IC 152

Computing & Data Science

Lab 10

Mohit Verma

B20215

Instructor- Dr.Padmanabhan Rajan

Questions & Answers

Q4) Revisiting pandas, File I/O, string processing and data visualization.

1. Import the File afi-top-100-quotes.csv" as a pandas data frame.

Ans 1 (a) The output is shown-

```
In [5]: df
Out[5]:
      #  ...  YEAR
0     1  ...  1939
1     2  ...  1972
2     3  ...  1954
3     4  ...  1939
4     5  ...  1942
..  ...  ...   ...
95   96  ...  1987
96   97  ...  1942
97   98  ...  1987
98   99  ...  1939
99  100  ...  1997

[100 rows x 4 columns]
```

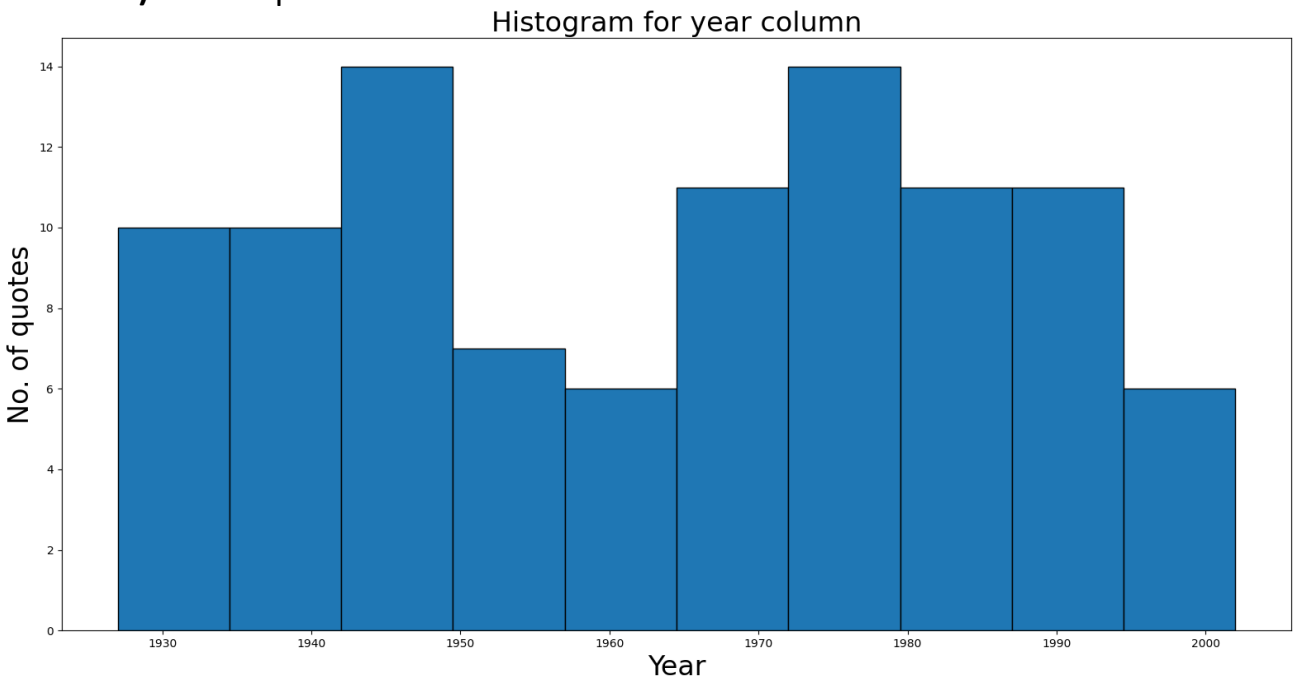
The code for output is shown below-

```
from collections import Counter
import pandas as pd
import math
import matplotlib.pyplot as plt

# Question 1
quotes=pd.read_csv("afi-top-100-quotes.csv")
df=pd.DataFrame(quotes)
```

Q2) Plot a histogram of the `YEAR' column by considering bin size = 10 years.

Ans 2) The output is shown below-



The code for output is shown below-

```
15
16  #Question 2
17  year=list(df["YEAR"])
18  plt.hist(year,bins=10,edgecolor='black')
19  plt.xlabel("Year",fontsize=24)
20  plt.ylabel("No. of quotes",fontsize=24)
21  plt.title("Histogram for year column",fontsize=24)
22
```

Q3) Print the movies that have appeared more than once in the descending order of their number of occurrences.

Ans 3) The output is shown below **–**

```
pandas.io.formats.html
                      MOVIE  No. of Occurence
0                 CASABLANCA                 6
1          GONE WITH THE WIND                 3
2               JERRY MAGUIRE                 2
3                THE GRADUATE                 2
4  A STREETCAR NAMED DESIRE                   2
5                SUNSET BLVD.                 2
6           THE WIZARD OF OZ                  2
```

The code for output is shown below-

```
22
23  # Question 3
24  count=df.value_counts("MOVIE").reset_index(name="No. of Occurence")
25  count=pd.DataFrame(count)
26  print(count[(count["No. of Occurence"]>1)].reset_index(drop=True))
27
```

Q4) Store the `MOVIE' column in a separate variable called “movies" and then change the datatype of its elements to string.

Ans 4) The output is shown below-

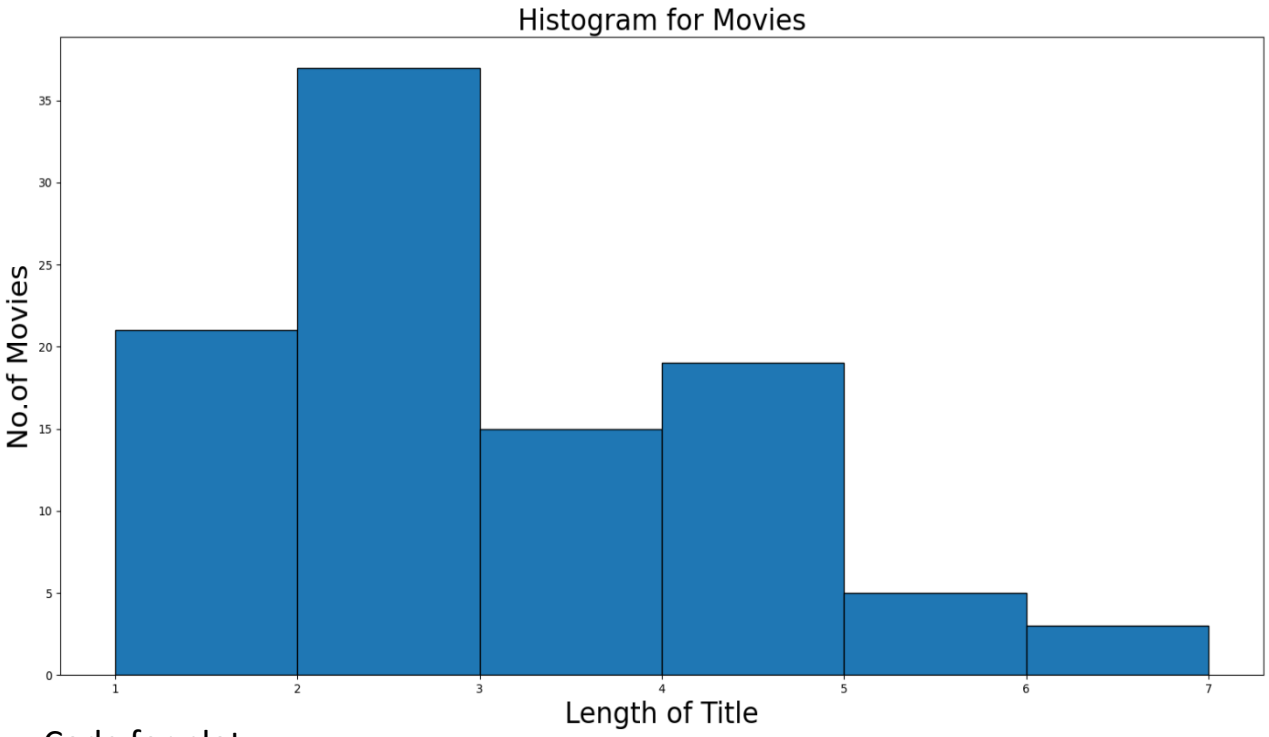
```
In [7]: movies
Out[7]:
0       GONE WITH THE WIND
1            THE GODFATHER
2        ON THE WATERFRONT
3          THE WIZARD OF OZ
4                CASABLANCA
               ...
95               MOONSTRUCK
96     YANKEE DOODLE DANDY
97           DIRTY DANCING
98       WIZARD OF OZ, THE
99                 TITANIC
Name: MOVIE, Length: 100, dtype: string
```

The code for output is –

```
27
28  # Question 4
29  movies=df["MOVIE"].astype("string")
30
```

Q5) Split the movie titles using string split() and the number of words for each of the titles and then plot a histogram of the number of the words by considering unit bin size.

Ans 5) Histogram for no. of words -



Code for plot-

```
30
31  #Question 5
32  number=[]
33  for titles in movies:
34      b=list(titles.split(" "))
35      number.append(len(b))
36
37  plt.figure()
38  n = math.ceil((max(number) -min(number))/1)
39  plt.hist(number,bins=n,edgecolor='black')
40  plt.xlabel("Length of Title",fontsize=24)
41  plt.ylabel("No.of Movies",fontsize=24)
42  plt.title("Histogram for Movies",fontsize=24)
43
```

6) From the split movie titles, find all the unique words and write them in a File along with their number of occurrences after sorting the words alphabetically. Please add the following header to the file-

Unique words Number of occurrences

Ans 6) The output is shown below-

```
verma@LAPTOP-L92N3PA1 ~
$ cd "C:\Users\verma\OneDrive\Desktop\LAB10"

verma@LAPTOP-L92N3PA1 /cygdrive/c/Users/verma/OneDrive/Desktop/LAB10
$ cat "Count_Title.txt"
Unique words       Number of occurrences
_____________________________________
13                          1
2001                        1
2                           1
42nd                        1
a                           5
about                       1
adventures                  1
afternoon                   1
airplane                    1
all                         2
american                    1
and                         2
animal                      2
annie                       1
apes                        1
apocalypse                  1
apollo                      1
auntie                      1
beyond                      1
blvd                        2
bonnie                      1
caddyshack                  1
caesar                      1
casablanca                  6
chinatown                   1
citizen                     1
clyde                       1
cool                        1
cowboy                      1
crackers                    1
dancing                     1
dandy                       1
day                         2
dead                        1
dearest                     1
desert                      1
desire                      2
dirty                       2
dog                         1
done                        1
doodle                      1
dr                          2
dracula                     1
dreams                      1
driver                      1
et                          1
eve                         1
extraterrestrial            1
falcon                      1
few                         1
```

```
et                      1
eve                     1
extraterrestrial        1
falcon                  1
few                     1
field                   1
forest                  1
forrest                 1
frankenstein            1
funny                   1
girl                    1
godfather               2
golden                  1
goldfinger              1
gone                    3
good                    1
graduate                2
grand                   1
green                   1
gump                    1
gun                     1
hall                    1
hand                    1
harry                   2
have                    2
heat                    2
him                     1
holmes                  1
hot                     1
hotel                   1
house                   1
ii                      1
impact                  1
in                      1
it                      1
jaws                    1
jazz                    1
jerry                   2
judgment                1
kane                    1
king                    1
knute                   1
kong                    1
lambs                   1
lampoons                1
league                  1
like                    1
little                  1
lord                    1
love                    1
luke                    1
madre                   1
maguire                 2
maltese                 1
mame                    1
man                     1
marathon                1
men                     1
```

```
man                     1
marathon                1
men                     1
met                     1
midnight                1
mommie                  1
moonstruck              1
named                   2
national                1
naughty                 1
network                 1
night                   1
nineties                1
no                      1
not                     1
now                     2
odyssey                 1
of                      13
on                      2
own                     1
oz                      3
part                    1
planet                  1
poets                   1
poltergeist             1
pond                    1
pride                   1
psycho                  1
rings                   1
rockne                  1
rocky                   1
sally                   1
scarface                1
sense                   1
shane                   1
she                     1
sherlock                1
shining                 1
sierra                  1
silence                 1
singer                  1
sixth                   1
society                 1
some                    1
sons                    1
soylent                 1
space                   1
star                    1
story                   1
strangelove             1
street                  2
streetcar               2
sudden                  1
sunset                  2
taxi                    1
terminator              2
the                     32
their                   1
```

```
the                     32
their                   1
titanic                 1
to                      1
top                     1
towers                  1
treasure                1
two                     1
voyager                 1
wall                    1
wars                    1
waterfront              1
when                    1
white                   1
wind                    3
with                    3
wizard                  3
wrong                   1
yankee                  1
yankees                 1

verma@LAPTOP-L92N3PA1 /cygdrive/c/Users/verma/OneDrive/Desktop/LAB10
$
```

The code for output is shown below-

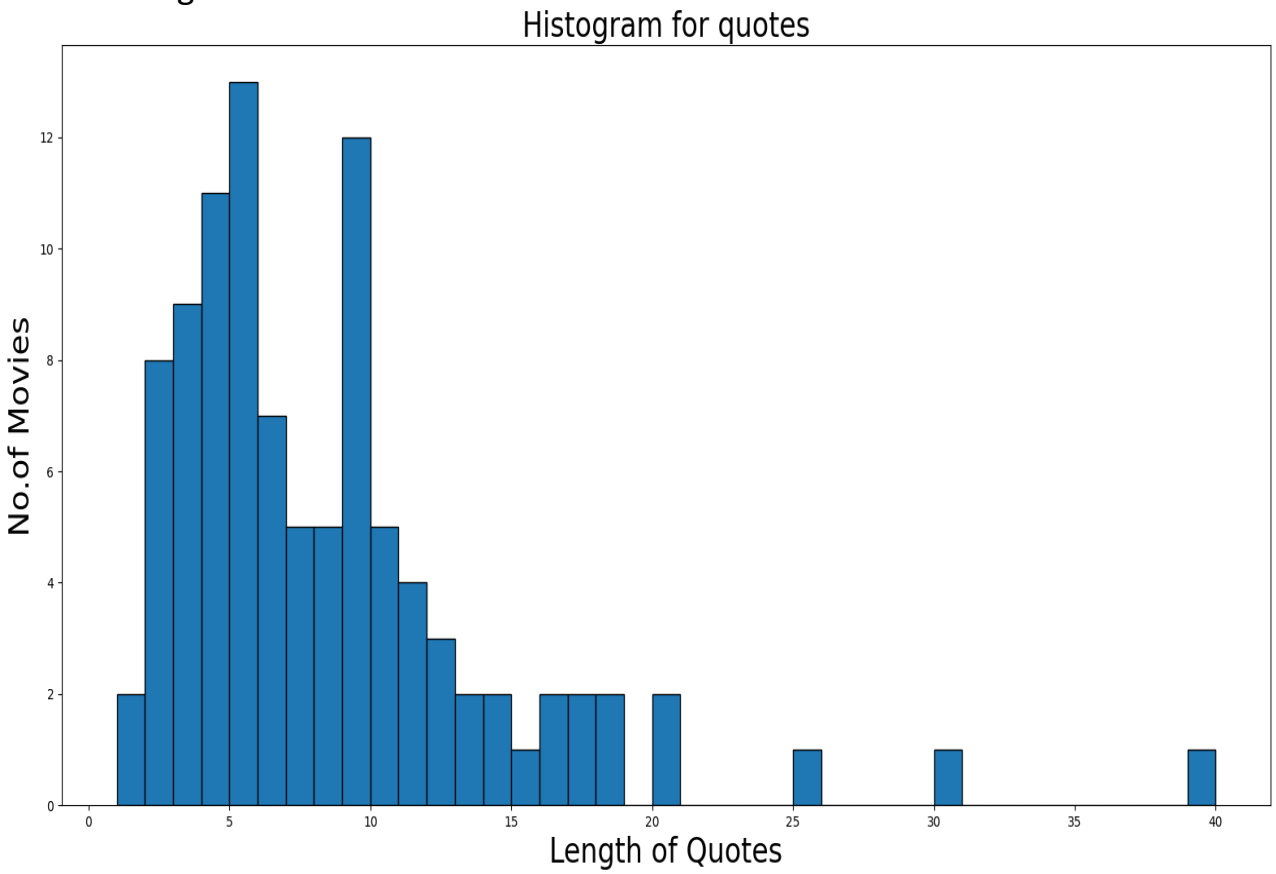
```
44
45  # Question 6
46  words=[]
47  for titles in movies:
48      b=list(titles.split())
49      for i in b:
50          words.append(i)
51  punctuations = '''!()-[]{};:'"\,<>./?@#$%^&*_~'''
52  words=sorted(words)
53  count1=[]
54  for i in words:
55    no_punct = ""
56    for char in i:
57     if char not in punctuations:
58         no_punct = no_punct + char
59    count1.append(no_punct.lower())
60  count2=Counter(count1)
61  l1=[]
62  l2=[]
63  for m in count2.items():
64          l1.append(m[0])
65          l2.append(m[1])
66
67  head="""Unique words       Number of occurrences\n_________
68  _____________________________\n"""
69  out1=open("Count_Title.txt","w")
70  out1.writelines(head)
71  for u,r in zip(l1,l2):
72      out1.writelines('{0:25}{1}\n'.format(u,str(r)))
73  out1.close()
74
```

Q7) Repeat 4 - 7 for the `QUOTE' column. Instructions for submitting Q-4: Create a pdf File by adding your codes and plots and upload to it to Moodle (the same way you submitted for Lab-7).

Ans) Quotes converted into string-

```
In [5]: quotes
Out[5]:
0                  Frankly, my dear, I don't give a damn.
1            I'm gonna make him an offer he can't refuse.
2       You don't understand! I coulda had class. I co...
3       Toto, I've a feeling we're not in Kansas anymore.
4                            Here's looking at you, kid.
                            ...
95                                        Snap out of it!
96      My mother thanks you. My father thanks you. My...
97                         Nobody puts Baby in a corner.
98      I'll get you, my pretty, and your little dog, ...
99                            I'm the king of the world!
Name: QUOTE, Length: 100, dtype: string
```

The histogram is shown below-



```
verma@LAPTOP-L92N3PA1 /cygdrive/c/Users/verma/OneDrive/Desktop/LAB10
$ cat "Count_Quotes.txt"
Unique words        Number of occurrences
_____________________________________________
a                           23
about                       1
adrian                      1
after                       1
again                       1
ahead                       1
aint                        2
airplanes                   1
alive                       2
all                         6
alone                       1
always                      3
am                          3
an                          2
and                         10
another                     2
any                         1
anymore                     2
ape                         1
are                         2
arent                       1
armor                       1
as                          4
ask                         2
at                          2
ate                         1
attica                      2
away                        1
baby                        2
back                        4
badges                      4
banks                       1
banquet                     1
baseball                    1
bay                         1
be                          7
beans                       1
beast                       1
beautiful                   1
beauty                      1
become                      1
been                        2
beginning                   1
behind                      1
best                        1
better                      1
big                         1
bigger                      1
blow                        1
boat                        1
bond                        2
box                         1
boys                        2
```

```
theyre                  1
think                   1
this                    4
tibbs                   1
tight                   1
time                    1
to                      20
today                   1
toga                    2
together                1
tomorrow                1
too                     1
top                     1
toto                    1
towns                   1
tried                   1
truth                   1
trying                  1
understand              1
up                      2
usual                   1
vista                   1
wait                    2
walking                 2
walks                   1
want                    1
war                     1
was                     2
wasnt                   1
watson                  1
we                      5
well                    4
were                    2
weve                    1
what                    6
which                   1
whistle                 1
whos                    1
why                     1
will                    1
win                     1
wire                    1
with                    3
witness                 1
word                    1
world                   3
ya                      1
yet                     1
yo                      1
you                     25
youngster               1
your                    7
youre                   7
yourself                1
youve                   3
```

Only first and last pages have been shown due to immense length of file.

The code for output is shown below-

```
#Question 6
quotes=df["QUOTE"].astype("string")

number1=[]
for g in quotes:
    b1=list(g.split(" "))
    number1.append(len(b1))
plt.figure()
n = math.ceil((max(number1) -min(number1))/1)
plt.hist(number1,bins=n,edgecolor='black')
plt.xlabel("Length of Quotes",fontsize=24)
plt.ylabel("No.of Movies",fontsize=24)
plt.title("Histogram for quotes",fontsize=24)

words1=[]
for x in quotes:
    b3=list(x.split())
    for i in b3:
        words1.append(i)
punctuations = '''!()-[]{};:'"\,<>./?@#$%^&*_~'''
count3=[]
for i in words1:
  no_punct = ""
  for char in i:
   if char not in punctuations:
       no_punct = no_punct + char
  count3.append(no_punct.lower())

count4=Counter(sorted(count3))
del count4['']
l3=[]
l4=[]
for m in count4.items():
        l3.append(m[0])
        l4.append(m[1])

head="""Unique words       Number of occurrences\n_____________________
_____________\n"""
out2=open("Count_Quotes.txt","w")
out2.writelines(head)
```

```
count4=Counter(sorted(count3))
del count4['']
l3=[]
l4=[]
for m in count4.items():
        l3.append(m[0])
        l4.append(m[1])

head="""Unique words       Number of occurrences\n_____________________
_____________\n"""
out2=open("Count_Quotes.txt","w")
out2.writelines(head)
for u,r in zip(l3,l4):
    out2.writelines('{0:25}{1}\n'.format(u,str(r)))
out2.close()
```

...