1 Tabularazor News

This report covers the methods and the results of the data analysis performed on the website of Tabularazor Inc. (`https://news.tabularazor.org`). This newspaper is considered a large national news-paper, covering news on nearly everything. Tabularazor wants to guarantee the anonymity of their journalists and has therefore issued a meta-data analysis to examine what the (minimal) information provided on the website still might disclose about its employees. The analysis covered in this report is aimed at answering the following questions:

- Are there couples amongst the employees of Tabularazor Inc. and if so, are they still together?
- Did any of the employees of Tabularazor Inc. have a child during their employment?
- What is the number of expected holidays for employees of Tabularazor Inc.?

In this report, the structure of Tabularazor Inc.'s website is described first, followed by the description of the process of "web-scraping" using the website's domain. Analysing the extracted data, the research questions mentioned above are answered. It then summarizes the findings of the analysis in the conclusions section and, lastly, elaborates on the limitations of the performed data analysis.

2 Gathering Information

2.1 Analyzing the Website Structure

In order to perform a data analysis that generate results to answer the research questions, it is important to find suitable data. The website of Tabularazor Inc. is set up in a logical tree-like structure. The homepage (root) states all years in which articles have been published, ranging from 2012 to 2019. The names are listed as, e.g. *Articles in year 2012*, and can be accessed through their corresponding HTML $<a>$ tag element, linking the source page with the destination. The $<a>$ tag element contains an attribute *href*, which specifies the link's destination. The link destination for *Articles in year 2012* is then represented by for example *./2012.html*, which is an extension and when placed behind the root URL (`https://news.tabularazor.org`) the page *Articles in year 2012* can be accessed.

This tree like structure of the website is displayed in Figure 1, where each year page contains the links to the month pages of that year, and each month page contains the links to the articles published within this month, arranged by date. The linkages (or edges) are represented by the corresponding $<a>$ tag elements that can be retrieved from the node to move down a level. Article names are defined by their publishing date followed by the title of that article, for example: /01/01/2012/dolore-velit-modi-dolor-quaerat-dolore-dolore.html. The date mentioned at the beginning of an article is arranged in a day/month/year (DD/MM/YYYY) format. The article pages show its Title, the Author of the article, the publishing date and time, and the article text itself.

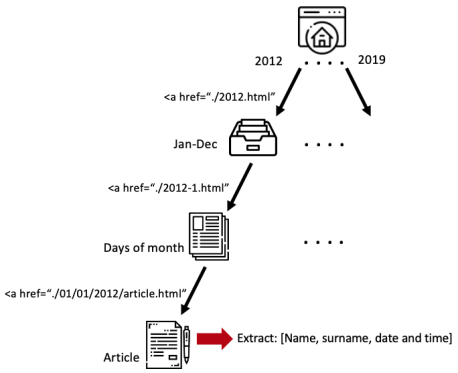

Figure 1: Example of accessing an article on the first of January 2012

2.2 Performing Web Scraping

Based on the structure of the website domain, a process of "web scraping" can be performed, which is the term used for extracting data from websites. In performing web scraping, one needs to find out what pages on the website contain information that is valuable, and with that, what data needs to be extracted from these pages that is needed to answer to the research question.

The data that is identified to be relevant for the analysis and, therefore, retrieved using web-scraping is: the *Name* of the author, the publishing *Date*, and publishing *Time* of each article. In order to do so, every single article page that can be identified within the domain of the Tabularazor inc. website is identified and the relevant data is retrieved. This task is performed by a Python based algorithm. Starting off, a request is made to access the homepage by going to the root URL (*https://news.tabularazor.org/*), after obtaining access all the <*a*> tag elements of the root are extracted and put together in a list of links to the years. In accessing the months per year, the root domain with the added *href* of that year is added to search for the month *hrefs* on that given page. Arriving at the month, all the <*a*> tag elements can again be extracted and put together in a list representing the months for that specific year (*href* of months 01/12). This will be done again for the days, and eventually for the articles themselves. Finally, when an article page is reached, the Name, *Date*, and *Time* are displayed in italics, as these elements have different semantic meaning than the article itself, they describe information about the creation of the article. These elements can be accessed within the HMTL <*i*> tag and the three elements that need to be retrieved are respectively stored in the class 'author', 'date', 'time'.

At the end of the web-scraping process, the retrieved elements are stored as strings in a data-frame and exported as comma-separated values (.csv). The format of the results are displayed in Table 1. As depicted in the Table, 328.360 articles in total are found with the first article being published on the 1st of January 2012 and the last on the 31st of December 2019. The resulting table and information are used for the data analysis.

Table 1: Result of scraping all articles of Tabularazor Inc.

Article number	Name	Date	Time
1	Jaye Shimek	2012-01-01	15:17
...	...	...	...
328.360	Lester Preiss	2019-12-31	17:57

An visualization of the retrieved data is portrayed in Figure 2, where the days on which an author has published at least one article are depicted by blue dots.

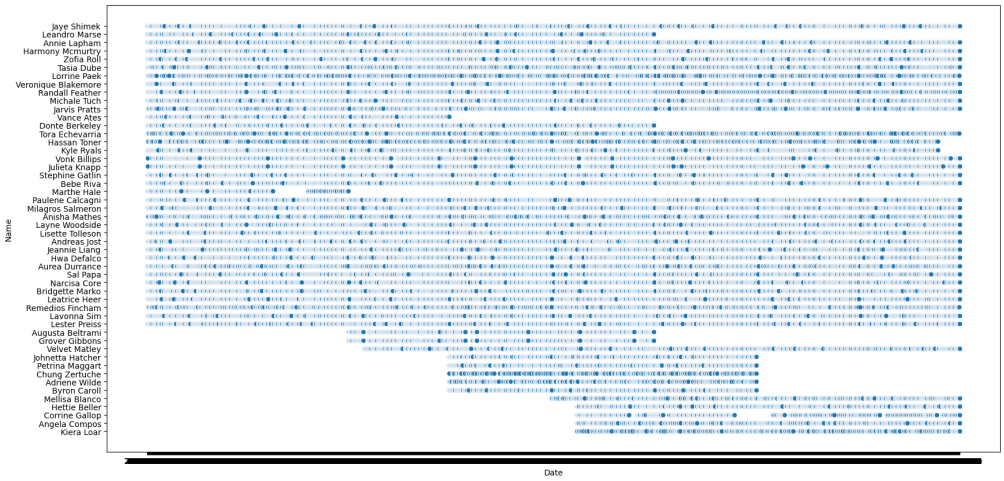

Figure 2: Author publications over all time

3 Analyzing the Data

For the data analysis the data set retrieved from scraping the website is used as a basis and can be further manipulated for specific research. The attributes per article from the data-set are displayed in Figure 1.

3.1 Are there couples amongst the employees and are they still together?

In checking for couples amongst employees, the hypotheses of couples spending their holidays together is tested. An author is identified to be on a holiday when that author did not publish any articles during a given week.

The correlation between authors having full weeks off is calculated, as this reduces chance of finding correlations between individual days off such as weekends or National Holidays.

Upon inspection of the provided data the highest correlations per year were retrieved, clearly indicating very high correlations between two couples over a number of years. These two potential couples are Julieta Knapp & Vonk Billips and Augusta Beltrami & Grover Gibbons. The correlation of these two potential couples over time is listed in Table 2

Table 2: Correlation of days off for the suggested couples

Couple	2012	2013	2014	2015	2016	2017	2018	2019
V. Billips & J. Knapp	0,92586	0,97433	0,95218	0,96901	0,94437	0,36848	0,15637	0,02995
G. Gibbons & A. Beltrami	0,00000	0,00000	0,91882	0,97084	0,90990	0,00000	0,00000	0,00000

In this table Vonk and Julieta show strong correlation from 2012 until 2016, indicating a high possibility of a relationship. Having a consecutive three weeks off at the same time for five times during this period. However, from 2016 on, their correlation decreases, possibly indicating the end of their relationship.
As visible in Figure 2, Grover and Augusta both did not start working at Tabularazor until 2014 and simultaneously left the company in 2017. Therefore, the correlation values of those years are marked as zero. During the years they worked at the company (2014, 2015 and 2016), their correlation is rather high, suggesting a high chance of a relationship based upon simultaneous holidays, sharing three weeks off at four different instances.

3.2 Did any of the employees have a child during their employment?

Analyzing the data to find whether someone had a child, focus is laid upon maternity leave. For European Member States, the minimum required time of maternity leave an employer should provide to its employees is 14 weeks. Therefore, the analysis checked for a consecutive number of at least 14 weeks in which an employee did not work but resumed working afterwards. This leaves out the employees that had a child and then did not return to their work, however, due to large uncertainty in this prediction we only focus on the group resuming work after expected maternity leave of minimally 14 weeks and a maximum of 20 weeks. Upon visual inspection of Figure 2, we see two large periods of absence for Marthe Hallen and Corrine Gallop.
As displayed in Figure 3, Marthe Hallen works an average of 5 days per week and is then absent for 18 weeks. When returning to her work, she continues to work an average of three days per week before she eventually stops working. For Corinne Gallop, a leave of 19 weeks is identified after which she returns to work an average of four days per week instead of the five days before her expected maternity leave. Both potentially started working less days per week to take care of their new-borns. Marthe eventually stopped working, possibly to focus on raising her child.

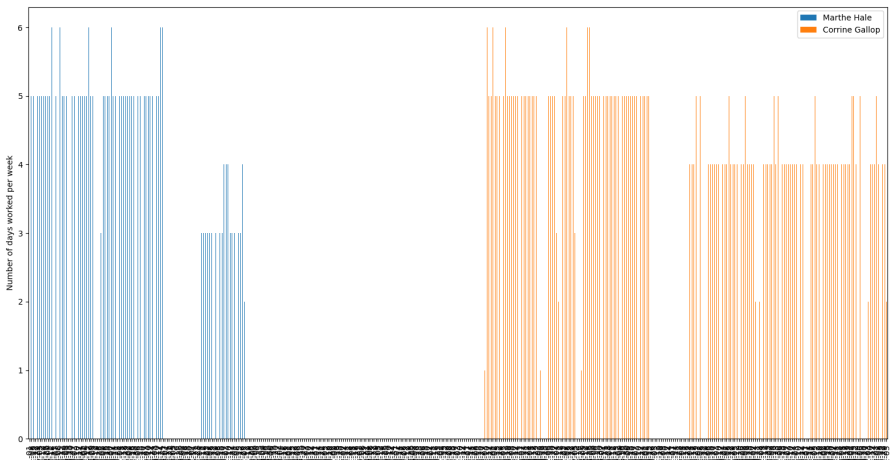

Figure 3: Potential Maternity Leave Absence and Decrease in Days Worked per Week

3.3 What is the number of expected holidays for employees of Tabularazor Inc.?

In estimating the number of holidays per year, we analyse the employees that have worked from the January 1st 2012 until December 31st 2019, to have the largest and most consistent data-set possible. This is done by only selecting employees that had less than 1350 days off in this period. A day on which an author did not publish an article is marked as a day off, however, it is important to note that days off also contain weekends and National Holidays (on a non-weekend day) for which an employee does not need to take a day off. In order to predict the number of holidays for an employee, the number of days per year on which an employee does not publish an article are added up. Then, the number of weekend days and National Holidays are subtracted from the resulting number. From 2012 until 2019, there were 835 weekend days and 51 additional National Holidays. The days off remaining were then visualized in Figure 2.

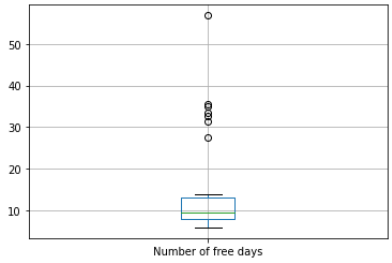

Figure 4: Amount of expected free days over all time for 30 full-timers

The box contained within the box-plot has a relatively small box containing the 2nd and 3rd Quartile, containing the median near 10 days off. Therefore, we assume that the number of holidays for a full-time employee at Tabularazor is 10 days.

4 Conclusions

This chapter provides a summarized answer to all research questions mentioned in the introduction. The expected couples are Julieta Knapp Vonk Bilips as well as Augusta Beltrami Grover Gibbons. The first are possibly no longer together and the latter do no longer work at the company. Marthe Hallen and Corrine Gallop are expected to have had a child during their employment at Tabularazor Inc., based upon a leave period comparable to maternity leave and a reduction of working days after this period of absence. With a reasonable amount of certainty, employees of Tabularazor Inc. are expected to have 10 holidays per year.

5 Limitations of the Analysis

The data analysis described within this report is based upon two types of data, namely date and names. Therefore, there is large uncertainty as some of the identified patterns might result from mere coincidence.
Partially caused by the limited data availability, many assumptions had to be made. For example, the maternity leave is expected to range in between 14 and 20 weeks, whilst applicant law is unknown as the country in which Tabularazor Inc. is based is unknown. Furthermore, it could be possible for employees to take a long holiday and come to the decision that they do no longer want to spend their lives working. This is also a possible explanation for the leaves and reduction of working days, which are now identified as having had a child. Adding to that, it is difficult to take part time workers into account when calculating the number of holidays given to an employee. In Dutch Labour Law, most collective labour agreements specify the number of holidays by a multiplier of the hours worked in a week. Part time employees therefore do not get as many holidays as full-time employees as they already have more days off.