

**Faculty of Natural Sciences**

**School of Computer Science and Mathematics**

**Visualisation for Data Analytics (CSC-40048)**

**COURSEWORK 2**

**Group 10**

|  |  |  |
| --- | --- | --- |
| **No.** | **Student Number** | **Student Name** |
| 1 | 24007706 | Benjamin Mensa-Bonsu |
| 2 | 23045694 | Kelvin Thumbi Macharia |
| 3 | 23033965 | Shaokun Guo |

**Question 1 and Question 2.**

**Dataset Information.**

As a group we decided to focus on the health domain, specifically, issues pertaining to Mental Health. Mental health significantly impacts overall well-being and quality of life, making it a critical aspect of healthcare. We downloaded a *Mental Health Dataset* from Kaggle.com. The dataset had 292,364 rows and 17 columns. It contains various attributes such as Timestamp, Gender, Country, Occupation, Self-employed (Are you self-employed?), family history (Do you have a family history of mental illness?), treatment (Have you sought treatment for a mental health condition?), Days Indoors, Growing Stress, Changes in Habit, Mental Health History (Do you personally have a history of mental health illness?), Mood Swings, Coping Struggles, Work Interest, Social weakness, mental health Interview (Would you bring up a mental health issue with a potential employer in an interview?) and care options . With the exception of Timestamp which had a datetime data type, all other attributes were of the string data type. This dataset was chosen because it encompasses a broad spectrum of linguistic, psychological, geographical and behavioural characteristics, making it appropriate for the analysis of topics connected to mental health. It provides valuable insights into mental health that can be used to build predictive models to identify or predict mental health outcomes based on textual data.

**Question 3.**

**Data Pre-Processing:**

An essential first step in data analysis is data preprocessing. Prior to additional processing and analysis, raw data must be cleaned, transformed, and arranged. Reliable insights, quality, correctness, and usability of data are ensured by proper preprocessing. Approximately 45% of a data scientist's time is dedicated to data processing [1].

Importance of Data Preprocessing

1. Data Quality Improvement: Preprocessing finds and fixes mistakes, anomalies, and inconsistencies in the data.
2. Improved Accuracy: Analysis and modelling outcomes are more accurate when the data is clean and preprocessed.
3. Resource Efficiency: Proper preprocessing reduces the time and resources needed for analysis.
4. Enhanced Usability and Portability: In a data landscape dominated by unstructured data, data preprocessing becomes important to facilitate seamless analysis and ensure data's ease of use and portability. [2]

During the data preprocessing, we did the following;

**Extracting Responses from the United Kingdom only.**

The dataset was from 35 countries. Since the dataset was big, we concentrated on responses from the UK only which had 51,404 data objects. This is closer to the required dataset size of 40,000 rows. We proceeded to check for null values and duplicates.

**Checking for Null Values**

Missing data points within a dataset are referred to as null values, and they can have an effect on the accuracy and dependability of data analysis.

Managing null values is important for a number of reasons.

1. Null values cause gaps in the dataset, which makes it difficult to obtain a complete picture of the information at hand and contributes to data incompleteness.
2. Analysing data that contains null values can distort findings and impair the precision of analytical models and forecasts.
3. Null values restrict the value that can be extracted from the dataset and impede the interpretation of data by making it challenging to obtain significant insights and conclusions from the data.

There are two approaches to handling null values in data:

1. Imputation: This involves filling missing values using methods like mean, median, or predictive modeling to estimate what the missing data could be.
2. Deletion: If the number of null values is small and won't greatly affect the overall data analysis, the null values can be removed by deleting the rows containing these null values.

For our case, we chose to use deletion, this is because the null values in our dataset was only 1.517% which is small. This does not have a significant impact on our dataset.

**Checking for Duplicates**

Handling duplicates is dealing with records that are exactly the same. It's important to manage duplicates because they can mess up the analysis results and take up unnecessary space. By getting rid of duplicates, we make sure our data is clean and reliable for making decisions. The dataset had 48 rows as duplicates which represented only 0.09% of the dataset, so we deleted all the duplicates.

**Removing Timestamp**

We also deleted the timestamp column which only represented the date of data collection and was not needed for analysis. After the data-preprocessing the new dataset had 50576 rows and 16 columns.

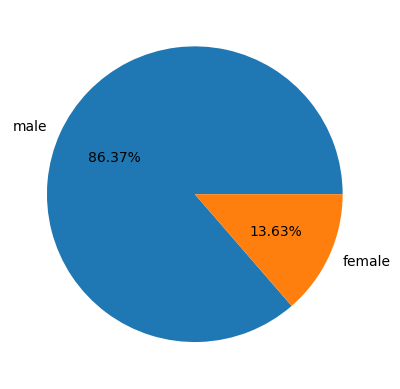
**Question 4**

For the visualizations, we used the pandas, Matplotlib.plt and Seaborn Libraries.

**Gender Distribution**

*The code:*

****



In our dataset**,** 86.4% were male and only 13.6% were female. All the respondents were from the United Kingdom

**Occupation Distribution by Gender**

*The code:*

****

A graph of people with their age

Description automatically generated with medium confidenceAmong women, students represent the largest group at 22.7%, closely followed by housewives at 22.3% out of the total 13.6% of women surveyed. Conversely, most men fall into the category of "Housewives (men)" at 22.8%, with men working in corporate roles following closely behind at 21.8%.

**Gender disparities in Mental Health Cases**

How does mental health problems affect males and females? Research shows that males and females are affected by mental health issues, but not equally. About one in five women and one in eight men in England suffer from common mental health problems [3]. To test this theory in our dataset, we compared the Mental Health History against the two genders.

A graph of a bar graph

Description automatically generated with medium confidence

**Results:** From our data, the females display a 36.2% rate of uncertain responses ("Maybe"), while men exhibit a 31.8% rate. This indicates a significant level of uncertainty among individuals regarding their mental health history. However, males have a slightly higher percentage of "No" responses at 36.2%, suggesting a lower reported prevalence of mental health issues compared to females, who have a rate of 33.2%. Additionally, males show a slightly higher percentage of "Yes" responses at 32.1%, indicating a slightly higher prevalence of confirmed mental health issues compared to females, who have a rate of 30.6%. This percentages are so close suggesting no major difference of effects of mental health in both genders.

**Mental health history by gender and proportions of individuals seeking treatment.**

Considering that mental health affects both women and men at nearly equal rates in our data, we conducted an in-depth analysis to explore how each gender seeks treatment. Our approach involved creating bar plots that juxtaposed mental health history against treatment-seeking behavior, with a specific focus on gender-based comparisons.

The data shows mental health history by gender and their percentages for seeking treatment reveals several key insights. When examining the "Yes" category for seeking treatment, females and males have percentages of 19.1% and 14.9%, respectively. While both genders show a need for treatment due to confirmed mental health histories, females exhibit a slightly higher percentage of seeking treatment unlike men where despite them having mental History, majority chose not to seek treatment.

A graph of different colored bars

Description automatically generated with medium confidence

In terms of the "No" category for seeking treatment, both genders show varying percentages: females at 20.7% and males at 16.8%. This indicates that a significant portion of females seeking treatment, do not have a history of mental health issues.

Moving to the "Maybe" category for seeking treatment, females have a higher proportion at 22.6% compared to males at 14.8%. This suggests that a larger percentage of females seeking treatment are uncertain or potentially facing mental health issues compared to males.

Research into men's attitudes towards seeking help for mental health, particularly depression, underscores the impact of traditional masculine norms [4]. These norms, characterized by traits like strength, self-reliance, emotional control, and avoidance of vulnerability, often discourage men from seeking assistance [4]. The fear of being perceived as weak or unmanly due to seeking support contributes to the stigma surrounding mental health among men [5]. This stigma can lead to reluctance in discussing mental health issues and seeking professional help [5].

Interestingly, exposure to mental health services can lead to shifts in men's attitudes towards help-seeking [5]. Men who utilize mental health services may develop a more positive perspective on mental health and recognize the importance of addressing their well-being [5]. Additionally, support groups tailored to men offer a secure setting for discussing vulnerabilities, exchanging stories, and enhancing self-worth [5]. These organizations are crucial in enabling men to question social standards around mental health and masculinity and to seek assistance.

**Recommendations**

Reducing the stigma attached to asking for help is the key to improving males seeking help behaviors [5]. Men have to realize that asking for assistance is a show of strength rather than weakness. Empowering men to recognize their mental health needs and promoting professional support without fear are crucial steps [5]. Additionally, healthcare providers, especially general practitioners, should receive training to address the impact of masculine norms on mental health and facilitate positive help-seeking behaviors among men [5].

**Comparison of Mental Health History in different Occupation**

How do mental health histories vary among different occupations? To answer this question, we compared the mental health history against each occupation present in our dataset.

*The code:*



**A graph showing mental health history

Description automatically generated**

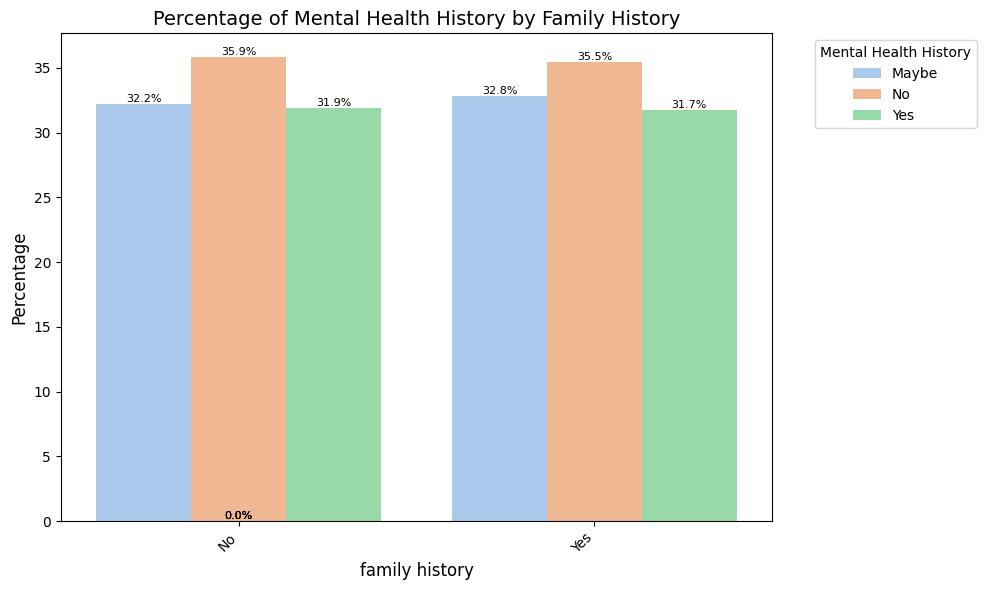
**Results:** The graph shows the percentage of mental health history by occupation reveals intriguing insights into the prevalence of mental health issues within different professional sectors. In the business sector, approximately 25.9% of individuals have been uncertain of their mental health history (Maybe), while 32.5% do not have a mental health history (No), and 41.6% have a confirmed mental health history (Yes). Similarly, in corporate roles, roughly 31.6% are uncertain of their mental health history (Maybe), 30.5% report no history (No), and 37.9% confirm a mental health history (Yes). When it comes to "Maybe," "No," and "Yes," housewives have percentages of 30.7%, 33.7%, and 35.6%, respectively. 27.3%, 35.6%, and 37.1% of workers in other occupations select "Maybe," "No," and "Yes." The percentages that students display for "Maybe," "No," and "Yes" are 29.3%, 33.1%, and 37.6%.

These results show the disparities in mental health histories among various occupational categories, emphasizing the necessity of providing specialized mental health interventions and support systems for particular occupations.

#Ben Add evidence

**Comparison of Individuals mental health history based on family Mental history.**

To check how family mental health background influences individuals' mental health perspectives, plotted Family history against individual mental health history.



**Results:** This graph sheds light on how family background influences individuals' mental health perspectives. For individuals without a family history of mental health issues, the breakdown reveals interesting insights. Approximately 31.9% of this group have a mental health history, while about 35.9% are uncertain (Maybe) about their mental health history. Surprisingly, around 32.2% do not have a mental health history, indicating that family history has little influence on their mental health status.

Conversely, for those with a family history of mental health issues, the percentages showcase a delicate picture. Approximately 31.7% of individuals in this category do not have a mental health history, challenging the assumption of direct hereditary impact on mental health. Furthermore, roughly 32.8% of respondents are unsure (Maybe) about their mental health history, whereas about 35.5% have a proven mental health history (Yes).

These results demonstrate the complexity of mental health histories and imply that, although familial history can be significant, it is not the only factor in determining an individual's mental health. This idea is supported by a number of studies and research findings, which also highlight how diverse mental health determinants are. For instance, a study titled “The interplay of family history of depression and early trauma: associations with lifetime and current depression in the German national cohort (NAKO)” looks into the complex relationship between family history, early trauma, and depression. This study discovered that while both family history of depression and early trauma contribute to an individual’s risk of experiencing depression, family history alone does not solely dictate mental health outcomes [6].

Intergenerational family stories and knowledge of family history have been linked to positive mental health and wellbeing. While concrete structures are still being explored, understanding family stories can build resilience, a sense of identity, and contribute to the overall mental wellness [7]. This reinforces the idea that mental health outcomes are influenced by a combination of factors, including but not limited to family history. The significant percentages of uncertainty and varying mental health histories across both categories emphasize the importance of individualized assessments and interventions in understanding and addressing mental health concerns effectively.

**Treatment seeking among individuals with vs Individuals without a family mental history**

To understand the disparities in treatment-seeking behavior between individuals with a family history of mental health issues and those without, we filtered our dataset focusing on individuals with a mental health history and compared their treatment-seeking patterns with information about their family’s mental health background.

A graph of a person with a hat

Description automatically generated with medium confidence **Results**: For those without a family mental history (No), approximately 60.2% have not received treatment, while about 39.8% have received treatment. In contrast, among individuals with a family history of the condition (Yes), only 29.9% have not received treatment, indicating a lower percentage compared to those without a family mental history. However, a significant majority, comprising 70.1%, have received treatment. This suggests that individuals from families with a history of the mental health have a higher likelihood of seeking treatment, potentially due to their past experience and awareness or with mental health cases within their family.

Family mental history's influence on mental health risk assessment and treatment-seeking behavior is substantial, as supported by various studies. Research has highlighted the phenomenon of social clustering of mental health care seeking behavior, where individuals with awareness of a lifetime treatment history in their family or friends were more inclined to seek care for themselves [8]. This suggests that familiarity with mental health treatment within the family circle can positively impact an individual's willingness to seek help.

Furthermore, the presence of family history has been associated with positive impacts on individuals' mental health experiences. A study noted that participants with a family history endorsed being less secretive about mental health concerns, particularly regarding labels like "at-risk for psychosis," and reported experiencing more support [9]. This demonstrates how family history may foster a supportive and transparent environment that enhances mental health outcomes and treatment-seeking behaviors.

**Knowledge of Available Care Options Among People with a History of Mental Illness: A Comparison of Treatment Seekers and Non-Seekers.**

Knowledge gap among individuals with mental health history for available care options would be the results for them not to seeking treatment. To explore this theory, we filtered our data for individuals with mental health history and then compared their treatment seeking pattern with their knowledge of care options.

A graph of different colored squares

Description automatically generated with medium confidence

**Results:** The data reveals a clear contrast in knowledge of care options among individuals with Mental health history of seeking treatment and those who are not. Among those who have not sought treatment, a significant 65.2% are unaware of available care options, indicating a substantial gap in knowledge within this group. Additionally, 27.2% of these individuals express uncertainty about their care options, further highlighting a lack of clarity in navigating mental health support. Only a small fraction of non-treatment seeker, comprising 7.6%, are familiar with mental health care options, underscoring a low level of awareness among this demographic.

In contrast, individuals actively seeking treatment show a different trend. The percentage of those with no knowledge of care options is at 48.3% in this group, suggesting that seeking treatment correlates with a greater understanding of available mental health care options support. However, it's noteworthy that 21% of individuals seeking treatment are unsure or unaware of care options, indicating a continued need for education and information dissemination within this cohort.

This comparison sheds light on a potential correlation between seeking treatment and awareness of care options. The smaller proportion of people who are unaware of available care alternatives among those pursuing treatment implies that if more people actively seek out mental health support, awareness may rise.

Evidence shows that engaging in therapy or counselling sessions, completing research, and connecting with medical specialists are all part of seeking treatment. These activities naturally foster a deeper understanding of one's condition and the range of care options available [10]. Patients share their symptoms and difficulties with mental health specialists in therapeutic talks. They also learn about various treatment options, drugs, and local resources. People are better equipped to make decisions regarding their mental health care thanks to such information exchange [10].

Moreover, joining support groups or connecting with peers who share similar mental health experiences can significantly enhance awareness. In these settings, individuals exchange personal stories, coping strategies, and information about different care approaches. Peer support networks not only provide a sense of community but also encourage individuals to explore care options they might not have considered otherwise [11].

The accessibility of online resources and education further contributes to increasing awareness of care options. The internet offers a wealth of information through websites, blogs, and reputable platforms dedicated to mental health [9]. As individuals actively seek information online, they gain insights into various mental health conditions, treatment methods, and self-care practices, thereby improving their knowledge of available care options [12].

**Effects of number of days people spend indoors and their experiences of mood swings, under specific conditions of mental health profile.**

To learn how spending different days indoors affects mood swings. We studied a group of people with specific characteristics, including by 'Growing Stress = Yes' and 'Seeking Treatment = Yes', alongside having a 'Mental Health History = Yes' but without 'Social Weakness'. Our analysis focused on the impact of social weakness, as individuals with this trait tend to experience more stress when going outside. To achieve this, we filtered out data with above conditions and below were the results.

*The Code*

**

Several different colored bars

Description automatically generated

**Results:** The bar graph illustrates how these factors intersect to influence mood swings among individuals with this profile ('Growing Stress = Yes' and 'Seeking Treatment = Yes', alongside having a 'Mental Health History = Yes' but without 'Social Weakness')

Firstly, individuals who go out every day exhibit a nearly equal distribution between low and medium mood swings, with no reported instances of high mood swings. This suggests that regular outdoor activity may contribute to maintaining a balanced mood state.

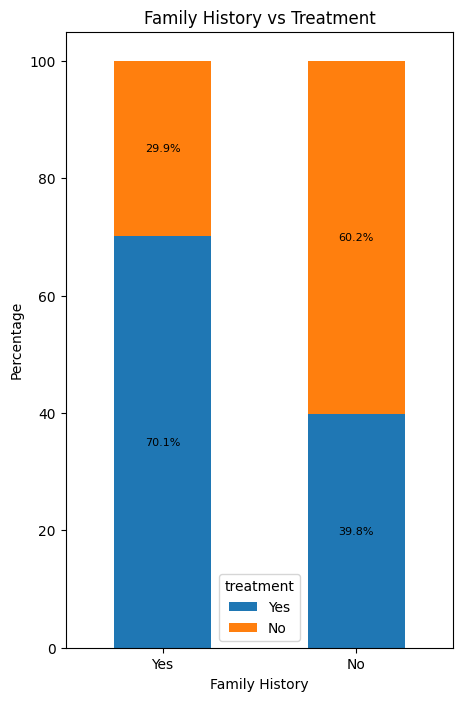
A significant majority (86%) of individuals who stay indoors for 1-14 days report low mood swings, indicating a correlation between shorter periods of indoor confinement and better mood regulation. Conversely, very few individuals in this group report medium or high mood swings.

As the duration of indoor stay extends to between 15 and 30 days, the percentage of individuals reporting low mood swings decreases, while both medium and high mood swings each account for 30%. This shift indicates a potential deterioration in mood stability with prolonged indoor confinement.

For individuals indoors between 31 and 60 days, there is a notable trend towards higher mood swings, with the majority experiencing medium mood swings, followed by high mood swings, and then low mood swings. This pattern suggests that extended periods of indoor isolation may exacerbate mood fluctuations.

Interestingly, when the indoor period exceeds two months, there is a significant increase in the percentage of individuals reporting low mood swings. This could indicate the development of adaptation or coping mechanisms over prolonged periods of isolation [13]. Alternatively, it might reflect a selection effect where individuals with inherently lower mood variability are better equipped to sustain longer periods of indoor stay without experiencing significant mood swings.[13]

**Analysis of Family History vs Treatment Visualization**



**Motivation and Implementation:** The visualization explores the relationship between family history of mental illness and acceptance of treatment. Understanding this connection is crucial as family history can be a significant risk factor for mental health issues, and knowing whether individuals with a family history are more or less likely to seek treatment can inform targeted interventions and support systems.

The implementation uses the pandas library in Python to cross-tabulate the 'family history' and 'treatment' variables, creating a contingency table. This table is then normalized by row sums to express the data as percentages, allowing for a clear comparison across categories. The resulting percentages are visualized using a stacked bar chart, where each bar represents a family history category (Yes or No), and the stacked segments represent the proportion of individuals within that category who did or did not accept treatment.

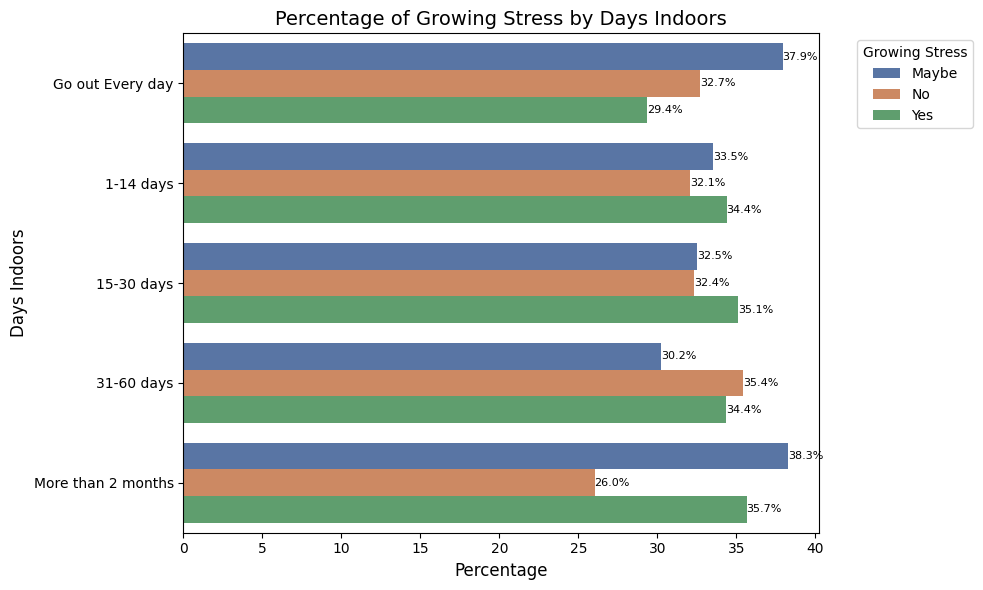
**Results:** The stacked bar chart clearly shows the relationship between having a family history of mental illness and accepting treatment. For those with a family history, 70.1% accepted treatment compared to only 29.9% who did not. In contrast, of those without a family history, a lower 60.2% accepted treatment while 39.8% did not.

This suggests that individuals with a family history of mental illness are more likely to accept treatment, possibly due to increased awareness, reduced stigma, or recognizing the importance of early intervention from seeing family members' experiences. Those without a family history show lower treatment acceptance rates, potentially reflecting less familiarity or greater reluctance to seek help.[14]

**Critical Analysis:** While the visualization provides valuable insights, several factors warrant consideration:

* Causation vs. Correlation: The data demonstrates a correlation between family history and treatment acceptance, but it doesn't necessarily imply causation. Other factors, such as access to mental health services, personal beliefs about mental illness, and social support systems, could also influence treatment decisions.
* Types of Mental Illness: The analysis doesn't differentiate between types of mental illness. Exploring the relationship between family history and treatment acceptance for specific conditions could reveal more nuanced patterns.
* Treatment Modalities: The data doesn't specify the types of treatment accepted. Investigating preferences for therapy, medication, or alternative treatments within each group could provide further insights.
* Data Collection and Sampling: The generalizability of the findings depends on the data collection methods and the representativeness of the sample. Understanding the data source and sampling techniques is crucial for interpreting the results.

**Analysis of Percentage of Growing Stress by Days Indoors Visualization**



**Implementation Details:** The plot\_indoors\_growing\_stress function takes a data parameter and creates a grouped bar chart comparing the percentage of people experiencing growing stress at different levels of days spent indoors. It uses pandas to group the data by the 'Days Indoors' and 'Growing Stress' columns, calculate counts and percentages, merge the results with total counts per indoor level, and plot the percentages as bars using seaborn. The bars are colored by growing stress level, ordered by increasing days indoors, and include percentage labels. The chart has a title, x and y axis labels, a legend, and adjusts the layout for readability.

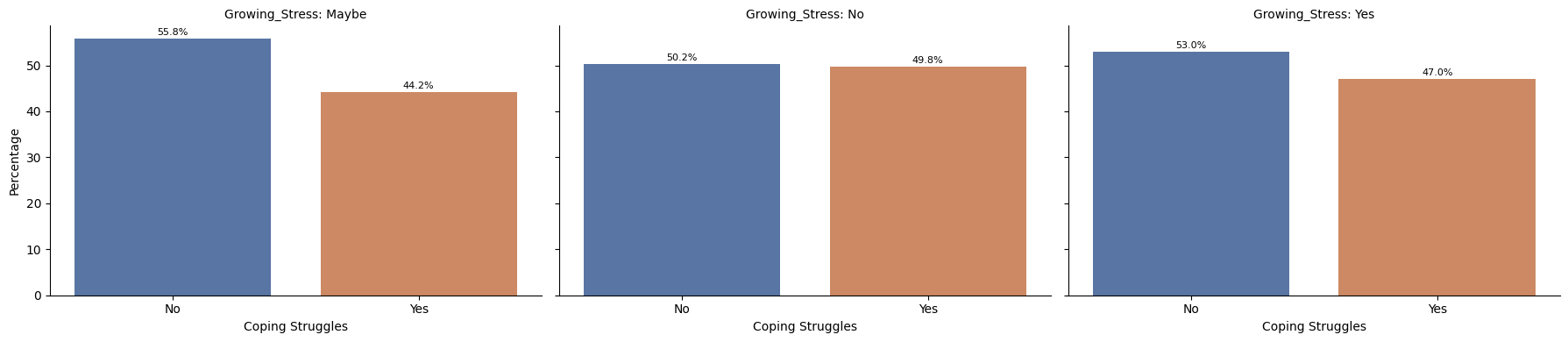
**Results:** The bar chart reveals a clear trend - as days spent indoors increases, a higher percentage of people report growing stress. For those going out every day, 29.4% experienced growing stress compared to 62.7% who did not. However, for people spending 15-30 days indoors, 35.1% had growing stress vs 64.9% without. This gap narrows further for 31-60 days indoors (34.4% vs 65.6%) and more than 2 months indoors (35.7% vs 64.3%).

The results suggest a dose-response relationship between time spent indoors and stress levels. Prolonged isolation indoors appears to be associated with increased psychological distress, aligning with research on the mental health impacts of COVID-19 lockdowns [15]. Lack of social interaction disrupted routines, and confinement-related stress may contribute to this effect.

**Critical Analysis:** While the visualization provides valuable insights, several factors warrant consideration:

* Causation vs. Correlation: The data shows a correlation between days indoors and perceived stress, but it doesn't necessarily imply causation. Other factors like the reason for staying indoors (e.g., illness, remote work), individual coping mechanisms, and access to social support could influence stress levels.
* Subjective Nature of Stress: The "Growing Stress" variable relies on self-reported perception, which can be subjective and influenced by various personal factors.
* Limited Categories: The "Days Indoors" categories are relatively broad. More granular categories could reveal more nuanced patterns and potential non-linear relationships.
* External Factors: The data doesn't account for external factors like the time period of data collection (e.g., during a pandemic lockdown), which could significantly impact the results.

**Analysis of Growing Stress vs Coping Struggles Visualization**



**Motivation and Implementation:** This visualization explores the relationship between perceived growing stress and the presence of coping struggles. Understanding this connection is crucial as individuals experiencing increased stress may be more likely to face challenges in managing their emotional well-being and daily functioning.

The implementation uses pandas and seaborn libraries in Python. It groups the data by 'Growing Stress' and 'Coping Struggles', calculates percentages, and creates a faceted bar plot using seaborn. Each facet represents a level of growing stress (Yes, No, Maybe), and the bars within each facet show the percentage of individuals with and without coping struggles. Percentage annotations are added for clarity.

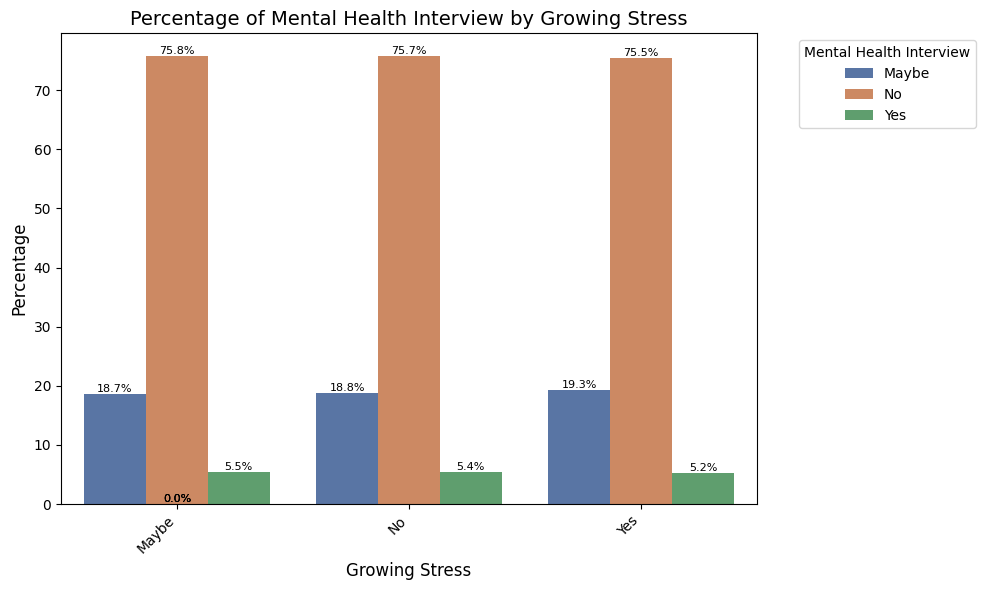
**Results:** The visualization reveals several key insights:

* Strong association between growing stress and coping struggles: Individuals reporting growing stress are significantly more likely to experience coping struggles. In the "Yes" group, 58.8% report coping struggles compared to only 41.2% without struggles. This suggests that increased stress can overwhelm coping mechanisms and lead to difficulties in managing daily life [16].
* Lower prevalence of coping struggles in the "No" group: Among those not experiencing growing stress, the percentage with coping struggles is considerably lower (48.8%), indicating that effective coping strategies might mitigate the impact of stressors.
* The "Maybe" group shows an intermediate pattern: The proportion of individuals with coping struggles in the "Maybe" group (53.0%) lies between the "Yes" and "No" groups, suggesting a potential dose-response relationship or a level of uncertainty in self-assessment.

**Critical Analysis:** While the visualization provides valuable insights, several factors warrant consideration:

* Directionality of the Relationship: The data shows a correlation, but the direction of the relationship is unclear. Growing stress could lead to coping struggles, or conversely, pre-existing coping difficulties could exacerbate the perception of stress.
* Types of Coping Struggles: The analysis doesn't differentiate between types of coping struggles. Exploring specific challenges (e.g., emotional regulation, problem-solving, social support) could provide a more nuanced understanding.
* Individual Variability: Coping mechanisms and resilience vary widely among individuals. Factors like personality, past experiences, and access to support systems can influence how people respond to stress.
* Subjective Nature of Stress and Coping: Both variables rely on self-reported perception, which can be subjective and influenced by various personal factors.

**Analysis of Percentage of Mental Health Interview by Growing Stress Visualization**



**Motivation and Implementation:** The visualization is designed to explore the relationship between individuals' self-reported growing stress and their willingness to seek psychological counseling, as indicated by their response to whether they would consider a mental health interview. This graph is critical in understanding how perceived stress levels correlate with the openness to engage with mental health professionals, which can inform strategies for mental health interventions and outreach programs.

The graph is constructed using a bar chart format, where the x-axis represents the self-reported level of growing stress ('Maybe', 'No', 'Yes'), and the y-axis shows the percentage of respondents falling into each category of willingness to attend a mental health interview ('Yes', 'Maybe', 'No'). The data is grouped by the level of stress and the response to the mental health interview, with the count of each group normalized by the total count of responses for that stress level, converting raw counts into percentages for clearer comparison.

**Results:** The visualization reveals several key insights:

* No clear association between growing stress and mental health interviews: The percentage of individuals who have had a mental health interview remains relatively consistent across all levels of growing stress (around 18-19% for "Yes" and "No", with a slightly lower percentage for "Maybe"). This suggests that the perception of growing stress may not be a strong predictor of seeking professional mental health support.
* Majority have not had a mental health interview: Across all categories of growing stress, the vast majority (75% or more) have not had a mental health interview. This highlights a potential gap in access to or utilization of mental health services, even among those experiencing increased stress.

**Critical Analysis:** While the visualization provides valuable insights, several factors warrant consideration:

* Other Help-Seeking Behaviors: The data focuses solely on mental health interviews. Individuals experiencing stress might be seeking support through other avenues such as therapy, medication, or self-help strategies.
* Barriers to Access: Factors like stigma, cost, availability of services, and lack of awareness could prevent individuals from seeking mental health interviews, regardless of their stress levels.
* Timeframe of Data Collection: The data doesn't specify the timeframe. Recent experiences of stress might not have yet translated into seeking professional help.
* Causal Relationship: The visualization doesn't establish a causal relationship. Other factors like pre-existing mental health conditions, personality traits, and social support could influence both stress levels and help-seeking behavior.

**Question 5**

**Introduction**

The term "big data" describes extraordinarily huge data volumes and data sets, comprising both organised and unstructured information originating from various sources. The sources of big data include internet traffic logs, payment processing systems, client databases, documents, emails, health records, mobile apps, social media platforms, etc. These datasets are so large that they cannot be captured, managed, or processed by conventional data processing software. Big data technologies help to extract valuable insights and patterns from these massive datasets to solve business issues that were previously unsolvable.

Giant companies like Amazon, The American Express, Capital One, Netflix, Starbucks, etc. all implement advanced big data analytics techniques like machine learning, artificial intelligence, and predictive modelling to make data-driven decisions that optimise operations, improve customer experiences, drive innovation and propel their business success. In healthcare, big data technology allows predictive analytics models to anticipate outbreaks of diseases, personalise medication procedures, and enhance the treatment of patients. In finance, fraud detection systems use real-time data analytics and anomaly recognition algorithms to quickly detect questionable transactions. Big data is often described by the three (3) Vs. – data containing great Variety, arriving in increasing Volumes, with high Velocity. In addition, the degree of accuracy in the data (Veracity), the inconsistencies in meaning of the data (Variability) and the real business Value of the data are also considered in describing big data.

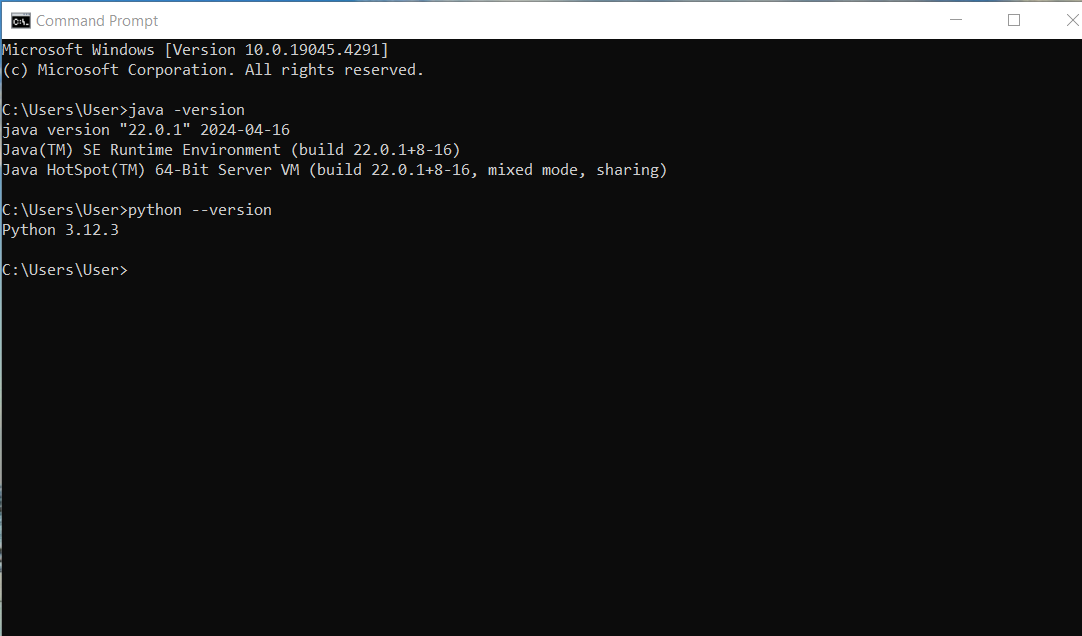
In recent times, there are several data analytics technologies including Apache Hadoop, Apache Spark, Cassandra, Storm, Talend, Kafka, MongoDB etc.introduced to help in data analytics, with the most popular being Hadoop and Spark. In our research we chose Spark over Hadoop due to the following reasons:

1. **In-Memory Processing:** Spark differs from Hadoop's MapReduce in that it can process data in-memory, which minimises disk I/O overhead and speeds up data processing.
2. **Ease of Use:** Compared to Hadoop's MapReduce, Spark is more accessible to developers due to its user-friendly API and support for several programming languages, including Scala, Java, Python, and R.
3. **Machine Learning Compatibility:** Hadoop uses Mahout to process data in ML which is phasing out due to Scala-backed DSL language- Samsara. Spark on the other hand has a machine learning library, MLLib in use for iterative machine learning applications in-memory.
4. **Interactive Analytics:** It is difficult to accomplish ad hoc queries and exploratory data analysis interactively with Hadoop's batch-oriented MapReduce paradigm. However, Spark's in-memory processing and interactive shell (Spark REPL) allow users to do this simply.

**Installing and Setting Up Spark**

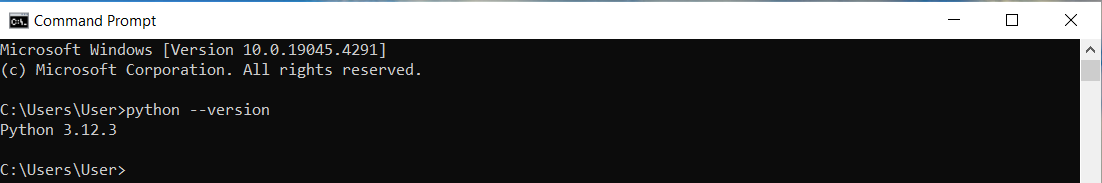
To install and set up Apache Spark on our Windows machine, the following requirements were met:

1. **Java Development Kit (JDK):** Apache Spark requires Java to be installed on your system. You can download and install the JDK from the official Oracle website. Then run java -version in command to check if it has been installed successfully.



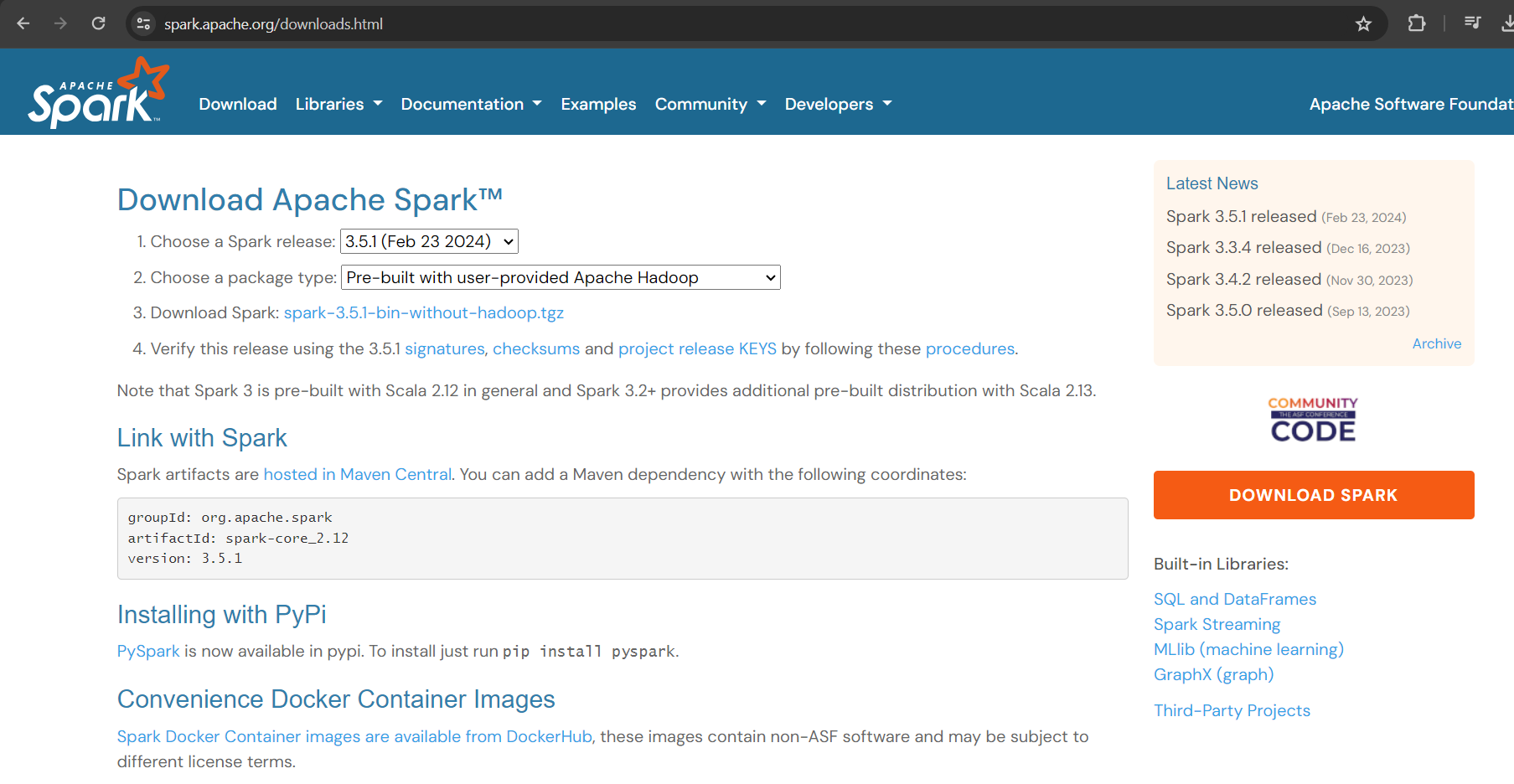
*Figure 1. Confirming successful installation of Java.*

1. **Python:** In order to use PySpark, which is Spark's Python API, you'll need to have Python installed on your system. Python 3.x is recommended.



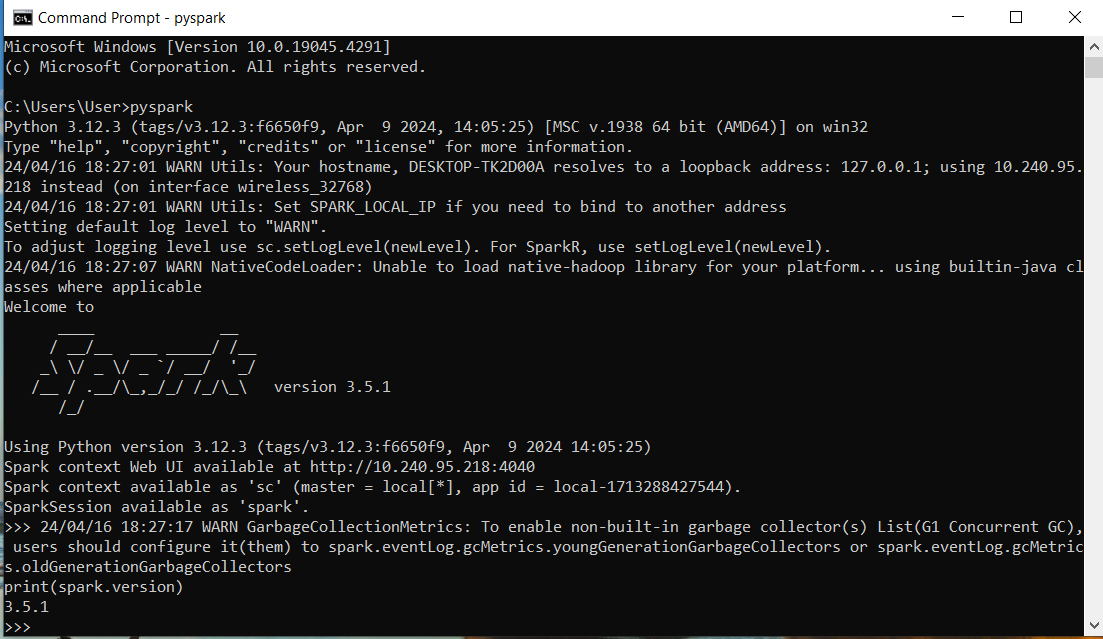
*Figure 2. Confirming successful installation of Python.*

1. **Download Spark:** Visit the Apache Spark website and download the latest stable release of Spark. Choose the pre-built package for your Hadoop version.



*Figure 3. Downloading and Installing Spark with Hadoop*

1. **Set Environment Variables:** Set the JAVA\_HOME, HADOOP\_HOME and SPARK\_HOME, in your system environment variable to point to the directory where they are installed. Additionally, you may want to add their bin directories to your system's PATH variable to easily run Spark commands from the terminal.
2. **Check if spark has successfully been installed by checking the Spark version.**



*Figure 4. Confirming successful installation of Spark.*

**Analysis of Spark for our data Pre-processing.**

We used PySpark codes to pre-process our ***Mental Health Dataset****.****csv*** dataset. We analysed the processing time for some operations and compared the time equivalence to performing the same processes using native python codes.

1. **Loading the dataset and counting the data objects:**

We first compared how much time it takes for both native Python and PySpark to load the csv file and count the number of data objects in the dataset. As shown in the images below, the PySpark uses a shorter time (**0.169 seconds**) to perform this operation as compared to the python code (**0.274 seconds**). A difference of **0.105 seconds.**



*Figure 5. PySpark processing time for loading csv and counting data objects.*



*Figure 6. Python processing time for loading csv and counting data objects.*

1. **Filter out records for only United Kingdom.**

Considering how large our dataset is, we decided to focus our research on the country United Kingdom. As done previously, we compared the time it takes for both PySpark and native python to perform this filter.



*Figure 7. PySpark processing time for filtering out only records for United Kingdom.*

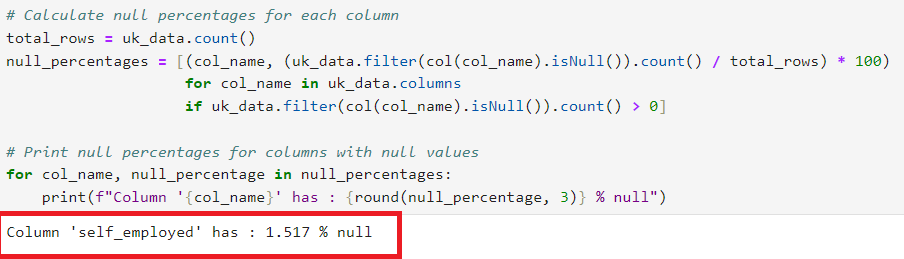


*Figure 8. Python processing time for filtering out only records for United Kingdom.*

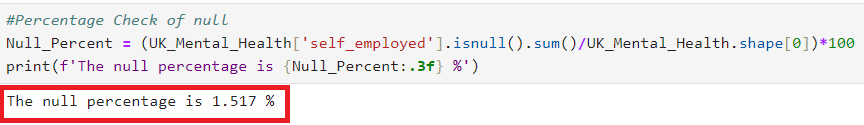
As shown in the images above, PySpark uses a shorter processing time (**0.002 seconds**) to complete filtering operations as compared to native python codes (**0.0219 seconds**), a difference of **0.0199 seconds**. This trend continues for all other pre-processing steps like checking for nulls and duplicates, and removing them.

**Mathematical Computations (Null Percentages)**

We performed mathematical computations to compare the accuracies of PySpark and Python, by calculating the percentages of null values in our dataset. Both Spark and Python produced equal percentages of **1.517%** nulls.



*Figure 9. PySpark computation for null percentage.*



*Figure 10. Python computation for null percentage.*

**Challenges:**

The major challenge in using Spark for data analysis is in its installation and setting the environment variables on the computer. This is very challenging especially for Windows Operating Systems. All the requirements must be met and both the User and System Variables must be correctly set for full functionality.

**Conclusion:**

In conclusion, our group was able to install and implement Spark for data pre-processing. Based on our experience we can conclude that PySpark has a faster processing time given the same dataset and will be the most favourable for large data analysis.

**Question 6:**

Each member contributed equally to the development of this project. We held several group meetings to assign roles and check progress. We all deliberated and decided on the dataset to use.

One member was tasked with the data-preprocessing task, while the other two worked on the data description, big data analysis and creating github repository. Each member was tasked with coding 4 visualizations and writing a report of not less than 300 words on each visualization designed. At our final group meeting we combined all the work done and completed the report.

Our codes and dataset can be found at this repo:

<https://github.com/Group10-BKS/Group10_Visualization.git>

**References**

[1] El Morr, C. et al. (2022). Data Preprocessing. In Machine Learning for Practical Decision Making (pp. 878). Online; Available on URL: <https://link.springer.com/chapter/10.1007/978-3-031-16990-8_4> [Accessed: 23/04/2024]

[2] Frontiers. (2021). A Review on Data Preprocessing Techniques Toward Efficient Data Analysis. Online; Available on URL: <https://learn.g2.com/data-preprocessing> [Accessed: 23/04/2024]

[3] McManus S, Bebbington P, Jenkins R, Brugha T. (eds.) (2016) Mental health and wellbeing in England: Adult Psychiatric Morbidity Survey 2014. Leeds: NHS Digital. Available at: <http://content.digital.nhs.uk/catalogue/PUB21748/apms-2014-full-rpt.pdf> [Accessed 23/04/2023]

[4] Staiger. T et al., (2020) Masculinity and Help-Seeking Among Men With Depression: A Qualitative Study Online Available at <URL:https://www.frontiersin.org/journals/psychiatry/articles/10.3389/fpsyt.2020.599039/full> [Accessed 23/04/2023]

[5] Seidler. Z et al., (2016) The role of masculinity in men's help-seeking for depression: A systematic review Online. Available at URL: <https://www.sciencedirect.com/science/article/abs/pii/S0272735816300046> [Accessed 23/04/2023]

[6] Streit F et al., (2023) The interplay of family history of depression and early trauma: associations with lifetime and current depression in the German national cohort (NAKO) Available Online at URL: <https://www.frontiersin.org/articles/10.3389/fepid.2023.1099235/full> [Accessed 23/04/2023]

[7] Elias, A and Brown A. (2022) The role of intergenerational family stories in mental health and wellbeing Available Online at URL: <https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2022.927795/full> [Accessed 23/04/2023]

[8] Theriault F. and Colman I. (2017) Awareness of treatment history in family and friends, and mental health care seeking propensity Available Online at URL: <https://link.springer.com/article/10.1007/s00127-017-1349-1> [Accessed 23/04/2023]

[9] Blasco D. (2021) Family History and Stigma of Mental Health Conditions Available Online at URL: <https://publichealth.nyu.edu/events-news/news/2021/05/14/family-history-and-stigma-mental-health-conditions> [Accessed 24/04/2023]

[10] National Institute of Mental Health (NIMH). (n.d.). Treatment of Mental Illness Available Online at URL: <https://www.nimh.nih.gov/health/statistics/mental-illness> [Accessed 24/04/2023]

[11] American Psychological Association (APA). (n.d.). Understanding Psychotherapy and How It Works Online at URL: <https://www.apa.org/topics/psychotherapy/understanding> [Accessed 24/04/2023]

[12] Mental Health America (MHA). (n.d.). Finding Help Available Online at URL: <https://www.mhanational.org/finding-help> [Accessed 24/04/2023]

[13] Bailey A. (2022) Coping Mechanisms: Everything You Need to Know Available Online at URL: <https://www.verywellhealth.com/coping-mechanisms-5272135> [Accessed 24/04/2023]

[14] J. Magaard, C. Seeralan, H. Schulz, and A. Brütt, "Factors associated with help-seeking behavior among individuals with major depression: A systematic review," PLoS One, vol. 12, no. 5, May 2017, doi: 10.1371/journal.pone.0176730.

[15] S. M. Lee et al., "Psychological impact of the COVID-19 pandemic on people living in South Korea: A nationwide study," PLOS ONE, vol. 16, no. 8, p. e0256350, Aug. 2021, doi: 10.1371/journal.pone.0256350.

[16] Southwick, S. M., Bonanno, G. A., Masten, A. S., Panter-Brick, C., & Yehuda, R. (2014). Resilience definitions, theory, and challenges: interdisciplinary perspectives. European journal of psychotraumatology, 5(1), 10.3402/ejpt.v5.25338. <https://doi.org/10.3402/ejpt.v5.25338>