Dylan Francis

DAAN 682
08FEB2026

Homework_4:

Code:

```
#
========S2=============================================================
=
# DAAN: 682: Data Analytics Programming in Python
# Author: Dylan Francis
# Title: Homework_4: Data Cleaning, Processing, and Manipulating with Pandas
#
#
# Homework Questions:
#1.1 Explore the datasets. (10 points)
#1.2 Find and handle missing values in the data. (It is your choice how you handle the missing
data.) ( 20 points)
#1.3 Explore the variable column and convert the "variable” column to dummy variables and
join the dummies to the data. (20 points)
#1.4 Convert the "one” column into 3 bins. (20 points)import numpy as np

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. """
#2.1 Find out how many unique words in s. (10 points)
#2.2 Which word appears the most? (10 points)
#2.3 How many words start with ‘t’ (case-insensitive). (10 points).

import pandas as pd
from pandas import Series, DataFrame
import numpy as np
import os


path = r"C:\Users\dylan\OneDrive\Documents\GRAD_SCHOOL\DAAN_682\HOMEWORK_4"
os.chdir(path)
data_set = pd.read_csv("Assignment4_data.csv")
```

```
#print(data_set)

#1.1
print(f'The has:',data_set.shape[0], 'rows, and', data_set.shape[1], 'columns')
print(f'Here is the number of empty cells for each column:\n', data_set.isnull().sum(axis =0))
print('The toal number of empty cells is:',data_set.isnull().values.sum())

#1.2
data_fill = data_set.fillna(data_set.mean(numeric_only=True))

#1.3
dummies = pd.get_dummies(data_fill['variable']).astype(int)
data_with_dummies = data_set[['one','two','three','four', 'five']].join(dummies)
print(data_with_dummies)

#1.4
data_with_dummies["one_bin"] = pd.qcut(data_with_dummies["one"], q=3,
labels=['low','med','high'])
print(data_with_dummies)
data_with_dummies[data_with_dummies.columns[0]] =
data_with_dummies[data_with_dummies.columns[-1]]
data_with_dummies = data_with_dummies.drop(columns=data_with_dummies.columns[-1])
print(data_with_dummies)

#2.1
words_list = s.casefold().split()
unique_words = set(words_list)
number_of_unique_words = len(unique_words)
print(f"The number of unique words in the text is: {number_of_unique_words}")

#2.2
dict_top_words={}

for word in words_list:
   if word in dict_top_words:
      dict_top_words[word]+=1
   else:
      dict_top_words[word]=1

sorted_dict = dict(sorted(dict_top_words.items(), key=lambda item: item[1], reverse=True))
first_key = next(iter(sorted_dict), None)
first_value = next(iter(sorted_dict.values()), None)

print(f"The word thst appears the most is: '{first_key}' with a quantity of: {first_value}.")
```

```
# highest_key = max(dict_top_words, key=dict_top_words.get)
# print(f"The key with the highest value is: {highest_key}")

#2.3
letter_looking_for = "t"
counter = 0
s_edited = s.split()
for i in s_edited:
   if i.startswith(letter_looking_for) == True:
      counter +=1
print(f"The number of words that start with the letter {letter_looking_for} is: {counter}")
```

Output from the Terminal

```
%runfile
C:/Users/dylan/OneDrive/Documents/GRAD_SCHOOL/DAAN_682/HOMEWORK_4/Homework_
4.py --wdir
The has: 200 rows, and 6 columns
Here is the number of empty cells for each column:
 one        5
two         3
three       1
four        6
five        4
variable    0
dtype: int64
The toal number of empty cells is: 19
     one   two  three  four  five  A1  A2  B1  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
..    ...   ...    ...   ...   ... ..  ..  ..  ..
195  63.0   3.0  -30.0 -24.0 -59.0   1   0   0   0
196  97.0 -48.0  -61.0 -25.0 -21.0   0   0   1   0
197 -93.0 -75.0  -18.0 -67.0 -58.0   0   0   1   0
198  54.0 -66.0  -80.0  92.0  62.0   1   0   0   0
199  82.0  53.0  -77.0  79.0  97.0   0   0   0   1

[200 rows x 9 columns]
     one   two  three  four  five  A1  A2  B1  B2 one_bin
```

```
0   -92.0 -76.0  -33.0   3.0 -13.0   0   0   0   1    low
1   -21.0  76.0   38.0  -6.0  80.0   0   0   1   0    med
2    -2.0 -47.0  -34.0 -86.0 -66.0   1   0   0   0    med
3   -76.0  43.0    7.0 -40.0 -42.0   1   0   0   0    low
4    44.0  37.0   -7.0 -14.0  30.0   1   0   0   0   high
..    ...   ...    ...   ...   ... .. .. .. .. ..    ...
195  63.0   3.0  -30.0 -24.0 -59.0   1   0   0   0   high
196  97.0 -48.0  -61.0 -25.0 -21.0   0   0   1   0   high
197 -93.0 -75.0  -18.0 -67.0 -58.0   0   0   1   0    low
198  54.0 -66.0  -80.0  92.0  62.0   1   0   0   0   high
199  82.0  53.0  -77.0  79.0  97.0   0   0   0   1   high

[200 rows x 10 columns]
     one   two  three  four  five  A1  A2  B1  B2
0    low -76.0  -33.0   3.0 -13.0   0   0   0   1
1    med  76.0   38.0  -6.0  80.0   0   0   1   0
2    med -47.0  -34.0 -86.0 -66.0   1   0   0   0
3    low  43.0    7.0 -40.0 -42.0   1   0   0   0
4   high  37.0   -7.0 -14.0  30.0   1   0   0   0
..   ...   ...    ...   ...   ... .. .. .. .. ..
195 high   3.0  -30.0 -24.0 -59.0   1   0   0   0
196 high -48.0  -61.0 -25.0 -21.0   0   0   1   0
197  low -75.0  -18.0 -67.0 -58.0   0   0   1   0
198 high -66.0  -80.0  92.0  62.0   1   0   0   0
199 high  53.0  -77.0  79.0  97.0   0   0   0   1

[200 rows x 9 columns]
The number of unique words in the text is: 109
The word thst appears the most is: 'the' with a quantity of: 14.
The number of words that start with the letter t is: 22
```