KausikChattapadhyay_DSC540_Week7-8Ex

DSC 540 Week 7-8

Kausik Chattapadhyay

Activity : For this assignment you need to complete 8 of the following exercises against this data.

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

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

Chapter 10

• Grouping with Dicts/Series

• Grouping with Functions

• Grouping with Index Levels

• Split/Apply/Combine

• Cross Tabs

Chapter 11

• Convert between string and date time

• Generate date range

• Frequencies and date offsets

• Convert timestamps to periods and back

• Period Frequency conversions

In [1]:
```python
# Load the necessary libraries.

import numpy as np
import pandas as pd
```

In [2]:
```python
# Reading the csv file and exploring contents

candyDF = pd.read_excel('candyhierarchy2017.xlsx')
candyDF.head()
```

```
/Users/kausik/opt/anaconda3/lib/python3.9/site-packages/openpyxl/worksheet/_read_only.py:79: UserWarning: Unknown extension is not supported and will be removed
  for idx, row in parser.parse():
```

Out[2]:

	Internal ID	Q1: GOING OUT?	Q2: GENDER	Q3: AGE	Q4: COUNTRY	Q5: STATE, PROVINCE, COUNTY, ETC	Q6 | 100 Grand Bar	Q6 | Anonymous brown globs that come in black and orange wrappers\t(a.k.a. Mary Janes)	Q6 | Any full-sized candy bar	Q6 | Black Jacks	...	Q8: DESPAIR OTHER	Q9: OTHER COMMENTS	Q10: DRESS	Unnamed: 113	Q11: DAY	Q12: MEDIA [Daily Dish]	Q12: MEDIA [Science]	Q12: MEDIA [ESPN]	Q12: MEDIA [Yahoo]	Coordin
0	90258773	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
1	90272821	No	Male	44	USA	NM	MEH	DESPAIR	JOY	MEH	...	NaN	Bottom line is Twix is really the only candy w...	White and gold	NaN	Sunday	NaN	1.0	NaN	NaN	(84
2	90272829	NaN	Male	49	USA	Virginia	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
3	90272840	No	Male	40	us	or	MEH	DESPAIR	JOY	MEH	...	NaN	Raisins can go to hell	White and gold	NaN	Sunday	NaN	1.0	NaN	NaN	(75
4	90272841	No	Male	23	usa	exton pa	JOY	DESPAIR	JOY	DESPAIR	...	NaN	NaN	White and gold	NaN	Friday	NaN	1.0	NaN	NaN	(70

5 rows × 120 columns

In [3]:
```python
candyDF.columns
```

Out[3]:
```
Index(['Internal ID', 'Q1: GOING OUT?', 'Q2: GENDER', 'Q3: AGE', 'Q4: COUNTRY',
       'Q5: STATE, PROVINCE, COUNTY, ETC', 'Q6 | 100 Grand Bar',
       'Q6 | Anonymous brown globs that come in black and orange wrappers\t(a.k.a. Mary Janes)',
       'Q6 | Any full-sized candy bar', 'Q6 | Black Jacks',
       ...
       'Q8: DESPAIR OTHER', 'Q9: OTHER COMMENTS', 'Q10: DRESS', 'Unnamed: 113',
       'Q11: DAY', 'Q12: MEDIA [Daily Dish]', 'Q12: MEDIA [Science]',
       'Q12: MEDIA [ESPN]', 'Q12: MEDIA [Yahoo]', 'Click Coordinates (x, y)'],
      dtype='object', length=120)
```

In [4]:
```python
# Renaming partial columns

candyDF = candyDF.rename(columns = {'Q1: GOING OUT?' : 'going_out', 'Q2: GENDER' : 'gender', 'Q3: AGE': 'age',
                                    'Q4: COUNTRY' : 'country', 'Q5: STATE, PROVINCE, COUNTY, ETC' : 'area',
                                    'Q10: DRESS' : 'dress', 'Q11: DAY': 'day','Q12: MEDIA [Daily Dish]' : 'media_DailyDish',
                                    'Q12: MEDIA [Science]': 'media_Science', 'Q12: MEDIA [ESPN]' : 'media_ESPN',
                                    'Q12: MEDIA [Yahoo]': 'media_Yahoo'})
candyDF.columns
```

Out[4]:
```
Index(['Internal ID', 'going_out', 'gender', 'age', 'country', 'area',
       'Q6 | 100 Grand Bar',
       'Q6 | Anonymous brown globs that come in black and orange wrappers\t(a.k.a. Mary Janes)',
       'Q6 | Any full-sized candy bar', 'Q6 | Black Jacks',
       ...
       'Q8: DESPAIR OTHER', 'Q9: OTHER COMMENTS', 'dress', 'Unnamed: 113',
       'day', 'media_DailyDish', 'media_Science', 'media_ESPN', 'media_Yahoo',
       'Click Coordinates (x, y)'],
      dtype='object', length=120)
```

In [5]:
```python
# Dropping non required columns

candyDF.drop(columns = ['Internal ID','Unnamed: 113', 'Click Coordinates (x, y)'], inplace = True)
candyDF.shape
```

Out[5]:
```
(2460, 117)
```

In [6]:
```python
# Handling null values

candyDF.dropna(subset = ['going_out', 'gender', 'age', 'country', 'area'], how = 'all', inplace = True)
candyDF.reset_index(drop = True, inplace = True)
candyDF.shape
```

Out[6]:
```
(2435, 117)
```

In [7]:
```python
# Formatting Columns
# Going Out Column

candyDF.going_out = candyDF.going_out.fillna('Not Sure')
candyDF.going_out.unique()
```

Out[7]:
```
array(['No', 'Not Sure', 'Yes'], dtype=object)
```

In [8]:
```python
# Gender Column

candyDF.gender.value_counts()
```

Out[8]:
```
Male                  1467
Female                 839
I'd rather not say      83
Other                   30
Name: gender, dtype: int64
```

In [9]:
```python
# Adding NaN genders to type 3 - I'd rather not say, as it seems to be similar to unknown or NA
candyDF[candyDF.gender == "I'd rather not say"].shape
#checking for spaces in text - found none
candyDF.gender = candyDF.gender.fillna("I'd rather not say")
candyDF.gender.value_counts()
```

Out[9]:
```
Male                  1467
Female                 839
I'd rather not say      99
Other                   30
Name: gender, dtype: int64
```

In [10]:
```python
# Lets look at the country column and format

candyDF.country.unique()
```

Out[10]:
```
array(['USA ', 'USA', 'us', 'usa', nan, 'canada', 'Canada', 'Us', 'US',
       'Murica', 'United States', 'uk', 'United Kingdom', 'united states',
       'Usa', 'United States ', 'United staes',
       'United States of America', 'UAE', 'England', 'UK', 'canada ',
       'Mexico', 'United states', 'u.s.a.', 'USAUSAUSA', 'america', 35,
       'france', 'United States of America ', 'U.S.A.', 'finland',
       'unhinged states', 'Canada ', 'united states of america',
       'US of A', 'Unites States', 'The United States', 'North Carolina ',
       'Unied States', 'Netherlands', 'germany', 'Europe', 'Earth', 'U S',
       'u.s.', 'U.K. ', 'Costa Rica', 'The United States of America',
       'unite states', 'U.S.', 46, 'cascadia', 'Australia',
       'insanity lately', 'Greece', 'USA? Hard to tell anymore..',
       "'merica", 'usas', 'Pittsburgh', 45, 'United State', 32, 'France',
       'australia', 'A', 'Can', 'Canae', 'New York', 'Trumpistan',
       'Ireland', 'United Sates', 'Korea', 'California', 'Japan', 'USa',
       'South africa',
       'I pretend to be from Canada, but I am really from the United States.',
       'Usa ', 'Uk', 'Iceland', 'Germany', 'Canada`', 'Scotland', 'UK ',
       'Denmark', 'United Stated', 'France ', 'Switzerland',
       'Ahem....Amerca', 'UD', 'Scotland ', 'South Korea', 'New Jersey',
       'CANADA', 'Indonesia', 'United ststes', 'America',
       'The Netherlands', 'United Statss', 'endland', 'Atlantis',
       'murrika', 'USA! USA! USA!', 'USAA', 'Alaska', 'united States ',
       'soviet canuckistan', 'N. America', 'Singapore', 'USSA', 'China',
       'Taiwan', 'Ireland ', 'hong kong', 'spain', 'Sweden', 'Hong Kong',
       'U.S. ', 'Narnia', 'u s a', 'United Statea', 'united ststes', 1,
       'subscribe to dm4uz3 on youtube', 'United kingdom',
       'USA USA USA!!!!', "I don't know anymore", 'Fear and Loathing'],
      dtype=object)
```

In [11]:
```python
candyDF.country.value_counts(dropna = False).sort_values(ascending = False)
```

Out[11]:
```
USA                      699
United States            497
usa                      217
Canada                   179
Usa                      139
                        ...
United kingdom             1
USA USA USA!!!!            1
I don't know anymore       1
France                     1
Fear and Loathing          1
Name: country, Length: 129, dtype: int64
```

In [12]:
```python
candyDF.country = candyDF.country.fillna('Unknown')
```

In [13]:
```python
set([x for x in candyDF.country if 'u' in str(x)])   # unique values with 'u'
```

Out[13]:
```
{'Australia',
 'Europe',
 'I pretend to be from Canada, but I am really from the United States.',
 'Murica',
 'Pittsburgh',
 'South Korea',
 'South africa',
 'Trumpistan',
 'australia',
 'murrika',
 'soviet canuckistan',
 'subscribe to dm4uz3 on youtube',
 'u s a',
 'u.s.',
 'u.s.a.',
 'uk',
 'unhinged states',
 'unite states',
 'united States ',
 'united states',
 'united states of america',
 'united ststes',
 'us',
 'usa',
 'usas'}
```

In [14]:
```python
USA = [x for x in candyDF.country if (('u' in str(x) or 'U' in str(x)) and
        'ingdom' not in str(x) and 'urope' not in str(x) and 'stralia' not in str(x) and 'South Korea' not in str(x)
        and 'South africa' not in str(x) and 'uk' not in str(x))]

candyDF.country = candyDF.country.replace(to_replace = USA, value = 'USA')
candyDF.country.unique()
```

Out[14]:
```
array(['USA', 'canada', 'Canada', 'uk', 'United Kingdom', 'England',
       'canada ', 'Mexico', 'america', 35, 'france', 'finland', 'Canada ',
       'North Carolina ', 'Netherlands', 'germany', 'Europe', 'Earth',
       'Costa Rica', 46, 'cascadia', 'Australia', 'insanity lately',
       'Greece', "'merica", 45, 32, 'France', 'australia', 'A', 'Can',
       'Canae', 'New York', 'Ireland', 'Korea', 'California', 'Japan',
       'South africa', 'Iceland', 'Germany', 'Canada`', 'Scotland',
       'Denmark', 'France ', 'Switzerland', 'Ahem....Amerca', 'Scotland ',
       'South Korea', 'New Jersey', 'CANADA', 'Indonesia', 'America',
       'The Netherlands', 'endland', 'Atlantis', 'Alaska', 'N. America',
       'Singapore', 'China', 'Taiwan', 'Ireland ', 'hong kong', 'spain',
       'Sweden', 'Hong Kong', 'Narnia', 1, 'United kingdom',
       "I don't know anymore", 'Fear and Loathing'], dtype=object)
```

In [26]:
```python
# removing duplicates , wrongly or differently names and updating

candyDF.country = candyDF.country.replace(to_replace = ['america','Ahem....Amerca',"'merica",'North Carolina ',
                'cascadia', 'New York','A','California','New Jersey','America','Alaska','N. America'],
                                          value ='USA')
canada = [x for x in candyDF.country if 'anada' in str(x).strip() or 'ANADA' in str(x) or 'Can' in str(x)]

candyDF.country = candyDF.country.replace(to_replace = canada, value = 'Canada')
candyDF.country = candyDF.country.replace(to_replace = ['uk', 'United kingdom', 'England','endland'],
                                          value = 'United Kingdom')
candyDF.country = candyDF.country.replace(to_replace = 'germany', value = 'Germany')
candyDF.country = candyDF.country.replace(to_replace = 'The Netherlands', value = 'Netherlands')
candyDF.country = candyDF.country.replace(to_replace = 'australia', value = 'Australia')
candyDF.country = candyDF.country.replace(to_replace = [1,"I don't know anymore",32,45,35,46,'Fear and Loathing','insanity lately'],
                                          value = 'Others')
candyDF.country.value_counts()
```

Out[26]:
```
USA               2111
Canada             227
United Kingdom      18
Germany             10
Others               8
Netherlands          7
Australia            7
Japan                5
Scotland             4
Mexico               4
Ireland              3
Switzerland          3
China                2
Denmark              2
Sweden               2
South Korea          1
Taiwan               1
Hong Kong            1
Singapore            1
Ireland              1
hong kong            1
Atlantis             1
spain                1
Indonesia            1
South africa         1
Scotland             1
France               1
Iceland              1
Korea                1
France               1
Greece               1
Costa Rica           1
Earth                1
Europe               1
finland              1
france               1
Narnia               1
Name: country, dtype: int64
```

In [27]:
```python
# Grouping Dataset to 3 Countries - USA, Canada, Others

# creating a dataset Other from existing one which we will merge later

other = [x for x in candyDF.country.unique()]
other.remove('USA')
other.remove('Canada')
other
```

Out[27]:
```
['United Kingdom',
 'Mexico',
 'Others',
 'france',
 'finland',
 'Netherlands',
 'Germany',
 'Europe',
 'Earth',
 'Costa Rica',
 'Australia',
 'Greece',
 'France',
 'Ireland',
 'Korea',
 'Japan',
 'South africa',
 'Iceland',
 'Scotland',
 'Denmark',
 'France ',
 'Switzerland',
 'Scotland ',
 'South Korea',
 'Indonesia',
 'Atlantis',
 'Singapore',
 'China',
 'Taiwan',
 'Ireland ',
 'hong kong',
 'spain',
 'Sweden',
 'Hong Kong',
 'Narnia']
```

In [28]:
```python
candyDF.country = candyDF.country.replace(to_replace = other, value = 'Other')
candyDF.country.value_counts()
```

Out[28]:
```
USA       2111
Canada     227
Other       97
Name: country, dtype: int64
```

In [29]:
```python
candyDF.columns
```

Out[29]:
```
Index(['going_out', 'gender', 'age', 'country', 'area', 'Q6 | 100 Grand Bar',
       'Q6 | Anonymous brown globs that come in black and orange wrappers\t(a.k.a. Mary Janes)',
       'Q6 | Any full-sized candy bar', 'Q6 | Black Jacks',
       'Q6 | Bonkers (the candy)',
       ...
       'Q6 | York Peppermint Patties', 'Q7: JOY OTHER', 'Q8: DESPAIR OTHER',
       'Q9: OTHER COMMENTS', 'dress', 'day', 'media_DailyDish',
       'media_Science', 'media_ESPN', 'media_Yahoo'],
      dtype='object', length=117)
```

In [30]:
```python
# Converting Datatype
candyDF = candyDF.astype({'going_out':'category', 'gender':'category', 'country':'category', 'dress':'category',
                          'day':'category'})
candyDF.describe(include = 'category')
```

Out[30]:

	going_out	gender	country	dress	day
count	2435	2435	2435	1714	1733
unique	3	4	3	2	2
top	No	Male	USA	White and gold	Friday
freq	2038	1467	2111	1080	1089

In [31]:
```python
# Method to Convert 4 Columns into one

def melt1(row):
    for c in data.columns:
        if row[c] == 1:
            return c
```

In [32]:
```python
# Checking Media columsn which we will merge into one

data = candyDF[candyDF.columns[-4:]]
data
```

Out[32]:

	media_DailyDish	media_Science	media_ESPN	media_Yahoo
0	NaN	1.0	NaN	NaN
1	NaN	NaN	NaN	NaN
2	NaN	1.0	NaN	NaN
3	NaN	1.0	NaN	NaN
4	NaN	1.0	NaN	NaN
...	...	...	...	...
2430	NaN	NaN	NaN	NaN
2431	NaN	1.0	NaN	NaN
2432	NaN	1.0	NaN	NaN
2433	NaN	NaN	NaN	NaN
2434	1.0	NaN	NaN	NaN

2435 rows × 4 columns

In [33]:
```python
new_col = data.apply(melt1, axis = 1)

# Adding newly created column
candyDF['media_preference'] = new_col

#dropping old columns
candyDF.drop(columns =['media_DailyDish','media_Science','media_ESPN','media_Yahoo'], inplace = True)

candyDF.media_preference.value_counts(dropna = False)
```

Out[33]:
```
media_Science      1361
None                824
media_ESPN           99
media_DailyDish      84
media_Yahoo          67
Name: media_preference, dtype: int64
```

In [34]:
```python
# Getting personnal info and questionaire columns into separate dataframes

personal_info_cols = candyDF.columns[:6]
questionare_cols = candyDF.columns[5:]
candyDF.columns
```

Out[34]:
```
Index(['going_out', 'gender', 'age', 'country', 'area', 'Q6 | 100 Grand Bar',
       'Q6 | Anonymous brown globs that come in black and orange wrappers\t(a.k.a. Mary Janes)',
       'Q6 | Any full-sized candy bar', 'Q6 | Black Jacks',
       'Q6 | Bonkers (the candy)',
       ...
       'Q6 | Whatchamacallit Bars', 'Q6 | White Bread',
       'Q6 | Whole Wheat anything', 'Q6 | York Peppermint Patties',
       'Q7: JOY OTHER', 'Q8: DESPAIR OTHER', 'Q9: OTHER COMMENTS', 'dress',
       'day', 'media_preference'],
      dtype='object', length=114)
```

In [35]:
```python
responses = len(questionare_cols) - candyDF[questionare_cols].isna().sum(axis =1)
candyDF['responses'] = responses
candyDF.head(3)
```

Out[35]:

	going_out	gender	age	country	area	Q6 | 100 Grand Bar	Q6 | Anonymous brown globs that come in black and orange wrappers\t(a.k.a. Mary Janes)	Q6 | Any full-sized candy bar	Q6 | Black Jacks	Q6 | Bonkers (the candy)	...	Q6 | White Bread	Q6 | Whole Wheat anything	Q6 | York Peppermint Patties	Q7: JOY OTHER	Q8: DESPAIR OTHER	Q9: OTHER COMMENTS	dress	day	media_preference	respo
0	No	Male	44	USA	NM	MEH	DESPAIR	JOY	MEH	DESPAIR	...	DESPAIR	DESPAIR	DESPAIR	Mounds	NaN	Bottom line is Twix is really the only candy w...	White and gold	Sunday	media_Science	
1	Not Sure	Male	49	USA	Virginia	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	None	
2	No	Male	40	USA	or	MEH	DESPAIR	JOY	MEH	MEH	...	DESPAIR	DESPAIR	DESPAIR	Reese's crispy crunchy bars, 5th avenue bars, ...	NaN	Raisins can go to hell	White and gold	Sunday	media_Science	

3 rows × 115 columns

In [38]:
```python
candyDF['age'].unique()
```

Out[38]:
```
array([44, 49, 40, 23, nan, 53, 33, 43, 56, 64, 37, 59, 48, 54, 36, 45,
       25, 34, 35, 38, 58, 50, 47, 16, 52, 63, 65, 41, 27, 31, 61, 46, 42,
       62, 29, 39, 32, 28, 69, 67, 30, 22, 26, 51, 70, 24, 18, 19, 57, 60,
       66, 12, 55, 72, 21, 11, 9, 68, 20, 6, 10, 71, 90, 13, 99, 7, 88,
       39.4, 74, 102, 17, 15, 8, 75, 'See question 2', 14, 100, 76, 77,
       73, 70.5, 1, 4], dtype=object)
```

In [39]:
```python
# Data Type Conversion

candyDF.drop_duplicates(inplace=True)
s = pd.to_numeric(candyDF['age'],downcast='float',errors='ignore')
s=pd.to_numeric(s,downcast='float',errors='coerce')
candyDF['age'].unique()
candyDF.replace(candyDF['age'],s,inplace=True)
candyDF['age'].replace(['old enough','45-55','24-50','?','no','Many','hahahahaha','older than dirt','Enough',
                        'See question 2','old','ancient','old enough'],np.nan,inplace=True)

candyDF['age'].replace(['5u','46 Halloweens.','sixty-nine','Over 50','OLD','MY NAME JEFF','59 on the day after Halloween',
                        'your mom' 'I can remember when Java was a cool new language','60+'],np.nan,inplace=True)

candyDF['age'].replace([312,1000,'Old enough','your mom','I can remember when Java was a cool new language'],np.nan,inplace=True)

pd.to_numeric(candyDF['age']).head()
```

Out[39]:
```
0    44.0
1    49.0
2    40.0
3    23.0
4     NaN
Name: age, dtype: float64
```

In [40]:
```python
nam=candyDF.columns
nam
```

Out[40]:
```
Index(['going_out', 'gender', 'age', 'country', 'area', 'Q6 | 100 Grand Bar',
       'Q6 | Anonymous brown globs that come in black and orange wrappers\t(a.k.a. Mary Janes)',
       'Q6 | Any full-sized candy bar', 'Q6 | Black Jacks',
       'Q6 | Bonkers (the candy)',
       ...
       'Q6 | White Bread', 'Q6 | Whole Wheat anything',
       'Q6 | York Peppermint Patties', 'Q7: JOY OTHER', 'Q8: DESPAIR OTHER',
       'Q9: OTHER COMMENTS', 'dress', 'day', 'media_preference', 'responses'],
      dtype='object', length=115)
```

In [ ]: