Peyton Bailey

Data Preparation and Exploration - D599

Task One – March 2, 2025

**Part I: Data Profiling**

**A. Profile data by doing the following:**

**1. Review the data dictionary in the attached "Employee Turnover Considerations and Dictionary" document and do the following:**

**a. Describe the general characteristics of the initial dataset (e.g., rows, columns).**

The initial dataset is a flat-file titled “Employee Turnover Dataset” that contains 10199 rows and 16 columns. It provides descriptive information within an organization regarding the employees who have left the company either voluntarily or involuntarily. This data includes the age, job type, commute distance, salary, tenure, as well as the number of annual professional development hours completed. Each employee record has a unique identifier, which is the employee number. There are also noticeable amounts of missing data within certain categories, such as “Annual Professional Development Hours.”

**b. Indicate the data type and data subtype for *each* variable.**

**c. Provide a sample of observable values for *each* variable.**

1. Employee Number
    a. Categorical
    b. Nominal
    c. 1, 2, 3
2. Age
    a. Numeric

b. Discrete

c. 28, 33, 22

3. Tenure

    a. Numeric

    b. Discrete

    c. 6, 2, 1

4. Turnover

    a. Categorical

    b. Nominal

    c. Yes, Yes, No

5. Hourly Rate

    a. Numeric

    b. Continuous

    c. $24.37, $24.37, $22.52

6. Hours Weekly

    a. Numeric

    b. Discrete

    c. 40, 40, 40

7. Compensation Type

    a. Categorical

    b. Nominal

    c. Salary, Salary, Salary

8. Annual Salary

a. Numeric

b. Continuous

c. 50689.6, 50689.6, 46841.6

9. Driving Commuter Distance

a. Numeric

b. Discrete

c. 89, 89, 35

10. Job Role Area

a. Categorical

b. Nominal

c. Research, Research, Information_Technology

11. Gender

a. Categorical

b. Nominal

c. Female, Female, Prefer Not to Answer

12. Marital Status

a. Categorical

b. Nominal

c. Married, Married, Single

13. Number of Companies Previously Worked

a. Numeric

b. Discrete

c. 3.0, 6.0, 1.0

14. Annual Professional Dev Hours
    a. Numeric
    b. Discrete
    c. 7.0, 7.0, 8.0
15. Paycheck Method
    a. Categorical
    b. Nominal
    c. Mail Check, Mail Check, Mailed Check
16. Text Messages Opt-In
    a. Categorical
    b. Nominal
    c. Yes, Yes, Yes

**Part II: Data Cleaning and Plan**

**B. Inspect the dataset through data cleaning techniques for *all* duplicate entries, missing values, inconsistent entries, formatting errors, and outliers and do the following:**

1. **Explain how you inspected the dataset for *each* of the quality issues listed in part B.**
2. **List your findings for *each* quality issue listed in part B.**

**Duplicate Entries**

The ".duplicated" function was used to find duplicate entries, passing the subset equal to 'Employee Number' as an argument. The output showed that there were 198 duplicated rows.

**Inconsistent Entries**

The ".unique()" feature was used to view each unique entry on each categorical variable. The ".describe()" feature was used on numerical variables. The ".unique()" output showed string

inconsistency amongst the Job Role Area and Paycheck Method categories. For example, within the Job Role Area category, the following entries were found for Information Technology:

1. Inforrmation_Technology,
2. InformationTechnology,
3. Information Technology

More than likely, this was a human error during the data entry stage. However, this needed to be fixed before analysis to produce accurate results. Using the "describe()" feature on the quantitative categories also revealed some inaccuracies. The Driver Commuter Distance and Annual Salary categories had negative values, which we know is impossible. Upon further inspection, I discovered that the Annual Salary entries with negative values were from the same Job Role Area, "Marketing." Also, the Hourly Rate category outputted an error because it was an object data type instead of float.

**Formatting Errors**

The ".dtypes()" feature was used to inspect each datatype. The output showed the previously mentioned object data type for the Hourly Rate category.

**Missing Data**

The ".isna().sum()" function gives a total number of null values per category. The "Num Companies Previously Worked" category had 663 null values, "Annual Professional Dev Hours" had 1947 null values, and "Text Message Opt-In" had 2258 null values.

**C. Discuss which data cleaning techniques you used to correct *all* the data quality issues you identified by doing the following:**

1. **Describe how you modified the dataset after identifying *each* quality issue in part B.**

**Duplicated Values**

The duplicated values were modified using the “drop_duplicates” function.

**Formatting Errors**

The values from the “Hourly Rate” category were treated using the str.strip function to remove the “$” and then converted into a float data type using the “astype” function.

**Inconsistent Entries**

The “str.replace” function was used to clean the text data to eliminate unnecessary white space and/or characters. “Information_Technology” was replaced with “Information Technology”. This was applied within the Job Role Area and Paycheck Method categories. The negative values from the “Annual Salary” and “Driver Commuter Distance” categories were temporarily replaced with null values.

**Null Values**

The null values imputed from the negative values within the “Annual Salary” category were filtered out and replaced with “Hourly Rate” multiplied by 2,080. The value 2,080 was computed from 40 hours in a week multiplied by 52 weeks in a year. The null values from the “Number of Companies Previously Worked” category were replaced by the median of 4 using the “fill na” function. The null values from the “Driver Commuter Distance” category were replaced by the median of 49. All values over 147 were considered outliers. There were 205 outliers, also replaced with a median of 49. The null values from the “Annual Professional Dev Hours” category were replaced with the median, 15. Finally, the null values from the “Text Message Opt-In” category were replaced with the mode, which was “Yes”.

2. **Discuss why you chose the specific data cleaning techniques you used to clean the quality issues listed in part B.**

The histogram for the “Number Companies Previously Worked” category was skewed to the right. In this case, the appropriate measure of center is the median, which was 4, used to replace all null values within this category. The “Driver Commuter Distance” histogram was also skewed right with very noticeable outliers. Using the describe function, I calculated the IQR (3Q -1Q) and utilized the outlier rule of 1.5 multiplied by IQR above Q3 or below Q1. There were no outliers below Q1 since all negative values had previously been removed. The highest reasonable value was calculated to be 147, meaning anything above would be considered an outlier. The “Annual Professional Dev Hours” category had a uniform distribution when plotted on a histogram. In this case, the mean and median are relatively equal, with the mean calculated as 14.94 and the median as 15. All null values in this category were replaced with 15 using the “fill na” function. The “Text Message Opt-In” category was the only categorical variable with null values. Those values were treated with an imputation of “Yes”, the mode of the category.

3. **Describe** two **or more advantages to your data cleaning approach specified in part C1.**

This data cleaning approach aimed to ensure better data quality, facilitating a more substantial and accurate analysis. Fixing inconsistent text data within certain variables, such as “Job Role Area,” provided consistency across the entire category. For example, Information Technology is now one unique variable instead of three. Furthermore, removing inaccurate data because it is logically impossible, such as negative values in the “Driver Commuter Distance” and “Annual Salary,” allowed us to observe more accurate descriptive statistic calculations. Finally, replacing inaccurate data using calculations based on domain knowledge and utilizing univariate analysis to treat null values ensures a higher quality dataset crucial to the success of the subsequent phases of the data life cycle.

**4. Discuss two or more limitations to your data cleaning approach specified in part C1.**

Even though this data-cleaning approach is essential for the success of further analysis, it can be highly time-consuming. When dealing with companies that have strict deadlines, this is something that we will have to consider as data professionals. Although it would not be beneficial to skip this step, working towards the most efficient data-cleaning approach would be ideal. Furthermore, even though univariate analysis was used to treat null values, this does not guarantee 100 percent accuracy. Even though it will likely be close to the actual values, there is a chance that some of them will be inaccurate because they are based on averages. Also, filling in the null values within the "Text Message Opt-In category with "Yes" affects the distribution of "Yes" vs "No" values.