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

1. **Rows and Columns**: The dataset includes data for each employee, with rows representing individual employees and columns representing different attributes of those employees. The total number of rows would correspond to the number of employees, which isn't specified in the document but should be evident in the dataset. (WGU D599 Task 1, 2024)
2. **Columns**: There are 16 distinct attributes (columns) defined in the data dictionary.
3. **Rows:** 10,199
    - **Data Completeness**:
        i. The dataset has missing values in several columns:
            1. **NumCompaniesPreviouslyWorked**: 9,534 non-null values (665 missing).
            2. **AnnualProfessionalDevHrs**: 8,230 non-null values (1,969 missing).
            3. **TextMessageOptIn**: 7,933 non-null values (2,266 missing).
    - **Data Types**:
        i. The column **HourlyRate** appears to be stored as an object (string) due to the presence of a dollar sign, which might need conversion to a numeric type for analysis.
        ii. Categorical columns include **CompensationType**, **JobRoleArea**, **Gender**, **MaritalStatus**, and **PaycheckMethod**.
    - **Turnover**:
        i. **Turnover** is a binary categorical field indicating if an employee has left the company. Its values are "Yes" or "No."
    - **Salary Insights**:
        i. **AnnualSalary** has a wide range of values, with some potential outliers like very high salaries.
    - **Potential Derived Data**:
        i. **AnnualSalary** might be computed based on **HourlyRate** and **HoursWeekly**, but this would require validation.
        ii. The relationship between **Turnover** and attributes like **CompensationType** or **JobRoleArea** could provide interesting insights.
    - **Outliers**:
        i. **DrivingCommuterDistance** has high values (e.g., 89), which may represent employees commuting significant distances or potential outliers.
        ii. **HourlyRate** includes values like $88.77, which are unusually high compared to typical hourly wages.
    - **Categorical Distribution**:
        i. **Gender** includes "Prefer Not to Answer," indicating inclusivity in data collection.
        ii. **JobRoleArea** includes a variety of roles such as "Research," "Information_Technology," and "Sales."

Next Steps:

1. **Data Cleaning**:
    - Convert **HourlyRate** to numeric after removing the dollar sign.
    - Handle missing values in columns like **NumCompaniesPreviouslyWorked**, **AnnualProfessionalDevHrs**, and **TextMessageOptIn**.
2. **Exploratory Data Analysis (EDA)**:
    - Investigate relationships between **Turnover** and other factors, such as **Age**, **Tenure**, and **JobRoleArea**.
    - Visualize distributions for numeric columns to identify trends and outliers.
3. **Hypothesis Testing**:
    - Determine if **Tenure** and **Professional Development Hours** significantly impact turnover rates.
    -

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

Variable	Data Type	Subtype	Notes
EmployeeNumber	Categorical	Nominal	A unique numeric identifier for each employee.
Age	Numeric	Discrete	Represents the employee's age in years
Tenure	Numeric	Discrete	Number of years the employee has worked at the company
Turnover	Categorical	Binary	Indicates whether the employee has left the company.
HourlyRate	Numeric	Continuous	Stored as text due to the dollar sign, needs conversion to float for analysis
HoursWeekly	Numeric	Discrete	Weekly hours worked.
CompensationType	Categorical	Nominal (Categorical)	Indicates whether the employee is salaried or hourly
AnnualSalary	Numeric	Continuous	Annual salary value, likely derived for hourly employees
DrivingCommuterDistance	Numeric	Continuous	Distance from home to work in miles.

JobRoleArea	Categorical	Nominal (categorical)	Job area of employee
Gender	Categorical	Nominal (categorical)	Includes categories like “Male,” “Female,” and “Prefer Not to Answer.”
MaritalStatus	Categorical	Nominal (Categorical)	Marital status of the employee
NumCompaniesPreviously Worked	Numeric	Discrete	Includes decimal values despite representing discrete entities; has missing values
PayCheckMethod	Categorical	Nominal (categorical)	How the employee receives their paycheck
TextMessageOptIn	Categorical	Binary (Yes / No)	Indicates if the employee opted in for text messages; includes missing values
AnnualProfessionalDevHrs	Numeric	Continuous	Total hours spent on professional development; has missing values

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

1. **Employee Number**
    - Example Values: `1, 2, 3`
2. **Age**
    - Example Values: `25`, `32`, `45`
3. **Tenure**
    - Example Values: `1.5`, `3.0`, `10.2`
4. **Turnover**
    - Example Values: `Yes`, `No`
5. **Compensation Type**
    - Example Values: `Salaried`, `Hourly`
6. **Hourly Rate**
    - Example Values: `15.50`, `20.00`, `45.75`
7. **Hours Weekly**
    - Example Values: `20`, `40`, `60`
8. **Annual Salary**
    - Example Values: `32000`, `52000`, `85000`
9. **Driving Commuter Distance**
    - Example Values: `5.0`, `15.2`, `30.0`
10. **Job Role**

    - Example Values: `Software Engineer`, `Data Analyst`, `Project Manager`
11. **Gender**
    - Example Values: `Male`, `Female`, `Prefer not to answer`
12. **Marital Status**
    - Example Values: `Single`, `Married`, `Divorced`
13. **Number of Companies Worked**
    - Example Values: `0`, `2`, `5`
14. **Annual Professional Development Hours**
    - Example Values: `10.5`, `40.0`, `80.0`
15. **Paycheck Method**
    - Example Values: `Direct Deposit`, `Check`
16. **Text Message Opt-In**
    - Example Values: `Yes`, `No`

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

1. Duplicate Entries

- **Finding**: The dataset contains **99 duplicate rows**. These need to be removed to ensure the integrity of the analysis. (WGU D599 Task 1, 2024)

2. Missing Values

- **Columns with Missing Values**:
    - `NumCompaniesPreviouslyWorked`: **665 missing values** (~6.5% of the dataset).
    - `AnnualProfessionalDevHrs`: **1969 missing values** (~19% of the dataset).
    - `TextMessageOptIn`: **2266 missing values** (~22% of the dataset).
- **Impact**: Missing values in these columns could affect any analysis involving them. Imputation or exclusion strategies should be applied.

3. Inconsistent Entries

- **Findings**:
    - `PaycheckMethod`:
        - Variants like `Mail Check`, `Mailed Check`, `Direct_Deposit`, `DirectDeposit`, etc., suggest inconsistent formatting.
    - `JobRoleArea`:
        - Variants like `Information_Technology`, `InformationTechnology`, and `Information Technology` are inconsistent.
    - Other categorical columns like `Gender` and `MaritalStatus` are consistent.

4. Formatting Errors

- **Findings**:
  - `HourlyRate`: This value is stored as a string because it contains a dollar sign (`$`). It needs to be converted to numeric for analysis.
  - `AnnualSalary`: Contains negative values that are invalid and require investigation.
  - `DrivingCommuterDistance`: Contains negative values that are nonsensical and need correction.

5. Outliers

- **Findings**:
  - `AnnualSalary`:
    - Maximum value: 339,950.4 (appears valid for high-level roles).
    - Minimum value: -33,326.4 (invalid and requires investigation).
  - `DrivingCommuterDistance`:
    - Maximum value: 950 miles (plausible but unusual).
    - Minimum value: -275 miles (invalid).

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.

1. Duplicate Entries:

- **Issue**: Duplicate entries inflate the dataset size and may bias the analysis.
- **Action**: Used `drop_duplicates()` to remove any duplicate rows.
- **Rationale**: Ensures each employee is represented only once, maintaining the integrity of the dataset.

2. Missing Values:

- **NumCompaniesPreviouslyWorked**:
  - **Action**: Filled missing values with the median.
  - **Rationale**: The median is robust to outliers and accurately reflects the central tendency of discrete values.
- **AnnualProfessionalDevHrs**:
  - **Action**: Filled missing values with the mean.
  - **Rationale**: Professional development hours are continuous data, and the mean provides a balanced estimate when missing values are evenly distributed.
- **TextMessageOptIn**:
  - **Action**: Filled missing values with the mode.
  - **Rationale**: This is binary data (Yes/No), and the mode reflects the most common value, minimizing distortion.

3. Inconsistent Entries:

- **PaycheckMethod**:
  - **Issue**: Entries like "Mail Check" and "Mailed Check" represent the same method but appear as distinct categories.
  - **Action**: Standardized to "Mail Check."
  - **Rationale**: Ensures uniformity for accurate analysis and prevents unnecessary category splitting.
- **JobRoleArea**:
  - **Action**: Normalized variations like "Information_Technology" to "Information Technology."
  - **Rationale**: Consistent formatting avoids errors in categorical analysis.

4. Formatting Errors:

- **HourlyRate**:
  - **Issue**: The column included dollar signs and was stored as text.
  - **Action**: Removed dollar signs and converted the column to numeric.
  - **Rationale**: Ensures the column is suitable for numeric calculations, such as computing mean or aggregations.

5. Outliers:

- **AnnualSalary** and **DrivingCommuterDistance**:
  - **Issue**: Outliers such as negative values are nonsensical and could skew analysis.
  - **Action**: Removed rows with negative values in these columns.
  - **Rationale**: Negative salaries or distances are invalid and removing these entries prevents their impact on statistical measures.

6. Save Cleaned Data:

- **Action**: Saved the cleaned dataset to a new file for further analysis.
- **Rationale**: Preserves the original data while ensuring future analysis uses a corrected, reliable dataset.

2. Discuss why you chose the specific data cleaning techniques to clean the quality issues listed in Part B. (WGU D599 Task 1, 2024)

1. **Duplicate Entries**
   - **Technique Used**: Removed duplicates using `drop_duplicates()`.
   - **Reason**: Duplicate records can skew analyses by inflating counts or introducing redundancy. Removing them ensures that each record represents a unique employee.
2. **Missing Values**

- `NumCompaniesPreviouslyWorked`: Replaced missing values with the **median**.
  - **Reason**: The median is robust against outliers and represents a typical value for this numeric field.
- `AnnualProfessionalDevHrs`: Replaced missing values with the **mean**.
  - **Reason**: The mean reflects the overall average time spent on professional development.
- `TextMessageOptIn`: Replaced missing values with the **mode**.
  - **Reason**: The mode captures the most frequent response, aligning with general trends when actual values are unavailable.

3. **Inconsistent Entries**
   - **Standardized `PaycheckMethod` and `JobRoleArea` values** using `replace()`.
     - **Reason**: Normalizing these values ensures consistency, preventing issues in grouping or filtering during analysis.
4. **Formatting Errors**
   - **I removed $ from HourlyRate and converted it** to numeric.
     - **Reason**: This conversion ensures the column can be used in numerical analyses, such as averages or correlations.
   - **Corrected Column Names**: Trimmed whitespace from column names.
     - **Reason**: Consistent column names prevent errors in referencing fields programmatically.
5. **Outliers**
   - **Removed rows with negative values** in `AnnualSalary` and `DrivingCommuterDistance`.
     - **Reason**: Negative values are invalid in these contexts and likely due to data entry errors. Removing them ensures the integrity of the data.
6. **Final Output**
   - **Saved the cleaned dataset** to a new file.
     - **Reason**: Preserving the cleaned version separately ensures reproducibility and prevents accidental overwriting of the raw data.

3. Describe two or more advantages to your data cleaning approach specified in part C1. (WGU D599 Task 1, 2024)

Advantages of the Data Cleaning Approach

1. **Systematic and Reproducible Process**:
   - By breaking the cleaning into structured steps (duplicates, missing values, inconsistencies, formatting errors, and outliers), the approach is easy to follow and replicate.
   - Saving the cleaned dataset ensures the raw data remains untouched, maintaining data provenance.
2. **Targeted Handling of Issues**:
   - Each quality issue is addressed using appropriate techniques:
     - **Duplicate removal ensures that** no redundant records skew the analysis.

        - **Missing Value Imputation** is tailored to the nature of each column (e.g., median for numerical history, mode for categorical preferences).
        - **Categorical Standardization** ensures consistent groupings and reduces errors in downstream analysis.
    - These targeted methods preserve as much data as possible, minimizing unnecessary information loss.

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

    1. **Assumptions in Missing Value Imputation**:
        - **Limitation**: Inputing missing values using the median, mean, or mode assumes that these represent the best estimates for the missing data. This might not always reflect the actual values, potentially introducing bias.
        - **Impact**: For example, filling `NumCompaniesPreviouslyWorked` with the median might overlook unique work histories or trends in the data.
    2. **Outlier Handling by Removal**:
        - **Limitation**: Simply removing rows with negative values in `AnnualSalary` and `DrivingCommuterDistance` could discard potentially valuable records that might have been incorrectly entered or need further investigation.
        - **Impact**: While this ensures clean data, it also reduces the dataset size, which could be significant if many outliers exist.

WGU D599 Task 1. (2024). *Employee Turnover Dataset* [Dataset]. Western Governors University. https://tasks.wgu.edu/student/012474228/course/33280017/task/4435/overview