FinalProject_Step2_PuppalaSucharitha

2022-08-06

Importing all the required libraries.

1. How to import and clean my data

Importing the data

Once the data is entered into Excel, CSV etc. file format we need a way to get the data file into a dataframe in R.Importing the data to R programming helps us to read the data from the external files, write data to the external files and can access the files from ooutside R environment.Some of the file formats that can be imported into R environment to read the data and perform data analysis are CSV,XML, xlsx, JSON, webdata.

Here for my project I have collected the data in a CSV(Comma Separated Values) file. For importing the data to a dataframe we use the "read.csv()" function.

Data read before cleaning the datasets.

b.) Clean the data

- For cleaning the data we start with removing the duplicates which can be done using the distinct(),unique(), duplicated() functions.
- From the Caffeine data set I have checked for the duplicates if any present and removed the duplicates.
- From the Coffee survey data set also i have checked for the duplicates if any present and removed the duplicates.
- From the Coffee chain data set, I have extracted the data that belongs to the Utah state in USA as the Coffee Survey data set has the survey data that has been collected from the students of Utah state.From the extracted part of the Utah data I have checked for the duplicates if any present and removed. *From the CoffeeConsumption data set, I have checked for the duplicates if any present and removed the duplicates.
- Here I have checked for the column names if any that required modification. I found the existing column names are good for doing the analysis part.
- During data preparation and cleaning we have to check for missing values which is the typical problem. For this we have to convert all the missing values to NA format, as in r programming the missing values are usually represented by 'NA'. In all the four data sets I have checked for the missing values if any present.
- I have also checked for any rows with missing values if any for all the four data sets
- I have also checked if any modification of the columns classes is required or not for all the four data sets

2. What does the final data set look like?

After cleaning the data the following are the final datasets that are required for the project.

```
##                          drink Volume..ml. Calories Caffeine..mg.   type
## 1                 Costa Coffee    256.9937        0           277 Coffee
## 2 Coffee Friend Brewed Coffee    250.1918        0           145 Coffee
## 3           Hell Energy Coffee    250.1918      150           100 Coffee
## 4          Killer Coffee (AU)    250.1918       0           430 Coffee
## 5                Nescafe Gold    250.1918       0            66 Coffee
## 6            Espresso Monster    248.4174     170           160 Coffee
```

```
##   Area.Code    Ddate  Market  Market.Size           Product Product.Type
## 1       970 1/1/2012 Central Major Market Decaf Irish Cream       Coffee
## 2       719 2/1/2012 Central Major Market Decaf Irish Cream       Coffee
## 3       720 3/1/2012 Central Major Market Decaf Irish Cream       Coffee
## 4       303 4/1/2012 Central Major Market Decaf Irish Cream       Coffee
## 5       720 5/1/2012 Central Major Market Decaf Irish Cream       Coffee
## 6       719 6/1/2012 Central Major Market Decaf Irish Cream       Coffee
##      State  Type Caffeine..mg. Budget.Cogs Budget.Margin Budget.Profit
## 1 Colorado Decaf             2         100           140           110
## 2 Colorado Decaf             2         100           140           110
## 3 Colorado Decaf             2         100           140           110
## 4 Colorado Decaf             2         100           150           120
## 5 Colorado Decaf             2         110           150           120
## 6 Colorado Decaf             2         130           180           140
##   Budget.Sales Coffee.Sales Cogs Inventory Margin Marketing Number.of.Records
## 1          240          234   95       821    139        26                 1
## 2          240          232   95       809    137        26                 1
## 3          240          234   95       799    139        26                 1
## 4          250          245  100       822    145        28                 1
## 5          260          256  104       871    152        29                 1
## 6          310          301  123       947    178        34                 1
##   Number.Of.Records Profit Total.Expenses
## 1                 1    101             38
## 2                 1     99             38
## 3                 1    101             38
## 4                 1    105             40
## 5                 1    112             40
## 6                 1    132             46
```

```
##   Do.you.drink.coffee.daily.
## 1                        YES
## 2                        YES
## 3                        YES
## 4                         NO
## 5                         NO
## 6                        YES
##   How.many.coffee.you.drink.daily..Starbucks.Grande.cup..
## 1                                                       2
## 2                                                       2
## 3                                                       3
## 4                                                       1
## 5                                                       0
## 6                                                       3
##                              Why.do.you.drink.coffee
## 1                                        Study Stress
## 2                            Refreshing every morning
## 3                                       Living habits
## 4                                       Living habits
## 5                                        Study Stress
## 6 Study Stress;Living habits;Refreshing every morning
##   Do.you.think.coffee.works.for.you.
## 1                                 No
## 2                                Yes
## 3                              Maybe
## 4                                 No
## 5                                 No
## 6                                Yes
```

```
##         country totCons2019 perCapitaCons2016
## 1 United States       27310              9.26
## 2       Germany        8670             12.13
## 3         Japan        7551                NA
## 4        France        6192             11.90
## 5         Italy        5469             13.00
## 6        Russia        4820                NA
```

3. What do you not know how to do right now that you need to learn to import and cleanup your dataset?

Initially I was confused of where to start for cleaning the data sets. But with the help of the text book readings and the weekly assignments I was able to complete the cleanup of the data sets. I felt the cleaning up of data sets and making it ready for the analysis is the toughest part for data analyzing. But once the data is ready and cleaned we can do the required analysis.

4. Discuss how you plan to uncover new information in the data that is not self-evident.

- In order to uncover new information in the data that is not self evident,from all the data sets collected I would like to create new variables as per the analysis requirement, or join separate data frames from the different data sets and create new summary information for doing the required analysis.
- I have created a new column in the Coffee chain data set with Caffeine values of different types of drinks which helps in the future analysis.

5. What are different ways you could look at this data to answer the questions you want to answer?

- As per the survey data I have collected I would like to know weather coffee really helps in changing the mood with help of the data visualizations available.
- I would like to go through all the different variables in the survey data to know which categories of people i.e like students, others drink more amount of coffee, this can be visualized with the help of histograms as it will be easy in identifying the result in a quick way.
- With regard to the additional data,I have collected some caffeine values of few types of drinks that are not available in the data sets collected. The new data is added to the cleaned data set of the coffee chain data set.

6. Do you plan to slice and dice the data in different ways, create new variables, or join separate data frames to create new summary information? Explain.

- In answering the research questions for this project I have to select particular columns from all the different data sets I have collected, and this process of selecting specific columns or rows of data based on some criteria is known as slicing.
- From the Caffeine data set I would use the drink and caffeine columns for the analysis.
- From Coffee survey data set I would use the column "Why do you drink Coffee".
- From the Coffee Chain data set I have added a new column with the caffeine values.

7. How could you summarize your data to answer key questions?

- To summarize the data I would use the summary() function for all the data sets, as obtaining the summary of the whole data set helps in understanding the data set clearly. To use the summary() I have imported the Dplyr package.
- The format of the result depends on the data type of the column. If the column is a numeric variable, mean, median, min, max and quartiles are returned. If the column is a factor variable, the number of observations in each group is returned.
- I would like to get the summary of the individual columns to know the statistical details of the data.
- With help of the summary details obtained I would like to combine all the data obtained from all the data sets in arriving the final conclusion.Like the Caffeine content among the drinks, the coffee consumption in different states.

8. What types of plots and tables will help you to illustrate the findings to your questions?

- Graphical data analysis helps in knowing the properties of the data that is plotted.
- There are many ways in plotting graphs in R. I would like to use the histograms, line graphs and scatter plots.
- I would like to use the caffeine and drink columns and plot a histogram and scatter plot from the caffeine data set as this helps in understanding the question of the project.
- I would use the Why do you drink Coffee and Do drinking coffee works for you? columns from the coffee Survey data set and plot a histogram to know the data more clearly.
- I would also use the plots of each individual column like a single variable plot for better understanding the data, where necessary.

9. Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

- In order to determine the relationship between the dataset variables we use the regression analysis. This regression analysis helps to understand how dependent variables change when one of the independent variable is changed and other independent variables are kept constant, This helps in building a model and hepls in forecasting the values with respect to the change in one of the independent variables.
- In this project I would like to use Linear Regression model as most of my project is to know the relationship between two variables.

10. Could summary statistics at different categorical levels tell you more?

- Summary statistics at different categorical levels is mostly about building tables,and calculating percentages or proportions of the data variable selected.
- This helps in knowing about the individual variable that contains values but there will not be any known relationship among them.