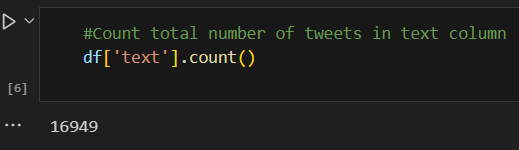


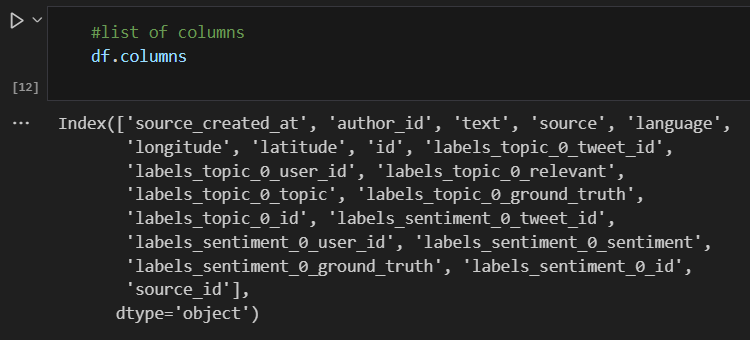
We are using a JSON file

—----------------------------------------------------------------------------------------------------------------



Totally 16949 Lines  
More rows provide a larger and more diverse set of training data for our NLP model. This can lead to a more accurate and effective model for sentiment analysis and classification. Also more data allows for a more comprehensive understanding of customer sentiment towards Thameslink's maintenance services. This deeper insight can lead to more effective strategies for improvement.

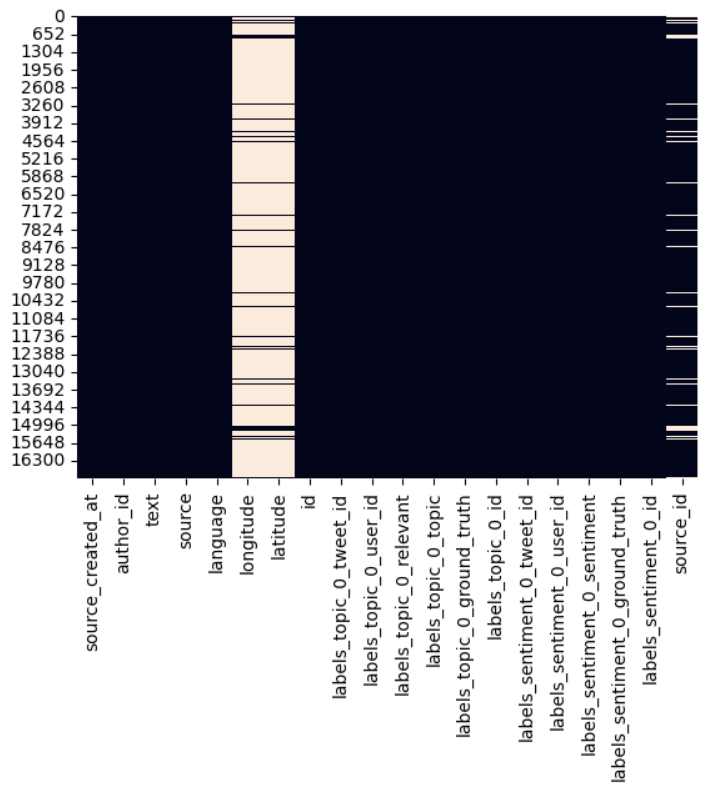
—----------------------------------------------------------------------------------------------------------------



Totally 20 columns to to detail our findings.

With an expanded set of features, we can perform a more detailed analysis. This can result in more fine-tuned recommendations for maintenance improvements, aligning closely with the project's aim of using NLP techniques to enhance maintenance strategies. Also having more columns offers a broader set of variables to use in prompt engineering. This enables you to ask users specific questions or prompt them for feedback in a targeted manner, supporting the project's aim of automating action recommendations

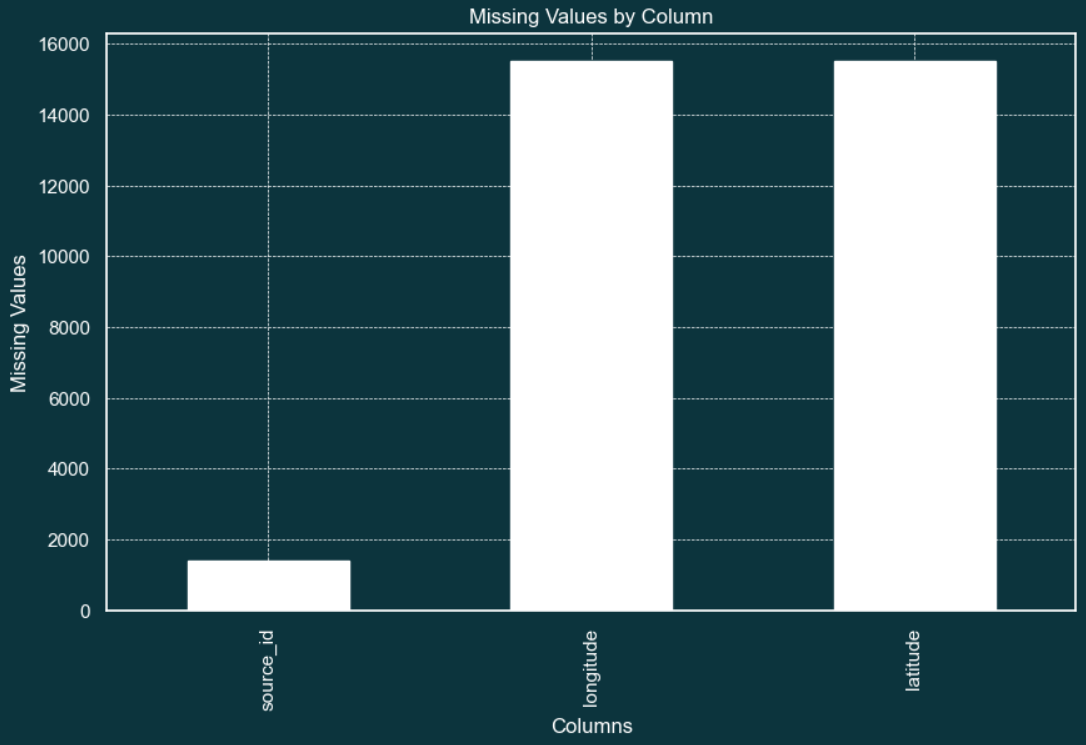
—----------------------------------------------------------------------------------------------------------------



3 Columns have missing values

Addressing minimal missing data allows for a finer analysis of customer sentiments. This leads to interventions, quicker issue resolution, and minimized disruptions for Thameslink services.

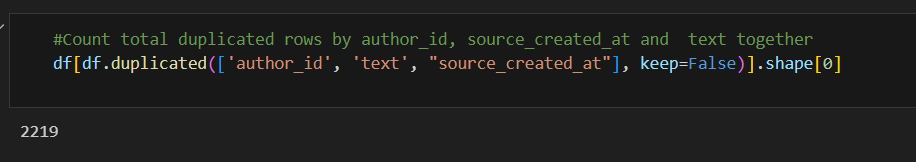
—----------------------------------------------------------------------------------------------------------------

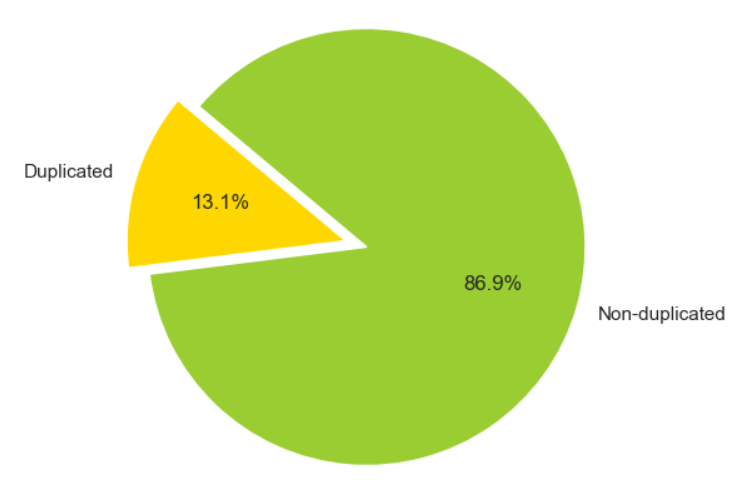


Longitude and Latitude columns are not able to be used, where we cannot gain any finding from GPS values.

Missing values in Longitude and Latitude columns mean that there is a lack of geographical context for those particular tweets. This can potentially limit your ability to analyze location-specific sentiments or issues related to Thameslink's maintenance services. Without this information, it may be more challenging to design precise and location-specific maintenance interventions.

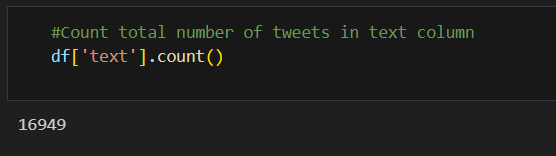
—----------------------------------------------------------------------------------------------------------------





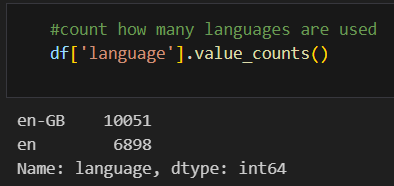
Totally 2219 duplicated rows depending on time, writer and text. Which is equal to 13.1% of all data.  
Duplicates may overemphasize certain sentiments or feedback, potentially leading to skewed insights about customer feedback related to Thameslink's maintenance services. This could result in an overemphasis on certain maintenance concerns, potentially leading to suboptimal recommendations.

—----------------------------------------------------------------------------------------------------------------



Totally 16949 tweets so we havent got any missing area on one of the important column

—----------------------------------------------------------------------------------------------------------------



Used language is english so we can make our research just on one language

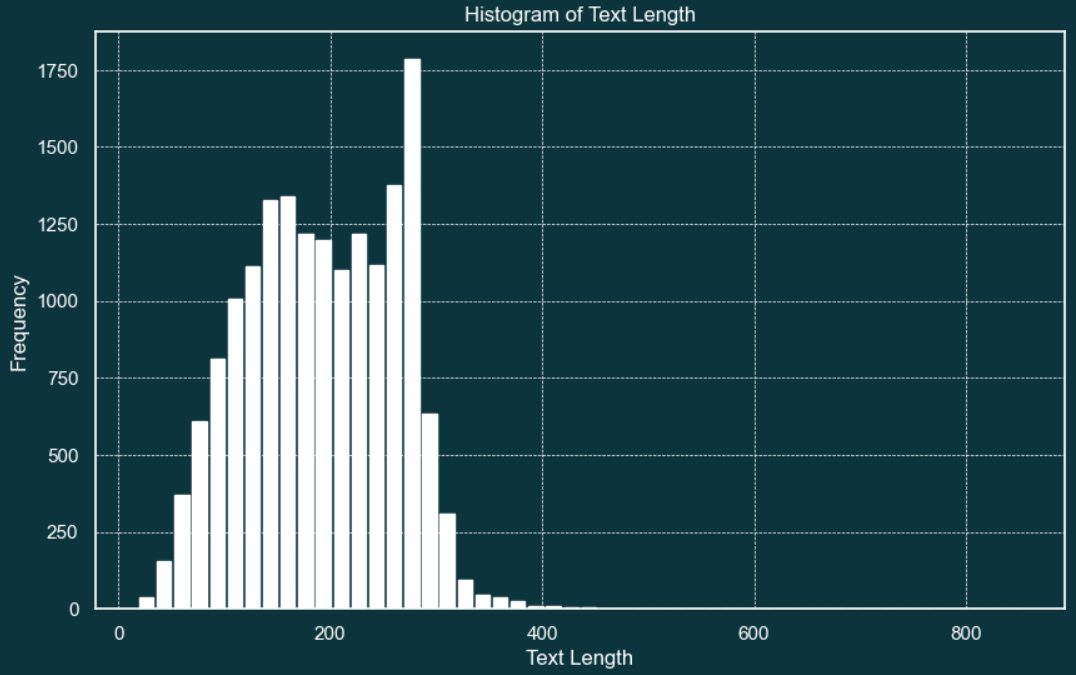
Analyzing data in a single language (English) simplifies the NLP processing. It reduces the complexity of language detection and translation, allowing for more focused analysis on English tweets. With a consistent language (English), model training and tuning become more efficient.

—----------------------------------------------------------------------------------------------------------------



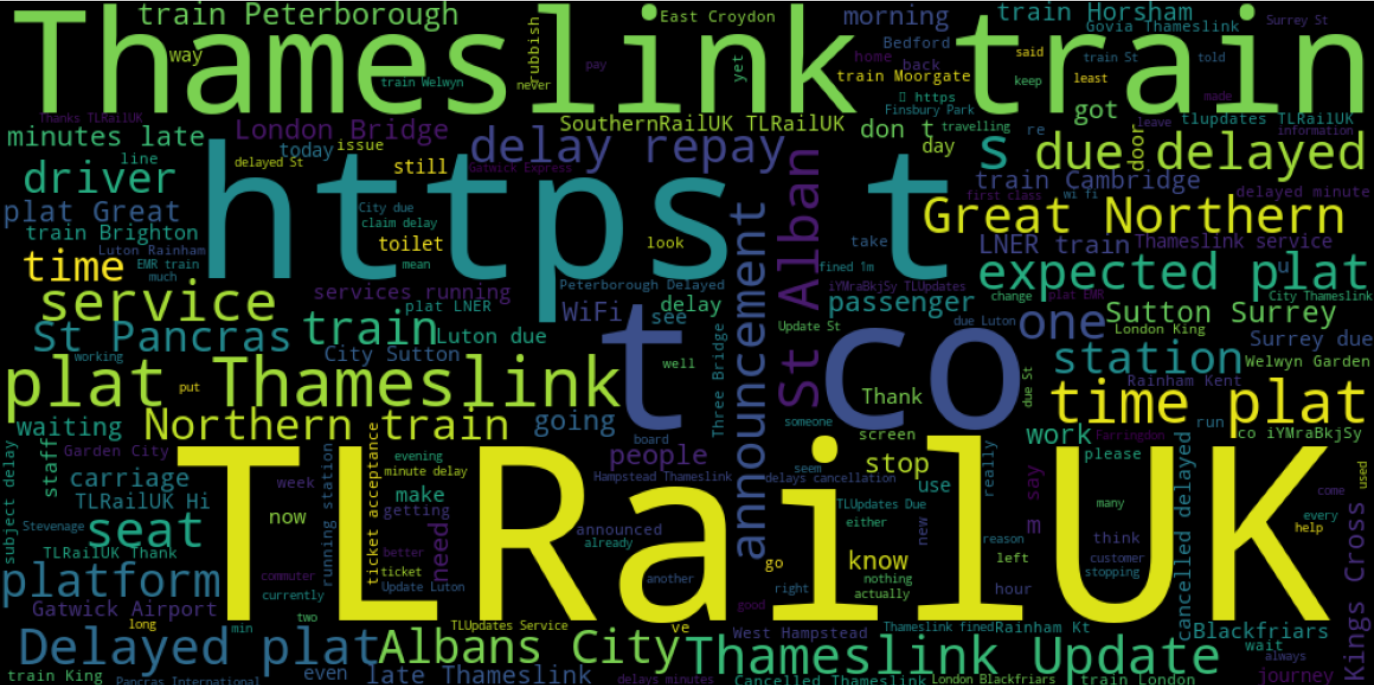
There is a positive linear correlation between the number of tweets in one day and the number of tweets in the next day. This means that an increase in tweets on a given day is associated with an increase in tweets the next day, and vice versa. Even though there are outliers, it indicates that we can use the data reliably.

—----------------------------------------------------------------------------------------------------------------



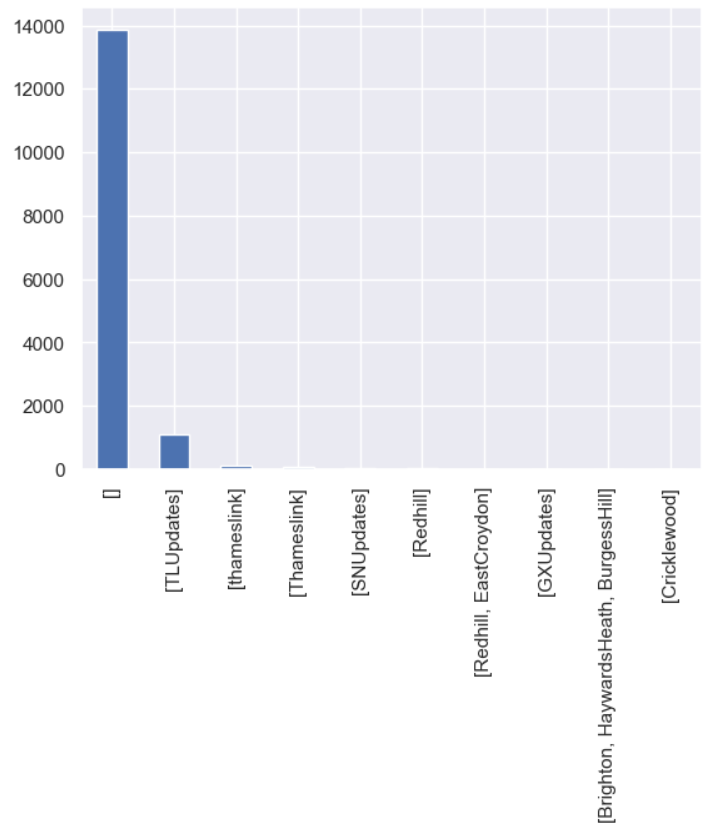
Length of tweets are controlled because longer tweets may contain more complex information and long tweets may need to be truncated or segmented, while very short ones might require additional context to be meaningful.

—----------------------------------------------------------------------------------------------------------------



Most used words are checked because unusual or unexpected frequent words may indicate issues in data collection or preprocessing. Also most frequent words helps in understanding the dominant topics, sentiments, or themes present in the tweets

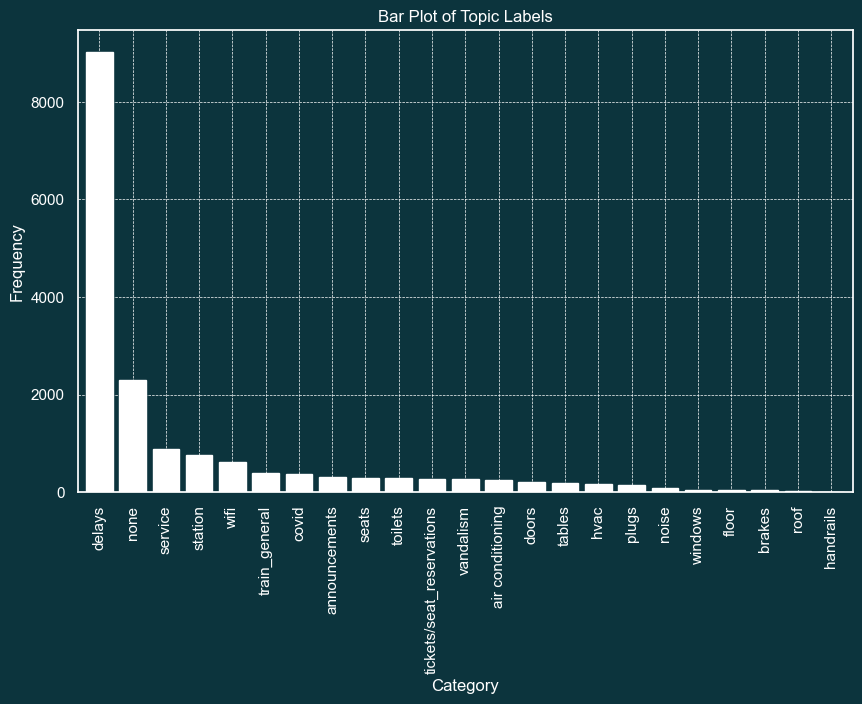
—----------------------------------------------------------------------------------------------------------------



Hashtags often encapsulate the central theme or topic of a tweet. Analyzing popular hashtags helps in understanding the problems or discussions within your dataset.

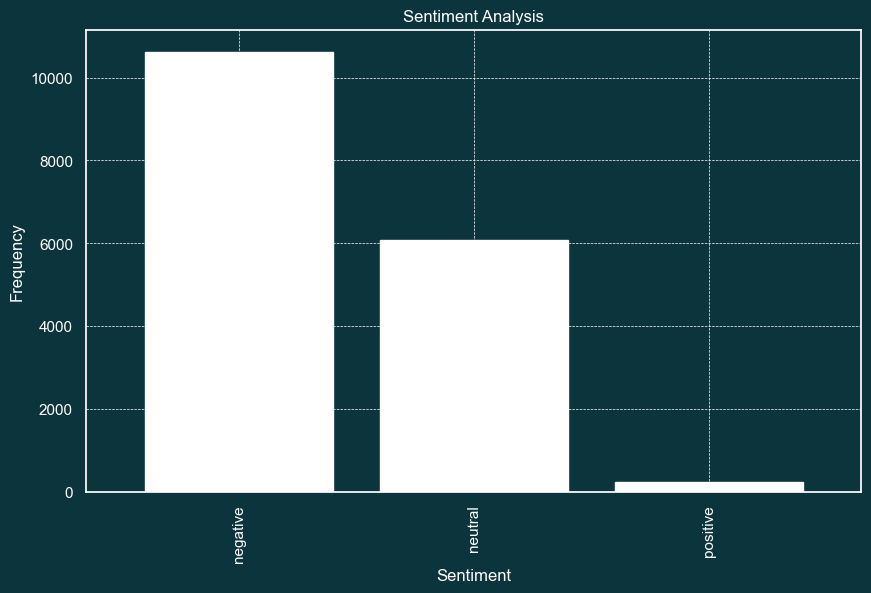
We have an empty hashtag because we still didn’t prepare the data.

—----------------------------------------------------------------------------------------------------------------



—----------------------------------------------------------------------------------------------------------------

Topic column would help us to detect the problem directly by categorisation of the tweet.



Analyzing sentiment in tweets allows you to gauge customer sentiment towards Thameslink's maintenance services. This feedback can be instrumental in identifying areas that require improvement or adjustment. Also it helps in quickly identifying tweets expressing strong negative sentiment. These may indicate urgent maintenance issues that need immediate attention.