DSC540-T301_2237-1_Samanta_Rajib_Week_5 and_6

July 16, 2023

[41]:
```
# DSC540-T301_2237-1 Data Preparation(2237-1)
# Assignment: Week 05 and 06 - Exercise
# Author by:  Rajib Samanta
# Date: 2023-07-16
```

[42]:
```
# Imports
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import requests
import io
import warnings

# Ignore all warning messages
warnings.filterwarnings("ignore")
```

[43]:
```
#Chapter 7

## Filter out missing data
## Fill in missing data
## Remove duplicates
## Transform data using either mapping or a function
## Replace values
## Discretization and Binning
##  Manipulate Strings
```

[44]:
```
#Set working directory
import os
directory = '/Users/rajibsamanta/Documents/Rajib/College/Sem5 2023/Week 5 & 6'
# Set the working directory
os.chdir(directory)
print(os.getcwd())
```

```
/Users/rajibsamanta/Documents/Rajib/College/Sem5 2023/Week 5 & 6
```

[45]:
```
df_ch_data= pd.read_csv("candyhierarchy2017.csv",encoding='latin-1' )

# Check the first 5 records.
```

```
df_ch_data.head(5)
```

[45]:
```
   Internal ID Q1: GOING OUT? Q2: GENDER Q3: AGE Q4: COUNTRY  \
0     90258773            NaN        NaN     NaN         NaN
1     90272821             No       Male      44         USA
2     90272829            NaN       Male      49         USA
3     90272840             No       Male      40          us
4     90272841             No       Male      23         usa

  Q5: STATE, PROVINCE, COUNTY, ETC Q6 | 100 Grand Bar  \
0                              NaN               NaN
1                               NM               MEH
2                         Virginia               NaN
3                               or               MEH
4                         exton pa               JOY

  Q6 | Anonymous brown globs that come in black and orange wrappers\t(a.k.a.
Mary Janes)  \
0                                                NaN
1                                           DESPAIR
2                                                NaN
3                                           DESPAIR
4                                           DESPAIR

  Q6 | Any full-sized candy bar Q6 | Black Jacks  … Q8: DESPAIR OTHER  \
0                           NaN              NaN  …               NaN
1                           JOY              MEH  …               NaN
2                           NaN              NaN  …               NaN
3                           JOY              MEH  …               NaN
4                           JOY          DESPAIR  …               NaN

                                Q9: OTHER COMMENTS      Q10: DRESS  \
0                                              NaN             NaN
1  Bottom line is Twix is really the only candy w…  White and gold
2                                              NaN             NaN
3                          Raisins can go to hell  White and gold
4                                              NaN  White and gold

  Unnamed: 113 Q11: DAY Q12: MEDIA [Daily Dish] Q12: MEDIA [Science]  \
0          NaN      NaN                     NaN                   NaN
1          NaN   Sunday                     NaN                   1.0
2          NaN      NaN                     NaN                   NaN
3          NaN   Sunday                     NaN                   1.0
4          NaN   Friday                     NaN                   1.0

  Q12: MEDIA [ESPN] Q12: MEDIA [Yahoo] Click Coordinates (x, y)
0               NaN                NaN                      NaN
```

```
1               NaN                NaN                (84, 25)
2               NaN                NaN                     NaN
3               NaN                NaN                (75, 23)
4               NaN                NaN                (70, 10)

[5 rows x 120 columns]
```

[46]:
```
# Print the dimension of the dataframe
print("Size of DataFrame: ", df_ch_data.shape )
```

```
Size of DataFrame:  (2460, 120)
```

[48]:
```
# count of missing values in each column of the dataframe
print(df_ch_data.isna().sum())

##--> Below list shows how many rows has null value for each column
```

```
Internal ID                    0
Q1: GOING OUT?               110
Q2: GENDER                    41
Q3: AGE                       84
Q4: COUNTRY                   64
                             ...
Q12: MEDIA [Daily Dish]     2375
Q12: MEDIA [Science]        1098
Q12: MEDIA [ESPN]           2361
Q12: MEDIA [Yahoo]          2393
Click Coordinates (x, y)     855
Length: 120, dtype: int64
```

[49]:
```
# Set the 'Internal ID' column as the index
df_ch_data.set_index(['Internal ID'], inplace=True)
# Print the first few rows of the dataframe
print(df_ch_data.head())
```

```
            Q1: GOING OUT? Q2: GENDER Q3: AGE Q4: COUNTRY  \
Internal ID
90258773               NaN        NaN     NaN         NaN
90272821                No       Male      44         USA
90272829               NaN       Male      49         USA
90272840                No       Male      40          us
90272841                No       Male      23         usa

            Q5: STATE, PROVINCE, COUNTY, ETC Q6 | 100 Grand Bar  \
Internal ID
90258773                                   NaN                NaN
90272821                                    NM                MEH
90272829                              Virginia                NaN
90272840                                    or                MEH
```

```
90272841                            exton pa                JOY

           Q6 | Anonymous brown globs that come in black and orange
wrappers\t(a.k.a. Mary Janes)  \
Internal ID
90258773                                                     NaN
90272821                                                DESPAIR
90272829                                                    NaN
90272840                                                DESPAIR
90272841                                                DESPAIR

           Q6 | Any full-sized candy bar Q6 | Black Jacks  \
Internal ID
90258773                              NaN              NaN
90272821                              JOY              MEH
90272829                              NaN              NaN
90272840                              JOY              MEH
90272841                              JOY          DESPAIR

           Q6 | Bonkers (the candy)  … Q8: DESPAIR OTHER  \
Internal ID                           …
90258773                       NaN    …               NaN
90272821                   DESPAIR    …               NaN
90272829                       NaN    …               NaN
90272840                       MEH    …               NaN
90272841                       MEH    …               NaN

                                          Q9: OTHER COMMENTS  \
Internal ID
90258773                                                     NaN
90272821     Bottom line is Twix is really the only candy w…
90272829                                                     NaN
90272840                                Raisins can go to hell
90272841                                                     NaN

               Q10: DRESS Unnamed: 113 Q11: DAY Q12: MEDIA [Daily Dish]  \
Internal ID
90258773                NaN          NaN      NaN                     NaN
90272821     White and gold          NaN   Sunday                     NaN
90272829                NaN          NaN      NaN                     NaN
90272840     White and gold          NaN   Sunday                     NaN
90272841     White and gold          NaN   Friday                     NaN

           Q12: MEDIA [Science] Q12: MEDIA [ESPN] Q12: MEDIA [Yahoo]  \
Internal ID
90258773                    NaN               NaN                NaN
90272821                    1.0               NaN                NaN
90272829                    NaN               NaN                NaN
```

```
90272840                    1.0                NaN                NaN
90272841                    1.0                NaN                NaN

           Click Coordinates (x, y)
Internal ID
90258773                        NaN
90272821                  (84, 25)
90272829                        NaN
90272840                  (75, 23)
90272841                  (70, 10)

[5 rows x 119 columns]
```

[50]:
```
# Drop rows where all values are missing
df_ch_data.dropna(axis=0, how='all', inplace=True)

# Print the dimensions of the modified dataframe
print("Size of DataFrame after removal of rows having all values missing  : ",␣
 ↪df_ch_data.shape )
## --> before droping missing value the Size of DataFrame was:  (2460, 120) ,␣
 ↪now :  (2439, 119)
## --> around 21 rows and 1 column has been dropped.
```

```
Size of DataFrame after removal of rows having all values missing  :  (2439,
119)
```

[51]:
```
#Count the no of duplicate rows and drop the duplicates
df_ch_data.duplicated().sum()
print("Duplicate recod count in current dataframe   : ", df_ch_data.
 ↪duplicated().sum() )
df_ch_data.drop_duplicates(inplace=True)
print("Size of DataFrame after removal of duplicaes : ", df_ch_data.shape )

## --> Duplicate rows count : 15
```

```
Duplicate recod count in current dataframe   :  15
Size of DataFrame after removal of duplicaes :  (2424, 119)
```

[52]:
```
# Dropping columns where more than 50% records are NaN
df_ch_data.dropna(thresh=df_ch_data.shape[0]*0.5, how='all', axis=1,␣
 ↪inplace=True)

# Dropping rows where more than 50% records are NaN
df_ch_data.dropna(thresh=df_ch_data.shape[1]*0.5, how='all', axis=0,␣
 ↪inplace=True)

print("Size of DataFrame after removal such records : ", df_ch_data.shape )
```

```
## Now the data frame size is : (1781, 112)
```

```
Size of DataFrame after removal such records :  (1781, 112)
```

[53]:
```
## --> In the provided example, we apply binning to the "age" variable,␣
 ↪categorizing it into nine distinct age groups. After creating these␣
 ↪intervals,
##.    we then assign bin labels to each group, giving meaningful names to␣
 ↪identify and represent the ranges effectively.
# Define the bin edges
bin_edges = [0, 18, 30, 40, 50, 60, 70, 80, 90, 99, 100]

# Define the bin labels
bin_labels = ['00-18', '19-30', '31-40', '41-50', '51-60', '61-70',␣
 ↪'71-80','81-90', '91-99', '100 +']

# Convert the 'Q3: AGE' column to numeric
df_ch_data['Q3: AGE'] = pd.to_numeric(df_ch_data['Q3: AGE'], errors='coerce')
# Create a new column called "Age Group"
df_ch_data['Age Group'] = pd.cut(df_ch_data['Q3: AGE'], bins=bin_edges,␣
 ↪labels=bin_labels)
# Print the counts of each Age Group
print(df_ch_data['Age Group'].value_counts(sort=False))

## --> The age groups were defined based on specific bin edges: 0-18, 19-30,␣
 ↪31-40, 41-50, 51-60, 61-70, 71-80, 81-90,91-99 and 100 +
##    The age group with the highest number of respondents is 41-50,␣
 ↪demonstrating significant participation in this category. Following closely,␣
 ↪the 31-40 age group also exhibits substantial representation among the␣
 ↪survey participants.
##   On the other end of the spectrum, the age groups 81-90 and 91-100 have the␣
 ↪fewest respondents, indicating limited participation in these higher age␣
 ↪ranges.
```

```
00-18     46
19-30    179
31-40    537
41-50    547
51-60    304
61-70     93
71-80     16
81-90      3
91-99      1
100 +      1
Name: Age Group, dtype: int64
```

[54]:
```
#. Chapter 8
## Create hierarchical index
## Combine and Merge Datasets (you will have to either create a new dataset␣
 ↪from your existing data or create a relationship between the data I have␣
 ↪provided)
## Reshape
## Pivot the data
```

[55]:
```
#extract the names of the different candies

# Create a dictionary of candy names
candy_names = {x:x.split('Q6 | ')[1] for x in df_ch_data.columns if x.
 ↪startswith('Q6 |')}

# Rename the columns of the dataframe
df_ch_data.rename(columns=candy_names, inplace=True)

# Reset the index of the dataframe
df_ch_data.reset_index(inplace=True)

# Print the first few rows of the modified dataframe
print(df_ch_data.head())
```

```
   Internal ID Q1: GOING OUT? Q2: GENDER  Q3: AGE Q4: COUNTRY  \
0     90272821             No       Male     44.0        USA
1     90272840             No       Male     40.0         us
2     90272841             No       Male     23.0        usa
3     90272852             No       Male      NaN        NaN
4     90272854             No       Male     33.0     canada

  Q5: STATE, PROVINCE, COUNTY, ETC 100 Grand Bar  \
0                               NM           MEH
1                               or           MEH
2                         exton pa           JOY
3                              NaN           JOY
4                          ontario           JOY

  Anonymous brown globs that come in black and orange wrappers\t(a.k.a. Mary
Janes)  \
0                                          DESPAIR
1                                          DESPAIR
2                                          DESPAIR
3                                          DESPAIR
4                                          DESPAIR

  Any full-sized candy bar Black Jacks  …  Vicodin Whatchamacallit Bars  \
0                      JOY         MEH  …  DESPAIR              DESPAIR
1                      JOY         MEH  …      JOY                  JOY
```

```
2                      JOY     DESPAIR  …      JOY                  JOY
3                      JOY         NaN  …  DESPAIR                  JOY
4                      JOY     DESPAIR  …      MEH              DESPAIR

  White Bread Whole Wheat anything York Peppermint Patties      Q10: DRESS  \
0     DESPAIR              DESPAIR                 DESPAIR  White and gold
1     DESPAIR              DESPAIR                 DESPAIR  White and gold
2     DESPAIR              DESPAIR                     JOY  White and gold
3     DESPAIR              DESPAIR                     JOY             NaN
4     DESPAIR              DESPAIR                 DESPAIR  Blue and black

  Q11: DAY Q12: MEDIA [Science] Click Coordinates (x, y) Age Group
0   Sunday                  1.0                 (84, 25)     41-50
1   Sunday                  1.0                 (75, 23)     31-40
2   Friday                  1.0                 (70, 10)     19-30
3      NaN                  1.0                 (75, 23)       NaN
4   Friday                  1.0                  (55, 5)     31-40

[5 rows x 114 columns]
```

[56]:
```python
# Drop rows with missing values in the 'Q2: GENDER' or 'Q3: AGE' columns
df_ch_data.dropna(subset=['Q2: GENDER', 'Q3: AGE'], inplace=True)

# Print the number of rows and columns of the modified dataframe
print(df_ch_data.shape)
```

```
(1726, 114)
```

[57]:
```python
# Set the index of the dataframe to the 'Q2: GENDER' and 'Q3: AGE' columns
df_ch_data.set_index(['Q2: GENDER', 'Q3: AGE'], inplace=True)

# Print the first few rows of the modified dataframe
print(df_ch_data.head())
```

```
                    Internal ID Q1: GOING OUT? Q4: COUNTRY  \
Q2: GENDER Q3: AGE
Male       44.0        90272821             No        USA
           40.0        90272840             No          us
           23.0        90272841             No         usa
           33.0        90272854             No      canada
           40.0        90272858             No      Canada

                   Q5: STATE, PROVINCE, COUNTY, ETC 100 Grand Bar  \
Q2: GENDER Q3: AGE
Male       44.0                                 NM           MEH
           40.0                                 or           MEH
           23.0                          exton pa            JOY
           33.0                           ontario            JOY
           40.0                           Ontario            JOY
```

```
                  Anonymous brown globs that come in black and orange
wrappers\t(a.k.a. Mary Janes)  \
Q2: GENDER Q3: AGE
Male       44.0                                                DESPAIR
           40.0                                                DESPAIR
           23.0                                                DESPAIR
           33.0                                                DESPAIR
           40.0                                                DESPAIR

                  Any full-sized candy bar Black Jacks Bonkers (the candy)  \
Q2: GENDER Q3: AGE
Male       44.0                        JOY         MEH             DESPAIR
           40.0                        JOY         MEH                 MEH
           23.0                        JOY     DESPAIR                 MEH
           33.0                        JOY     DESPAIR             DESPAIR
           40.0                        JOY         MEH                 MEH

                  Bonkers (the board game)  …  Vicodin  \
Q2: GENDER Q3: AGE                          …
Male       44.0                    DESPAIR  …  DESPAIR
           40.0                    DESPAIR  …      JOY
           23.0                    DESPAIR  …      JOY
           33.0                        MEH  …      MEH
           40.0                        MEH  …  DESPAIR

                  Whatchamacallit Bars White Bread Whole Wheat anything  \
Q2: GENDER Q3: AGE
Male       44.0                DESPAIR     DESPAIR              DESPAIR
           40.0                    JOY     DESPAIR              DESPAIR
           23.0                    JOY     DESPAIR              DESPAIR
           33.0                DESPAIR     DESPAIR              DESPAIR
           40.0                    MEH     DESPAIR              DESPAIR

                  York Peppermint Patties      Q10: DRESS Q11: DAY  \
Q2: GENDER Q3: AGE
Male       44.0                   DESPAIR  White and gold   Sunday
           40.0                   DESPAIR  White and gold   Sunday
           23.0                       JOY  White and gold   Friday
           33.0                   DESPAIR  Blue and black   Friday
           40.0                   DESPAIR  Blue and black   Sunday

                  Q12: MEDIA [Science] Click Coordinates (x, y) Age Group
Q2: GENDER Q3: AGE
Male       44.0                    1.0                (84, 25)     41-50
           40.0                    1.0                (75, 23)     31-40
           23.0                    1.0                (70, 10)     19-30
           33.0                    1.0                 (55, 5)     31-40
```

```
          40.0                    1.0                (76, 24)     31-40

[5 rows x 112 columns]
```

[58]:
```
#Create a new index with candy names
new_index = pd.Index(list(candy_names.values()),name='candy')

# Reindex the DataFrame with the new index
new_df_ch_data = df_ch_data.reindex(columns=new_index)

# Display the first few rows of the updated DataFrame
new_df_ch_data.head()
```

[58]:
```
candy                100 Grand Bar  \
Q2: GENDER Q3: AGE
Male       44.0               MEH
           40.0               MEH
           23.0               JOY
           33.0               JOY
           40.0               JOY

candy                Anonymous brown globs that come in black and orange
wrappers\t(a.k.a. Mary Janes)  \
Q2: GENDER Q3: AGE
Male       44.0                                                DESPAIR
           40.0                                                DESPAIR
           23.0                                                DESPAIR
           33.0                                                DESPAIR
           40.0                                                DESPAIR

candy                Any full-sized candy bar Black Jacks Bonkers (the candy)  \
Q2: GENDER Q3: AGE
Male       44.0                          JOY         MEH              DESPAIR
           40.0                          JOY         MEH                  MEH
           23.0                          JOY     DESPAIR                  MEH
           33.0                          JOY     DESPAIR              DESPAIR
           40.0                          JOY         MEH                  MEH

candy                Bonkers (the board game) Bottle Caps Box'o'Raisins  \
Q2: GENDER Q3: AGE
Male       44.0                       DESPAIR     DESPAIR       DESPAIR
           40.0                       DESPAIR         MEH       DESPAIR
           23.0                       DESPAIR         MEH       DESPAIR
           33.0                           MEH         JOY           MEH
           40.0                           MEH         MEH       DESPAIR

candy                Broken glow stick Butterfinger  … Three Musketeers  \
```

```
Q2: GENDER Q3: AGE                                  …
Male       44.0              DESPAIR      DESPAIR  …              JOY
           40.0              DESPAIR          MEH  …          DESPAIR
           23.0              DESPAIR          MEH  …              JOY
           33.0                  JOY          JOY  …              JOY
           40.0              DESPAIR          JOY  …              MEH

candy              Tolberone something or other Trail Mix Twix  \
Q2: GENDER Q3: AGE
Male       44.0                            JOY   DESPAIR  JOY
           40.0                            JOY       MEH  JOY
           23.0                            JOY   DESPAIR  JOY
           33.0                            MEH   DESPAIR  JOY
           40.0                            JOY   DESPAIR  JOY

candy              Vials of pure high fructose corn syrup, for main-lining into
your vein  \
Q2: GENDER Q3: AGE
Male       44.0                                              DESPAIR
           40.0                                              DESPAIR
           23.0                                                  MEH
           33.0                                                  JOY
           40.0                                                  MEH

candy               Vicodin Whatchamacallit Bars White Bread  \
Q2: GENDER Q3: AGE
Male       44.0     DESPAIR              DESPAIR     DESPAIR
           40.0         JOY                  JOY     DESPAIR
           23.0         JOY                  JOY     DESPAIR
           33.0         MEH              DESPAIR     DESPAIR
           40.0     DESPAIR                  MEH     DESPAIR

candy              Whole Wheat anything York Peppermint Patties
Q2: GENDER Q3: AGE
Male       44.0                 DESPAIR                DESPAIR
           40.0                 DESPAIR                DESPAIR
           23.0                 DESPAIR                    JOY
           33.0                 DESPAIR                DESPAIR
           40.0                 DESPAIR                DESPAIR

[5 rows x 103 columns]
```

[59]:
```
## --> The data reshaping process involves utilizing the stack() method to␣
 ↪transform the DataFrame by stacking its columns.
##   Subsequently, we reset the index, resulting in a new DataFrame with a␣
 ↪multi-level index.
```

```
##   This new DataFrame comprises two columns: one column retains the original␣
 ↪column names, and the other column contains the corresponding values␣
 ↪extracted from the original DataFrame.
```

[60]:
```
# Reshaping the DataFrame to long format for easier analysis
ldata = new_df_ch_data.stack().reset_index().rename(columns={0: 'candy_liking'})
ldata[:10]
```

[60]:
```
  Q2: GENDER  Q3: AGE                                                candy  \
0       Male     44.0                                       100 Grand Bar
1       Male     44.0  Anonymous brown globs that come in black and o…
2       Male     44.0                          Any full-sized candy bar
3       Male     44.0                                       Black Jacks
4       Male     44.0                            Bonkers (the candy)
5       Male     44.0                          Bonkers (the board game)
6       Male     44.0                                       Bottle Caps
7       Male     44.0                                     Box'o'Raisins
8       Male     44.0                                  Broken glow stick
9       Male     44.0                                       Butterfinger

  candy_liking
0          MEH
1      DESPAIR
2          JOY
3          MEH
4      DESPAIR
5      DESPAIR
6      DESPAIR
7      DESPAIR
8      DESPAIR
9      DESPAIR
```

[61]:
```
# Chapter 10
## Grouping with Dicts/Series
## Grouping with Functions
## Grouping with Index Levels
## Split/Apply/Combine
## Cross Tabs
```

[62]:
```
# Reset the index of the Dataframe
df_ch_data.reset_index(inplace=True)
df_ch_data.head()
```

[62]:
```
  Q2: GENDER  Q3: AGE  Internal ID Q1: GOING OUT? Q4: COUNTRY  \
0       Male     44.0     90272821             No         USA
1       Male     40.0     90272840             No          us
2       Male     23.0     90272841             No         usa
```

```
3       Male     33.0     90272854             No      canada
4       Male     40.0     90272858             No      Canada

  Q5: STATE, PROVINCE, COUNTY, ETC 100 Grand Bar  \
0                               NM           MEH
1                               or           MEH
2                         exton pa           JOY
3                          ontario           JOY
4                          Ontario           JOY

  Anonymous brown globs that come in black and orange wrappers\t(a.k.a. Mary
Janes)  \
0                                          DESPAIR
1                                          DESPAIR
2                                          DESPAIR
3                                          DESPAIR
4                                          DESPAIR

  Any full-sized candy bar Black Jacks  …  Vicodin Whatchamacallit Bars  \
0                      JOY         MEH  …  DESPAIR               DESPAIR
1                      JOY         MEH  …      JOY                   JOY
2                      JOY     DESPAIR  …      JOY                   JOY
3                      JOY     DESPAIR  …      MEH               DESPAIR
4                      JOY         MEH  …  DESPAIR                   MEH

  White Bread Whole Wheat anything York Peppermint Patties      Q10: DRESS  \
0     DESPAIR              DESPAIR                 DESPAIR  White and gold
1     DESPAIR              DESPAIR                 DESPAIR  White and gold
2     DESPAIR              DESPAIR                     JOY  White and gold
3     DESPAIR              DESPAIR                 DESPAIR  Blue and black
4     DESPAIR              DESPAIR                 DESPAIR  Blue and black

  Q11: DAY Q12: MEDIA [Science] Click Coordinates (x, y) Age Group
0   Sunday                 1.0                 (84, 25)     41-50
1   Sunday                 1.0                 (75, 23)     31-40
2   Friday                 1.0                 (70, 10)     19-30
3   Friday                 1.0                  (55, 5)     31-40
4   Sunday                 1.0                 (76, 24)     31-40

[5 rows x 114 columns]
```

[63]:
```
print("Size of DataFrame  : ", df_ch_data.shape )
```

```
Size of DataFrame  :  (1726, 114)
```

[64]:
```
# Dropping records where either country or State/Province/City field have␣
 ↪missing data
```

```
df_ch_data.dropna(subset=['Q4: COUNTRY','Q5: STATE, PROVINCE, COUNTY,␣
 ↪ETC'],inplace=True)
df_ch_data.shape

## --> Now, we have the opportunity to map the original column names to new␣
 ↪column names.
##  After this mapping, we proceed to group the DataFrame based on the new␣
 ↪column names using the 'groupby' function, where the 'mapping' dictionary is␣
 ↪passed as an argument.
```

[64]:
```
(1706, 114)
```

[65]:
```
# Define a dictionary to map the original column names to new column names
mapping = {
'Q4: COUNTRY': 'COUNTRY',
'Q5: STATE, PROVINCE, COUNTY, ETC': 'LOCATION'
}

# Group the DataFrame by columns using the mapping dictionary and axis=1
by_column = df_ch_data.groupby(mapping, axis=1)
by_column.describe()
## --> United States (USA) displayed the highest participation, with total of␣
 ↪518 respondents contributing to the survey where total is : 1706.
##  The survey covered 417 distinct locations, encompassing states and cities.
##   The state of California stood out with the highest level of participation,␣
 ↪garnering 105 responses, reflecting its active involvement in the survey.
##  These statistics underscore the widespread reach and engagement of the␣
 ↪survey, reflecting valuable insights from a diverse range of respondents␣
 ↪across different countries, states, and cities.
```

[65]:

	count	unique	top	freq
Q4: COUNTRY	1706	87	USA	518
Q5: STATE, PROVINCE, COUNTY, ETC	1706	417	California	105

[66]:
```
# Returns the first 10 rows of the ldata DataFrame, which is created by␣
 ↪stacking the columns

ldata[:10]
```

[66]:

	Q2: GENDER	Q3: AGE	candy
0	Male	44.0	100 Grand Bar
1	Male	44.0	Anonymous brown globs that come in black and o…
2	Male	44.0	Any full-sized candy bar
3	Male	44.0	Black Jacks
4	Male	44.0	Bonkers (the candy)
5	Male	44.0	Bonkers (the board game)
6	Male	44.0	Bottle Caps

```
7       Male     44.0                                    Box'o'Raisins
8       Male     44.0                                  Broken glow stick
9       Male     44.0                                       Butterfinger

  candy_liking
0          MEH
1      DESPAIR
2          JOY
3          MEH
4      DESPAIR
5      DESPAIR
6      DESPAIR
7      DESPAIR
8      DESPAIR
9      DESPAIR
```

[67]:
```
# Used to find the count of each candy marked as JOY, MEH, and DISPAIR.
pd.crosstab([ldata['candy']],ldata.candy_liking,margins=True)
```

[67]:
```
candy_liking                                      DESPAIR    JOY    MEH  \
candy
100 Grand Bar                                          82    845    724
Abstained from M&M'ing.                               671    213    591
Anonymous brown globs that come in black and or…     1043    173    448
Any full-sized candy bar                               15   1502    200
Black Jacks                                           768     85    598
…                                                       …      …      …
Whatchamacallit Bars                                  271    826    493
White Bread                                          1404     42    201
Whole Wheat anything                                 1248    111    300
York Peppermint Patties                               221   1067    407
All                                                 56205  62155  51493

candy_liking                                          All
candy
100 Grand Bar                                        1651
Abstained from M&M'ing.                              1475
Anonymous brown globs that come in black and or…     1664
Any full-sized candy bar                             1717
Black Jacks                                          1451
…                                                       …
Whatchamacallit Bars                                 1590
White Bread                                          1647
Whole Wheat anything                                 1659
York Peppermint Patties                              1695
All                                                169853
```

```
[104 rows x 4 columns]
```

[68]:
```
# counting the number of times each candy is marked as "JOY", "MEH", or␣
 ↪"DISPAIR" in the crosstab_df dataframe
crosstab_df_ch_data = pd.crosstab([ldata['candy']],ldata.candy_liking)
print(crosstab_df_ch_data[crosstab_df_ch_data['JOY'] ==␣
 ↪crosstab_df_ch_data['JOY'].max()])
```

```
candy_liking              DESPAIR   JOY  MEH
candy
Any full-sized candy bar       15  1502  200
```

[69]:
```
## --> The output shows the count of candy liking categories (DESPAIR, JOY,␣
 ↪MEH) for each candy,
##     and the candy with the highest count of JOY category is "Any full-sized␣
 ↪candy bar" with 1502 counts.
```

[70]:
```
# Chapter 11
## Convert between string and date time
## Generate date range
## Frequencies and date offsets
## Convert timestamps to periods and back
## Period Frequency conversions
```

[71]:
```
df_bbch_2016= pd.read_excel("BOING-BOING-CANDY-HIERARCHY-2016-SURVEY-Responses.
 ↪xlsx"
                            )

# Check the first 5 records.
df_bbch_2016.head()
```

[71]:
```
                 Timestamp  \
0 2016-10-24 05:09:23.033
1 2016-10-24 05:09:54.798
2 2016-10-24 05:13:06.734
3 2016-10-24 05:14:17.192
4 2016-10-24 05:14:24.625

  Are you going actually going trick or treating yourself? Your gender:  \
0                                                 No                Male
1                                                 No                Male
2                                                 No              Female
3                                                 No                Male
4                                                Yes                Male

  How old are you? Which country do you live in?  \
0             22.0                         Canada
1             45.0                            usa
```

```
2             48.0                            US
3             57.0                           usa
4             42.0                           USA

  Which state, province, county do you live in?  [100 Grand Bar]  \
0                                       Ontario              JOY
1                                            il              MEH
2                                      Colorado              JOY
3                                            il              JOY
4                                  South Dakota              MEH

   [Anonymous brown globs that come in black and orange wrappers]  \
0                                           DESPAIR
1                                               MEH
2                                           DESPAIR
3                                               MEH
4                                           DESPAIR

   [Any full-sized candy bar]  [Black Jacks]  …  \
0                         JOY            MEH  …
1                         JOY            JOY  …
2                         JOY            MEH  …
3                         JOY            MEH  …
4                         JOY        DESPAIR  …

  Please estimate the degree(s) of separation you have from the following
celebrities [JK Rowling]  \
0                                     3 or higher
1                                     3 or higher
2                                     3 or higher
3                                     3 or higher
4                                     3 or higher

  Please estimate the degree(s) of separation you have from the following
celebrities [JJ Abrams]  \
0                                            2.0
1                                     3 or higher
2                                     3 or higher
3                                     3 or higher
4                                     3 or higher

  Please estimate the degree(s) of separation you have from the following
celebrities [Beyoncé]  \
0                                     3 or higher
1                                     3 or higher
2                                     3 or higher
3                                     3 or higher
```

```
4                                       3 or higher

  Please estimate the degree(s) of separation you have from the following
celebrities [Bieber]  \
0                                       3 or higher
1                                       3 or higher
2                                       3 or higher
3                                       3 or higher
4                                       3 or higher

  Please estimate the degree(s) of separation you have from the following
celebrities [Kevin Bacon]  \
0                                       3 or higher
1                                       3 or higher
2                                       3 or higher
3                                       3 or higher
4                                       3 or higher

  Please estimate the degree(s) of separation you have from the following
celebrities [Francis Bacon (1561 - 1626)]  \
0                                       3 or higher
1                                       3 or higher
2                                       3 or higher
3                                       3 or higher
4                                       3 or higher

  Which day do you prefer, Friday or Sunday?  \
0                                     Friday
1                                     Friday
2                                     Sunday
3                                     Sunday
4                                     Friday

  Do you eat apples the correct way, East to West (side to side) or do you eat
them like a freak of nature, South to North (bottom to top)?  \
0                                     South to North
1                                       East to West
2                                       East to West
3                                     South to North
4                                       East to West

  When you see the above image of the 4 different websites, which one would you
most likely check out (please be honest).  \
0                 Science: Latest News and Headlines
1                 Science: Latest News and Headlines
2                 Science: Latest News and Headlines
3                 Science: Latest News and Headlines
```

```
4                                                  ESPN

   [York Peppermint Patties] Ignore
0                                NaN
1                                NaN
2                                NaN
3                                NaN
4                                NaN

[5 rows x 123 columns]
```

[72]:
```
# Retrieve the data type of the Timestamp column
dataTypeObj = df_bbch_2016.dtypes['Timestamp']
print(dataTypeObj)
```

```
datetime64[ns]
```

[73]:
```
#Adding day of the week field as extra column in dataframe
day_of_week = df_bbch_2016['Timestamp'].apply(lambda x: x.strftime('%A'))

# Inserting day_of_week as 2nd column in dataframe
df_bbch_2016.insert(loc=1,column='Day_of_Week',value=day_of_week)


#
df_bbch_2016.head
```

[73]:
```
<bound method NDFrame.head of                    Timestamp Day_of_Week  \
0    2016-10-24 05:09:23.033      Monday
1    2016-10-24 05:09:54.798      Monday
2    2016-10-24 05:13:06.734      Monday
3    2016-10-24 05:14:17.192      Monday
4    2016-10-24 05:14:24.625      Monday
…                        …         …
1254 2016-10-29 16:53:52.516    Saturday
1255 2016-10-30 06:53:54.735      Sunday
1256 2016-10-30 11:06:10.827      Sunday
1257 2016-10-30 16:07:26.539      Sunday
1258 2016-10-30 17:06:45.660      Sunday

     Are you going actually going trick or treating yourself? Your gender:  \
0                                                    No                Male
1                                                    No                Male
2                                                    No              Female
3                                                    No                Male
4                                                   Yes                Male
…                                                     …                 …
1254                                                 No              Female
```

```
1255                                                  No               Male
1256                                                  No               Male
1257                                                  No               Male
1258                                                 Yes             Female

     How old are you? Which country do you live in?  \
0                22.0                         Canada
1                45.0                            usa
2                48.0                             US
3                57.0                            usa
4                42.0                            USA
…                …                                …
1254             52.0                            USA
1255             33.0                  united states
1256              NaN                            NaN
1257             48.0                         canada
1258             44.0                             Us

     Which state, province, county do you live in?  [100 Grand Bar]  \
0                                          Ontario              JOY
1                                               il              MEH
2                                         Colorado              JOY
3                                               il              JOY
4                                      South Dakota              MEH
…                                                 …              …
1254                                            TX              JOY
1255                                     minnesota              JOY
1256                                           NaN              JOY
1257                                            BC              NaN
1258                                            Nh              JOY

      [Anonymous brown globs that come in black and orange wrappers]  \
0                                                  DESPAIR
1                                                      MEH
2                                                  DESPAIR
3                                                      MEH
4                                                  DESPAIR
…                                                    …
1254                                               DESPAIR
1255                                               DESPAIR
1256                                                   MEH
1257                                               DESPAIR
1258                                                   MEH

      [Any full-sized candy bar]  …  \
0                            JOY  …
1                            JOY  …
```

```
2                            JOY  …
3                            JOY  …
4                            JOY  …
…                          …  …
1254                         JOY  …
1255                         JOY  …
1256                         JOY  …
1257                         JOY  …
1258                         JOY  …

     Please estimate the degree(s) of separation you have from the following
celebrities [JK Rowling]  \
0                                          3 or higher
1                                          3 or higher
2                                          3 or higher
3                                          3 or higher
4                                          3 or higher
…                                               …
1254                                       3 or higher
1255                                Actually, that's me.
1256                                              NaN
1257                                              1.0
1258                                       3 or higher

     Please estimate the degree(s) of separation you have from the following
celebrities [JJ Abrams]  \
0                                                 2.0
1                                          3 or higher
2                                          3 or higher
3                                          3 or higher
4                                          3 or higher
…                                               …
1254                                       3 or higher
1255                                       3 or higher
1256                                              NaN
1257                                              2.0
1258                                       3 or higher

     Please estimate the degree(s) of separation you have from the following
celebrities [Beyoncé]  \
0                                          3 or higher
1                                          3 or higher
2                                          3 or higher
3                                          3 or higher
4                                          3 or higher
…                                               …
1254                                       3 or higher
```

```
1255                                          3 or higher
1256                                                 NaN
1257                                          3 or higher
1258                                          3 or higher

     Please estimate the degree(s) of separation you have from the following
celebrities [Bieber]  \
0                                             3 or higher
1                                             3 or higher
2                                             3 or higher
3                                             3 or higher
4                                             3 or higher
…                                                    …
1254                                          3 or higher
1255                                          3 or higher
1256                                                 NaN
1257                                          3 or higher
1258                                          3 or higher

     Please estimate the degree(s) of separation you have from the following
celebrities [Kevin Bacon]  \
0                                             3 or higher
1                                             3 or higher
2                                             3 or higher
3                                             3 or higher
4                                             3 or higher
…                                                    …
1254                                                 2.0
1255                                          3 or higher
1256                                                 NaN
1257                                                 2.0
1258                                          3 or higher

     Please estimate the degree(s) of separation you have from the following
celebrities [Francis Bacon (1561 - 1626)]  \
0                                             3 or higher
1                                             3 or higher
2                                             3 or higher
3                                             3 or higher
4                                             3 or higher
…                                                    …
1254                                          3 or higher
1255                                Actually, that's me.
1256                                                 NaN
1257                                          3 or higher
1258                                          3 or higher
```

```
     Which day do you prefer, Friday or Sunday?  \
0                                         Friday
1                                         Friday
2                                         Sunday
3                                         Sunday
4                                         Friday
…                                          …
1254                                      Friday
1255                                      Friday
1256                                      Sunday
1257                                      Sunday
1258                                      Sunday

     Do you eat apples the correct way, East to West (side to side) or do you
eat them like a freak of nature, South to North (bottom to top)?  \
0                                      South to North
1                                        East to West
2                                        East to West
3                                      South to North
4                                        East to West
…                                                 …
1254                                     East to West
1255                  Sinusoidally around the equator
1256                     nne to east to nnw to s to n
1257                                     East to West
1258                                     East to West

     When you see the above image of the 4 different websites, which one would
you most likely check out (please be honest).  \
0                    Science: Latest News and Headlines
1                    Science: Latest News and Headlines
2                    Science: Latest News and Headlines
3                    Science: Latest News and Headlines
4                                                 ESPN
…                                                 …
1254                 Science: Latest News and Headlines
1255                 Science: Latest News and Headlines
1256                 Science: Latest News and Headlines
1257                 Science: Latest News and Headlines
1258                                       Daily Dish

      [York Peppermint Patties] Ignore
0                                  NaN
1                                  NaN
2                                  NaN
3                                  NaN
4                                  NaN
```

```
…                                   …
1254                                  NaN
1255                                  NaN
1256                                  NaN
1257                                  NaN
1258                                  NaN

[1259 rows x 124 columns]>
```

[74]:
```
# Print the date/Time range of data
print(f" Data collection Start Date/Time : {df_bbch_2016['Timestamp'].min()}")
print(f" Data collection End Date/Time   : {df_bbch_2016['Timestamp'].max()}")
```

```
Data collection Start Date/Time : 2016-10-24 05:09:23.033000
Data collection End Date/Time   : 2016-10-30 17:06:45.660000
```

[75]:
```
## converting the timestamp column into periods based on the time of the day,␣
 ↪specifically dividing it into four segments: Morning, Afternoon, Evening,␣
 ↪and Night.
# Convert timestamps to periods and back
session=pd.cut(df_bbch_2016.Timestamp.dt.hour,[0,6,12,18,23],
               ␣
 ↪labels=['Night','Morning','Afternoon','Evening'],include_lowest=True)

session.value_counts()
## --> This will provide valuable insights into the distribution of survey␣
 ↪responses across different times of the day, helping us understand patterns␣
 ↪and trends related to response rates and preferences during various segments␣
 ↪of the day.
```

[75]:
```
Morning      586
Afternoon    362
Night        218
Evening       93
Name: Timestamp, dtype: int64
```