**Nicole Gallo**

**D599 – Data Preparation and Exploration**

**June 26, 2025**

**TCN1 Task 1: Data Cleaning and Profiling**

**Part I: Data Profiling**

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 dataset**

The Employee Turnover Dataset contains 10,199 rows and 16 columns of data. Each row represents an individual employee. Other important data attributes include demographic, employment, and compensation (type and annual salary). This data is used to understand the “employee turnover” within the company.

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

Below is a list of the variables in the dataset with their corresponding data type and subtype:

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Data Type** | **Subtype** |
| EmployeeNumber | Integer | Nominal (ID) |
| Age | Integer | Discrete Numeric |
| Tenure | Integer | Discrete Numeric |
| Turnover | Object | Nominal (Categorical - Yes/No) |
| HourlyRate | Float | Continuous Numeric |
| HoursWeekly | Integer | Discrete Numeric |
| CompensationType | Object | Nominal (Categorical) |
| AnnualSalary | Float | Continuous Numeric |
| DrivingCommuterDistance | Integer | Discrete Numeric |
| JobRoleArea | Object | Nominal (Categorical) |
| Gender | Object | Nominal (Categorical) |
| MaritalStatus | Object | Nominal (Categorical) |
| NumCompaniesPreviouslyWorked | Float | Discrete Numeric |
| AnnualProfessionalDevHrs | Float | Continuous Numeric |
| PaycheckMethod | Object | Nominal (Categorical) |
| TextMessageOptIn | Object | Nominal (Categorical - Yes/No) |

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

A sample for each variable:

* EmployeeNumber: 5270
* Age: 47
* Tenure: 15
* Turnover: 'No'
* HourlyRate: 41.07
* HoursWeekly: 40
* CompensationType: 'Salary'
* AnnualSalary: 85425.6
* DrivingCommuterDistance: 12
* JobRoleArea: 'Human Resources'
* Gender: 'Male'
* MaritalStatus: 'Divorced'
* NumCompaniesPreviouslyWorked: 8.0
* AnnualProfessionalDevHrs: 11.0
* PaycheckMethod: 'Mail Check'
* TextMessageOptIn: 'Yes'

**Part II: Data Cleaning and Plan**

1. Inspect the dataset through data cleaning techniques
2. Explain how you inspected the dataset for each of the quality issues listed in part B.

**Duplications** – used df.duplicated().sum() to find the number of duplicate rows

**Missing values** – used df.isnull().sum() to find the number of missing values

**Formatting errors** – standardized text fields using Python to remove any unnecessary symbols and space characters

**Inaccurate data** – reviewed unique values in object columns; used Python to confirm inaccurate data, such as negative values *df['DrivingCommuterDistance'] < 0*

**Outliers** – used IQR (interquartile range) method to calculate upper and lower bounds; values outside of 1.5xIQR were flagged as outliers

1. List your findings for each quality issue listed in part B.

**Duplications** – 99 duplicate rows were found

A screenshot of a computer code

AI-generated content may be incorrect.

**Missing values** – NumCompaniesPreviouslyWorked (665), AnnualProfessionalDevHrs (1969), TextMessageOptIn (2266)

A screenshot of a computer

AI-generated content may be incorrect.

**Formatting errors** – HourlyRate contained formatting issues ($ and extra whitespace) as a currency number data type – data type is showing HourlyRate as ‘object’ instead of ‘float’ because of the $ in the entries

A screenshot of a computer

AI-generated content may be incorrect.

**Inaccurate data** – reviewed unique values in object columns; used Python to confirm inaccurate data, such as negative values *df['DrivingCommuterDistance'] < 0*

* Negative Commuting Distances – e.g., *df['DrivingCommuterDistance'] < 0*

A screenshot of a computer code

AI-generated content may be incorrect.

A table with numbers and numbers

AI-generated content may be incorrect.

* Inconsistent naming convention - PaycheckMethod “Mail Check” and “Direct Deposit”, and JobRoleArea “Human Resources” and “Information Technology”

A screenshot of a computer program

AI-generated content may be incorrect.

* Miscalculation of AnnualSalary, showing negative salary

A screenshot of a computer code

AI-generated content may be incorrect.

**Outliers** – using IQR (interquartile range) to find outliers in DrivingCommuterDistance

e.g., *lower\_bound = Q1 - 1.5 \* IQR*

*upper\_bound = Q3 + 1.5 \* IQR*

A screenshot of a computer

AI-generated content may be incorrect.

1. Discuss which data cleaning techniques you used
2. Describe how you modified the dataset after identifying each quality issue listed in part B.

**Duplications** – used df.drop\_duplicates() to remove 99 duplicated rows

**A screenshot of a computer program

AI-generated content may be incorrect.**

**Missing values** – used df.isnull().sum() to find the number of missing values

Before:

A screenshot of a computer

AI-generated content may be incorrect.

After (NumCompaniesPreviouslyWorked):

A screenshot of a computer

AI-generated content may be incorrect.

After (AnnualProfessionalDevHrs):

A screenshot of a computer

AI-generated content may be incorrect.

After (TextMessageOptIn):

A screenshot of a computer

AI-generated content may be incorrect.

**Formatting errors** – HourlyRate contained formatting issues ($ and extra whitespace) as a currency number data type – e.g., *df.columns = df.columns.str.strip()*

Before:

A screenshot of a computer

AI-generated content may be incorrect.

After:

A screenshot of a computer program

AI-generated content may be incorrect.

**Inaccurate data** – negative commuting distances, PaycheckMethod and JobRoleArea showing the same entry but different version of the word, AnnualSalary not matching HourlyRate x HoursWeekly x 52

* Negative Commuting Distances – e.g., *df['DrivingCommuterDistance'] < 0*

A screenshot of a computer code

AI-generated content may be incorrect.

* Standardizing Paycheck Method – Inconsistencies between “Mail Check” and “Direct Deposit”

e.g., *df['PaycheckMethod'] = df['PaycheckMethod'].replace({...})*

A screenshot of a computer

AI-generated content may be incorrect.

* Standardizing Job Role Area – Inconsistencies between “Human Resources” and “Information Technology”

*df['JobRoleArea'] = df['JobRoleArea'].replace({*

*'Humanresources': 'Human Resources',...*

*})*

A screenshot of a computer

AI-generated content may be incorrect.

* AnnualSalary not matching HourlyRate x HoursWeekly x 52

e.g., *df.loc[hourly\_mask, 'AnnualSalary'] = df.loc[hourly\_mask, 'HourlyRate'] \* df.loc[hourly\_mask, 'HoursWeekly'] \* 52*

A screenshot of a computer

AI-generated content may be incorrect.

**Outliers** – removing outliers with outlier capping using *.clip()*

A screenshot of a computer program

AI-generated content may be incorrect.

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

These specific data cleaning techniques were used to ensure data accuracy and consistency for better analysis:

* Removing duplicates, .drop\_duplicates(), eliminates redundancy that could skew analysis
* Median imputation is less affected by outliers and keeps distributions intact and mode imputation maintains consistency for categorical fields, according to Syed Burhan Ahmed in [When to Use Mean, Median, and Mode for Handling Missing Values in Data?](https://www.linkedin.com/pulse/when-use-mean-median-mode-handling-missing-values-data-ahmed-tebje/)
* Text and formatting standardization improves grouping initatives and any future visualization needs.
* Recalculating AnnualSalary ensures available employee data is accurate, consistent and up-to-date
* To handle outliers in the DrivingCommuterDistance column, I used the Interquartile Range (IQR) method to identify any extreme values. Instead of removing these rows from the dataset, I chose to apply outlier capping or clipping (also known as Winsorizing), which caps values above the upper threshold. This method preserves all rows and avoids reducing the dataset size.

1. Describe **two** or more advantages to your data cleaning approach specified in part C1.
2. It preserves the dataset size by imputing or capping rather than deleting records altogether.
3. It improves consistency across categorical and standardized fields.
4. The logical corrections ensure that data reflects realistic values without the need for manual editing.
5. Discuss **two** or more limitations to your data cleaning approach specified in part C1.
6. Annual salary recalculation will rely on other values being accurate in order to produce an accurate calculation.
7. **Outlier capping may hide extremely high, yet accurate values; for example, long commuter distances or high salaries.**

**Part III: Submission**

D. Submit your findings by doing the following:

1.  Provide a data cleaning report as a document file that includes responses to task prompts. **NicoleGallo\_TCN1 Task 1 Data Cleaning and Profiling\_attempt2.docx**

2.  Provide the annotated code you used to detect and mitigate the data quality as an executable script file. R files and Python script files are accepted. **NicoleGallo\_D599\_Task1\_attempt3.py**

3.  Provide a copy of the cleaned dataset as a CSV file. **NicoleGallo\_Employee\_Turnover\_Cleaned.csv**

4.  Panopto video: <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=ba031a08-2f2b-4ba3-9f3b-b30e00f5f9e8>

**Sources**

E. The only external source I used beyond WGU resources was the following to learn more about imputation:

Ahmed, S. B. (n.d.). When to use mean, median, and mode for handling missing values in data? LinkedIn. <https://www.linkedin.com/pulse/when-use-mean-median-mode-handling-missing-values-data-ahmed-tebje/>