Assignment 4

Question 1:

Upload Assignment4_data.csv into R. Please perform the following steps:

1. Explore the datasets using the ‘describe’ method in pandas. (10 points)

In [1]:

```
import os
import pandas as pd
os.chdir("E:/GoogleDrive/PSU/DAAN862/Course contents/Lesson 4")
data = pd.read_csv("Assignment4_data.csv")
```

In [2]:

```
data.columns
```

Out[2]:

```
Index(['one', 'two', 'three', 'four', 'five', 'variable'], dtype='object')
```

In [3]:

```
data.shape
```

Out[3]:

```
(200, 6)
```

In [4]:

```
data.head()
```

Out[4]:

	one	two	three	four	five	variable
0	-92.0	-76.0	-33.0	3.0	-13.0	B2
1	-21.0	76.0	38.0	-6.0	80.0	B1
2	-2.0	-47.0	-34.0	-86.0	-66.0	A1
3	-76.0	43.0	7.0	-40.0	-42.0	A1
4	44.0	37.0	-7.0	-14.0	30.0	A1

```
In [5]: data.describe()
```

Out[5]:

	one	two	three	four	five
count	195.000000	197.000000	199.000000	194.000000	196.000000
mean	-2.656410	2.208122	2.095477	-2.829897	-2.612245
std	67.489135	53.116759	101.864120	87.098996	84.158719
min	-363.000000	-100.000000	-100.000000	-576.000000	-821.000000
25%	-54.500000	-44.000000	-60.000000	-52.500000	-42.250000
50%	0.000000	-1.000000	-9.000000	-8.000000	-1.000000
75%	52.000000	45.000000	45.000000	44.000000	54.250000
max	97.000000	97.000000	832.000000	728.000000	99.000000

2.Determine how many missing values are in the data. (It is your choice how you handle the missing data.) (20 points)

```
In [6]: data.isnull().sum()
```

```
Out[6]: one         5
        two         3
        three       1
        four        6
        five        4
        variable    0
        dtype: int64
```

```
In [7]: data = data.fillna(data.mean())
```

```
In [8]: data.isnull().sum()
```

```
Out[8]: one         0
        two         0
        three       0
        four        0
        five        0
        variable    0
        dtype: int64
```

3.Explore the variable comlumn and Convert the "variable” column to dummy variables and join the dummies to the data. (20 points)

```
In [9]: data.variable.value_counts()
```

```
Out[9]: A1    59
        B2    50
        A2    46
        B1    45
        Name: variable, dtype: int64
```

```
In [10]: variable_dummy = pd.get_dummies(data.variable, prefix = "var")
```

```
In [11]: data_dummies = data.iloc[:, :5].join(variable_dummy)
         data_dummies.head()
```

Out[11]:

	one	two	three	four	five	var_A1	var_A2	var_B1	var_B2
0	-92.0	-76.0	-33.0	3.0	-13.0	0	0	0	1
1	-21.0	76.0	38.0	-6.0	80.0	0	0	1	0
2	-2.0	-47.0	-34.0	-86.0	-66.0	1	0	0	0
3	-76.0	43.0	7.0	-40.0	-42.0	1	0	0	0
4	44.0	37.0	-7.0	-14.0	30.0	1	0	0	0

Varify if convert to dummies correctly

```
In [12]: data_dummies.iloc[:, 5:].sum()
```

```
Out[12]: var_A1    59
         var_A2    46
         var_B1    45
         var_B2    50
         dtype: int64
```

4.Convert the "one” column into 3 bins. (20 points)

```
In [13]: data['one_bin'] = pd.cut(data.one, 3)
         data['one_bin'].value_counts()
```

```
Out[13]: (-56.333, 97.0]        152
         (-209.667, -56.333]     46
         (-363.46, -209.667]      2
         Name: one_bin, dtype: int64
```

Question 2:

Use the following speech by the Rev. Dr. Martin Luther King, Jr:

s = “I am happy to join with you today in what will go down in history as the greatest demonstration for freedom in the history of our nation. Five score years ago, a great American, in whose symbolic shadow we stand today, signed the Emancipation Proclamation. This momentous decree came as a great beacon light of hope to millions of Negro slaves who had been seared in the flames of withering injustice. It came as a joyous daybreak to end the long night of their captivity. But one hundred years later, the Negro still is not free. One hundred years later, the life of the Negro is still sadly crippled by the manacles of segregation and the chains of discrimination. One hundred years later, the Negro lives on a lonely island of poverty in the midst of a vast ocean of material prosperity. One hundred years later, the Negro is still languishing in the corners of American society and finds himself an exile in his own land. So we have come here today to dramatize a shameful condition."

In [14]:

```
s = '''I am happy to join with you today in what will go down in history as the greatest demonstration for freedom in the history of our nation. Five score years ago, a great American, in whose symbolic shadow we stand today, signed the Emancipation Proclamation. This momentous decree came as a great beacon light of hope to millions of Negro slaves who had been seared in the flames of withering injustice. It came as a joyous daybreak to end the long night of their captivity. But one hundred years later, the Negro still is not free. One hundred years later, the life of the Negro is still sadly crippled by the manacles of segregation and the chains of discrimination. One hundred years later, the Negro lives on a lonely island of poverty in the midst of a vast ocean of material prosperity. One hundred years later, the Negro is still languishing in the corners of American society and finds himself an exile in his own land. So we have come here today to dramatize a shameful condition.'''
```

1. Find out how many unique words in s. (10 points)

In [15]:

```
s_lower = s.lower()
```

In [16]:

```
s_list = s_lower.split(" ")
```

In [17]:

```python
for i in range(len(s_list)):
    if not s_list[i][-1].isalpha():
        print(s_list[i])
        s_list[i] = s_list[i][0:(len(s_list[i]) - 1)]
```

```
nation.
ago,
american,
today,
proclamation.
injustice.
captivity.
later,
free.
later,
discrimination.
later,
prosperity.
later,
land.
condition.
```

In [18]:

```python
len(set(s_list))
```

Out[18]:

```
107
```

In [19]:

```
s_list
```

```
Out[19]: ['i',
          'am',
          'happy',
          'to',
          'join',
          'with',
          'you',
          'today',
          'in',
          'what',
          'will',
          'go',
          'down',
          'in',
          'history',
          'as',
          'the',
          'greatest',
          'demonstration',
          'for',
          'freedom',
          'in',
          'the',
          'history',
          'of',
          'our',
          'nation',
          'five',
          'score',
          'years',
          'ago',
          'a',
          'great',
          'american',
          'in',
          'whose',
          'symbolic',
          'shadow',
          'we',
          'stand',
          'today',
          'signed',
          'the',
          'emancipation',
          'proclamation',
          'this',
          'momentous',
          'decree',
          'came',
          'as',
          'a',
          'great',
          'beacon',
          'light',
          'of',
          'hope',
          'to',
          'millions',
          'of',
          'negro',
          'slaves',
          'who',
          'had',
          'been',
          'seared',
          'in',
          'the',
          'flames',
          'of',
          'withering',
          'injustice',
          'it',
          'came',
          'as',
          'a',
          'joyous',
          'daybreak',
          'to',
          'end',
          'the',
          'long',
          'night',
          'of',
          'their',
          'captivity',
          'but',
          'one',
          'hundred',
          'years',
          'later',
          'the',
          'negro',
          'still',
          'is',
          'not',
          'free',
          'one',
          'hundred',
          'years',
          'later',
          'the',
          'life',
          'of',
          'the',
          'negro',
          'is',
          'still',
          'sadly',
          'crippled',
          'by',
          'the',
          'manacles',
          'of',
          'segregation',
           'and',
           'the',
           'chains',
           'of',
           'discrimination',
           'one',
           'hundred',
           'years',
           'later',
           'the',
           'negro',
           'lives',
           'on',
           'a',
           'lonely',
           'island',
           'of',
           'poverty',
           'in',
           'the',
           'midst',
           'of',
           'a',
           'vast',
           'ocean',
           'of',
           'material',
           'prosperity',
           'one',
           'hundred',
           'years',
           'later',
           'the',
           'negro',
           'is',
           'still',
           'languishing',
           'in',
           'the',
           'corners',
           'of',
           'american',
           'society',
           'and',
           'finds',
           'himself',
           'an',
           'exile',
           'in',
           'his',
           'own',
           'land',
           'so',
           'we',
           'have',
           'come',
           'here',
            'today',
            'to',
            'dramatize',
            'a',
            'shameful',
            'condition']
```

2. Which word appears the most? (10 points)

```
In [20]: s_dict = {i: s_list.count(i) for i in set(s_list) }
```

```
In [21]: max(s_dict, key = s_dict.get)

Out[21]: 'the'
```

```
In [22]: s_dict['the']

Out[22]: 14
```

Approach 2

```
In [23]: s_series = pd.Series(s_list)
```

```
In [24]: s_series.value_counts().head()

Out[24]: the      14
         of       12
         in        8
         a         6
         negro     5
         dtype: int64
```

3. How many words start with ‘t’. (10 points).

```
In [25]: count = 0
```

```
In [26]: for i in s_list:
             if i.startswith('t'):
                 count += 1
```

```
In [27]: count

Out[27]: 23
```