

UNIVERSITY OF MACEDONIA  
POSTGRADUATE PROGRAM  
DEPARTMENT OF APPLIED INFORMATICS

**Analyzing trends in Data Science and Data Analytics  
Using the Stack Overflow platform**

M.Sc. THESIS  
Of  
Margaritopoulos Stergios

October 2023

**Analyzing trends in Data Science and Data Analytics  
using the Stack Overflow platform**

**Margaritopoulos Stergios**

B.Sc.in Business Administration

University of Macedonia

M.Sc. Thesis

submitted as a partial fulfillment of the requirements for

THE DEGREE OF MASTER OF SCIENCE IN APPLIED INFORMATICS

Supervisor: Dr Abatzoglou Apostolos

Approved by examining board on 30 October 2023

Prof Alexander Chatzigeorgiou Prof Stylianos Xinogalos Prof Apostolos Abatzoglou



## **Abstract**

In recent years, the fields of data science and data analytics have experienced a remarkable surge, primarily fueled by the ever-increasing demand for valuable insights derived from data. Professionals in this domain are increasingly turning to community-driven platforms such as Stack Overflow to facilitate knowledge sharing and collaborative problem-solving. This master's thesis embarks on the exploration of the dynamic landscape of data science and data analytics, conducting an analysis of the prevailing and actively discussed questions within the Stack Overflow platform. By closely examining these inquiries and identifying emerging trends, this study aims to offer invaluable insights that have the potential to empower not only data analytics practitioners but also enrich the broader community.

## **Keywords**

Data analytics, Data science, Stack Overflow, NLP, QA websites, Big Data, SEDE, Topic Modeling, Word Cloud

# Table of contents

<b>Chapter 1: Introduction.....</b>	<b>4</b>
1.1 Aim and objectives.....	4
1.2 Thesis layout.....	4
<b>Chapter 2: Theoretical Background.....</b>	<b>6</b>
2.1 Data Science.....	6
2.2 Big Data and Data analytics.....	7
2.3 Natural Language Processing (NLP).....	8
2.4 Question Answering (QA) websites.....	9
2.5 Stack Overflow Platform.....	10
<b>Chapter 3: Related Work.....</b>	<b>12</b>
3.1 Related Work.....	12
<b>Chapter 4: Analysis.....</b>	<b>15</b>
4.1 Data Collection.....	15
4.2 Excel Analysis.....	17
4.3 Orange Data Mining Software.....	20
4.4 Line Charts and Histograms with Orange.....	23
4.5 Topic Modeling.....	25
4.6 Word Cloud for question titles.....	28
4.7 Word Cloud for question tags.....	32
<b>Chapter 5: Conclusions, limitations &amp; future work.....</b>	<b>38</b>
5.1 Research limitations.....	38
5.2 General Conclusions.....	38
5.2.1 Temporal Analysis.....	38
5.2.2 Topic Modeling.....	39
5.2.3 Word Cloud.....	40
5.3 Used tools.....	40
5.4 Future work.....	40
<b>References.....</b>	<b>42</b>

# Chapter 1: Introduction

## 1.1 Aim and objectives

In today's rapidly evolving digital landscape, the field of data science and data analytics has become integral to organizational success. As businesses strive to make the most out of data for informed decision-making, understanding the prevalent trends in this domain is crucial. This thesis aims to delve into the discussions related to the domains of data science and data analytics within the Stack Overflow platform. Among its primary objectives is the examination of the progression of these discussions across time, in order to discover whether the fields of data science and data analytics remain popular during the years.

The study involves the analysis of data to uncover trends, prevalent topics, and frequently used tools, all with the purpose of comprehending the interests and resources that drive the data science and data analytics community on Stack Overflow. In doing so, it also seeks to identify potential areas for future research. Furthermore, it assesses the role of Stack Overflow as a knowledge-sharing platform in advancing these fields.

## 1.2 Thesis layout

Chapter 1 serves as an introduction to the thesis, outlining its primary aim and core objectives.

Chapter 2 offers the necessary theoretical foundation for this thesis, laying the groundwork for its practical focus. It covers the evolution of data science, how it relates to other fields like statistics and introduces the concept of "Big Data" and its core aspects (volume, variety, velocity, value, and veracity).

Additionally, it delves into the realm of Natural Language Processing (NLP), a crucial domain of research and application that explores how computers can effectively comprehend and manipulate natural language for various purposes. Emphasis is given to the significance of NLP in the context of Big Data principles while introducing essential concepts such as topic modeling, with a focus on the widely-used Latent Dirichlet Allocation (LDA) method.

It also explains what QA websites are, the benefits they provide, and how they foster online communities. It specifically focuses on Stack Overflow, a widely-used QA site in the tech community, describing how it works and why it remains successful over time.

In Chapter 3, the existing literature and research relevant to the thesis topic is reviewed. It provides a critical context summarizing key findings, methodologies, and insights from prior research. By examining existing studies, frameworks, and practical applications, the aim is to identify gaps and potential areas for innovation.

Chapter 4 encompasses the complete data process, spanning from the initial data collection phase to subsequent analysis and visualization. It commences with an in-depth description of the data collection process, highlighting the use of SQL to gather the essential data and setting the stage for a profound understanding.

Moving forward, this chapter delves into the distribution of questions and answers over the years, effectively conveying this information through the use of excel-generated charts.

To further explore the data, Orange, a data mining software, is introduced. In parallel, essential background information on the LDA algorithm is provided to facilitate a comprehensive exploration of text analysis, using Orange to craft informative line charts and histograms. Two word clouds are generated from the dataset to reveal the most relevant question titles and tags and, by extension, the most commonly used programming tools within the Stack Overflow community.

Chapter 5 presents the key conclusions drawn from the data analysis. It also highlights the limitations that stem from the data collection criteria. Moreover, it points toward potential areas for future research and exploration in the field.

# Chapter 2: Theoretical Background

## 2.1 Data Science

The concept of data science has a long and evolving history that dates back several decades. It is a field that has emerged at the intersection of various disciplines, driven by the need to use the power of data for insights, decisions, and innovation. In this chapter, we will delve into the evolution of data science, exploring its roots and the factors that contributed to its emergence as a distinct and interdisciplinary domain.

The term "data science" first appeared in literature in 1974 when Peter Naur, a Danish computer scientist, introduced it in his book "Concise Survey of Computer Methods". However, its foundations were laid even earlier in 1968 when Naur mentioned another term, "datalogy". These early references indicate that the idea of data science had been brewing in the minds of scholars and practitioners for some time.

Yet, the essence of data analysis predates the formalization of data science. In 1962, John Tukey, an American mathematician and statistician, expressed a shift in his perspective in "The Future of Data Analysis", stating, "I have come to feel that my central interest is in data analysis". This sentiment foreshadowed the evolving landscape of data-related fields [3].

The relationship between statistics and data science is profound [4]. In 1997, Chien-Fu Jeff Wu proposed renaming statistics to "Data Science" and suggested calling statisticians "Data Scientists". This renaming was prompted by the recognition that data science encompasses much more than traditional statistics [5].

But why the need for a new term? As Vasant Dhar highlights, data science deals with increasingly unstructured data, including text, images, and videos. Beyond statistical tools, data analysis requires insights from sociology, linguistics, and other disciplines. Moreover, advancements in technologies such as markup languages and tags have enabled computers to actively participate in sense-making processes. This evolution brings forth complex issues spanning business, law, ethics and machine learning [6].

As we navigate through the historical roots and contemporary facets of data science, it becomes evident that it is more than just another form of statistics. It is an interdisciplinary field that synthesizes statistics, informatics, computing, communication, sociology, and management.

A definition from the book "Data Science: A Comprehensive Overview" by Cao, L. (2017) regarding data science is as follows: "Data science is a new interdisciplinary field that synthesizes and builds upon statistics, informatics, computing, communication, management, and sociology. It aims to study data and its environments, including domains and other contextual aspects such as organizational and social aspects, in order to transform data into insights and decisions by following a data-to-knowledge-to-wisdom thinking and methodology".

In the same book, the following formula describes the integration of various sciences within the field of data science:



Data Science = Statistics + Informatics + Computing + Communication + Sociology + Management + Data + Environment + Thinking [5].

To date, the demand for data science expertise continues to be critical in various industries. One reason behind this demand is the imperative to analyze vast and complex datasets, a phenomenon widely recognized as "big data".

## 2.2 Big Data and Data analytics

The term "**Big Data**" has emerged as a response to the increasing growth of data, which appears to outpace the ability to process it effectively [13]. Moreover, the cost of storing such data has become significantly more affordable than in the past, and this trend is expected to continue and possibly accelerate in the future [14].

This vast scale of data extends far beyond petabytes and exabytes, encompassing not only the volume but also the speed and diverse data types that organizations must contend with. In 2001, Doug Laney introduced the 3V model as a way to characterize big data, focusing on three fundamental attributes: volume, variety, and velocity.

Over time, many organizations and experts have expanded upon the original model, extending it into the 4V or even the 5V model. This evolution has introduced two new dimensions to the equation: "value" and "veracity". The addition of "value" underscores the critical role that big data plays in an organization's ability to extract meaningful insights. By analyzing data effectively, organizations can gain invaluable information regarding customer behavior, deliver personalized services, and solve previously obscure problems.

Meanwhile, "veracity" forms the fifth dimension, focusing on the accuracy and truthfulness of data. In the vast volume of data, veracity becomes a critical concern. Uncertainty can spread through data for various reasons, such as legal complexities, privacy concerns, data duplications, and more. These elements add a layer of complexity to the data environment, highlighting the need for data reliability [15].

To address the challenges of the 4 or 5 V's of Big Data, organizations in the field of data analytics recognize that having more data doesn't always equate to having more useful information. The challenge of dealing with an overwhelming amount of data that the system can't handle isn't a new problem; it has been a concern in earlier approaches. This challenge still persists in today's era of big data analytics. Therefore, preprocessing plays a crucial role in enabling computers, platforms, and analysis algorithms to manage input data effectively.

Traditional data preprocessing methods, such as compression, sampling, and feature selection, are expected to be efficient in the age of big data. In the realm of data analytics, these preprocessing techniques serve as the foundation for extracting valuable insights from the vast volume of data [16].

## 2.3 Natural Language Processing (NLP)

Natural Language Processing (NLP) represents a domain of research and practical application dedicated to exploring how computers can effectively comprehend and manipulate natural language, whether in text or speech, for beneficial purposes. NLP researchers endeavor to gain insights into how humans understand and utilize language, with the ultimate goal of developing suitable tools and techniques that enable computer systems to grasp and manipulate natural languages to achieve desired objectives.

The foundations of NLP are rooted in various fields, including computer and information sciences, linguistics, mathematics, electrical and electronic engineering, artificial intelligence, robotics, and psychology. The scope of NLP applications is extensive, spanning diverse areas of study [20].

Over the initial decades of computational linguistics, researchers grappled with the challenge of encoding the vocabularies and rules of human languages into computer systems. This proved to be exceptionally challenging due to the inherent variability, ambiguity, and context-dependent interpretations within human languages. For instance, a word like "star" could signify an astronomical entity or a famous person, and it could function as both a noun and a verb. Similarly, the interpretation of a headline like "Teacher strikes idle kids" hinged on multiple factors, including word assignments as nouns, verbs, or adjectives and the sentence's grammatical structure.

The landscape of Natural Language Processing (NLP) underwent a transformative shift, commencing in the 1980s and more prominently in the 1990s. This transformation was steered by researchers who shifted from rule-based systems to constructing models based on extensive empirical language data. Statistical or corpus-based NLP was a pioneering application of big data principles, preceding the broader acknowledgment of the power of machine learning and the introduction of the term "big data".

A pivotal revelation arising from this statistical approach to NLP was that even simple methods employing words, part-of-speech (POS) sequences (such as categorizing a word as a noun, verb, or preposition), or basic templates could yield noteworthy results when trained on substantial datasets. Many text and sentiment classifiers still rely solely on distinct word sets, often referred to as the "bag of words" approach, without taking into account sentence and discourse structure or meaning. Enhancing performance beyond these fundamental baselines proved to be a complex task. Nevertheless, contemporary high-performing systems leverage advanced machine learning techniques and exhibit a comprehensive understanding of linguistic structure.

Today, we have access to high-performance NLP tools that can identify not only syntactic and semantic information but also discern discourse context [21].

Within the expansive landscape of NLP, a significant breakthrough has been the development of topic modeling techniques. These methods offer the means to unveil latent topics within extensive collections of documents. They find applications across diverse domains, including the digital humanities. One of the renowned approaches to topic

modeling is Latent Dirichlet Allocation (LDA), a generative probabilistic model known for its effectiveness and, critically, its interpretability. This ability to explain results is pivotal for understanding the underlying themes within large text corpora.

But what is LDA, and how does it operate in the realm of NLP? LDA leverages the principles of Dirichlet processes, named after mathematician Peter Gustav Lejeune Dirichlet. These processes are central to probability theory, manifesting as realizations representing probability distributions. In the context of topic modeling, we're dealing with the identification of groups of words that are statistically significant within a text corpus. Here's where Dirichlet models come into play, helping describe the patterns of word co-occurrence that signify thematic similarity. It employs Bayesian inference to incorporate prior knowledge about the distribution of random variables. Essentially, LDA is a generative probabilistic model that operates by estimating the likelihood of word occurrences within topics and documents. It then utilizes these probabilities to discern latent topics and classify them into documents, revealing the hidden thematic structures that lurk within vast collections of text.

The incorporation of topic modeling, such as LDA, within the realm of NLP has opened up new horizons for the automated understanding of the content, allowing us to dive deeper into the semantic nuances within text data [12],[17].

## **2.4 Question Answering (QA) websites**

The internet has revolutionized the way individuals generate and seek information and knowledge. Now, expressing a need is as simple as entering keywords into a search engine, which promptly presents numerous webpages, varying in relevance, for the user's consideration. However, these search results often fall short of offering precise solutions to the user's problem, necessitating the time-consuming review of numerous pages without a guarantee of finding the desired answer. Question Answering (QA) websites represent a fresh approach to acquiring the desired knowledge more swiftly and efficiently [18].

Question Answering websites serve as a platform for users to engage in knowledge exchange and sharing. Participants within these online forums can be classified into three primary categories: those who solely ask questions, those who exclusively provide answers, and those who both inquire and respond. When a user posts a question on a QA platform, their primary objective is to acquire knowledge about a specific topic, as they actively seek out an expert within the community who can supply the desired information. In this context, those posing questions effectively explore a particular subject, with the community's experts taking on the pivotal role of being the primary source of information, effectively supplanting conventional sources like documents or databases.

While the prospect of obtaining immediate responses to specific information needs is appealing, QA websites also come with risks due to the absence of guaranteed information quality. A significant contrast between user-generated content and traditional content lies in the wide variability in content quality present in user-generated content, as opposed to the more consistent quality typically found in traditional sources. This spectrum encompasses content of exceptionally high quality as well as content of notably lower quality.

The significance of content quality within community-driven question and answer websites has been acknowledged and explored. Research has revealed a strong connection between the quality of questions and the quality of answers. Well-crafted questions typically lead to high-quality responses, while poorly articulated ones often result in less satisfactory answers. Notably, high-quality questions tend to capture more user attention, leading to increased attempts to provide answers and a higher likelihood of receiving the best solutions in a shorter time. Consequently, these well-posed questions not only enhance the overall appeal and utility of QA websites but also significantly contribute to the effective problem-solving process and the continuous enrichment of collective knowledge within the community.

Notable QA websites include Quora, well known for its wide-ranging subject matter expertise and Yahoo Answers, which has been a longstanding platform for diverse inquiries. However, one platform stands out in the domain of technical problem-solving and programming expertise and this is Stack Overflow [18],[19].

## **2.5 Stack Overflow Platform**

Established in 2008, Stack Overflow has emerged as a leading platform for question and answer (Q&A) interactions, garnering special recognition within the developer community. Over the years, it has evolved into a thriving ecosystem, drawing an astonishing 100 million monthly visitors and hosting a repository of 21 million questions [7],[8].

This platform offers users a plethora of interactive features, enabling them to seek answers to their inquiries and contribute solutions to existing problems. Additionally, it employs a voting system, allowing users to express their judgments on questions and answers, thus determining their perceived value within the community.

Stack Overflow currently categorizes each question with user-defined tags. A question must have 1-5 tags. A tag is a word (e.g., "java") or a phrase (e.g. "data visualization") that describes the technical term that the question revolves around.

Much like other online communities, Stack Overflow operates on a self-sustaining knowledge hierarchy. This hierarchy is internally established and recognized, supported by a system of trust. Users accumulate reputation points through their activities on the platform, leading to privileges within the community. Reputation reflects one's contributions and is earned by receiving recognition for the quality of questions or answers provided (note that comments do not contribute to reputation points). Points are awarded based on the number of upvotes received for a question (5 points) or an answer (10 points), with downvotes subtracting points. Moreover, having one's answer accepted as the best solution earns 15 points, and participating in questions with associated bounties can yield varying amounts of reputation points. These bounties are placed on questions as incentives to encourage responses from the community.

Stack Overflow's model of reputation and community engagement has not only established it as a reliable knowledge-sharing hub but has also introduced a unique gamification aspect to

the platform. As users accrue reputation points, they unlock privileges that enable them to participate in moderating content, contributing to the platform's self-policing mechanisms. This system encourages expertise and active involvement while fostering a sense of community and cooperation among developers and technology enthusiasts. Stack Overflow's exceptional growth and the dynamics of its reputation system continue to make it a remarkable case study in online community building and knowledge exchange [9],[10].

# Chapter 3: Related Work

## 3.1 Related Work

Datasets from QA websites have been used in numerous studies. Stack Overflow - the most vibrant community where IT enthusiasts actively exchange programming knowledge - annually conducts a survey and releases the anonymized findings for public access, enabling further examination. Consequently, a multitude of scholars have extracted valuable insights from the publicly accessible Stack Overflow Developer Survey (SODS) data [11].

Tahmooresi, Heydarnoori, and Aghamohammadi (2020) conducted a study that used data from StackOverflow posts to shed light on Python's prevailing trends. Their objectives encompassed defining the primary discussion areas within the Python community, exploring the interests of Python developers and their evolutionary patterns over time, and presenting the various technological offerings associated with Python.

In pursuit of the first goal, the authors began by extracting tags from StackOverflow that were linked to each programming language, and subsequently, they subjected the Python-related posts to a preprocessing stage. This preprocessing involved filtering out posts with zero or negative scores, code snippets, HTML tags, and common English stop words. Finally, the authors applied the Porter stemming algorithm to tokenize the posts and employed the LDA algorithm, a topic modeling technique, to categorize the results.

For their second objective, the researchers divided time into three-month intervals, during which the resulting clusters of topics were discussed and categorized by experts into 12 distinct clusters. They then applied the concept of "impact" as defined by Barua et al. (2014) to these clusters. This concept measured the significance of a topic within an interval relative to other topics.

Regarding the third objective, the study delved into Python technologies, encompassing software solutions, packages, libraries, and frameworks. To accomplish this, Tahmooresi, Heydarnoori, and Aghamohammadi (2020) leveraged the word2vec approach, which assigns a vector to each word in the corpus, thus organizing technologies and uncovering implicit relationships between them.

The findings from the first goal, ranked by priority, revealed the prominence of Python's standard features, encompassing data structures, regex/string manipulation, generators, and list comprehension. Subsequently, web programming and scientific programming emerged as the second and third most popular areas, respectively. Other areas included OS/multitasking/message queues, databases, data formats/serialization, networking, desktop programming, performance, testing, gaming, and, finally, device programming/IOT.

In terms of temporal trends, discussions concerning Python standard features and web programming displayed a declining share compared to other areas. Scientific programming, on the other hand, exhibited rapid growth, surpassing web programming. Additionally, areas such as multitasking, message queues, data formats, and serialization have seen increased

attention since the beginning of 2015. Meanwhile, packaging, library versioning, and installation have remained stable since that time.

The article also introduced various technology alternatives in a matrix format, although it elaborated on only two examples. The first recommendation was to replace Visual Studio, a C# IDE, with PyCharm, a popular Python IDE. The second suggestion was more intricate: replacing Maven with Virtualenv, pip, and requirements.txt. This combination was advocated for scenarios where multiple applications rely on different versions of the same package on a single machine [1].

Kochhar (2016) conducted a survey centered on data mining from StackOverflow, specifically focusing on software testing. This study encompassed five key questions, exploring the themes within testing, the most prominent testing topics, the temporal trends in testing discussions, the prevalence of testing-related subjects in mobile web development, and the primary technical challenges faced in the area of software testing.

The data collection process commenced with the acquisition of necessary data from StackOverflow, comprising a dataset containing 38,000 questions and involving over 25,000 distinct users. To classify topics, the LDA algorithm was employed. Similar to the survey mentioned earlier, a preprocessing stage occurred before applying LDA. This preprocessing involved tasks such as filtering out undesired posts and eliminating noise, ensuring the dataset's quality.

The resulting discussion categories included "test framework", "database", "client-server", "login", "XML build", "threads", "forms", and "image processing". Among these, "test framework", "database", and "client-server" emerged as the most prevalent discussion topics.

For determining the hottest topics, Kochhar extracted the top 2000 most-viewed questions. This step was deemed important because many developers may not actively post on StackOverflow but often read and benefit from questions that have already been resolved, thereby increasing the value of questions with high view counts.

When analyzing temporal trends, the dataset was divided into six-month subsets to provide sufficient data for the LDA algorithm. Notably, dominant topics consistently engaged developers from January 2009 to December 2014, with "test framework" experiencing an upswing in discussions from January 2012 to December 2014.

To measure the prevalence of testing discussions within mobile development, a subset of mobile development questions from a prior survey by Bajaj, Pattabiraman, and Mesbah (2014) was utilized. This subset contained 2,434 questions. The analysis of testing trends within this subset revealed a steady increase in the percentage of testing-related discussions among the mobile development community from July 2012 to December 2014.

Finally, the study identified technical challenges faced by developers through the collection and qualitative analysis of the fifty most significant questions. The importance of these questions was determined based on factors such as user upvotes, downvotes, the number of comments, answers provided, and the number of users who marked the questions as

favorites. Key themes that emerged included basic testing for novice developers, app testing (e.g., testing email functionality or on various mobile devices), best testing practices shared by experienced developers with novices, automation test frameworks and reusable testing modules, and techniques for database testing [2].

The survey conducted by Dada, Obaido, Sanusi, and their colleagues in 2022 is another example of research that utilized data from Stack Overflow to gain insights into various job roles within the IT industry and the corresponding skill sets needed for these roles.

Their data analysis was carried out using Python programming libraries, including pandas, collections, NumPy, and matplotlib. The analysis centered around the diverse roles present in the dataset. The survey question regarding IT roles allowed respondents to specify multiple roles. For instance, one respondent had three distinct roles: 1) designer, 2) front-end developer and 3) mobile developer.

To streamline the analysis, they separated these distinct roles. In total, 23 unique roles were identified, and the frequencies of each role were computed. JavaScript emerged as the predominant programming language across 16 out of the 23 surveyed job roles, particularly those within the software development domain. On the other hand, individuals in marketing predominantly reported HTML/CSS as their most-used programming language, although HTML/CSS also ranked as the second-most used language in 11 instances. For roles like database administrator, data engineer, and data analyst, SQL stood out as the primary programming language. Remarkably, SQL was also among the top three most-used languages in 17 other roles. Additionally, Python held the distinction of being the primary programming language for professionals engaged in data science, academic research, and scientific fields. Conversely, some programming languages like C++ and Hypertext Preprocessor (PHP) were frequently mentioned but were among the less commonly used languages [11].

In light of the recent technological advancements and the growing importance of data utilization, possessing programming skills is becoming increasingly crucial for prospective IT job candidates. Among the respondents of Dada's, Obaido's, Sanusi's et al. survey, an impressive 67% reported extensive usage of JavaScript, making it the dominant language for programming, scripting, and markup tasks. Meanwhile, Python has garnered significant recognition, particularly within the scientific realm. It is renowned for its prevalence in high-performance scientific applications and has become a staple among scientists, academic researchers, and data engineers. This popularity is largely attributed to Python's exceptional performance and ease of use [11].



# Chapter 4: Analysis

## 4.1 Data Collection

Access to Stack Overflow data can be accomplished through three primary options. Firstly, a copy of the Stack Overflow data dump can be obtained via archive.org. Archive.org is a platform that provides free access to collections of digitized materials, websites, software applications, music, audiovisual content, and print materials. The Stack Overflow data dump offers a comprehensive snapshot of the platform's content, allowing for offline analysis and research.

The second option involves utilizing the Stack Exchange Application Programming Interface (API). Within the extensive network of online question-and-answer communities, including Stack Overflow, Stack Exchange offers an API. An API, an abbreviation for Application Programming Interface, comprises a set of rules and protocols that enable different software applications to communicate with one another. In the context of Stack Exchange, an API provides programmatic access to the wealth of data generated by these communities.

The third option is the Stack Exchange Data Explorer (SEDE), an online tool for those who prefer a more interactive approach to data access and analysis. SEDE allows users to interact with the underlying database of Stack Exchange sites through the use of custom SQL queries. SQL, or Structured Query Language, is widely used for managing and manipulating relational databases. With SEDE, specific SQL queries can be crafted to extract, transform, and analyze data sourced from the millions of questions, answers, and discussions found on Stack Exchange sites.

For this thesis, the third option, the Stack Exchange Data Explorer (SEDE), was chosen as the primary method for data collection. However, to effectively use the power of SEDE and formulate SQL queries that accurately extract the necessary data, it was paramount to develop a comprehensive understanding of the underlying database structure and the intricacies of Stack Exchange sites. Stack Exchange database is composed of various tables, each designed to store specific types of information. To navigate and extract data successfully, it's essential to understand the relationships between these tables and the meanings of specific fields within them.

At the core of the Stack Exchange database schema lies the "Posts" table. This table serves as a central repository for storing a wide array of content, including questions, answers, comments, and more. Each entry within the "Posts" table is uniquely identified by an "Id", distinguishing it as either a question or an answer. In the SQL query created, specific fields are extracted from the "Posts" table using the SELECT command. These fields include q.Id, q.Title, q.Tags, q.CreationDate, a.Id, a.Body, and a.CreationDate.

To connect questions with their corresponding answers, a LEFT JOIN operation is utilized between the "Posts" table. In this operation, the "Posts" table is aliased as "q" for questions, and another instance is aliased as "a" for answers. This relationship is established through the condition q.Id = a.ParentId. Essentially, this SQL operation matches each answer to its parent question based on their respective IDs.

Within the query, a WHERE clause is employed to filter the data according to specific criteria. This clause incorporates several conditions: the tags “data-science”, “data-analytics”, “data-mining” and “data-visualization” are used to ensure that only questions relevant to these subjects will be included in the dataset. The approach of including multiple tags for filtering is worth noting, as it is a deliberate decision to ensure comprehensive coverage. This decision takes into account that tags are user-generated, and, in some cases, questions related to data science may be categorized under various tags, such as data analytics or data mining. Likewise, questions concerning data visualization may also pertain to data science. By including multiple relevant tags, the aim is to capture a broader range of relevant content within the chosen topics.

Another condition, q.PostTypeId = 1, is applied to narrow the focus to question posts. By specifying that q.PostTypeId must be equal to 1, the query ensures that only questions from the database are retrieved. Similarly, the condition a.PostTypeId = 2 is implemented to concentrate on answer posts. When a.PostTypeId is equal to 2, the query narrows down its scope to answers.

Last but not least, to maintain temporal relevance, the query also includes YEAR(a.CreationDate) >= 2018. This condition restricts the analysis to answers created in 2018 or later by examining the creation date of answers (a.CreationDate). This constraint ensures that the focus remains on recent answers within the chosen topics, providing current insights into data-related discussions.

Below is the complete SQL query:

```
SELECT
  q.Id AS QuestionId,
  q.Title AS QuestionTitle,
  q.Tags AS QuestionTags,
  q.CreationDate AS QuestionCreationDate,
  a.Id AS AnswerId,
  a.Body AS AnswerBody,
  a.CreationDate AS AnswerCreationDate
FROM
  Posts q
LEFT JOIN
  Posts a ON q.Id = a.ParentId
WHERE
  (q.Tags LIKE '%data-analytics%' OR
   q.Tags LIKE '%data-science%' OR
   q.Tags LIKE '%data-mining%' OR
   q.Tags LIKE '%data-visualization%')
AND
  q.PostTypeId = 1 -- Question posts
AND
  a.PostTypeId = 2 -- Answer posts
AND
  YEAR(a.CreationDate) >= 2018 -- Limit answers to 2018 and beyond
ORDER BY
  q.CreationDate DESC;
```

## 4.2 Excel Analysis

The SEDE gives users the option to download the results of SQL queries in the form of CSV files. Subsequently, these CSV files can be conveniently transformed into excel files (XIs) for further analysis. The resultant excel file conforms to the following structure:

QuestionId	QuestionTitle	QuestionTags	QuestionCreationDate	AnswerId	AnswerBody	AnswerCreationDate
7711698	What is the order in catboost's select_features mean?	<python><data-science><catboost><rf>	16/9/2023 16:42	77116947	<p>The ordering of features depends on the algorithm parameter <code>l</code>	16/9/2023 17:54
77116715	pandas groupby middle heading	<python><pandas><group-by><data-science>	16/9/2023 7:06	77116812	<h3>By pattern</h3><p>To group by string/pattern match, you can i	16/9/2023 7:37
77109331	Is there a method to access specific node and relationship data within a graph catbo4j</cypher><graph-databases><graph-data-scien<	<graph><cypher><graph-databases><graph-data-scien<	15/9/2023 2:42	77109531	<p>You can get extended information about a specific native name<	15/9/2023 4:03
77108657	Trying to run a markdown code on my notebook and having some issues what do</data-science><markdown><data-analysis><data-si<	<data-science><markdown><data-analysis><data-si<	14/9/2023 22:15	77112963	<p>Well, First: Ensure that you've executed the Markdown cell cont	15/9/2023 13:54
77098700	How do I prevent NA values from plotting in ggplot2 boxplot?	<ggplot2><data-science><boxplot><na>	13/9/2023 15:57	77099484	<p>If only column 'yield_grain' is NA, these values should be remov	13/9/2023 18:03
77091827	Can I export Postgres query results from DataGrip to parquet format?	<gresql><intellij-idea><data-science><parquet><data<	12/9/2023 18:34	77091996	<p>Export to parquet format is still a feature request for now on Da	12/9/2023 19:02
77091101	Encountered 'MemoryError' while splitting a Pandas DataFrame column with str.thon</pandas><apache-spark><data-science><parq<	<python><pandas><apache-spark><data-science><parq<	12/9/2023 16:36	77091594	<p>If I got you right, you want to get date and time as separate colu	12/9/2023 17:57
77085842	I'm having a problem on data science (pandas, csv) not sure it's my laptop or whato</pandas><dataframe><read_csv><graph-data-sci<	<pandas><dataframe><read_csv><graph-data-sci<	12/9/2023 11:13	77088632	<p>Try: <code>df = pd.read_csv('Movies.csv', skip_blank_lines=True)not II</code>	12/9/2023 11:20
77078918	How to write a conditional formatting rule for this specific situation?	<excel-formula><data-science><ms-office><data-ar<	11/9/2023 4:30	77079138	<ul><li>Select cell A2</li><li>Hold Shift Key click G12 to select data i	11/9/2023 5:40
77077831	My dataframe keeps making columns with name 'Unnamed'	<python><pandas><dataframe><data-science>	10/9/2023 20:51	77077917	<p><strong>Solution</strong></p><p>To drop multiple columns wh	10/9/2023 21:16
77069310	About the inputs of the Wasserstein Distance W1	<python><scipy><data-science><probability-distributic<	7/9/2023 14:02	77084375	<p>Not sure about SciPy and how they compute Wasserstein Distan	11/9/2023 19:41
77042466	ML Classification for categorical data	<machine-learning><data-science><classification><i<	5/9/2023 11:53	77045424	<p>First of all you made a good job. You've already done the first th	5/9/2023 14:31
77043737	Bulk replace nan in one column based on existing data from others	<python><pandas><data-science><bigdata>	5/9/2023 10:32	77043806	<p>No need to loop over dataframe, just <code>ca href="https://pandas.p</code>	5/9/2023 10:41
77040461	How to use nba_api to find all player seasons in which a player has averaged x st</python><pandas><dataframe><data-science><nba-a<	<pandas><dataframe><data-science><nba-a<	4/9/2023 20:31	77040925	<p>The reason why your code loops &quot;infinitely&quot; is probi	4/9/2023 21:44
77039622	BeautifulSoup doesn't return the proper HTML	<web-scraping><beautifulsoup><python-requests><di<	4/9/2023 17:21	77039639	<p>The content of the website you're trying to access is loaded thr<	4/9/2023 17:24
77038831	How to add new filter to a DAX query	<erbi><data-science><dax><powerbi><desktop><daxst<	3/9/2023 19:18	77034091	<p>Try the following: <code>&lt;/p&gt;&lt;pre&gt;code&gt;SumValueForTable1[Record &lt;</code>	3/9/2023 20:37
77014530	Overhead in parallel tasks in C++	<memory><parallel-processing><data-science><autoco<	31/8/2023 9:26	77015786	<blockquote><p>I can see that at the beginning all 64 CPUs are load	31/8/2023 12:21
77005059	A question about using the loc method to create a new column based on existing</pandas><dataframe><data-science><data-manipulat<	<pandas><dataframe><data-science><data-manipulat<	30/8/2023 4:51	77005181	<p>When using a Series in an assignment, pandas performs index a	30/8/2023 5:28
77004255	Time Series Long to Wide Format R?	<p><dataframe><time-series><data-science>	20/8/2023 0:49	77004410	<p>Double pivot:</pre>class="lang-r prettyprint-override"><code>	30/8/2023 1:08
76999712	counting the number of values separated by a set difference	<p><database><data-science>	29/8/2023 11:17	76999982	<p>Okay, so this is what I've come up with:</pre><code>dataSN	29/8/2023 11:55
76982111	How to set the equations to implement a Kalman filter for predicting the foot tra</statistics><data-mining><kalman-filter><lin</>	<statistics><data-mining><kalman-filter><lin</>	26/8/2023 8:17	76992582	<p>You must have some relations between the data in order to for<	26/8/2023 11:57
76979630	Running batch predictions with fine-tuned PALM model in Vertex AI always thro</artificial-intelligence><google-cloud-vertex</>	<artificial-intelligence><google-cloud-vertex</>	25/8/2023 18:09	77003676	<p>In case anyone stumbles on this question... From my ticket to G<	29/8/2023 21:09
76978380	web scraping showing access denied	<web-scraping><beautifulsoup><python-requests><di<	25/8/2023 15:03	76978566	<p>You have to copy the full curl from your browser and see if it wo<	25/8/2023 15:28
76967724	Datetime column deforms when it is converted to parquet file	<python><dataframe><datetime><data-science><data-manipulat<	24/8/2023 8:58	76970298	<p>Your approach with the assignment to the time zone was already<	24/8/2023 14:23
76965601	How to scrape news articles with (irregular) paragraphs?	<python><web-scraping><beautifulsoup><data-scien<	23/8/2023 10:41	76961112	<p>You can chain <code>&lt;/code&gt; to get a list of all the para&lt;</code>	23/8/2023 11:55
76959391	Standardize python numbers	<python><data-science>	23/8/2023 7:55	76959557	<p>This should work for you. I tested it with all your use cases:</p>	23/8/2023 8:21
76957864	Neo4j GDS: Cannot use Cypher to make a complex projection	<neo4j><cypher><graph-data-science>	23/8/2023 2:20	76962783	<p>Can you check the installed GDS version? (RETURN gdsversion())	23/8/2023 3:50
76948894	Edit filter function on spark	<spark><pyspark><apache-spark-sql><data-science><i<	21/8/2023 21:47	76950005	<p>For every column applying <code>&lt;code&gt;groupBy&lt;/code&gt; &amp;amp; <code>&lt;code</code></code>	22/8/2023 3:55
76948658	GDS 2.4.4 Failed to load with Neo4j 5.11.0 Analytics cluster	<neo4j><graph-data-science>	21/8/2023 20:56	76949160	<p>Try modifying <code>&lt;code&gt;neo4j.conf&lt;/code&gt; on your <code>&lt;/code&gt;</code></code>	21/8/2023 23:01
76948997	Iterating through columns to generate countplot   seaborn	<science><visualization><plotly><express><explorator<	21/8/2023 16:03	76947990	<p>Here is the amended <code>&lt;code&gt;get_count_plot&lt;/code&gt; <code>&lt;/pre&gt;&lt;pre&gt;</code></code>	21/8/2023 16:13
76941008	Scrapy Problem - When I ran scrapy file, there is no output	<scraping><scrapy><web-crawler><extract><data-mi<	20/8/2023 19:24	76941730	<p>This is my try with the scraping of the website of audible:</pre></p>	20/8/2023 23:46
76939671	BeautifulSoup methods	<python><3.x><web-scraping><data-mining>	20/8/2023 13:44	76945698	<p>Based on error you used <code>&lt;code&gt;find_all&lt;/code&gt; method (or <code>&lt;/si</code></code>	21/8/2023 13:22
76934142	How to pass only necessary features to pipeline after SelectKBest	<data-science><feature-engineering><ml><ops><data<	19/8/2023 8:06	76938946	<p>In your pipeline, the <code>&lt;code&gt;SelectKBest&lt;/code&gt; step will be re&lt;</code>	20/8/2023 10:21

The dataset contains 15,046 rows, each representing with an answer. This is in line with the SQL query, which targeted Stack Overflow questions with at least one answer since 2018. It's worth noting that sometimes, the "QuestionId" and "QuestionTitle" fields are the same. This happens because certain questions have received multiple answers.

Excel software provides a "Remove Duplicate" option, which can be applied separately to the "QuestionTitle" and "AnswerBody" columns. This action will validate the singularity of all answers or track any possible duplicate answers by the users. Secondly, it can measure the number of distinct questions.

After using the "Remove Duplicate" function on the "AnswerBody" column, there is one duplicate answer. This might happen when users post identical responses or accidentally submit their answer more than once. In contrast, applying "Remove Duplicate" to the "QuestionTitle" column revealed 3.800 duplicate values and 11.246 unique question titles.

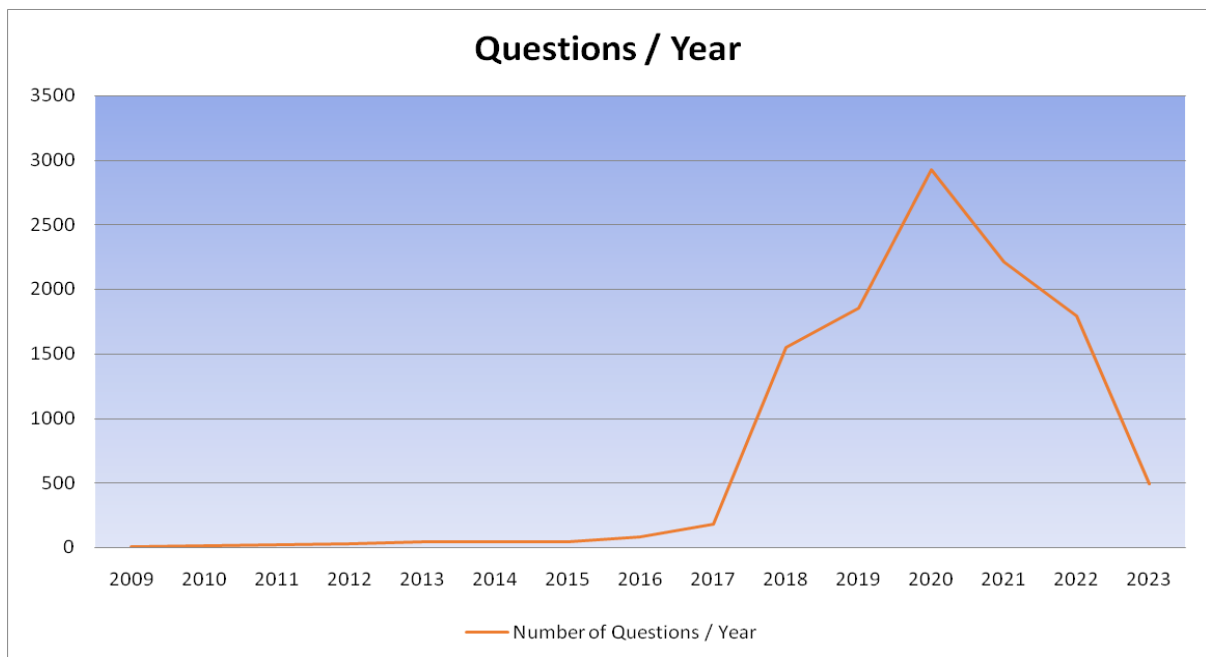
	Questions	Answers
Total Number	15.046	15.046
Unique Number	11.246	15.045
Percentage	74,74%	99,99%

Following the identification of duplicate values in the "QuestionTitle" column, the next step involves visualizing and examining the creation date of these 11,246 questions over time.

In this section, it is important to mention again that the selected dataset covers questions from 2009 to 2023. However, a specific criterion was applied: only questions that had at least one answer posted between 2018 and 2023 were included. This filtering was essential to focus on questions that remained relevant or generated ongoing discussion during this specific time frame. It is crucial to acknowledge that the dataset may exhibit certain limitations, primarily due to the omission of questions from the earlier years (2009 to 2018) that may not have received answers within the designated time frame. This omission is understandable, as it is not uncommon for questions posed a decade earlier to remain unanswered or to lack recent engagement.

Therefore, while these questions might have been numerous in the early years of Stack Overflow, they may not be present in our dataset, leading to a potential underrepresentation of historical trends. Despite these limitations, the trends observed in questions from 2018 to 2023 are valuable and can provide highly accurate insights.

To facilitate this analysis, a Line Chart can be generated to gain insights into the temporal distribution of these questions.

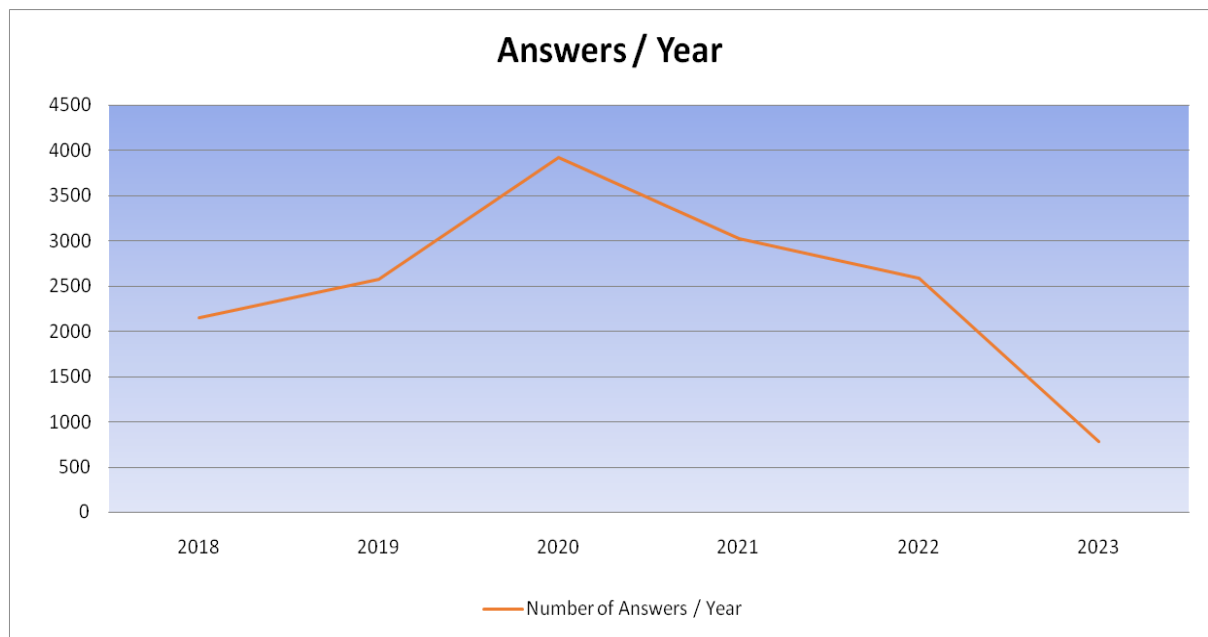


The Line Chart indicates little activity from 2009 to 2017, which aligns with the SQL query's requirement of at least one answer from 2018 to the present. Consequently, there is a relatively low number of active Stack Overflow posts, particularly concerning the selected tags, during the earlier years. However, a trend emerges from 2017, in which the creation of questions keeps increasing dramatically and reaches a peak in 2020. Subsequently, there is a decline in 2021 and a further reduction in 2022.

The figure below shows the percentage difference of questions created per year from the peak year of 2020:

Year	Questions Created	Percentage difference from 2020 Peak
<b>2020</b>	<b>2924</b>	-
2017	180	93,80%
2018	1544	47,10%
2019	1848	36,70%
2021	2209	24,40%
2022	1792	38,70%
2023	490	83,20%

This same analysis can be employed to visualize the activity in terms of answers:



The Line chart of answers per year exhibits a pattern quite similar to that of the Questions. Notably, there's a substantial increase in answers from 2018 to 2020, leading to a peak year with a remarkable difference from the rest. This peak corresponds with that of the question creation during the same year, underscoring a heightened level of user engagement and responsiveness to the community's questions. However, in the subsequent years from 2020 to 2023, there is a marked decline in answer activity.

The figure below shows the percentage difference of answers created per year from the peak year of 2020:

Year	Answers Created	Percentage difference from 2020 Peak
2020	3926	-
2018	2153	45,20%
2019	2578	34,30%
2021	3031	22,80%
2022	2582	34,20%
2023	775	80,30%

### 4.3 Orange Data Mining Software

For further analysis and visualization, Orange data mining software was chosen. Orange is a versatile and user-friendly data mining and machine learning software tool that allows individuals and organizations to extract valuable insights from their data. It is renowned for its intuitive visual interface, making it accessible to both data science newcomers and professionals. With its drag-and-drop functionality and a wide array of built-in widgets, Orange offers a seamless experience for data preprocessing, exploration, modeling, and visualization.

One of its standout features is its interactive and visual approach to data analysis. Users can simply connect widgets in a visual workflow to perform various data mining tasks, such as data cleaning, feature selection, classification, and clustering. This visual approach not only simplifies complex data analysis but also enhances transparency, allowing users to understand and interpret their results more easily.

Orange offers a rich library of widgets that cover a broad spectrum of data analytics and machine learning techniques. It provides a comprehensive toolkit whether someone is interested in exploring data distributions, building predictive models, or conducting advanced text analysis. Furthermore, Orange's flexibility extends to its compatibility with various data formats and its ability to seamlessly integrate with popular Python libraries making it a powerful choice for those seeking both simplicity and depth in their data mining endeavor [12].

## Data



File



CSV File Import



Datasets



SQL Table



Data Table



Paint Data



Data Info



Rank



Edit Domain



Color



Feature Statistics



Save Data

## Transform



Data Sampler



Select Columns



Select Rows



Transpose



Merge Data



Concatenate



Select by Data Index



Unique



Aggregate Columns



Group by



Pivot Table



Apply Domain



Preprocess



Impute



Continuize



Discretize



Randomize



Purge Domain



Melt



Formula



Create Class

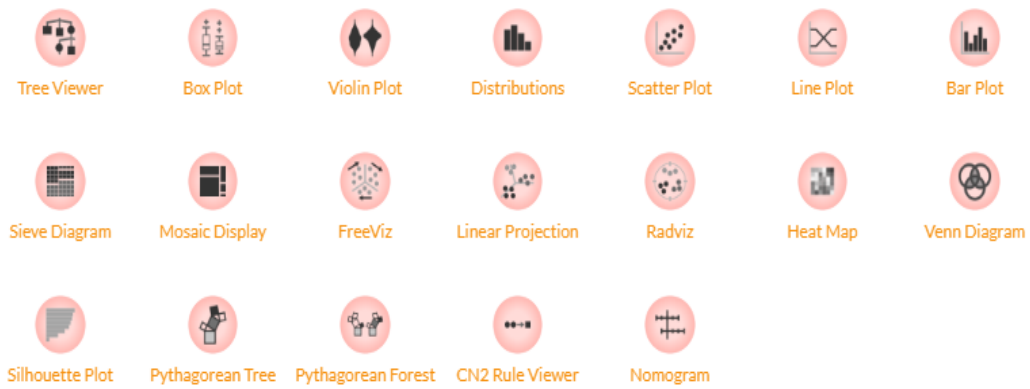


Create Instance

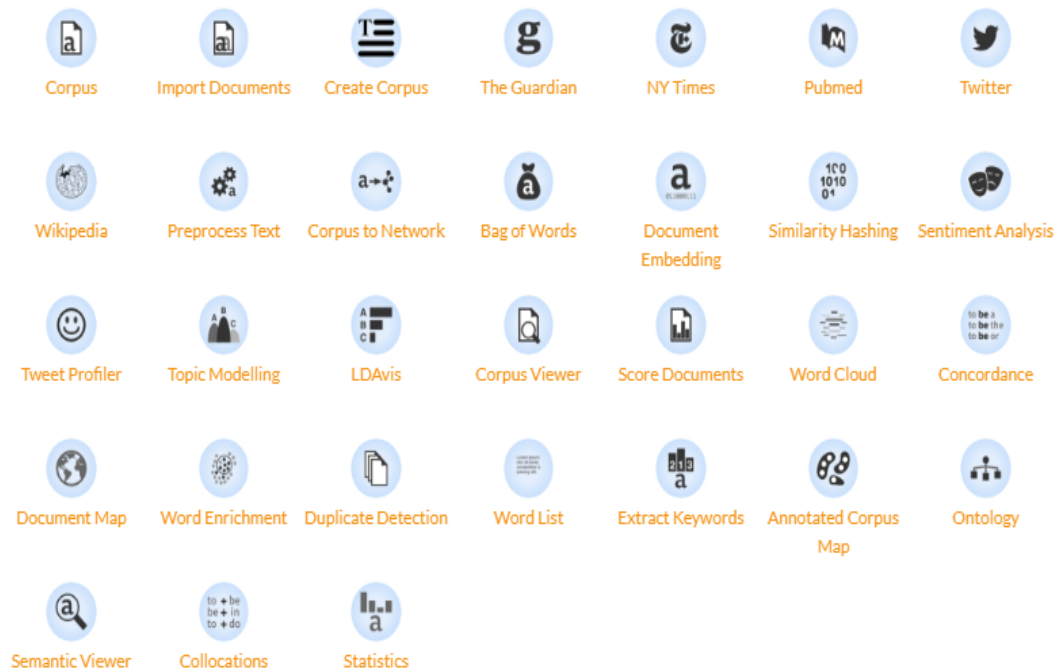


Python Script

## Visualize



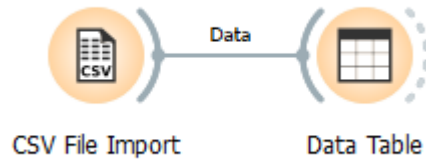
## Text Mining





## 4.4 Line Charts and Histograms with Orange

The initial step in utilizing Orange for data analysis involves importing data into the software. Orange facilitates the import of CSV files through the "CSV File Import" widget. These files can be effortlessly transformed into structured data tables using the "Data Table" widget.



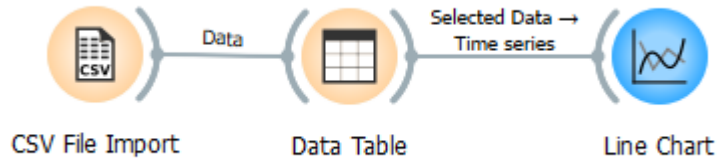
The data table in Orange environment:

QuestionTitle	QuestionTags	AnswerBody	QuestionId	QuestionCreationDate	AnswerId	AnswerCreationDate
What is the ord...	<python><r><...>	The orderin...	77118698	2023-09-16 16:42:54	77118947	2023-09-16 17:54:55
pandas groupb...	<python><pan...	<b>By pattern...</b>	77116715	2023-09-16 07:06:04	77116812	2023-09-16 07:37:32
Is there a meth...	<neo4j><cyph...	You can ge...	77109331	2023-09-15 02:42:37	77109531	2023-09-15 04:03:32
Is there a meth...	<neo4j><cyph...	You can als...	77109331	2023-09-15 02:42:37	77110640	2023-09-15 08:02:37
Is there a meth...	<neo4j><cyph...	To get nod...	77109331	2023-09-15 02:42:37	77114030	2023-09-15 16:33:17
Trying to run a ...	<jupyter-noteb...	Well, First: ...	77108657	2023-09-14 22:15:28	77112963	2023-09-15 13:54:57
How do I preve...	<r><ggplot2>...>	If only colu...	77098700	2023-09-13 15:57:14	77099484	2023-09-13 18:03:37
Can I export Po...	<postgresql><i...	Export to p...	77091827	2023-09-12 18:34:54	77091996	2023-09-12 19:02:03
Encountered '...	<python><pan...	If I got you ...	77091101	2023-09-12 16:36:41	77091594	2023-09-12 17:57:55
I'm having a pr...	<python><pan...	Try: df = p...	77088582	2023-09-12 11:13:21	77088632	2023-09-12 11:20:38
How to write a ...	<excel><excel-...	Selec...	77078918	2023-09-11 04:30:23	77079138	2023-09-11 05:40:34
My dataframe k...	<python><pan...	S...	77077831	2023-09-10 20:51:04	77077917	2023-09-10 21:16:42
My dataframe k...	<python><pan...	What happ...	77077831	2023-09-10 20:51:04	77077964	2023-09-10 21:32:49
About the inpu...	<python><scip...	Not sure a...	77060310	2023-09-07 14:02:05	77084375	2023-09-11 19:41:00
ML Classificatio...	<python><ma...	First of all y...	77044266	2023-09-05 11:53:18	77045424	2023-09-05 14:31:57
Bulk replace na...	<python><pan...	No need to...	77043737	2023-09-05 10:32:06	77043806	2023-09-05 10:41:29
How to use nba...	<python><pan...	The reason ...	77040461	2023-09-04 20:31:27	77040925	2023-09-04 22:44:23
BeautifulSoup ...	<python><we...	The conten...	77039622	2023-09-04 17:21:16	77039639	2023-09-04 17:24:43
How to add ne...	<powerbi><da...	Try the foll...	77033831	2023-09-03 19:18:13	77034091	2023-09-03 20:37:55
Overhead in pa...	<c++><memo...	...	77014530	2023-08-31 09:26:57	77015786	2023-08-31 12:21:20
A question abo...	<pandas><dat...	When usin...	77005059	2023-08-30 04:51:28	77005181	2023-08-30 05:28:53
Time Series Lon...	<r><dataframe...	Double-piv...	77004355	2023-08-30 00:49:17	77004410	2023-08-30 01:08:36
Time Series Lon...	<r><dataframe...	Here is ano...	77004355	2023-08-30 00:49:17	77004426	2023-08-30 01:14:25
Time Series Lon...	<r><dataframe...	zea...	77004355	2023-08-30 00:49:17	77004571	2023-08-30 02:06:16
counting the n...	<r><database>...	Okay, so th...	76999712	2023-08-29 11:17:18	76999982	2023-08-29 11:55:37
How to set the ...	<machine-lear...	You must h...	76982111	2023-08-26 08:17:33	76992582	2023-08-28 11:57:05
Running batch ...	<python><dat...	In case any...	76979630	2023-08-25 18:09:16	77003676	2023-08-29 21:09:33
web scraping s...	<python><we...	You have t...	76978380	2023-08-25 15:03:10	76978566	2023-08-25 15:28:08
Datetime colu...	<python><dat...	Your appro...	76967724	2023-08-24 08:58:07	76970298	2023-08-24 14:23:47

As previously mentioned, SQL query aimed to extract all questions from StackOverflow that encompassed at least one of the four selected tags and had received at least one answer within the past five years. The previous analysis in Excel provided valuable insights into the progression of questions and answers over time.

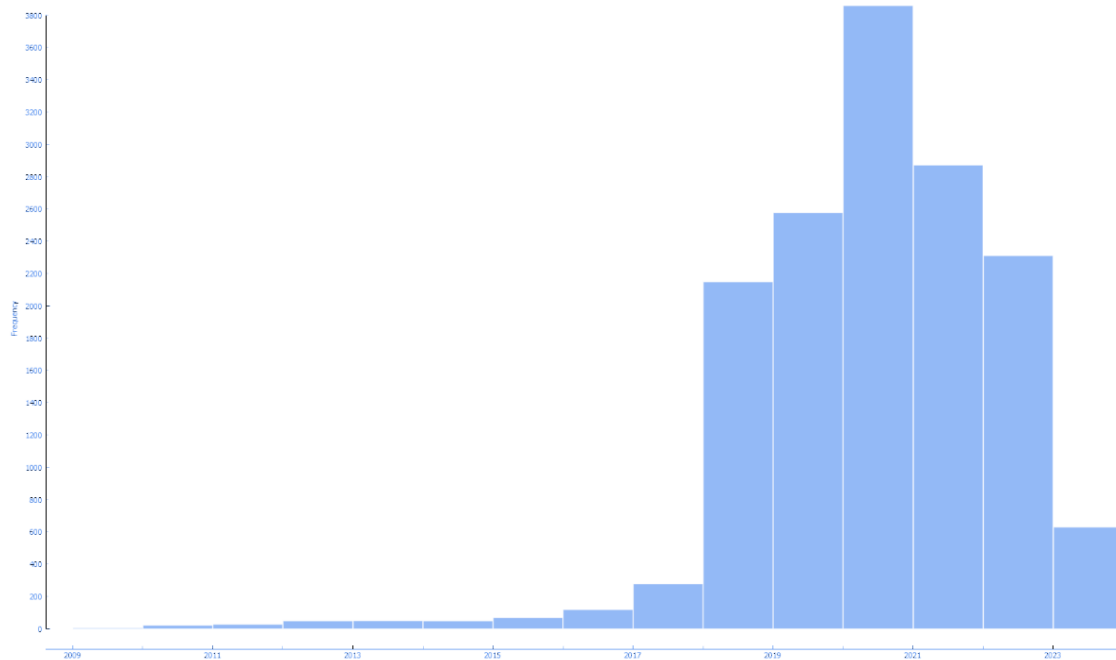
Similar to Excel, Orange offers powerful tools for data visualization, allowing users to create Line Charts that provide a visual representation of how the number of questions has evolved

over the years. A line chart can be created by simply adding the Time add-on, and then connecting the "Data Table" widget to the "Line Chart".

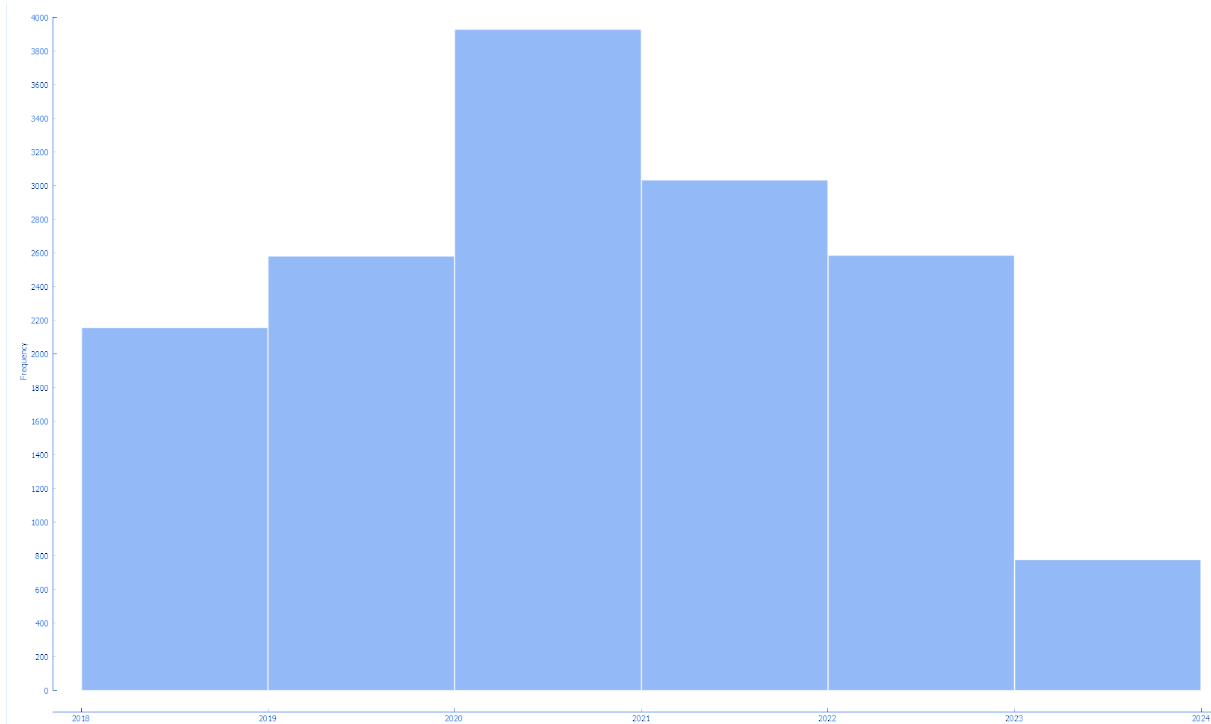


Another widget, the "Distribution", can be utilized to construct a histogram graph, enabling the determination of the specific years in which these questions were created. Utilizing the "Distribution" widget, a corresponding histogram depicting the dates of answer posts can also be generated.

**Histogram for questions created over the years**



**Histogram for answers created over the years**



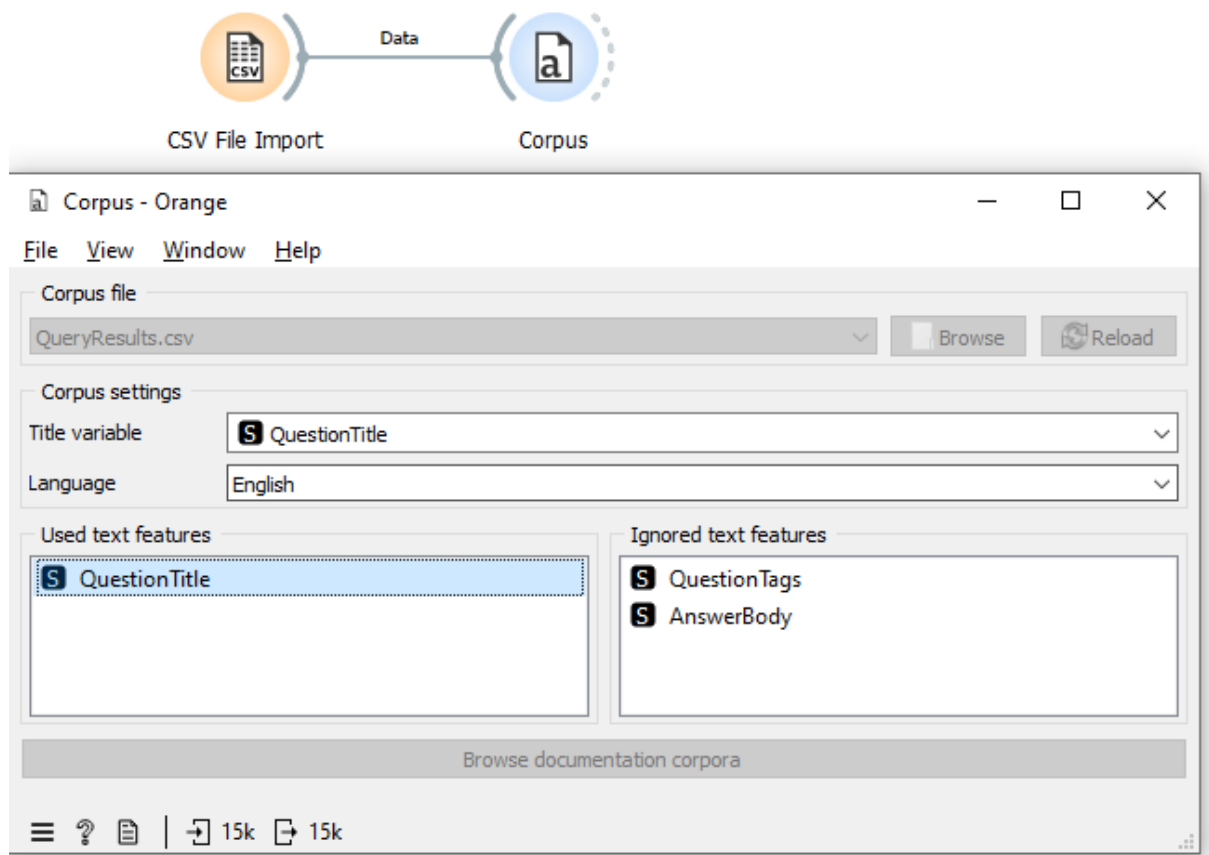
## 4.5 Topic Modeling

Line Charts and Histograms are valuable tools for analyzing the distribution of questions and answers over the years. They provide essential insights into the progress of questions-answers on Stack Overflow over the years. However, these visualizations, while informative, focus on quantitative aspects and don't dive into the content of the questions or answers themselves. To address this critical aspect of analysis, the utilization of topic modeling tools becomes imperative.

Topic modeling techniques enable a more in-depth exploration of the content within the questions, unveiling the underlying themes and recurring subjects in the data. This is especially advantageous when seeking a more abstract and thematic perspective on the content.

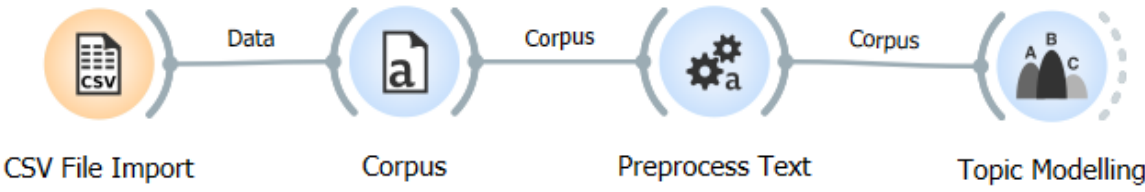
With the text add-on, Orange provides tools for text analysis tasks, allowing to preprocess, analyze, and visualize textual data efficiently.

To complete the topic modeling process, it is essential to connect the "CSV File Import" widget to the "Corpus" widget. To ensure the production of valuable results and trends, special attention should be given to the "QuestionTitle" variable, which is thoughtfully selected within the Corpus settings:

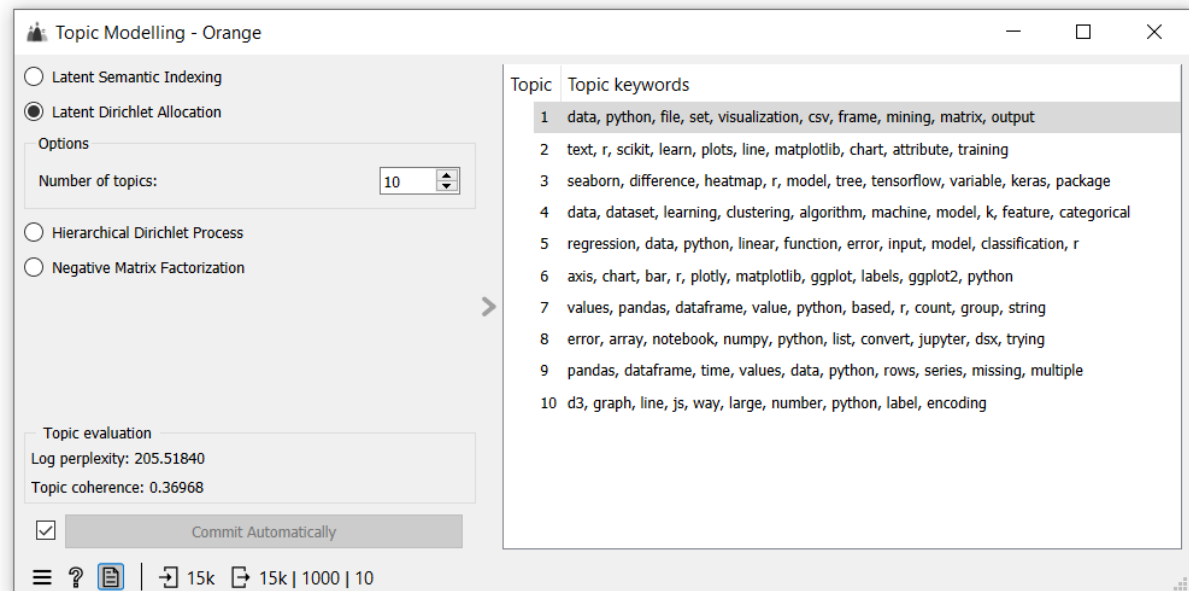
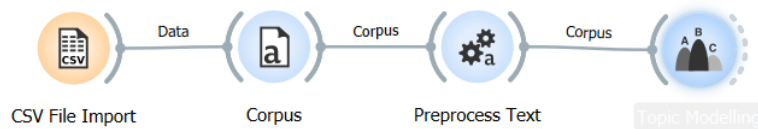


For a thorough and effective topic modeling analysis, it is crucial to preprocess the data accurately. The "Preprocess Text" widget empowers users with the ability to perform various data transformations. This includes options to either retain or remove stopwords, numbers, HTML tags, or URLs from the text.

Following preprocessing, the next step involves connecting the "Preprocess Text" widget to the "Topic Modeling" widget:



The "Topic Modeling" widget provides users with a range of options, from Latent Semantic Indexing to Latent Dirichlet Allocation (LDA) and Hierarchical Dirichlet Process. To identify the topic keywords and gain a better understanding of the ten most prevalent topics found in Stack Overflow questions, Latent Dirichlet Allocation (LDA) is chosen. The results are as follows:



After analyzing these ten topics, we can draw several conclusions about their content:

**Topic 1:** This topic primarily focuses on data-related tasks, with a particular emphasis on Python. The presence of words such as "file", "set", "visualization" and "csv" indicates discussions related to data management, analysis, and visualization using Python.

**Topic 2:** This topic revolves around text analysis and machine learning, with a strong focus on Python. Words like "text", "scikit", "learn" and "matplotlib" suggest discussions related to natural language processing (NLP) and data visualization tasks in Python.

**Topic 3:** This topic is closely associated with data visualization and analysis, specifically utilizing Python libraries like Seaborn and Matplotlib. The inclusion of terms such as "seaborn", "heatmap" and "matplotlib" implies conversations regarding the creation of data visualizations and charts.

**Topic 4:** This topic encompasses discussions related to machine learning and data analysis. The terms "data", "dataset", "learning", "clustering" and "algorithm" indicate a broad exploration of machine learning aspects, including data preprocessing and model development.

**Topic 5:** This topic seems to be focused on regression analysis, particularly using Python. Words like "regression", "linear", "function" and "classification" point to discussions on regression modeling and classification tasks in Python.

Topic 6: This topic is centered around data visualization using Python, with terms like "axis", "chart", "bar", "plotly" and "matplotlib" suggesting conversations related to the creation of various charts and graphs.

Topic 7: This topic revolves around data manipulation using Python's Pandas library. Terms like "values", "pandas", "dataframe" and "count" indicate discussions on working with data frames and performing data-related operations.

Topic 8: This topic appears to involve error handling and debugging in Python. Terms like "error", "array", "notebook", "numpy" and "jupyter" imply discussions related to identifying and resolving errors in Python code.

Topic 9: This topic appears to be related to time series data analysis and data manipulation using Python's Pandas library. Terms like "pandas", "dataframe", "time" and "values" suggest discussions on handling time series data.

Topic 10: This topic focuses on web development, particularly using D3.js to create interactive web graphics. Terms like "d3", "graph", "line" and "js" indicate discussions on web-based data visualization and web development using JavaScript.

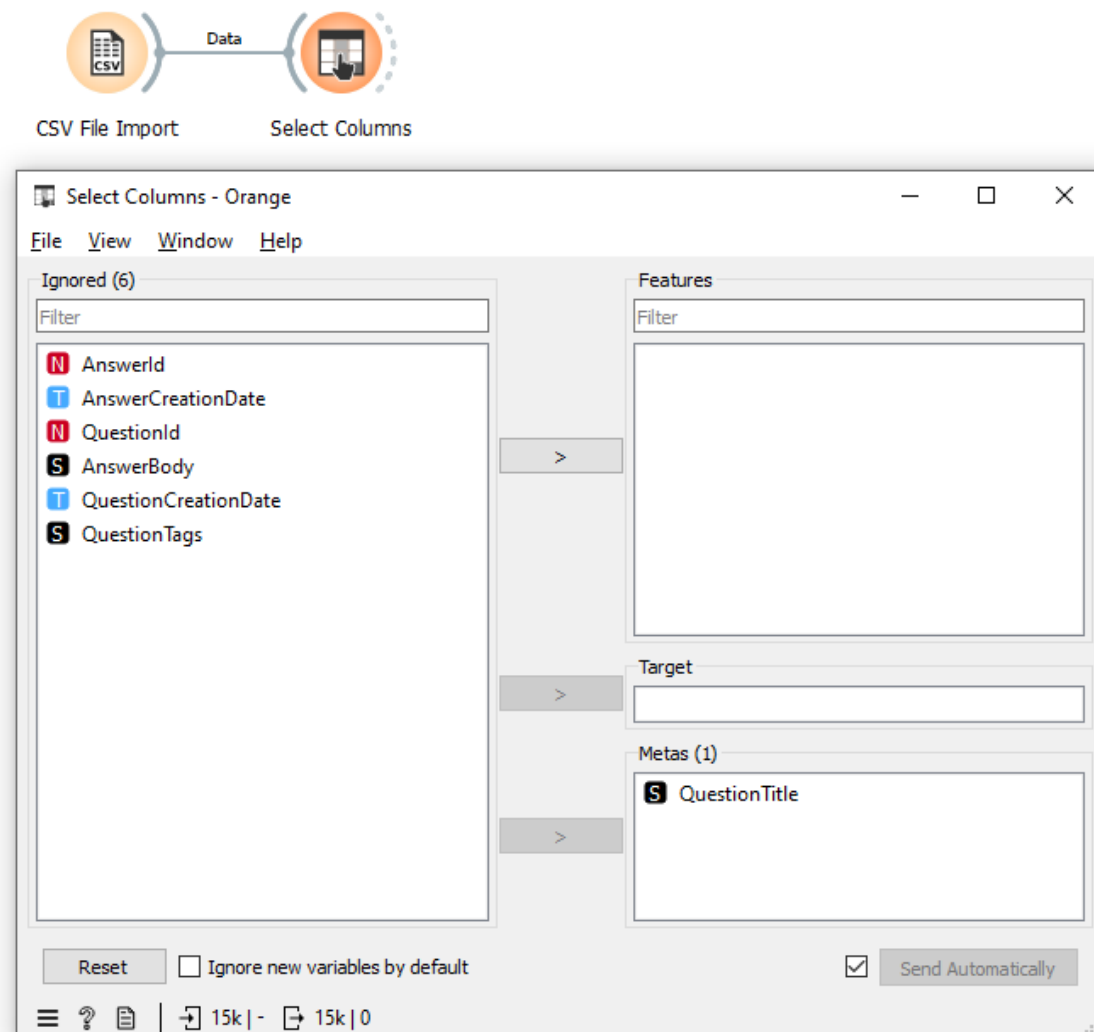
These topics show that people who work with data have different interests. Some topics focus on machine learning and predicting things, while others are about studying data that changes over time. Python's role is optimal in most of these topics because it's used for both analyzing data and making visuals like charts. This indicates how versatile Python is for solving different data problems. But, besides Python, there are also discussions about other tools. Some talk about using Javascript for showing data on the web, and others mention R, which is a statistics language still widely used in data analysis.

## 4.6 Word Cloud for question titles

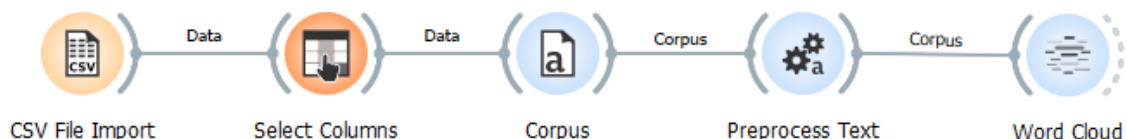
Another widely used and effective method for analyzing text is the generation of word clouds. These word clouds provide visual representations of the most frequently occurring words or terms within a body of text. They are valuable tools for quickly identifying and visually presenting the primary themes and trends in textual data, offering a more nuanced perspective on the evolving discussions within the Stack Overflow community.

A word cloud generated from the question titles can effectively depict the key words and topics that pique the interest of the Stack Overflow community. For optimal results, it's crucial to isolate the text column to be analyzed. Otherwise, keywords from other text columns, such as "QuestionTags" and "AnswerBody" may be inadvertently included in the results, making it challenging to draw meaningful conclusions.

Columns can be isolated by utilizing the "Select Column" widget to select specific columns from the CSV file.



By connecting the "Select Column" widget to the "Corpus" and subsequently to "Preprocess Text" and finally to "Word Cloud" it becomes possible to effectively isolate and analyze these textual elements.







Weight	Word
2475	data
1897	python
1640	pandas
1482	using
1460	column
1308	dataframe
1159	r
1143	values
1119	plot
768	value
706	columns
629	chart
624	multiple
588	error
506	based
490	one
490	function
487	two
486	list
478	rows
476	create
442	time
441	axis
433	different
429	matplotlib
422	bar
415	file
406	model
405	use
386	dataset
377	line
374	get
371	graph
367	plotly
363	x
353	series
352	way
341	add
330	find
325	number

Notably, the word "data" carries the most significant weight (2475), which is unsurprising, given that all questions and selected tags are closely related to data.

An intriguing finding is that "Python" is the second most commonly searched word (weight 1897), followed by "pandas" (weight 1640), a Python library renowned for its data manipulation and analysis capabilities. This highlights the pivotal role of Python and its associated tools in data applications. Python's versatility, extensive library support, and user-friendly syntax make it a top choice for data scientists and analysts worldwide.

Additionally, the presence of "R" as the seventh most prominent word in the word cloud signifies its sustained popularity among data professionals. R is known for its statistical analysis capabilities and visualization tools, making it indispensable for many data scientists. The coexistence of both Python and R within the top word rankings underlines the diversity and richness of the data science and analytics landscape, where practitioners often leverage the strengths of both languages.

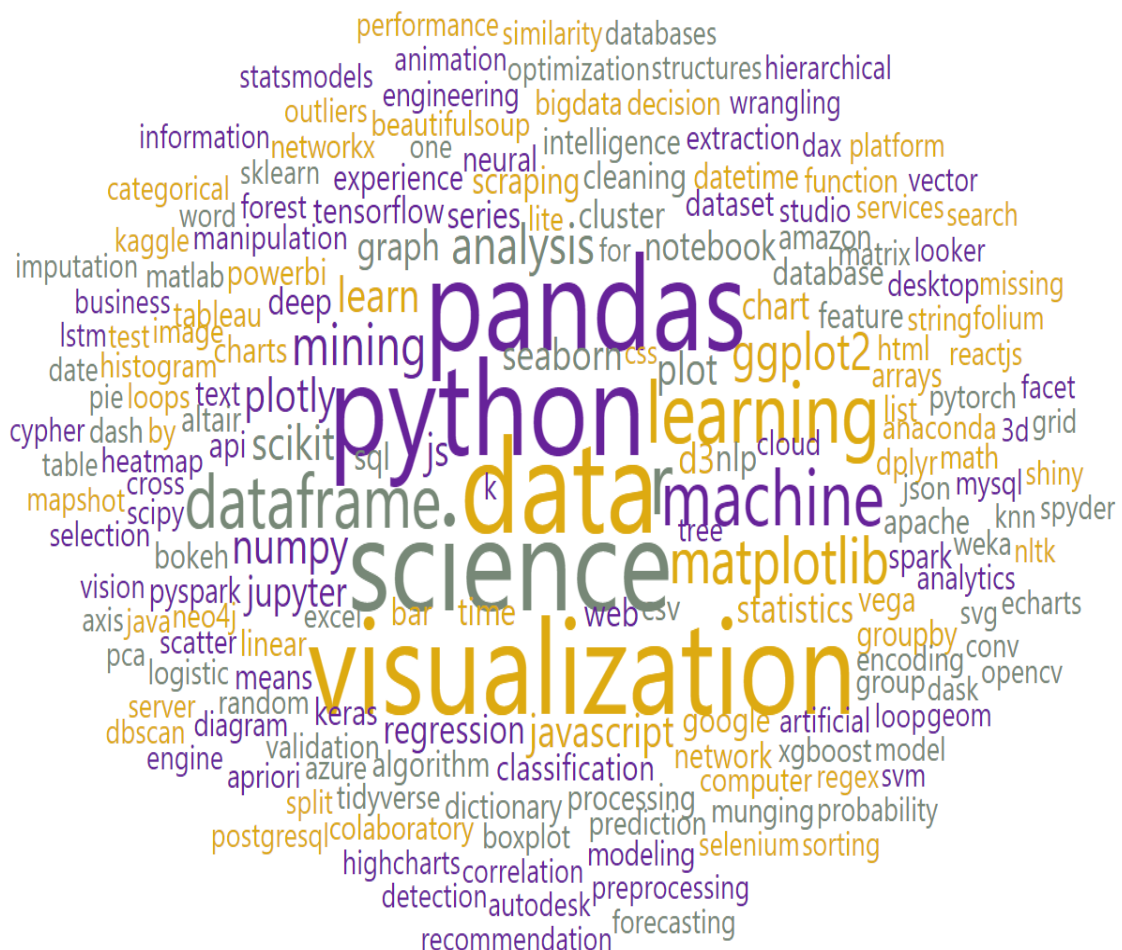
Terms like "column", "dataframe", "values", "value" and "plot" round out the top 10 words in the word cloud. These terms underscore their significant relevance in the data science and data analytics domains. "Column" and "dataframe" are fundamental components of data structures, integral for organizing and manipulating data. "Values" and "value" likely reflect the importance of data values and their analysis in this context. "Plot" suggests the

widespread utilization of data visualization techniques, emphasizing the importance of conveying insights through visual representations.

## 4.7 Word Cloud for question tags

Following the word cloud analysis of question titles, insights have been gathered regarding the most frequently used terms within the Stack Overflow community. The next step is to explore another facet of the dataset: the tags. A word cloud can be generated, employing the same methodology as used for the question titles analysis to reveal the most prevalent tags. This approach ensures continuity and consistency, offering a clear picture of the current dominant themes and topics within the Stack Overflow community.

Below is the word cloud for the QuestionTags:



Again, the inclusion of the "Words and Weights" section is essential for a more in-depth understanding of the data at hand.

Weight	Word
16689	data
9942	python
9414	science
5287	visualization
4157	pandas
2372	r
2205	learning
1925	machine
1759	dataframe
1289	matplotlib
1050	mining
950	analysis
845	ggplot2
722	numpy
711	scikit
706	learn
664	plotly
557	plot
556	javascript
541	js
504	seaborn
437	d3
413	jupyter
381	graph
352	statistics
342	regression
330	notebook
319	chart
296	web
294	nlp
271	time
267	series
265	cluster
230	scraping
222	deep
210	sql

It's expected that "data", "science" and "visualization" are among the top 5 tags, as all of them were included in the SQL query.

One particularly significant finding from this word cloud is that "python" is the second most frequently used tag, surpassing all others except "data" itself. "Pandas" holds the 5th position, with nearly 1800 more mentions than the 6th most common tag, which happens to be the statistical programming tool "R". "Machine learning", "dataframe" and "matplotlib" complete the 10 most common tags

Notably, out of the 10 most frequently used tags, two of them, "python" and "R" are full-fledged programming languages, underscoring their significance in the data-related discussions. Moreover, two other tags, "matplotlib" and "pandas" are Python libraries, demonstrating the integral role of Python in data science.

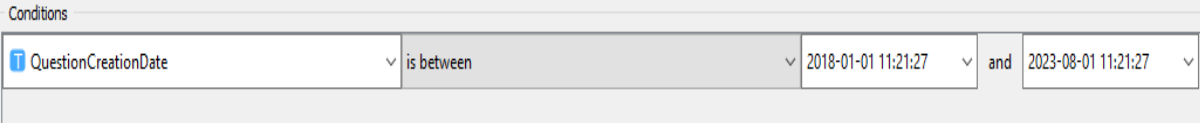
## 4.9 Python and R: Trends analysis

The word cloud analysis has underscored the significance of Python and R as the primary tools for data-related tasks. Expanding on these findings, a more in-depth analysis of the trends surrounding these tools offers users the chance to track the evolution of Python and R over the years.

Given the interactive nature of word clouds, there is the capability to isolate and analyze all data containing each programming language, allowing users to observe trends spanning multiple years.

It's important to acknowledge a potential limitation in this kind of analysis. The selected dataset encompasses questions from 2009 to 2023; however, it's important to note that it specifically includes questions with answers from the period of 2018 to 2023. When investigating trends, it's crucial to focus on this 2018-2023 timeframe because earlier years' data may have limited utility. This is because questions tagged with our selected programming languages and libraries may be excluded due to the absence of answers within the 2018-2023 period. The exclusion of such questions is logical, as many years may have passed between their creation and the 2018-2023 period. Additionally, even within the 2018-2023 timeframe, there may be questions without answers, which would also be absent from our dataset. However, despite these considerations, we can effectively identify and analyze trends for questions created during the specified 2018-2023 period, making the most of the available data.

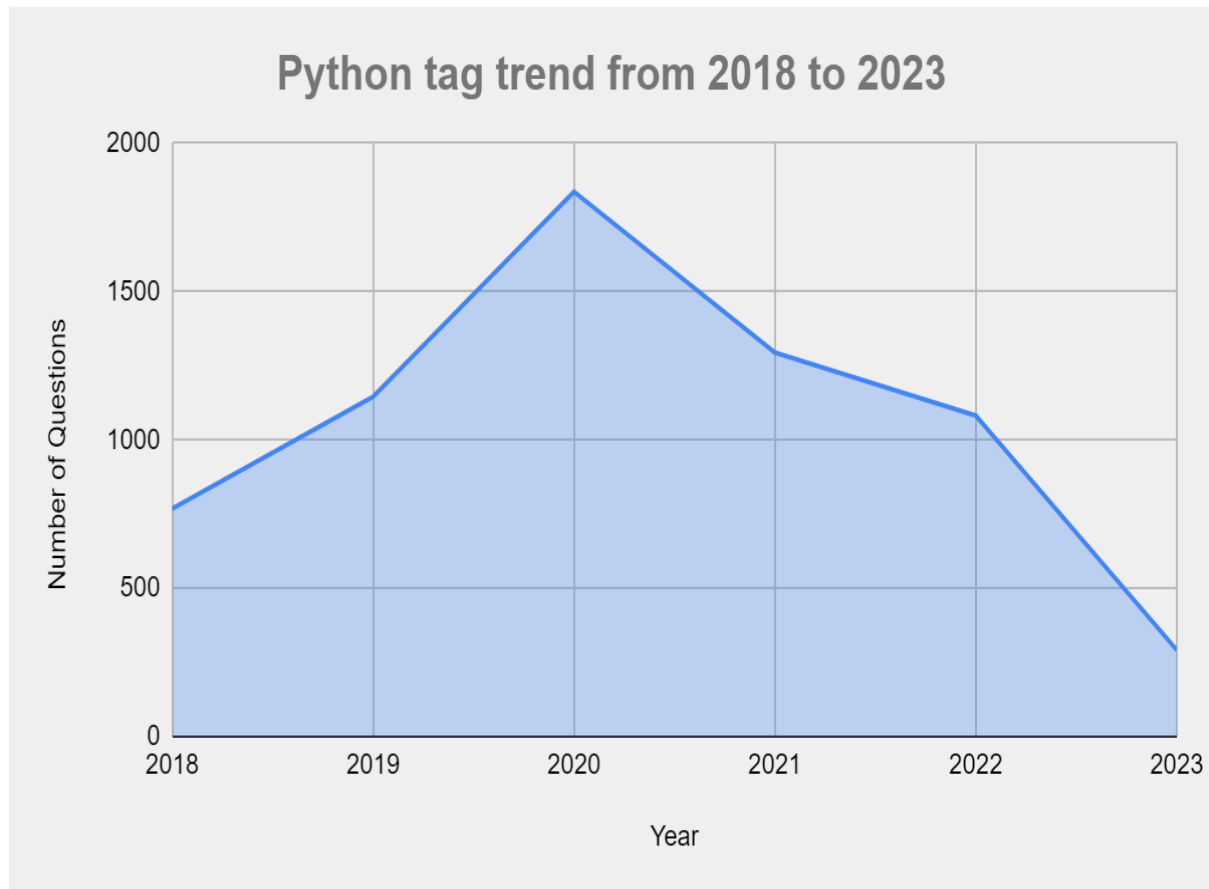
To isolate the data for questions created between 2018 and 2023, the Word Cloud can be connected with the "Select Rows" widget. The condition for selection is straightforward:



Conditions				
QuestionCreationDate	is between	2018-01-01 11:21:27	and	2023-08-01 11:21:27



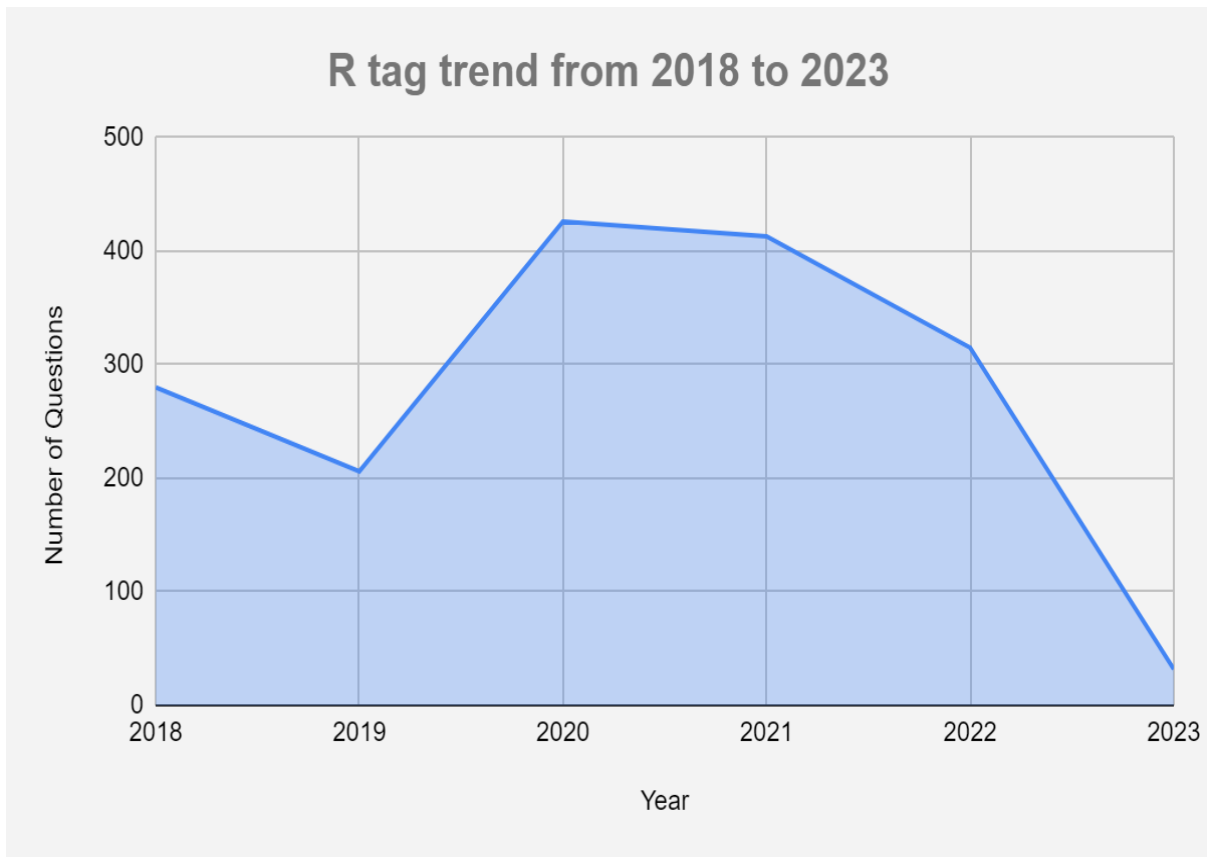
Following the removal of duplicates, the new table consists of 6,420 unique questions. The trend analysis over time is as follows:



The analysis of Python trends over the years reveals a pattern. From 2018 to 2020, Python experienced a robust upward trend, with a significant increase in questions related to the language. In 2018, there were 768 Python-related questions, and this number steadily rose to 1,837 in 2020, marking a substantial increase in community engagement with Python. However, as years progress from 2020 to 2023 there is a different trajectory. The number of Python-related questions declined, reaching 1,295 in 2021, 1,082 in 2022, and 292 in 2023. This decrease suggests a shift in community interest or a possible saturation of certain Python-related topics.

Comparing these Python trends with the broader activity on Stack Overflow in terms of questions and answers, some interesting insights emerge. From 2018 to 2020, the number of Python questions was on the rise, mirroring a period of heightened user engagement. This period also saw a substantial increase in the number of answers provided, reinforcing the idea that the community was highly responsive to Python-related queries during this time. However, starting from 2021, we witness a decline in both the number of Python-related questions and answers, which coincides with the overall trend.

Below the similar analysis for the programming language R:



From 2018 to 2020, R exhibited a steady increase in questions related to the language. In 2018, there were 280 R-related questions, and this number rose to 426 in 2020, indicating a growing level of interest and engagement with R. However, similar to the Python trend analysis, a different trajectory emerges from 2021 to 2023 for R. During this period, the number of R-related questions declined, with 413 questions in 2021, 315 in 2022, and a significant drop to 32 in 2023.

When comparing Python and R trends, some similarities and differences become apparent. Both of them experienced a period of growth from 2018 to 2020, indicating increased interest in these languages. However, both languages experienced a decline in the number of questions from 2021 onward, aligning with the overall trend of questions and answers in Stack Overflow. Python consistently maintained higher engagement levels throughout the analysis period, reinforcing its dominant position as the preferred language within the Stack Overflow community.

# Chapter 5: Conclusions, limitations & future work

In this chapter, conclusions are drawn based on the various analyses conducted in this thesis. The insights have been derived from an examination of Stack Overflow data, which was obtained through SQL queries, processed in excel, analyzed using Orange data mining software and specifically natural language processing tools.

## 5.1 Research limitations

The limitations of this thesis arise from its specific data extraction criteria. The dataset includes Stack Overflow questions spanning the years 2009 to 2023, with the condition that each question must have received at least one answer between 2018 and 2023.

This filtering process led to the exclusion of questions created both before and during the specified 2018-2023 timeframe, provided they did not receive answers within that period. Notably, questions created prior to 2018, especially those from 2009 to 2018, might be more likely to lack answers within the 2018-2023 time frame due to the substantial time gap between their creation and the analysis period's onset.

As a result, the dataset may not present a comprehensive view of trends from 2009 to 2018 due to these inclusion criteria. Nevertheless, it remains a valuable resource for understanding evolving trends and community interests from 2018 onwards.

## 5.2 General Conclusions

### 5.2.1 Temporal Analysis

The temporal analysis focused on the creation dates of questions and answers. While the dataset spanned from 2009 to 2023, the analysis concentrated on questions created between 2018 and 2023 to highlight contemporary trends. This filtering approach was used to ensure relevance and ongoing discussion.

The analysis revealed that there was limited activity on Stack Overflow from 2009 to 2017, in line with the SQL query's criteria. However, starting from 2017, the creation of questions and answers increased significantly, reaching a peak in 2020. There was a decline in both questions and answers in 2021 and 2022, with a further reduction in 2023.

This reduction in activity from 2020 to 2023 is significant because it serves as a reminder that trends in user engagement can be transient, and they may not always follow a linear or predictable pattern. It underlines the importance of continuously monitoring and analyzing user activity to stay attuned to the evolving needs and interests of the community.



The percentage difference analysis from the peak year of 2020 showed the extent of the increase and subsequent decline in both questions and answers.

For questions, the percentage difference from 2020 peaked at 93.80% in 2017.

For answers, the percentage difference from 2020 peaked at 80.30% in 2023. Notably, the year 2023 is still ongoing, and it remains uncertain whether the total year difference will surpass the 45.20% observed in 2018.

## 5.2.2 Topic Modeling

The topic modeling analysis has unveiled ten distinctive topics within the content of questions and answers, providing valuable insights into the areas of interest within the Stack Overflow community.

Python emerges as a central and recurring theme in several of these topics. This prominence emphasizes Python's versatility and widespread adoption in various data-related contexts. It is evident that it plays a pivotal role in addressing diverse data challenges. Python's multifaceted applications are evident, spanning from data visualization using libraries like Seaborn and Matplotlib to machine learning and text analysis employing tools such as scikit-learn. This trend is also consistent with the findings from the related work by Tahmooresi, Heydarnoori, and Aghamohammadi in 2020.

However, Python is not the sole focus. Another significant topic that surfaces is related to web development, particularly the use of JavaScript and D3.js for creating interactive web graphics. This reflects the Stack Overflow community's interest in web-based data visualization and web development using JavaScript.

Additionally, the prominence of machine learning and data analysis-related terms within the identified topics signifies the community's strong interest in this domain. Terms such as "data", "dataset", "learning", "clustering", and "algorithm" are indicative of the community's active engagement in discussions related to machine learning, data preprocessing, and model development.

The observed decline in the volume of questions and answers pertaining to data science, data visualization, data mining, and data analytics over time may suggest a notable shift in the community's focus towards machine learning and web development. This shift, away from other data-related topics, happens together with a concurrent rise in questions and answers about machine learning and web development. These trends collectively indicate that machine learning and web development are gaining prominence as subjects of interest and expertise within the Stack Overflow community.

### 5.2.3 Word Cloud

The word cloud analysis of tags unveils significant trends within the Stack Overflow community. Among the most commonly used tags, "python" and "R" stand out as programming languages, showcasing their pivotal roles in data-related discussions. Python, in particular, demonstrates its versatility and broad applications, from data analysis to machine learning. Additionally, Python's libraries, such as "pandas" and "matplotlib" feature prominently, emphasizing their essential contributions to data science tasks and data visualization.

Furthermore, "machine-learning" emerges as a highly frequent tag, indicating a substantial interest and engagement in the field. This highlights the Stack Overflow community's dedication to exploring the intricacies of machine learning, from data preprocessing to model development. The prominence of "machine-learning" in the tag cloud not only underscores this trend but also reaffirms the conclusions drawn from the topic modeling analysis.

This alignment between the insights from the tags word cloud and the findings of the topic modeling section underscores the community's consistent and substantial focus on the domain of machine learning and further validates its significance in data-related conversations.

## 5.3 Used tools

Python stands as the dominant player, with a multitude of libraries extensively utilized. Among these, "Pandas" takes center stage, enabling data manipulation and analysis with unparalleled ease. Similarly, "Matplotlib" and "Seaborn" have become the go-to libraries for data visualization, offering the means to create informative and compelling charts and graphs. Python's versatility extends to "Jupyter" a powerful interactive computing environment that facilitates data exploration and visualization, aligning with the project's objective of analyzing trends and developments in data science.

R is a solid second choice, maintaining its relevance over time. R's influence extends to the "ggplot2" library, renowned for its comprehensive data visualization capabilities.

Within the context of web-based data visualization and related development, "JavaScript" emerges as an integral tool. This aligns with the shared interest in web development uncovered by the analysis. JavaScript's role in crafting interactive web graphics, coupled with libraries such as "D3.js" highlights its relevance in enhancing the data visualization landscape.

## 5.4 Future work

The temporal analysis uncovered a significant decline in questions and answers related to data science and data analytics, particularly from 2021 onward. In contrast, the topic modeling revealed the presence of more than one topic centered around machine learning. The terms "machine" and "learning" also prominently feature among the top 10 tags used in questions related to data science and data analytics.

This convergence of evidence suggests a possible shift towards machine learning within the Stack Overflow community. Future research can delve deeper into this machine learning trend. A comprehensive study could aim to uncover the reasons behind it, whether it's driven by technological advancements, evolving industry demands, or educational initiatives. Exploring the nature of questions and answers related to machine learning, along with their temporal trends, can provide valuable insights into the community's evolving needs and interests.

For Python, researchers can investigate the specific challenges, issues, and queries that data scientists and analysts encounter. This analysis could encompass topics such as optimizing Python code for data processing, addressing common pitfalls, and exploring advanced techniques in data manipulation and visualization. By pinpointing areas of difficulty and interest, this research can contribute to the development of comprehensive solutions and resources tailored to the Python community's needs.

Similarly, an in-depth exploration of the R ecosystem can provide insights into the challenges faced by statisticians and data analysts. This could encompass topics like advanced statistical modeling, data visualization with "ggplot2" and best practices in R package development. By addressing the unique requirements of R users, future research can enhance the toolkit available to the community.

Future research could also explore the intersection of these areas. This could involve analyzing machine learning libraries, frameworks, and tools within the Python and R ecosystems. Researchers could identify the most commonly used libraries for machine learning tasks, understand the specific challenges users encounter, and assess the effectiveness of different approaches. By discovering the synergies between machine learning and these programming languages, research can provide valuable guidance and best practices for data scientists and analysts seeking to leverage Python and R for machine learning projects.

# References

1. Tahmooresi, H., Heydarnoori, A., & Aghamohammadi, A. (2020). An Analysis of Python's Topics, Trends, and Technologies Through Mining Stack Overflow Discussions. *arXiv preprint arXiv:2004.06280*.
2. Kochhar, P. S. (2016, September). Mining testing questions on stack overflow. In *Proceedings of the 5th International Workshop on Software Mining* (pp. 32-38).
3. <https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/>
4. [A Brief History of Data Science - DATAVERSITY](#)
5. Cao, L. (2017). Data science: a comprehensive overview. *ACM Computing Surveys (CSUR)*, 50(3), 1-42.
6. Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, 56(12), 64-73.
7. [Empowering the world to develop technology through collective knowledge - Stack Overflow](#)
8. Moutidis, I., & Williams, H. T. (2021). Community evolution on stack overflow. *Plos one*, 16(6), e0253010.
9. Sengupta, S., & Haythornthwaite, C. (2020). Learning with comments: An analysis of comments and community on Stack Overflow.
10. Barua, A., Thomas, S. W., & Hassan, A. E. (2014). What are developers talking about? an analysis of topics and trends in stack overflow. *Empirical software engineering*, 19, 619-654.
11. Dada, O. A., Obaido, G., Sanusi, I. T., Aruleba, K., & Yunusa, A. A. (2022). Hidden gold for it professionals, educators, and students: Insights from stack overflow survey. *IEEE Transactions on Computational Social Systems*, 10(2), 795-806.
12. <https://orangedatamining.com/>
13. Brady, H. E. (2019). The challenge of big data and data science. *Annual Review of Political Science*, 22, 297-323.
14. Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National science review*, 1(2), 293-314.
15. Sagiroglu, S., & Sinanc, D. (2013, May). Big data: A review. In *2013 international conference on collaboration technologies and systems (CTS)* (pp. 42-47). IEEE.
16. Tsai, C. W., Lai, C. F., Chao, H. C., & Vasilakos, A. V. (2015). Big data analytics: a survey. *Journal of Big data*, 2(1), 1-32.
17. <https://www.analyticsvidhya.com/blog/2021/06/part-2-topic-modeling-and-latent-dirichlet-allocation-lda-using-gensim-and-sklearn/>
18. Baltadzhieva, A., & Chrupala, G. (2015). Question quality in community question answering forums: a survey. *Acm Sigkdd Explorations Newsletter*, 17(1), 8-13.
19. Anderson, A., Huttenlocher, D., Kleinberg, J., & Leskovec, J. (2012, August). Discovering value from community activity on focused question answering sites: a case study of stack overflow. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 850-858).
20. Chowdhary, K., & Chowdhary, K. R. (2020). Natural language processing. *Fundamentals of artificial intelligence*, 603-649.
21. Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science*, 349(6245), 261-266.