Jalen Jeudy

November 3, 2021

STAT 1129

Midterm

Github Link:

Question 1

In [1]:

unsorted **=** [2,4,6,8,4,5,2,1,9,0,4,6,7,4,3,2,1,9,10,3,7,9,6,0,1,3,5,6,7,8,9,10,2,3,6,8,9,10,6,7,4,3]

count **=** {}

unsorted**.**sort()

In [2]:

**for** number **in** unsorted:

**if** number **in** count:

count[number] **+=** 1

**else**:

count[number]**=**1

print("Output")

print()

**for** key, value **in** count**.**items():

print(key, ':', value)

print()

print('Explanation:')

print()

**for** key, value **in** count**.**items():

print('Here', key, 'appears', value, 'times')

Output

0 : 2

1 : 3

2 : 4

3 : 5

4 : 5

5 : 2

6 : 6

7 : 4

8 : 3

9 : 5

10 : 3

Explanation:

Here 0 appears 2 times

Here 1 appears 3 times

Here 2 appears 4 times

Here 3 appears 5 times

Here 4 appears 5 times

Here 5 appears 2 times

Here 6 appears 6 times

Here 7 appears 4 times

Here 8 appears 3 times

Here 9 appears 5 times

Here 10 appears 3 times

In [3]:

**import** matplotlib

**import** matplotlib.pyplot **as** plt

**import** numpy **as** np

In [4]:

x = count.keys()

y = count.values()

plt.xlabel('Item')

plt.ylabel('Prevalence')

plt.bar(x,y)

plt.show()

Chart, bar chart

Description automatically generated

In [5]:

**import** json

**with** open("Item\_Prevalence.json", "w") **as** out:

json.dump(count, out)

rd = open("Item\_Prevalence.json", "r")

rd.read()

Out[5]:

'{"0": 2, "1": 3, "2": 4, "3": 5, "4": 5, "5": 2, "6": 6, "7": 4, "8": 3, "9": 5, "10": 3}'

Question 2

In [6]:

**import** pandas **as** pd

In [7]:

df **=** pd**.**read\_json('data.json')

In [8]:

df**.**head(1)

Out[8]:

Graphical user interface, text, application

Description automatically generated

In [9]:

*#change the name of the timestamp column*

df**.**rename(columns**=**{'timestamp': 'date'}, inplace**=True**)

In [10]:

*#delete unimportant columns*

df **=** df**.**drop(['attachments', 'title'], axis**=**1)

In [11]:

*#format the datatime*

pd**.**to\_datetime(df['date'])

Out[11]:

0 2019-01-06 19:24:34

1 2019-02-25 03:57:45

2 2020-01-30 14:29:00

3 2020-04-16 18:36:23

4 2020-10-25 00:59:47

5 2021-04-16 19:03:12

Name: date, dtype: datetime64[ns]

In [12]:

print(df**.**shape)

(6, 2)

In [13]:

df **=** df**.**set\_index('date')

post\_counts **=** df['data']**.**resample('MS')**.**size()

post\_counts

Out[13]:

date

2019-01-01 1

2019-02-01 1

2019-03-01 0

2019-04-01 0

2019-05-01 0

2019-06-01 0

2019-07-01 0

2019-08-01 0

2019-09-01 0

2019-10-01 0

2019-11-01 0

2019-12-01 0

2020-01-01 1

2020-02-01 0

2020-03-01 0

2020-04-01 1

2020-05-01 0

2020-06-01 0

2020-07-01 0

2020-08-01 0

2020-09-01 0

2020-10-01 1

2020-11-01 0

2020-12-01 0

2021-01-01 0

2021-02-01 0

2021-03-01 0

2021-04-01 1

Freq: MS, Name: data, dtype: int64

In [14]:

**import** matplotlib

**import** matplotlib.pyplot **as** plt

**from** matplotlib.pyplot **import** figure

**import** numpy **as** np

In [15]:

x **=** post\_counts**.**index

y **=** post\_counts

figure(figsize**=**(20,10))

plt**.**xlabel('Date')

plt**.**ylabel('Number of Posts')

plt**.**bar(x, y, width**=**7)

plt**.**show()

Chart, histogram

Description automatically generated