

Web Traffic Analysis - Predicting Blog Post Performance

Christian Frank (#473088)
February 26, 2021



Wirtschaftsinformatik: Big Data & Data Science
Dr. Marcel Graus
FOM - Hochschule für Oekonomie & Management
WS 2020

This paper will analyze blog post performance with data taken by Plausible website analytics, enhance it with categorization and sentiment analysis, and aim to make predictions on future blog post performance.



This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/> or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.

Contents

List of Figures	III
List of Code Examples	IV
List of Abbreviations	V
1 Web Traffic Analysis	5
1.1 Web Analytics	5
1.2 Social Media and Loneliness	5
1.3 Gender-neutral Pronouns	6
1.4 Cultural Bias	6
2 Data Sources and Research Methods	8
2.1 The Blog	8
2.2 Plausible	8
2.3 Count	9
2.4 Web Traffic KPIs	9
2.5 Statistical Methods	10
2.6 Sentiment Analysis	11
3 Data Exploration	12
3.1 Overall Access	12
3.2 Access by Country	12
3.3 Access by Operating System	13
3.4 Access by Referral	14
3.5 Access by Source	15
3.6 Categories	15
3.7 Blog Post Sentiment	17
3.8 Data Source	18
4 Data Analysis	19
4.1 Blog Post Data Overview	19
4.2 Data Wrangling	19
4.3 Blog Post Category Correlation	20
4.4 Blog Post Sentiment Correlation	21
4.5 Discarded graphs	23
4.6 Results	23
4.7 Outlook	23
4.8 Data Source	24
5 Summary	25
References	26

List of Figures

1	Plausible Summary	9
2	Overall Access	12
3	Access by Country	13
4	Access by Operating System	13
5	Access by Browser	14
6	Access by Referral	14
7	Access by Source	15
8	Raw Blog Post Data	19
9	Category Analysis	21
10	Sentiment Analysis - Climate	22
11	Sentiment Analysis - Covid-19	22

List of Code Examples

1	Vader Example	17
2	Vader Sentiment Compound	20

List of Abbreviations

CCPA	California Consumer Privacy Act
CTR	Click-Through-Rate
GDPR	General Data Protection Regulation
KPI	Key Performance Indicator
NLP	Natural Language Processing
PECR	Privacy and Electronic Communications Regulations
RSS	Really Simple Syndication
SMM	Social Media Marketing
SQL	Structured Query Language
URL	Uniform Resource Locator
VADER	Valence Aware Dictionary and sEntiment Reasoner

1 Web Traffic Analysis

1.1 Web Analytics

In this paper, I will look at my blog's traffic data, analyze that data, and then lay the foundation to predict a blog post's performance based on its content.

Web analytics is the domain of the big search engines, and Google Analytics and Google AdSense are the market leaders, followed by Microsoft Bing Web Analytics. In all of Social Media and Social Media Marketing, web analytics plays a crucial role in evaluating a website's performance and forms the basis of automated advertising placement.

Advertising, as much as we might dislike it, pays for most of the content we consume.

Page views, bounce rate, and unique visitors are key metrics to evaluate a website and the currency that fuels the internet. Every marketer or web site owner will use these metrics to analyze performance and identify areas for growth; many tools for analysis have become available in the last couple of years, some of them open-source, some closed-source.

Generally speaking, more traffic can potentially lead to more business opportunities. It is mandatory for a commercial website to monitor its web analytics data daily and act immediately on any anomaly.

However, a blog does not necessarily have commercial interests and might be an outlet for personal interests or interactions. Why would we want to look at web analytics anyway?

1.2 Social Media and Loneliness

In the current COVID-19 pandemic, social distancing is a key element in containing the virus's spread. Social distancing over a long period of time can increase loneliness and significantly affect people's health negatively, according to a recent study conducted by the American Psychological Association.¹

In the study, the researchers formulate the hypothesis that an increase in perceived support from others can offset loneliness during the required isolation.

One element to offset the effects of loneliness is increased interaction on social media and virtual meetings with video. Social media interaction includes reading blogs - the

¹See *Luchetti, M. (2020): The trajectory of loneliness in response to COVID-19. [8]*

more engaging a blog post is, the more chances it has to reach people to whom it will be entertaining or otherwise beneficial.

Thus this analysis aims to get an answer for my blog on the question "On which subject(s) should I post to increase my reach?". We will be using visualization and correlation as the primary means of analysis to increase the value of the blog for others.

1.3 Gender-neutral Pronouns

As we move towards a more inclusive and gender-fluid society, it's time to rethink the usage of gendered pronouns in scientific texts. Two well-known professors from UCLA, Abigail C. Saguy and Juliet A. Williams, argue that it makes a lot of sense to use singular they/them instead: "The universal singular they is inclusive of people who identify as male, female or nonbinary."² Throughout this paper, I'll attempt to follow their suggestion and invite my readers to do the same in future articles, and support an inclusive approach through gender-neutral language. Thank you!

1.4 Cultural Bias

As we start to rely on data more and more to make business decisions, train our machine learning models and make predictions, we need to pay more attention to the cultural bias in the data that we use.³ The study quoted in this article makes a case for data sets that respect the context they were created in and their creators' privacy rights.⁴ As a side note, it also points out that data curated with more respect will also most likely be more expensive.

The content in my blog and the people who read it are influenced by the fact that I am White and male. This bias affects both source (content) and destination (analysis) and thus does not distort the results.

Other than that is the data in use in this paper utterly free from cultural bias.

I have collected it for myself, and it is based on my content. The data also fully respects the blog readers' privacy as I do not collect any personally-identifiable data points; I assume that it would satisfy the study's quest for purposefully curated data.

²Saguy, A. (2020): Why We Should All Use They/Them Pronouns. [12]

³See Johnson, K. (2020): AI research survey finds machine learning needs a culture change. [7]

⁴See Paullada, A. (2020): Data and its (dis)contents: A survey of data-set development and use in machine learning research. [9]

However, the data is not free from personal bias - as you follow along, please feel free to interpret the data in different ways and draw your own conclusions; I will include links to all the data throughout the text.

2 Data Sources and Research Methods

2.1 The Blog

The blog in question is my personal web log that I started a couple of years ago, initially on the now-defunct Google+ platform. It's now on a hosted WordPress instance provided by a local service provider running on a web server farm in Strasbourg. A move from Google+ to Blogger instead of WordPress might have been the more comfortable choice and allowed for more consistency; however, I made a clean break between the hosting platforms. The original Google+ data is lost. I maintain some Google+ connections on Diaspora*, though.

I use the blog to complement my other social media whenever I feel the need to express myself in slightly longer texts. I try to post at least once per week and cross-post the blog entries to my other channels, private and professional; for small essays, I use Medium to publish them.

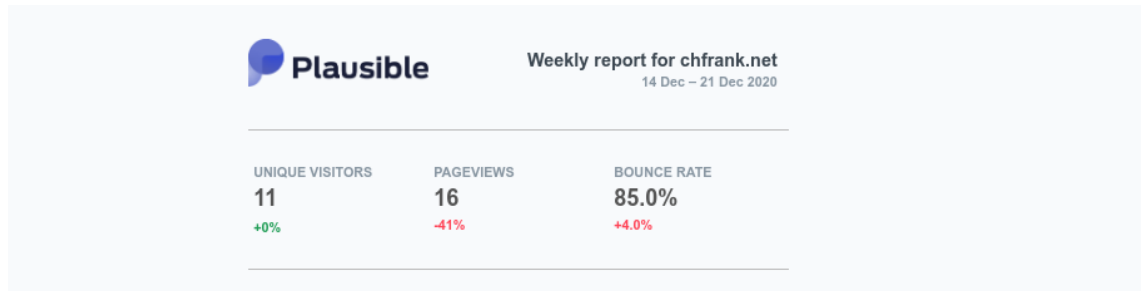
Although I do cross-post, I never duplicate entries, and thus there is no dilution of data; the data from every post is unique, and all content is only published once.

During the current and past COVID-19 lockdowns, I wrote a running week-by-week commentary, affecting the traffic analysis outcome, as we'll see later.

2.2 Plausible

The data I'll be using was collected using GDPR-compliant web analytics by Plausible, a lightweight and open-source website analytics tool. Plausible provides an easy-to-use webhook to integrate its analytics with almost any kind of website framework; it is ad-free and offered as Software-as-a-Service with a subscription model.

Plausible does not use any tracking cookies and claims to be fully compliant with GDPR, CCPA, and PECR. In addition to the web analytics data, I will complement Plausible's data with additional data from the blog.

Figure 1: Plausible Summary

In a previous paper, I have covered Plausible in more detail and compared it with other web analytics tools⁵; I will not go any further into details of the tool itself in this paper.

In case you're interested, you can find up-to-date raw web traffic data from the blog here: [Plausible Analytics](#).

2.3 Count

To analyze and visualize the data, I'll be using data notebooks from Count. Like Jupyter notebooks that combine (Python) Code and Text, the data notebooks combine (SQL) Data and Text in a pretty ingenious way. Data notebooks support data-driven decision making, and they are currently in open beta.⁶

2.4 Web Traffic KPIs

In marketing categories, a blog belongs to inbound marketing, as it tries to offer engaging content and create value for the visitors but does not reach out by itself.

Unlike outbound email campaigns, for example, that ask the visitor to view a particular website, a blog relies on its content and the willingness of the visitor to actively choose the site for a visit, for example, by being pointed to a post from a Google search result or a tweet.

In WordPress, there is an option to subscribe to a blog to get a notification on new posts; WordPress also provides the ability to subscribe to a blog's RSS feed. Even though, as this requires active user interaction and interest, blogs are considered an inbound channel.

⁵See *Frank, C. (2020): Usefulness of open-source tools for web analytics in E-Marketing.*[4]

⁶See *Count.co (2020): About Count.*[2]

In a recent paper on inbound marketing, Yvonne Romes identifies a couple of important KPIs for inbound marketing, a couple of which I will summarize here based on her paper:⁷

- Page Views
- Bounce Rate
- Visit Duration
- Unique Visitors

Page Views is the number of clicks a specific page has received; on a blog, more page views indicate more engaging content.

Bounce Rate describes the rate of users that leave the site without selecting another link; a high bounce rate can indicate a lack of engaging or interesting content.

Visit Duration is the amount of time a unique visitor spends on the website; for a blog that mainly offers content to read, a longer duration most likely indicates higher engagement.

A unique visitor is a visitor that can be differentiated from another visitor. Unlike many other platforms, Plausible does not use tracking cookies to identify individual visitors but relies on publicly available information, such as an IP address, to differentiate them. Even though the metric is less accurate with Plausible than with other platforms, it's still an important metric, and as before, on a blog, more visitors usually indicate higher engagement.

In this paper, these are the four metrics that I will focus on.

2.5 Statistical Methods

I will base the analysis on the excellent work and great documentation available at Towards Data Science, a platform on Medium to exchange ideas and to expand the understanding of data science.⁸

Vital elements in data science are statistics and linear algebra.

"From an academic perspective, understanding linear algebra is paramount to having a strong knowledge of specific topics within computer vision and deep learning."⁹

⁷See *Romes, Y. (2020): 10 Inbound KPIs, die jetzt auch Personaler kennen sollten.*[11]

⁸See *TDS Editors (2018): About Towards Data Science.*[14]

⁹*Alake, R. (2020): 6 Questions Asked By Machine Learning Enthusiasts.*[1]

An excellent reference for using linear algebra in data science and machine learning is the book "Mathematics for Machine Learning" by Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong.¹⁰

This paper will stay relatively simple and only look at two-dimensional data to identify possible correlations between measurements. I am acutely aware of the fact that "correlation does not imply causation"¹¹ and will make sure that our findings match real-world scenarios and experiences.

Statistics is the other bedrock of data science. In this paper, I will measure central tendency through arithmetic mean, median, and mode of our data.¹²

The actual data set is small, as the blog does not have a lot of traffic, so any method that relies on a high number of data points will not work. For the analysis and predictions of future performance, I will depend mainly on visualization techniques, tables, and verbal interpretation of the data.

2.6 Sentiment Analysis

The blog already has categories for its content, so there's no need to extract content information from the posts to categorize them manually. There is, however, no information in the blog data in regards to the sentiment of the post.

Especially concerning the current pandemic, I think it can make a lot of difference if a blog post is either upbeat or somewhat fatalistic, so I will attempt to identify the individual entries' sentiment.

To do this, I will use sentiment analysis. Sentiment analysis is a technique from the field of Natural Language Processing and fits into Contextual Semantic Search.

Sentiment analysis aims to identify a text message's sentiment as either positive, negative, or neutral.¹³ With the categories already present, there's no need to go deeper and add intent or context to the analysis; it will suffice to enrich the post data with sentiment information.

There are several ready-made sentiment analysis libraries available for both R and Python. Since I am a bit more familiar with Python and functional programming than with R, I will concentrate on using Python in the next chapters.

¹⁰See *Deisenroth, M.P. (2020): Mathematics for Machine Learning.*[3]

¹¹See *Singh, S. (2020): Why correlation does not imply causation.*[13]

¹²See *Yildirim, S. (2020): 10 Must-Know Statistical Concepts for Data Scientists.*[15]

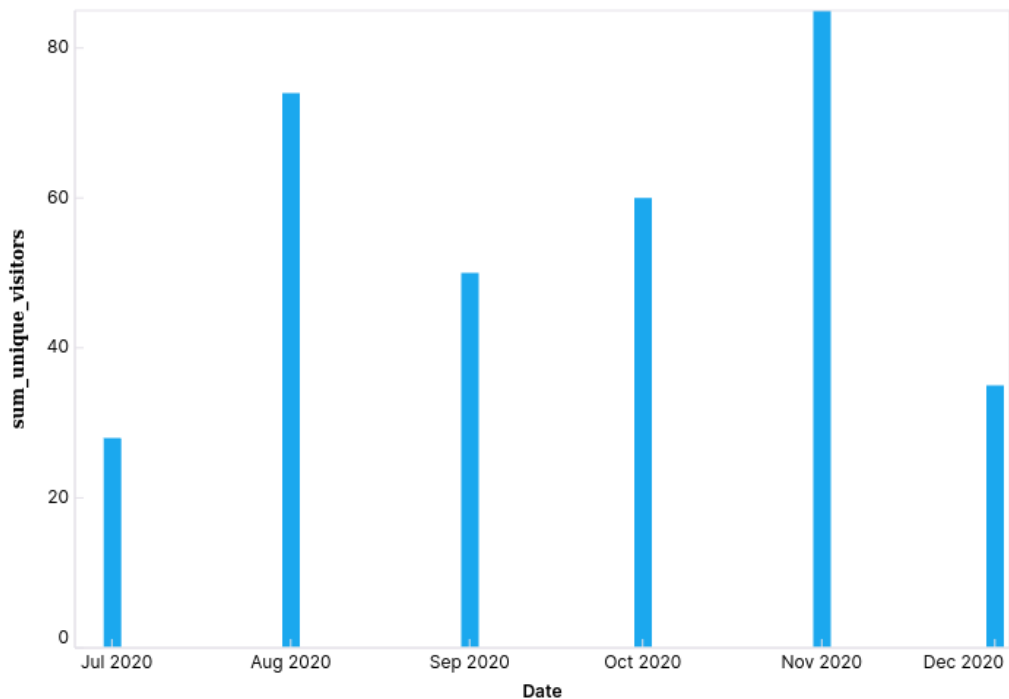
¹³See *Gupta, S. (2018): Sentiment Analysis: Concept, Analysis and Applications.*[5]

3 Data Exploration

3.1 Overall Access

I enabled Plausible web analytics on the blog only six months ago. Let's start the analysis with an overall view of people accessing the blog during the last months:

Figure 2: Overall Access



Looking at the number of unique visitors, we can see a spike in August, most likely for the coverage of the preparation for the latest global climate strike and in November, when the second lockdown started. I would assume that December's figures are probably a bit too low as I have collected the data before the end of the month.

3.2 Access by Country

Where do all the visitors of the blog come from? Let's first have a look at the distribution by country:

Figure 3: Access by Country

Country	sum_unique_visitors
Germany	215
United States	35
United Kingdom	20
Switzerland	16
Canada	7

2 columns, 26 rows

Not surprisingly, most visitors access the blog from Germany, my home country, immediately followed by the United States and the United Kingdom. I write the blog in English so that visitor distribution makes total sense. Looking at the access number from the German-speaking countries alone, I might get a higher number of visitors if the blog was in German. Still, I would lose all visitors from countries other than Germany, Austria, and Switzerland. For now, I will keep writing blog posts in English.

3.3 Access by Operating System

The blog's visitors not only originate from countries, but also from operating systems:

Figure 4: Access by Operating System

Operating_system	sum_visitors
Windows	93
Android	91
iOS	64
GNU/Linux	52
Mac	24

2 columns, 7 rows

and browsers:

Figure 5: Access by Browser

Browser	sum_unique_visitors
Chrome	202
Safari	61
Firefox	41
Microsoft Edge	11
Samsung Browser	5

2 columns, 9 rows

Chrome is still the undisputed market leader in browsing, and I know that the WordPress theme in use for the blog displays well on Chrome.

Looking at the operating system, we can deduct that there is about equal access from mobile devices (Android + iOS) and Desktop devices (Windows + Linux + Mac). I will need to make sure that the WordPress theme I use is well suited for mobile devices.

3.4 Access by Referral

In addition to the number of visitors, there are other interesting KPIs, especially visit duration and bounce rate. Let's look at these KPIs first by referral:

Figure 6: Access by Referral

Source	sum_unique_visitors	sum_visit_duration	sum_bounce_rate
Direct / None	290	85	79
Twitter	28	52	89
LinkedIn	28	10	83
Facebook	5	0	80

4 columns, 4 rows

We can see that direct access to the blog has the highest number of unique visitors, the most prolonged visit duration, and the lowest bounce rate. This is not surprising; people who actively seek out the blog most likely have the highest interest in its content.

My cross-posts to Twitter, LinkedIn, and Facebook yield a lot fewer visits, with Facebook being the outlier, generating almost no visitors and no visit duration. In the future, I

might review cross-posting to Facebook and stop the practice if the numbers do not improve.

I did not expect the cross-posts to LinkedIn to yield any visits and am pleasantly surprised; without looking at the data, I would have assumed that access from Facebook would be much higher than from LinkedIn.

3.5 Access by Source

Second, let's look at these KPIs by source:

Figure 7: Access by Source

Source	sum_unique_visitors ↓	sum_visit_duration	sum_bounce_rate
Direct	189	97	75
dlvr.it	115	63	86
Google	36	2	92
DuckDuckGo	6	16	83
Bing	5	101	80

4 columns, 9 rows

As expected, we can see the same picture here, direct access to the blog has the highest number of unique visitors, the most prolonged visit duration, and the lowest bounce rate.

However, access from the search engines is a lot less than from cross-posting through dlvr.it, which I did not expect. To improve this, I might need to review the Google Search Console settings and pay closer attention to tags and keywords.

Just by looking at the data, I was already able to gain some insights into the visitors' behavior and identify some possible improvements for the future.

3.6 Categories

As a next data point, we'll want to look at the content of the blog posts. First, we'll have a look at the list of categories taken from the blog:

- buddhism
- climateaction
- climateemergency

- climatestrike
- cloudnative
- electric-cars
- fridaysforfuture
- kubernetes
- leavonoonebehind
- movies
- music
- politics
- rancher
- socialmedia
- travel
- university
- youtube

I have omitted two, cologne and general, from the blog's categories, as they are present on all posts. These categories are at the same time hashtags for Diaspora*, another social network that I cross-post to; Diaspora* is part of the Fediverse and does not have the same reach as the more popular commercial networks and thus did not appear in the statistics above. Interactions on Diaspora* are far fewer than on the other networks, but they tend to be more thoughtful and exciting.

To prepare for the analysis, I have decided to reduce the number of categories down to four, based on the distribution of the posts in the WordPress dashboard:

- Climate and politics (1)
- Cloud (2)
- Social Media and culture (3)
- Covid-19 (4)

Climate and politics will include all posts regarding our major crisis, the climate emergency, and the climate justice movement.

Cloud will include all posts in regards to cloud-native technology.

Social Media and culture will include posts on YouTube, music, and movies.

Covid-19 will cover all posts related to the other crisis, the current global pandemic.

If there's a blog post without a category, I will encode the missing value with 0.

These four content categories should help us in the next chapter analyzing the individual blog post performance and possibly arrive at more insights and finally, at some predictions.

3.7 Blog Post Sentiment

The final data point to explore is the sentiment of the post.

From the many available libraries for Python I've chosen VADER (Valence Aware Dictionary and sEntiment Reasoner) based on the description of the algorithm and its results in an analysis and ranking of texts by H.P. Lovecraft.¹⁴

Vader is "specifically attuned to analyzing social media posts"¹⁵ and thus the perfect choice to analyze blog posts; Vader can also be used just with a few lines of code, making it an ideal library for casual analysis such as this one.

With an extensive data set, there would have been the option to create and train a more specific NLP machine learning model to identify the post's sentiment. For a small data set like the blog, I felt it would be more prudent to use a ready-made library.

Let's look at an example invocation:

Example 1: Vader Example

```
from vaderSentiment.vaderSentiment
    import SentimentIntensityAnalyzer

blogPost = "My first ever entirely virtual Christmas and
    it went quite well! And so will New Years, I hope."

analyzer = SentimentIntensityAnalyzer()

print(analyzer.polarity_scores(blogPost))
```

¹⁴See Pocs, M. (2018): Lovecraft with Natural Language Processing.[10]

¹⁵Hutto, C. (2018): VADER Sentiment Analysis.[6]

The text in the example is the first sentence of the most recent post and Vader will show the following sentiment rating:

```
{'neu': 0.74, 'compound': 0.68, 'neg': 0.0, 'pos': 0.26}
```

We know that compound is the normalized, weighted composite score of the other three values (positive, negative, and neutral) from the documentation. I'll be using the sentiment compound values to enrich our data.¹⁶ The compound will have values between 1 (most positive) and -1 (most negative), and there won't be any missing values; a value of 0 will indicate a completely neutral text.

3.8 Data Source

For the data analysis and visualization in this chapter, I used a Count data notebook; you can find the raw data and all charts from this chapter here: [WTA Chapter 3](#)

¹⁶See *Hutto, C. (2018): VADER Sentiment Analysis*. [6]

4 Data Analysis

4.1 Blog Post Data Overview

Before enriching the data for analysis, let's have a look at the raw web analytics data and KPIs for the various posts:

Figure 8: Raw Blog Post Data

Page_Url	sum_unique_visitors	sum_page_views ↓	sum_bounce_rate
/wordpress/2020/08/16/travel/	21	32	88
/wordpress/2020/10/10/danni-bleibt/	13	28	94
/wordpress/2020/11/17/the-greens/	14	23	75
/wordpress/2020/08/03/anti-sexist-...	14	22	80
/wordpress/2020/07/26/perseveran...	12	19	75

4 columns, 33 rows

I have excluded all top-level access, such as from "/wordpress/". Visits by accessing the blog itself and then scrolling down can unfortunately not be attributed to a specific post and thus not be enriched with category or sentiment data.

4.2 Data Wrangling

Adding category information from the blog to the raw blog post data from above was relatively easy and a straight-forward exercise with vi.

Text from blog posts on a WordPress blog, however, cannot easily be extracted. WordPress offers an export function, but that will result in an archive in XML format, which is not suitable for text analysis.

After several tests, I found that the best way to extract the text from the actual posts was to use a text-based browser, such as Lynx, Links, or Elinks; I chose Elinks for the task.

We already have the post URL in the blog post data above, so extracting was done with a simple shell script:

```
eLinks https://.../2020/08/16/travel/ -dump > travel
```

...

As a the next step I created a small Python script to analyze the text and return the compound sentiment value:

Example 2: Vader Sentiment Compound

```
from vaderSentiment.vaderSentiment
    import SentimentIntensityAnalyzer

analyzer = SentimentIntensityAnalyzer()

filenames = [ "anti-sexist-social-club",
               "biden-harris",
               ...
               "xfce" ]

for filename in filenames:
    fd = open(filename)
    blogPost = fd.read()
    sentiment = analyzer.polarity_scores(blogPost)
    print(filename + " " +
          str(int(sentiment['compound'] * 10)))
```

This script will return the blog post's compound sentiment value from the VADER analysis, one row per post.

```
anti-sexist-social-club 9
```

```
biden-harris 6
```

```
...
```

```
xfce 9
```

Adding the sentiment to the raw data again was a task for vi and sed.

As a result, we now have a data set of all blog posts, the associated KPI values, their categories, and their sentiment.

4.3 Blog Post Category Correlation

Now that we've enriched the blog post data with category and sentiment information let's look at our KPIs by category. As a reminder, the encoding was as follows:

- Climate and politics: 1
- Cloud: 2
- Social Media and culture: 3
- Covid-19: 4

Figure 9: Category Analysis

Category ↑	sum_unique_visitors	sum_page_views	avg_bounce_rate
1	136	243	83.667
2	3	4	100
3	29	40	59.25
4	66	107	85.429

4 columns, 4 rows

The top-performing category with the highest number of unique visitors, the most page views, and the lowest bounce rate is Climate and Politics; the second-best performing category in all KPIs is Covid-19. The two other categories fall way behind - my posts regarding either Cloud or Social Media don't seem to be interesting to my readers.

4.4 Blog Post Sentiment Correlation

For the two best performing categories, Climate and Politics, and Covid-19, let's have a look at the KPIs broken down by sentiment, Climate first.

The sentiment is a compound value from the NLP analysis with ranges between -10 and 10, from a total negative to a full positive sentiment.

Figure 10: Sentiment Analysis - Climate

Sentiment ↓	sum_unique_visitors	sum_page_views	avg_bounce_rate
9	20	29	90
7	8	11	100
6	14	23	70
3	20	35	51
-2	14	25	70
-3	17	34	92.667
-5	13	28	94
-6	4	7	100
-7	7	9	100
-8	8	20	94.5
-9	11	22	94.333

4 columns, 11 rows

If we compare the number of Unique Visitors for a Climate post with positive and negative sentiment, we cannot observe a significantly large difference (62 vs. 74). When we look at the number of Page Views (98 vs. 145), we can see a slight difference, favoring climate action posts with a more negative sentiment. There seems to be no influence from the posts' sentiment on the average Bounce Rate.

Now let's do the same analysis for posts regarding Covid-19 and the current pandemic.

Figure 11: Sentiment Analysis - Covid-19

Sentiment ↓	sum_unique_visitors	sum_page_views	avg_bounce_rate
9	13	32	84
8	4	7	83
6	28	40	94
0	9	10	86
-4	12	18	73

4 columns, 5 rows

Now there's a visible difference - both Unique Visitors (45 vs. 21) and Page Views (79 vs. 28) are significantly higher for blog posts on Covid-19 with a more positive sentiment.

Again, there seems to be no influence from the posts' sentiment on the average Bounce Rate.

4.5 Discarded graphs

There were a couple of graphs that I discarded during the analysis. I attempted to look at the blog post's KPIs over time or by publishing day, with no significant results. Equally uninteresting were my attempts to correlate data on the visitor's origins with the posts' performance - I had hope for insights such as "Visitors from Austria as more interested in Climate Change than visitors from Switzerland" but alas, the data would offer no such hints.

4.6 Results

But there was enough information in the data to make the exercise worthwhile and yield a couple of actionable results:

- Focus on Climate, Politics, and Covid-19
 - Focus on the dangers of climate change
 - Be upbeat about the outlook on Covid-19
- Improve Google Search settings
- Remove Facebook referrals
- Switch to the German language
- Improve access for mobile devices

4.7 Outlook

Coming back to the original question for the paper "On which subject(s) should I post to increase my reach?" we now have a definite answer.

According to the data, posting more content on Climate Change and Covid-19 should lead to more visitors' engagement.

I'll keep that in mind for the future and change the content that I post accordingly.

Cross-posting to Facebook is an already established automated procedure, so I'll keep doing that for now. Also, I like posting in English, so I'll continue this practice as well.

I will work on the other two recommendations to improve Google search and mobile usability.

4.8 Data Source

Again you can find the raw data and all charts from this chapter in the data notebook here: [WTA Chapter 4](#)

5 Summary

From the analysis, we can conclude that it does make sense to focus the posts on my blog on the two major crises of our time, the Climate Emergency and the COVID-19 pandemic.

The result is not entirely surprising, as these are the most talked-about issues on all media and are on everybody's mind all the time. However, having the assumptions being proved by data science through a thorough data analysis helps a lot and will guide me in further postings and the blog's development.

The power is in the data - even for a small blog (and thus a small data set), analyzing the data is a worthwhile thing to do and will lead to exciting and actionable results.

As we saw, Plausible is a unique open-source tool for web analytics. There are many other open-source tools available to support data science and analysis; there is also a big community around these tools and the subject of Big Data and Data Science.

In this paper, I was merely able to scratch the surface, but I hope I could provide you with at least some valuable insights and pointers to start with; all available raw data is in the two data notebooks and the sidebar of my blog.

Happy Analysis!

References

- [1] R. Alake. (2020) 6 questions asked by machine learning enthusiasts. [Access 2020-12-30]. [Online]. Available: <https://towardsdatascience.com/6-questions-asked-by-machine-learning-enthusiasts-d2931eabee>
- [2] Count.co. (2018) About count. [Access 2020-12-30]. [Online]. Available: <https://count.co/about>
- [3] M. P. Deisenroth, A. Faisal, and C. S. Ong. (2020) Mathematics for machine learning. [Access 2020-11-07]. [Online]. Available: <https://mml-book.com>
- [4] C. Frank. (2020) Usefulness of open-source tools for web analytics in emarketing. [Access 2020-09-07]. [Online]. Available: <https://storage.googleapis.com/bucket.chfrank.net/Open%20Source%20Web%20Analytics.pdf>
- [5] S. Gupta. (2018) Sentiment analysis: Concept, analysis and applications. [Access 2020-12-30]. [Online]. Available: <https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17>
- [6] C. Hutto. (2020) Vader sentiment analysis. [Access 2020-12-30]. [Online]. Available: <https://github.com/cjhutto/vaderSentiment/blob/master/README.rst>
- [7] K. Johnson. (2020) Ai research survey finds machine learning needs a culture change. [Access 2020-12-30]. [Online]. Available: <https://venturebeat.com/2020/12/26/ai-research-survey-finds-machine-learning-needs-a-culture-change/>
- [8] M. Luchetti. (2020) The trajectory of loneliness in response to covid-19. [Access 2020-12-30]. [Online]. Available: <https://doi.apa.org/fulltext/2020-42807-001.html>
- [9] A. Paullada. (2020) Data and its (dis)contents: A survey of dataset development and use in machine learning research. [Access 2020-12-30]. [Online]. Available: <https://arxiv.org/abs/2012.05345>
- [10] M. Pocs. (2020) Lovecraft with natural language processing. [Access 2020-12-30]. [Online]. Available: <https://towardsdatascience.com/lovecraft-with-natural-language-processing-part-1-rule-based-sentiment-analysis-5727e774e52>
- [11] Y. Romes. (2020) 10 inbound kpis, die jetzt auch personaler kennen sollten. [Access 2020-12-30]. [Online]. Available: <https://www.omt.de/inbound-marketing/10-inbound-kpis-die-jetzt-auch-personaler-kennen-sollten/>

- [12] A. Saguy and J. Williams. (2020) Why we should all use they/them pronouns. [Access 2020-05-20]. [Online]. Available: <https://blogs.scientificamerican.com/voices/why-we-should-all-use-they-them-pronouns/>
- [13] S. Singh. (2018) Why correlation does not imply causation? [Access 2020-12-30]. [Online]. Available: <https://towardsdatascience.com/why-correlation-does-not-imply-causation-5b99790df07e>
- [14] TDS Editors. (2018) About towards data science. [Access 2020-12-30]. [Online]. Available: <https://towardsdatascience.com/about-towards-data-science-d691af11cc2f>
- [15] S. Yildirim. (2020) 10 must-know statistical concepts for data scientists. [Access 2020-12-30]. [Online]. Available: <https://towardsdatascience.com/10-must-know-statistical-concepts-for-data-scientists-645619783c08>