Take at home assignment for Holidaycheck_ag

Language used for this Project: Python3

In [1]:

```
#Basic imports of libraries which we will be using throughout this project


import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import pprint
import json
import requests
from os import makedirs
from os.path import join, exists
from datetime import date, timedelta
import time
from datetime import datetime
from matplotlib.dates import DateFormatter
import matplotlib.dates as mdates
pd.options.display.max_colwidth = 500
pd.options.display.max_rows = 500
%matplotlib inline
```

In [2]:

```
#A function which will be used to request the data from the guardian API
#parameters:
#apikey as obtained from the API, (pre-defined as it is unique per user), in other case of a job related project the key would in a separate file in a hashed way
#query_term: the terms for which we will be querying the api  (predefined for the purposes of task - although can be modified)
#page: which number of page to fetch from the result set, starting from 1 always
#fromdate: the date to begin fetching results from, we will request only from the day as per instructions to avoid unnecessary requests and data overload
#format_ : the format of the returned results
def query_guardian(apikey = '47267fdd-e0bb-4537-b3d1-a40678e080b5',query_term = "Justin%20AND%20Trudeau", page = 1, fromdate = '2018-01-01', format_ = 'json'):
    #define the guardian url to request data for, with all parameters set
    guardianurl = f'https://content.guardianapis.com/search?page={page}'\
    f'&q={query_term}'\
    f'&from-date={fromdate}'\
    f'&api-key={apikey}' \
    f'&page={page}'
    r = requests.get(guardianurl) #do the reequest
    return r.json() #return the results in json format

response = query_guardian() #first query of the guardian api
data =  pd.DataFrame(response['response']['results']) #create a dataframe with the initial returned results to handle and process it

#function which will be used to fetch all the pages for the specific response
def fetch_pages(dataframe, pages):
    #iterate from page 2 (as we already have page 1) until the last page from the returned results
    #we obtain the number of pages from the response at pages key
    for i in range(2,pages):
        time.sleep(2) #make a pause of 2 seconds among consecutive requests to avoid having our IP blocked
        response = query_guardian(page=i) #query with the next page the api
        #generate a new dataframe for the current page, and append it to the previously obtained results
        dataframe = pd.concat([dataframe,pd.DataFrame(response['response']['results'])])
    return dataframe

n_data = fetch_pages(data, response['response']['pages']) #collect the rest of the data from next pages

n_data.reset_index(inplace=True) #reset the index, will help us store the data in local file
n_data.to_json(r'guardian.json') #export the data into local file to avoing asking for the data again
```

In [3]:

```
#read the local json file which maintains our results
data=pd.read_json(r'guardian.json')
data.head() #take a glance of the dataset, by looking at its 5 rows, to see what data we have available
```

Out[3]:

	index	id	type	sectionId	sectionName	webPublicationDate	webTitle

	index	id	type	sectionId	sectionName	webPublicationDate	webTitle
0	0	world/2021/oct/26/canada-cabinet-reshuffle-melanie-joly-anita-anand-justin-trudeau	article	world	World news	2021-10-26T16:50:47Z	Justin Trudeau names women to top posts in Canada cabinet reshuffle
1	1	world/2021/sep/21/justin-trudeau-wins-third-election-victory	article	world	World news	2021-09-21T17:02:39Z	Justin Trudeau secures a third victory in an election ‘nobody wanted’
2	2	world/2021/sep/10/canada-election-rivals-force-justin-trudeau-on-to-defensive-in-leaders-debate	article	world	World news	2021-09-10T04:04:17Z	Canada election: rivals force Justin Trudeau on to defensive in leaders’ debate
3	3	world/2021/sep/20/justin-trudeaus-bid-for-third-term-in-balance-as-canada-goes-to-polls	article	world	World news	2021-09-20T05:00:03Z	Justin Trudeau’s bid for third term in balance as Canada goes to polls
4	4	global-development/2021/oct/29/trudeau-government-canada-indigenous-children	article	global-development	Global development	2021-10-29T22:34:34Z	Trudeau files last-ditch appeal against billions for Indigenous children

In [4]:

```python
#function to process the data and extract the number of articles per day
#parameters: the dataset as obtained previously, in DataFrame format
def number_of_articles(data):

    data['date'] = pd.to_datetime(data.webPublicationDate,format='%Y/%m/%d').dt.date #create a new column with only the date of the article in YEAR-MONTH-DAY FORMAT
    df = pd.DataFrame(data.groupby('date').size()) #group the data by date and count them
    df = df.rename({0:'No. of articles'},axis=1) #rename the columns as per instructions
    return df #return the dataframe
```

In [5]:

```
#Question 2: Count how many articles about Justin Trudeau have been posted
since 01.01.2018 until today:
q2_df = number_of_articles(data) #use the previously defined function to
process the data from 2018-01-01
q2_df #show the results
```

Out[5]:

date	No. of articles
2018-01-01	1
2018-01-03	1
2018-01-10	1
2018-01-13	1
2018-01-19	2
...	...
2021-11-19	3
2021-11-23	1
2021-11-24	1
2021-11-25	1
2021-11-30	2

602 rows × 1 columns

In [6]:

```
#question 3: Calculate the average of all days for the above-mentioned period
from “No. of articles”.
print('The average number of articles per day is: ',q2_df['No. of
articles'].mean()) #use pandas.mean() function to calculate the average
number of articles per day
```

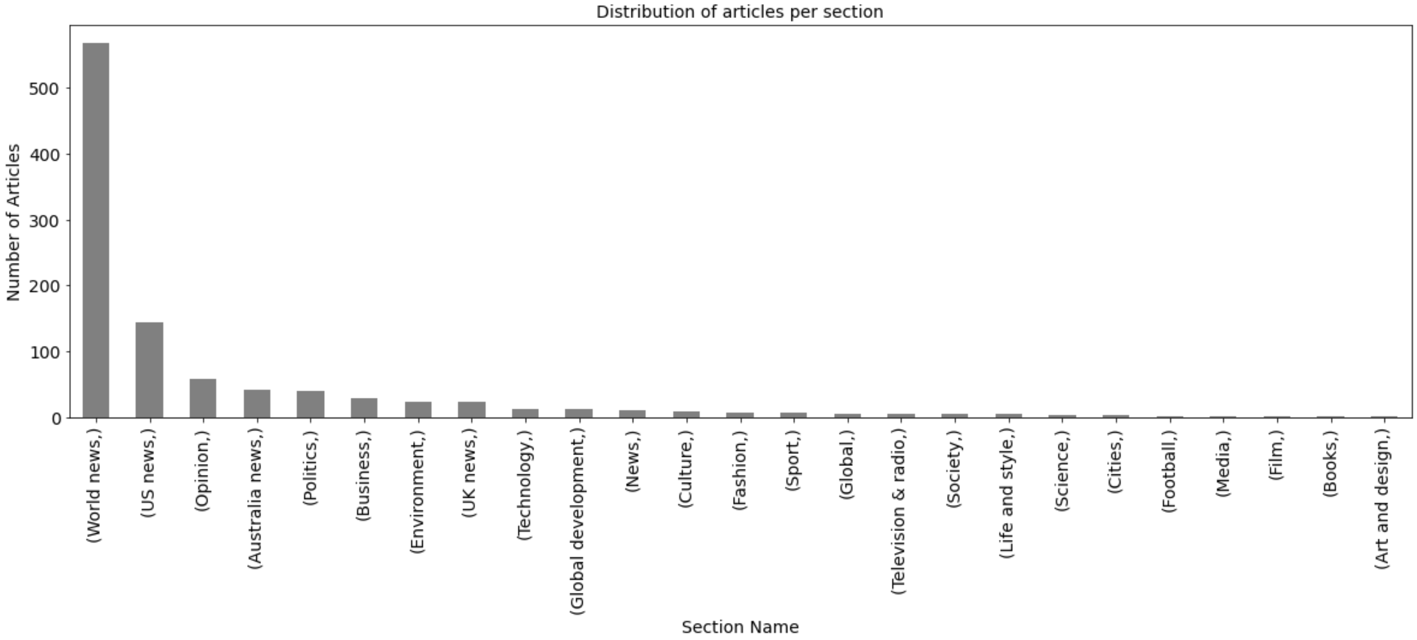
```
The average number of articles per day is:  1.6943521594684385
```

In [7]:

```
#question 4: In which section are most articles written?
q4 = pd.DataFrame(data.sectionName) #extract the section names form the data
q4.value_counts().plot(kind='bar',color='gray',figsize=(20,6), fontsize=14)
#count the values of each section and then plot it in descending order
plt.xlabel('Section Name', fontsize=14) #set label of X axis
plt.ylabel('Number of Articles', fontsize=14) #set label of Y axis
plt.title('Distribution of articles per section', fontsize=14) #set title of
figure
```

Out[7]:

```
Text(0.5, 1.0, 'Distribution of articles per section')
```

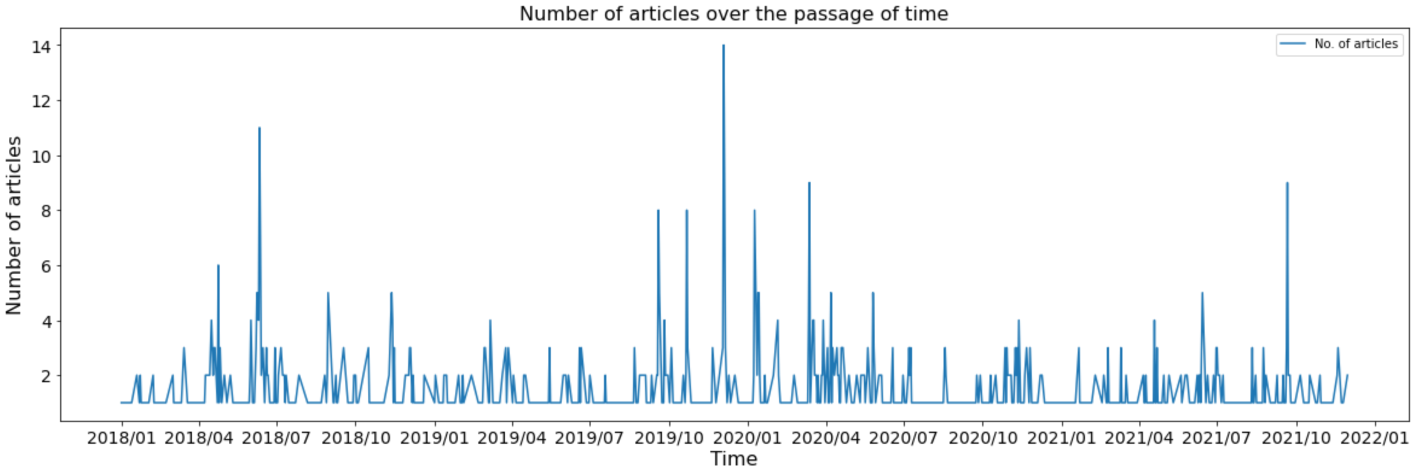


As we can see from the figure above, where sections are on X axis and Number of articles are on Y axis, World news is the section with the highest number of written articles

In [8]:

```
#Question 5: Show the evolution of the "No. of articles" over time for the above period.

#a function that will help us visualise the time series analysis we need for the
#evolution of number of articles since the start date, input is the data in dataframe format
#and output is a figure visualising the evolution of articles
def time_series_analysis(df):
    print('In the following figure, we can see the evolution of number of articles over the passage of time')
    print('X axis is representing the time, in ascending order (from start date of interest) until today')
    print('Y axis is showing the number of articles for each corresponding day')
    df['date'] = (df.index) #create new feature at the data frame of question 2, which will be the date
    df['date']=df['date'].apply(lambda x: x.strftime("%Y %m")) #transform the date into YEAR-MONTH format, to help us visualise it in lucid way


    ax = plt.gca() # get current axis

    df.plot(kind='line', y='No. of articles', ax=ax,figsize=(20,6),fontsize=14) #plot the data in timeseries manner
    ax.xaxis.set_major_formatter(DateFormatter("%Y/%m")) #format the dates in X axis in Year/Month format
    ax.xaxis.set_major_locator(mdates.MonthLocator(interval=3)) #Show every three months
    plt.xlabel('Time',fontsize=16) #set the label of X axis
    plt.ylabel('Number of articles',fontsize=16) #set the label of Y axis
    plt.title('Number of articles over the passage of time',fontsize=16) #set the title of the figure


    plt.show()


time_series_analysis(q2_df) #call the function to make the time series analysis for Question 5, by using the results of question 2 as instructed
```

```
In the following figure, we can see the evolution of number of articles over the passage o
f time
X axis is representing the time, in ascending order (from start date of interest) until to
day
Y axis is showing the number of articles for each corresponding day
```



In [ ]:

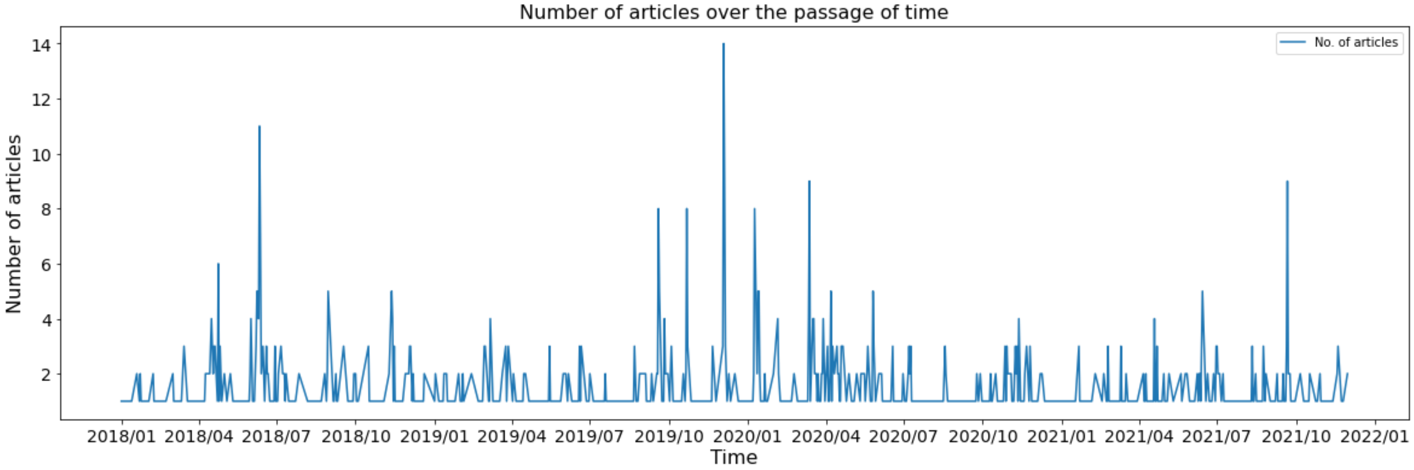
```

```

Question 6: Are there any unusual events in the time series under investigation?

In [9]:

```
#get the statistics of the data regarding the number of articles per day
q2_df.describe(), time_series_analysis(q2_df) #call the function to make the time series analysis for Question 5
```

```
In the following figure, we can see the evolution of number of articles over the passage o
f time
X axis is representing the time, in ascending order (from start date of interest) until to
day
Y axis is showing the number of articles for each corresponding day
```



Out[9]:

```
(        No. of articles
 count        602.000000
 mean           1.694352
 std            1.329156
 min            1.000000
 25%            1.000000
 50%            1.000000
 75%            2.000000
 max           14.000000,
 None)
```

According to the above figure, we can see the number of articles per day (Y axis = number of articles), over the passage of time (X axis = time). More precisely, we can see from the graph that the majority of days returned from guardian api have 1 article (which can be confirmed from the statistics presented below the graph) which essentially, let us know that 75% of days have 2 or less articles, consequently days with 8 or more seem to be unusual, and we could claim thath they are outliers in our distribution.

Question 7: If so, show these. Why are these unusual? (Define for yourself what you want to show by ordinary or unusual).

We could characterize as an unusal event, dates that there is significant increase of number of articles compared to the rest of the days. More precisely, we can see from the statistics of the data, that the average articles per day are ~1.7 we could claim that we have 2 subcateogies of unusual events.

1. extremely unusual event (8 or more articles per day)
2. unusual event -as upper quartile is more than 2 articles-, we will select from 2 up to 8 articles per day) Based on that assumption, we are going to focus initially on the first category, of extermely unusual events

In [10]:

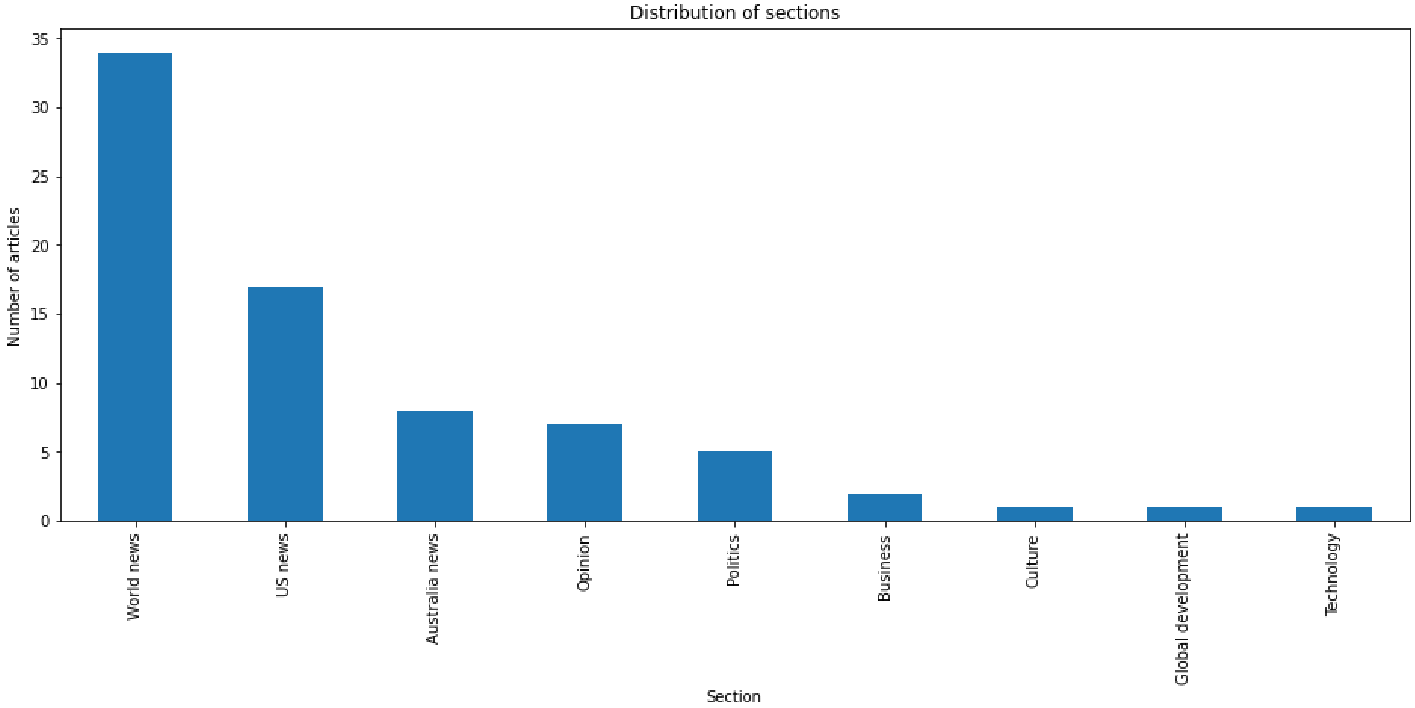
```
indeces = q2_df[q2_df['No. of articles']>=8].index #get the indeces of extremely unusual events, to extract those rows
q6_df = data[data.date.isin(indeces)] #extract data related to extremely unusual events
```

In [11]:

```
q6_df = q6_df[['type','sectionName','webTitle','date']].sort_values(by='date') #check the distributios of sections
q6_df.groupby('sectionName')['date'].count().sort_values(ascending=False).plot(kind='bar',figsize=(16,6))
```

```
#plot the distrbution of sections
plt.title('Distribution of sections')
plt.xlabel('Section')
plt.ylabel('Number of articles')
```

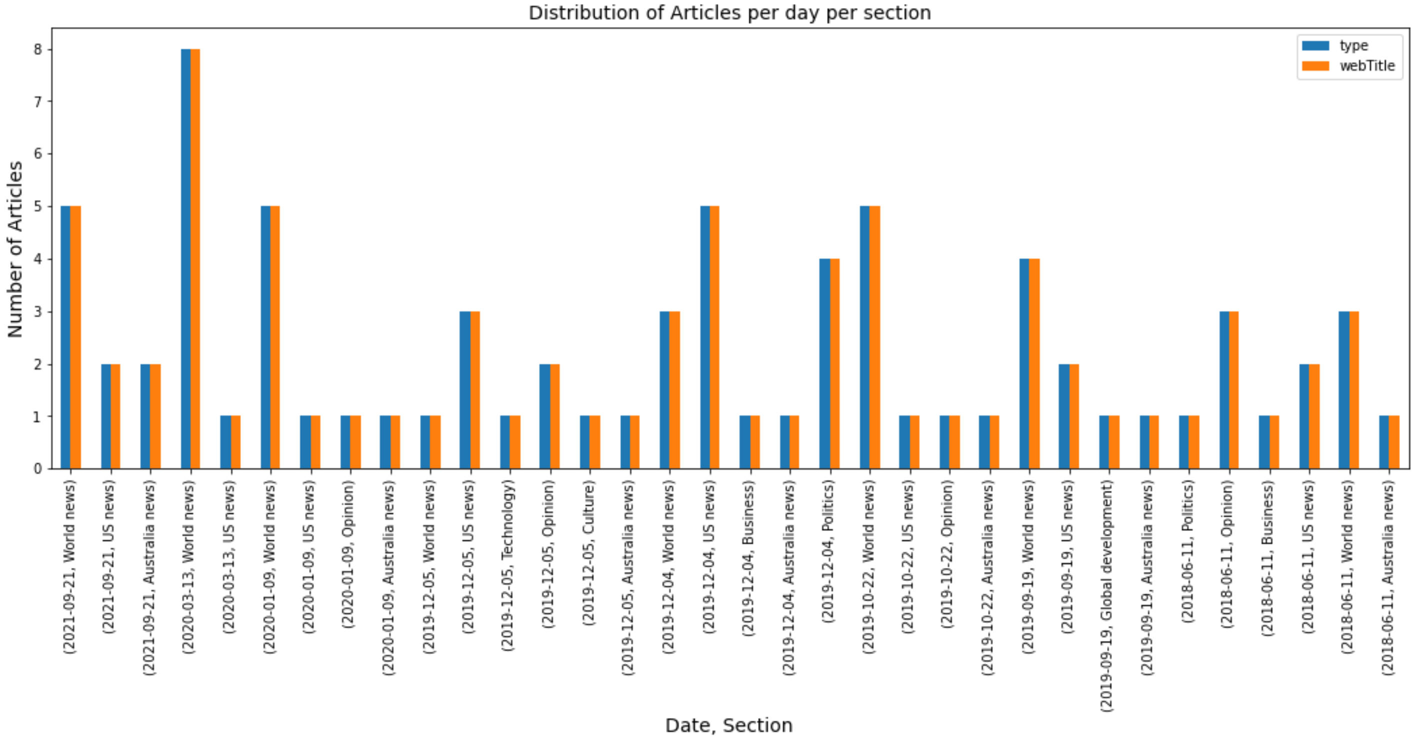
Out[11]:
```
Text(0, 0.5, 'Number of articles')
```



In the extremely unusual events, we can see that the articles stem from various News sections in first place (World News, US News, Australia News), while the second more usual sections are opinion and politics

In [12]:
```
#group data per day, per section and count the number of articles, and plot
them
q6_df.groupby(['date','sectionName']).count().sort_values(by='date',ascending=F
(18,6))
plt.title('Distribution of Articles per day per section', fontsize=14)
plt.ylabel('Number of Articles', fontsize=14)
plt.xlabel('Date, Section', fontsize=14)
```

Out[12]:
```
Text(0.5, 0, 'Date, Section')
```



In the above figure, we can see a distribution of the sections (to which articles belong) for every date that falls into our unusual assumption, ordered by date

- 2018-06-11 : 3 World News + 3 Opinion
- 2019-10-22 : 5 World News
- 2019-12-04 : 4 US news + 4 Politics
- 2019-12-05 : 4 US news
- 2020-01-09 : 5 World News
- 2020-03-13 : 8 World News
- 2021-09-21 : 5 World News

From the above, we can see that in every occasion there is some kind of hot event happening regardom Justin Trudeau as there are many articles steming from news section, as previously mentioned, however, we will explore all categories together.

In [13]:

```
#create groups by date
groups = q6_df.sort_values(by='date').groupby('date')
#extract the names of the groups, to use them for analysiis
names = []
for name,group in groups:
    names.append(name)
```

Question 8: Based on question one. Show the cause of the unusual event.

Lets now investigate each date individually, so we can draw some conclusions for each day

In [14]:

```
#a function that will be used for analysis, showing the date alongside with
```

```
article titles, section and type
def analysis(idx):
    print(names[idx])
    display(q6_df[q6_df.date == names[idx]]
[['type','sectionName','webTitle']].sort_values(by='sectionName'))
```

In [15]:

```
analysis(0) #analysis for first unusual date
```

```
2018-06-11
```

	type	sectionName	webTitle
900	article	Australia news	Morning mail: Trump meets Kim, child protection failings, Spain takes migrants
913	liveblog	Business	UK manufacturing output shrinks; Markets shrug off G7 debacle - as it happened
746	article	Opinion	Trump, Merkel, Macron: the G7 photos worth a thousand words | Hannah Jane Parkinson
714	article	Opinion	Canada and America are cousins. We don't stab each other in the back | Margaret MacMillan
633	article	Opinion	Trump is a bully who thought Canada was weak. He was wrong about us | Jen Gerson
883	liveblog	Politics	Nigel Farage interviewing Arron Banks on LBC – as it happened
821	article	US news	Kim Jong-un and Trump 'to discuss permanent peace-keeping' at Singapore summit
537	article	US news	Trudeau 'stabbed us in back' on trade, says Trump chief economic adviser
899	article	World news	North Korea summit: US president says 'we will be fine' as meeting nears
709	article	World news	Q&A: how damaging was Donald Trump's G7 blow-up?
658	article	World news	'Prepare for the worst': souring Canada-US relations fuel worries of trade war

At 2018-06-11, we can see that the articles are distributed accross many categories, however we can identify some issues clearly,

1. USA - Canada relations seem to be in tangible point
2. There is concern about North Korea
3. Topics stemming form G7, where Justin Trudeau belongs to, as Canadian prime minister,

Moreover, at liveblogs we can see that the topics of discussion are not directly linked with Justin Trudeau and perhaps we need the actual text to present a clear-cut answer about its corellation with Justin Trudeau.

In [16]:

```
analysis(1)  #analysis for second unusual date
```

```
2019-09-19
```

	type	sectionName	webTitle
217	article	Australia news	Morning mail: climate strike, Trudeau blackface, bird extinctions
356	article	Global development	Commonwealth ministers look to revitalise progress on gender equality
119	article	US news	US briefing: Greta Thunberg, Justin Trudeau and a Trump whistleblower
120	article	US news	US briefing: Greta Thunberg, Justin Trudeau and a Trump whistleblower
167	article	World news	Thursday briefing: Trudeau apologises for 'brownface' picture
151	article	World news	How will Justin Trudeau's blackface photos affect Canada's election?

	type	sectionName	webTitle
108	article	World news	Justin Trudeau brownface: Canada PM apologises after image emerges
431	article	World news	Jacinda Ardern mistakes Japan for China during Tokyo visit

At 2019-10-22, we can see that there is only one main topic,

1. Canadian Elections

but again, we can see some irrelevant stuff such as the one entitled as 'Tuesday briefing: Johnson – let's get on with the WAB'

In [17]:

```
analysis(2)   #analysis for third unusual date
```

```
2019-10-22
```

	type	sectionName	webTitle
466	article	Australia news	Morning mail: Brexit fast-track rejected, farmers' drought demands, reality show payout
210	article	Opinion	The Guardian view on the Canadian election: a win for Trudeau, but not a triumph | Editorial
310	article	US news	US briefing: Trudeau's narrow win, GOP disunity and ocean acidification
234	article	World news	Trudeau faces rough road as Canada's minority parties lay out their conditions
373	article	World news	Tuesday briefing: Johnson – let's get on with the WAB
231	liveblog	World news	Canada election 2019: 'We'll govern for everyone' says Trudeau, after narrow win – as it happened
187	article	World news	Canada elections: Trudeau wins narrow victory to form minority government
160	article	World news	Justin Trudeau's victory is a death knell for Canada's fledgling far-right

At 2019-12-04, we can see that there is no direct linkage of the articles with Justin Trudeau, so, we could assume that references to Canadian PM are part of role he holds as a Nato member. More precisely, topics that are being discussed in those articles are

1. About an impeachment report vote
2. Nato summit issue (Nato leaders joking on USA president)

while the trend on liveblog continues, with the corresponding titles not to be able to be linked with Justin Trudeau directly

In [18]:

```
analysis(3)   #analysis for fourth unusual date
```

```
2019-12-04
```

	type	sectionName	webTitle
303	article	Australia news	Morning mail: Trump snubs Nato, Taylor inquiry call, Wilderness Society questions
809	liveblog	Business	M&G suspends property fund amid Brexit uncertainty and retail crisis - business live
347	article	Politics	Andrew Sparrow's election briefing: Trump visit ends without jeopardy for Johnson
343	article	Politics	What's the joke? Mugged off Trump sulks his way through Nato summit | John Crace
494	article	Politics	PM's Operation Avoid Trump goes off almost without a hitch | Marina Hyde

	type	sectionName	webTitle
345	liveblog	Politics	Boris Johnson denies joking about Donald Trump at Nato reception and not taking him seriously – as it happened
528	article	US news	US briefing: impeachment, Nato summit and Kamala Harris drops out
492	liveblog	US news	House intelligence committee votes to pass impeachment report – as it happened
334	article	US news	Trump cuts short Nato summit after fellow leaders' hot-mic video
489	liveblog	US news	House intelligence committee votes to pass impeachment report – as it happened
315	article	US news	Footage appears to show world leaders joking about Trump at Nato summit
425	article	World news	How does Nato look at the age of 70? It's complicated
390	article	World news	Long-term damage from logging hits ability of Canada's forests to regenerate
526	article	World news	Macron clashes with both Erdoğan and Trump at Nato summit

At 2019-12-05, there are news about Nato and Donald Trump, and american election as they were getting closer (took place 2020-11) alongside with a wide variety of articles such as technological and corporate (e.g. Facebook),

Consequently, we can not really interpret the spike on number of articles regarding Justin Trudeau in this occasion

In [19]:
```
analysis(4)  #analysis for fifth unusual date
```

```
2019-12-05
```

	type	sectionName	webTitle
664	article	Australia news	Inside the hate factory: how Facebook fuels far-right profit
307	article	Culture	Trevor Noah: Trump realized 'all the cool kids at school are laughing' at him
372	article	Opinion	What's it like to stand stark naked on the world stage? Ask Donald Trump | Richard Wolffe
490	article	Opinion	Nato is not braindead. But it does need a shot of adrenaline | Michael H Fuchs
453	article	Technology	Monetising hate: covert enterprise co-opts far-right Facebook pages to churn out anti-Islamic posts
313	article	US news	Joe Biden targets Trump's Nato sore spot with video mash-up of mockery
311	article	US news	US briefing: Trump's Nato flounce, impeachment and George Zimmerman
394	article	US news	John Kerry endorses Joe Biden in 2020 Democratic primary race
505	article	World news	Thursday briefing: Pique Trump – after farcical exit, back to election

At 2020-01-09 there is one main topic of interest, which is the crash of an Iranin airplane, and its related ton Justin Trudeau, as we can see clearly there is a statement from the PM himself, but also, a lot of mentions about western leaders, where Canada is key member

In [20]:
```
analysis(5)  #analysis for sixth unusual date
```

```
2020-01-09
```

	type	sectionName	webTitle
994	liveblog	Australia news	'If you are told to leave, leave,' Daniel Andrews warns – as it happened

	type	sectionName	webTitle
144	article	Opinion	Dear Justin Trudeau, a beard will only make it look like something has gone wrong in your life\r\n
592	liveblog	US news	Congress to vote on curbing president's war powers – as it happened
177	article	World news	Iran crash: plane shot down by accident, western officials believe
434	article	World news	Iran plane crash: Missile strike and engine failure being explored
94	article	World news	Justin Trudeau: Canada 'will not rest' until it gets answers about plane crash
461	article	World news	Catastrophic failure of Ukraine jet in Iran suggests missile strike
293	article	World news	Australia echoes western leaders in alleging Iran accidentally downed Ukraine plane

At 2020-03-13, it is the period that COVID became part of our lives, and we can see that there the following topics in the articles type

1. Justin Trudeau's wife caught the virus and PM got in quarantine
2. A lot of discussion about coronavirus

While the liveblog responses, are concerned about financial markets and Donald Drump actions against COVID, thus they could be characterised irrelevant

In [21]:
```
analysis(6)  #analysis for seventh unusual date
```

```
2020-03-13
```

	type	sectionName	webTitle
955	liveblog	US news	Trump has 'no plans' for coronavirus test despite contact with infected Bolsonaro aide – as it happened
166	article	World news	Coronavirus pandemic reaches world leaders and disrupts global sporting events
67	article	World news	Justin Trudeau in self-isolation after wife Sophie tests positive for coronavirus
454	article	World news	'Do not let this fire burn': WHO warns Europe over Covid-19
204	article	World news	Friday briefing: F1 non-starter, Canada PM's wife has Covid-19
265	article	World news	Coronavirus latest: 13 March at a glance
400	article	World news	Coronavirus latest: 13 March at a glance
182	liveblog	World news	Markets fall again as global Covid-19 cases near 130,000 – as it happened
46	article	World news	Justin Trudeau announces sweeping steps to tackle coronavirus in Canada

At 2021-09-21, we can see that in the articles type there is one main topic

1. Win of Justin Trudeau at Canadian Elections, which can be considered as very hot topic and would expected the spike

While the liveblog ones, again seem irrelevant to Justin Trudeau

Daily automated job

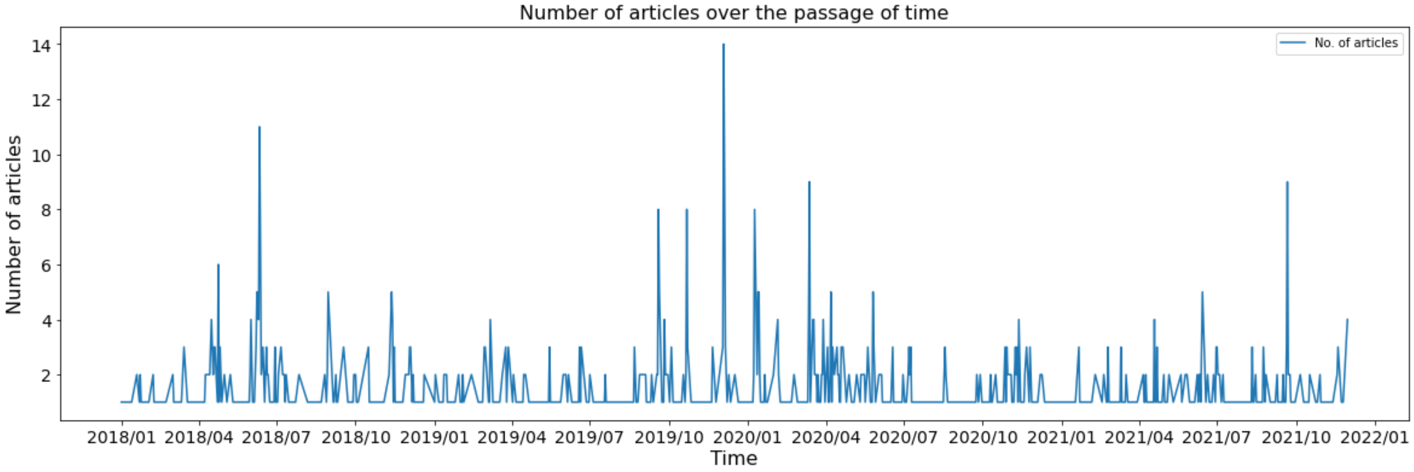
Due to timing constraints, we will create the job based on the assumption that the user will be using the same local machine, provided we have had more time the approach we would follow, would be to store the data in a SQLite Database, then update them with the daily results and present the data the same way, below you can see the function which will be doing the job, however a separate script has been created to present it as whole, and it is explained in latter cells

In [22]:

```
def daily_update(data):
    data=pd.read_json(r'guardian.json') #read the pre-stored data
    new_data = data.copy(deep=True) #create a new copy of them, to avoid potential bugs or issues which will lead to a loss of data
    today = str(datetime.now())[:10] #get (everyday's) today's date
    print(f"Before we make request for day {today}, we have {new_data.shape[0]} entries") #print the size of dataset prior to todays request
    query = query_guardian(fromdate=today) #request from the guardian api to fetch articles about today
    query_df = pd.DataFrame(query['response']['results']) #create dataframe from todays request
    if query['response']['total'] >0: #check if there articles today
        print(f"There are {query['response']['total']} New articles today!\n")
        #check if there are more than 10 articles today, as the page size is 10, then we need to fetch more pages
        if query['response']['total'] > 10:
            query_df = fetch_pages(query_df,query['response']['pages']) #fetch next pages until we fetch them all

        new_data = new_data.append(query_df) #append to the entire data-set today's responses
        new_data.reset_index(inplace=True) #reset the index, will help us store the data in local file
        new_data.to_json(r'\guardian.json') #export the data into local file to avoing asking for the data again
        print(f"After todays {today} request we have {pd.read_json(r'guardian.json').shape[0]} entries total") #print the size of dataframe after todays request



    df_numberofarticles = number_of_articles(new_data) #use the function to count todays articles
    time_series_analysis(df_numberofarticles) #present the time series analysis (Question 5)



daily_update(data) #call the function for daily update
```

```
Before we make request for day 2021-11-30, we have 1020 entries
There are 2 New articles today!
```

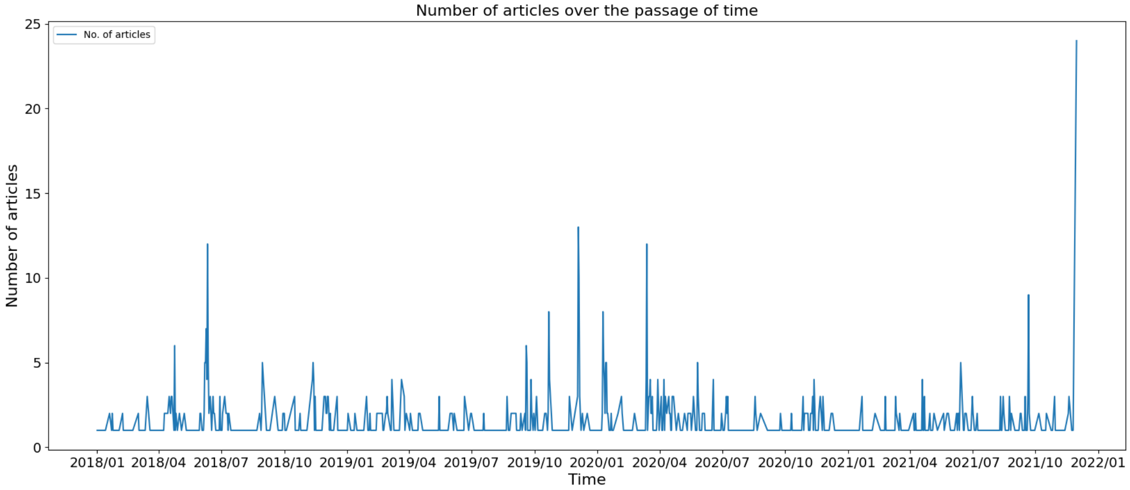
```
After todays 2021-11-30 request we have 1020 entries total
In the following figure, we can see the evolution of number of articles over the passage o
f time
X axis is representing the time, in ascending order (from start date of interest) until to
day
Y axis is showing the number of articles for each corresponding day
```



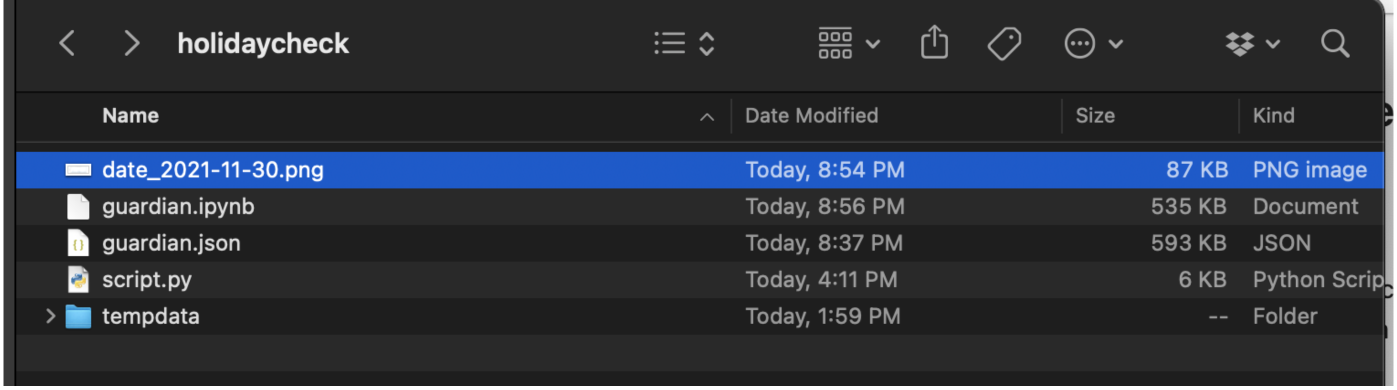
In the next lines we are going to present the commands -in a terminal window- used for the daily automaded job, using cron

What is cron though?

Cron is a job scheduler, executable from command line (command line utility), which will allow us to run the script with the required functions to obtain thedaily number of articles, and then present it to the user As we want it daily, we have decided to execute the script at 23:59 every day, in order to fetch all the articles from that particular day The script is attached and named script.py, which will save a new image in the format date_YEAR-MM-DD.png for every day, as shown below



Below, we can see how the finder looks like after we have executed the script (though for testing purposes not at the given time of the day)



Name	Date Modified	Size	Kind
date_2021-11-30.png	Today, 8:54 PM	87 KB	PNG image
guardian.ipynb	Today, 8:56 PM	535 KB	Document
guardian.json	Today, 8:37 PM	593 KB	JSON
script.py	Today, 4:11 PM	6 KB	Python Scrip
tempdata	Today, 1:59 PM	--	Folder

Commands executed in command line for job scheduling with cron

pip install crontab : command to install crontab

pwd : command to get the absolute path

crontab -e : command to create a cron job, immediately followed by I button on keyboard : this way we enter in cron insert mode

59 23 * python3 path/to/script.py : execute every day at 23:59, the file in path

press ESC : to exit from crontab, and the cron job has been created

crontab -l :to verify that the job has been created

```
In [ ]:
```