|  |
| --- |
| Initial Steps followed: |
|  |

|  |
| --- |
| 1) Import db into MS Access using the MDB file provided along with the dataset and following instructions outlined in the corresponding instructions PDF file. |
|  |

|  |
| --- |
| 2) Filter the data for apples by selecting value for attribute "COMMOD (Commodity type)" as 'AP'. |
|  |

|  |
| --- |
| 3) Export the result set into ".xlsx" format. |
|  |

|  |
| --- |
| 4) Import the xlsx file into R Studio. |
|  |

|  |
| --- |
| Issues occurred: |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| The dataset was showing garbage values when imported into R. Following is the screenshot of the issue: |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| <In the document, this place will have a screenshot> |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| Steps to resolve the issue:- |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| 1) Copy the text from Samples.txt for 2014 data. |
|  |

|  |
| --- |
| 2) Open a blank CSV file. |
|  |

|  |
| --- |
| 3) Under the "Data" tab, click in "import from text". |
|  |

|  |
| --- |
| 4) Select the Samples.txt file under the "Import from text option", give delimiter as "|" |
|  |

|  |
| --- |
| 5) Add appropriate headers for each column. |
|  |

|  |
| --- |
| 6) Save the CSV file with an appropriate name. |
|  |

|  |
| --- |
| 7) Import the CSV into R Studio using import dataset option. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| Using the above steps we get 4 csv files: Apples\_Sample14.csv, Apples\_Results14.csv for 2014 data & Sampleapples04.csv, ResultsApples04.csv for 2004 data. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| 2014 Samples |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| 1) Source\_ID is blank for 158 out of the177 rows of Apple samples. Source\_ID not used in any of the research questions. The metadata accompany the dataset mentioned that Source\_ID is used to make a sample unique, however all of the samples already have unique values in the sample\_pk attribute, which is the primary key in incremental numbers to identify each sample. Therefore, the Source\_ID column was removed from the dataset. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| 2) Country field is used only when the Origin attribute has a value of '2' (i.e., the variety is imported from another country). However, for our samples, there is only one sample which is imported from "New Zealand" and corresponds to only one row in the dataset. The remaining 176 rows were blank for the Country attribute. For the sanity purpose, we replace the country code with "US" for '1' and "New Zealand" for '2' directly into the "Origin" attribute, and removed the Country variable. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| 3) GROWST- State postal code for location of sample grower. 163/177 blanks. PACKST- 118/177 blanks. DISTST- 34/177 blanks. These fields are not used in any of the research questions, and are thus removed |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| 2014 Results |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| 1) CONFMETHOD2 is blank for all rows ( 38792 rows total), and is thus removed from the cleaned data. |
|  |

|  |
| --- |
| 2) ANNOTATEis blank for all rows ( 38792 rows total), and is thus removed from the cleaned data. |
|  |

3) QUANTITATE is blank for all rows ( 38792 rows total), and is thus removed from the cleaned data.