

In []:

```
"""
DSC540 Week 7 and 8 assignment Exercise
"""

from __future__ import print_function
from itertools import zip_longest

import csv
import logging
import sys
import numpy as np
import pandas as pd
import random
import thinkplot
import thinkstats2
import datetime
import regression
import statsmodels.formula.api as smf
import statsmodels.api as sm
import matplotlib.pyplot as plt
import math
from bs4 import BeautifulSoup
import pandas as pd
import sqlite3
%matplotlib inline

def ReadData(filename):
    # File was throwing some encoding error while trying to read with default utf-8 encoding
    # Using latin-1 encoding to fix issue
    # Missing values or spaces would be replaced by np.nan
    # I have downloaded the data and will use local file (So Much Data Candy, Seriously (Ng, 2017))
    na_values = ["", " ", None, "Missing", "NA"]
    df = pd.read_csv(filename, encoding='latin-1', na_values=na_values)
    df.head()
    return df

# To perform 2 tasks out of below list Chapter 7
# 1. Filter out missing data
# 2. Fill in missing data
# 3. Remove duplicates
# 4. Transform data using either mapping or a function
# 5. Replace values
# 6. Discretization and Binning
# 7. Manipulate Strings
# 8. Noticed that missing Values are automatically populated as NaN

def ExerciseChapter_7(df):
    df.set_index(['Internal ID'], inplace=True)
    df.head()

    # It shows that around 21 rows dropped as it didn't had any value populated.
    # Check if there are any duplicate records.
    print(df.duplicated().sum())

    df.drop_duplicates(inplace=True)
    df.shape
    # print(df)

    # There are few columns for which more than 50% records have missing values i.e. NaN.
    # I don't think those columns would be of any use, hence dropping those columns.
    # Dropping columns where more than 50% records are NaN
    df.dropna(thresh=df.shape[0]*0.5, axis=1, inplace=True)
    df.shape

    # Dropping rows where more than 50% records are NaN
    df.dropna(thresh=df.shape[1]*0.5, axis=0, inplace=True)
    df.shape
    # print(df)

    # Country names are missing at some places and also in different cases like US, us and some places usa
    df['Q4: COUNTRY'].unique()
    # There are too many 129 countries with varying names. I will just try to correct country name for US.
    # First changing all contry names to uppercase.
    df['Q4: COUNTRY'] = df['Q4: COUNTRY'].str.upper()
    df['Q4: COUNTRY'].unique()
    print("df['Q4: COUNTRY'].unique().size Value :", df['Q4: COUNTRY'].unique().size)

    usa_names_map = {
        'USA': 'USA',
        'USA ': 'USA',
        'US': 'USA',
        'UNITED STATES': 'USA',
        'U.S.': 'USA',
        'UNITED STATES OF AMERICA': 'USA',
        'US OF A': 'USA',
        'U.S.A.': 'USA',
        'USAUSAUSA': 'USA',
    }
```

```

'THE UNITED STATES': 'USA',
'UNIED STATES': 'USA',
'USA! USA! USA!': 'USA',
'USAA': 'USA',
'AMERICA': 'USA',
'UNHINGED STATES': 'USA',
'THE UNITED STATES OF AMERICA': 'USA',
'UNITE STATES': 'USA',
'UNITED SATES': 'USA',
'UNITES STATES': 'USA',
'UNITED STSTES': 'USA',
'UNITED STATES ': 'USA',
'UNITED STATES OF AMERICA ': 'USA',
'U S': 'USA',
'UNITED STATED': 'USA',
'USA USA USA!!!!': 'USA' }
df['Q4: COUNTRY'] = df['Q4: COUNTRY'].map(usa_names_map).fillna(df['Q4: COUNTRY'])
df['Q4: COUNTRY'].unique()

print("df['Q4: COUNTRY'].unique().size Value :",df['Q4: COUNTRY'].unique().size)

return df

# To perform below 2 tasks Chapter 8
# 1. Create hierarchical index
# 2. Reshape
def ExcerciseChapter_8(df):
    candy_names = {x:x.split('Q6 | ')[1] for x in df.columns if x.startswith('Q6 | ')}

    #candy_names = [ {x:x.split('Q6 | ')[1]} for x in df.columns if x.startswith('Q6 | ')]
    # #candy_names
    df.rename(columns=candy_names,inplace=True)
    df.reset_index(inplace=True)
    #df.set_index(list(candy_names.values()),inplace=True)
    #df.head()

    df.dropna(subset=['Q2: GENDER','Q3: AGE'],inplace=True)
    df.shape

    # Created Hierachial Data using columns Gender and Age
    df.set_index(['Q2: GENDER','Q3: AGE'],inplace=True)
    df.head()

    new_index = pd.Index(list(candy_names.values()),name='candy')
    new_df = df.reindex(columns=new_index)
    new_df.head()

    #print('new_df', new_df)

    ldata = new_df.stack().reset_index().rename(columns={0: 'candy_liking'})
    ldata[:10]

    yield df
    yield ldata

# To perform below 2 tasks Chapter 10
# 1. Grouping with Dicts/Series
# 2. Cross Tabs
def ExcerciseChapter_10(df, ldata):
    df.reset_index(inplace=True)
    df.head()
    df.shape

    print("ldata Value :",ldata)
    print("df2 Value :",df)

    # dropping records where either country or State/Province/City field have missing data
    df.dropna(subset=['Q4: COUNTRY','Q5: STATE, PROVINCE, COUNTY, ETC'],inplace=True)
    df.shape

    mapping = {
        'Q4: COUNTRY': 'COUNTRY',
        'Q5: STATE, PROVINCE, COUNTY, ETC': 'LOCATION'
    }

    by_column = df.groupby(mapping, axis=1)
    print('by_column Value :', by_column)

    print("by_column.describe() Value :", by_column.describe())

    # Now I would like to find count of each candy marked as JOY , MEH and DISPAIR. I will use CrossTab functio
    # I also want to find most liked Candy.
    # For this analysis , I will use dataframe created in Chapter 8 Exercise i.e. "ldata"
    ldata[:10]

    # Now Running crosstab function over candy and candy_liking column.
    pd.crosstab([ldata['candy']],ldata.candy_liking,margins=True)
    crosstab_df = pd.crosstab([ldata['candy']],ldata.candy_liking)

```

```

print(crosstab_df[crosstab_df['JOY'] == crosstab_df['JOY'].max()])

# To perform below 2 tasks Chapter 11
# 1. Convert between string and date time
# 2. Convert timestamps to periods and back
def ExerciseChapter_11():
    na_values = ["", " ", None, "Missing", "NA"]
    df_2016 = pd.read_excel("BOING-BOING-CANDY-HIERARCHY-2016-SURVEY-Responses.xlsx", na_values=na_values)
    df_2016.head()
    print("df_2016 Value :", df_2016)

    # Checking Data type of Timestamp column.
    dataTypeObj = df_2016.dtypes['Timestamp']
    print("dataTypeObj Value :", dataTypeObj)

    day_of_week = df_2016['Timestamp'].apply(lambda x: x.strftime('%A'))
    # Adding day_of_week as 2nd column in dataframe
    df_2016.insert(loc=1, column='Day_of_Week', value=day_of_week)
    df_2016.sample(5)

    print(f" Data collected is between date range {df_2016['Timestamp'].min()} - {df_2016['Timestamp'].max()}")

    # Convert timestamps to periods and back
    session = pd.cut(df_2016.Timestamp.dt.hour,
                    [0, 6, 12, 18, 23],
                    labels=['Night', 'Morning', 'Afternoon', 'Evening'],
                    include_lowest=True)
    print("session.value_counts() Value:", session.value_counts())

def main():
    print("Inside Main function")
    # function to read the So Much Data Candy, Seriously (Ng, 2017) data
    df = ReadData("candyhierarchy2017.csv")
    print("df.shape Value:", df.shape)
    print("df.isna().sum() Value:", df.isna().sum())

    # Perform 2 tasks out of below list Chapter 7
    df = ExerciseChapter_7(df)

    # Perform 2 tasks out of below list Chapter 8
    result = ExerciseChapter_8(df)

    df = next(result)
    ldata = next(result)
    print("next(result).1 Value:", df)
    print("next(result).2 Value :", ldata)

    # Perform 2 tasks out of below list Chapter 10
    ExerciseChapter_10(df, ldata)

    # Perform 2 tasks out of below list Chapter 11
    ExerciseChapter_11()

if __name__ == "__main__":
    main()

```

```

Inside Main function
df.shape Value: (2460, 120)
df.isna().sum() Value: 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
35
df['Q4: COUNTRY'].unique().size Value : 79
df['Q4: COUNTRY'].unique().size Value : 61
next(result).1 Value: Internal ID Q1: GOING OUT? Q4: COUNTRY \
Q2: GENDER Q3: AGE
Male      44      90272821      No      USA
          40      90272840      No      USA
          23      90272841      No      USA
          33      90272854      No     CANADA
          40      90272858      No     CANADA
...
Female    26      90314022      No      USA
Male      24      90314359      No      USA
Female    33      90314580      No      USA
          26      90314634      No      USA
          66      90314802      No      USA

```

Q5: STATE, PROVINCE, COUNTY, ETC 100 Grand Bar \

Q2: GENDER	Q3: AGE		
Male	44	NM	MEH
	40	or	MEH
	23	exton pa	JOY
	33	ontario	JOY
	40	Ontario	JOY
...	
Female	26	Michigan	JOY
Male	24	MD	JOY
Female	33	New York	MEH
	26	Tennessee	MEH
	66	Pennsylvania	DESPAIR

Anonymous brown globs that come in black and orange wrappers\t(a.k.a. Mary Janes) \

Q2: GENDER	Q3: AGE	
Male	44	DESPAIR
	40	DESPAIR
	23	DESPAIR
	33	DESPAIR
	40	DESPAIR
...		...
Female	26	MEH
Male	24	DESPAIR
Female	33	DESPAIR
	26	DESPAIR
	66	DESPAIR

Any full-sized candy bar Black Jacks Bonkers (the candy) \

Q2: GENDER	Q3: AGE			
Male	44	JOY	MEH	DESPAIR
	40	JOY	MEH	MEH
	23	JOY	DESPAIR	MEH
	33	JOY	DESPAIR	DESPAIR
	40	JOY	MEH	MEH
...	
Female	26	JOY	DESPAIR	MEH
Male	24	MEH	DESPAIR	DESPAIR
Female	33	JOY	NaN	NaN
	26	JOY	DESPAIR	MEH
	66	JOY	DESPAIR	DESPAIR

Bonkers (the board game) ... \

Q2: GENDER	Q3: AGE	
Male	44	DESPAIR ...
	40	DESPAIR ...
	23	DESPAIR ...
	33	MEH ...
	40	MEH ...
...		...
Female	26	MEH ...
Male	24	MEH ...
Female	33	NaN ...
	26	JOY ...
	66	DESPAIR ...

Vials of pure high fructose corn syrup, for main-lining into your vein \

Q2: GENDER	Q3: AGE	
Male	44	DESPAIR
	40	DESPAIR
	23	MEH
	33	JOY
	40	MEH
...		...
Female	26	MEH
Male	24	MEH
Female	33	NaN
	26	MEH
	66	DESPAIR

Vicodin Whatchamacallit Bars White Bread \

Q2: GENDER	Q3: AGE			
Male	44	DESPAIR	DESPAIR	DESPAIR
	40	JOY	JOY	DESPAIR
	23	JOY	JOY	DESPAIR
	33	MEH	DESPAIR	DESPAIR
	40	DESPAIR	MEH	DESPAIR
...	
Female	26	MEH	JOY	MEH
Male	24	JOY	DESPAIR	MEH
Female	33	NaN	JOY	DESPAIR
	26	JOY	MEH	DESPAIR
	66	JOY	DESPAIR	MEH

Whole Wheat anything York Peppermint Patties \

Q2: GENDER	Q3: AGE		
Male	44	DESPAIR	DESPAIR
	40	DESPAIR	DESPAIR
	23	DESPAIR	JOY
	33	DESPAIR	DESPAIR

	40	DESPAIR	DESPAIR
...	
Female	26	MEH	JOY
Male	24	DESPAIR	MEH
Female	33	MEH	JOY
	26	DESPAIR	MEH
	66	DESPAIR	JOY

Q10: DRESS Q11: DAY Q12: MEDIA [Science] \

Q2: GENDER	Q3: AGE			
Male	44	White and gold	Sunday	1.0
	40	White and gold	Sunday	1.0
	23	White and gold	Friday	1.0
	33	Blue and black	Friday	1.0
	40	Blue and black	Sunday	1.0
...	
Female	26	White and gold	Friday	1.0
Male	24	White and gold	Friday	NaN
Female	33	Blue and black	Friday	1.0
	26	Blue and black	Friday	1.0
	66	White and gold	Sunday	NaN

Click Coordinates (x, y)

Q2: GENDER	Q3: AGE	
Male	44	(84, 25)
	40	(75, 23)
	23	(70, 10)
	33	(55, 5)
	40	(76, 24)
...		...
Female	26	(68, 39)
Male	24	NaN
Female	33	(70, 26)
	26	(67, 35)
	66	(19, 26)

[1743 rows x 111 columns]

next(result).2 Value :	Q2: GENDER	Q3: AGE		candy \
0	Male	44	100 Grand Bar	
1	Male	44	Anonymous brown globs that come in black and o...	
2	Male	44	Any full-sized candy bar	
3	Male	44	Black Jacks	
4	Male	44	Bonkers (the candy)	
...	
171509	Female	66	Vicodin	
171510	Female	66	Whatchamacallit Bars	
171511	Female	66	White Bread	
171512	Female	66	Whole Wheat anything	
171513	Female	66	York Peppermint Patties	

candy_liking	
0	MEH
1	DESPAIR
2	JOY
3	MEH
4	DESPAIR
...	...
171509	JOY
171510	DESPAIR
171511	MEH
171512	DESPAIR
171513	JOY

[171514 rows x 4 columns]

ldata Value :	Q2: GENDER	Q3: AGE		candy \
0	Male	44	100 Grand Bar	
1	Male	44	Anonymous brown globs that come in black and o...	
2	Male	44	Any full-sized candy bar	
3	Male	44	Black Jacks	
4	Male	44	Bonkers (the candy)	
...	
171509	Female	66	Vicodin	
171510	Female	66	Whatchamacallit Bars	
171511	Female	66	White Bread	
171512	Female	66	Whole Wheat anything	
171513	Female	66	York Peppermint Patties	

candy_liking	
0	MEH
1	DESPAIR
2	JOY
3	MEH
4	DESPAIR
...	...
171509	JOY
171510	DESPAIR
171511	MEH
171512	DESPAIR
171513	JOY

[171514 rows x 4 columns]

df2 Value :	Q2: GENDER	Q3: AGE	Internal ID	Q1: GOING OUT?	Q4: COUNTRY \
0	Male	44	90272821	No	USA
1	Male	40	90272840	No	USA
2	Male	23	90272841	No	USA
3	Male	33	90272854	No	CANADA
4	Male	40	90272858	No	CANADA
...
1738	Female	26	90314022	No	USA
1739	Male	24	90314359	No	USA
1740	Female	33	90314580	No	USA
1741	Female	26	90314634	No	USA
1742	Female	66	90314802	No	USA

	Q5: STATE, PROVINCE, COUNTY, ETC	100 Grand Bar	\
0	NM	MEH	
1	or	MEH	
2	exton pa	JOY	
3	ontario	JOY	
4	Ontario	JOY	
...	
1738	Michigan	JOY	
1739	MD	JOY	
1740	New York	MEH	
1741	Tennessee	MEH	
1742	Pennsylvania	DESPAIR	

	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
...	...
1738	MEH
1739	DESPAIR
1740	DESPAIR
1741	DESPAIR
1742	DESPAIR

	Any full-sized candy bar	Black Jacks	...	\
0	JOY	MEH	...	
1	JOY	MEH	...	
2	JOY	DESPAIR	...	
3	JOY	DESPAIR	...	
4	JOY	MEH	...	
...	
1738	JOY	DESPAIR	...	
1739	MEH	DESPAIR	...	
1740	JOY	NaN	...	
1741	JOY	DESPAIR	...	
1742	JOY	DESPAIR	...	

	Vials of pure high fructose corn syrup, for main-lining into your vein \
0	DESPAIR
1	DESPAIR
2	MEH
3	JOY
4	MEH
...	...
1738	MEH
1739	MEH
1740	NaN
1741	MEH
1742	DESPAIR

	Vicodin	Whatchamacallit	Bars	White Bread	Whole Wheat	anything	\
0	DESPAIR		DESPAIR	DESPAIR		DESPAIR	
1	JOY		JOY	DESPAIR		DESPAIR	
2	JOY		JOY	DESPAIR		DESPAIR	
3	MEH		DESPAIR	DESPAIR		DESPAIR	
4	DESPAIR		MEH	DESPAIR		DESPAIR	
...	
1738	MEH		JOY	MEH		MEH	
1739	JOY		DESPAIR	MEH		DESPAIR	
1740	NaN		JOY	DESPAIR		MEH	
1741	JOY		MEH	DESPAIR		DESPAIR	
1742	JOY		DESPAIR	MEH		DESPAIR	

	York Peppermint Patties	Q10: DRESS	Q11: DAY	Q12: MEDIA [Science]	\
0	DESPAIR	White and gold	Sunday	1.0	
1	DESPAIR	White and gold	Sunday	1.0	
2	JOY	White and gold	Friday	1.0	
3	DESPAIR	Blue and black	Friday	1.0	
4	DESPAIR	Blue and black	Sunday	1.0	
...	
1738	JOY	White and gold	Friday	1.0	
1739	MEH	White and gold	Friday	NaN	

1740	JOY	Blue and black	Friday	1.0
1741	MEH	Blue and black	Friday	1.0
1742	JOY	White and gold	Sunday	NaN

Click Coordinates (x, y)

0	(84, 25)
1	(75, 23)
2	(70, 10)
3	(55, 5)
4	(76, 24)
...	...
1738	(68, 39)
1739	NaN
1740	(70, 26)
1741	(67, 35)
1742	(19, 26)

[1743 rows x 113 columns]

by_column Value : <pandas.core.groupby.generic.DataFrameGroupBy object at 0x000002B221E53640>

by_column.describe() Value :

COUNTRY Q4: COUNTRY	1723	56	USA	1490
LOCATION Q5: STATE, PROVINCE, COUNTY, ETC	1723	423	California	107
candy_liking	DESPAIR	JOY	MEH	
candy				
Any full-sized candy bar	16	1517	201	

df_2016 Value :

0	2016-10-24 05:09:23.033	Timestamp \
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	
...	...	
1254	2016-10-29 16:53:52.516	
1255	2016-10-30 06:53:54.735	
1256	2016-10-30 11:06:10.827	
1257	2016-10-30 16:07:26.539	
1258	2016-10-30 17:06:45.660	

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	Canada
1	45	usa
2	48	US
3	57	usa
4	42	USA
...
1254	52	USA
1255	33	united states
1256	NaN	NaN
1257	48	canada
1258	44	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]	[Black Jacks]	...	\
0	JOY	MEH	...	
1	JOY	JOY	...	
2	JOY	MEH	...	
3	JOY	MEH	...	
4	JOY	DESPAIR	...	
...	
1254	JOY	MEH	...	
1255	JOY	DESPAIR	...	
1256	JOY	NaN	...	
1257	JOY	DESPAIR	...	
1258	JOY	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
1258	3 or higher

Please estimate the degree(s) of separation you have from the following celebrities [JJ Abrams] \

0	2
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
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
1255	3 or higher
1256	NaN
1257	2
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 123 columns]
dataTypeObj Value : datetime64[ns]
Data collected is between date range 2016-10-24 05:09:23.033000 - 2016-10-30 17:06:45.660000
session.value_counts() Value: Morning      586
Afternoon    362
Night        218
Evening       93
Name: Timestamp, dtype: int64
```