Preliminary Data Report

SEAGATE EMPLOYEE CHURN PROJECT

Submitted by:

Charles Biddle Porter
charles.biddleporter@colorado.edu

Duc Hoang
duc.hoang@colorado.edu

Harshit Gole
harshit.gole@colorado.edu

Kirti Vatsh
kirti.vatsh@colorado.edu

Manan Dhanteja
Manan.dhanteja@colorado.edu

TABLE OF CONTENTS

REPORT OBJECTIVE

The objectives of preliminary data analysis are to edit the data to prepare it for further analysis, describe the key features of the data, and summarize the results. The goal of this report is to maximize our insights into the Seagate dataset.

The primary objective of this preliminary data analysis is to meticulously prepare the Seagate dataset for in-depth examination by editing and refining the data, ensuring its readiness for subsequent analysis phases. This entails a comprehensive description of the dataset's key features and a summary of the initial findings, aiming to uncover and highlight critical insights that inform our understanding of the dataset's characteristics and underlying patterns.

This report aims to maximize insights derived from the Seagate dataset, facilitating a more profound comprehension of its intricacies, and fostering informed decision-making processes. By achieving a thorough preliminary analysis, we set the foundation for identifying trends, challenges, and opportunities within the dataset, thereby enhancing the effectiveness of further analyses and strategic initiatives.

DATA DESCRIPTION

A. METADATA

The dataset is a detailed collection of employee data with a scope that includes various job-related and demographic variables. The dataset has 25,995 rows corresponding to individual observations or employee entries. The columns represent job-related attributes such as job code, job title, comp ratio, tenure, and base pay. There are a total of 22 columns in the dataset. The data was formatted in an Excel file and then converted into a CSV file used in R for data analysis.

B. DATA ANALYSIS TOOLS/SOFTWARE

Our team's data examination process is strategically designed, employing a combination of tools for comprehensive analysis. Excel is our initial tool, providing a broad view of the data and helping us understand its structure and key characteristics. This step is crucial as it flags any significant issues or patterns that require further investigation. R, on the other hand, is our go-to for in-depth analysis, especially for descriptive statistics and data aggregation. Its robust capabilities allow us to perform complex statistical analyses and aggregate data precisely, leading to a deeper understanding of our dataset's underlying trends and insights. Excel and R are the pillars of our data examination process, showcasing our team's expertise and the value of our analytical workflow.

A parallel analysis shall be undertaken with Python for the following end goals:

1. Corroborate the preliminary results from the descriptive analytics of Excel and R to satisfy the deliverables of employee churn rate and its financial impact on the company.
2. Depict these results via data visualizations and append baseline actionable HR insights to them.
3. Mold the cleaned raw data into sets of training, validation, and testing data models.
4. Python is instrumental in using predictive analytics and machine learning techniques. We use it to fine-tune the data models until the desired actionable HR insights are achieved. This process continues until we confidently forecast the headcount for a two-year hiring pipeline, meeting our deliverables.

C. DATA EXPLORATION

1. A unique value count of categorical values

   The dataset encompasses a diverse range of job-related classifications. It contains 191 job titles and is organized into 11 job categories. Employees are further grouped into four distinct job groups. In terms of remuneration, there are 33 different pay levels. The

workforce is spread across 82 locations in 27 different countries. Gender diversity is acknowledged with four different identified groups. Employee status is bifurcated into two categories: Terminated and Active. In capturing generational diversity, individuals are segmented into six groups, which includes accounting for null values. The tenure of employees is categorized into six distinct buckets. Finally, the dataset details five termination types alongside 38 unique termination reasons.

2. Frequency count of categories

The dataset encompasses a diverse range of job-related classifications. It contains 191 job titles and is organized into 11 job categories. Employees are further grouped into four distinct job groups. In terms of remuneration, there are 33 different pay levels. The workforce is spread across 82 locations in 27 different countries. Gender diversity is acknowledged with four different identified groups. Employee status is bifurcated into two categories: Terminated and Active. In capturing generational diversity, individuals are segmented into six groups, which includes accounting for null values. The tenure of employees is categorized into six distinct buckets. Finally, the dataset details five termination types alongside 38 unique termination reasons.

There are 19060 males and 6917 females, with 18 records that have not been provided or are unknown. The top five job titles which have the highest number of records are:

- Sr engineer - 2446 records
- Engineering Specialist IV - 2224 records
- Engineering Specialist III - 2190 records
- Staff Engineer - 1899 records
- Sr Engineering Specialist - 1792 records

Millennials constitute the largest group within the workforce, totaling 111,147 individuals. Following them is Generation X, with 10,675 members. Baby Boomers are also represented, though in a smaller number of 3,191. Generation Z, the youngest cohort, is present with 966 individuals. Additionally, there are 15 individuals from the Silent Generation.

The tenure categories show the distribution of the employees' tenure with the largest group of employees falls within the '10 - 20 Years' range, with 7,306 individuals. This suggests that a significant portion of the workforce has a long-standing history with the company. The '5 - 10 Years' range also has a notable representation with 5,226 employees.

The '1 - 3 Years' category includes 3,813 individuals, followed closely by those with '20+ Years' at 5,582, indicating a substantial number of very experienced employees. The '3 - 5 Years' tenure bracket has 2,686 employees. The least represented group is those with less than one year of service, numbering 1,382.

3. Histograms of Continuous value columns

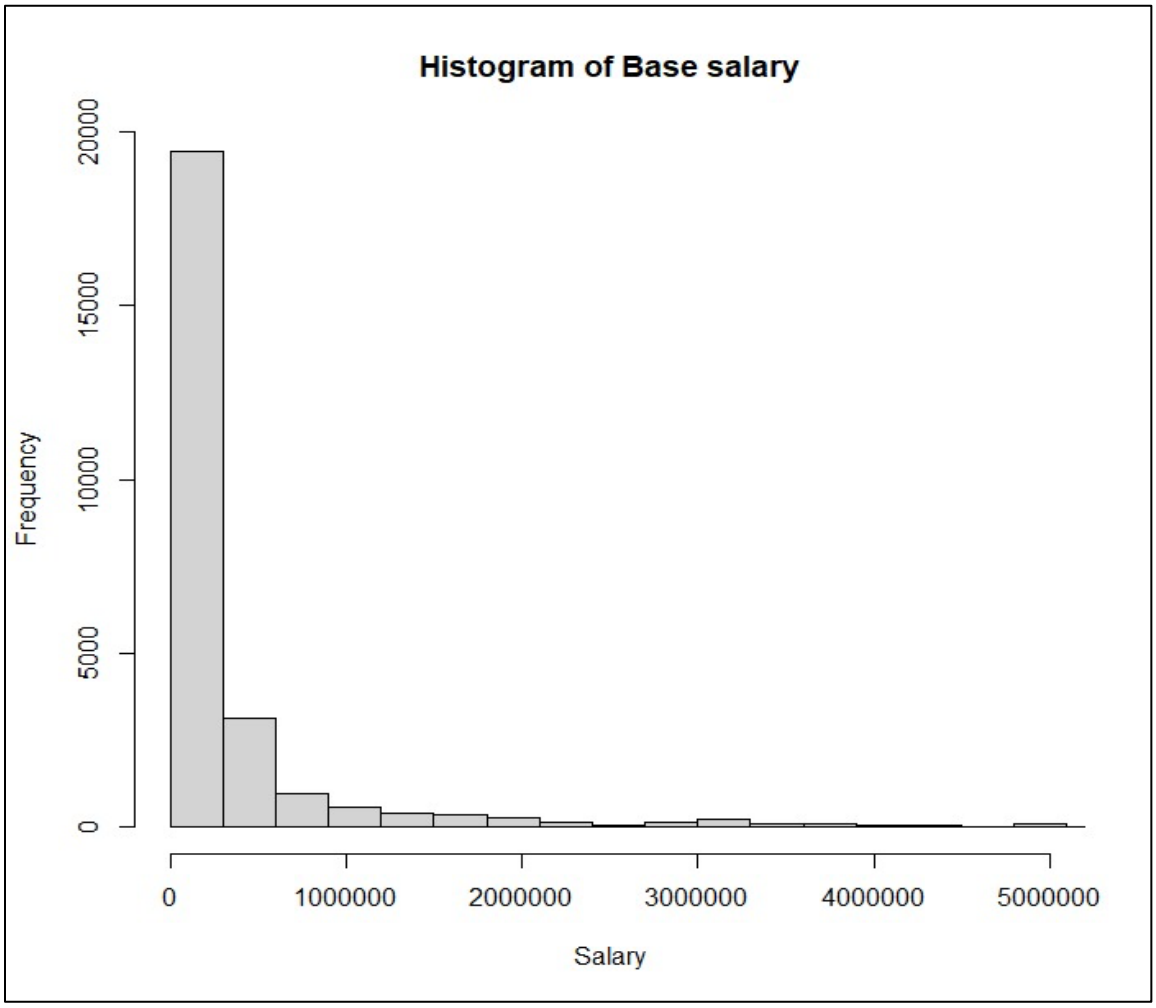

Figure 1: Histogram of Base Salary

The histogram shows that the base salary distribution is not uniform and is likely not normally distributed, with a concentration of employees earning lower salaries and a long tail extending towards higher salaries. This indicates that the data is heavily skewed to the left. here's a significant peak in the 0 to 300,000, suggesting that the most common salaries are close to the minimum of the range which is smaller than 300,000.

The histogram of tenure indicates that the largest frequency of employees has tenure lengths on the lower end of the scale, with a peak frequency for employees with less than

10 years of staying. As the tenure length increases, there is a clear downward trend in frequency, showing that fewer employees have longer tenures. The histogram displays a left-skewed distribution of tenure.

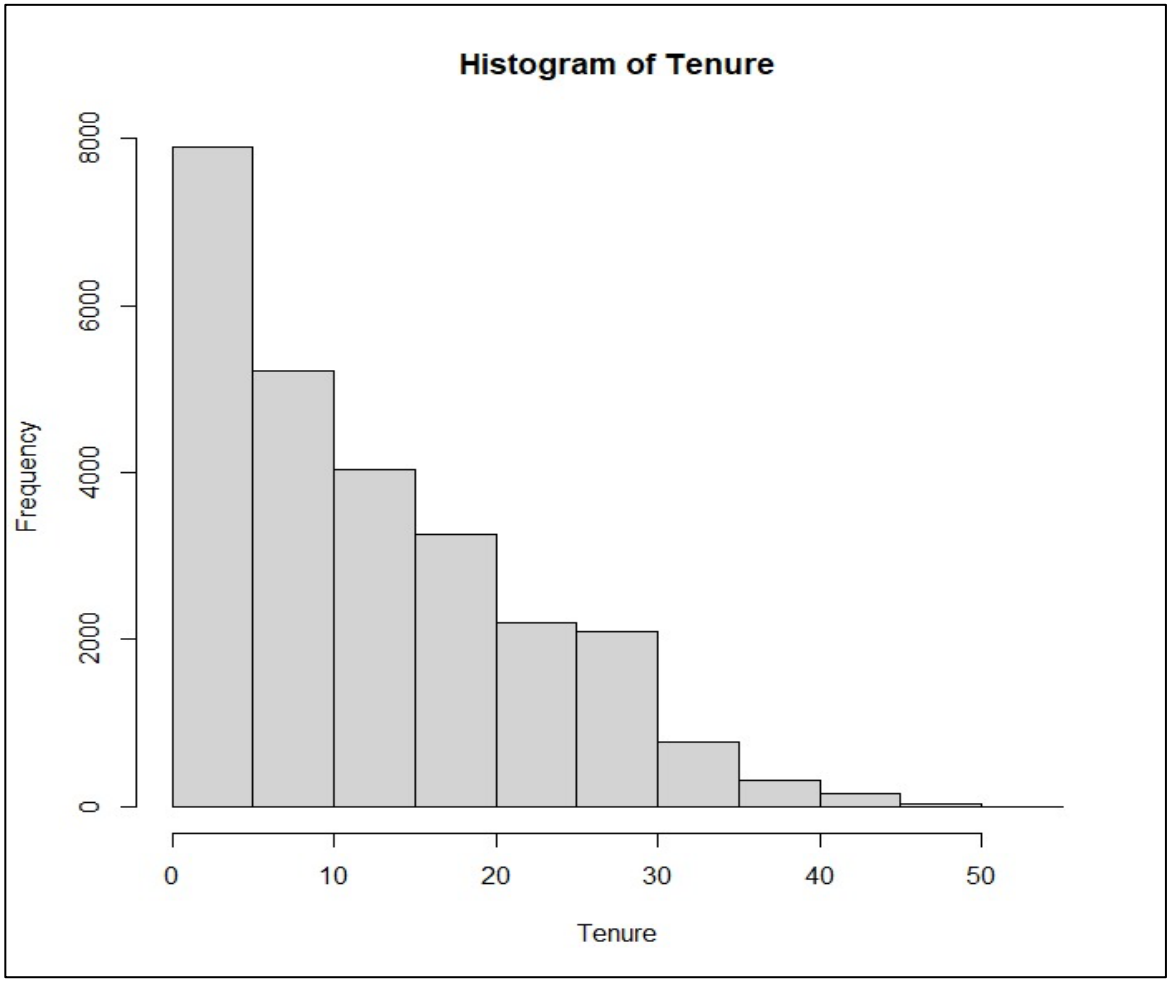

Figure 2: Histogram of Tenure

4. Variance of continuous value columns

   The variance for Base salary column is 882674086035. This variance seems to be extremely high due to the base salary is not normalized to a standardized currency.

   A variance of 92 in the tenure data suggests moderate spread in the lengths of service among employees. This figure indicates that, on average, the individual tenure lengths vary from the mean tenure by roughly 9.6 years, considering that the standard deviation is the square root of the variance.

5. Scatterplot of continuous variables

The scatter plot of Base Salary versus Tenure shows a wide dispersion of base salary values at lower levels of tenure, with a concentration of data points towards the bottom of the salary scale, which suggests that a significant number of employees with less tenure tend to have lower base salaries. As tenure increases, the spread of salary values becomes less dense and more variable, with some data points. However, the plot does not show a clear, definitive trend or correlation between tenure and base salary, indicating that factors other than tenure may have a stronger influence on salary levels or due to heavily skewed base salary which is not normalized to standardized currency.

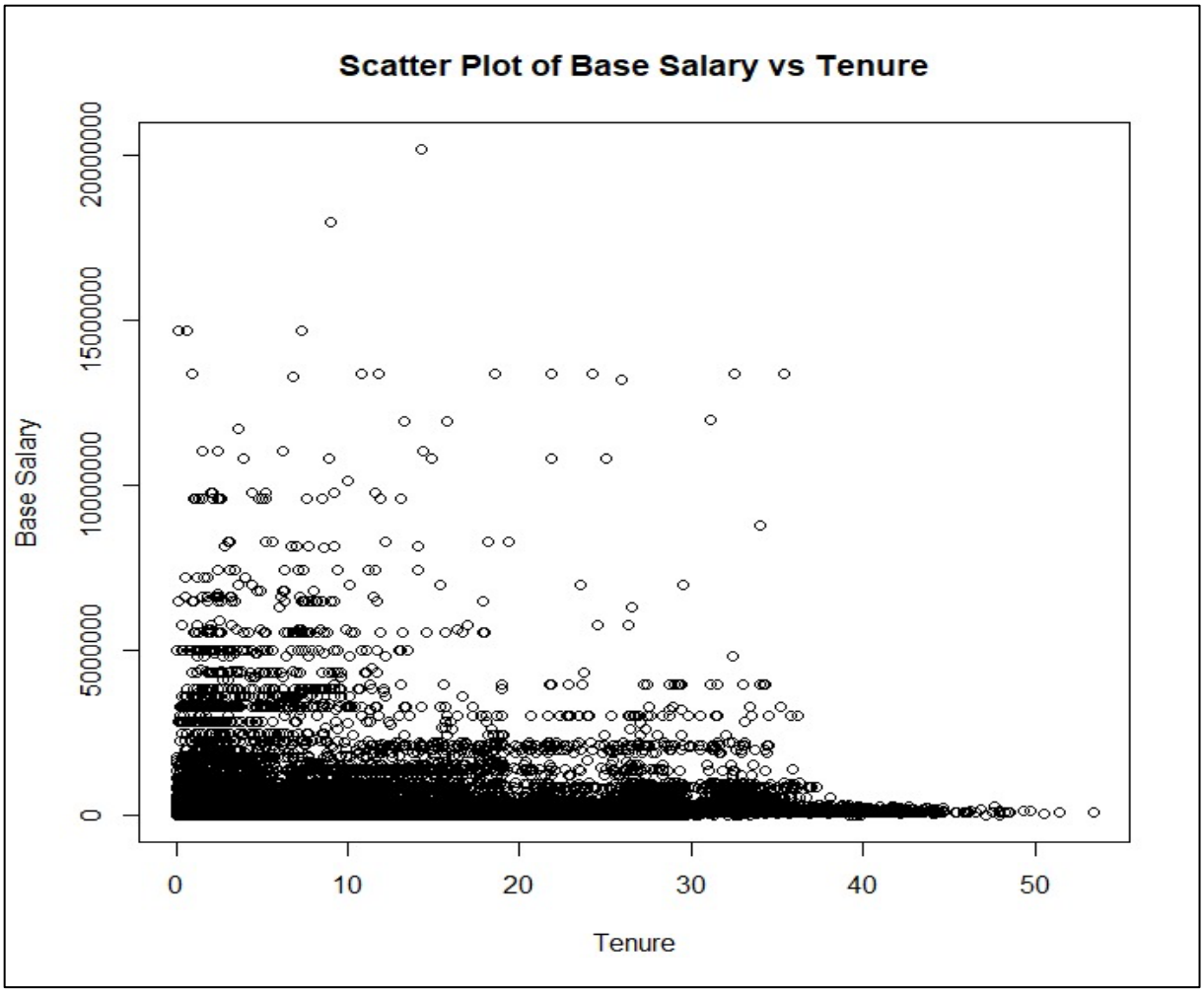

Figure 3: Scatterplot of Salary vs Tenure

6. Correlation Heatmap

A correlation heatmap was utilized to conduct a detailed analysis within an HR dataset, employing Python due to the computational intensity that surpassed Excel's capabilities. This visual representation elucidates the relationships between various HR metrics, highlighting their key correlations. Notably, the heatmap unveiled a robust negative correlation of -0.72 between GenX and Millennial employees, suggesting a distinct generational gap. Positive correlations were observed between Base Pay Mid-Point Annualized USD and specific geographic locations, with a +0.59 correlation with India and a +0.50 correlation with Japan, indicating regional salary differentials. Furthermore, a negative correlation of -0.44 was found between Tenure Bucket 20+ years and Millennial employees, aligning with the generational age differences.

In contrast, a positive correlation of +0.43 between Comp Ratio and France suggests a specific pay structure alignment in the French context. The analysis also highlighted increased correlation values with longer tenure buckets, indicating more pronounced tenure-related trends. Singapore's negative correlations with the US and Thailand and the Cost to Replace Employee Multiplier suggest geographical and financial complexities in employee replacement strategies. The correlation heatmap is shown in the next page.

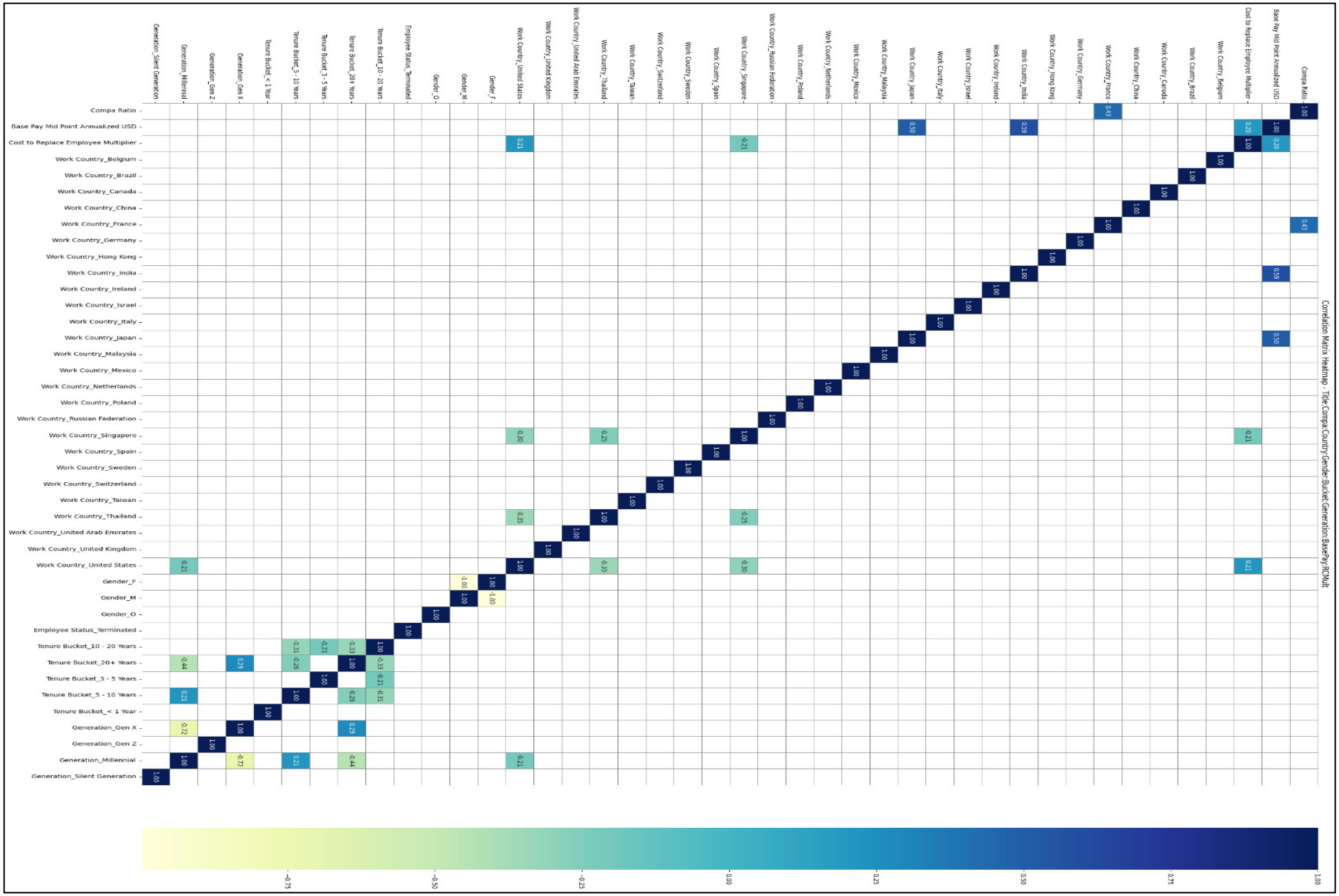

Figure 4: Correlation Heatmap

7. Descriptive Statistics

The dataset's descriptive statistics offer insights into various employee-related attributes across 25,995 records. We have a wide range of unique identifiers for the categorical data, such as Job Code and Job Title. The Tenure variable reveals an average employee tenure of roughly 12.08 years, with a standard deviation of about 9.62 years, suggesting diversity in the workforce's employment duration. In addition, the Comp Ratio, a measure comparing actual salary to the market rate of the salary range for a position, shows an average value of approximately 0.90 with a median of 0.89. These figures are close, indicating a relatively even distribution around the market rate. In terms of the Cost to Replace Employee Multiplier, which reflects the relative expense of replacing an employee, the average is reported at 1.02, with a median of 1.25. This suggests that the typical cost to replace an employee is about 1.02 times their annual salary, although the range of values could vary, as indicated by the data. Null values are noted in several columns, including Termination Date, Work Structure, Termination Type, and Termination Reason, indicating incomplete data. For example, the Termination Type has a relatively high frequency of null values, which could impact the filtering process of involuntary terminations.

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew
Annonymized ID	1	25995	36705.08	25162.70	33063.00	35078.88	28086.37	1111.0	8.7564e+04	8.6453e+04	0.50
Job Code*	2	25995	1908.11	1178.56	1640.00	1848.72	1322.48	1.0	4.5710e+03	4.5700e+03	0.40
Job Title*	3	25995	98.91	51.03	109.00	98.34	80.06	1.0	1.9100e+02	1.9000e+02	0.02
Job Function*	4	25995	10.07	4.70	7.00	9.60	1.48	1.0	2.2000e+01	2.1000e+01	0.56
Job Category*	5	25995	3.47	2.41	4.00	3.12	1.48	1.0	1.1000e+01	1.0000e+01	1.20
Job Group*	6	25995	3.02	0.93	3.00	3.15	0.00	1.0	4.0000e+00	3.0000e+00	-0.99
Compa Ratio	7	25995	0.90	0.74	0.89	0.89	0.13	0.0	1.6830e+01	1.6830e+01	12.88
Pay Level*	8	25990	12.89	6.04	12.00	13.01	4.45	1.0	3.2000e+01	3.1000e+01	0.08
Work Location*	9	25995	46.83	27.13	34.00	47.08	35.58	1.0	8.2000e+01	8.1000e+01	0.04
Work Country*	10	25995	20.43	7.14	24.00	21.51	4.45	1.0	2.7000e+01	2.6000e+01	-0.97
Work Region*	11	25995	1.76	0.55	2.00	1.75	0.00	1.0	3.0000e+00	2.0000e+00	-0.05
Gender*	12	25995	2.73	0.44	3.00	2.79	0.00	1.0	4.0000e+00	3.0000e+00	-1.07
Employee Status*	13	25995	1.55	0.50	2.00	1.56	0.00	1.0	2.0000e+00	1.0000e+00	-0.20
Termination Date*	14	14327	1180.23	636.07	1272.00	1202.26	773.92	1.0	2.1610e+03	2.1600e+03	-0.24
Tenure	15	25995	12.08	9.62	9.88	11.05	10.02	0.0	5.3410e+01	5.3410e+01	0.83
Tenure Bucket*	16	25995	3.77	1.48	4.00	3.78	1.48	1.0	6.0000e+00	5.0000e+00	0.10
Base Pay Mid Point Annualized USD	17	25995	406152.71	939507.36	149593.60	198139.21	138493.82	0.0	2.0190e+07	2.0190e+07	6.44
Generation*	18	25995	2.77	1.13	2.00	2.84	1.48	1.0	6.0000e+00	5.0000e+00	-0.05
Work Structure*	19	17048	2.01	0.30	2.00	2.00	0.00	1.0	3.0000e+00	2.0000e+00	0.19
Termination Type*	20	14327	2.95	1.94	3.00	2.94	2.97	1.0	5.0000e+00	4.0000e+00	0.04
Termination Reason*	21	14326	19.58	11.22	27.00	19.94	4.45	1.0	3.7000e+01	3.6000e+01	-0.49
Cost to Replace Employee Multiplier	22	25995	1.02	0.35	1.25	1.05	0.00	0.5	1.2500e+00	7.5000e-01	-0.81
...23	23	0	NaN	NA	NA	NaN	NA	Inf	-Inf	-Inf	NA

Figure 5: Descriptive Statistics for the dataset

D. DATA ISSUES

1. Outliers

   Our investigation into the dataset was thorough, revealing the presence of outliers. Specifically, we noticed that the Base Salary values for certain employees were significantly higher than the market rates posted online for identical roles. This discrepancy led to an inflation of both the mean and standard deviation of our base salary data. Upon further investigation, we assessed the compensation ratios (comp ratios) of these outliers and found them to be modest, which did not align with the substantial differences in pay observed. Our analysis led us to hypothesize that the reported salaries might not have been exchanged to USD as initially assumed but rather in the local currencies of the employees' respective locations. By reevaluating the salaries under this assumption and converting them to USD using appropriate foreign exchange rates, we found that the adjusted salary figures were consistent with both market expectations and the observed comp ratios. This adjustment corrected the disparities and aligned the data with our initial expectations. We also identified an issue with the Base Salary column: approximately 5% of the total observations are recorded as zero values. Such outliers can potentially skew the mean and standard deviation towards lower values.

2. Null Values

   Upon analysis, we identified null values in several columns, including Pay Level, Termination Date, Work Structure, Termination Type, and Termination Reason. These null values, particularly in Termination Type and Termination Reason, could present challenges during the data cleaning. For accurate analysis, it is essential to filter out specific categorical variables, such as involuntary terminations due to position eliminations. If not addressed, these omissions could skew our dataset, affecting the validity of subsequent analyses. It will be crucial to devise a strategy to handle these null entries to maintain the integrity of our data.

3. Sanity Checks

   The sanity check revealed no issues. Most active employees lack a termination date, indicating ongoing employment with the company. A few active employees do have termination dates listed; however, each of these entries is accompanied by a corresponding termination reason.

4. Duplicates

   There are no duplicates in the dataset.

CONCLUSION

In conclusion, our preliminary examination of the Seagate dataset has yielded valuable insights and discoveries. We have uncovered a broad spectrum of job titles, categories, groups, and pay levels, accompanied by an expansive distribution of work locations spanning numerous countries. The dataset reflects gender and generational diversity, with a notable predominance of males in the workforce. Tenure among employees spans a wide range, with a significant portion exhibiting prolonged service within the company. While termination types and reasons are thoroughly recorded, the presence of null values in specific columns calls for careful management to preserve the integrity of the data.

Our statistical analysis has indicated an exceptionally high variance in the base salary column, highlighting the necessity for standardizing salaries to a common currency for accurate comparison. The scatter plot analysis has not demonstrated a direct positive correlation between tenure and base salary, suggesting that additional variables may be at play in determining salary figures. The absence of anomalies in sanity checks affirms the data's reliability, and no duplication instances were detected in the dataset.

This preliminary report lays the groundwork for more intricate analyses to follow. With an understanding of these initial findings, we are better equipped to facilitate informed decision-making and delve deeper into the nuances of the data.