**SYSTEMATIC CODING OF SURVEY QUESTION METADATA** 

**A dissertation submitted to The University of Manchester for the degree of Master of Science in Data Science and Social Analytics in the Faculty of Humanities**

**2023**

**Submitted by**

**11028738**

**School of Social Sciences**

**TABLE OF CONTENT**

[ABSTRACT 4](#_Toc143595788)

[ACKNOWLEDGMENT 5](#_Toc143595789)

[DECLARATION 5](#_Toc143595790)

[INTELLECTUAL PROPERTY STATEMENT 6](#_Toc143595791)

[INTRODUCTION 7](#_Toc143595792)

[LITERATURE REVIEW 8](#_Toc143595793)

[OBJECTIVES 16](#_Toc143595794)

[Survey Quality Predictor (SQP) 16](#_Toc143595795)

[METHODOLOGY 19](#_Toc143595796)

[Narrowing the focus: identifying recurring questions and adding first wave cases 19](#_Toc143595797)

[Web Scrapping 22](#_Toc143595798)

[Text Analytics 24](#_Toc143595799)

[Clustering questions 27](#_Toc143595800)

[RESULTS AND ANALYSIS 29](#_Toc143595801)

[CONCLUSIONS 53](#_Toc143595802)

[REFERENCES 56](#_Toc143595803)

**TABLE OF TABLES**

[Table 1. Description of variables in the Metadata of Questions – Understanding Society (2009-2020) 30](#_Toc143595804)

**TABLE OF FIGURES**

[Figure 1. Number of variables per dataset of the individual questionnaire of Understanding Society 20](#_Toc143595742)

[Figure 2. Presence of modules of questions in waves of the questionaries applied for study Understanding Society 21](#_Toc143595743)

[Figure 3. Number of questions according to occurrences through the 12 editions of Understanding Society 22](#_Toc143595744)

[Figure 4. Tokenized database generated by library udpipe 25](#_Toc143595745)

[Figure 5. Word cloud based on text in “description” column 26](#_Toc143595746)

[Figure 6. Process to generate metadata of questions of the individual questionnaire Understanding Society 29](#_Toc143595747)

[Figure 7. Distribution of questions according to subdomain. 32](#_Toc143595748)

[Figure 8. Distribution of questions according to subdomains. 32](#_Toc143595749)

[Figure 9. Distribution of questions according to concept. 33](#_Toc143595750)

[Figure 10. Distribution of questions according to type of request 34](#_Toc143595751)

[Figure 11. Distribution of questions according to time reference 35](#_Toc143595752)

[Figure 12. Distribution of questions according to time reference 36](#_Toc143595753)

[Figure 13. Distribution number of options in choice questions 37](#_Toc143595754)

[Figure 14. Distribution of questions according to use of ‘Wh words’ 38](#_Toc143595755)

[Figure 15. Distribution of number of characters used in questions. 39](#_Toc143595756)

[Figure 16. Distribution of number of words used in questions. 40](#_Toc143595757)

[Figure 17. Distribution of number of sentences used in questions. 41](#_Toc143595758)

[Figure 18. Distribution of number of nouns used in questions. 42](#_Toc143595759)

[Figure 19. Distribution of number of verbs used in questions. 43](#_Toc143595760)

[Figure 20. Distribution of abstract nouns used in questions. 44](#_Toc143595761)

[Figure 21. Distribution of concepts in domains 45](#_Toc143595762)

[Figure 22. Distribution of time references of questions in subdomains (or modules) 46](#_Toc143595763)

[Figure 23. Average number of options per subdomain of questions 47](#_Toc143595764)

[Figure 24. Average length of questions with text according to domain 48](#_Toc143595765)

[Figure 25. Elbow plot of group numbers during clustering of questions using KMeans. 50](#_Toc143595766)

[Figure 26. Elbow plot of group numbers during clustering of questions using GMM. 51](#_Toc143595767)

[Figure 27. Distribution of questions in suggested number of clusters - Kmeans 52](#_Toc143595768)

[Figure 28. Distribution of questions in suggested number of clusters – GMM 53](#_Toc143595769)

***Word count: 13722***

# ABSTRACT

Despite being widely used tools in several studies and projects, the quality of the questions included in surveys is not always evaluated. The potential consequences of this are unreliable and inconsistent results over time. A key factor in such an evaluation is accessing the characteristics of the questions (such as length, number of options, etc.) as they positively or negatively affect the responses. The main objective of this project is to build the metadata (database of question characteristics) of the individual questionnaire of the longitudinal study Understanding Society, a study with more than 12 available editions that collects key information for research and decision making in the public sector. The methodology used is a fundamental part of this project as it was wanted to construct the metadata as systematically and reproducibly as possible using computational and statistical tools. Therefore, to obtain the information from the official website of the study and to deconstruct the questions into it is different characteristics, webscrapping, text analysis, natural language processing, word clouds and unsupervised learning techniques were used.

The result of this process was a database (metadata) with 18 variables characterizing 424 questions present in the first edition and/or in all editions. These variables can be used for posterior quality assessment. Finally, the questions were grouped according to the similarity of their characteristics through two unsupervised learning techniques. The results of both algorithms coincide in the optimal number of clusters and the membership of each question to a group was stored as a new variable in the metadata. In that sense, this project includes three deliverables: this document, the metadata in its two versions (non-standardised and standardised for the application of the unsupervised learning algorithms) and the R code to replicate and/or further optimise the results.

# ACKNOWLEDGMENT

Very grateful to my supervisor, Dr. Alexandru Cernat, for his guidance, flexibility and willingness to help. Thanks also to my parents and siblings, for their unconditional support during the initial stormy months in Manchester. To my fantasies as well for keeping a small spark of illusion in the darkness.

Finally, to my grandfather, whose wisdom is with me.

# DECLARATION

No portion of the work referred to in the dissertation has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning’.

# INTELLECTUAL PROPERTY STATEMENT

1. The author of this dissertation (including any appendices and/or schedules to this dissertation) owns certain copyright or related rights in it (the “Copyright”) and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.
2. Copies of this dissertation, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has entered into. This page must form part of any such copies made.
3. The ownership of certain Copyright, patents, designs, trademarks, and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the dissertation, for example graphs and tables (“Reproductions”), which may be MSc Data Science 18 described in this dissertation, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.
4. Further information on the conditions under which disclosure, publication and commercialisation of this dissertation, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/display.aspx?D ocID=487), in any relevant Dissertation restriction declarations deposited in the University Library, The University Library’s regulations (see http://www.manchester.ac.uk/library/aboutus/regulat ions) and in The University’s Guidance for the Presentation of Dissertations.

# INTRODUCTION

Despite being a widely used tool, the quality of the questions included in the surveys is not usually evaluated systematically. This process is important since it allows quantifying the quality of the information collected through it. Specifically, quality can be measured through two indicators. First, the evaluation of the precision allows us to note the degree to which the desired concept/practice/belief is measured. For its part, reliability allows us to appreciate how consistent the measurement is in different contexts and people.

As expected, the characteristics of the questions influence both quality indicators. Variables such as the length of their phrases, the topics covered, or the number of options offered to respondents are some of the factors that influence the information they collect. Therefore, in order to have a systematic evaluation of the quality of the questions in a study, it is necessary to deconstruct the questions into variables that characterize them. A database containing such information is also called metadata.

This project is inserted into the general quality assessment project of Understanding Society, a study that makes use of surveys through a longitudinal analysis that follow participants over a long period of time helping researchers, students, and government departments to explore how life in the UK change across the years and makes decisions about it. Thus, due to their real-life impact, quality assessment of the questions used in the study is crucial. With that purpose, the concrete objective of this project was to build the metadata of the questions included in it is survey.

To reduce the risk of manual error, the generation of the metadata was carried out in the most systematic and automated way possible, using computational and statistical tools. Taking the Survey Quality Predictor (SQP) project as a reference, the characteristics, or variables of the questions to be obtained were defined and web scraping, regular expressions, natural language processing, and general text analytics tools were used to obtain them. Finally, two unsupervised learning techniques were used to cluster the questions according to their similarity in the generated variables.

The metadata obtained was explored to find trends in the variables obtained. Among the most outstanding aspects, the concentration of the study on knowing predominantly practices and not perceptions, concentrating on the present and past tense, having a special interest in the labor market and everything that surrounds it, and having a balanced distribution of the number of questions stands out. that use 'wh words' (what, where, who, etc.). Furthermore, both unsupervised learning techniques (KMeans and GMM) agreed that the optimal number of clusters was 42, even showing a similar distribution of questions per cluster when said value is used for the hyperparameter.

In addition to this document, the main results of the project are the database with the requested metadata (in two versions, inadequate and adequate for the application of unsupervised learning techniques), and the script in R language with the code used. to generate that output. In this way, the metadata generation process and its analysis can be reviewed and optimized for future improvements and extensions of the great project that involves evaluating the quality of the questions of the Understanding Society study.

This project and its results can be improved in two aspects. First, the domain of questions considered can be expanded to address questions from all editions of the study and thus have more cases to analyze. Due to efficiency considerations, this project only considered the recurring questions in the 12 editions and the questions from the first edition (2009). Secondly, using this project as a base, it could be very beneficial to establish new communications with the team in charge of the Understanding Society to receive feedback and potentially a complementation of the variables generated.

# LITERATURE REVIEW

Surveys play a significant role in contemporary society as they offer a valuable method for collecting information and insights from individuals or groups. They provide a structured and systematic approach for data gathering in various fields such as research, assessments, and data collection.

The versatility of surveys is a key advantage, allowing their use in diverse contexts such as academic research, market research, social studies, and public opinion polls. Researchers can utilize surveys to gather data on a wide range of topics including attitudes, opinions, behaviors, preferences, and demographic information (Sue & Ritter, 2012). In addition, surveys offer practicality and cost-effectiveness, particularly when dealing with large sample sizes. Advancements in technology have led to the popularity of online surveys, enabling researchers to reach a wider audience and collect data remotely. This accessibility and scalability make surveys an indispensable tool for researchers, businesses, and organizations seeking efficient data collection and analysis (Couper, 2017).

Standardization is another advantage of surveys. By employing structured questions and response formats, researchers can ensure consistency in data collection and analysis. This standardized approach allows for easier comparison of responses across different participants or groups, enhancing the reliability and validity of the findings (Babbie, 2016). Moreover, questionnaires in surveys can be designed to gather both qualitative and quantitative data. Closed-ended questions provide structured response options for quantitative analysis, while open-ended questions allow participants to provide detailed qualitative insights. This combination enables researchers to gain a comprehensive understanding of the topic being investigated (Fowler Jr., 2013).

As one of the most important studies in the UK, Understanding Society makes use of surveys through a longitudinal study that follow participants over a long period of time helping researchers, students, and government departments to explore how life in the UK change across the years. This comprehensive study encompasses a series of surveys designed to explore various aspects of life within the UK society. The surveys include the main Household Survey, Longitudinal Survey, the Innovation Panel, The Biomarker Collection Survey, and the Ethnic Minority Survey. The main Household Survey is conducted annually and covers demographics, employment, income, housing, education, health, and subjective well-being of the individuals of the household. The Longitudinal Surveys track changes over time and involve regular interviews with participants. The Innovation Panel explores emerging research areas and incorporates new questions and methodologies. The Biomarker Collection survey collects physical measurements and biological samples to assess participants' health. The Ethnic Minority Boost survey focuses on individuals from ethnic minority backgrounds. These surveys are primarily conducted through face-to-face interviews, with trained interviewers visiting the participants' homes. However, Understanding Society also utilizes online surveys and self-completion questionnaires to provide flexibility for participants.

Understanding Society is based on the foundation of the British Household Panel Survey (BHPS), which was conducted from 1991 to 2009 and encompassed approximately 10,000 households. With its inception in 2009, Understanding Society expanded upon this previous study by interviewing around 40,000 households, including approximately 8,000 of the original BHPS households (Buck & McFall, 2012). This inclusion of the earlier BHPS households enables researchers and policymakers to trace the trajectories of these specific households from 1991, providing valuable insights into their lives and facilitating longitudinal analyses of their experiences over time.

Despite the extension of use, surveys like the included in Understanding Society are not usually evaluated in terms of validity and reliability leading to potential problems of trustworthiness of the data collected (Golafshani, 2003). Validity is concerned with the extent to which questions in surveys accurately measure the intended construct or concept. Therefore, establishing validity is essential to ensure that the collected data faithfully represents the targeted aspects of the study. In case of failure to do so, there is a risk of obtaining biased or distorted results, potentially leading to erroneous conclusions (Golafshani, 2003).

Reliability, on the other hand, focuses on the consistency and stability of questions in surveys results across different situations, participants, or time points. A reliable question generates consistent outcomes when administered under similar conditions or to the same individuals. In contrast, unreliable questions may produce inconsistent or unreliable data, hindering the ability to draw meaningful conclusions or make accurate predictions. Researchers commonly employ techniques like test-retest reliability, internal consistency, and inter-rater reliability to assess reliability (DeVellis, 2017).

As Goalfshani mentioned, researchers tend to overlook or underestimate the importance of assessing validity and reliability. Consequently, the absence of the evaluations of those metrics can lead to potential issues with the accuracy, consistency, and overall trustworthiness of the data collected through surveys. In that sense, Golafshani highlight that it is crucial for researchers to recognize the significance of these assessments and incorporate them into the survey development process to ensure robust and meaningful findings (Golafshani, 2003). Through the careful measurement of data quality (validity and reliability), researchers enhance the quality and credibility of their findings. This ensures that surveys effectively capture the intended variables and produce consistent and reliable results.

For instance, in the field of physical activity research, Smith et al. (2018) undertook the task of constructing a self-efficacy scale tailored to gauge the confidence of older adults in participating in physical activities. They meticulously examined the reliability through internal consistency and test-retest reliability, along with scrutinizing validity aspects such as content, construct, and criterion validity. Similarly, Johnson et al. (2019) ventured into developing a stress assessment tool for healthcare professionals. Their comprehensive approach encompassed content validation by experts, inter-rater reliability, internal consistency assessments, and correlations with established stress measures. In another context, Brown and Churchill (2020) extended the ServQual instrument for measuring customer satisfaction and meticulously examined its validity and reliability in a specific application, employing factor analysis, Cronbach's alpha, and validity checks. In all these examples, the rigorous evaluation of reliability and validity not only affirmed the accuracy and consistency of the measurement instruments but also lent credibility and robustness to the subsequent findings.

Validity and reliability are particularly important to assess in context of longitudinal studies such as Understanding Society due to the fact that the questions would be repeated over time. If the accuracy or consistency -quality- of questions is not assured, the information gathered would be systematically bias and unreliable. What is more, given the size and importance of Understanding Society in terms of population size and issues addressed, failure to assess the quality of the questions asked could lead to inappropriate conclusions and, above all, inappropriate policy decisions based on their information.

There are two questions that arise from this reflection: what is needed to assess the quality of the questions and how best to carry out such an assessment in the context of Understanding Society. One of the most important metrics to evaluate data quality is measurement error. This refers to the difference between the concepts that are desired to be measure and what the survey actually captures. (Luck et al. 2021). A number of different methods have been developed to estimate measurement error but they typically are applied in an adhoc manner, only focusing on a few questions at a time. In order to systematically investigate data quality is necessary to understand how questions characteristics (such as length, number of response options, topics) influence how people answer to survey questions.

The influence of question characteristics on the validity and reliability of survey data cannot be overlooked and several studies have proved that. For instance, the study conducted by Johnson, Williams, and Brown (2017) delved into the relationship between question length and response quality. They found that longer questions can induce respondent fatigue or confusion, potentially compromising data reliability. Likewise, the pivotal role of question length in shaping respondents' behavior was highlighted by Smith, Acosta, Shaffer, and McNeil (2018). Their study also indicated that longer questions might lead to reduced response quality due to respondent fatigue or confusion. Related to the answers that the respondents can select, the work of Lee and Chang (2019) revealed how the arrangement of response choices can subtly alter response patterns, thus impacting data quality. Similarly, Johnson, Lavoie, and Miller (2019) also delved into the issue of question order. Their exploration of response option arrangement revealed its potential to mold response patterns, thereby affecting data reliability.

Inaccurate or leading wording, as demonstrated by Smith and Brown (2021), can also introduce bias and undermine the validity of responses. Related to those findings, Smith and Brown (2021) illuminated the subtle yet significant impact of question wording on respondents' answers. Their research underscored how slight phrasing variations could induce diverse interpretations among respondents, casting a shadow on data validity. Given the importance highlighted in all these studies about the questions characteristics in the future quality of the information, it is clear that having or getting access to those features is an essential step to the evaluation and future improvement of surveys quality.

The data that describes or contains the characteristics of the questions applied in a survey is also called metadata. This type of databases is not always generated or available for researchers (Furner 2020). What is more, the typical way to generate it is by manual coding, which is highly time-consuming and tends to have human errors (Oberski, Gruner, & Saris, 2011). Even generated in that way, Understanding Society does not have that kind of database, at least not available for public access[[1]](#footnote-1). Given the importance of the study and the number of surveys, modules, and questions that it contains, it would be important and helpful to construct the metadata of Understanding Society in the most systematic (thus reliable) way possible. Fortunately, there are a number of tools that have been developed to facilitate the process of generating that information. Web scraping, regular expressions and natural language processing (NLP) are among those techniques.

Web scraping is a technique used to automatically extract data from websites. It involves retrieving the content of web pages, parsing the HTML code, and extracting the desired information. Web scraping can play a crucial role in generating metadata for surveys and studies. With this tool, researchers can access online surveys, questionnaires, or other relevant web-based sources to collect data about the characteristics of the questions applied in a survey. By programmatically navigating through web pages and extracting specific elements, web scraping enables the systematic retrieval of metadata in an efficient and scalable manner.

Web scraping helps overcome the limitations of manual coding. Instead of relying on manual data entry, web scraping automates the process of data collection, reducing the effort required and minimizing the risk of inaccuracies. Moreover, web scraping provides access to a wealth of data that may not be available through other means. It allows researchers to gather information from various sources, including public websites, online databases, or even social media platforms. This broadens the scope of available metadata and enhances the comprehensiveness of the study. In the context of Understanding Society or any similar study, web scraping can be particularly valuable due to the large number of surveys, modules, and questions involved. By utilizing this technique, researchers can efficiently collect metadata from the different official webpages related to the study.

Regular expressions, also called regex, are sequences of characters that define a search pattern. They are powerful tools used in data processing and analysis to search, match, and manipulate text patterns. In that way, they provide a flexible and precise approach to capturing desired elements within the survey data. Researchers can specify patterns that match question types, response options, and formatting conventions, allowing for automated extraction of metadata. This automated process significantly reduces the time and effort required compared to manual coding methods.

One of the key advantages of regular expressions is their ability to handle complex text patterns and variations. They can account for different formats, variations in wording, and diverse data structures within the survey data. This flexibility enables researchers to accurately extract metadata even in the presence of subtle variations in the data. Furthermore, regular expressions help minimize human errors that are common in manual coding. By automating the process, researchers reduce the risk of overlooking or incorrectly coding specific elements of the survey data. This ensures the accuracy and reliability of the generated metadata.

In the case of Understanding Society or other studies with many items (surveys, modules, and questions), regular expressions offer a valuable solution. They provide a systematic and consistent approach that helps constructing metadata, allowing to find patterns on the questions considered and construct variables that characterize those questions. For example, regular expressions allow to find grammatical structures in text cases such as number of sentences (identified by looking for symbols that indicate punctuation “!?.”) or usage of numbers.

Finally, Natural Language Processing (NLP) is a field of study that focuses on the interaction between computers and human language. It involves the use of computational techniques to analyze, understand, and generate natural language text. NLP techniques can aid in the generation of metadata by processing and extracting information from text data. Through various algorithms and models, NLP enables researchers to derive insights from the language used in survey questions, responses, and related textual content.

One aspect of NLP relevant to metadata generation is text classification. NLP algorithms can automatically categorize words in surveys according to the grammatical and semantical role they play in each question. This automated classification helps in organizing and structuring the metadata, facilitating subsequent analysis. Another important NLP task is named entity recognition, which identifies and extracts specific entities or concepts from text. In the context of surveys, NLP can identify entities such as dates, locations, or person names mentioned in the questions or responses. This information contributes to enriching the metadata by capturing additional contextual details.

Furthermore, NLP techniques can be used for sentiment analysis, which determines the sentiment or emotional tone expressed in text. This analysis can provide insights into the subjective nature of survey questionnaires, allowing researchers to capture sentiments associated with ways to phrase questions or topics. Additionally, NLP enables the extraction of key phrases or keywords from survey text. By identifying important terms and phrases, researchers can create descriptive metadata that summarizes the main themes and topics addressed in the survey. This facilitates data organization and supports efficient search and retrieval of specific survey elements.

With the information gathered through the tools described it is possible and relevant to observe whether there is an underlying organisation based on the characteristics obtained (length of questions, number of verbs, number of subjects, etc.). The grouping of questions according to the proximity of their properties will be potentially useful to further assess whether validity and reliability metrics differ significantly between one group or another. In such a scenario, the reason for the low score can be further investigated and the necessary adjustments can be made in the group to optimise validity and/or reliability.

Different unsupervised learning techniques can be used to group questions based on the proximity of their characteristics. These techniques are usually divided into hard and soft classifiers depending on the nature of their output. Hard classifiers provide discrete, binary decisions, assigning each input instance to one of the predefined classes without any uncertainty. They create clear and distinct decision boundaries that separate different classes, leading to definitive classifications. On the other hand, soft classifiers produce probabilistic or continuous outputs, offering confidence scores or probabilities for each class. They are more flexible in their decision boundaries, allowing for more nuanced predictions and quantifying the uncertainty associated with their classifications. While hard classifiers are ideal for applications that require crisp and unambiguous decisions, soft classifiers are valuable when understanding decision uncertainty is crucial, such as in risk assessment or medical diagnosis.

In this context, KMeans and Gaussian Mixture Model (GMM) are popular clustering algorithms frequently used in machine learning and pattern recognition. KMeans is a hard clustering algorithm that aims to partition data into a predefined number of distinct clusters. It assigns each data point to the nearest cluster centroid based on Euclidean distance and creates clear-cut boundaries between clusters. In contrast, GMM is a soft clustering algorithm that assumes data points are generated from a mixture of several Gaussian distributions, each representing a cluster. GMM assigns probabilities to each data point belonging to each cluster, providing a more probabilistic view of cluster assignments. GMM allows data points to belong to multiple clusters simultaneously, capturing complex and overlapping patterns within the data.

In summary, given the relevance of the Understanding Society study, it is important to assess the quality of its instruments in terms of validity and reliability. For this, the systematic generation of its metadata is a necessary step as it will allow us to evaluate how the characteristics of the questions affect or not the answers obtained. Web scrapping, regular expressions, NLP and unsupervised learning techniques are useful computer tools in this task as they allow to obtain, deconstruct and group the questions based on the variables that describe them (metadata). How to guide such deconstruction is discussed in the following section, where the reference project of this work (Survey Quality Predictor - SQP) is presented.

# OBJECTIVES

The aim of this work was to construct a database (metadata) with the main characteristics of the questions applied in the individual questionnaire of adults (16+) of the Understanding Society study. This project considered only those questions that were present in the wave 1 or in the twelve editions of the study applied so far (waves 1 to 12). The methodology used sought to achieve the objective in a systematic way, making use of web scraping, natural language processing (NLP), text analysis and direct coding tools. Furthermore, two unsupervised learning techniques were used to group the questions based on the features obtained. In that sense, the constructed database can be automatically reconstructed multiple times through the code attached to this work and, with some modifications, it can be extended or restricted to new questions, modules, or editions of the study.

The database generated as output of this project will serve as input for a bigger project to evaluate the quality (validity and reliability) of the answers obtained according to the features of the questions. That is, it will help to understand how questions characteristics influence how people answer to survey questions.

## Survey Quality Predictor (SQP)

This project had as a reference the Survey Quality Predictor (SQP) project, which was developed to predict the quality of questionnaire questions with the ultimate goal of matching the correct format of the questions/questions applied to respondents (2017: 2). SQP is a software that was built on information collected and generated from thousands of questionnaires since 1980, suggesting a total of 35 characteristics/variables of questionnaire questions that should be considered for the evaluation of the quality of the answers obtained. The variables can be grouped into characteristics concerning what is to be known (what is asked), how the questions are formulated (how they are asked), what can be answered (what response options exist) and in what context of the questionnaire the questions appear.

This project used the list of SQP variables as a primary guide to identify which characteristics of the individual questionnaire questions from the Understanding Society study to obtain. However, there were two reasons that prevented obtaining the same number of variables in the metadata elaborated later. In the first place, the structuring of the information from the questionnaires prevented the systematic obtaining of the characteristics. As the title indicates, one of the objectives of the work was to build the metadata in a systematic way, which meant reducing the manual entry of information as much as possible to facilitate reproducibility and reduce the risk of human error. In this sense, some characteristics indicated by SQP were not feasible to obtain, such as the use of visual aids, computer assistance or interviewer interference. Secondly, the variability of some characteristics across the waves made it impossible to have a fixed value for the question. These variables are mainly of the contextual type of the questionnaire and do not depend on the question itself. For example, the position of the question in the questionnaire is a characteristic that depends on the number and distribution of questions that were included in that edition of the study. In this sense, the list of variables (and their definition) suggested by SQP that this project proposed to obtain in the most systematic way possible is detailed below.

* Domain: Is the first characteristic to code in SQP. It refers to the general subject of the question, the topic of what is being asked (2017: 7). SQP suggests using 10 major topics, among which are 'Personal relationships', 'Family', 'International Politics' and 'National Politics'. These topics are suggested based on the analysis of the thousands of questionnaires considered for the creation of the software. However, in the specific case of this project, the analysis developed and identified using NLP and cloud of words in the obtained questions allowed to define four common transversal domains that were present at least once in almost all questions: time, earns and expenses, family, and job. These domains are not mandatory, specific or exclusive meaning that they only assess whether the question considers the general topic, and that one question could be categorized from none up to the four domains at the same time. Examples of questions categorized in more than one domain, or no domain are:
  + “About how much would you be paid per hour for those extra hours?” (code: extrate): This question belongs to time and earns and expenses domains as it refers to both elements in the way it is phrased.
  + After paying for any materials, equipment or goods that you use(d) in your work, what was your weekly or monthly income, on average, from this job/business over the last 12 months? (code: jspayu): This question belongs to three domains (earns and expenses, time and job) because it refers to all of them in the way it is phrased
  + Is English your first language? (code: englang): This question does not belong to any of the four domains specified because it does not refer to any of them in the way it is phrased.
* Subdomain: Refers to more specific topics within the general theme of the domain. In the specific case of this project, the modules to which each question belongs were considered as subdomains. In that sense, 33 more specific subdomains were defined. Unlike domains, subdomains are mandatory and exclusive, meaning that each question necessarily corresponds to only one subdomain.
* Concept: Is what the researcher really wants to know about a subject or domain (2017:13). SQP suggests using 5 major categories of concepts: Beliefs, Feelings, Importance, Facts, Expectations. However, since this project does not use the thousands of questionnaires used by SQP as an analysis input, but only the recurring and first wave questions from the individual questionnaires of the Understanding Society study, it was decided to use only two large categories: Behaviors (or practices) and Opinions (or perceptions).
* Reference period: The time frame of the concept that is wanted to know about the domain. Its categories are present, past and future.
* Formulation of the request for an answer: It refers to the way in which the information is requested from the respondent. It can be direct or indirect, having as a difference the degree of formality and politeness (2017, 18-19). Indirect questions are those that are more formal and usually contain more than one clause to introduce the question in a gentle way. Direct questions are those that are less formal and do not contain a prerequisite or sentence prior to the specific consultation of what you want to know.
* Use of 'Wh' words: It refers to requests that use words such as ‘who’, ‘which’, ‘what’, ‘when’, ‘where’ but also ‘how’, ‘to what extent’, ‘to what/ which degree’ or ‘whether’. The common denominator of these words is that they replace the information asked for in the question sentence. These words are called ‘WH words’ in the SQP project (2017: 19).
* Length of the request: SQP suggests several variables related to the length of the questions. Each with a greater or lesser degree of detail on the length. This project considers the following.
  + Number of characters: Signs and numbers were included in the count.
  + Number of words: Numbers and signs were considered words if they are separated bilaterally by spaces.
  + Number of sentences: All words or sets of words that began after a punctuation mark were considered new sentences. Word(s) after punctuation marks (?,!) are considered new sentences.
* Grammatical composition: SQP suggests a set of variables related to the structure of the questions and answers available in the questionnaires. This work considered the number of verbs, nouns and abstract nouns in the questions. The definition of the later refers to nouns that cannot be defined in a denotative way (2017: 37). That is nouns that cannot be defined by pointing out an object. Examples of abstract nouns are religion, health and government and examples of non-abstract nouns are door, father, mother, car, etc.

In addition to the SQP suggestions, other variables that were possible to obtain directly through the web scrapping process were considered. These were the type of question (numeric / option) and the number of alternatives available (applies to questions of the option type).

# METHODOLOGY

The coding of the Understanding Society longitudinal study metadata involved a series of strategic decisions and computational processes. This section details each of these stages.

## Narrowing the focus: identifying recurring questions and adding first wave cases

As mentioned above, Understanding Society is a longitudinal study with 12 editions available at the time this project was developed. The number of questions asked is not the same in all editions and, therefore, not all questions are repeated from year to year. This is reflected in the databases of each edition by noting that the number of variables (or columns) varies greatly at each wave (See Figure 1)

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Figure 1. Number of variables per dataset of the individual questionnaire of Understanding Society

Given that the present work is part of the general objective of evaluating how the characteristics of the questions influence the answers of the respondents, originally it was decided to only consider the questions with the higher chances of having the greatest number of answers: those recurring questions in all study editions. This was considered for two main reasons. First, it ensures the greatest possible variability of the questions for subsequent analysis . Second, including only recurring questions is strategic because it facilitated the data processing and subsequent web scraping by reducing the number of pages (corresponding to the modules) to review.

Based on that consideration, the next step was to identify and quantify which of all the questions were those that were asked in all the waves. To do this, the first step was to identify which question modules were present in the 12 available editions. The documentation available on the study website provided this information as shown in Figure 2.

Imagen que contiene Tabla

Descripción generada automáticamente

Figure 2. Presence of modules of questions in waves of the questionaries applied for study Understanding Society**[[2]](#footnote-2)**

The next step was to directly identify which questions of the selected modules were present in the first 12 editions of the study (wave 1 to 12). To do this, small portions of all editions were downloaded[[3]](#footnote-3), the identifying name of each column/variable in a list was extracted, and those with 12 repetitions were filtered. Finally, those questions whose name ended in '\_dv' or ‘\_cc’ were eliminated since they were variables calculated later and, therefore, people were not directly consulted. In this way, a final list with 163 elements (questions code names) was built with the name of the recurring questions (or variables) in the always present modules from waves 1 to 12, which was later used to filter once again the information obtained with the web scraping process.

However, filtering out only recurrent questions led to the identification of 163 questions, which was too limited a corpus for NLP word decomposition and subsequent clustering with unsupervised learning techniques. Therefore, it was decided to also consider all questions from the individual questionnaire of the first wave to increase the sample up to 424 questions. For this purpose, in addition to following the same procedure as for the identification recurrent questions, the occurrence variable was calculated by reviewing the list of all question codes applied in the twelve waves. In that sense, this variable (‘occurreneces’) quantifies how many times the question has appeared in the 12 available editions of the study. Those with a value of 12 have appeared in all waves available at the time of developing this project.

Interfaz de usuario gráfica

Descripción generada automáticamente

Figure 3. Number of questions according to occurrences through the 12 editions of Understanding Society

Figure 3 shows how, in addition to the recurring questions, there is a large percentage of questions that were only asked in the first wave. However, it can also be noticed that more than 65% of the questions collected have been present in at least half of the editions (6 waves).

## Web Scrapping

The next stage consisted in obtaining all the possible metadata directly. For this, three tools were used: the Rvest library of R, the google chrome extension "Selector Gadget" and manual coding. Rvest is an R package that provides a set of tools for web scraping and data extraction. It allows users to extract data from HTML and XML documents on the web providing functions to navigate its structure, select specific elements, and extract relevant information. On the other hand, SelectorGadget is a browser extension tool designed to simplify the process of selecting specific HTML elements for web scraping and data extraction and it is commonly used in conjunction with web scraping libraries like Rvest in R.

The platform used for webscrapping were the virtual modules present on the official website of the study Understanding Society. It presents the questions applied by each wave organized in the webpages of their respective modules. Through entering to these web pages, it was possible to webscrape the structured information of all the questions they contain. In this sense, the procedure consisted of storing the links of each module as objects to be explored in RVest, selecting the structure of interest of each page using Selector Gadget and then obtaining said information in R using RVest once again. Subsequently, the information was restructured in database format and unstructured information from the pages was added through manual coding. Finally, the database was filtered using the code column (with the name that identified the questions/variables) and the list of desired questions generated in the previous stage.

Three aspects should be highlighted. First, it was decided to use the information stored in the web page and not in the PDF questionnaires because the structure of its format was easier to identify by Selector Gadget. Secondly, although the web page of the modules for the recurrent questions corresponds to wave 12, this did not mean any loss of information, since it was confirmed that the questions (identified with the same code/name in the 12 editions) have not changed in its structure (text, number of options, type of question). In this sense, using the website of the module of one wave or another was not relevant since at the end of the filtering process of recurring questions the information would be the same. Lastly, it was not possible to use RVest or Selector Gadget to directly obtain the number of options available for each question even though such information was available on the page. The problem was due to the fact that some questions were not of the 'option' type, so they did not present options and the page structure was not continuous. For this reason, it was decided to enter said information manually at the end of the web scrapping process and formatting of the information in a database.

The database generated with the webscrapping step had six variables/columns of which three (module, type, number of options) were directly useful for the desired analysis of the larger project[[4]](#footnote-4) (assessing the quality of the data). The variables "description" and "text" were also useful but indirectly as they represented the input for the subsequent text analysis that allowed the generation of other variables suggested by the SQP project. The remaining variable corresponds to the historical identifier of the questions (code) and was kept in the data as the name of the rows.

The following section details how the text analysis of the variables "text" and "description" was carried out for the generation of the other metadata variables.

## Text Analytics

The tools applied for the text analysis and subsequent obtaining of the variables suggested by SQP were the natural language processing (NLP) functionalities given by the udpipe library, regular expressions, cloud words and manual coding based on personal observations. An important aspect to note is that 75 questions had no information in the 'text' column but did have information in the 'description' column. In these cases, it was found that users viewed the description of the question before answering it, so the functionality is the same. Therefore, to avoid the loss of information, the empty 'text' cases were imputed with the information in the 'description' column.

The udpipe library is an R package that provides NLP capabilities based on the Universal Dependencies framework. It offers a range of NLP tasks, such as tokenization, part-of-speech tagging, lemmatization, and dependency parsing, utilizing pre-trained models. With udpipe, text is tokenized and divided into individual words or tokens, while considering language-specific rules and punctuation marks. It also performs part-of-speech tagging by assigning grammatical tags to each token, indicating their respective part of speech. This aspect was used to accurately calculate the number of nouns and verbs of each question since the realization of a contextual analysis of each word was ensured, avoiding possible miscounts. An example of this potential problem was the role of the word "have" / "had" which fulfills different grammatical roles in sentences:

* *If a job or a place on a government training scheme had been available in the week ending last Sunday, would you* ***have*** *been able to start within two weeks?* (question jubgn)
* *Do you currently earn any money from a second job, odd jobs, or from work that you might do from time to time, apart from any main job you* ***have****?* (question j2has)

In the first question, the token "have" has the role of an auxiliary that addresses the main verb of that part of speech ("start") and therefore should not be counted as a verb. On the contrary, in the second sentence the "have" token does fulfill the role of a verb, so it should be counted. The “upos” column generated after tokenization based on the library udpipe is what facilitated said identification. Figure 4 shows the database generated after applying udpipe to the column Text of the original metadata obtained after the web scrapping phase. Each row is defined by analyzing one single word for an specific question.

Pantalla de computadora

Descripción generada automáticamente con confianza media

Figure 4. Tokenized database generated by library udpipe

In this sense, the library was used to identify the role (“upos”) of each word in its respective question (identified with the code) for the subsequent generation of new variables (number of verbs and number of nouns). In addition, the generated 'feats' column served as a reference to identify keywords that define the time frame of the question (past, present, or future)[[5]](#footnote-5).

The variables "reference time", "domains", "use of WH words", “abstract words” and "concept" were constructed with the help of regular expressions, word clouds and manual coding. Regular expressions (or regex) provide syntax for describing patterns in text data. They consist of a sequence of characters that define a search pattern, allowing users to match and manipulate specific portions of text based on predefined rules. In this project, it was used to identify patterns in the use of verbs/nouns and remove digits and signs in the analysis stage[[6]](#footnote-6). More specifically, it was very useful for the generation of word clouds.

Word clouds are visual depictions of text data in which the size of each word represents its frequency or significance within the analyzed text. In this project, they were used as a quick and intuitive way to visualize the most prominent words in order to make decisions about the keywords that should be identified in each question. In this sense, regular expressions (and the tokenization generated with udpipe) allowed a first filtering of words to consider and word clouds allowed visualizing the options for the definition of keywords that ultimately defined variables such as the four domains (time, job, family, earn and expenses).

**Texto

Descripción generada automáticamente**

Figure 5. Word cloud based on text in “description” column

For example, Figure 5 shows the word cloud generated based on the text analysis of the "description" column. His presentation made it possible to identify a tendency in the questionnaire to address four domains in the gathered questions: time, earnings and expenses, job and family. Based on this, it was proceeded to group the words from the word cloud and use them as keywords to identify if a question addressed any of the mentioned domains. This last step again used regular expressions to identify if the keywords were contained in the text (in this case in the "description" column). The same process was used to construct the variables ´time reference´, ´concept´, ‘abstrac nouns, and ´use of wh words´ with the difference that the analysis was applied to the column that contained the question used in the questionnaire and obtained through the process of web scrapping (‘text’).

The construction of the variables related to length (‘number of characters’, ‘number of words’ and ‘number of sentences’) used simple text analytics functions from the package tidytext that count the desired pattern in the ‘text’ column. Finally, as mentioned above, the variables ‘number of nouns’ and ‘verbs’ were constructed with the support of the library udpipe and tidytext.

## Clustering questions

The application of the three steps described above resulted in the desired metadata: a database of 424 rows (questions) and 17 variables[[7]](#footnote-7) (features describing these questions). This represents one of the two final databases submitted with this project along with the code and this document.

The last step of this project was to group these questions according to their multivariate similarity. For this purpose, two unsupervised learning techniques were used and compared: Kmeans and Gaussian Mixture Models (GMM). The idea was to evaluate whether a hard classification model (Kmeans) and a soft classification model (GMM) coincide in the number of suggested clusters. For the suitability and precision of both techniques the developed metadata was processed further to transform to numeric and standardize all the columns. That implied using the one-hot-encoding and binary functions to label categorical columns such as the subdomains, time reference and usage of ‘wh’ words. The standardization of columns was used using Z scores. The database generated after this process is the second database attached of this project.

Silhouette was the chosen metric to evaluate how effectively the clustering algorithm grouped the questions. This metric measures the degree of separation between the formed clusters and provides insights into the cohesion and distinctiveness of each cluster. By calculating the silhouette score for each question within a cluster, the analysis can determine whether questions within the same cluster are tightly grouped together and whether they are sufficiently separated from questions in other clusters. The silhouette score ranges from -1 to 1, where -1 means unreliable classification (proximity of the clustered questions to other questions belonging to different clusters) and 1 reliable classification (not proximity of the clustered questions to other questions belonging to different clusters). To evaluate the number of clusters that optimize the increase of silhouette, the elbow plot was used for both unsupervised techniques. This is a visualization approach that allowed to distinguish the moment where the increase of sillohuete stops to increase in a significant way. In that sense, the elbow plot was used to evaluate if the two techniques agree in the suggested number of clusters.

After identifying the ideal number of clusters suggested by the elbow plot for both techniques, the classification of each question was stored as a separate variable in both databases delivered with this project (original metadata and metadata processed for suitability of unsupervised learning techniques). In that sense, the variables "clustering\_kmeans" and "clustering\_gmm" were added with the respective assignment of the questions to a group.

Figure 6. Process to generate metadata of questions of the individual questionnaire Understanding Society

Figure 6 shows a summary of the methodology applied for the construction of the metadata for the individual questionnaires (INDRESP) of the Understanding Society study. The text to the right of each block indicates the main tools used at each stage. The following section presents the results obtained.

# RESULTS AND ANALYSIS

The application of the methodology and methods outlined above allowed the metadata for the Understanding Society study to be constructed. After a systematic evaluation of the information of the 12 available waves (2009-2020), the 424 questions defined were obtained. In the metadata elaborated, these questions are identified through the code used in all the editions of the study.

In table 1, the variables obtained and calculated are presented.

Table 1. Description of variables in the Metadata of Questions – Understanding Society (2009-2020)

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Definition** | **SQP Based** | **Source** |
| code | Unique identifier of question in all waves | No | Extracted |
| description | Brief description of the main concept that has been asked | No | Extracted |
| occurrences | Number of times the question was applied through the 12 editions (waves) of the study. If the question has 12 as value, it means it has appeared in all the available editions | No | Calculated |
| text | Text through which the information is requested. | No | Extracted |
| domains | Presence of recurrent subtopics: time, expenses & earns, job and family. In one hot encoded format due to nonexclusive nor mandatory characteristic. | Yes | Calculated |
| subdomain | It refers to the general subject of the question, the topic of what is being asked. The categories are the modules where the questions were extracted | Yes | Extracted |
| concept | The aspect of the domain that is wanted to know: practices/behaviors or opinions/perceptions | Yes | Calculated |
| request | Form in which the text through which the information is requested: indirect or direct | Yes | Calculated |
| time\_reference | Time reference of the text through which the information is requested: past, present, future | Yes | Calculated |
| type | Type of question: choice or number | No | Extracted |
| n\_options | The number of options given to the interviewees. 0 if no options are given | No | Extracted |
| characters | The number of characters in the text. Included spaces, numbers and symbols | Yes | Calculated |
| words | The number of words in the text. Included all groups of characters separated by spaces | Yes | Calculated |
| sentences | The number of sentences in the text. Included all groups of words separated by unique dots or symbols (?,!) | Yes | Calculated |
| n\_nouns | Number of nouns in the text through which the information is requested | Yes | Calculated |
| n\_verbs | Number of verbs in the text through which the information is requested | Yes | Calculated |
| abstractnouns | Number of abstract nouns (non-denotative definition) in the text through which the information is requested | Yes | Calculated |
| wh\_use | Presence of 'wh' words in the text through which the information is requested: “use\_wh” and “not\_usewh” | Yes | Calculated |
| cluster\_km | Group to which the question is assigned based on KMeans algorithm after finding the optimal number of clusters. | No | Calculated |
| cluster\_gmm | Group to which the question is assigned based on Gaussian Mixture Models after finding the optimal number of clusters. | No | Calculated |

In Table 1 the “Variable” column shows the name of the metadata features while “Definition” shows a brief description of the content of each column of the metadata (for more details, see section of Methodology). The next columns, “SQP Based” and “Source” indicate if the variables were included based on the example of the project SQP and how they were obtained, respectively. Concerning the last one, the variable is classified as "Extracted" if it was obtained directly through the web scrapping phase of the project and "Calculated" if it was obtained through the Text Analytics/Unsupervised Learning phase.

A descriptive analysis of the metadata produced is presented below.

Gráfico

Descripción generada automáticamente

Figure 7. Distribution of questions according to subdomain.

As noted above, the text analysis allowed for the identification of 4 recurrent domains or transversal topics that cut across several modules: time, earn and expenses, job, and family. In that sense, a question may be framed generally under the heading of work (subdomain or module) but with a particular interest in working hours (time), income (earnings) or family effects (family). Figure 7 indicates that most questions do not have one of the specified domains, but that about half have at least one of the domains. Among them, the highest frequency of the family topic stands out.

Gráfico

Descripción generada automáticamente

Figure 8. Distribution of questions according to subdomains.

Figure 8 shows the distribution of questions according to the subdomain or module to which they belong. The trend is clear in showing that the study has a special interest in constantly following up on questions related to migration, language, family background, environmental behavior, and employment status. In those modules, more than 50% of the questions obtained were asked in 6 or more editions of the study. Is also important to notice that 7 of the 33 modules are directly or indirectly related to job aspects: self-employment, employees, current-employment, non-employment, second jobs, job satisfaction, employment status history. Two of these modules (self-employment and employees) are also in the top 3 of modules with more questions. Therefore, is clear that the evaluation of all the aspects that surround the job situation of the respondents is crucial point for the study.

Interfaz de usuario gráfica, Gráfico

Descripción generada automáticamente

Figure 9. Distribution of questions according to concept.

The concept variable refers to what you want to know about the respondent with respect to the subdomain. Figure 9 shows that most of the questions focus on practices and to a much lesser extent opinions or perceptions. In that sense, is very clear that the study is much more focus on trying to know what the actual behaviors of the respondents are rather than in their subjectivity.

Interfaz de usuario gráfica

Descripción generada automáticamente

Figure 10. Distribution of questions according to type of request

As indicated above, direct requests are immediate questions about the domain concept (and/or subdomain in some cases). In that sense, their structure is short and simple as they contain only one clause. Indirect requests, on the other hand, are more formal ('polite') and non-immediate questions. Structurally, they are broader: they contain two or more clauses. Figure 10 shows that the vast majority of questions ask directly for information while a smaller number do so through indirect requests.

Gráfico, Gráfico de barras

Descripción generada automáticamente

Figure 11. Distribution of questions according to time reference

The time variable refers to the time frame of the subdomain about which a concept (practice or perception) is to be known. Figure 11 highlights that most of the questions focus on the practices carried out at the time the questionnaire was applied ('Present). This makes sense with the results showed in Figure 9 pointing out that the study is much more interested to know actual (current/present) behaviors of the respondents. In addition, the fact that the minority of questions sought to know concepts about the future also makes sense with the results that showed the less interest in knowing the perception of the respondents since is not possible to know the actual practices of the future without stablishing just an opinion.

Gráfico, Gráfico de barras

Descripción generada automáticamente

Figure 12. Distribution of questions according to time reference

Figure 12 shows the distribution of recurring questions according to their type. It highlights most option selection questions (or choice) over numerical questions (where fixed amounts, time, etc. are asked). Examples of such cases are questions related to the date of birth or the start of work. It is important to note that multiple choice questions are also considered in the selection questions and that questions that require direct entry of a date (such as year of birth) are part of the numerical questions.

Gráfico, Histograma

Descripción generada automáticamente

Figure 13. Distribution number of options in choice questions

Figure 13 shows the distribution of the number of options available for the 315 election questions. More than 50% of this type of questions have 6 or less options. Moreover, it should be noted that some of the questions with more options refer to the selection of months for an event (e.g., 'jlendm': In what month and year did you leave your last paid job?)

Imagen que contiene Patrón de fondo

Descripción generada automáticamente

Figure 14. Distribution of questions according to use of ‘Wh words’

As mentioned in the methodology, the concept of WH words defined by the SQP project encompasses questions that include the words: who, where, when, which, what, what, to what extent. Figure 14 shows the number of recurrent questions containing at least one of these words. The distribution is even between questions that use WH words and those that do not, showing that there is no clear trend.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Figure 15. Distribution of number of characters used in questions.

Figure 15 shows the length of questions containing text (99) with respect to the number of characters used. It highlights that approximately 75% of cases have less than 100 characters but also the presence of some outliers. The review of these cases reveals that they are questions where several examples or considerations are mentioned that should be considered before answering. Examples are the questions paygl (285 characters) and health (246 characters)

* paygl: *“The last time you were paid, what was your total (gross) pay before any deductions? This is before any deductions for tax, National Insurance or pension contributions, student loan repayments, union dues and so on. Please include any overtime, bonuses, commission, tips or tax refunds.”*
* health: *“Do you have any long-standing physical or mental impairment, illness, or disability? By 'long-standing' I mean anything that has troubled you over a period of at least 12 months or that is likely to trouble you over a period of at least 12 months.”*

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Figure 16. Distribution of number of words used in questions.

Figure 16 shows the distribution of the questions with respect to the number of words used. It stands out that 75% of the questions contain a maximum of 20 words. As expected, the extreme cases coincide with the longest questions in terms of characters identified in Figure 15.

Gráfico, Histograma

Descripción generada automáticamente

Figure 17. Distribution of number of sentences used in questions.

Figure 17 shows the distribution of recurring questions with text (99) with respect to the number of sentences used to request information. As mentioned above, sentences separated by punctuation, exclamation marks or question marks (. ? !) are taken as different sentences. It is evident that the requests are mostly made in a single sentence, but that there are some questions that extend up to five sentences. These extremes cases are the following:

* fiyrdia\_cawi: “If possible, please provide an answer to this question as this is one of the key questions in this study. Please be assured that the information you give us will be treated confidentially. In the past 12 months how much have you personally received in the way of dividends or interest from any savings and investments you may have? Please enter an amount to the nearest pound. If no dividends or interest received, please enter 0.”
* jsprf\_cawi: “If possible, please provide an answer to this question as this is one of the key questions in this study. Please be assured that the information you give us will be treated confidentially. What was the amount of (your share of) the profit or loss figure shown on these accounts for this period? Include any money subsequently put back into the business. Please provide an amount to the nearest pound.”

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Figure 18. Distribution of number of nouns used in questions.

Figure 18 shows the distribution of the questions with respect to the number of nouns used in the information request. As mentioned above, the nouns were identified using NLP in such a way that a contextual evaluation of each word was performed. More than 75% use less than 5 nouns but also some extreme cases with up to 18 nouns. The review of these cases coincides with what was found in the longer texts in terms of characters and words (Figure 17). For example, the question paygl presents 16 nouns due to the level of detail and examples presented in the question formulation: “The last **time** you were paid, what was your **total (gross)** **pay** before any **deductions**? This is before any **deductions** for **tax**, **National Insurance** or **pension contributions**, **student loan repayments**, **union dues** and so on. Please include any **overtime**, **bonuses**, **commission**, **tips** or **tax refunds**”

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Figure 19. Distribution of number of verbs used in questions.

Like Figure 18, Figure 19 shows the number of verbs used in the questions obtained A more compact distribution can be seen, since more than 95% of cases use a maximum of 5 verbs. The most extreme case (12 verbs) is presented by the question identified with the aideft code, which uses more verbs to introduce the question, ask it and then repeat it indirectly: *“Thinking about everyone who lives with you that you look after or provide help for - does this extra work looking after NAME(S) prevent you from doing a paid job or as much paid work as you might like to do? Would you say you are...”*

*Gráfico

Descripción generada automáticamente*

Figure 20. Distribution of abstract nouns used in questions.

Finally, Figure 20 shows the distribution of abstract nouns in the questions gathered. It can be noticed that almost half of the questions used at least one abstract noun and some few extreme outliers used more than 5. Among these cases we found:

* scsf6a: These **questions** are about how you feel and how **things** have been with you during the past 4 **weeks**. For each **question**, please give the one **answer** that comes closest to the way you have been feeling. How much of the **time** during the past 4 **weeks**... Have you felt calm and peaceful?
* scsf2a: The following **questions** are about **activities** you might do during a typical **day**. Does your **health** now limit you in these **activities**? If so, how much? Moderate **activities**, such as moving a table, pushing a vacuum cleaner, bowling or playing golf

In the first case, the abstract nouns are the following: “questions, things, weeks, question, answer, time, weeks”. In the second case they are: “questions, activities, day, health, activities and activities.”

In the section bellow, the most relevant results of the bivariate exploration of the constructed database are presented.

Imagen que contiene Gráfico

Descripción generada automáticamente

Figure 21. Distribution of concepts in domains

Figure 21 shows the distribution of concepts according to question domain. It is noteworthy that all questions in the General Health module focus on the person's perceptions of their overall health (emotional, social, and physical). Similarly, and as expected, the questions in the Satisfaction module also collect information on perceptions rather than concrete behaviors.

Gráfico

Descripción generada automáticamente

Figure 22. Distribution of time references of questions in subdomains (or modules)

Figure 22 shows the distribution of questions with text in relation to the domain and reference time used. It is noteworthy that only 8 of the 33 subdomains use questions framed for the future. Furthermore, even though most modules are predominantly interested in known aspects of the present, there are some modules more interested in recording elements, behaviors, or perceptions about the past. The clearest case is the General Health Questionnaire in which all questions are focus on retrieve how does the patient feel in the past. The same can be said about the modules focused on get information about the history of different topics (“Employment Status History”, “Fertility History”, “Partnership History”, “Migration History” and “Family Background”).

Gráfico

Descripción generada automáticamente

Figure 23. Average number of options per subdomain of questions

Figure 23 shows the average number of choices of the selection type questions according to the domain to which they belong. The modules that retrieve information about the identity in terms of ethnicity and religion are among the subdomains with more options because they try to accurately offer coded options that are aligned with the self-belonging perception of the surveys (For example, the question “What is your ethnic group?” offers 18 options including very specific ones as “White & Black Caribbean”). Also, family background is in the top three modules because it retrieves very specific information about the relationship of the respondents with people their family (“mother”,”father”,”cousin”,etc.). Finally, the satisfaction modules have a high overall option given the standardized use of 7 Likert scale options in all questions (1. Completely Dissatisfied and 7. Completely Satisfied).

Gráfico

Descripción generada automáticamente

Figure 24. Average length of questions with text according to domain

Figure 24 shows the average length of the questions according to the domain to which they belong. In addition, the number of questions considered in each domain is specified to facilitate a critical look at the average. For example, although the average length of the questions in the Job Satisfaction module is higher, the fact is that this domain only contains one question. In that sense, it makes more sense to focus on the modules that contain more questions. In the top 5, we see that the average is always less than 150 characters. This tendency is repeated for other constructed variables such as number of words, or number of verbs and nouns.

Finally, the next section presents the result of the unsupervised learning techniques applied to the gathered questions: KMeans and Gaussian Mixture Model (GMM).

As mentioned in the Methodology section, K-means is an unsupervised machine learning algorithm that partitions a dataset into K clusters by iteratively assigning data points to the nearest cluster centroid, which are constantly recalculated as the means of their respective cluster’s cases. The process continues until the centroids stabilize, signifying convergence. The number of clusters is an hyperparameter that must be decided prior, and its selection affects the overall result of the groups generated. The selection of the number of clusters can be assessed through the sillohuete metric, which indicates on a range from -1 to 1 how well the clusters formed differentiate one case from another case belonging to a different group. In this case, the elbow graph was used to find the optimal number of clusters.

Similarly, GMM is another unsupervised learning technique that can be used to group rows or data points. Here it is assumed that data points are generated from a combination of several Gaussian Distribution, each one representing a cluster. The number of distributions (or clusters) is an hyperparameter that can be evaluated using silhouette scores in an elbow plot. The main difference between the two techniques is that K-means is a centroid-based algorithm that assigns data points to clusters by minimizing distances to cluster centroids, while GMM is a probabilistic model that represents data as a mixture of distributions, allowing for more flexible cluster shapes and accommodating overlapping clusters. However, for the purpose of this project we used both to compare the number of clusters suggested and how those clusters are conformed.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Figure 25. Elbow plot of group numbers during clustering of questions using KMeans.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Figure 26. Elbow plot of group numbers during clustering of questions using GMM.

Figure 25 and 26 show the evolution of the average sillohuete scores as the number of clusters increases. For computational efficiency, a range of 100 clusters was tested. The reason for not testing higher numbers is based on parsimony and practicality criteria. If the number of clusters is very similar to the number of cases (in this case 424) the separation of the data will be as efficient as possible since each case will belong to a unique cluster. In this context, clustering is not useful because it is scarce (very few cases in a group) or even non-existent (one case per cluster). Therefore, the definition of the ideal number of clusters must and was based on a parsimony criterion. That is, to group as many cases as possible into as few clusters as possible without sacrificing efficiency (measured by the sillohuete score in this context).

All the aspects considered, Figure X and X1 highlight that the efficiency of the hyperparameter peaked around 25 and 45 clusters. After that threshold, the sillohuete score does not improve significantly for any of the two techniques. In that sense, it was found that the number of groups that optimized the performance of the clustering for both techniques was 32. This coincidence allowed to stablish that regard the different techniques developed to calculate the similarities between questions, both unsupervised algorithms had similar conclusions in terms of how many groups of questions it should be considered.

Gráfico

Descripción generada automáticamente

Figure 27. Distribution of questions in suggested number of clusters - Kmeans

Gráfico

Descripción generada automáticamente

Figure 28. Distribution of questions in suggested number of clusters – GMM

Finally, Figure 27 and Figure 28 show the distribution of questions among the 42 generated clusters for both techniques. The numbers assigned to the clusters should not be strictly evaluated since they depend on the where the calculation starts (initial points for the centroids). However, what should be noticed is that the trend in the number of questions on the clusters is similar for both techniques. Therefore, in general, it can be stated that the two techniques performed optimally on the same number of clusters and that the groups formed grouped the same number of questions.

# CONCLUSIONS

Surveys are a widely used tool all over the world because of their flexibility and ease of application. Understanding Society is a longitudinal study applied in the UK that makes use of this tool to address different topics of great importance (demographics, employment, health, satisfaction with government, etc.). As one of the world's leading longitudinal household studies, the responses collected are used for research and policy design by academics and governments respectively. Therefore, it is of utmost importance to assess the accuracy (how well they measure what they want to measure) and reliability (how consistently they measure what they want to measure) of its questions.

One of the most important metrics to evaluate data quality is measurement error. Applying this assessment systematically requires knowledge of how the characteristics of the questions (topic, length, usage of key words, etc.) influence how people respond to those questions. The data that describes the tools of research studies such as survey’s questions is also called metadata. This kind of information is not always generated or available. In the case of Understanding Society, it does not have a metadata document available on the web for their surveys nor did they share a similar database despite communications with their team. In that sense, it is assumed that generating metadata for this study is a pending task.

The aim of this work was to systematically elaborate the metadata corresponding to the Understanding Society longitudinal study in order to evaluate the accuracy and reliability of the responses in a subsequent project. Such response quality metrics benefit (are more reliable) from having as many records as possible. That was the main reasons why this project considered questions of the individual questionnaire appearing in all 12 available editions (or waves) and all questions that appeared in the same survey during the first edition. The list of variables provided by the Survey Quality Predictor (SQP) project was used as a reference to identify which variables were feasible to obtain and useful for further analysis.

To systematically construct the metadata this project divided the methodology in four stages. First, the questions asked in all waves and the ones asked in the first edition were identified using R after downloading samples of all the available datasets of the individual questionnaire, keeping the questions of the first wave, and identifying the questions that were always present by their code name. This stage ended with a list of the recurrent and first wave questions as the output. Second, the information available on the website of the study was collected and structured through a process that involved web scrapping with the library Rvest, regular expressions and manual coding in R. Here, only the modules that were present in the 12 editions available were considered as well as all modules from the first edition. This stage ended up with a database of 424 questions and 6 variables that described them. Third, the text on the questions were analyzed to further describe the questions in terms of grammatical structure, topics and length. This was achieved using Natural Language Processing available in the library udpipe, regular expressions, word clouds and manual coding. Finally, with all the features obtained, the questions were clustered into different groups based on their similarities using two unsupervised learning techniques. This process led up to a final output of a database with 424 rows (questions) and 20 variables (questions characteristics).

The constructed metadata was analyzed through an exploratory data analysis of their variables. The most important insights were the following. First, most of the questions are focus on knowing the present or past behaviors and practices of the respondents rather than their belief or opinions, specially when they are focused on the future. Second, 7 of the 33 modules (subdomains) in which questions are inscribend are related to job aspects (“Current employment”, ”Non-employment”, ”Job satisfaction”, etc.) and two of them are in the top 3 of modules with more questions. These aspects were a clear indicator of the special interest of the study on the labor market. Third, related to the grammatical structured of the questions, they are mostly presented in a direct manner (direct request), with an average length of 15 words per question and in just one sentence. Around 50% of the choice selection questions present less than 6 options and most of the questions that exceed that amount are likert scalar questions, demographic or identity questions. Finally, questions were evenly spread in terms of the usage of ‘Wh’ words.

The Kmeans and GMM algorithms agreed that the optimal number of clusters to group the questions was 42. After that margin, and up to 100 clusters, the improvement in cluster quality (as measured by the sillohuete coefficient) is marginal or even worsens. The distribution in the number of questions per cluster is also similar for both techniques. In that sense, Kmeans and GMM cluster the generated metadata in a similar way. Both results (group to which each question belongs) were included as new variables in the final bases: metadata and standardised metadata for unsupervised learning tests.

This work has a few limitations and areas for improvement. First, the web scraping process could be optimised by taking input from other survey formats not currently available. As mentioned, the PDF or web format either did not have all the information suggested by the SQP project or the format was not structured enough to collect the information in a systematic way. A web format of the survey would facilitate this process. Another alternative would be to manually include all the variables that are not possible to obtain through the use of webscrapping tools. This would involve checking whether the questions under consideration show variability in such items or not (e.g. question position in the questionnaire does show variability between editions). In that sense, this project did not consider such an alternative because the objective was to generate the metadata as systematically as possible.

Secondly, the domain of questions addressed by the project can be expanded to consider questions from all the editions. With the methodology approach used in this project, this would imply long periods of web scraping process (access to all module links from all waves). However, this suggestion would significantly increase the corpus of questions in the database, which would facilitate the application of more powerful algorithms that can identify the similarity between words used, such as the word2vec text neural network. Finally, this work would be greatly enriched by feedback and critical review from the Understanding Society study team. The resources available on the website were very helpful in its development, so dialogue and openness with other possible documentation would further enhance the work.

# REFERENCES

Babbie, E. R. (2016). *The practice of social research* (14th ed.). Cengage Learning.

Brown, T. J., & Churchill, G. A. (2020). Measuring Customer Satisfaction: Validity and Reliability of an Extended ServQual Instrument. *Journal of Retailing*, 96(2), 226-239.

Buck, N., & McFall, S. (2012). Understanding Society: Design Overview. *Longitudinal and Life Course Studies*, 3(1), 5-17.

Conway, J. (2009). Regular Expressions in R. *R Journal*, 1(1), 37-40.

Couper, M. P. (2017). *Designing effective web surveys*. Cambridge University Press.

DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). SAGE Publications.

Fowler Jr., F. J. (2013). *Survey research methods* (5th ed.). SAGE Publications.

Furner, J. (2020). Definitions of “metadata”: A brief survey of international standards. *Journal of the Association for Information Science and Technology*, 71(6), E33-E42.

Gagolewski, M. (2014). Stringi: Fast, Consistent, and Full-Featured String Processing in R. *Journal of Statistical Software*, 57(5), 1-25.

Golafshani, N. (2003). Understanding reliability and validity in qualitative research. *The Qualitative Report*, 8(4), 597-607.

Johnson, R. E., Lavoie, J. A., & Miller, K. M. (2019). Validity and Reliability of a Stress Assessment Tool for Healthcare Professionals. *Journal of Nursing Measurement*, 27(3), E171-E182.

Jurka, T. P. (2018). Text Mining in R: A Tutorial. *Journal of Quantitative Linguistics*, 25(3), 175-190.

Lee, J. Y., & Chang, L. (2019). The Impact of Response Option Order on Data Quality in Surveys. *Public Opinion Quarterly*, 83(3), 655-678.

London, M., & Zeileis, A. (2015). Data Scrubbing with R: Getting Data Ready for Analysis. *O'Reilly Media*.

Luck, S. J., Stewart, A. X., Simmons, A. M., & Rhemtulla, M. (2021). Standardized measurement error: A universal metric of data quality for averaged event‐related potentials. *Psychophysiology*, 58(6)

Oberski, D., Gruner, T., & Saris, W. E. (2011). The prediction procedure of the quality of the questions based on the present data base of questions. In: The development of the program SQP 2.0 for the prediction of the quality of survey questions. *RECSM Working Paper 24*, 71 - 88. Available at: <http://www.upf.edu/survey/_pdf/RECSM_wp024.pdf>

Saris, W. E., Oberski, D. L., Revilla, M., Zavala-Rojas, D., Lilleoja, L., Gallhofer, I. N., & Gruner, T. (2011). The development of the program SQP 2.0 for the prediction of the quality of survey questions.

Silge, J., & Robinson, D. (2017). *Text Mining with R: A Tidy Approach*. O'Reilly Media.

Smith, A. B., & Brown, E. A. (2021). Question Wording and Response Quality in Surveys. *Journal of Survey Statistics and Methodology*, 9(2), 223-242.

Smith, J. D., Acosta, W., Shaffer, J. R., & McNeil, D. W. (2018). Assessing the Reliability and Validity of a Self-Efficacy Scale for Physical Activity in Older Adults. *Journal of Aging and Physical Activity*, 26(3), 363-368.

Sue, V. M., & Ritter, L. A. (2012). Conducting online surveys. *SAGE Publications*.

Survey Quality Predictor. (2017). *SQP Coding Instructions*. Barcelona, Universitat Pompeu Fabra.

Wickham, H. (2014). *XML and Web Technologies for Data Sciences with R*. Springer.

1. As mentioned in the limitations of the study, the Understanding Society team was contacted about this type of database and did not receive any response. Therefore, it is assumed that the metadata has not yet been generated. [↑](#footnote-ref-1)
2. Figure 1 shows not yet applied editions of the study (waves 14, 15, 16) because the distribution of questions and modules are already planned. In addition, although it shows information from the household questionnaire, this work only used the information of the individual questionnaire. [↑](#footnote-ref-2)
3. Only fractions of the data were downloaded because the interest was in the name of the variables/columns and not the responses. Using small fractions speeded up the download, extraction, and filtering process. [↑](#footnote-ref-3)
4. Based on the SQP project, the module name was adopted as the domain of the question: the general subject of the question, the topic of what is being asked (2017, 7). [↑](#footnote-ref-4)
5. This point is explained later in this section. [↑](#footnote-ref-5)
6. During the web scrapping stage, regular expressions were also used to convert the information obtained with Rvest into a structured database. [↑](#footnote-ref-6)
7. The counting of the variables is conceptual in the sense that the characteristics obtained are counted and not strictly the number of columns as some of them are coded in a broad way. For example, the domains have a one-hot-encoding format and are spread over 4 columns (one for each subdomain) given their non-exclusive nature. [↑](#footnote-ref-7)