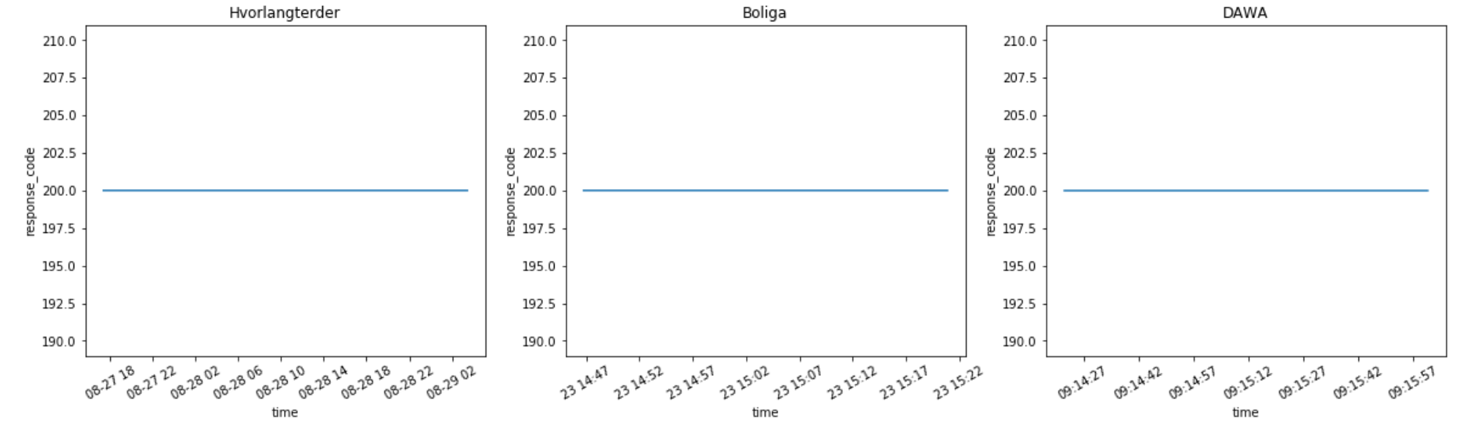
**Analysis of Scraping logs**

Three datasets were obtained through web scraping from Boliga.dk, Hvorlangterder.dk and DAWA.dk. This section will analyze the logs created in the scraping process, with the intention of providing evidence of the data quality. The analysis the scraping logs was done according to Snorres demand on analyzing logs.

**Response Code**

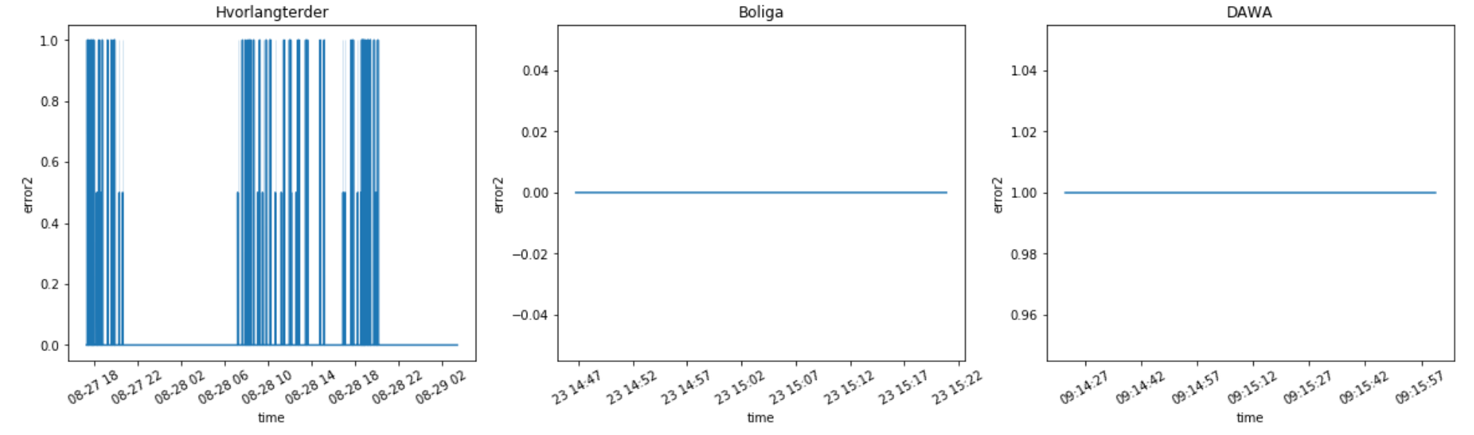
When parsing data from websites, the goal is to receive a response code ‘200’, expressing a successful request. Fig XA displays response codes for the duration of the web scrapings. Every single request received a response code 200. Further evaluation of the response codes was not necessary.



XA

**Error Codes**

Figure XB plots errors codes received throughout the duration of the scrapings. Boliga and DAWA received no errors, while hvorlangterder received a total of 80 errors. The error code returns the same response in all 80 cases: **('Connection aborted.', RemoteDisconnected('Remote end closed connection without response'))**

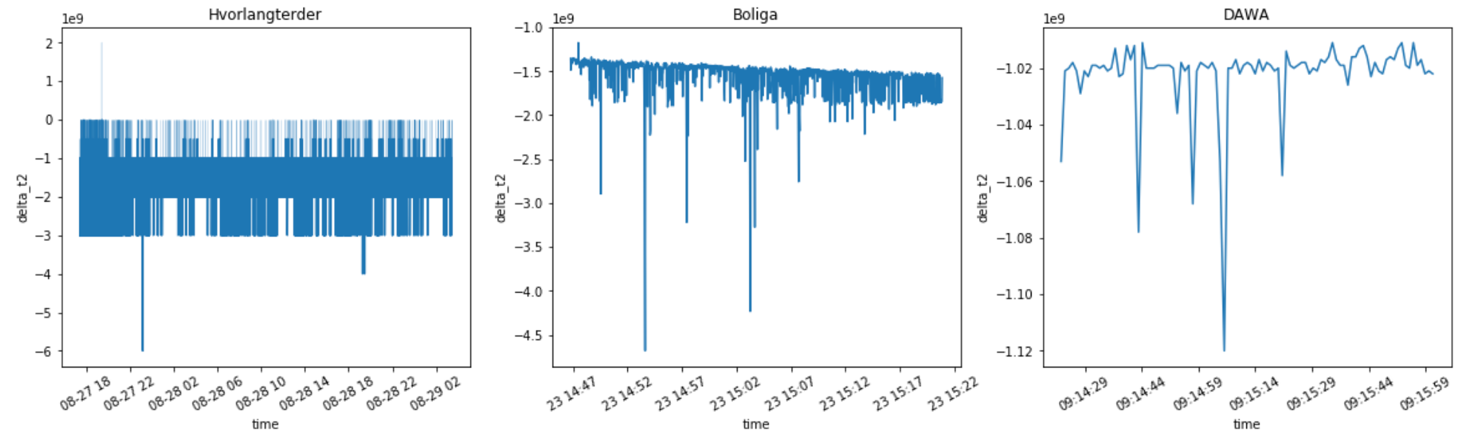


XB

To understand the errors, a closer manual analysis of the log was performed on requests receiving errors. The errors occur in random intervals and are subsequently followed by another request to an identical URL. Therefore, we assessed that these errors were due to connection problems or server traffic. Since the errors were corrected by a following request, we assessed it would not affect the research.

**Server response times**

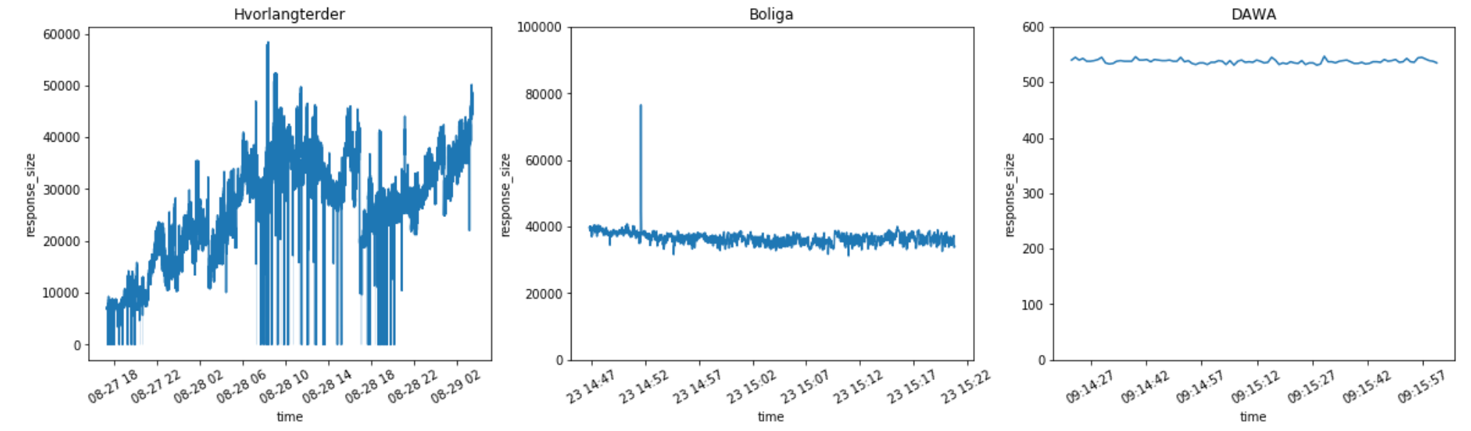
An indicator of poor data quality is varying the response times from the server being scraped. Figure XC plots changes in response times for the duration of the web scrapings. The 3 data sources all resulted in few variations of response times. Responses times differing from usual within the logs, have been assessed to be caused by two either a slow server time or connection, or due to larger quantities of data being requested. To validate the quality of data, manual control of the returned result and the webpage result was performed, without the notice of significant corruptions in the data to be utilized in the research.

****

XC

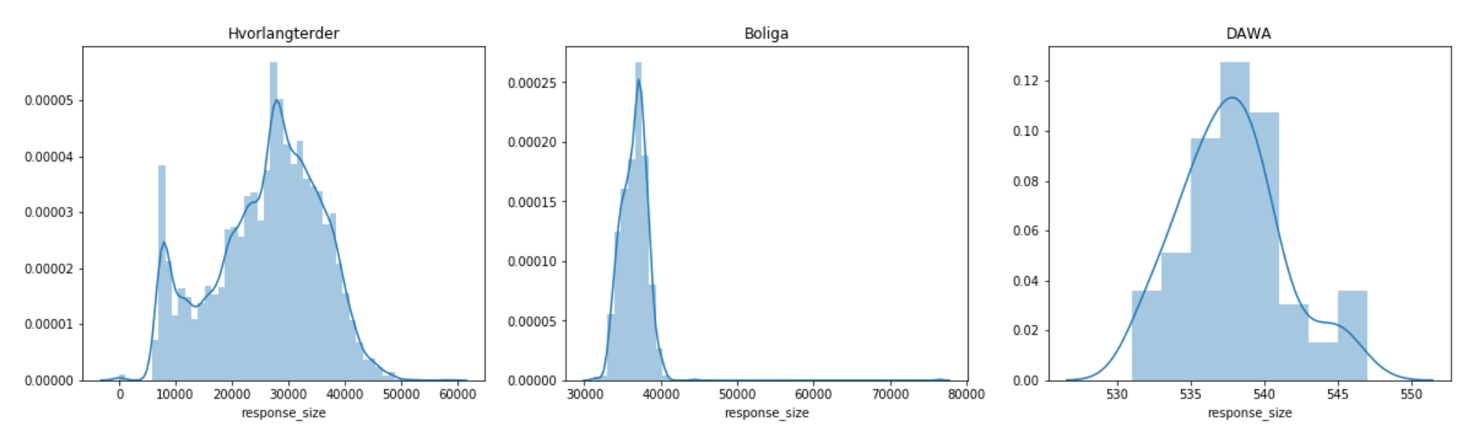
**Response sizes**

To validate the response quality, an assessment of the response size was performed. Figure XD plots response sizes throughout the duration of the web scrapings. Boliga contains a single noticeable large response size within short time, while hvorlangterder response sizes vary throughout the scrape. The DAWA scrape resulted in indifferent response sizes, and therefore needed no further investigation.

****

XD

With a further analysis the distribution of sample sizes in figure XE, the Boliga and DAWA response sizes seem to be close to normally distributed, except for two single cases of large sample sizes in the Boliga log. By manually assessing the returned output of the Boliga scrape, these two requests were found to return a larger amount of data than usual. The two requests were made to two pages, where few properties contain nested dictionary attributes approximately 59 times larger than the average size of dictionaries within the ‘image’ data column. These dictionaries are image descriptions, and are not utilized within this research. These larger responses were deemed as not polluting the results of the research.

****

XE

Figure XE depicts that the hvorlangerder response sizes are a bit left scewed, while Figure XD shows that the distribution of large sizes happens during the middle of the scrape. With a further glimpse into the log, we observed two things. First, all response sizes of 0 are related to the errors, displayed in figure XB. The variance of responses could be explained by the type of data retrieved from the scrape. The request was set to return names, addresses, distances and other information, which sizes are varying in nature. When extracting the value from every single key of the dictionary output, no errors or missing values were found. Further control of samples of large and small sizes were performed without discovering flaws. The response sizes from hvorlangterder are found to vary by nature, which may explain the difference.

**Assessment of the scraping logs**

The individual analyses of the scraping logs showed minor suspicious traits which were assessed manually through the output or logs. The most concerning analysis being difference in response sizes and response times from the hvorlangterder scrape. This scrape consumed large amounts of time and computing power, as every single observation from the Boliga data, had to be input for requesting a response. In an extensive research, these problems could be handled by gaining direct access to the website instead of scraping and would benefit by running on a server with greater computing power. The results of the data collection were evaluated to be satisfying, and not corrupting the results of our research, as fairly reasonable explanations for noticeable dissimilarities within the logs were found.

**Critique of Scraping logs using scraping\_class**

In the process of analyzing the scraping logs, minor flaws from the scraping library named ‘scarping\_class’ were detected. The returned columns ‘t’ and ‘delta\_t’ contained non-generic formats of epoch time. This was fixed by applying string manipulation and datetime conversions.