**WRITTEN TEST for DATA QUALITY ANALYST @ TURBOLAB TECHNOLOGIES**

Swaroop N C

E-mail : ncswaroop1997@gmail.com

Mobile : +91 99 1616 0739

LinkedIn : https://www.linkedin.com/in/swaroop-n-c-a09499131/

Date : 09-05-2022

Instructions: 3 hours to complete

- All three questions should be attempted.

```
1 import numpy as np
2 import pandas as pd
```

# QUESTION 1

Part 01. **Create a Pandas series with 100 random dates as it falls between 01-01-20 to 01-01-21**

(hint: use **Pandas date_range** function).

Like:- _ 2020-01-06 2020-06-11 2020-02-18

Part 02. **Dedupe it and calculate the number of duplicates and convert it to percentage.**

Then by **using regex, filter values where either the month is 02,05,09 OR the date is 01,04,07** - the apply function should not be used.

Finally, **calculate the percentage of values filtered for the month condition and the date condition**.

# SOLUTION TO Q 1

```
1 import random
```

```
1 start_date='1/1/2020'
2 end_date='1/01/2021'
3 n=100
```

```
1 All_dates=pd.date_range(start=start_date, end=end_date)
```

```
All_dates
```

```
DatetimeIndex(['2020-01-01', '2020-01-02', '2020-01-03', '2020-01-04',
               '2020-01-05', '2020-01-06', '2020-01-07', '2020-01-08',
               '2020-01-09', '2020-01-10',
               ...
               '2020-12-23', '2020-12-24', '2020-12-25', '2020-12-26',
               '2020-12-27', '2020-12-28', '2020-12-29', '2020-12-30',
               '2020-12-31', '2021-01-01'],
              dtype='datetime64[ns]', length=367, freq='D')
```

```
df= pd.DataFrame(random.choices(x,k=n))
df
```

| | 0 |
|---|---|
| **0** | 2020-01-23 |
| **1** | 2020-08-25 |
| **2** | 2020-05-12 |
| **3** | 2020-04-18 |
| **4** | 2020-07-24 |
| **...** | ... |
| **95** | 2020-03-17 |
| **96** | 2020-09-18 |
| **97** | 2020-10-04 |
| **98** | 2020-05-01 |
| **99** | 2020-09-06 |

100 rows × 1 columns

```
df.columns=['Date']
```

```
df
```

| | Date |
|---|---|
| **0** | 2020-01-23 |
| **1** | 2020-08-25 |
| **2** | 2020-05-12 |
| **3** | 2020-04-18 |

Dedupe it and calculate the number of duplicates and convert it to percentage.

... ...

```
df.nunique()
```

```
Date    88
dtype: int64
```

```
Num_Of_Duplicates = n - df.nunique()
Num_Of_Duplicates
```

```
Date    12
dtype: int64
```

```
Dup_Percent = ((Num_Of_Duplicates)/n)*100
Dup_Percent
```

```
Date    12.0
dtype: float64
```

Then by using regex, filter values where either the month is 02,05,09 OR the date is 01,04,07 - the apply function should not be used.

Finally, calculate the percentage of values filtered for the month condition and the date condition.

```
#COULD NOT ANSWER THIS PART
```

# QUESTION 2

This question has two parts,

a) Sentence Validator and

b) Name Reducer

a) Create a sentence validator function. The validator should return input if it is valid otherwise return False

Validation Criteria:

1. Start letter must be an uppercase letter and it should follow either a lowercase letter or a single whitespace.
2. All letters in the sentence except the start letter must be in lowercase.
3. The last character (aka terminal character) of the sentence must be any of the following:

. (dot) ? (question mark) ! (exclamation mark)

4. Words must be separated with a single whitespace.

If there is a hyphen between any two words then there should be one whitespace before and after that hyphen.

eg: Lab - 1 is valid, but Lab- 7 and Lab -7 are invalid

b) Write a reducer function to clean the sentences. Reducer takes output from the validator function as input and performs the following cleaning steps,

○ Removes terminal characters (see above validation criteria for list of allowed terminal characters)

○ Removes all duplicated word groups (see below examples) but keep its first occurence

○ Removes all leading and trailing whitespaces and hyphens

After completing the functions for part a & b.

Create a function check_and_clean which takes a sentence as input and validates (using validator function) and returns reduced string (output from reducer function) when the sentence is valid, "" otherwise.

Note: You are not allowed to use regex for this question.

Example 1: Input: Melo diagnostics melo Labs

Sentence is invalid, failed validation criteria 2 & 3

Output: invalid

Example 2:

Input: Melo diagnostics - southpark east 29th street - southpark east 29th street - free drug testing not offered. Sentence is valid

Word groups: Melo diagnostics southpark east 29th street free drug testing not offered

Output: Melo diagnostics - southpark east 29th street - free drug testing not offered

Example 3:

Input: Simple labs - covid test available - west side hospital - covid test available!

Sentence is valid

Word groups: Simple labs covid test available west side hospital

Output: Simple labs - covid test available - west side hospital

# SOLUTION TO Q 2

sentence validator

Validation Criteria:

1 **Start letter must be an uppercase letter** and **it should follow either a lowercase letter** or **a single whitespace**.

2 **All letters in the sentence except the start letter must be in lowercase.**

3 **The last character (aka terminal character) of the sentence must be any of the following: . (dot) ? (question mark) ! (exclamation mark)**

**Words must be separated with a single whitespace.** If there is a hyphen between any two words then there should be one whitespace before and after that hyphen.

eg: Lab - 1 is valid, but Lab- 7 and Lab -7 are invalid

```
def SentenceValidator(string):
  length = len(string)
  #Start letter must be an uppercase letter
  if (string[0].islower()):
    return False
  #last character of sentence must be  . (dot) ? (question mark) ! (exclamation mark)
  elif not(string[length-1] == '.' or string[length-1] == '?' or string[length-1] == '!
    return False
  #All letters in the sentence except the start letter must be in lowercase
  else:
    for ele in string[1:]:
      if ele.isupper():
        return False
  #Words must be separated with a single whitespace.
  #check for more than 2 consecutive whitespace
  for i in range(len(string)) :
    if (string[i]==' ' and string[i+1]==' ') :
      return False
  #If there is a hyphen between any two words then
  #there should be one whitespace before and after that hyphen.
  for i in range(len(string)) :
    if (string[i]=='-') :
      if not((string[i-1]==' ')and (string[i+1]==' ')):
        return False
  return string
```

b) Write a reducer function to clean the sentences. Reducer takes output from the validator function as input and performs the following cleaning steps,

○ **Removes terminal characters (see above validation criteria for list of allowed terminal characters)**

○ **Removes all duplicated word groups (see below examples) but keep its first occurence**

```
def reducer(ValidOp) :
  #Removes terminal characters
  ValidOp=ValidOp[:-1]
  #Removes all leading and trailing whitespaces
  ValidOp=ValidOp.strip()
  #Removes all leading and trailing HYPHENS
  ValidOp=ValidOp.strip("-")
  #Removes all duplicated word groups  but keep its first occurence
  l = ValidOp.split()
  k = []
  for i in l:
    if (ValidOp.count(i)>=1 and (i not in k) or i =='-'):
      k.append(i)
  ValidOp=' '.join(k)
  return ValidOp
```

check_and_clean

Create a function check_and_clean which takes a sentence as input and validates (using validator function) and returns reduced string (output from reducer function) when the sentence is valid, "" otherwise.

```
def check_and_clean() :
  stringcheck=str(input("Input String : "))
  if (SentenceValidator(stringcheck) != False ) :
    print("******** OUTPUT **********************")
    return (reducer(SentenceValidator(stringcheck)))
  else :
    print("******** OUTPUT **********************")
    return ("<invalid>")
```

```
#driverprogram
#test case 1 Melo diagnostics melo Labs
check_and_clean()
```

```
Input String : Melo diagnostics melo Labs
******** OUTPUT **********************
'<invalid>'
```

```
#driverprogram
#test case 3  : Simple labs - covid test available - west side hospital - covid test av
check_and_clean()
```

```
Input String : Simple labs - covid test available - west side hospital - covid test a
******** OUTPUT **********************
'Simple labs - covid test available - west side hospital -'
```

# Question 3:

A sample data of posts of random users are given in this link: click here to download Post_id, date of post and post caption details are available in the sample dataset.

Create a function that extracts posts older than 13/11/2021 and finds the 3 most frequently used special characters out of it. The function should return the 3 most frequently used special characters and the number of times they occurred in the filtered data.

Example

| post_id | date | caption |
|---|---|---|
| post #1 | 11/11/2021 | @bla bl@ bla! 23🔥 |
| post #2 | 15/11/2021 | Foo b@r foOB!a🙌 |
| post #3 | 12/11/2021 | 🔥 aerrt!! Qwe r rr$ |
| post #4 | 13/11/2021 | @momo bati$t@🔥 |

## Output

[ (@, 4), (!, 3), (🔥, 2) ]

Explanation The post #2 is eliminated since it is older than the date 13/11/2021. In the remaining 3 rows, the special character “@” occurred the most i.e, 4 times in the posts #1 and #4. The second most frequent special character is “!” which occurred 3 times and then “�” occurred twice. Since only the top 3 most frequent ones are required, the remaining special characters “�” and “$” are ignored.

```
#load dataset
df = pd.read_csv("Captions.csv")
df
```

| | post_id | date | caption |
|---|---|---|---|
| **0** | post #1 | 15/11/2021 | NaN |
| **1** | post #2 | 14/11/2021 | Skippers at the ground for a photo call but ca... |
| **2** | post #3 | 18/11/2021 | Kia ora everyone, We tried but it wasn't to be... |
| **3** | post #4 | 11/11/2021 | Teihorangi Walden 🔒 2022 |
| **4** | post #5 | 13/11/2021 | NaN |
| **...** | ... | ... | ... |
| **95** | post #96 | 17/11/2021 | Pink maomao, blue maomao, granddaddy hāpuku (N... |
| **96** | post #97 | 12/11/2021 | Hiking in Queenstown? Yes, please! 🥾 Now th... |
| **97** | post #98 | 11/11/2021 | Opportunity to engage: Gender and Sex Diverse ... |
| **98** | post #99 | 11/11/2021 | 7 years signed, sealed and delivered ❤️💙 |

extracting posts older than 13/11/2021

100 rows × 3 columns

```
df=df[df['date']<'13/11/2021']
df
```

| | post_id | date | caption |
|---|---|---|---|
| 3 | post #4 | 11/11/2021 | Teihorangi Walden 🔒 2022 |
| 5 | post #6 | 12/11/2021 | Lineup 🔥 |
| 6 | post #7 | 11/11/2021 | “Taking some Mospiration from the hedges” @sam... |
| 8 | post #9 | 12/11/2021 | A reduction in wild animals, vegetation recove... |
| 10 | post #11 | 12/11/2021 | Which witch will you watch first this weekend?... |
| 13 | post #14 | 11/11/2021 | Just returned from the best kingfish fishing w... |
| 14 | post #15 | 11/11/2021 | ICYMI \| Our short film 'The Mountain Storm' wa... |
| 16 | post #17 | 12/11/2021 | The PENN BATTLE III DX Reel has landed! Exclus... |
| 18 | post #19 | 11/11/2021 | Excited to be back @nrl_weststigers as an offi... |
| 20 | post #21 | 12/11/2021 | 💥🏏🙌🏆 It all comes down to this! A trans-Tasman... |
| 22 | post #23 | 12/11/2021 | Working on our fitness influencer poses. Swipe... |
| 23 | post #24 | 11/11/2021 | 🐂 vs 🐪 🙌 Team Milestones: 💛 Stephen Perofeta w... |
| 28 | post #29 | 12/11/2021 | NZKS is open for shopping this weekend! If you... |
| 31 | post #32 | 12/11/2021 | Twins Otukinekina and Valingi Kepu will join o... |
| 38 | post #39 | 11/11/2021 | Need warmth and performance? All the natural b... |
| 43 | post #44 | 12/11/2021 | 🌊🌊🌊🌊 |
| 46 | post #47 | 12/11/2021 | Grab a pack while the stock lasts! Get in quic... |
| 47 | post #48 | 11/11/2021 | Congratulations to the @volcanichillswinery Te... |
| 51 | post #52 | 11/11/2021 | 2 years ago today we launched Live Ocean. Afte... |
| 58 | post #59 | 12/11/2021 | Grateful 🖤🤍 |

```
def special_char_list(string):
    special_char = 0
    spec_char_list = []
    for i in range(len(string)):
        if(string[i].isalpha()):
          continue
        elif(string[i].isdigit()) :
          continue
        else :
          if (string[i] not in spec_char_list) :
            spec_char_list.append(string[i])
    return(spec_char_list)
```

| | | | |
|---|---|---|---|
| 76 | post #77 | 11/11/2021 | Summer holidays on your mind? 🐱 We've got you ... |

```
caption_text = ''
for row in df.caption :
  caption_text += row
```

| | | | |
|---|---|---|---|
| 83 | post #84 | 12/11/2021 | 🏒LIVE HOCKEY TOMORROW 🏒- Link in our bio for l... |

```
caption_text
```

ed deer, 11 chamois and one feral cat found near Burn Creek. Find out more: http s://bit.ly/3D0qPLK (link in bio) 📷: Jack Mace.Which witch will you watch first t his weekend? 🐭 The Witches (1990) & The Witches (2020) are both on Neon now!\u20 63 \u2063 Let us know Grand High Witch reigns supreme in our Instagram story poll s!\u2063 \u2063 \u2063 \u2063 \u2063 #TheWitches #RoaldDahl #AnjelicaHuston #AnneH athaway #RowanAtkinson #OliviaSpencer #ChrisRock #KidsMovies #NewMovie #Neon #Neon NZJust returned from the best kingfish fishing we have ever seen, huge schools of raging kingfish surface feeding like nothing we have ever seen. Totally awesomeICY MI | Our short film \'The Mountain Storm\' was released last night. It\'s a story of hunters banding together to save our Tahr. It\'s been a long time in the making so we hope you enjoy it. Link to the film in our bio. . . . #HuntLonger #HuntersEl ementThe PENN BATTLE III DX Reel has landed! Exclusive to your local Hunting & Fis hing New Zealand store this range of reels from 2500 through to 8000 size builds o ff the reputation of the legendary PENN BATTLE III Reels by adding some fantastic upgrades » Extra bearing under the spool for additional support » Upgraded wave sp ring design for more minute drag adjustment and wider range of useable drag » Impr oved gear durability with brass main and pinion gear on all sizes » EVA ball handl e knob for improved cranking comfort » DX range exclusive silver and black cosmeti c With the team at @pennfishing_anz we are celebrating the arrival by giving away two of these epic reels! All you need to do is tag your fishing mate in the commen ts and make sure you are following both Hunting & Fishing NZ & @pennfishing_anz to qualify! We will make the draw on Friday 26th November. If you miss out – you can drop into your local H&F store to check out the full range – Reels are priced from just $229.99, and we have a great assortment of combos across Softbaiting, Jiggin g, Topwater and Surf priced from $299.99Excited to be back @nrl_weststigers as an official ambassador for 2022 . Look foward to working with and coaching the junior s and helping develop some of our next stars . Also excited to be on debut attacki ng the commercial world also. #newbeginnings #welcomebackhome💥🎷🙌🏆 It all com es down to this! A trans-Tasman showdown to decide the ICC T20 World Cup! Who are you backing to win the Final? Don\'t miss out: @blackcapsnz v Australia - Monday 2.30am LIVE on Sky Sport 3* 📸 @photosportnzWorking on our fitness influencer pos es. Swipe for today’s workout #fitspo #gainztrain🐂 vs 🦌 🙌 Team Milestones: 💛 Stephen Perofeta will play his 50th game for Taranaki 🖤 Jacob Ratumaitavuki – Kn eepkens will play his 20th game for Taranaki (Blazer) 🚑 Unavailable due to injur y: · Fin Hoeata 😟 Season ending Injuries: · Jesse Parete · Mitch Brown · Reuben O’Neill · Johnny Faletagoa’i – MalaseNZKS is open for shopping this weekend! If yo u are popping in please bring your mask. Thank you! Happy paddling. 💪Twins Otuki nekina and Valingi Kepu will join our NRL squad in 2023 after securing extended co ntracts.Need warmth and performance? All the natural benefits of merino in a base layer you can take anywhere - whatever it looks like outside. Choose the original natural base layer. #stillwearingplastic #movetonatural🌊🌊🌊🌊Grab a pack while the stock lasts! Get in quick to your local Countdown, SuperValue, or Fresh Choice store near you.Congratulations to the @volcanichillswinery Team of the Year finali sts for the 2021 Forsyth Barr Bay of Plenty Sports Awards. 🔶 Mount Maunganui Rug by Club Womens’ 🔷 Upper Central Rugby League Stallions 🔶 Tauranga City AFC Wome n’s First Team 🔷 Bay of Plenty Senior Mens’ Volleyball @mountmaunganuisportsclub @uppercentralzone @tcafc_nz @volleyballbayofplenty 📸 @photographyjuliemaree (vol leyball) Help us cheer them on when we broadcast the awards live from 5:30pm Frida y, 19 November. #sportbop #getactivebop #sport #sports #sportsawards #whakatane #o potiki #tauranga #rotorua #bayofplentynz #womensrugby #womensrugbyleague #womensfo otball #mensvolleyball #rugbynz #rugbyleaguenz #footballnz #volleyballnz2 years ag o today we launched Live Ocean. After @theoceanrace I felt the ocean wasn’t infini te and wanted to do my part, so teamed up with my mate @blairtuke and what a journ ey it has been. @itsliveocean #itsliveoceanGrateful 🖤\U0001f90dGood luck and goo d racing to @lena_gasson who’s in the water for @lacurrent at the @iswimleague pla yoffs starting this weekend in the Netherlands 🙌 The kiwi power duo is completed by @coastelite coach @michaeljweston who’s also assisting the LA Current 💪 Best of luck team, go give ‘em some kiwi 🏆The backyard\'s not so bad 👊 Happy weekend adventurers – tag us in your pics and what you get up to, rain or shine. 📸 @dger l #WeatherAnything #GreatDayForItTom Saunders is your new Laser world champion 🎉

```
#nzlsailing #nzlsailingteam #sailing #yachting #laser #laserworldsEnjoyable few we
eks with @canterbury.cricket Excited to be in India ready for some Test cricket!
IN V NZ 📸 @photosportnzRipper first chamois for @fannymackdaddy Old as the hills,
but how old is it? #agematters #chamois #firstlite #gofartherstaylongerLET US OPEN
If you have been wondering when Auckland swimming pools will reopen, you're not th
e only ones. New Zealand swimmers and triathletes who are striving to represent ou
r country on the world stage are fearing the ongoing lockdown is jeopardising thei
r chances at the upcoming Birmingham Commonwealth Games. Here SwimTastic's CEO @Ma
rkBone comments on Newshub https://lnkd.in/eauQnhWE There was a smidgen of hope a
week ago that pools like SwimTastic would be open when Auckland and Waikato moved
to Alert level 3, step 2, but this was dashed by the Government at the 11th hour.
What is strange is that the Ministry of Health's public health director Dr Carolin
e McElnay said "we do look at these situations very specifically and then provide
```

```
1 spec_char_list=special_char_list(caption_text)
```

```
1 spec_char_list
```

```
':',
'/',
'📸',
'?',
'🐭',
'&',
'\u2063',
'|',
"'",
'»',
'$',
'💥',
'🏏',
'👐',
'🏆',
'*',
'📸',
''',
'🐂',
'🦌',
'💛',
'🚑',
'.',
'😐',
'🤽',
'🌊',
'🔶',
'🔷',
'\U0001f90d',
'💪',
'ς',
'🤜',
' ',
'_',
'🎺',
'I',
'N',
'Z',
'👏',
'🏈',
'😃',
'❤',
```

```
'+',
'🌞',
'ɛ',
'🙋',
' ',
'\u200d',
'♂',
'🏑',
'💚',
'🎥',
' ',
'🥾',
'\u2060',
'👏',
'⬆',
'💙']
```

```
i=0
count_num=0
count=[]
for sp_char in spec_char_list :
  for char in caption_text :
    if(sp_char == char) :
      count_num+=1
  count.append(count_num)
  count_num=0
```

```
len(count)
```

```
79
```

```
len(spec_char_list)
```

```
79
```

```
zip_ob = list(zip(spec_char_list,count))
zip_ob
```

```
(':', 10),
('/', 7),
('📷', 1),
('?', 7),
('🐭', 1),
('&', 6),
('\u2063', 7),
('|', 2),
("'", 10),
('»', 5),
('$', 2),
('💥', 1),
('🏏', 1),
('🙌', 3),
('🏆', 2),
('*', 1),
('📸', 5),
('’', 23),
```

```
('🐂', 1),
('🦌', 1),
('💛', 2),
('🚑', 1),
('•', 5),
('😐', 1),
('🥶', 1),
('🌊', 5),
('🔶', 2),
('🔷', 2),
('\U0001f90d', 1),
('💪', 1),
('ʿ', 1),
('🤜', 2),
(' ', 1),
('–', 1),
('🍻', 1),
('I', 2),
('N', 2),
('Z', 1),
('👏', 2),
('🩹', 1),
('😃', 1),
('♥', 4),
('+', 1),
('🌞', 1),
('E', 1),
('🙋', 1),
(' ', 1),
('\u200d', 1),
('♂', 1),
('🏑', 2),
('💚', 1),
('📽', 1),
(' ', 1),
('🥾', 1),
('\u2060', 4),
('👏', 1),

('⬆', 1),
```

```
1 sorted(zip_ob)
```

```
('·', 5),
('»', 5),
('\u200d', 1),
('-', 5),
('–', 1),
('ʿ', 1),
('’', 23),
('“', 2),
('”', 2),
('\u2060', 4),
('\u2063', 7),
('♂', 1),
('♥', 4),
('➡', 1),
('⬆', 1),
('', 7),
('E', 1),
```

```
('I', 2),
('N', 2),
('Z', 1),
('🌊', 5),
('🌞', 1),
('🌿', 1),
('🍓', 1),
('🎥', 1),
('🏆', 2),
('🏏', 1),
('🏑', 2),
(' ', 1),
(' ', 1),
(' ', 1),
('🐂', 1),
('🐍', 1),
('🐭', 1),
('🐳', 1),
('👏', 2),
('👏', 1),
('💙', 1),
('💚', 1),
('💛', 2),
('💥', 1),
('💪', 1),
('📷', 1),
('📸', 5),
('🔒', 1),
('🔗', 1),
('🔥', 1),
('🔶', 2),
('🔷', 2),
('🖤', 7),
('😀', 1),
('🙋', 1),
('🙌', 3),
('🚑', 1),
('\U0001f90d', 1),
('😕', 1),
('🤜', 2),
('🦵', 1),
('🦾', 1)]
```

```
1 res = list(reversed(sorted(zip_ob, key = lambda x: x[1])))
2 res
```

```
('📸', 5),
('»', 5),
('–', 5),
('\u2060', 4),
('🖤', 4),
('🙌', 3),
('🏑', 2),
('👏', 2),
('N', 2),
('I', 2),
('🤜', 2),
('🔷', 2),
('🔶', 2),
('💛', 2),
```

```
('⏳', 2),
('$', 2),
('|', 2),
('"', 2),
('”', 2),
('“', 2),
('💙', 1),
('⬆', 1),
('👏', 1),
('🥾', 1),
(' ', 1),
('📽', 1),
('💚', 1),
('♂', 1),
('\u200d', 1),
(' ', 1),
('🙋', 1),
('ᴇ', 1),
('🌞', 1),
('+', 1),
('😃', 1),
('📎', 1),
('z', 1),
('🎏', 1),
('—', 1),
(' ', 1),
('‘', 1),
('💪', 1),
('\U0001f90d', 1),
('🥶', 1),
('😟', 1),
('🚑', 1),
('🦌', 1),
('🐂', 1),
('*', 1),
('🏏', 1),
('💥', 1),
('🐭', 1),
('📷', 1),
('🌿', 1),
('🐛', 1),
('➡', 1),
('🔥', 1),
('🔒', 1)]
```

TOp 3 special characters

```
1 res[:3] #including space
```

```
[(' ', 1801), ('#', 68), ('.', 67)]
```

```
1 res[1:4] #exluding space space
```

```
[('#', 68), ('.', 67), (',', 49)]
```