accuracy and reliability of the statistics and analyses derived from the data.

## **Sanity Checks**

Discrepancies in several records were noted, especially when comparing Compa Ratio and Base Pay with Job Title and Job Function. These inconsistencies suggest potential data entry errors or systematic issues within the dataset, indicating that remuneration data may not consistently align with job classifications. Additionally, there are 19 cases where employees are marked as Active, yet their records include a termination type and reason, indicating anticipated terminations in 2024. Addressing these anomalies is essential for accurate representation of the organization's job hierarchy, compensation strategy and employment status.

In addition to the above-mentioned data quality issues, two critical aspects were thoroughly examined: Data Type Discrepancies and Temporal Inconsistencies. For Data Type Discrepancies, we verified that all data types, including dates, categorical data, and numerical values, were correctly identified, and formatted, ensuring they met the expected standards. In the case of Temporal Inconsistencies, we scrutinized the dataset for any chronological inconsistencies, ensuring that all records are aligned correctly with the dataset's timeline. Upon evaluation, no inconsistencies were found in these aspects, affirming the data's structural integrity and chronological accuracy.