AbhijitMandal_DSC540_Week7-8Ex

May 2, 2021

0.0.1 DSC 540 Week 7-8

Abhijit Mandal

0.0.2 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

0.0.3 Load the necessary libraries.

[115]:
```
import numpy as np
import pandas as pd
```

0.0.4 Reading the csv file and exploring contents

[116]:
```
candyDF = pd.read_excel('candyhierarchy2017.xlsx')
candyDF.head()
```

[116]:
```
   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]
```

[117]:
```
candyDF.columns
```

[117]:
```
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)
```

0.0.5 Renaming partial columns

[118]:
```
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
```

[118]:
```
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)
```

0.0.6 Dropping non required columns

[119]:
```
candyDF.drop(columns = ['Internal ID','Unnamed: 113', 'Click Coordinates (x,␣
↪y)'], inplace = True)
candyDF.shape
```

[119]:
```
(2460, 117)
```

0.0.7 Handling null values

[120]:
```
candyDF.dropna(subset = ['going_out', 'gender', 'age', 'country', 'area'], how␣
↪= 'all', inplace = True)
candyDF.reset_index(drop = True, inplace = True)
candyDF.shape
```

[120]:
```
(2435, 117)
```

0.0.8 Formatting Columns

[121]:
```
# Going Out Column
candyDF.going_out = candyDF.going_out.fillna('Not Sure')
candyDF.going_out.unique()
```

[121]:
```
array(['No', 'Not Sure', 'Yes'], dtype=object)
```

[122]:
```
# Gender Column
candyDF.gender.value_counts()
```

[122]:
```
Male                   1467
Female                  839
I'd rather not say       83
Other                    30
Name: gender, dtype: int64
```

[123]:
```
# 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()
```

[123]:
```
Male                  1467
Female                 839
I'd rather not say      99
Other                   30
Name: gender, dtype: int64
```

[124]:
```
# Lets look at the country column and format
candyDF.country.unique()
```

[124]:
```
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)
```

[125]:
```
candyDF.country.value_counts(dropna = False).sort_values(ascending = False)
```

[125]:
```
USA              699
United States    497
usa              217
Canada           179
Usa              139
                 …
```

```
USAUSAUSA          1
murrika            1
Canae              1
'merica            1
1                  1
Name: country, Length: 129, dtype: int64
```

[126]:
```
candyDF.country = candyDF.country.fillna('Unknown')
```

[127]:
```
set([x for x in candyDF.country if 'u' in str(x)])  # unique values with 'u'
```

[127]:
```
{'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'}
```

[128]:
```
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()
```

[128]:
```
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…America', '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)
```

[129]:
```
# removing duplicates , wrongly or differently names and updating
candyDF.country = candyDF.country.replace(to_replace = ['america','Ahem....
 ↪America',"'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.value_counts()
```

[129]:
```
USA                    2111
Canada                  227
United Kingdom           13
Germany                   7
Netherlands               6
Australia                 5
Japan                     5
Scotland                  4
Mexico                    4
germany                   3
Ireland                   3
Switzerland               3
Sweden                    2
uk                        2
China                     2
australia                 2
Denmark                   2
South africa              1
Fear and Loathing         1
Singapore                 1
England                   1
Taiwan                    1
```

```
I don't know anymore       1
France                     1
Greece                     1
46                         1
Costa Rica                 1
45                         1
35                         1
The Netherlands            1
Korea                      1
32                         1
hong kong                  1
United kingdom             1
Europe                     1
Atlantis                   1
Iceland                    1
spain                      1
endland                    1
Earth                      1
South Korea                1
finland                    1
Hong Kong                  1
France                     1
Ireland                    1
Indonesia                  1
Scotland                   1
insanity lately            1
france                     1
Narnia                     1
1                          1
Name: country, dtype: int64
```

0.0.9 Grouping Dataset to 3 Countries - USA, Canada, Others

[130]:
```
# 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
```

[130]:
```
['uk',
 'United Kingdom',
 'England',
 'Mexico',
 35,
 'france',
 'finland',
 'Netherlands',
```

```
 'germany',
 'Europe',
 'Earth',
 'Costa Rica',
 46,
 'Australia',
 'insanity lately',
 'Greece',
 45,
 32,
 'France',
 'australia',
 'Ireland',
 'Korea',
 'Japan',
 'South africa',
 'Iceland',
 'Germany',
 'Scotland',
 'Denmark',
 'France ',
 'Switzerland',
 'Scotland ',
 'South Korea',
 'Indonesia',
 'The Netherlands',
 'endland',
 'Atlantis',
 'Singapore',
 'China',
 'Taiwan',
 'Ireland ',
 'hong kong',
 'spain',
 'Sweden',
 'Hong Kong',
 'Narnia',
 1,
 'United kingdom',
 "I don't know anymore",
 'Fear and Loathing']
```

[131]:
```
candyDF.country = candyDF.country.replace(to_replace = other, value = 'Other')
candyDF.country.value_counts()
```

[131]:
```
USA       2111
Canada     227
```

```
Other       97
Name: country, dtype: int64
```

[132]:
```
candyDF.columns
```

[132]:
```
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)
```

0.0.10 Converting Datatype

[133]:
```
candyDF = candyDF.astype({'going_out':'category', 'gender':'category',␣
↪'country':'category', 'dress':'category', 'day':'category'})
candyDF.describe(include = 'category')
```

[133]:

	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

[134]:
```
# Method to Convert 4 Columns into one
def melt1(row):
    for c in data.columns:
        if row[c] == 1:
            return c
```

[135]:
```
# Checking Media columsn which we will merge into one
data = candyDF[candyDF.columns[-4:]]
data
```

[135]:

	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 x 4 columns]
```

[136]:
```
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)
```

[136]:
```
media_Science      1361
NaN                 824
media_ESPN           99
media_DailyDish      84
media_Yahoo          67
Name: media_preference, dtype: int64
```

0.0.11 Getting personnal info and questionaire columns into separate dataframes

[137]:
```
personal_info_cols = candyDF.columns[:6]
questionare_cols = candyDF.columns[5:]
candyDF.columns
```

[137]:
```
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)
```

[138]:
```
responses = len(questionare_cols) - candyDF[questionare_cols].isna().sum(axis =␣
 ↪1)
candyDF['responses'] = responses
candyDF.head(3)
```

[138]:
```
  going_out gender age country      area Q6 | 100 Grand Bar  \
0        No   Male  44     USA        NM                MEH
1  Not Sure   Male  49     USA  Virginia                NaN
2        No   Male  40     USA        or                MEH

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

  Q6 | Any full-sized candy bar Q6 | Black Jacks Q6 | Bonkers (the candy)  \
0                          JOY              MEH                   DESPAIR
1                          NaN              NaN                       NaN
2                          JOY              MEH                       MEH

   … Q6 | White Bread Q6 | Whole Wheat anything  \
0  …          DESPAIR                  DESPAIR
1  …              NaN                      NaN
2  …          DESPAIR                  DESPAIR

  Q6 | York Peppermint Patties  \
0                      DESPAIR
1                          NaN
2                      DESPAIR

                                    Q7: JOY OTHER Q8: DESPAIR OTHER  \
0                                          Mounds                NaN
1                                             NaN                NaN
2  Reese's crispy crunchy bars, 5th avenue bars, …                NaN

                                 Q9: OTHER COMMENTS           dress     day  \
0  Bottom line is Twix is really the only candy w…  White and gold  Sunday
1                                              NaN             NaN     NaN
2                         Raisins can go to hell  White and gold  Sunday

  media_preference responses
0    media_Science       108
1             None         0
2    media_Science       108

[3 rows x 115 columns]
```

0.0.12 Data Type Conversion

[139]:
```
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()
```

[139]:
```
0    44.0
1    49.0
2    40.0
3    23.0
4     NaN
Name: age, dtype: float64
```

[141]:
```
nam=candyDF.columns
nam
```

[141]:
```
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)
```

[ ]: