**Task 1: Data Cleaning and Profiling**

Krescens Kok

Department of Technology, Western Governors University

D599: Data Preparation and Exploration

Keiona Middleton

2025 September 6

**Part I: Data Profiling**

**A.  Review the data dictionary in the attached "Employee Turnover Considerations and Dictionary" document and provide profile data by doing the following within a document:**

**1.  Identify the dataset’s number of records (i.e., rows) and number of variables (i.e., columns).**

The dataset, Employee Turnover Dataset, is a csv file containing 10,199 rows and 16 variables/columns. The dataset was collected in hopes of identifying why employees may be leaving in order to design smarter employee retention strategies.

**2.  List each variable and indicate the variable’s data type (quantitative/numerical or qualitative/categorical) and data subtype (i.e., continuous/discrete or nominal/ordinal).**

Employee Number:

* Data Type: Qualitative/categorical
* Data Subtype: Nominal

Age:

* Data Type: Quantitative/numerical
* Data Subtype: Discrete

Tenure:

* Data Type: Quantitative/numerical
* Data Subtype: Discrete

Turnover:

* Data Type: Qualitative/categorical
* Data Subtype: Nominal

Hourly Rate:

* Data Type: Quantitative/numerical
* Data Subtype: Continuous

Hours Weekly:

* Data Type: Quantitative/numerical
* Data Subtype: Discrete

Compensation Type:

* Data Type: Qualitative/categorical
* Data Subtype: Nominal

Annual Salary:

* Data Type: Quantitative/numerical
* Data Subtype: Continuous

Driving Commuter Distance:

* Data Type: Quantitative /numerical
* Data Subtype: Continuous

Job Role:

* Data Type: Qualitative/categorical
* Data Subtype: Nominal

Gender:

* Data Type: Qualitative/categorical
* Data Subtype: Nominal

Marital Status:

* Data Type: Qualitative/categorical
* Data Subtype: Nominal

Number of Companies Worked:

* Data Type: Quantitative/numerical
* Data Subtype: Discrete

Annual Professional Development Hours:

* Data Type: Quantitative/numerical
* Data Subtype: Discrete

Paycheck Method:

* Data Type: Qualitative/categorical
* Data Subtype: Nominal

Text Message Opt-In:

* Data Type: Qualitative/categorical
* Data Subtype: Nominal

**3.  Identify a sample of observable values for each variable.**

Employee Number: 1,2,3

Age: 28,33,22

Tenure: 6,2,1

Turnover: Yes, No

Hourly Rate: $24.37, $22.52, $88.77

Hours Weekly: 40

Compensation Type: Salary

Annual Salary: 50689.6, 61443.2, 284080

Driving Commuter Distance: 89, 35, 12

Job Role: Research, Information\_Technology, Sales

Gender: Female, Male, Prefer Not to Answer

Marital Status: Married, Single, Divorced

Number of Companies Worked: 3,6,1

Annual Professional Development Hours: 7,8,19

Paycheck Method: Mail Check, Mailed Check, Direct\_Deposit

Text Message Opt-In: Yes, N/A, No

**Part II: Data Cleaning and Plan**

**B.  Using Python or R code data cleaning techniques do the following:**

**1.  Explain how you inspected the dataset to detect the following data quality issues:**

**2. Discuss your findings for each quality issue listed in part B1.**

**•   Duplicate Entries**

I inspected the data to detect duplicate records by performing the following code on the data frame:

**A screenshot of a computer

AI-generated content may be incorrect.**

The duplicated() python method returns a boolean series indicating whether each row is a duplicate or not (*Course |*, n.d.-c). By using the sum() function after the duplicated() function, it shows that there are 99 total rows that are duplicated in the data frame.

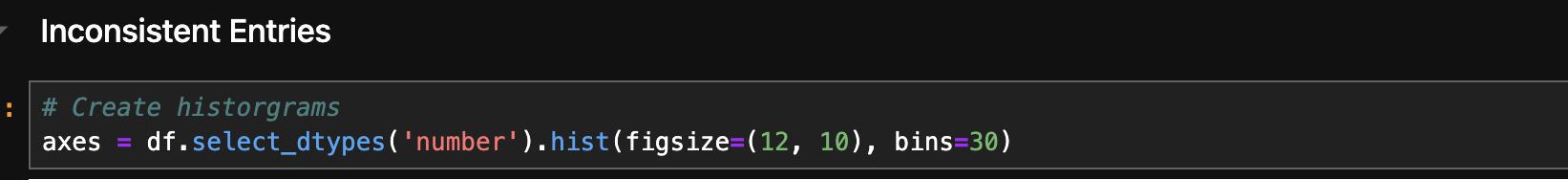
**•   Missing Values**

A screenshot of a computer

AI-generated content may be incorrect.I inspected the data to detect missing values by running the following:

The isnull() python method checks for NaN and null values, so by using the sum() function after the isnull() function, it displays the total count of null values for each column. Looking at the output, we can see that ‘TextMessageOptIn’ has the most null values.

**•   Inconsistent Entries and Formatting Errors**

I inspected the data to detect inconsistent entries by creating a histogram for the numeric columns. A histogram is similar to a bar graph but used for numbers instead. Using this graph can help us understand the distribution of columns and identify inconsistent entries.

**A group of blue graphs

AI-generated content may be incorrect.**

In order to generate the histograms, I first had to identify the numeric columns by using the select\_dtypes function and included ‘number,’ and then performed the hist function.

Looking at the results, most of the histograms look accurate, however, the ‘AnnualSalary’ and ‘DrivingCommuterDistance’ columns contain negative values. For both columns, it does not make sense for the values to be negative because employees don’t make a negative salary

Another method used to check for inconsistent entries and formatting errors for the non-numeric data is using the function ‘value\_counts.’ This method returns the unique values in the column and the count of each value.

**A screenshot of a computer

AI-generated content may be incorrect.**A black screen with white text

AI-generated content may be incorrect.**A screen shot of a computer

AI-generated content may be incorrect.A screenshot of a computer code

AI-generated content may be incorrect.**‘Turnover’, ‘CompensationType’, ‘TextMessageOptIn’, ‘Gender’, and ‘MaritalStatus’ look to be accurate and have no inconsistent entries or formatting errors.

**A screen shot of a computer

AI-generated content may be incorrect.**

**A screenshot of a computer screen

AI-generated content may be incorrect.**Here, we can see there were inconsistent entries for ‘Human Resources’, ‘HumanResources’, and ‘Human\_Resources’ as well as ‘Information Technology’, ‘InformationTechnology’, and ‘Information\_Technology.’

**A screenshot of a computer

AI-generated content may be incorrect.**Here, we can see inconsistent entries as well for ‘Mail Check’, ‘Mailed Check’, ‘MailedCheck’, and ‘Mail\_Check’ as well as ‘DirectDeposit’, ‘Direct\_Deposit’, and ‘Direct Deposit.’

Lastly, I noticed that the ‘AnnualSalary’ amount was inconsistent for some rows. I first calculated the ‘totalAmount’ by multiplying the ‘HourlyRate’, ‘HoursWeekly’, and 52 together, based on the data dictionary definition for ‘AnnualSalary.’ I then compared the values together and found that 2,149 rows were inconsistent.

A screenshot of a computer screen

AI-generated content may be incorrect.

**•   Outliers**

I inspected the data to detect outliers by creating boxplots for all the numeric columns. Boxplots are graphical distributions that display the overall distribution. These graphs display the median value and values known as ‘whiskers’ that are 1.5x the IQR from 1st quartile and the 3rd quartile (*Course |*, n.d.-c). Anything that is less than or greater than the whiskers is considered an outlier.

A screenshot of a computer screen

AI-generated content may be incorrect.In order to create the boxplots for the numeric columns, I used a function called select\_dtypes and selected ‘number’ for the data type in order to returns the numeric columns. Then using this variable, numeric\_cols, I used the matplotlib library to create boxplots for each of the columns. Looking at the results, it is evident that AnnualSalary and DrivingCommuterDistance both have outliers.

Zooming in closer to the ‘AnnualSalary’, the outliers that the boxplot revealed are around 300000 – 350000. After filtering the data frame to only include rows where the ‘AnnualSalary’ is greater than 300,000, we can see that the hourly rates are in the 90s, however, when calculating 90 (hourly rate) x 40 (number of hours worked weekly) x 52 (number of weeks in a year), this only amounts to 187,200. According to the data dictionary, the ‘AnnualSalary’ is the total amount of money an employee earns annually (52 weeks) based upon the hourly rate and the number of hours worked weekly. Based on this definition of the column, it can be concluded that these records are outliers.

**A screenshot of a computer screen

AI-generated content may be incorrect.A graph on a screen

AI-generated content may be incorrect.**

A screen shot of a graph

AI-generated content may be incorrect.Zooming in closer to the ‘DrivingCommuterDistance’, the outliers that the boxplot revealed are over 200 miles. As mentioned before, even though these values fall outside the typical range and are calculated as outliers based on the IQR, it is not necessarily irrelevant or erroneous. It is possible that employees are driving more than 200 miles to work; however, what is less possible is the record that shows the value of more than 800 miles. Since this value is so far away from the rest of the data, this is more likely to be an outlier.

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 using Python or R code after identifying each quality issue listed in part B1.

**Duplicate Entries:** In order to remove the duplicated rows, the function drop\_duplicates() is used to keep the first occurrence of the record and drop the rest. Looking at the screenshot below, after dropping the duplicated records, the count has dropped to 10,100, which is accurate because when looking for the duplications, there were a total of 99.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.**Missing Values:** In order to input null values, Python has a function called fillna(). For ‘NumCompaniesPreviouslyWorked’, 0 was used to replace the null values as shown below. After using the fillna() function, we can see that the column no longer has null values.

The same logic applies for ‘AnnualProfessionalDevHrs’, where I am assuming that null values mean 0 hours, so I used the fillna() function again to impute the missing data. Here, we can also see that after using the function, there were no more null values for the column.

A black screen with colorful text

AI-generated content may be incorrect.

For ‘TextMessageOptIn’, I chose to input the null values with ‘No’ by using the fillna() function. After using the function, we can see that there are no more null values.

A black screen with text

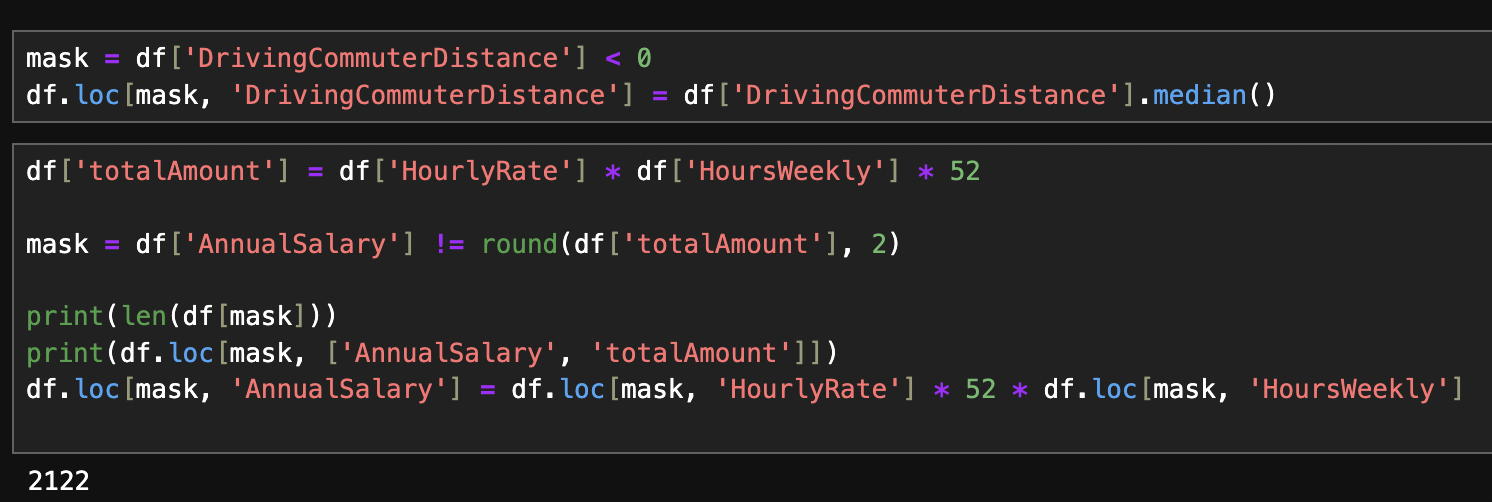
AI-generated content may be incorrect.

Running this code again, we can now see that there are no more missing values in the entire data frame.

A screenshot of a computer

AI-generated content may be incorrect.

**Inconsistent Entries and Formatting Errors:**

I modified the dataset by imputing the correct values for the negative values in ‘DrivingCommuterDistance’ as well as the negative/incorrect values in ‘AnnualSalary.’ For ‘DrivingCommuterDistance,’ I found all the rows that had a value of less than 0 and replaced it with the median value of that column. Similarly, with ‘AnnualSalary,’ I found all the rows where the calculation was inaccurate and replaced them with the correct calculation.

I modified the dataset by replacing the inconsistent categories to have the same spelling. I did this by remapping the values to one golden value using the replace() function in Python. The screenshot below checks to see if the value of the record is equal to the string, and if it is, it’ll replace it with the new string. After applying this to ‘JobRoleArea’, we can see now that the value counts look good and all the categories are different.

**A screenshot of a computer screen

AI-generated content may be incorrect.**

The same concept has been applied to ‘PaycheckMethod’.

**A screen shot of a computer

AI-generated content may be incorrect.**

**Outlier:**

I modified the dataset by imputing the correct values for ‘AnnualSalary.’ First, I calculated the IQR, took the 3rd quartile performed the calculation to identify the threshold for the values that were outliers (Q3 + 1.5 x IQR). After doing so, I found the values that were greater than the threshold and replaced them with the correct calculations I got for ‘AnnualSalary.’ This has already been fixed with the inconsistent entries before this step, however, this is how the value would be imputed if it was not fixed before this step.

A screenshot of a computer

AI-generated content may be incorrect.

Lastly, I modified the dataset by imputing the median values for ‘DrivingCommuterDistance.’ First, I found the records that were greater than 800 miles and calculated the median of the column. The median was used instead of the mean since it is robust to outliers and will not have a biased value. As seen in the screenshot below, there were 6 records with a distance greater than 800 miles. After imputing the data, there are no longer records with a distance greater than 800.

A screenshot of a computer program

AI-generated content may be incorrect.

2.  Discuss why you chose the specific data cleaning techniques you used to clean the quality issues listed in part B1.

**Duplicate Entries:** Duplicate records can bias results and distort insights because they could increase the frequency of certain observations. If left unaddressed, they could inflate statistics such as counts, averages, or distributions, leading to inaccurate conclusions. To address this, I removed exact duplicates and retained only one instance of each duplicated record. In some cases, it may be appropriate to aggregate duplicate entries by calculating averages, counts, or sums across the duplicated rows. However, this approach is not suitable for this dataset. For example, summing or averaging values such as age, hourly rate, or annual salary would not produce meaningful or accurate results, since these attributes are tied to an individual’s record and should not be artificially altered. Therefore, removing duplicates entirely was the most appropriate technique to preserve the integrity and accuracy of the dataset.

**Missing Values:** One of the methods of cleaning missing values is to remove the records that contain the missing data. However, according to the WGU Course material, eliminating the rows is suggested only when the null values are around or less than 5% of the data. For this dataset, this is not the case, as there are a total of 4,900 rows with a null value, equating to about 48% of the data. The next option is to replace the null values with estimated values, which can avoid issues of having missing data and avoid biases (*Course |*, n.d.-c).

There were 3 columns total with missing data, ‘NumCompaniesPreviouslyWorked’, ‘AnnualProfessionalDevHrs’, and ‘TextMessageOptIn.’ For ‘NumCompaniesPreviouslyWorked’, I decided to input the null values with 0 since there were no ‘0’ values to begin with. So, I am assuming that a null input means that the employee didn’t work at any previous companies. Using a filter, it is evident that there are no ‘0’ values.

A black rectangle with red text

AI-generated content may be incorrect.

For ‘AnnualProfessionalDevHrs’, I followed the same method to identify if there were any ‘0’ values using a filter.I also got an empty data frame as a result and so I am making an assumption that a null input means that the employee didn’t spend any hours on professional development in a year.

**A black rectangle with red text

AI-generated content may be incorrect.**

Lastly, for ‘TextMessageOptIn’, since there are only 2 possible values, ‘yes’ or ‘no’, I’m assuming that if the value is null, then the default is that the employee did not opt in for text messages, making the null values ‘no’.

**Inconsistent Entries and Formatting Errors:** In order to handle the negative values in ‘AnnualSalary’ and ‘DriverCommuterDistance’, I decided to impute those values with the correct calculation for ‘AnnualSalary’ and the median value for ‘DriverCommuterDistance’. I chose the median to decrease bias in the dataset since it is robust to outliers. Inputing the values is better than deleting the records so that the data can be preserved and we don’t lose valuable information.

For ‘JobRoleArea’ and ‘PaycheckMethod’, since there were categories that were the same thing, just a different way of spelling it or an added character, we can just replace the category names to match a certain way of spelling. For example, in ‘JobRoleArea’,

‘Human Resources’, ‘HumanResources’, and ‘Human\_Resources’ could all be replaced with ‘HumanResources.’ It doesn’t make sense to delete these records because it is obvious what category it is, so it is just a simple fix we can apply to the dataset.

**Outlier:** There were 2 columns with outliers that were found earlier: ‘AnnualSalary’ and ‘DrivingCommuterDistance.’ For ‘AnnualSalary’, I decided to impute the correct values by multiplying the ‘HourlyRate’ \* ‘HoursWeekly’ \* 52 to get the correct ‘AnnualSalary.’ Since this is something that can be calculated and fixed, it is better to fix it than to remove the records so that the dataset doesn’t lose valuable information.

For ‘DrivingCommuterDistance’, since this is something that cannot be calculated, I imputed the outliers with the median value of the ‘DrivingCommuterDistance’ column. Since the median is robust to outliers, it doesn’t introduce bias or drastically change the outcome of the dataset. This is better than removing the rows with outliers so that the dataset doesn’t lose valuable information. For the records that have an ‘outlier’ based on the formula of Q3 + 1.5 \* IQR, some of the records are believable that the employee does have to drive that far and therefore the record should be kept. For example, the threshold that was calculated was 158, however, based on the histogram from above, there is a big gap between ~350 and 800. It is also very unlikely that an employee is driving more than 800 miles to work, therefore, this is the outlier threshold I will be using to remove the outliers.

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

Removing the duplicates allows the data to be more accurate and cleaner. This also reduces bias in the dataset that could lead analysts to biased analysis.

The second advantage to my data cleaning approach was using median values to replace outliers. Since the median value is robust to outliers this prevents the data from being skewed and represents the central tendency. By doing so, I was able to keep the data rows instead of losing data that could be important to the analysis.

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

By dropping the duplicates, I could have been dropping important data as well. We don’t

always know the story behind the data, so it is possible that the employee quit at some point and got hired again. So the first instance of the employee number could be the previous position, while the second or third instance could be the most recent position the employee holds. This is a limitation, as I could have potentially deleted data that was important.

One of the data cleaning approaches I had was replacing null values with ‘0’ where I thought it made sense. By doing so, I could have been inputting false data that could lead to a misinterpretation of the data.

**References**

*Course |*. (n.d.-c). https://apps.cgp-oex.wgu.edu/learning/course/course-v1:WGUx+OEX0395+v01/block-v1:WGUx+OEX0395+v01+type@sequential+block@bd9a1e46ad734b79b411fb04fa43175e/block-v1:WGUx+OEX0395+v01+type@vertical+block@a0d92d52bf1542a798b0f2e6bc999a9e