Documentation of
Fileado KPIs Case Study Project

Course: Data Preparation

Submitted by Group 6:   Submitted to:

  Kristofar Stavrev   Assoc. Prof. Boryana Bogdanova, Ph. D.

  Radostina Georgieva

  Tamara Vasileva

  Vasil Penev

  Yuliya Kalanzhova

8th of February, 2022

Contents

1. Business understanding

The first step in the CRISP-DM methodology begins with gaining better understanding of how the business operates and building up a domain knowledge in the particular field. In the current case study, the company has decided that in order to increase profits and customer satisfaction levels, they want to track and analyse the website visitor's behaviour. This will allow the creation of sustainable marketing strategies which will attract new and loyal website visitors. All of this can be seen in Figure 1 – the stages of improvement in marketing when converting random visitors to loyal, returning readers - potentially being future customers.

Figure 1: The process of building up a loyal audience through data analysis and data driven marketing strategies
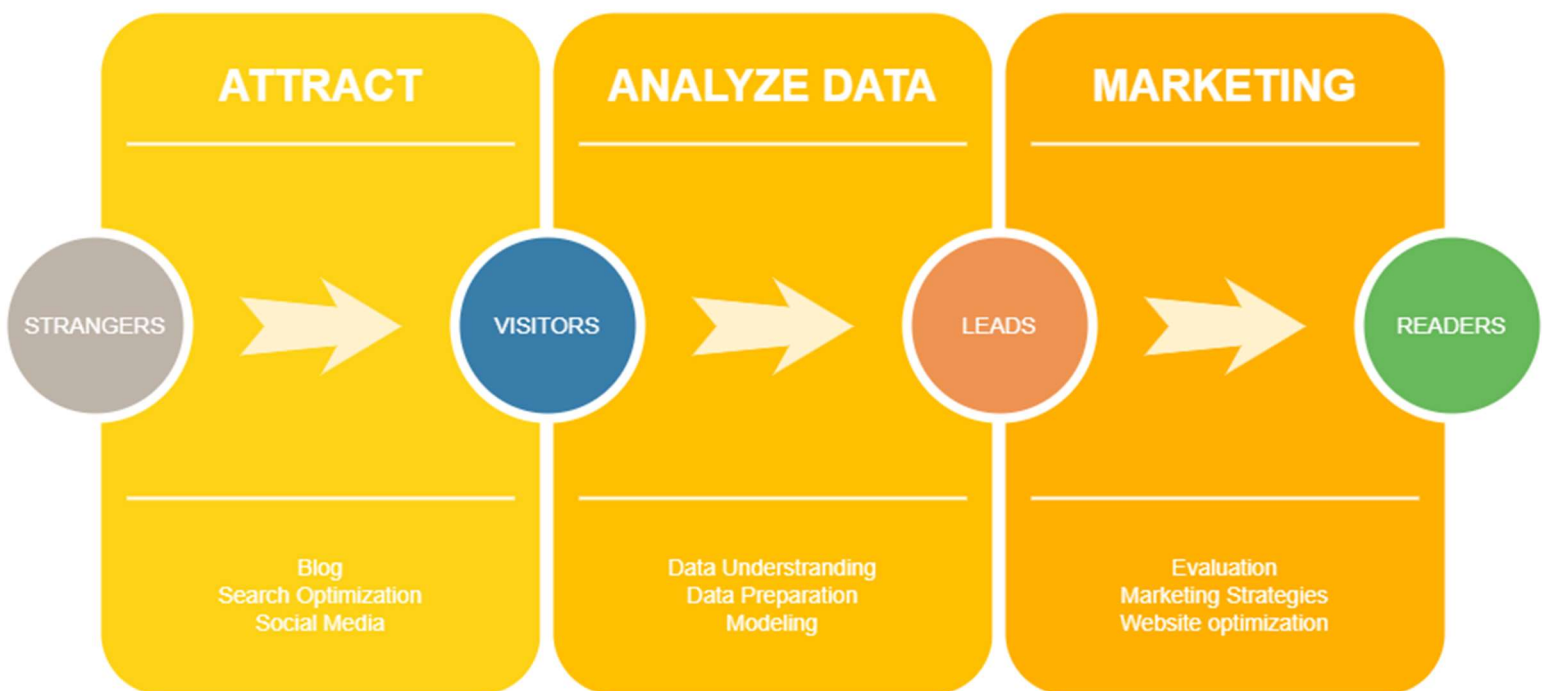
In the last decade more and more people spent time and search for information on the internet. When browsing through the webpages some of the websites gather more attention to their content compared to others. If business owners would like to become competitively viable, it is almost obligatory for them to know and serve their customers' needs by providing

services online, creating engaging blogs, making websites more efficient, investing efforts into SEO (Search Engine Optimisation) all of which contributes to being more discoverable and to improving social media engagement.

In the case of Fileado, a wealth management company, they seek to build a presence in the finance industry while gaining new and active users. This will in return generate more user data which will significantly improve the insight extraction process and the achievement of a better customer understanding. As of the creation of this project one of the more important components of their company website is the blog section where specialised experts are responsible for writing articles on business and financial topics. Therefore, the main goal of the project is to create a way to measure key performance indicators of the website activities and produce monthly and quarterly reports on them using automated data cleaning and analysis methods all done in the programming langue R. This will lead to gaining valuable insights on the Fileado website users and their interests, how they gained access or knowledge about its existence, help steer marketing efforts to increase the visits, and gain an overall competitive advantage.

The next points of this documentation are going to focus on the data understanding process, data preparation and finally presenting the cleaned information with various visualisation techniques which will later be used by management in analysing and creating marketing strategies.

2. Data understanding

The data that this project is going to be working with is stored on an external company server and can be accessed and downloaded at any time. The information can be found in 15 access log files and 15 error log files (for the purpose of the case study we are interested only in the access log files) and they need to be downloaded at least every 14 days. Each file represents a separate day and contains a complete list of all requests made to the web server. Each request to the web server is stored as a single line and there are the same number of columns for each line. In Figure 2 is shown the exact structure of the information and all the different columns that are contained in the raw data log files - IP address, date and time zone, page access method, path of requested file, HTTP version, server response code, bytes received, referrer

URL and additional information about the user agent. All of this will be used in some form or another when cleaning, filtering, formatting, and organising the data later in the project.

Figure 2: Data format explained
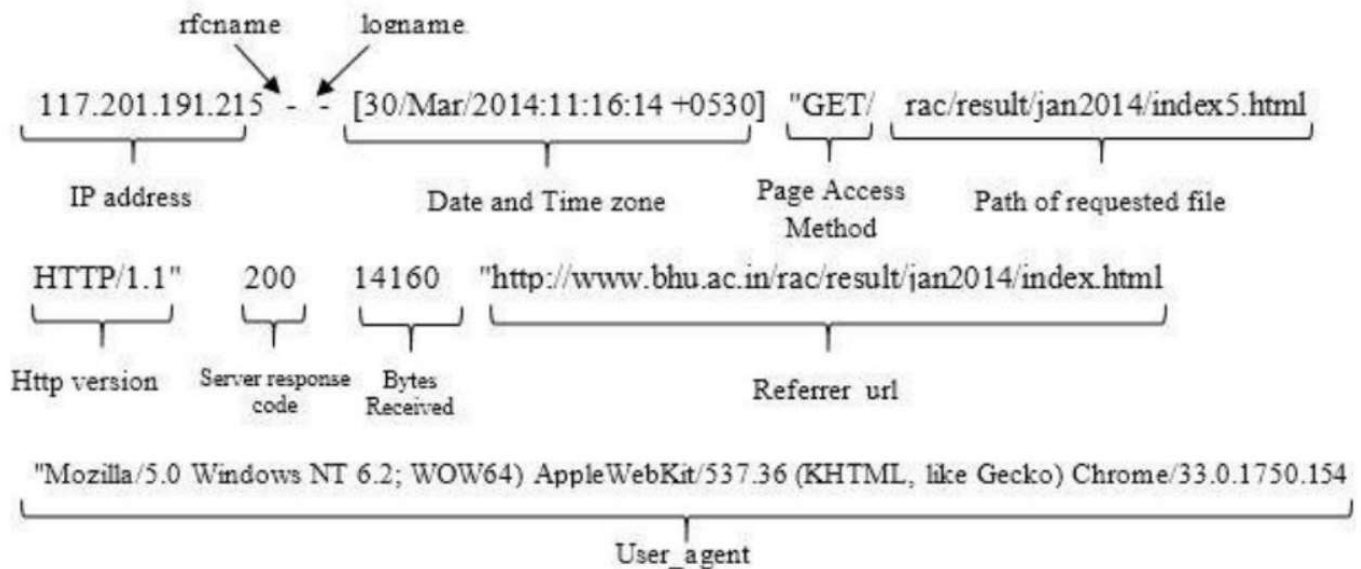
3. Data Preparation

Before we dive deeper into the process of data preparation it is important to mention the methodology used for collecting, cleaning, and visualising the information from the log files. The main intention was to create an automated, scalable, and long-term solution which allows for an easy way to include new data in the analysis and the dashboard of charts. For this purpose, our team decided that the best approach would be to create an artificial database file under the format of CSV that would contain the entirety of the cleaned and filtered data. Every time new log files are downloaded from the server they can simply be placed in a specific location and the script would automatically clean, organise, and append to the CSV database. The script would then proceed to update (or create) all the charts present on the dashboard. The first part of this whole process would be to loop through all the log files and extract them from the ".gz" format they are stored. Then in another loop the script loads the data (excluding current day) into a data frame and begins filtering by all the necessary criteria

– page access method should be "GET" without including any "static" or "debug" parameters, server response code should be 200, bytes received needs to be a minimum size of 3000, exclude the rows that contain keywords "bot", "crawler", "agent", "research", "scan", "data", "grab", "http" using the user agent field. In addition to that a mapping between article and author is made since that is not initially present in the data. The script also creates new columns that extract information about the visited page of the website. After all the necessary columns are created, data filtered and cleaned the script proceeds to save the data frame into the CSV database while also checking for any duplicate records. The steps are then repeated for every log file until all of them have been successfully imported in the database. One of the many benefits of this approach is that the artificially created database file can be used in the future for additional analysis, reports and even applying machine learning models for further extraction of insights and creating predictions.

The next step in the script is cleaning the environment and loading the now ready database in order to begin data aggregation. For the purpose of creating a better and more complete dashboard the aggregation is made on many different levels (total and distinct visits by day, week, month, year, visits per website sections, visits per articles, popularity of authors) and also includes the extraction of further information from the data about user journeys and drop-off sections.

4. Data Visualisation

After the aggregation is complete the script moves on to the final part – creating an interactive dashboard and filling it with visual representation of the data with various types of charts. The dashboard itself is created using the R library "flexdashboard" - a package that enables you to easily create flexible, attractive, interactive dashboards using R Markdown. The main type of charts that were used are:

- Column chart (as well as grouped/stacked column chart);
- Bar chart (as well as grouped/stacked bar chart);
- Grouped line chart;
- Doughnut chart;
- Data table.

They were created using the libraries "ggplot2" and "DT" (for the data table). All of them are completely user interactive and can be seen in the figures below.

4.1. Website Visits

Figure 3: Multi-Line chart for daily website visits – total and distinct
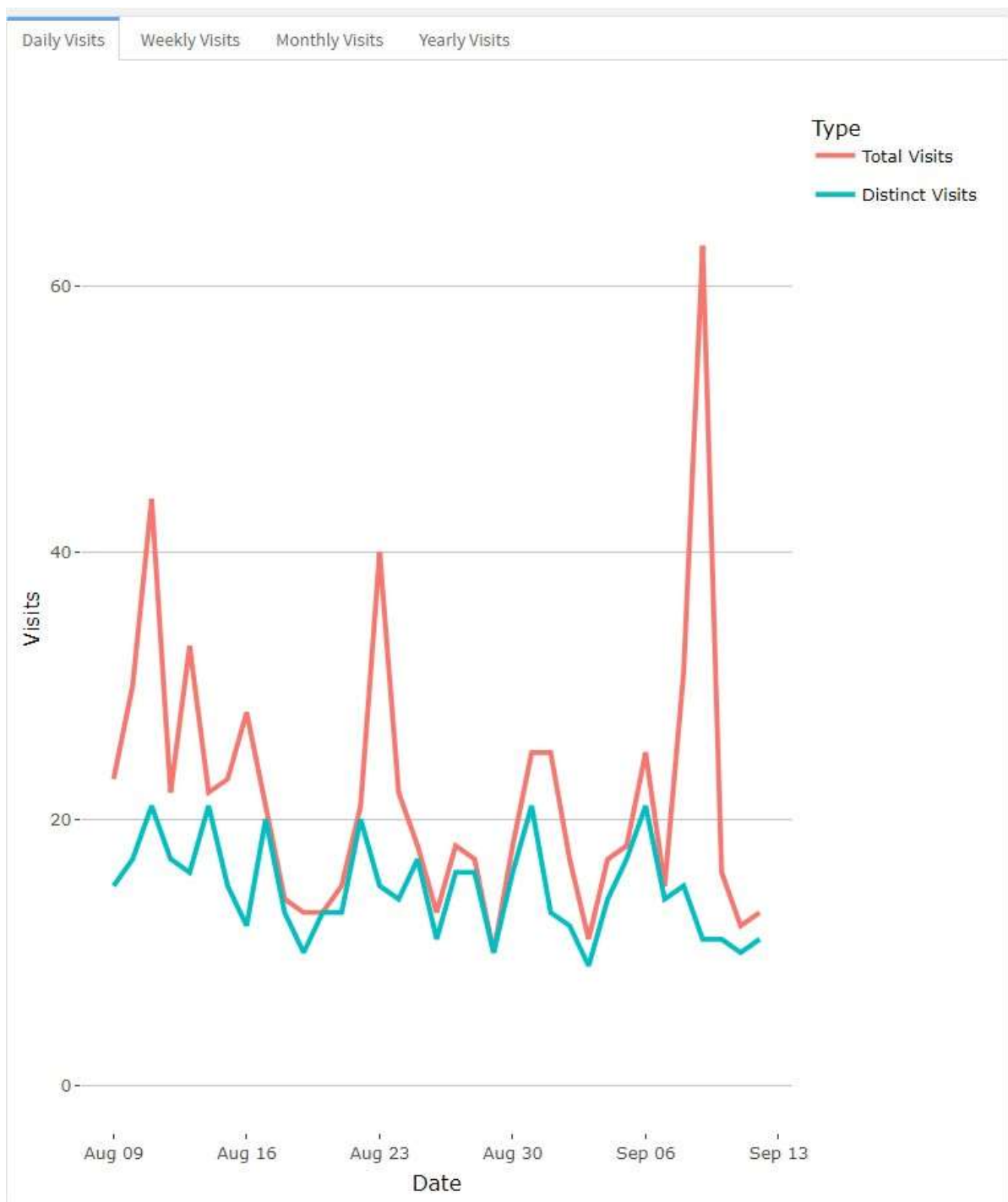
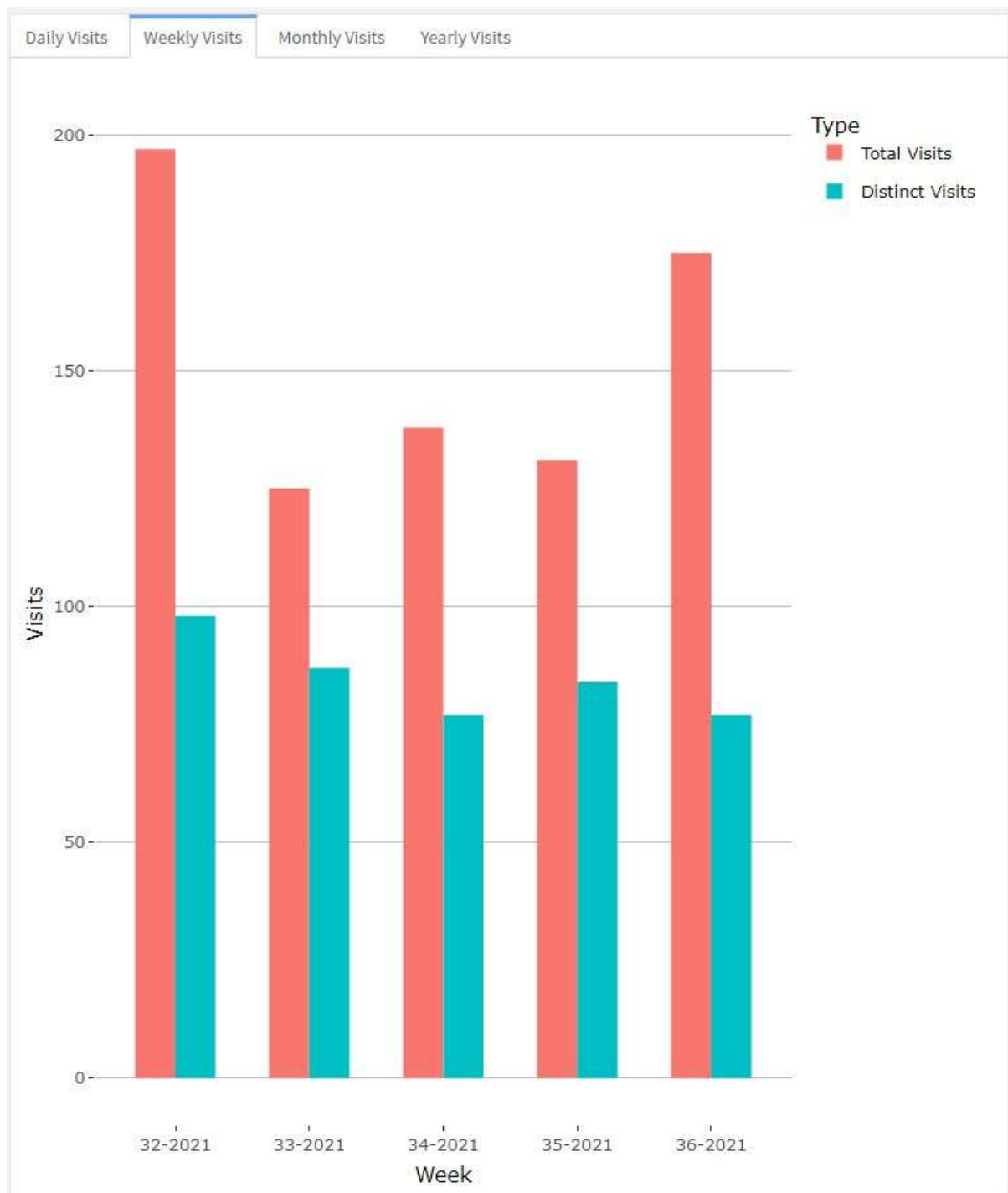**Figure 4: Grouped column chart for weekly website visits – total and distinct**

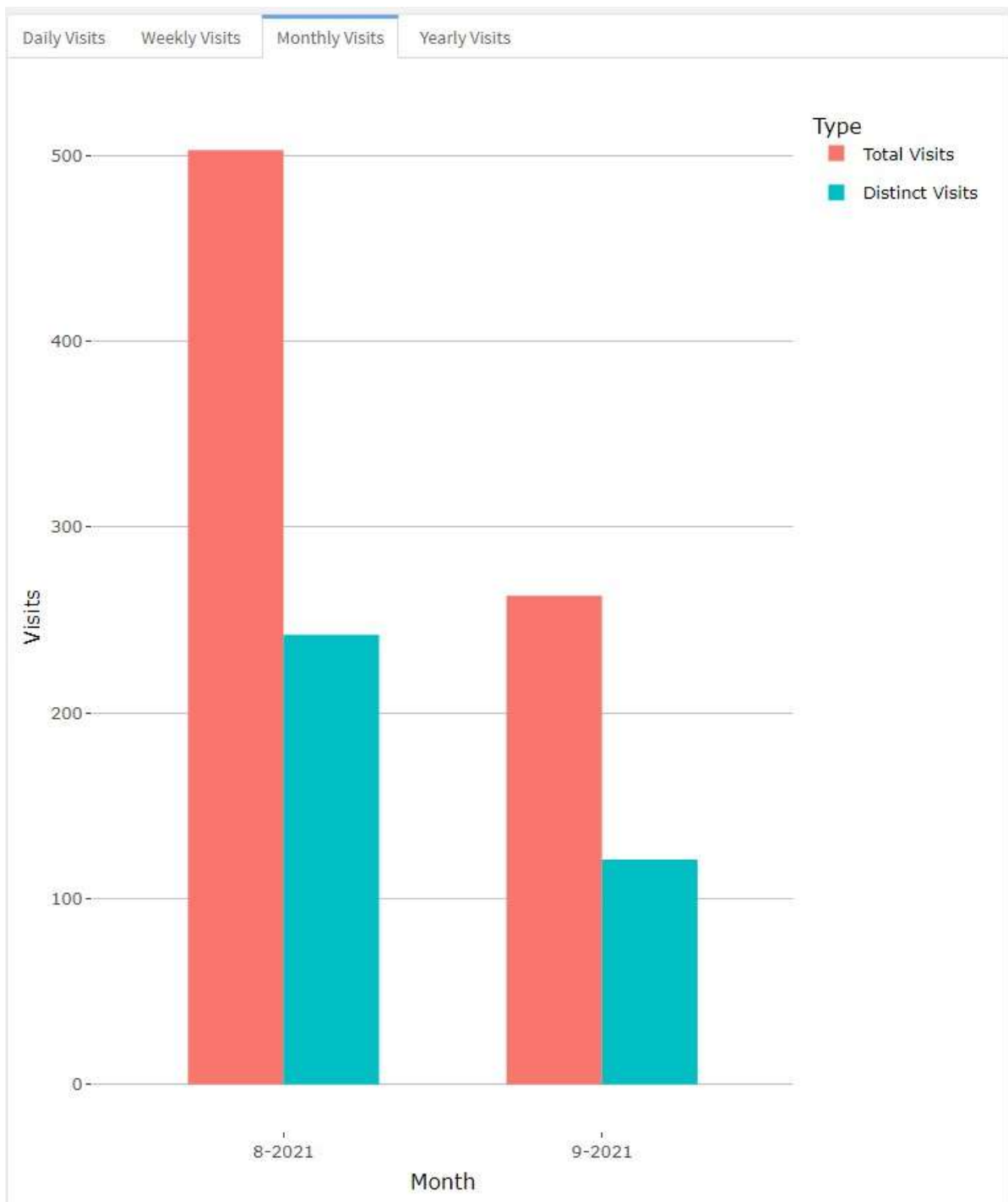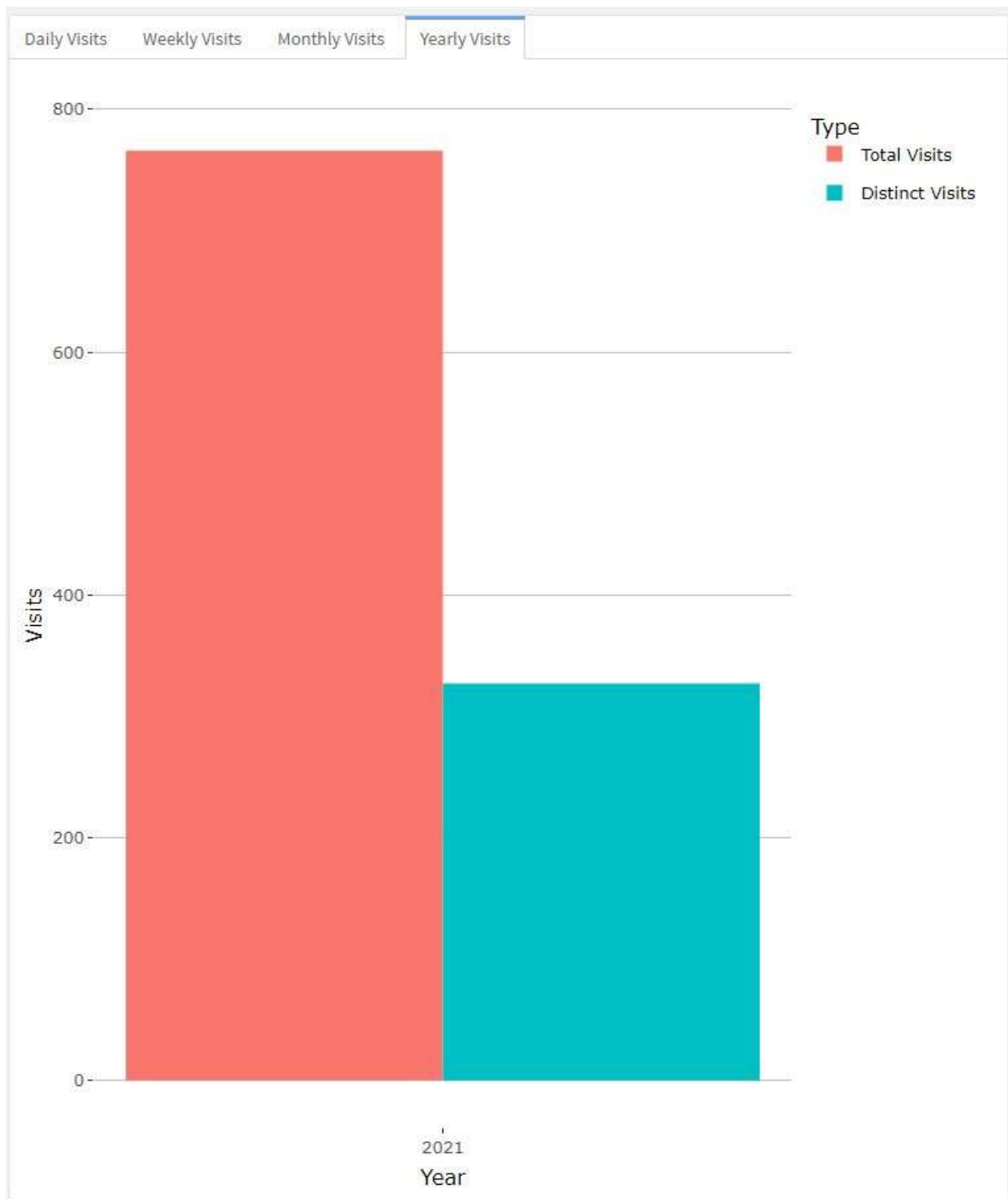**Figure 5: Grouped column chart for monthly website visits – total and distinct**

Figure 6: Grouped column chart for yearly website visits – total and distinct

4.2.   Total Visits by website sections

Figure 7: Stacked column chart for total weekly website visits by Sections
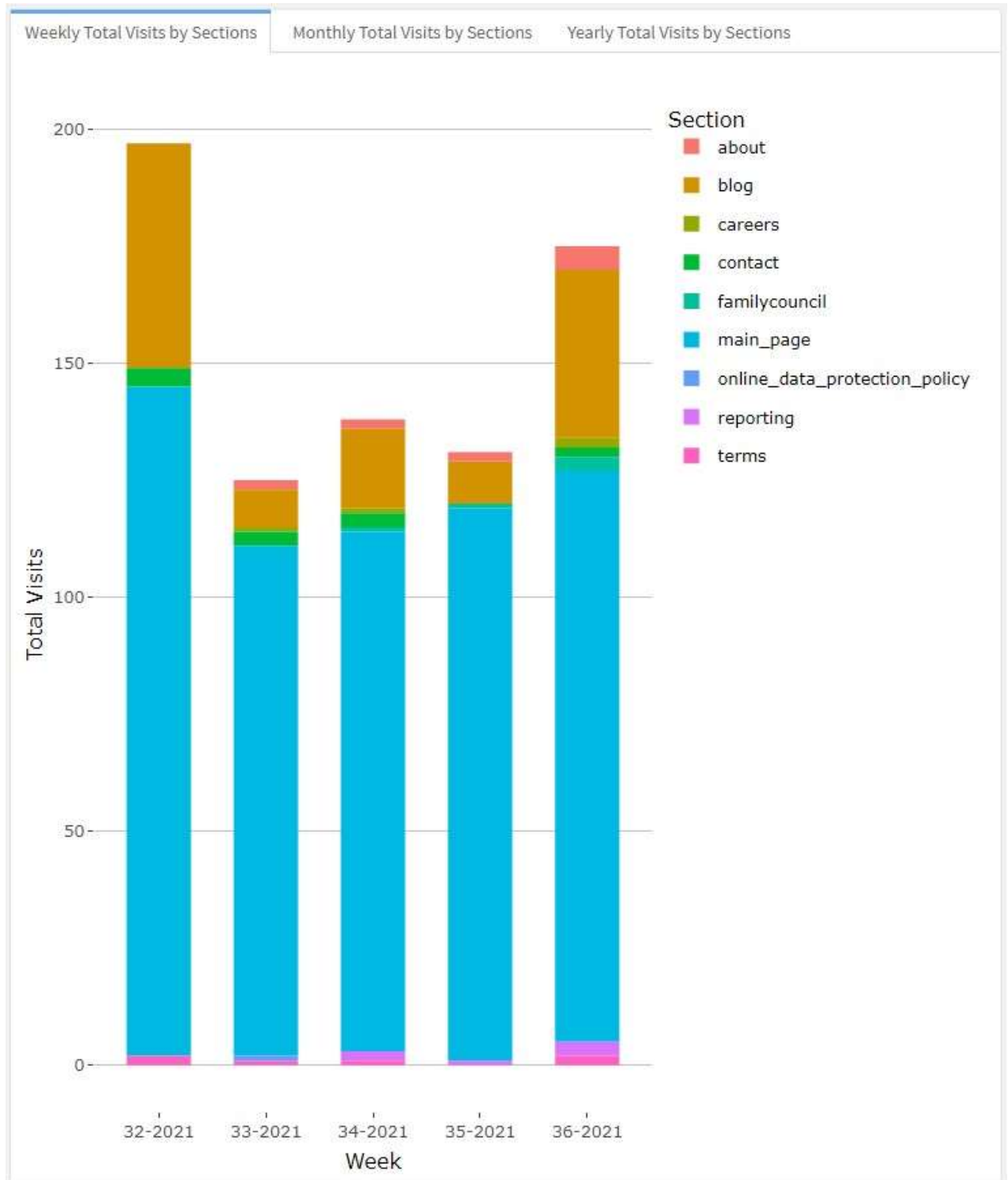
Figure 8: Stacked column chart for total monthly website visits by Sections
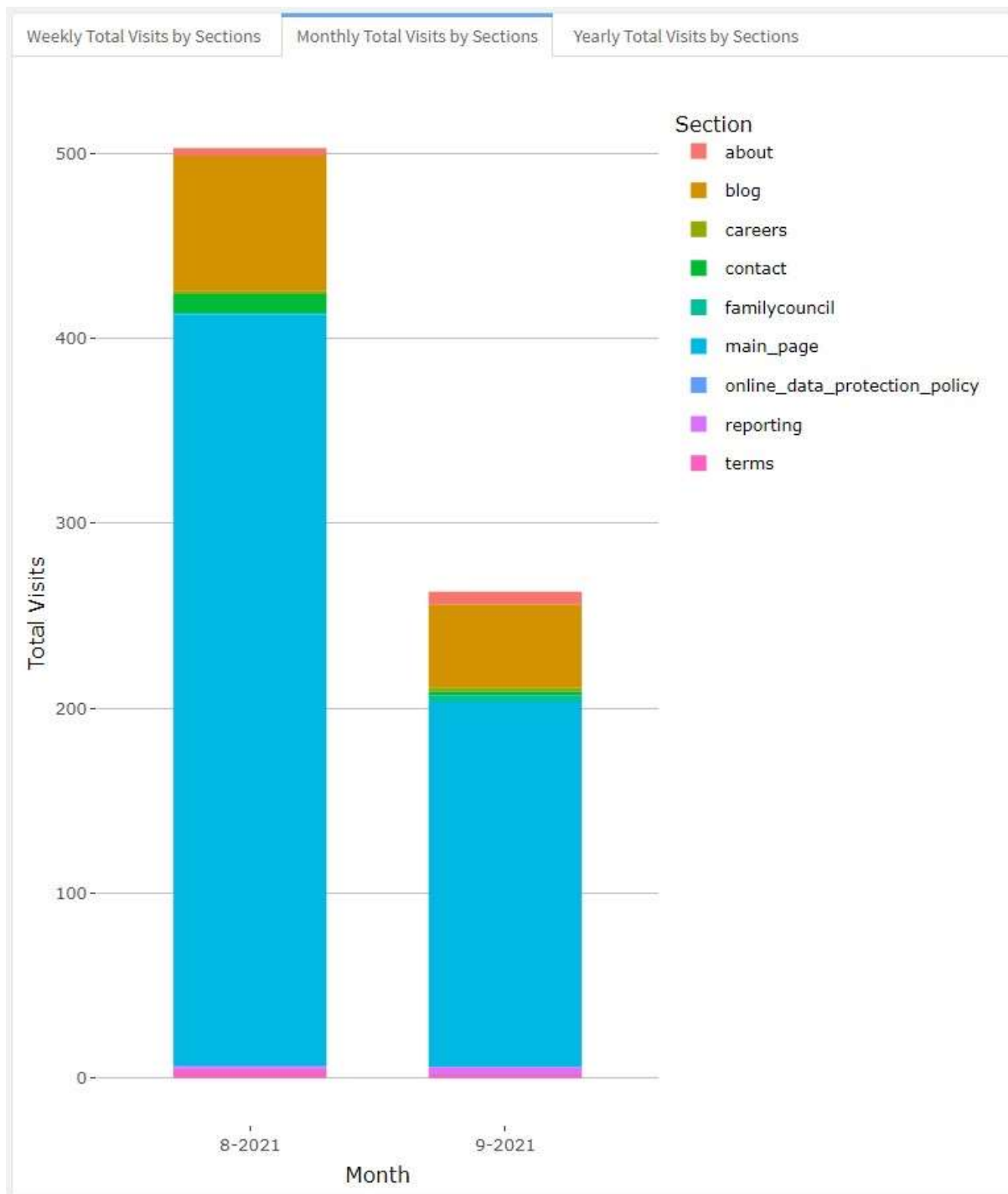
Figure 9: Stacked column chart for total yearly website visits by Sections
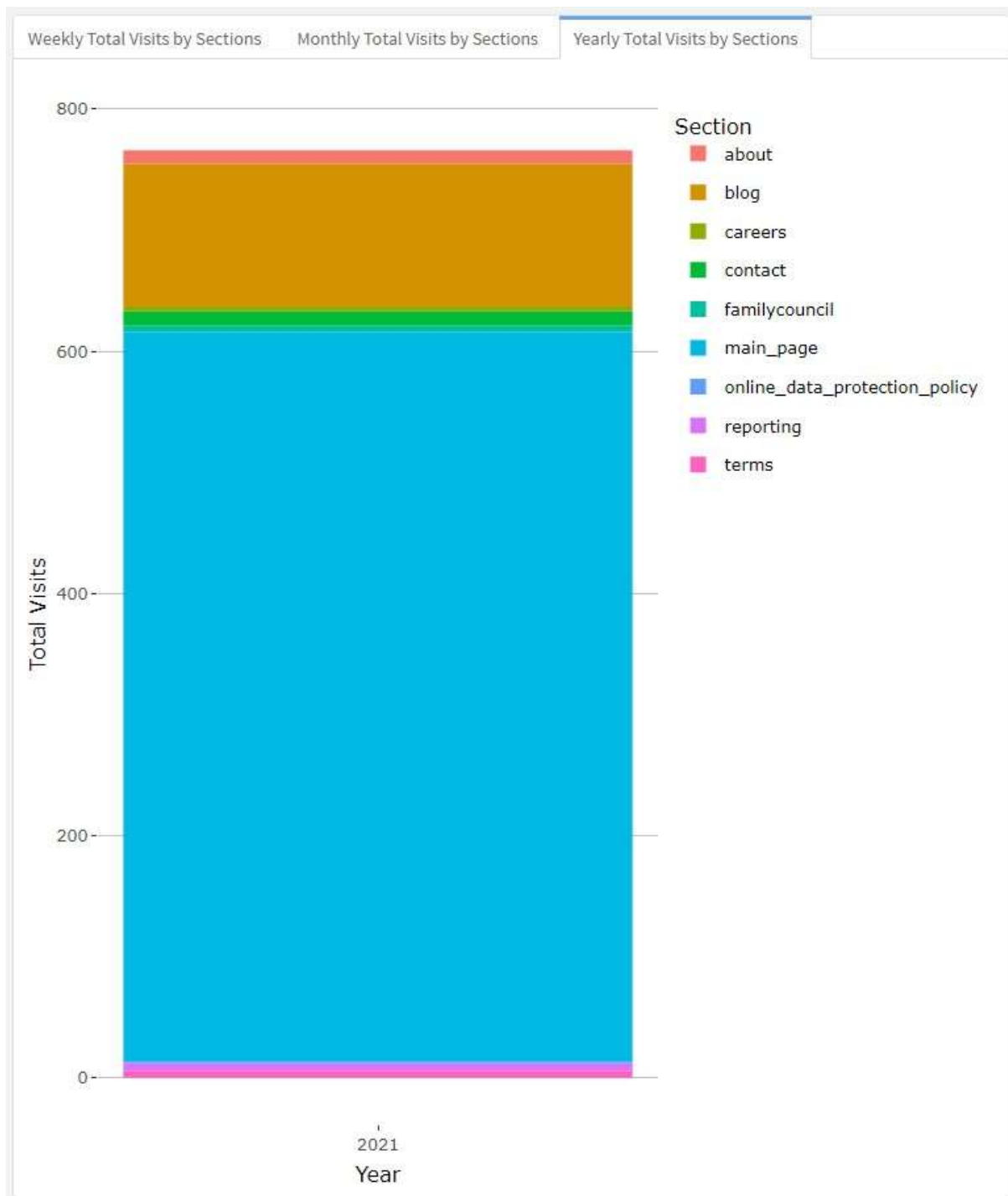
4.3. User behaviour

Figure 10 and 11: Bar chart for user drops by pages (left) and number of pages visited per user session (right). Drops are defined as the page on which the user ended their website browsing activity. The timeframe for one session is set to one calendar day.
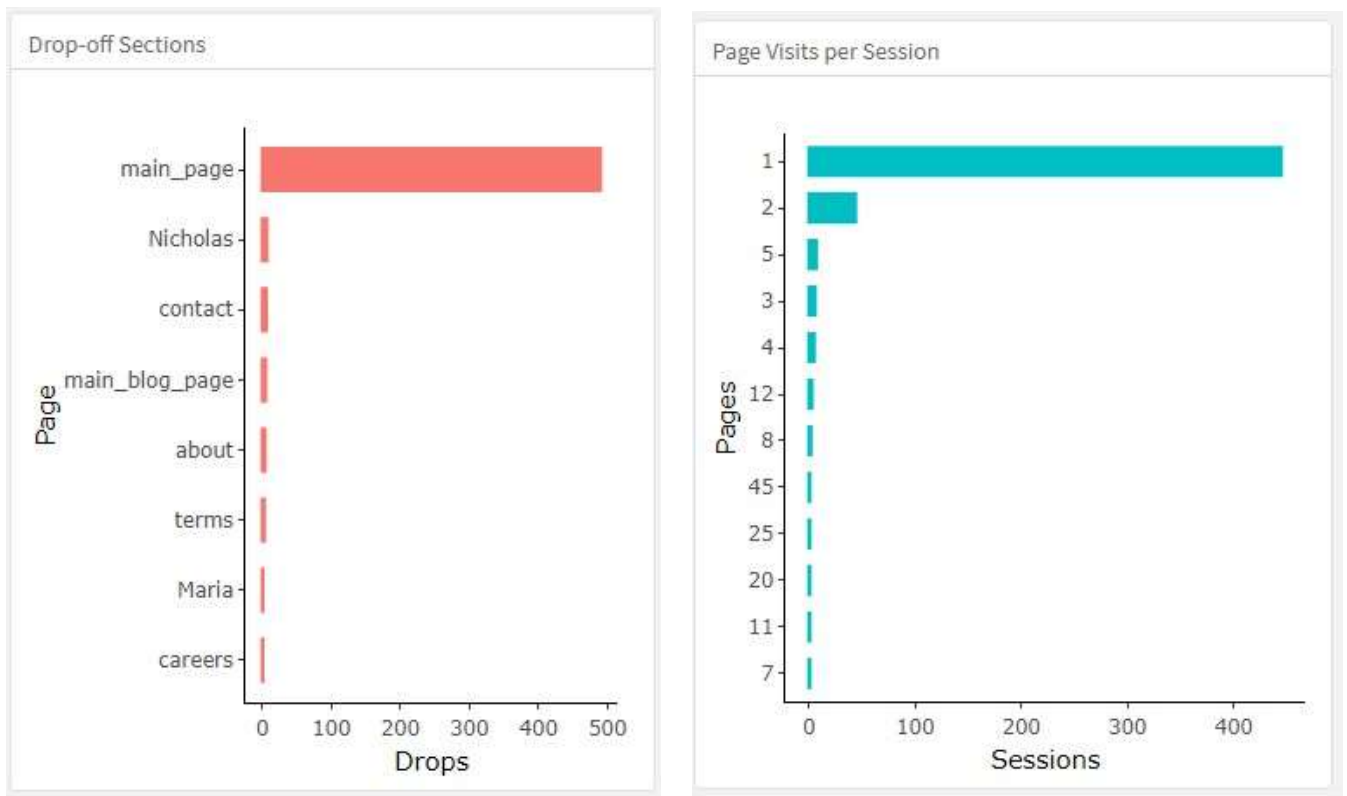
4.4. Blog section

Figure 12: Table with information about article popularity and its respective author

Total Article Visits

Copy	Print	CSV	Show 25 ▼ entries		Search:

	Page	Author	Total Visits	Distinct Visits
1	5-Reasons-that-Projects-Get-Delayed	Maria	2	2
2	Family-Wealth-Four-Types-Of-Capital	Nicholas	22	4
3	Financial-Branding-With-Videos-Enhance-Your-Website-Design	Rebeka	1	1
4	How-do-I-choose-an-asset-manager?	Zara	3	3
5	How-To-Build-An-Awesome-Team	Sherry	1	1
6	How-to-create-a-unique-logo	Rebeka	1	1
7	Keeping-Together-3-Gens	Nicholas	4	3
8	Let's-Celebrate-Project-Managers	Maria	1	1
9	People-dont-leave-their-jobs-they-leave-their-managers	Maria	1	1
10	Streamlining-the-Onboarding-Process	Sherry	3	3
11	Why-Should-People-Invest-In-Gold	Zara	1	1
12	Working-In-Pajamas-A-Guide-To-Productivity	Maria	1	1
13	Writing-Your-Companys-Values-for-Long-Term-Success	Sherry	3	3

Showing 1 to 13 of 13 entries

Previous 1 Next

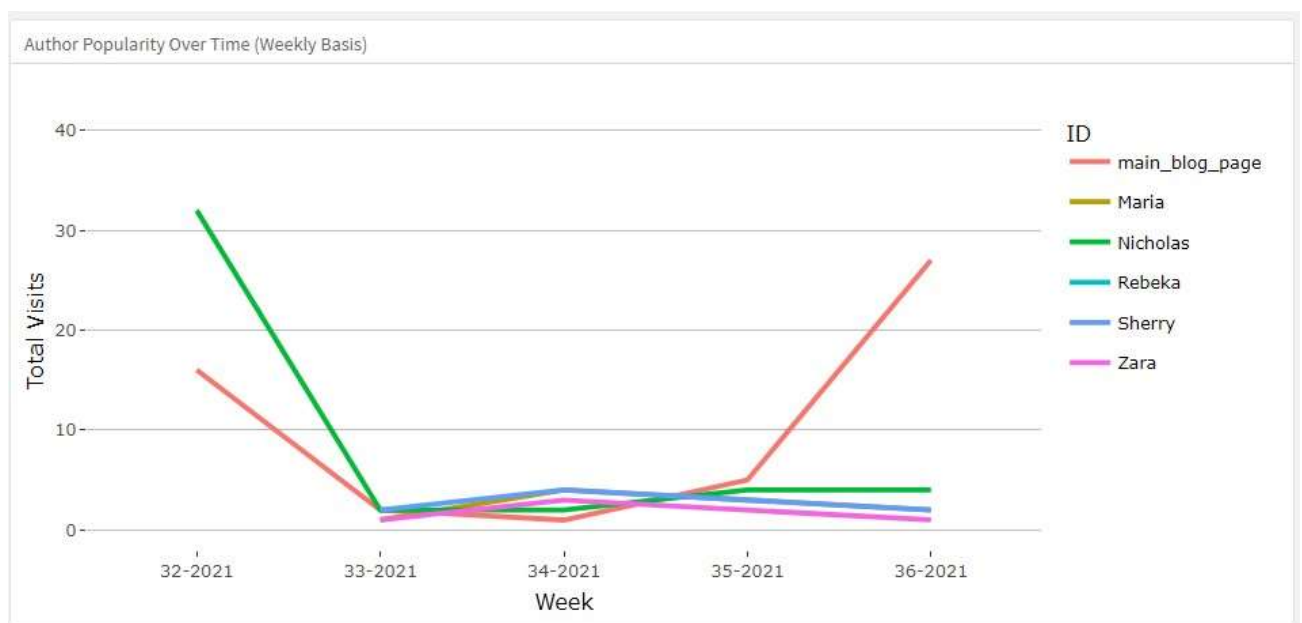**Figure 13: Doughnut chart for author popularity (also including the main blog page)**

Most Popular Authors by Total Visits Share
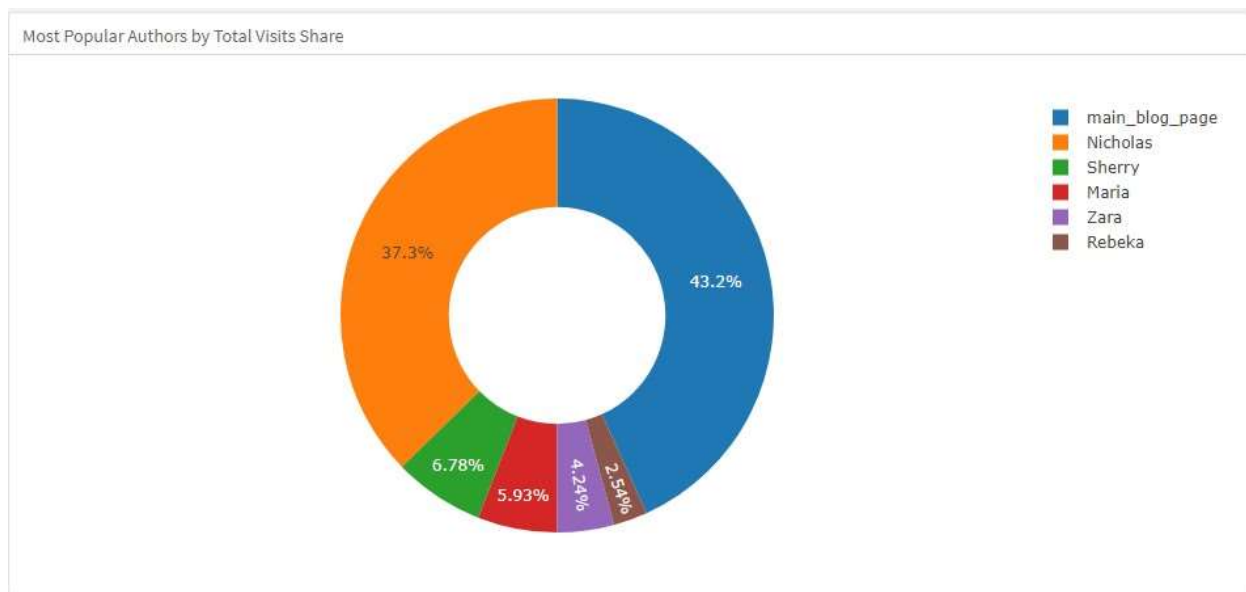
Figure 14: Multi-Line chart for author popularity over time on a weekly basis (also including the main blog page)

Author Popularity Over Time (Weekly Basis)

5. Data analysis using the dashboard

One thing that must be noted is that the current data provided covers from the 9th of August till 12th of September 2021. This explains why there is a sharp decline in visits for the month of September. There is an approximately two times the difference which indicates that if the trend continues September will end up with the same number of visits as the month of August. The most visited pages overall, are the main page and the blog section (this includes the main blog page, articles, and author profiles). The user journey can also indicate if the traffic that is being generated on the website is genuine and "healthy". Currently most of the generated sessions visit only a single page. In addition, the home page is also the section with the most amount of session drops. Using this information, it can be concluded that most of the website visitors land on the main page and leave without visiting the other parts of the site. Moving on to the "Blog Section" charts the first thing that can be noticed is that Nicholas is by far the most popular author amongst the other (also having the most visited article). Despite that most of this is a result of his significantly increased visits at the beginning of August. Following the first week there is a sharp decline in his visits never fully reaching the previous high.

6. Deployment

A software is not particularly useful unless the customer can access its results. Therefore in this final stage of the project it is important to descibe the steps necessary for the bussiness implementation. The complexity of this phase usually varies widely but in our case user friendliness has been a core focus during the development of the product. Detailed instructions and steps for deploying the code can be found below.

1. Place the three R scrips ("dashboard.Rmd", "data_aggregate.R", "data_clean.R") in a new and empty folder.

Name	Date modified	Type	Size
logs	04/02/2022 10:02	File folder	
dashboard.Rmd	03/02/2022 23:45	RMD File	5 KB
data_aggregate.R	03/02/2022 23:50	R File	7 KB
data_clean.R	03/02/2022 19:11	R File	7 KB

2. Create a new directory called "logs" and place the folders containing the server log files inside.

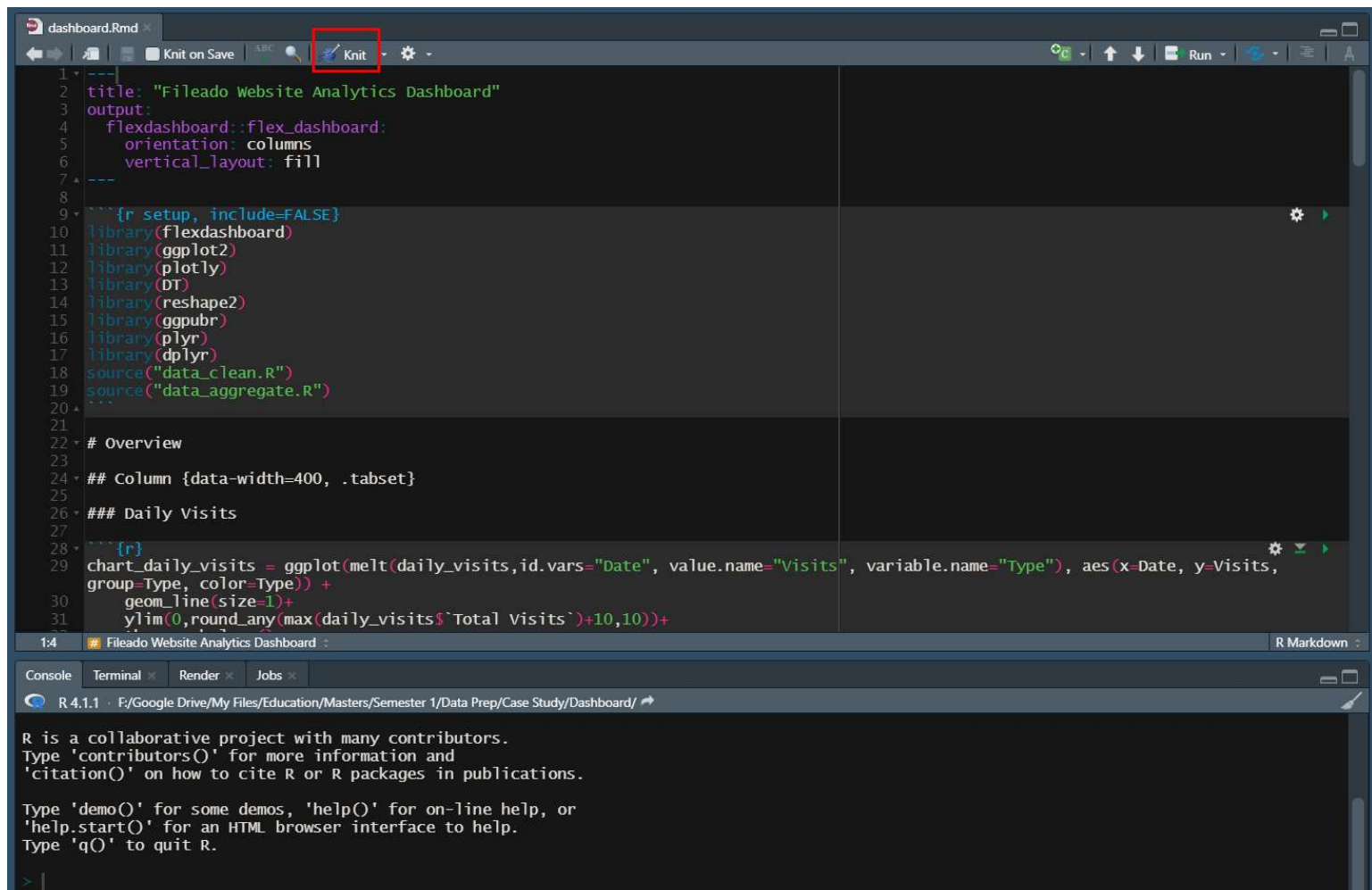| Name	Date modified	Type	Size
2021_08_23	28/11/2021 16:19	File folder	
2021_09_01	28/11/2021 15:11	File folder	
2021_09_13	28/11/2021 15:11	File folder	

3. In order to execute the R script, you must first install these libraries:
   - "stringr"
   - "lubridate"
   - "R.utils"
   - "dplyr"
   - "flexdashboard"
   - "DT"
   - "ggplot2"
   - "plotly"
   - "reshape2"
   - "ggpubr"
   - "plyr"

   To do so run this command in the RStudio console for each library → install.packages("stringr"). If you already have them installed, you can skip this part.

4. Open file "dashboard.Rmd" with RStudio and press the "Knit" button. The script takes about 30 – 120 seconds before it completes extracting, processing, and visualizing the data. The main reason for this is that the ".gz" extraction process takes a significant amount of time. After the script is done and the data is uploaded to the CSV database everything in the "logs" folder will be deleted in order to improve future runtime and memory usage!

The "Knit" button can be seen in the image below

When you see this icon in the Render tab the script is still loading.

In the below screenshot is shown how the Render tab looks when the script has finished working.
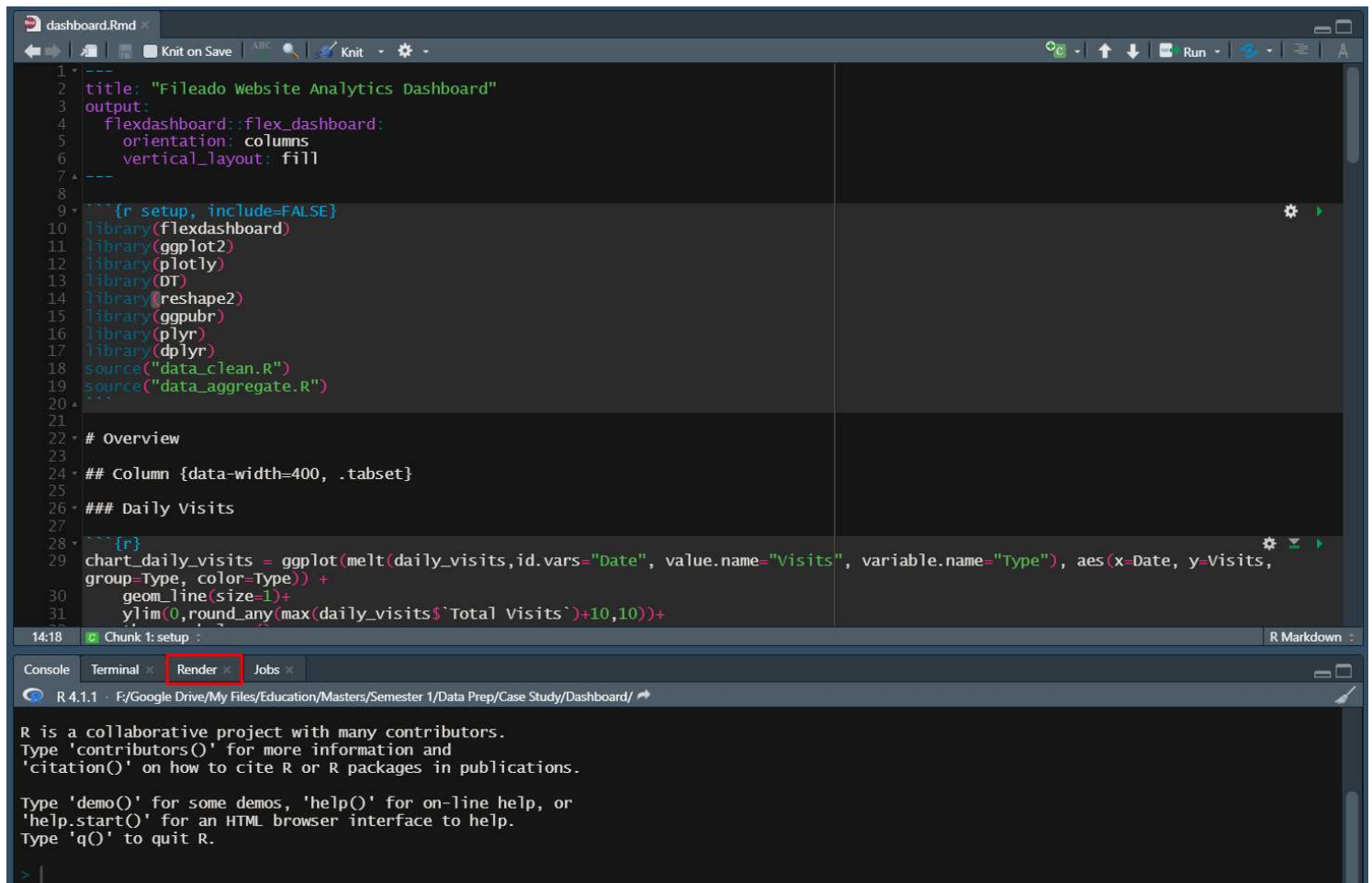
5. After the script is executed successfully a "db" folder is created in the same directory. Inside of it can be found the CSV database file named "db_raw_data.csv" containing all the cleaned, filtered and ordered data. In addition, in the root directory a new "dashboard.html" file was created. Open it to load the dashboard in your browser. From now on you can load the dashboard by opening the "dashboard.html" file. The script is only used when uploading new data into the database (new log files).

Name	Date modified	Type	Size
db_raw_data.csv	04/02/2022 10:46	Microsoft Excel C...	169 KB

Name	Date modified	Type	Size
db	07/02/2022 09:26	File folder	
logs	07/02/2022 17:38	File folder	
dashboard.html	04/02/2022 10:46	Chrome HTML Do...	6,040 KB
dashboard.Rmd	03/02/2022 23:45	RMD File	5 KB
data_aggregate.R	03/02/2022 23:50	R File	7 KB
data_clean.R	03/02/2022 19:11	R File	7 KB