Wiki Traffic Forecast

Project Report(MSDS 596)

Chaitanya Sharma Domudala (NetID: cd817)

Nov 21, 2020

Introduction

The Project centers around the comprehensive Exploratory Data Analysis of estimating the future values of multiple time series, as it has always been one of the most challenging problems in the field. All the more explicitly, we aim on the issue of guaging future web traffic for roughly 145,000 Wikipedia articles.

Sequential or temporal observations emerge in many key real-world problems, ranging from biological data, financial markets, weather forecasting, to audio and video processing. The field of time series encapsulates many different problems, ranging from analysis and inference to classification and forecast.

Dataset

The Dataset contains about 145k time series records. Each of these record represent a number of daily views of a different Wikipedia article, starting from July, 1st, 2015 up until December 31st, 2016:

train.csv holds the traffic data, where each column corresponds to a particular date and each row corresponds to a particular article

Exploratory Analysis & Visualization

Data Overview

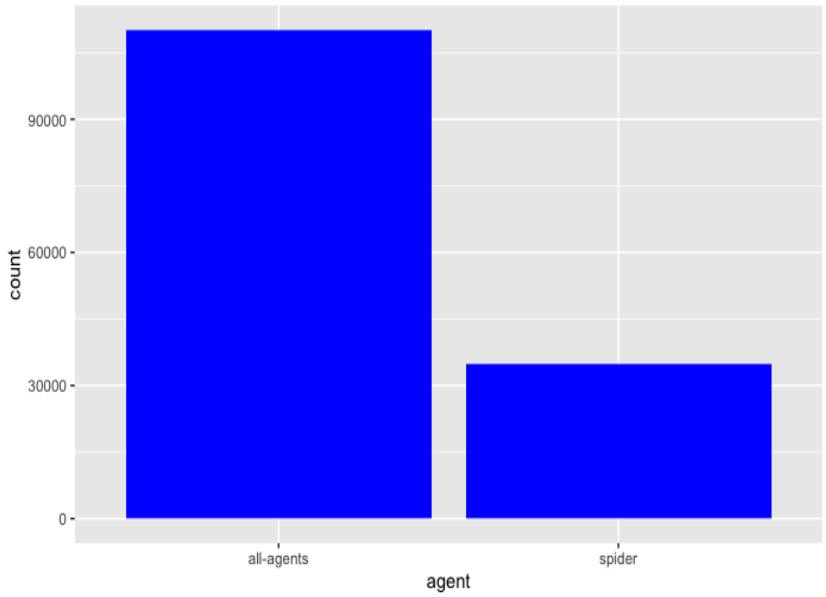
Firstly, looked into structure and content of the dataset. Then checked for missing values & outliers, that are needed to be removed prior to the analysis. To handle the data in a much easier way, the data was split into two parts : one with article

information and other with time series data from date columns. Further, briefly separated the article info. into data from various sources due to different formats of the URL's and included the type of access, agent. This implementation helped to filter the parameters & search for specific page articles.

rowname	article	locale	access	agent
<chr>	<chr>	<chr>	<chr>	<chr>
78239	File:Comedy_Central_2011_Logo.svg	wikmed	mobile-web	all-agents
74018	List_of_NCAA_Men's_Division_I_Basketball_champions	en	mobile-web	all-agents
25815	Claude_Onesta	fr	all-access	all-agents
50321	Karlheinz_BÃ¶hm	de	all-access	spider
99305	ÐžÑ‚\200Ñ\217Ð´_Ñ\201Ð°Ð¼Ð¾Ñ‚Ð±Ð¸Ð¾Ñ‚	ru	all-access	all-agents

5 rows

Few visualisation plots were charted to have a look how the different meta-parameters are distributed & conclusions were drawn.



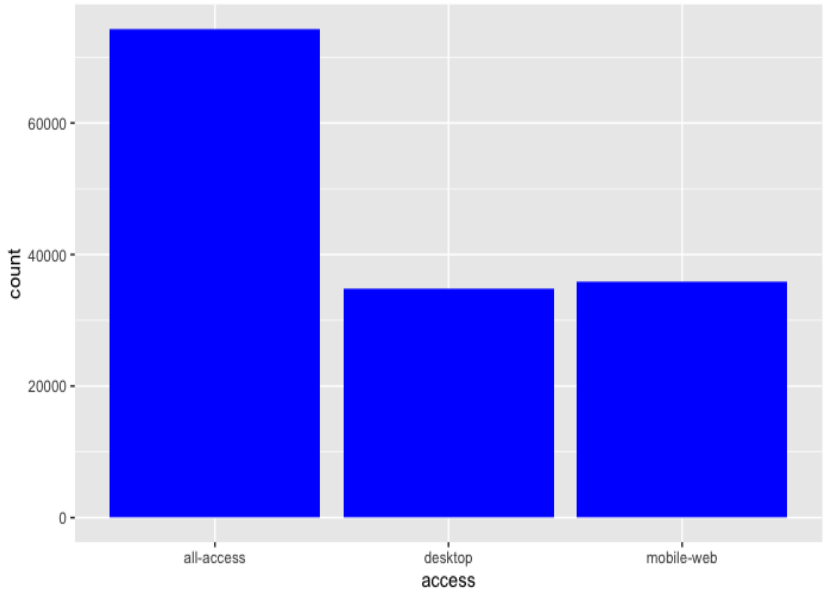
Analysis shows :

1) all-agents has total of 110150 pages.

2) spider has total of 34913 pages.

All-agent and Spider are two agents used for accessing page.

Pages with 'spider' agent is very less compared to that of 'all-agent'



Analysis shows :

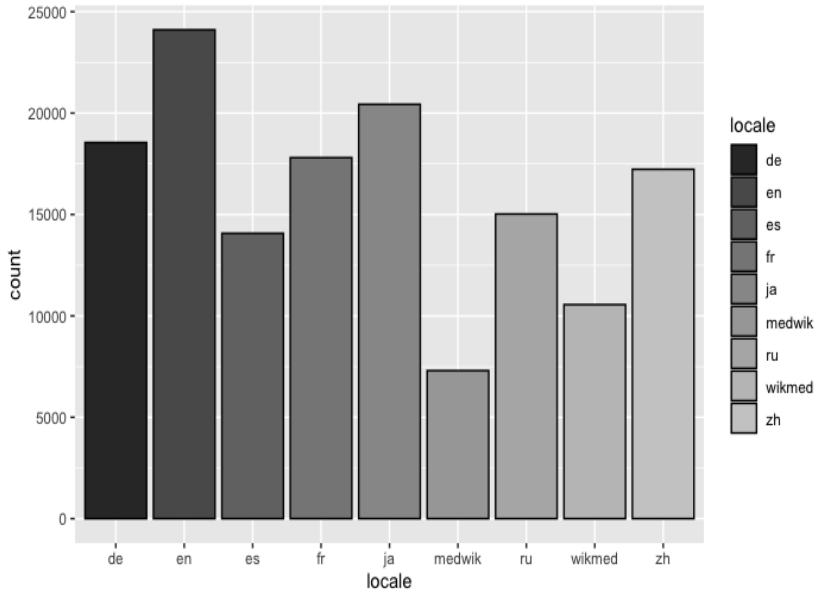
1) all-access has total of 74315 pages.

2) mobile-web has total of 35939 pages.

3) desktop has total of 34809 pages.

Page is mainly accessed from Three categories viz. all-access, Desktop access and Mobile-web.

More pages are from 'all-access' category compared to that of remaining categories.



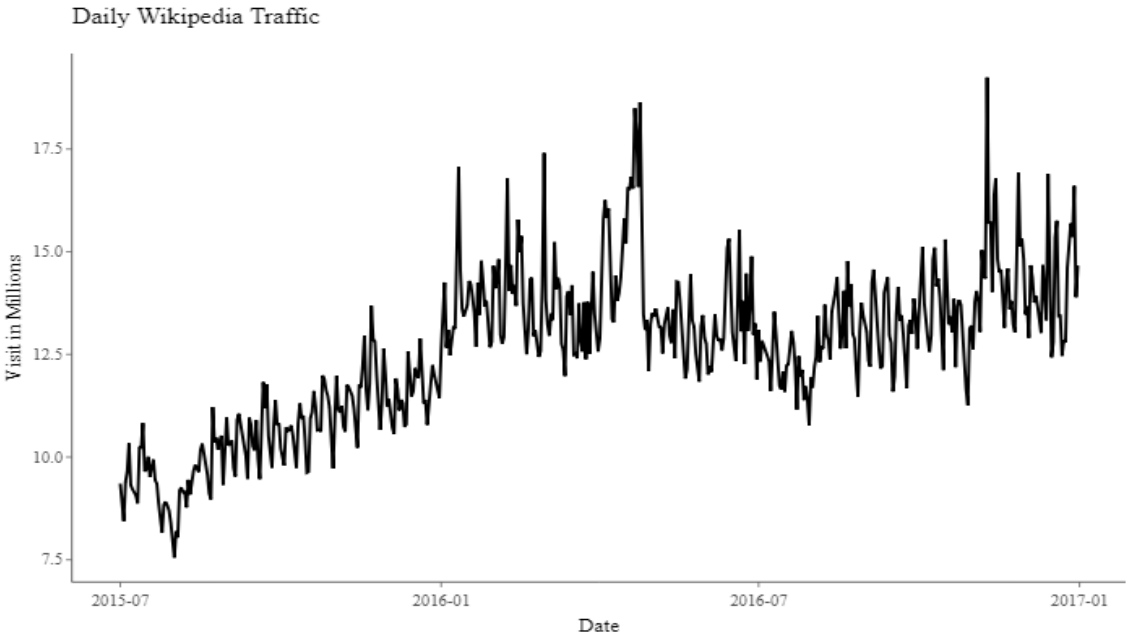
We find that our wikipedia data includes 7 languages: German, English, Spanish, French, Japanese, Russian, and Chinese. All of those are more frequent than the mediawiki and wikimedia pages.

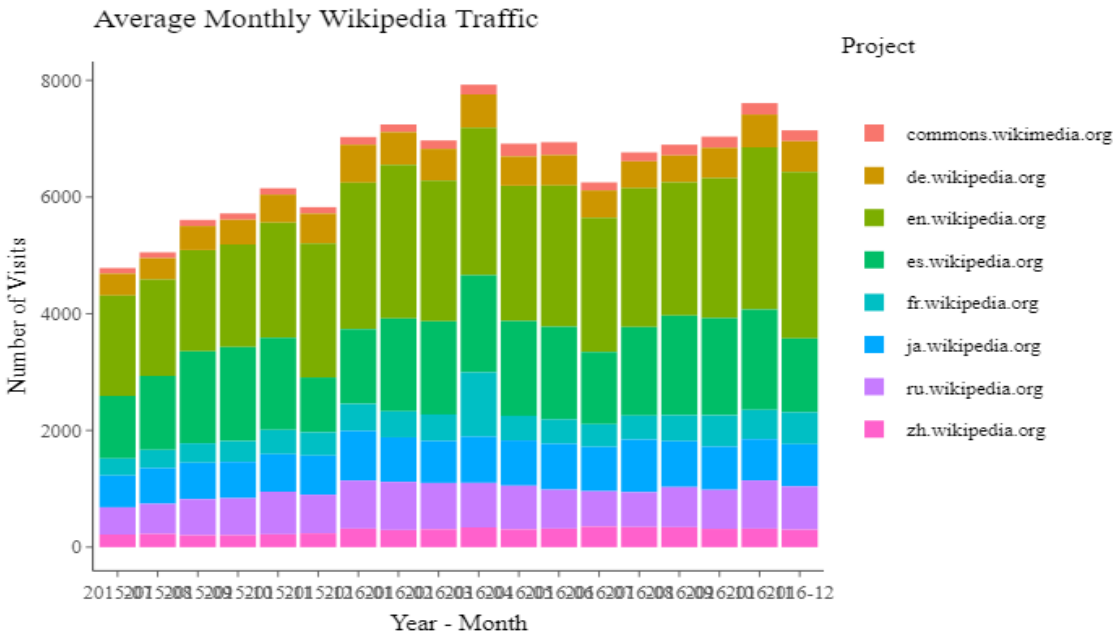
Since the most data was from wikipedia pages, created an individual data table for it including the visits each day, access type name of the search topic.

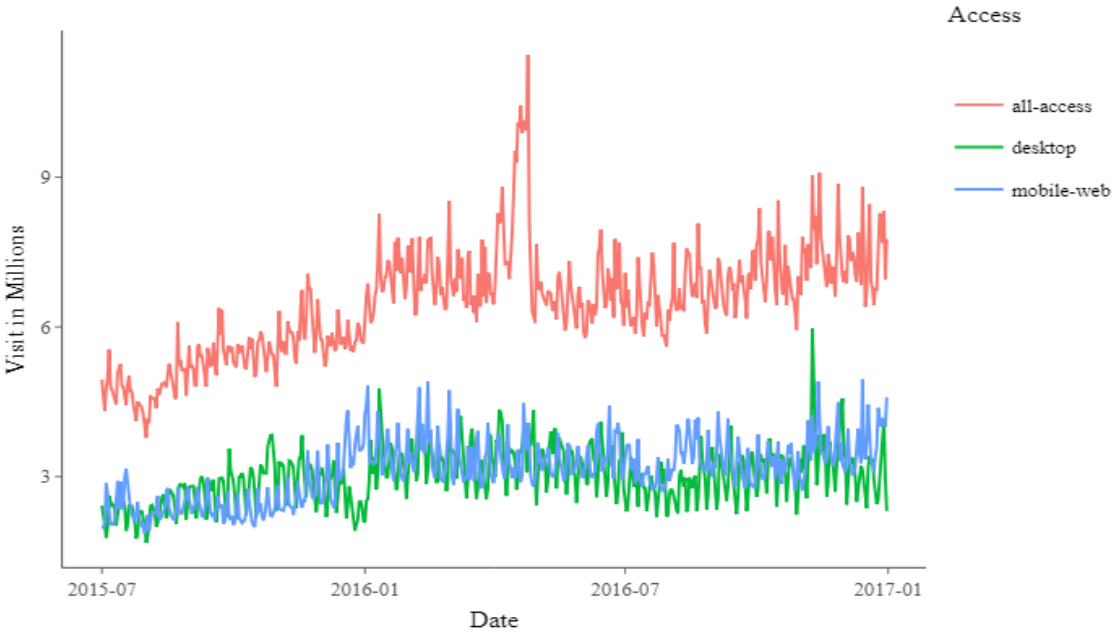
Date	Visit	Project	Access	Agent	Name
<chr>	<dbl>	<chr>	<chr>	<chr>	<chr>
2015-07-01	0	fr.wikipedia.org	all-access	spider	Sarah Stern
2015-07-01	26	de.wikipedia.org	mobile-web	all-agents	BjÃ¶rn Engholm
2015-07-01	6	fr.wikipedia.org	all-access	spider	Donald Trump
2015-07-01	17	zh.wikipedia.org	desktop	all-agents	å¾¼\220å\235å…ƒ
2015-07-01	68	en.wikipedia.org	desktop	all-agents	Jean Batten
2015-07-01	130	ru.wikipedia.org	all-access	all-agents	ÐšÐ°Ñ\201Ñ‚ÐµÑ\200Ð¾Ð²Ð°, Ð\220Ð½Ð½Ð° Ð Ð¾Ð¼Ð°Ð½Ð¾Ð²Ð½Ð°Ð²

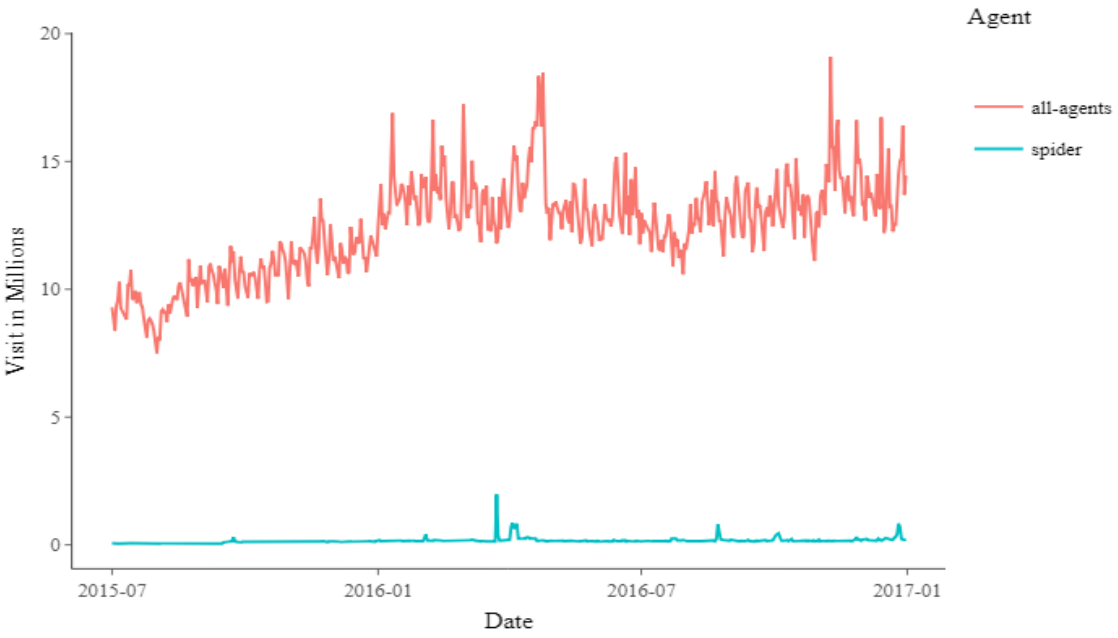
6 rows

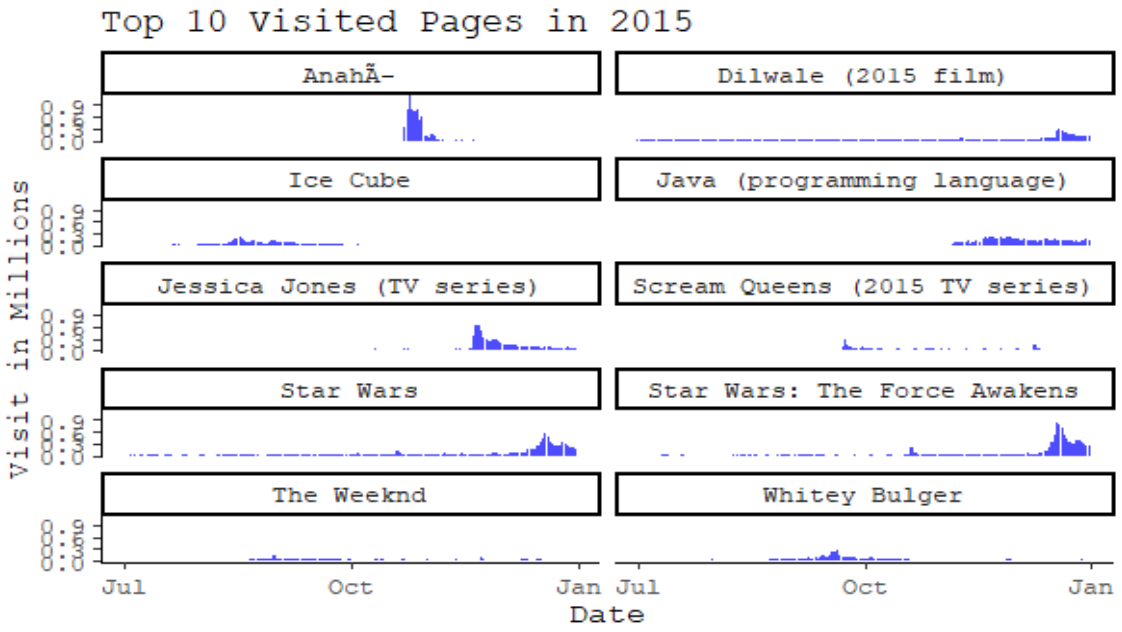
Various plots to extrapolate how the wikipedia data is spread out:

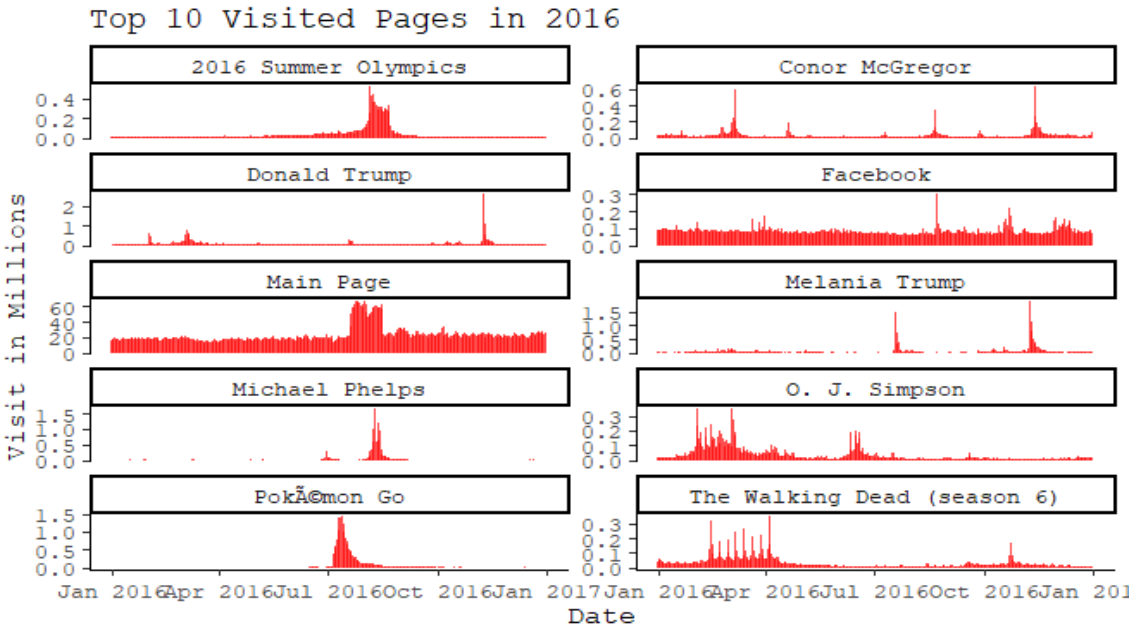

Daily Wikipedia Traffic
Visit in Millions
17.5
15.0
12.5
10.0
7.5
2015-07
2016-01
2016-07
2017-01
Date


Average Monthly Wikipedia Traffic
Number of Visits
8000
6000
4000
2000
0
Year - Month
Project
commons.wikimedia.org
de.wikipedia.org
en.wikipedia.org
es.wikipedia.org
fr.wikipedia.org
ja.wikipedia.org
ru.wikipedia.org
zh.wikipedia.org


Access
all-access
desktop
mobile-web
Visit in Millions
9
6
3
2015-07
2016-01
2016-07
2017-01
Date


Agent
all-agents
spider
Visit in Millions
20
15
10
5
0
2015-07
2016-01
2016-07
2017-01
Date

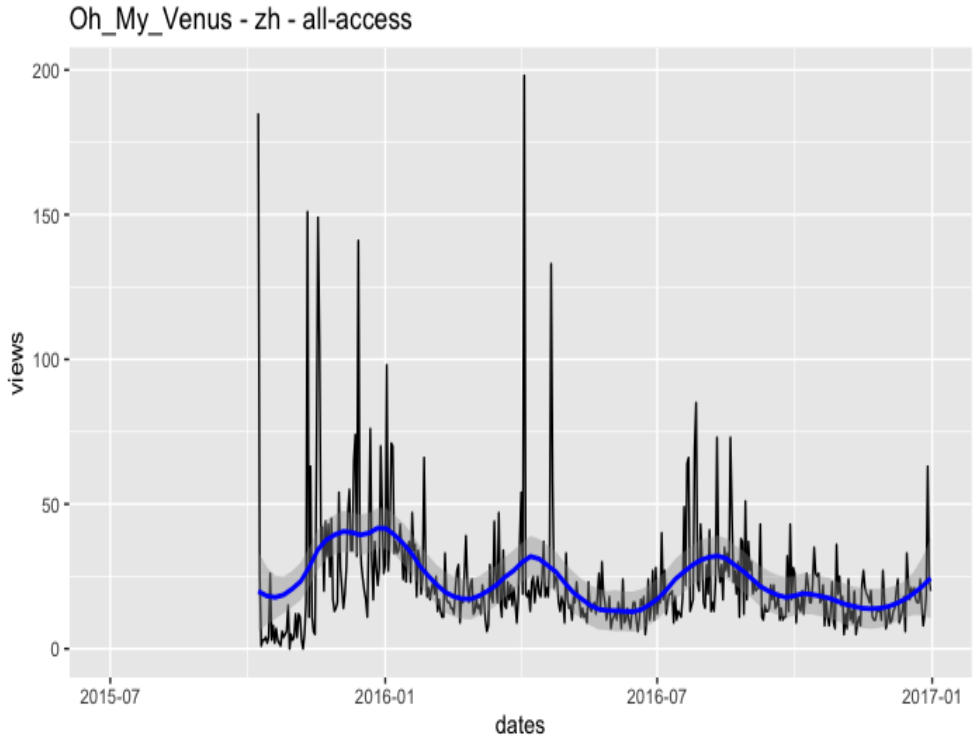

Top 10 Visited Pages in 2015
AnahÃ-
Dilwale (2015 film)
Ice Cube
Java (programming language)
Jessica Jones (TV series)
Scream Queens (2015 TV series)
Star Wars
Star Wars: The Force Awakens
The Weeknd
Whitey Bulger
Visit in Millions
Jul
Oct
Jan
Date


Top 10 Visited Pages in 2016
2016 Summer Olympics
Conor McGregor
Donald Trump
Facebook
Main Page
Melania Trump
Michael Phelps
O. J. Simpson
PokÃ©mon Go
The Walking Dead (season 6)
Visit in Millions
Jan 2016
Apr 2016
Jul 2016
Oct 2016
Jan 2017
Date

Time series Extraction & Analysis

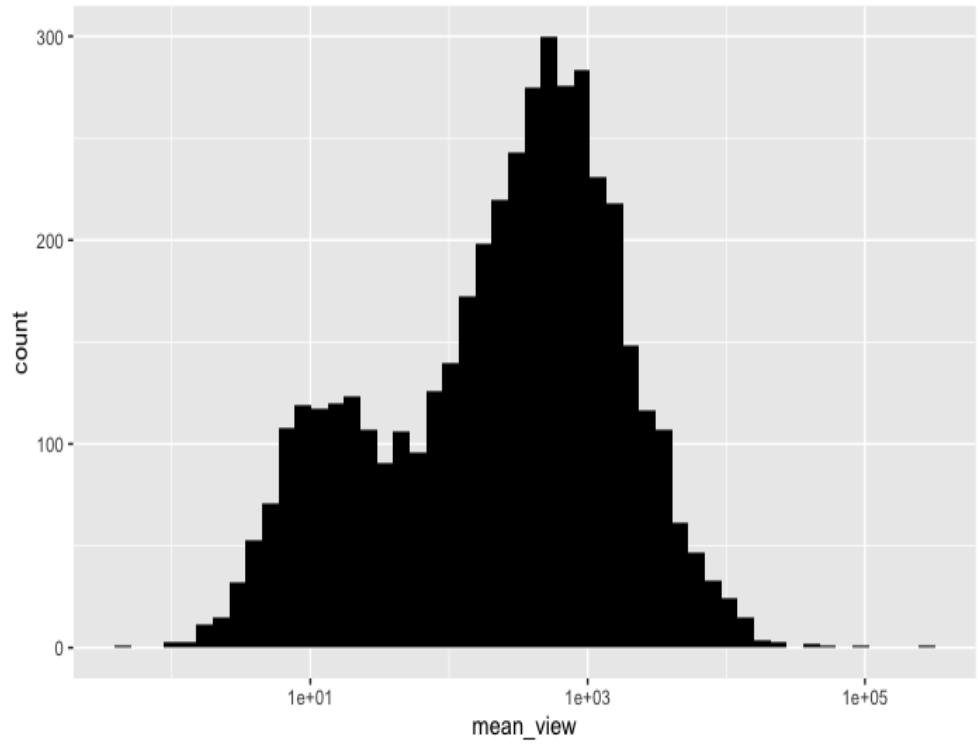
Built up a custom function to specifically extract time series data of any article and a custom plotting function to visualise specific article's time series data and extract its meta data.

The time series data of row number 112 is depicted below.

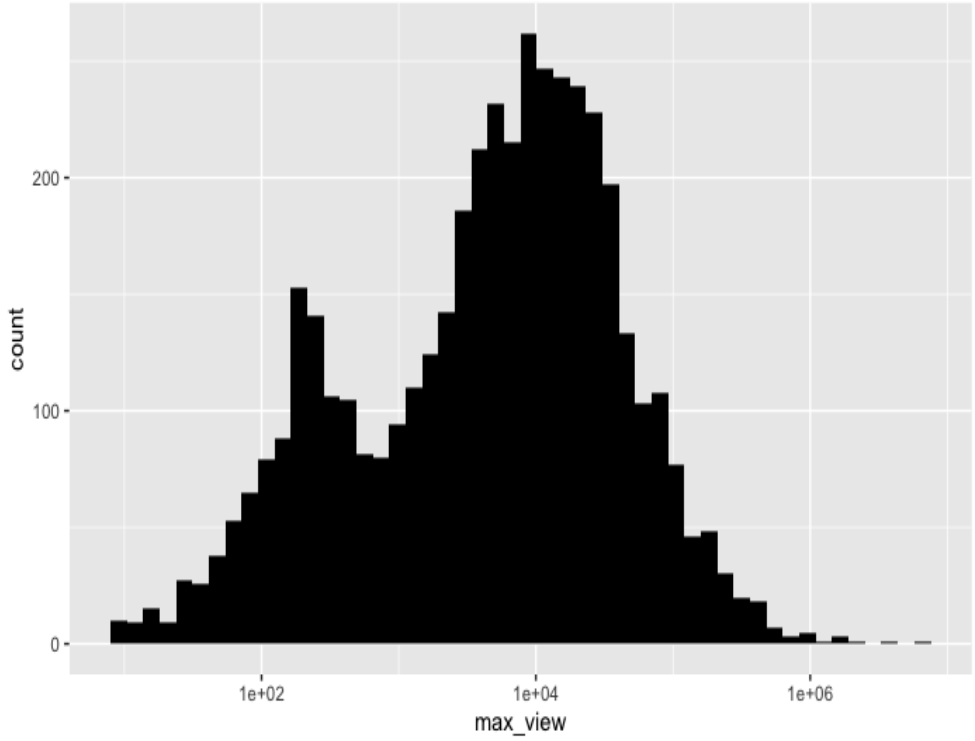


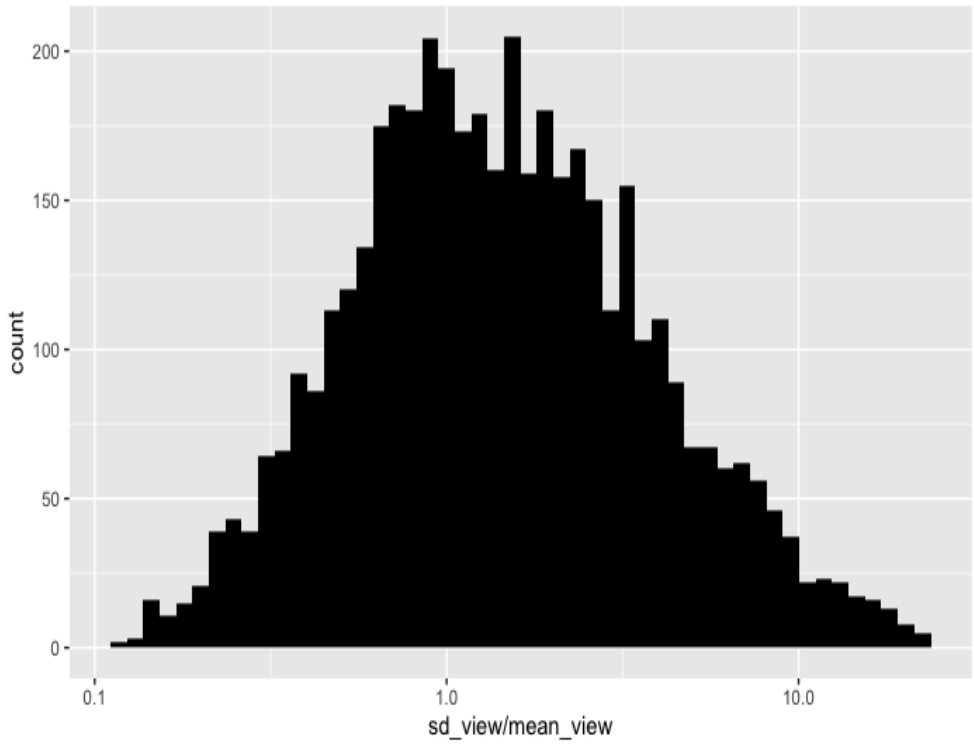
An extraction function is built to get basic set of parameters: Mean, Standard Deviation, Amplitude, and the slope of Naive Linear fit.

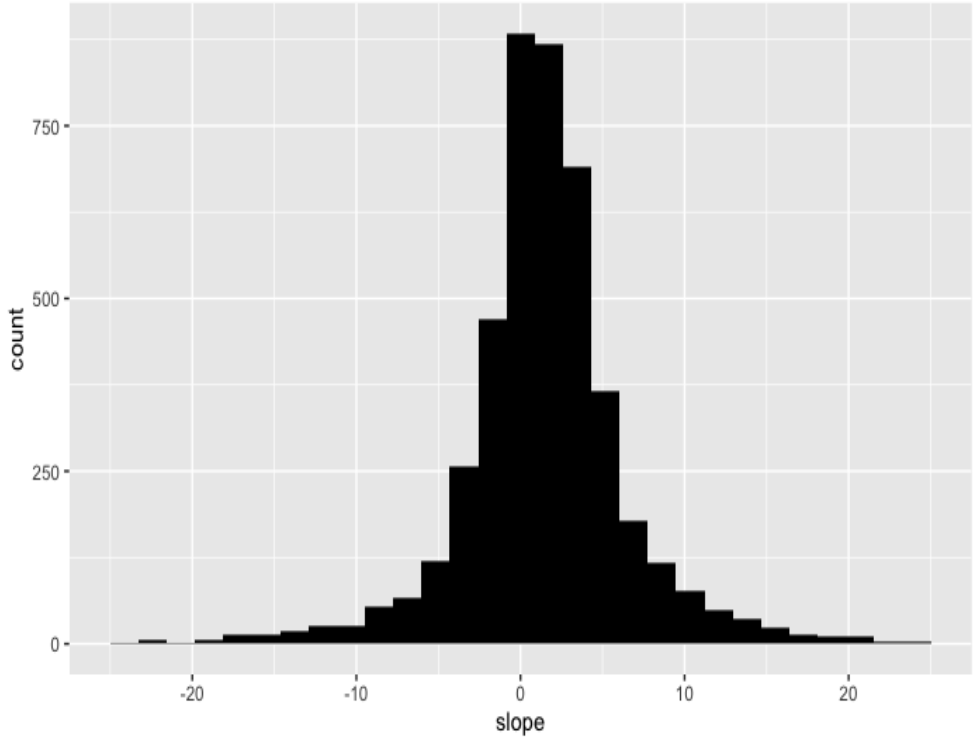
To explore the parameter space built, histograms are plotted as below:


300
200
100
0
count
1e+01
1e+03
1e+05
mean_view

.5 .5


200
100
0
count
1e+02
1e+04
1e+06
max_view


200
150
100
50
0
count
0.1
1.0
10.0
sd_view/mean_view


750
500
250
0
count
-20
-10
0
10
20
slope

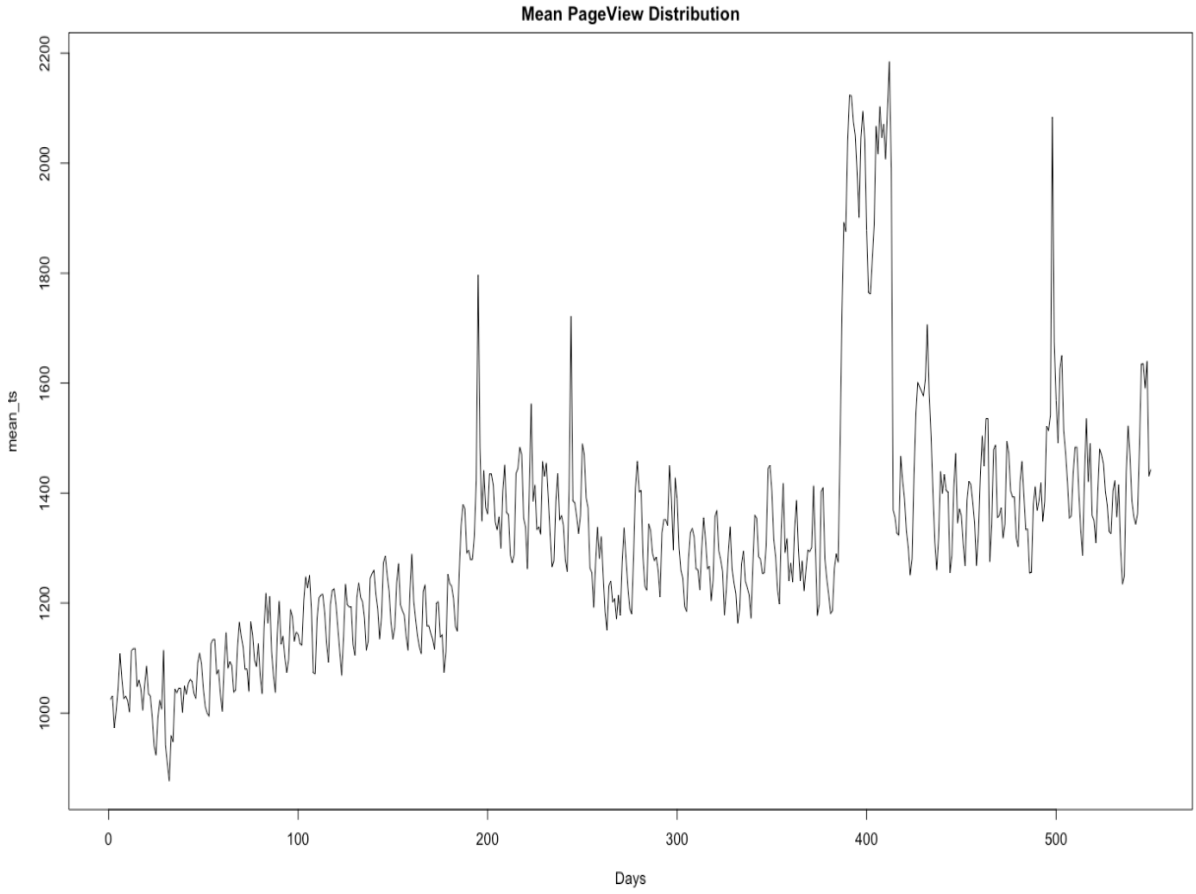
We found:

The distribution of average views is clearly multimodal, with peaks around 10 and 200-300 views. Something similar is true for the number of maximum views, although here the first peak (around 200) is narrow. The second peak is centred above 10,000.

The normal distribution of standard deviations is leaned toward higher values with larger numbers of spikes or stronger variability trends.

The slope curve is reasonably symmetric and centred above zero.

The distrbution of mean pageviews for all pages is shown below. This can tell us that the data is not stationary over time.



ACF for the mean pageviews data was plotted and We can clearly see that the pageviews are strongly correlated with their previous values.

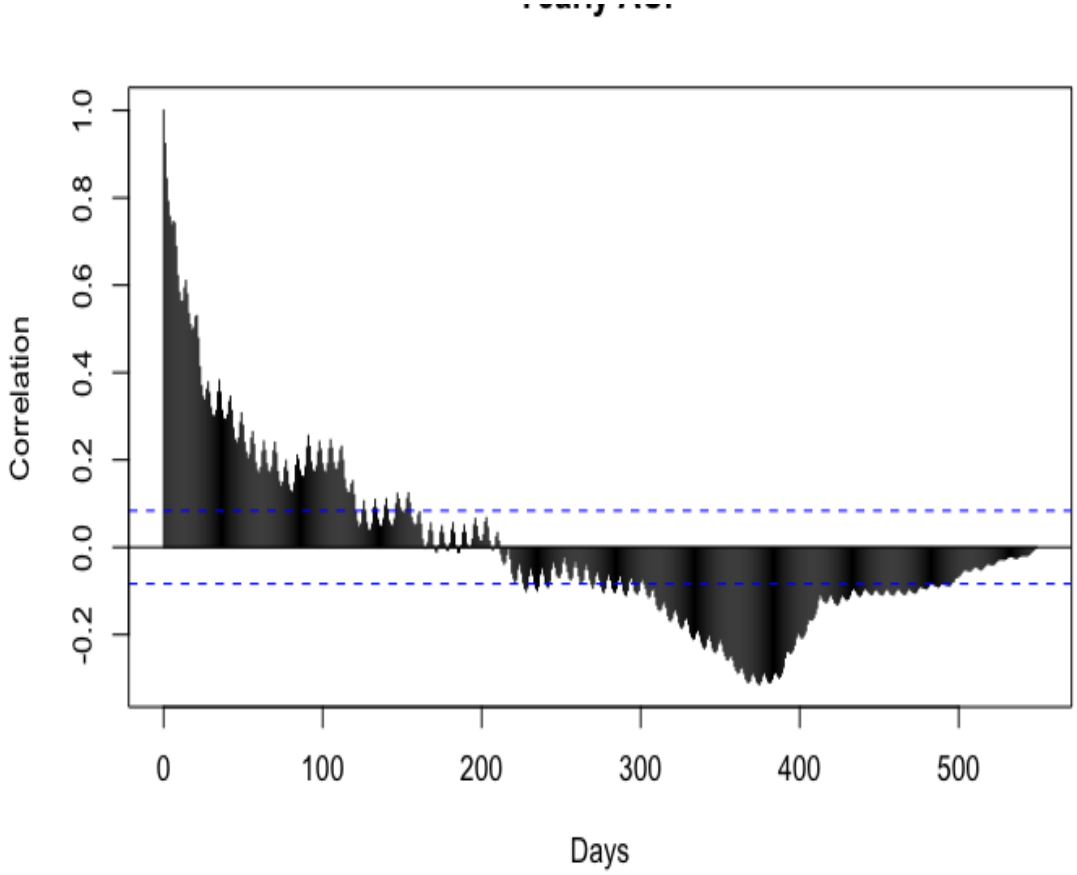


Figure 1: **Yearly_ACF**

Partial ACF on the complete timestamps as seen from below figure which shows there is a good correlation of partial residuals of the lagged values in the Yearly manner and it is not periodic as the peaks do not occur in timely fashion.

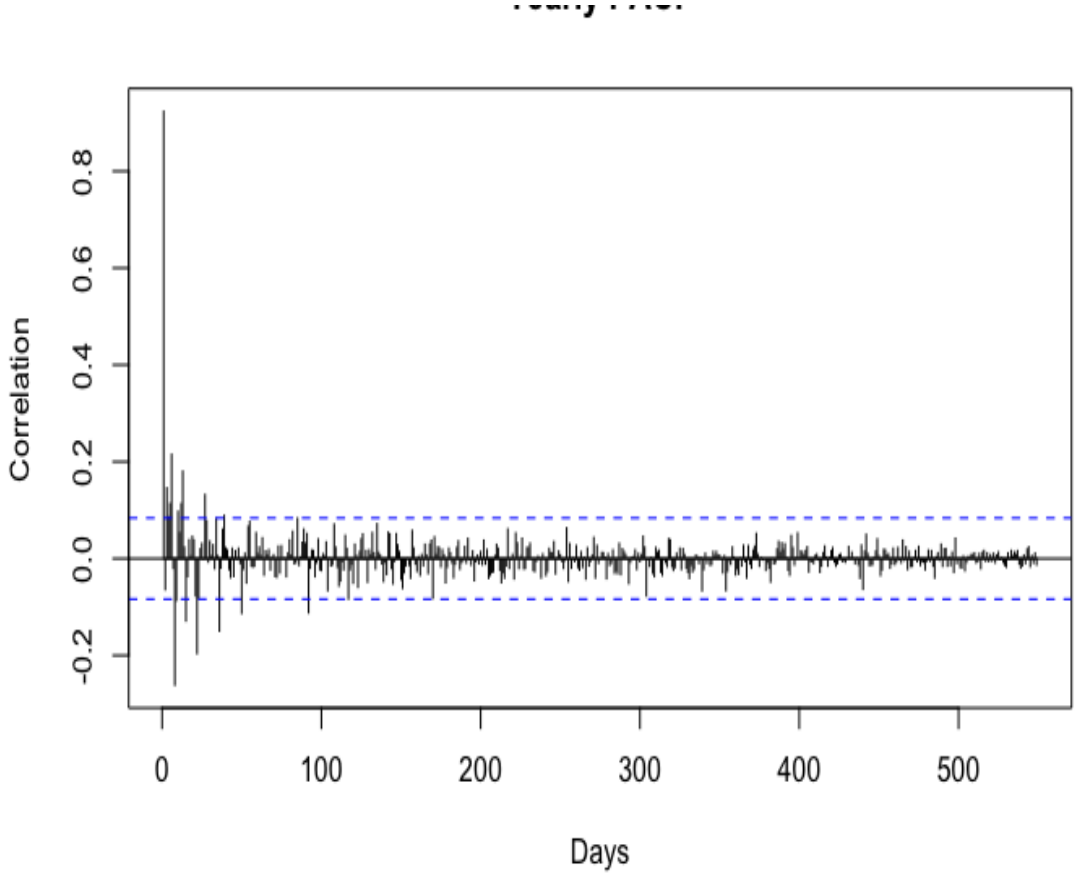


Figure 2: **Yearly_PACF**

Forecasting

For a sample of 145k articles most likely we will have to rely on an automatic mechanism to make our predictions.Therefore, our forecasting method will have to perform robustly for a range of different time series variabilities.

Our forecast period is 2 months. Simulated this period and assess our prediction accuracy by keeping a hold-out sample of the last 60 days. After making the predictions we compare the actual view counts to the forecasted ones.

A popular approach in time series forecasting is to use an autoregressive integrated moving average model; In short ARIMA model implemented using auto.arima function in R.

Using the insights, we can implement the frequency when turning our view counts

into a time series object (using the ts function). Note, that we also perform data cleaning and handle outliers using the tsclean tool. We wrap the modeling and visualising process into a customized function for predictions.

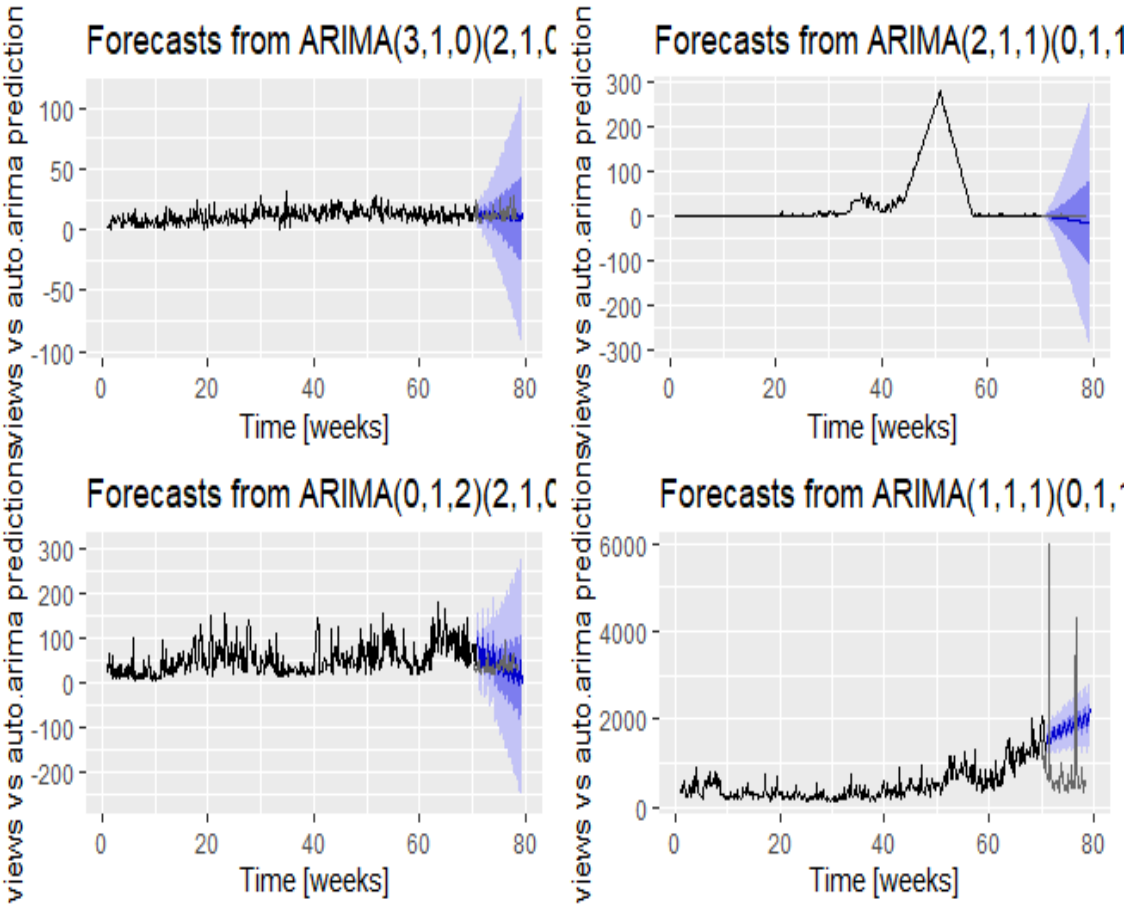


Figure 3: **Yearly_PACF**

Given that it's a fully automatic forecast the auto.arima tool performs well and provides us with a useful predictions of different pages.

References

https://www.kaggle.com/c/web-traffic-time-series-forecasting

https://towardsdatascience.com/forecasting-with-web-traffic-data-6681ff148df0

https://medium.com/swlh/wikipedia-web-traffic-time-series-forecasting-part-1-e43734adca3d