Gaby Masak

D599 – Data Preparation and Exploration

Task 1: Data Cleaning and Exploration

9/21/2024

Data Cleaning Report

**Part 1: Data Profiling**

At the turn of the century, technology companies started to institute major changes in their workplaces. The likes of Google and Meta designed large campuses that were a far cry from the traditional cubicle filled, fluorescent lit office buildings. Instead of stale break rooms, tech companies offered their employees private chefs and large, free pantries. Cubicles became open floor plans with fresh and inspired décor. Even services like exercise classes and dry cleaning were offered, free of charge. These tech companies realized the importance of employee retention with highly skilled workers. As it costs much less to retain an employee than it is to recruit, onboard, and train a replacement, these companies analyzed turnover to subsequently design and utilize better employee retention strategies.

Before analysis can be performed and relevant strategies developed, data must first be inspected and cleaned. This process includes reviewing the data dictionary to glean an understanding of the general characteristics of the dataset, as well as the data types and subtypes of each variable. The data dictionary provided by the multinational tech company features 35 variables, each providing a unique insight, and representing individual columns in the dataset. The rows are determined by the number of employees. The data for each employee will occupy a single row. Individual fields correspond with singular data points.

Profiling the data dictionary provided provides insights into the 35 variables listed. Both Categorical and Numeric data types are represented. Subtypes of each, ordinal and nominal, and discrete and continuous, are respectively present and defined. This is represented in Table 1 below.

Table 1: Description of Variables from Dictionary

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Data Type** | **Data Subtype** | **Sample Observations** |
| Age | Numeric | Discrete | 18-65 |
| Turnover | Categorical | Nominal | Yes, No |
| BusinessTravel | Categorical | Nominal | Non-Travel, Travel\_Rarely, Travel\_Frequently |
| DailyRate | Numeric | Continuous | 100-1500 |
| Department | Categorical | Nominal | Sales, Human Resources, Support, Software, etc |
| DistanceFromHome | Numeric | Discrete | 1-50 |
| Education | Numeric | Discrete | 1-5 |
| EducationalField | Categorical | Nominal | Medical, Marketing, Life Sciences, Other, etc |
| EmployeeCount | Numeric | Discrete | 1 |
| EmployeeNumber | Numeric | Discrete | 1-9999 |
| EnvironmentSatisfaction | Numeric | Discrete | 1-5 |
| Gender | Categorical | Nominal | Male, Female |
| HourlyRate | Numeric | Continuous | 20-200 |
| JobInvolvement | Numeric | Discrete | 1-4 |
| JobLevel | Numeric | Discrete | 1-5 |
| JobRole | Categorical | Nominal | Developer, Sales Executive, Human Resources, etc |
| JobSatisfaction | Numeric | Discrete | 1-4 |
| Marital Status | Categorical | Nominal | Single, Married, Divorced |
| MonthlyIncome | Numeric | Continuous | 1001-50996 |
| MonthlyRate | Numeric | Continuous | 1270-1523280 |
| NumCompaniesWorked | Numeric | Discrete | 0-8 |
| Over18 | Categorical | Nominal | Yes, No |
| OverTime | Categorical | Nominal | Yes, No |
| PercentSalaryHike | Numeric | Continuous | 0-49 |
| PerformanceRating | Numeric | Discrete | 1-4 |
| RelationshipSatisfaction | Numeric | Discrete | 1-4 |
| StandardHours | Numeric | Continuous | 80 |
| StockOptionsLevel | Numeric | Discrete | 1-4 |
| TotalWorkingYears | Numeric | Discrete | 1-40 |
| TrainingTimesLastYear | Numeric | Discrete | 1-6 |
| WorkLifeBalance | Numeric | Discrete | 1-4 |
| YearsAtCompany | Numeric | Discrete | 1-40 |
| YearsInCurrentRole | Numeric | Discrete | 1-39 |
| YearsSinceLastPromotion | Numeric | Discrete | 1-38 |
| YearsWithCurrentManager | Numeric | Discrete | 1-38 |

**Part 2: Data Cleaning and Plan**

No dataset is impervious to errors. Whether the errors are due to human error and entry or systemic issues within the data collection process, it’s crucial to implement robust data validation and cleaning procedures to ensure data quality and reliability. The provided dataset was checked and cleaned for duplicate entries, missing values, inconsistent entries, formatting errors, and outliers. Python scripts were utilized to automate these processes, leveraging libraries such as Pandas and NumPy to efficiently handle and rectify these issues.

Duplicate entries are easily identified with the pandas function “duplicated”, where the entire dataset is processed row-by-row and identical rows are counted and stored. After passing the dataset through the function, 298 duplicated rows were identified of the 10322 original rows. Missing data can be identified with the pandas function “isnull”, where the computer scores the dataset to identify rows with null values. This function shows one null value in the age column, two in the education field column, three in the gender column, one in the monthly income column, two in the month rate column, one in the number of companies worked column, one in the total working years column, two in the years since last promotion column, and 418 missing values in the training times last year column.

Inconsistent entries represent data that is outside of the expected range of observations, as defined by the data dictionary. Identifying these errors was accomplished by sorting the respective columns by both ascending and descending order. Errors in the age column include ages 12, 15, and 148, two negative employee count values, a negative monthly income value, a negative value in total working years, and a negative value in the years with current manager value.

Formatting errors represent erroneous spacing, capitalization, numeric values for categorical variables, or vice versa, as well as erroneous characters or misspellings. The approach to identify these errors was similar to that of inconsistent entries, with manual sorting of the columns. Python and manual sorting were also used to find spacing, spelling, and capitalization errors. In the Employee Turnover Dataset, there are formatting errors in the categorical Business Travel column, where three numerical values appear.

Outliers are identified by calculating the Z-scores of the data, where data points that possess a Z-score of greater than 3 are considered to be outliers. In Python, this can be achieved by using the “zscore” function in the stat package on each column, then filtering for the resulting values that exceed the absolute value of 3. The resulting output shows that there are outliers in the Age, DistanceFromHome, EmployeeCount, MonthlyIncome, MonthlyRate, TotalWorkingYears, and YearsWithCurrManager columns.

After these quality issues are identified, they can be either removed or corrected. Duplicate entries can be automatically removed with the pandas function “drop\_duplicates” with Python. Missing values can be replaced with the replace function, adding stipulations what to replace these values within specific columns. Due to the low number of inconsistent and missing entries, most of these values could be replaced manually. For example, replacing the age number 12 with 52, as considering that most adults start their careers in their early 20s and this individual was listed as having 31 total working years. The same logic was applied to correct the individuals marked as a 15 and 96 years old. Reciprocally, the negative one value in TotalWorkingYears was corrected to be 31 with respect to the employee’s age of 55 years old. Manual replacements also fixed values like the outlier in MonthlyRate, where 872214 was entered twice as 872214872214 or YearsInCurrentRole as 222, which was switched to 22. Similar errors were identified and fixed in the DistanceFromHome column.

The logic behind filling in some missing or inconsistent data entries was more involved. Usually, communication with stakeholders would guide this process more, but the absence of stakeholders in this project highlights the need to elevate logic and critical thinking when cleaning data. There were four missing entries in the BusinessTravel column, which could be either Non-Travel, Travel\_Frequently, or Travel\_Rarely. Several parameters were accessed for a possible correlation. No immediate correlation was found between columns like Department or EducationalField, as shown by the pivot table in Figure 1 below. Similar issues arose with the JobRole column, with three missing entries. A pivot table shows that the distribution between categories is seemingly random. Analysis resulted in locating the job role that was most prevalent based on travel, department, and educational field to determine the job role most likely to be substituted. The same could be stated for the Education Field missing entries, however, the availability of the “other” option was the most valid, as it is an ambiguous category, and there is no discernable correlation between EducationField and other variables.

|  |  |
| --- | --- |
| **Row Labels** | **Sum of Education** |
|  | **2** |
| **Research & Development** | **2** |
| Human Resources | 2 |
| **Non-Travel** | **10085** |
| **Hardware** | **1794** |
| Human Resources | 287 |
| Life Sciences | 309 |
| Marketing | 263 |
| Medical | 263 |
| Other | 371 |
| Technical Degree | 301 |
| **Human Resources** | **1567** |
| Human Resources | 251 |
| Life Sciences | 222 |
| Marketing | 269 |
| Medical | 257 |
| Other | 294 |
| Technical Degree | 274 |
| **Research & Development** | **1747** |
| Human Resources | 264 |
| Life Sciences | 283 |
| Marketing | 289 |
| Medical | 312 |
| Other | 297 |
| Technical Degree | 300 |
| (blank) | 2 |
| **Sales** | **1661** |
| Human Resources | 271 |
| Life Sciences | 233 |
| Marketing | 292 |
| Medical | 283 |
| Other | 250 |
| Technical Degree | 332 |
| **Software** | **1594** |
| Human Resources | 237 |
| Life Sciences | 236 |
| Marketing | 268 |
| Medical | 269 |
| Other | 311 |
| Technical Degree | 273 |
| **Support** | **1722** |
| Human Resources | 242 |
| Life Sciences | 350 |
| Marketing | 281 |
| Medical | 313 |
| Other | 278 |
| Technical Degree | 258 |
| **Travel\_Frequently** | **10068** |
| **Hardware** | **1667** |
| Human Resources | 277 |
| Life Sciences | 288 |
| Marketing | 303 |
| Medical | 282 |
| Other | 249 |
| Technical Degree | 268 |
| **Human Resources** | **1604** |
| Human Resources | 271 |
| Life Sciences | 299 |
| Marketing | 301 |
| Medical | 245 |
| Other | 275 |
| Technical Degree | 213 |
| **Research & Development** | **1558** |
| Human Resources | 300 |
| Life Sciences | 260 |
| Marketing | 233 |
| Medical | 268 |
| Other | 244 |
| Technical Degree | 253 |
| **Sales** | **1734** |
| Human Resources | 258 |
| Life Sciences | 310 |
| Marketing | 359 |
| Medical | 296 |
| Other | 281 |
| Technical Degree | 230 |
| **Software** | **1719** |
| Human Resources | 278 |
| Life Sciences | 289 |
| Marketing | 310 |
| Medical | 309 |
| Other | 252 |
| Technical Degree | 281 |
| **Support** | **1786** |
| Human Resources | 308 |
| Life Sciences | 289 |
| Marketing | 349 |
| Medical | 300 |
| Other | 299 |
| Technical Degree | 241 |
| **Travel\_Rarely** | **10162** |
| **Hardware** | **1715** |
| Human Resources | 280 |
| Life Sciences | 281 |
| Marketing | 288 |
| Medical | 320 |
| Other | 258 |
| Technical Degree | 288 |
| **Human Resources** | **1666** |
| Human Resources | 301 |
| Life Sciences | 293 |
| Marketing | 209 |
| Medical | 314 |
| Other | 295 |
| Technical Degree | 254 |
| **Research & Development** | **1749** |
| Human Resources | 277 |
| Life Sciences | 221 |
| Marketing | 298 |
| Medical | 337 |
| Other | 298 |
| Technical Degree | 318 |
| **Sales** | **1618** |
| Human Resources | 233 |
| Life Sciences | 265 |
| Marketing | 332 |
| Medical | 264 |
| Other | 261 |
| Technical Degree | 263 |
| **Software** | **1741** |
| Human Resources | 268 |
| Life Sciences | 359 |
| Marketing | 269 |
| Medical | 319 |
| Other | 266 |
| Technical Degree | 260 |
| **Support** | **1673** |
| Human Resources | 311 |
| Life Sciences | 263 |
| Marketing | 298 |
| Medical | 328 |
| Other | 270 |
| Technical Degree | 203 |
| **(blank)** | **5** |
| **Hardware** | **4** |
| Human Resources | 2 |
| Technical Degree | 2 |
| **Software** | **1** |
| Medical | 1 |
| **Grand Total** | **30322** |

Figure 1: Pivot Table of Select Variables in Turnover Dataset

Python allowed for some automation as all entries in the EmployeeCount column were set to equal one, removing the negative and outlier values. Automation also allowed for the empty cells in the TrainingTimesLastYear column to be filled with 0. The same logic also allowed for the inconsistent categorical “na” entries in the YearsWithCurrentManager to be updated to numerical values, namely zero. Columns with only one acceptable value, like Over18 and EmployeeCount could also make sure this rule is applied with Python.

Using Python to automate the data cleaning process possess large advantages, especially when using large datasets. It would take a long time for a human to sift through each of the 10,000 columns, but a computer can do so within milliseconds. This efficiency, paired with versatility, represent the biggest advantages of using Python. Python’s extensive library ensures that there are always resources and tools available to tackle any data cleaning challenge. There are, however, several limitations that should be acknowledged. Automated processes are not infallible and errors could still be present. Ultimately, they do not completely mitigate the need for a human with domain knowledge to make decisions regarding potential errors and omissions. Cleaning data with Python also possesses limitations with handling extremely large datasets, where performance issues can arise, and the need for continuous monitoring and updating of scripts to adapt to new data structures and requirements.

**Part 3: Submission**

Annotated code used to detect and mitigate the data quality:

D599DataCleaning.py

Included in the file submission is a link to the cleaned dataset as a CSV file:

[D599 Cleaned.csv](https://1drv.ms/u/s!AsWS8nj845pg3SN-ed0oHAygGRpZ?e=WksZVw)

Also included is a link the audiovisual Panopto presentation of the functionality of the code used to clean the data and comments on the programming environment:

<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=fdd81bbb-d65c-4641-96d4-b1fa0117fe2f>