Data Cleaning and Profiling

Trayvonious Pendleton

WGU

Course Number: D599

STUDENT ID: 011205284

Part 1 Data Profiling

***A1 Dataset Description***

**General Characteristics:**

* **Rows**: 10,199
* **Columns**: 16
* The dataset includes employees’ demographics, work-related characteristics, and compensation details.

# *A2 Data Types and Subtypes*

|  |  |  |
| --- | --- | --- |
| Variable | Type | Subtype |
| Employee Number | Numerical | Discrete |
| Age | Numerical | Continuous |
| Tenure | Numerical | Continuous |
| Turnover | Categorical | Nominal |
| Compensation Type | Categorical | Nominal |
| Hourly Rate | Numerical | Continuous |
| Hours Weekly | Numerical | Continuous |
| Annual Salary | Numerical | Continuous |
| Driving Commuter Distance | Numerical | Continuous |
| Job Role | Categorical | Nominal |
| Gender | Categorical | Nominal |
| Marital Status | Categorical | Nominal |
| Number of Companies Worked | Numerical | Discrete |
| Annual Professional Development Hours | Numerical | Continuous |
| Paycheck Method | Categorical | Nominal |
| Text Message Opt-In | Categorical | Nominal |

## *A3 Observables Values*

|  |  |
| --- | --- |
| Variable | Sample Values |
| Employee Number | 1,2,3 |
| Age | 25,28,43 |
| Tenure | 1,5,10 |
| Turnover | Yes, No |
| Compensation Type | Salary, Hourly |
| Hourly Rate | $30.00, $25.00 |
| Hours Weekly | 25,40 |
| Annual Salary | 50689.6, 46841.6 |
| Driving Commuter Distance | 89,35,10 |
| Job Role | Research, Sales, Manufacturing |
| Gender | Male, Female, prefer not to answer |
| Marital Status | Single, Married |
| Number of Companies Worked | 1,3,6 |
| Annual Professional Development Hours | 89.35, 15 |
| Paycheck Method | Mail Check, Direct Deposit |
| Text Message Opt-In | Yes, No |

### PART 2 Data Cleaning and Plan

***B1 Data Cleaning and Inspection***

|  |  |
| --- | --- |
| Issue | Technique |
| Duplicate Entries | Use duplicated() to find rows that are completely identical |
| Missing Values | Use isnull() and sum() to identify columns and counts of missing values |
| Inconsistent Entries | Use unique() to check for inconsistent formats |
| Formatting Errors | Use contains()to validate formats |
| Outliers | Use statistical methouds to detect values that deviate significantly |

*B2 Findings*

|  |  |  |
| --- | --- | --- |
| Issue | Column | Description |
| Duplicate Entries | Multiple Columns | **Found:** 99 duplicate rows. Duplicate records may inflate employee counts and distort retention analysis. |
| Missing Values | Annual Professional Development Hours, etc**.** | 665 missing valuesin **NumCompaniesPreviouslyWorked.** 1,969 missing values in **AnnualProfessionalDevHrs.** 2,266 missing values in **TextMessageOptIn.** Impact: Missing HR data may affect predictive modeling and retention strategy insights. |
| Inconsistent Entries | Paycheck Method | **JobRoleArea:** Different text formats for the same category(e.g., "Information\_Technology" vs. "Information Technology", "Human\_Resources" vs "Human Resources"). **PaycheckMethod:** Multiple variations(e.g., "Mail Check", "MailedCheck", "Direct\_Deposit", "Direct Deposit").Impact: Inconsistencies can cause errors in grouping and data aggregation. |
| Formatting Errors | Hourly Rate | |  | | --- | |  |  |  | | --- | | **Detected:** Values contain dollar signs ($), spaces, and text characters (e.g., "$24.37 "). **Impact:** Prevents salary calculations and statistical analysis without conversion to numeric values. | |
| Outliers | Annual Salary | **544 extreme values** detected in **AnnualSalary**. **245 extreme values** detected in **DrivingCommuterDistance**. **Impact:** Some values may be valid high salaries or long commutes, but they could also indicate incorrect data entry (e.g., misplaced decimal points). |

C1 Explanation of Modifications

1. **Handling Duplicate Entries**

|  |  |  |
| --- | --- | --- |
| Issue | Modification Applied | Reason |
| 99 duplicate rows | df.drop\_duplicates(inplace=True) | |  | | --- | |  |  |  | | --- | | Prevents inflated employee counts and ensures each record is unique. | |

1. **Handling Missing Values**

|  |  |  |  |
| --- | --- | --- | --- |
| Column | Missing Values | Modification Applied | Reason |
| NumCompaniesPreviouslyWorked | 665 | Replace with median value | Since employees typically work for a whole number of companies, median imputation is preferred over mean. |
| AnnualProfessionalDevHrs | 1969 | Replace with mean value | Employees may take various hours of training, so using the mean maintains a realistic distribution. |
| TextMessageOptIn | 2266 | Replace with “No” | Assumes missing values indicate that employees did not opt in for text messages. |

1. **Fixing Inconsistent Entries**

|  |  |  |  |
| --- | --- | --- | --- |
| Column | Inconsistent Values | Modification Applied | Reason |
| JobRoleArea | "Information\_Technology" vs "Information Technology" | Standardized by replacing underscores with spaces | Ensures uniform grouping. |
| JobRoleArea | "Human\_Resources" vs "Human Resources" | Standardized to "Human Resources" | Prevents category mismatches. |
| PaycheckMethod | "Mail Check" vs "MailedCheck" | Unified to "Mailed Check" | Maintains consistency in paycheck categories. |

1. **Fixing Formatting Errors**

|  |  |  |  |
| --- | --- | --- | --- |
| Column | Issue | Modification Applied | Reason |
| HourlyRate | Contains dollar signs ($) and spaces | Removes $ and convert to float | Ensures the column can be used for numeric analysis |

1. **Handling Outliers**

|  |  |  |  |
| --- | --- | --- | --- |
| Column | Outliers Detected | Modification Applied | Reason |
| AnnualSalary | 544 extreme values | Applied Interquartile Range (IQR) method to cap extreme values | Prevents unrealistic salary values. |
| DrivingCommuterDistance | 245 extreme values | Applied Interquartile Range (IQR) method to cap extreme values | Prevents errors from extreme commuting distances. |

***C2 Justification for Data Cleaning Techniques***

1. **Handling Duplicate Entries**

|  |  |  |
| --- | --- | --- |
| Issue | Technique Use | Justification |
| 99 duplicate rows | dr.drop\_duplicates() | Removing duplicates prevents inflated employee counts and ensures that turnover and retention metrics are accurate. Keeping duplicates could lead to incorrect workforce analytics and biased machine learning models. |

1. **Handling Missing Values**

|  |  |  |  |
| --- | --- | --- | --- |
| Column | Missing Values | Technique Used | Justification |
| NumCompaniesPreviouslyWorked | 665 | df.loc[:, 'NumCompaniesPreviouslyWorked'] = df['NumCompaniesPreviouslyWorked'].fillna(df['NumCompaniesPreviouslyWorked'].median()) | |  | | --- | |  |  |  | | --- | | Median imputation is used because the number of previous employers is a discrete variable and may contain extreme values. The median is less affected by outliers compared to the mean. | |
| AnnualProfessionalDevHrs | 1969 | df.loc[:, 'AnnualProfessionalDevHrs'] = df['AnnualProfessionalDevHrs'].fillna(df['AnnualProfessionalDevHrs'].mean()) | Mean imputation is used because professional development hours are typically continuous and normally distributed. This helps maintain an average training record per employee. |
| TextMessageOptIn | 2266 | df.loc[:, 'TextMessageOptIn'] = df['TextMessageOptIn'].fillna('No') | |  | | --- | |  |  |  | | --- | | Assumes that missing values indicate employees did not opt in for text messaging. This prevents unnecessary null handling in further analysis. | |

1. **Fixing Inconsistent Entries**

|  |  |  |  |
| --- | --- | --- | --- |
| Column | Inconsistencies Detected | Technique Used | Justification |
| JobRoleArea | "Information\_Technology" vs "Information Technology" "Human\_Resources" vs "Human Resources" | df.loc[:, 'JobRoleArea'] = df['JobRoleArea'].replace({...}) | |  | | --- | |  |  |  | | --- | | Standardizing these names ensures consistent grouping during analysis and prevents duplicate categories. | |
| PaycheckMethod | "Mail Check" vs "Mailed Check" vs "MailedCheck" | |  | | --- | |  |  |  | | --- | | df.loc[:, 'PaycheckMethod'] = df['PaycheckMethod'].replace({...}) | | This ensures that employees are **categorized correctly** for payroll processing. |

**4. Fixing Formatting Errors**

|  |  |  |  |
| --- | --- | --- | --- |
| Column | Issue | Technique Used | Justification |
| HourlyRate | Values contain dollar signs ($), spaces, and text | df.loc[:, 'HourlyRate'] = df['HourlyRate'].astype(str).str.replace(r'[\$ ]', '', regex=True).astype(float) | Converting text-based numbers to float allows correct salary calculations and prevents errors in financial modeling. |

1. **Handling Outliers**

|  |  |  |  |
| --- | --- | --- | --- |
| Column | Outliers Detected | Technique Used | Justification |
| AnnualSalary | 544 extreme values | df.loc[:, 'AnnualSalary'] = df['AnnualSalary'].clip(lower=lower\_bound, upper=upper\_bound) | |  | | --- | |  |  |  | | --- | | IQR (Interquartile Range) capping ensures extreme salaries do not skew salary trend analysis. This prevents misinterpretation of compensation trends. | |
| DrivingCommuterDistanceDrivingCommuterDistance | 245 extreme values | df.loc[:, 'DrivingCommuterDistance'] = df['DrivingCommuterDistance'].clip(lower=lower\_bound, upper=upper\_bound) | Long commutes may be valid but extreme values (e.g., 500 miles) suggest data entry errors. Using the IQR method prevents distortions in commute distance analysis. |

***C3 Advantages of the Data Cleaning Approach***

1. **Removing Duplicate Entries**

|  |  |
| --- | --- |
| Technique Used | Advantages |
| df.drop\_duplicates() | Prevents duplicate employee records, ensuring an accurate headcount.  Reduces data redundancy, leading to faster computation and better storage efficiency. Avoids overestimation of turnover rates, which could distort retention analysis. |

1. **Handling Missing Values**

|  |  |  |
| --- | --- | --- |
| Column | Technique Used | Advantages |
| NumCompaniesPreviouslyWorked | Median imputation | Handles skewed data without being influenced by outliers. Preserves the data distribution, making predictive models more reliable. |
| AnnualProfessionalDevHrs | Mean imputation | Maintains a realistic average of training hours, ensuring that training statistics remain consistent. Reduces information loss while keeping the dataset usable for modeling. |
| TextMessageOptIn | Mode imputation | Assumes a logical default value, reducing the impact of missing values on SMS communication analysis. Prevents null-related errors in categorical analysis and machine learning models. |

1. **Fixing Inconsistent Entries**

|  |  |  |
| --- | --- | --- |
| Column | Technique Used | Advantages |
| JobRoleArea | |  | | --- | |  |  |  | | --- | | Standardizing category names | | Ensures consistent grouping of job roles, improving data aggregation. Prevents category duplication (e.g., "Information Technology" and "Information\_Technology" being treated as separate roles). |
| PaycheckMethod | |  | | --- | |  |  |  | | --- | | Standardizing category names | | Reduces categorical noise, ensuring accurate payroll processing insights. Helps HR teams generate clear payroll reports without misclassified categories. |

**4. Fixing Formatting Errors**

|  |  |  |
| --- | --- | --- |
| Column | Technique Used | Advantages |
| HourlyRate | Removing $ symbols and spaces, converting to numeri | Ensures correct salary calculations, preventing errors in financial analysis. Converts the column into a numerical format, making it compatible with machine learning models. |

1. **Handling Outliers**

|  |  |  |
| --- | --- | --- |
| Column | Technique Used | Advantages |
| AnnualSalary | Applying the IQR capping method | Prevents skewed salary trends, ensuring realistic compensation analysis. Avoids incorrect assumptions about salaries that could influence budget planning and HR decisions. |
| DrivingCommuterDistance | Applying the IQR capping method | Prevents misleading commuting patterns, ensuring realistic travel distance insights. Avoids extreme values affecting transportation and location-based HR policies. |

***C4 Limitations of the Data Cleaning Approach***

1. **Removing Duplicate Entries**

|  |  |
| --- | --- |
| Technique Used | Limitations |
| df.drop\_duplicates() | If duplicate records contain minor differences (e.g., salary updates or job role changes), valid data might be removed. Some employees may have legitimate duplicate records (e.g., rehired employees appearing multiple times). |

1. **Handling Missing Values**

|  |  |  |
| --- | --- | --- |
| Column | Technique Used | Limitations |
| NumCompaniesPreviouslyWorked | Median Imputation | Does not preserve real work history—assumes that most employees had a similar number of past jobs, which may not be true. |
| AnnualProfessionalDevHrs | Mean Imputation | The mean is affected by outliers—if a few employees report extremely high training hours, it skews the imputed values. |
| TextMessageOptIn | Mode Imputation | Assumes missing values mean "No", which may not always be true—some employees might have forgotten to opt in. |

1. **Fixing Inconsistent Entries**

|  |  |  |
| --- | --- | --- |
| Column | Technique Used | Limitations |
| JobRoleArea | Standardizing names | Manually defining replacements may not cover all possible variations—new inconsistencies might appear later. |
| PaycheckMethod | Standardizing categories | Assumes all variations refer to the same method, which may not always be true. |

1. **Fixing Formatting Errors**

|  |  |  |
| --- | --- | --- |
| Column | Technique Used | Limitations |
| HourlyRate | Removing $ symbols and spaces, converting to float | If some values were entered incorrectly (e.g., missing decimal points), converting them to float does not correct the error. |

1. **Handling Outliers**

|  |  |  |
| --- | --- | --- |
| Column | Technique Used | Limitations |
| AnnualSalary | IQR Capping | This method assumes all extreme values are errors—but some high salaries may be valid compensation packages. |
| DrivingCommuterDistance | IQR Capping | Long commute distances might be real for remote employees, so treating them as outliers may remove valid data. |

**D1 Data Cleaning Report**

This is my data cleaning report.